Statistica Sinica **30** (2020), 741-762
doi:https://doi.org/10.5705/ss.202017.0468

OPTIMAL RATES AND TRADE-OFFS IN MULTIPLE TESTING

Maxim Rabinovich, Aaditya Ramdas,
Michael I. Jordan and Martin J. Wainwright

University of California, Berkeley

Abstract: Multiple hypothesis testing is a central topic in statistics. However, despite abundant research on the false discovery rate (FDR) and the corresponding Type-II error concept known as the false nondiscovery rate (FNR), we do not yet have a fine-grained understanding of the fundamental limits of multiple testing. The main contribution of this study is to derive a precise nonasymptotic trade-off between the FNR and FDR for a variant of the generalized Gaussian sequence model. Our approach is flexible enough to permit analyses of settings where the problem parameters vary with the number of hypotheses n, including various sparse and dense regimes (with $o(n)$ and $\mathcal{O}(n)$ signals). Moreover, we prove that the Benjamini–Hochberg and Barber–Candès algorithms are both rate-optimal up to constants across these regimes.

Key words and phrases: Minimax lower bounds, multiple testing.

1. Introduction

The problem of multiple comparisons has been a central topic in statistics ever since Tukey's influential book (Tukey (1953)). In broad terms, suppose we observe a sequence of n independent random variables, $X_1, \ldots, X_n$, of which some unknown subset are drawn from a null distribution, corresponding to the absence of a signal or effect. The remainder are drawn from a non-null distribution, corresponding to signals or effects. Within this framework, we can pose three problems of increasing hardness: the *detection* problem, testing whether or not there is at least one signal; the *localization* problem, identifying the positions of the nulls and signals; and the *estimation* problem, which returns estimates of the means and/or distributions of the observations. Note that these problems form a hierarchy of difficulty: identifying the signals implies that we know whether at least one exists, and estimating each mean implies we know which are zero and which are not. This study focuses on the problem of localization.

There are a variety of ways of measuring type-I errors for the localization

problem. For example, the *family-wise error rate* measures the probability of incorrectly rejecting at least one null, and the *false discovery rate* (FDR) is the expected ratio of incorrect rejections to total rejections. An extensive body of literature has developed around both metrics, resulting in algorithms geared towards controlling one or the other. Our focus is the FDR metric. Although the FDR has been widely studied, relatively little is known about the behavior of existing algorithms in terms of the corresponding Type-II error concept, namely the *false nondiscovery rate* (FNR). Indeed, it is only recently that Arias-Castro and Chen (2017), working with a version of the sparse generalized Gaussian sequence model, established asymptotic consistency for the FDR − FNR localization problem. Informally, in this framework, we receive n independent observations, $X_1, \dots, X_n$, of which $n^{1-\beta_n}$ are non-nulls. The remainder are nulls. The $n-n^{1-\beta_n}$ null variables are drawn from a centered distribution with tails decaying as $\exp(-(|x|^\gamma)/\gamma)$, whereas the non-nulls are drawn from the same distribution, shifted by $(\gamma r_n \log n)^{1/\gamma}$. Using this notation, Arias-Castro and Chen (2017) considered the setting with fixed problem parameters $r_n = r$ and $\beta_n = \beta$, and showed that when $r < \beta < 1$, all procedures must have risk FDR + FNR $\to 1$. They also showed that in the achievable regime $r > \beta > 0$, the Benjamini–Hochberg (BH) procedure is consistent, meaning that FDR + FNR $\to 0$. Finally, they proposed a new "distribution-free" method inspired by the knockoff procedure of Foygel Barber and Candès (2015), and showed that the resulting procedure is also consistent in the achievable regime.

These existing consistency results are asymptotic. However, to date, no studies have examined the important nonasymptotic questions that are of interest in comparing procedures. For instance, for a given FDR level, what is the best achievable FNR? What is the best nonasymptotic behavior of the risk FDR + FNR attainable in finite samples? In addition, and perhaps most importantly, nonasymptotic questions on whether procedures such as BC and BH are *rate-optimal* for the FDR+FNR risk, remain unanswered. The main contributions of this study are to develop techniques to address such questions, and then to use these techniques to solve the problems in the context of the sparse generalized Gaussians model.

Specifically, we establish the trade-off between the FDR and FNR in finite samples (and, hence, also asymptotically), and we use the trade-off to determine

We follow Arias-Castro and Chen (2017) in defining the FNR as the ratio of undiscovered to total non-nulls, which differs from the definition of Genovese and Wasserman (Genovese and Wasserman (2002)).

the best attainable rate for the FDR+FNR risk. Our theory is sufficiently general to accommodate sequences of parameters (r_n, β_n), and enabling it to reveal new phenomena that arise when $r_n - \beta_n = o(1)$. For a fixed pair of parameters (r, β) in the achievable regime $r > \beta$, our theory leads to an explicit expression for the optimal rate at which FDR+FNR can decay. In particular, defining the γ-"distance" $D_\gamma(a, b) := \left|a^{1/\gamma} - b^{1/\gamma}\right|^\gamma$ between pairs of positive numbers, we show that the equation

$$\kappa = D_\gamma(\beta + \kappa, r)$$

has a unique solution κ_*. Moreover, the combined risk of any threshold-based multiple testing procedure $\mathcal{I}$ is lower bounded as $\mathcal{R}_n(\mathcal{I}) \gtrsim n^{-\kappa_*}$. Furthermore, using a direct analysis, we prove that the Benjamini–Hochberg (BH) and the Barber–Candès (BC) algorithms both attain this optimal rate.

At the core of our analysis is a simple comparison principle. The flexibility of the resulting proof strategy allows us to identify a new critical regime in which $r_n - \beta_n = o(1)$. However, in this regime, the problem is infeasible, which means that if the FDR is driven to zero, then the FNR must remain bounded away from zero. Moreover, we are able to study challenging settings in which the fraction of signals is a constant $\pi_1 \in (0, 1)$ and not asymptotically vanishing. This corresponds to the setting $\beta_n = \log(1/\pi_1)/\log n$, such that $\beta_n \to 0$. Perhaps surprisingly, even in these regimes, the BH and BC algorithms continue to be optimal, although the best rate can weaken from polynomial to subpolynomial in the number of hypotheses n.

1.1. Related work

As described above, our work provides a nonasymptotic generalization of the recent work by Arias-Castro and Chen (2017) on asymptotic consistency in localization, using FDR + FNR as the notion of risk. Note that this notion of risk is distinct from the asymptotic Bayes optimality under sparsity (ABOS) studied by Bogdan et al. (2011) for Gaussian sequences, and more recently, by Neuvial and Roquain (2012) for binary classification with extreme class imbalance. The ABOS results concern a risk derived from the probability of incorrectly rejecting a single null sample (false positive; FP) and the probability of incorrectly failing to reject a single non-null sample (false negative; FN). Specifically, we have $\mathcal{R}_n^{\mathrm{ABOS}} = w_1 \cdot \mathrm{FP} + w_2 \cdot \mathrm{FN}$ for some pair of positive weights (w_1, w_2), which need not be equal. Because this risk is based on the error probability for a single sample, it is much closer to a misclassification risk or a single-testing risk than

it is to the ratio-based FDR + FNR risk examined here.

Using our notation, the work of Neuvial and Roquain (2012) can be understood as focusing on the setting $r = \beta$, a regime the authors refer to as the "verge of detectability." Furthermore, their performance metric is given by the Bayes classification risk, rather than the combination of FDR and FNR studied here. In comparison, our results provide additional insight into models that are close to the verge of detectability. Even when $\beta_n = \beta$ is fixed, we can provide quantitative lower and upper bounds on the FDR/FNR ratio as $r_n \to \beta$ from above. Moreover, these bounds depend on how quickly r_n approaches β. A further transition in rates occurs when $r = \beta$ exactly, for all n; however, we do not explore this case in depth. We suspect that our methods may offer sufficient precision to answer the nonasymptotic minimaxity questions posed by Neuvial and Roquain (2012) on whether any threshold-based procedure can match the Bayes optimal classification error rate, up to an additive error $\ll 1/\log n$.

For the special case of $\gamma = 2$, Ji and Jin (2012) and Ji and Zhao (2014) prove bounds for localization that are closely related to, but distinct from, our bounds on the overall risk. Both deal with a sparse high-dimensional regression. The former work proposes a new method for variable selection, called UPS, that has advantages over the lasso and subset selection methods in certain settings. The latter builds on the first to prove upper and lower bounds for multiple testing, using the so-called mFNR and mFDR. These metrics replace the expected ratio in the definitions of FDR and FNR (see definition (2.3) below) with a ratio of expectations—a modification that should lead to qualitatively similar behavior as n becomes large. The resulting bounds in both works can be used to recover our bounds up to polylogarithmic factors in the special case where $\gamma = 2$. The main advantage of their work, relative to ours, lies in how they handle the dependence between the p-values. Unlike our work, however, they do not establish the tradeoff between the FDR and FNR when both quantities can decay to zero at different rates; in addition, as mentioned, they only consider the case of $\gamma = 2$. Nor do they consider regimes where the sparsity and signal strength vary with n. Our results can handle this more general setting, which encompasses dense regimes with qualitatively different behavior from the more commonly investigated sparse one.

The above line of work is complementary to the well-known asymptotic results of Donoho and Jin (2004, 2015) on phase transitions in detectability using Tukey's higher-criticism statistic, which employs standard type-I and type-II errors for testing of the single global null hypothesis. Note that Donoho and Jin

use the generalized Gaussian assumption directly on the PDFs, whereas our assumption (2.5) is on the survival function. Just as in Arias-Castro and Chen (2017), Donoho and Jin consider the asymptotic setting with $r_n = r$ and $\beta_n = \beta$, which they sometimes call the RW (rare and weak) model. We are not aware of any nonasymptotic results for detection that are similar to those proposed here for localization.

Our study also complements work on estimation, the most notable result being the asymptotic minimax optimality of BH-derived thresholding for denoising an approximately sparse high-dimensional vector (Abramovich et al. (2006); Donoho and Jin (2006)). The relevance of our results to the minimaxity of BH for approximately sparse denoising problems lies primarily in the use of deterministic thresholds as a useful proxy for BH, as well as other procedures that determine their threshold in a manner that has complex dependence on the input data (Donoho and Jin (2006)). Unlike the strategy of Donoho and Jin (2006), which depends on establishing the concentration of the empirical threshold around the population-level value, we use a more flexible comparison principle. Deterministic approximations to optimal FDR thresholds are also studied by Chi (2007) and Genovese, Roeder and Wasserman (2006). Other related papers are discussed in Section 5, when discussing directions for future work.

The remainder of this paper is organized as follows. In Section 2, we provide some background on the multiple testing problem, as well as the particular model we consider. In Section 3, we provide an overview of our main results: the optimal trade-offs between the FDR and FNR, which imply lower bounds on the FDR + FNR risk, and optimality guarantees for the BH and BC algorithms. In Section 4, we prove our main results. We first focus on the lower bounds, and then provide matching upper bounds for the well-known and popular BH and BC algorithms for multiple testing with FDR control. The proofs of some technical lemmas are given in the online Supplementary Material.

2. Problem Formulation

In this section, we provide background and a precise formulation of the problem under study.

2.1. Multiple testing and the FDR

Suppose that we observe a real-valued sequence $X_1^n := \{X_1, \dots, X_n\}$ of n independent random variables. When the null hypothesis is true, X_i is assumed

to have a zero mean; otherwise, it is assumed that the mean of X_i is equal to some unknown number $\mu_n > 0$. The binary labels $\{H_1, \ldots, H_n\}$ indicate whether the null hypothesis holds for each observation; the setting $H_i = 0$ indicates that the null hypothesis holds. We define

$$\mathcal{H}_0 := \{i \in [n] \mid H_i = 0\}, \quad \text{and} \quad \mathcal{H}_1 := \{i \in [n] \mid H_i = 1\}, \tag{2.1}$$

corresponding to the *nulls* and *signals*, respectively. Our task is to identify a subset of indices that contains as many signals as possible, while not containing too many nulls.

More formally, a testing rule $\mathcal{I} : \mathbb{R}^n \to 2^{[n]}$ is a measurable mapping of the observation sequence X_1^n to a set $\mathcal{I}(X_1^n) \subseteq [n]$ of *discoveries*, where the subset $\mathcal{I}(X_1^n)$ contains those indices for which the procedure rejects the null hypothesis. There is no single unique measure of performance for a testing rule for the localization problem. We employ the FDR and FNR for this purpose. These can be viewed as generalizations of the type-I and type-II errors for single hypothesis testing.

We begin by defining the false discovery proportion (FDP) and the false nondiscovery proportion (FNP), respectively, as

$$\mathrm{FDP}_n(\mathcal{I}) := \frac{\mathrm{card}(\mathcal{I}(X_1^n) \cap \mathcal{H}_0)}{\mathrm{card}(\mathcal{I}(X_1^n)) \vee 1}, \text{and } \mathrm{FNP}_n(\mathcal{I}) := \frac{\mathrm{card}(\mathcal{I}(X_1^n)^c \cap \mathcal{H}_1)}{\mathrm{card}(\mathcal{H}_1)}. \tag{2.2}$$

Because the output $\mathcal{I}(X_1^n)$ of the testing procedure is random, both quantities are random variables. The FDR and FNR are given by taking the expectations of these random quantities; that is,

$$\mathrm{FDR}_n(\mathcal{I}) := \mathbb{E}\left[\frac{\mathrm{card}(\mathcal{I}(X_1^n) \cap \mathcal{H}_0)}{\mathrm{card}(\mathcal{I}(X_1^n)) \vee 1}\right], \text{and } \mathrm{FNR}_n(\mathcal{I}) := \mathbb{E}\left[\frac{\mathrm{card}(\mathcal{I}(X_1^n)^c \cap \mathcal{H}_1)}{\mathrm{card}(\mathcal{H}_1)}\right], \tag{2.3}$$

where the expectation is taken over the random samples X_1^n.

Note that our definitions of the FNP and FNR, which follow those of Arias-Castro and Chen (2017), differ from an alternative definition of the FNR_{alt}, where the denominator is set to the number of nonrejections. In general, however, the number of nonrejections will be close to n for any procedure with low FDR. Thus, in the sparse regime, the FNR_{alt} would trivially go to zero for any procedure that controls the FDR at any level strictly below one. Our definition is therefore better suited to studying transitions in difficulty in the multiple testing problem.

We measure the overall performance of a procedure in terms of its *combined risk*,

$$\mathcal{R}_n(\mathcal{I}) := \mathrm{FDR}_n(\mathcal{I}) + \mathrm{FNR}_n(\mathcal{I}). \tag{2.4}$$

Finally, when the testing rule $\mathcal{I}$ is clear from the context, we frequently omit an explicit reference to the dependence on the testing rule.

2.2. Tail generalized Gaussian model

In this paper, we describe the distribution of the observations for both nulls and non-nulls in terms of a *tail generalized Gaussian model*. Our model is a variant of the generalized Gaussian sequence model, studied in Arias-Castro and Chen (2017) and Donoho and Jin (2004); the only difference is that whereas a γ-generalized Gaussian has a density proportional to $\exp(-(|x|^\gamma)/\gamma)$, we focus on distributions with tails proportional to $\exp(-(|x|^\gamma)/\gamma)$. This alteration is in line with the asymptotically generalized Gaussian (AGG) distributions studied by Arias-Castro and Chen (2017), with the important caveat that our assumptions are imposed in a nonasymptotic fashion.

For a given degree $\gamma \geq 1$, a γ-tail generalized Gaussian random variable with mean zero, written as $G \sim \mathrm{tGG}_\gamma(0)$, has a survival function $\Psi(t) := \mathbb{P}\big(G \geq t\big)$ that satisfies the bounds

$$\frac{e^{(-|t|^\gamma)/\gamma}}{Z_\ell} \;\leq\; \min\{\Psi(t), 1-\Psi(t)\} \;\leq\; \frac{e^{(-|t|^\gamma)/\gamma}}{Z_u}, \quad t \in \mathbb{R}, \tag{2.5}$$

for some constants $Z_\ell > Z_u > 0$. (Note that $t \mapsto \Psi(t)$ is a decreasing function, and becomes smaller than $1 - \Psi(t)$ at the origin.) As a concrete example, a γ-tail generalized Gaussian with $Z_\ell = Z_u = 1$ can be generated by sampling a standard exponential random variable E and a Rademacher random variable ε, and then letting $G = \varepsilon\big(\gamma E\big)^{1/\gamma}$. We use the terminology "tail generalized Gaussian" because the survival function of a two-tail Gaussian random variable is of the order of $\exp(-|x|^2/2)$, whereas that of a Gaussian is of the order of $(1/\mathrm{poly}(x))\exp(-x^2/2)$. In particular, this observation implies that a tGG_2 random variable has tails that are equivalent to those of a Gaussian in terms of their exponential decay rates.

In terms of this notation, we assume that each observation X_i is distributed as

$$X_i \sim \begin{cases} \mathrm{tGG}_\gamma(0) & \text{if } i \in \mathcal{H}_0, \\ \mathrm{tGG}_\gamma(0) + \mu_n & \text{if } i \in \mathcal{H}_1, \end{cases} \tag{2.6}$$

where our notation reflects the fact that the mean shift μ_n is permitted to vary with the number of observations n. See Section 3.1 for further discussion of the scaling of the mean shift.

2.3. Threshold-based procedures

Following prior work (Arias-Castro and Chen (2017); Donoho and Jin (2004)), we restrict our attention to testing procedures of the form

$$\mathcal{I}(X_1^n) = \big\{i \in [n] \mid X_i \geq T_n(X_1^n)\big\}, \tag{2.7}$$

where $T_n(X_1^n) \in \mathbb{R}_+$ is a data-dependent threshold. We refer to such methods as *threshold-based procedures*. The BH and BC procedures both belong to this class. Moreover, from an intuitive standpoint, the observations are exchangeable in the absence of prior information, and we are testing between a single unimodal null distribution and a single positive shift of that distribution. In this setting, it is difficult to conceive of reasonable procedures that would reject the hypothesis corresponding to one observation, while rejecting a hypothesis with a smaller observation value.

In particular, as part of our argument, it will be important to analyze the performance metrics associated with rules of the form

$$\mathcal{I}_t(X_1^n) = \big\{i \in [n] \mid X_i \geq t\big\}, \tag{2.8}$$

where $t > 0$ is a prespecified (fixed, nonrandom) threshold. In this case, we adopt the notation $\mathrm{FDR}_n(t)$, $\mathrm{FNR}_n(t)$, and $\mathcal{R}_n(t)$ to denote the metrics associated with the rule $X_1^n \mapsto \mathcal{I}_t(X_1^n)$.

2.4. The BH and BC procedures

Arguably the most popular threshold-based procedure that provably controls FDR at a user-specified level q_n is the BH procedure. More recently, Arias-Castro and Chen (2017) proposed a method that we refer to as the BC procedure. Both algorithms are based on estimating the FDP_n that would be incurred at a range of possible thresholds, and then choosing the largest one possible (maximizing discoveries), while satisfying an upper bound linked to q_n (controlling FDR_n). Furthermore, they both only consider thresholds that coincide with one of the values X_1^n, which we denote as the set $\mathcal{X}_n = \big\{X_1, \ldots, X_n\big\}$. The data-dependent threshold for both can be written as

$$t_n\big(X_1, \ldots, X_n\big) = \min\big\{t \in \mathcal{X}_n \colon \widehat{\mathrm{FDP}}_n\big(t\big) \leq q_n\big\}. \tag{2.9}$$

The two algorithms differ in the estimator $\widehat{\mathrm{FDP}}_n\big(t\big)$ they use. The BH procedure assumes access to the true null distribution through its survival function Ψ and sets

$$\widehat{\mathrm{FDP}}_n^{\mathrm{BH}}(t) = \frac{\Psi(t)}{\#(X_i \geq t)/n}, \quad \text{for } t \in \mathcal{X}_n. \tag{2.10}$$

The BC procedure instead estimates the survival function $\Psi(t)$ from the data and, therefore, does not need to know the null distribution. This approach is viable when $\#(X_i \leq -t)/n$ is a good proxy for $\Psi(t)$, which our upper and lower tail bounds guarantee; more typically, the BC procedure is applicable when the null distribution is (nearly) symmetric, and the signals are shifted by a positive amount (as they are in our case). Then, the BC estimator is given by

$$\widehat{\mathrm{FDP}}_n^{\mathrm{BC}}(t) = \frac{[\#(X_i \leq -t) + 1]/n}{\#(X_i \geq t)/n}, \quad \text{for } t \in \mathcal{X}_n. \tag{2.11}$$

With these definitions in place, we are now ready to describe our main results.

3. Main Results

We now state our main results and examine their consequences. Our first main result (Theorem 1) characterizes the optimal trade-off between the FDR and FNR for any testing procedure. By optimizing this trade-off, we obtain a lower bound on the combined FDR and FNR of any testing procedure (Corollary 1). Our second main result (Theorem 2), shows that the BH procedure achieves the optimal FDR-FNR trade-off up to constants, and that the BC procedure almost achieves optimality. In particular, our result implies that, with the proper choice of target FDR, the BH and BC procedures can both achieve the optimal combined FDR-FNR rate (Corollary 2).

3.1. Scaling of sparsity and mean shifts

We study a sparse instance of the multiple testing problem, in which the number of signals is assumed to be small relative to the total number of hypotheses. In particular, motivated by related works on multiple hypothesis testing (Arias-Castro and Chen (2017); Donoho and Jin (2004, 2015); Jin and Ke (2016)), we assume that the number of signals scales as

$$\mathrm{card}(\mathcal{H}_1) = m_n = n^{1-\beta_n} \quad \text{for some } \beta_n \in (0,1). \tag{3.1}$$

Note that, to the best of our knowledge, all previous results in the literature assume that $\beta_n = \beta$ is actually independent of n. In this case, the sparsity assumption (3.1) implies that all but a polynomially vanishing fraction of the hypotheses are null. In contrast, as indicated by our choice of notation, the setup in this study allows for a sequence of parameters β_n that can vary with the

number of hypotheses n. As a result, our framework is flexible enough to handle relatively dense regimes (e.g., those with $n/\log n$ or even $\mathcal{O}(n)$ signals).

The non-null hypotheses are distinguished by a positively shifted mean $\mu_n > 0$. It is natural to parameterize this mean shift in terms of a quantity $r_n > 0$ via the relation

$$\mu_n = \big(\gamma r_n \log n\big)^{1/\gamma}. \tag{3.2}$$

As shown by Arias-Castro and Chen (2017), when the pair (β, r) are fixed such that $r < \beta$, the problem is asymptotically infeasible, meaning that there is no procedure such that $\mathcal{R}_n(\mathcal{I}) \to 0$ as $n \to \infty$. Accordingly, we focus on sequences (β_n, r_n), for which $r_n > \beta_n$. Furthermore, even though the asymptotic consistency boundary of $r < \beta$ versus $r > \beta$ is apparently independent of γ, we find that the rate at which the risk decays to zero is determined jointly by r, β, and γ.

3.2. Lower bound on any threshold-based procedure

In this section, we assume:

$$\beta_n \overset{(i)}{\geq} \frac{\log 2}{\log n} \quad \Longleftrightarrow \quad n^{1-\beta_n} \leq \frac{n}{2}, \quad \text{and} \tag{3.3a}$$

$$\max\left\{\beta_n, \frac{1}{\log^{(\gamma-1/2)/\gamma} n}\right\} \overset{(ii)}{<} r_n \overset{(iii)}{<} r_{\max} \quad \text{for some constant } r_{\max} < 1. \tag{3.3b}$$

Condition (i) requires that the proportion π_1 of non-nulls is at most $1/2$. Condition (ii) asserts that the natural requirement of $r_n > \beta_n$ is not sufficient, but further insists that r_n cannot approach zero too fast. The constants $\log 2$ and $(\gamma - 1/2)/\gamma$ are somewhat arbitrary and can be replaced, respectively, by $\log(1/(\pi_{\max}))$ for any $0 < \pi_{\max} < 1$ and $(\gamma - 1 + \rho)/\gamma$ for any $\rho > 0$. However, we fix their values in order not to introduce unnecessary extra parameters. With regard to condition (iii), although the assumption $r_n < 1$ is imposed because the problem becomes qualitatively easy for $r_n \geq 1$, the assumption that it is bounded away from one is a technical convenience that simplifies some of our proofs.

Our analysis shows that the FNR behaves differently depending on the closeness of the parameter r_n to the boundary of feasiblity given by β_n. In order to characterize this closeness, we define

$$r_{\min} = r_{\min}(\kappa_n) := \begin{cases} \beta_n + \kappa_n + \dfrac{\log(1/(6Z_\ell))}{\log n} & \text{if } \kappa_n \le 1 - \beta_n - \dfrac{\log(3/\log 16)}{\log n}, \\ 1 + \dfrac{\log(1/(24Z_\ell))}{\log n} & \text{otherwise.} \end{cases} \tag{3.4}$$

Here, κ_n is interpreted as the "exponent" of a target FDR rate q_n, in the sense that $q_n = n^{-\kappa_n}$. The rate q_n may differ from the actual achieved FDR_n, but it is nonetheless useful for parameterizing the quantities in our analysis. When we need to move between q_n and κ_n, we shall write $\kappa_n = \kappa_n(q_n) = \log(1/q_n)/\log n$ and $q_n = q_n(\kappa_n) = n^{-\kappa_n}$. For mathematical convenience, we wish to have the target FDR q_n be bounded away from one; therefore, we impose one further technical, but inessential assumption in this section:

$$q_n \le \min\left\{\frac{1}{24},\ \frac{1}{6Z_\ell}\right\} \quad \Longleftrightarrow \quad \kappa_n \ge \frac{\log \max\{24,\ 6Z_\ell\}}{\log n}. \tag{3.5}$$

The theorem that follows applies to all sample sizes $n > n_{\min,\ell}$ (the subscript ℓ denotes lower), where

$$n_{\min,\ell} := \min\left\{n \in \mathbb{N} : \exp\left(-\frac{n^{1-r_{\max}}}{24(Z_\ell \vee 1)}\right) \le \frac{1}{4}\right\} \tag{3.6}$$

$$= \left\lfloor [24(Z_\ell \vee 1)\log 4]^{1/(1-r_{\max})} \right\rfloor, \tag{3.7}$$

which is an explicit known function of the problem parameters, and can therefore be computed whenever the problem setting is fixed.

Finally, for $\gamma \in [1, \infty)$ and nonnegative numbers $a, b > 0$, we define the associated γ-"distance" as follows:

$$D_\gamma(a, b) := \left|a^{1/\gamma} - b^{1/\gamma}\right|^\gamma. \tag{3.8}$$

Our first main theorem states that for $r_n > r_{\min}(\kappa_n)$, the FNR decays as a power of $1/n$, with the exponent specified by the γ-distance.

Theorem 1. *Consider the γ-tail generalized Gaussian testing problem with sparsity β_n and signal level r_n, satisfying conditions* (3.3a)*, and* (3.3b) *and with sample size $n > n_{\min,\ell}$, from definition* (3.7)*. Then, for any choice of exponent $\kappa_n \in (0, 1)$ satisfying condition* (3.5)*, there exists a minimum signal strength $r_{\min}(\kappa_n)$ from definition* (3.4)*, such that any threshold-based procedure $\mathcal{I}$ that satisfies $\mathrm{FDR}_n(\mathcal{I}) \le n^{-\kappa_n}$ must have its* FNR *lower bounded as*

$$\mathrm{FNR}_n(\mathcal{I}) \ge \begin{cases} \dfrac{1}{32} & \text{if } r_n \in \left[\beta_n, r_{\min}\right], \\ c(\beta_n, \gamma) n^{-D_\gamma(\beta_n + \kappa_n, r_n)} & \text{otherwise,} \end{cases} \tag{3.9}$$

where $c(\beta_n, \gamma) := c_0 \exp\left(c_1 \beta_n^{(1-\gamma)/\gamma}\right)$, *with* (c_0, c_1) *being positive constants depending only on* (Z_ℓ, Z_u, γ).

The proof of this theorem is provided in Section 4.1. Note that the theorem holds for any choice of $\kappa_n \in (0,1)$. In the special case of constant pairs (β, r), this choice can be optimized to achieve the best possible lower bound on the risk $\mathcal{R}_n(\mathcal{I}) = \mathrm{FDR}_n(\mathcal{I}) + \mathrm{FNR}_n(\mathcal{I})$. Because we obtain this lower bound by optimizing the sum of the FDR and FNR lower bounds from Theorem 1, we want to balance the contributions from these two bounds. Doing so requires that we set the FDR rate κ equal to the corresponding FNR rate $D_\gamma(\beta + \kappa, r)$, which leads to a fixed-point equation for the overall rate, as summarized below.

Corollary 1. *When* $r > \beta$, *let* $\kappa_* = \kappa_*(\beta, r, \gamma) > 0$ *be the unique solution to the equation*

$$\kappa = D_\gamma(\beta + \kappa, r). \tag{3.10}$$

Then, the combined risk of any threshold-based multiple testing procedure $\mathcal{I}$ *is lower bounded as*

$$\mathcal{R}_n(\mathcal{I}) \gtrsim n^{-\kappa_*}, \tag{3.11}$$

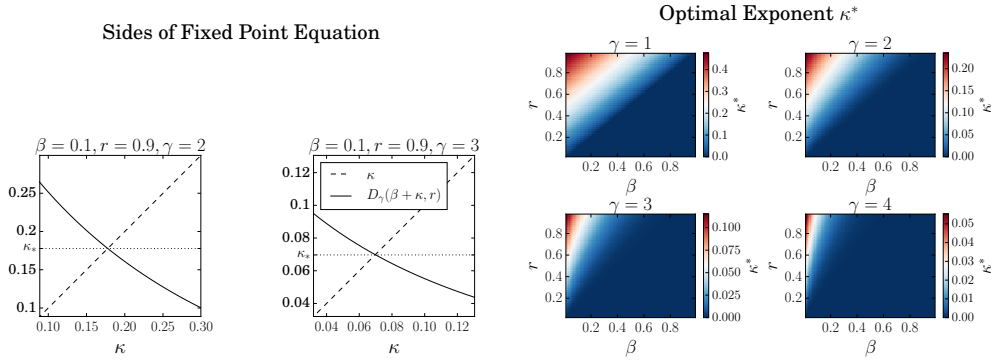
where $\gtrsim$ *denotes inequality up to a prefactor independent of* n.

The proof of this corollary is provided in the Supplementary Material, Section S3. Figure 1 shows the predictions in Corollary 1. In particular, panel (a) shows how the unique solution κ_* to equation (3.10) is determined for varying settings of the triple (r, β, γ). Panel (b) shows how κ_* varies over the interval $(0, 0.5)$, again for different settings of the triple (r, β, γ). As would be expected, the fixed point κ_* increases as a function of the difference $r - \beta > 0$.

3.3. Upper bounds for some specific procedures

Thus far, we have provided general lower bounds that can be applied to any threshold procedure. We now turn to the complementary question—how do these lower bounds compare to the results achievable by the BH and BC algorithms introduced in Section 2.3? Remarkably, we find that up to the constants defining the prefactor, both procedures achieve the minimax lower bound of Theorem 1.

We state these achievable results in terms of the fixed point κ_* from equation (3.10). Moreover, they apply to all problems with sample size $n > n_{\min,u}$ (the subscript u denotes upper), where



(a) (b)

Figure 1. Visualizations of the fixed-point equation (3.10). (a) Plots comparing the left- and right-hand sides of the fixed-point equation. (b) The optimal exponent κ^* as a function of r and β.

$$n_{\min,u} := \min\left\{n \in \mathbb{N} : \exp\left(-\frac{n^{1-r_{\max}}}{24}\right) \le \frac{1}{Z_u n}\right\}$$
$$= \min\left\{n \in \mathbb{N} : n \ \ge \ [24\log(Z_u n)]^{1/(1-r_{\max})}\right\}. \tag{3.12}$$

As in the case of (3.7), this lower bound on n is explicitly computable from the problem parameters.

In order to state our results cleanly, let us introduce the constants

$$c_{\mathrm{BH}} := \frac{Z_u}{36 Z_\ell}, \quad c_{\mathrm{BC}} := \frac{Z_u}{48 Z_\ell}, \quad \text{and} \quad \zeta := \max\left\{6Z_\ell, \ \frac{1}{6Z_\ell}\right\}, \tag{3.13}$$

and require in particular that $r_n \ge r_{\min}(\kappa_n(c_{\mathrm{A}} q_n))$ for algorithm $\mathrm{A} \in \{\mathrm{BH}, \mathrm{BC}\}$. Note that $c_{\mathrm{A}} < 1$ because $Z_\ell \ge Z_u$, by definition, and that the introduction of c_{A} into the argument of $r_{\min}$ only changes the minimum allowed value of r_n by a conceptually negligible amount of $\mathcal{O}(1/\log n)$.

Lastly, note that the BC procedure requires an additional mild condition that the number of non-nulls $n^{1-\beta_n}$ is large relative to the target FDR $q_n = n^{-\kappa_n}$ (otherwise, in some sense, the problem is too hard if there are too few non-nulls and a very strict target FDR). Specifically, we need that both quantities cannot simultaneously be too small, formalized by the assumption:

$$\exists n_{\min,\mathrm{BC}}, \quad \text{such that, for all} \quad n \ge n_{\min,\mathrm{BC}},$$
$$\text{we have} \quad \frac{3c_{\mathrm{BC}}}{4} \cdot \frac{q_n}{\log(1/q_n)} \cdot n^{1-\beta_n} \ \ge \ 1. \tag{3.14}$$

Note that when $r_n = r$ and $\beta_n = \beta$ are constants, this decay condition is satisfied

by $q_n = n^{-\kappa_*}$.

Our second main theorem delivers an optimality result for the BH and BC procedures, showing that under some regularity conditions, their performance achieves the lower bounds in Theorem 1, up to constant factors.

Theorem 2. *Consider the β_n-sparse γ-tail generalized Gaussian testing problem with target FDR level q_n, upper bounded as in condition* (3.5).

(a) Guarantee for BH procedure: Given a signal strength $r_n \geq r_{\min}(\kappa_n(c_{\mathrm{BH}} q_n))$ and sample size $n > n_{\min,u}$, as in condition (3.12)*, the BH procedure satisfies the bounds*

$$\mathrm{FDR}_n \leq q_n \quad \textit{and} \quad \mathrm{FNR}_n \leq \frac{2\zeta_{\mathrm{BH}}^{2\beta_n^{(1-\gamma)/\gamma}}}{Z_u} \cdot n^{-D_\gamma(\beta_n+\kappa_n, r_n)}, \textit{ where } \zeta_{\mathrm{BH}} := \frac{\zeta}{c_{\mathrm{BH}}}. \tag{3.15}$$

(b) Guarantee for BC procedure: Given a signal strength $r_n \geq r_{\min}(\kappa_n(c_{\mathrm{BC}} q_n))$ and sample size $n > \max\{n_{\min,\mathrm{BC}}, n_{\min,u}\}$, as in condition (3.14)*, the BC procedure satisfies the bounds*

$$\mathrm{FDR}_n \leq q_n \quad \textit{and} \quad \mathrm{FNR}_n \leq \frac{2\zeta_{\mathrm{BC}}^{2\beta_n^{(1-\gamma)/\gamma}}}{Z_u} \cdot n^{-D_\gamma(\beta_n+\kappa_n, r_n)} + q_n,$$
$$\textit{where } \zeta_{\mathrm{BC}} := \frac{\zeta}{c_{\mathrm{BC}}}. \tag{3.16}$$

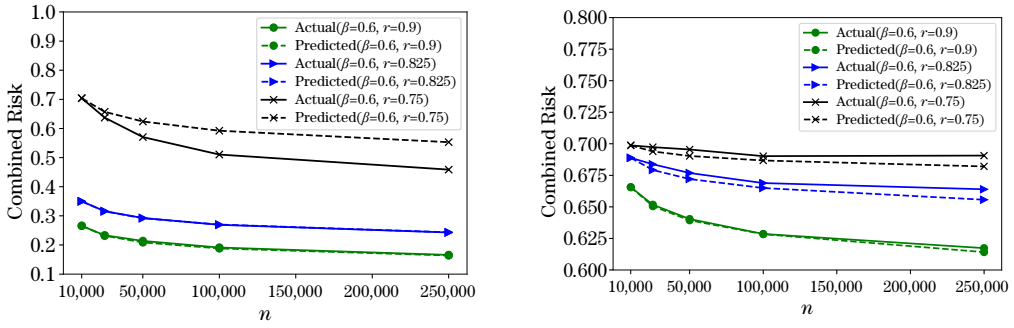
The proof of the theorem can be found in Section 4.2. For constant pairs (r, β), Theorem 2 can be applied with a target FDR proportional to $n^{-\kappa_*}$ to show that the BH and BC procedures both achieve the optimal decay of the combined FDR-FNR up to constant factors, as stated formally below.

Corollary 2. *For $\beta < r$ and $q_* = c_* n^{-\kappa_*}$, with $0 < c_* \leq \min\{1/24,\ 1/(6Z_\ell)\}$, the BH and BC procedures with target FDR q_* satisfy*

$$\mathcal{R}_n \lesssim n^{-\kappa_*}. \tag{3.17}$$

The proof of this corollary is given in the Supplementary Material, Section S5. To help visualize the result of the corollary, Figure 2 displays the results of simulations of the BH procedure that show the correspondence between its performance and the theoretically predicted rate of $n^{-\kappa_*}$.

Despite the optimality, Figures 1 and 2 suggest that the methods may not be practical. Although asymptotic consistency can be achieved when $r > \beta$, the convergence of the risk to zero can be extremely slow, exhibiting "nonparametric" rates far slower than $n^{-1/2}$. Figure 2 shows in particular that the decay to zero



(a) $\gamma = 1$ (b) $\gamma = 2$

Figure 2. Results of simulations comparing the predicted combined risk with the actual, experimentally observed risk for the BH procedure. Agreement is good and improves as the gap $(r - \beta)$ increases, which we believe occurs because the sampling error becomes a smaller fraction of the risk as the separation increases.

may be barely evident, even for sample sizes as large as $n = 250{,}000$, and with comparatively strong signals. The "nonparametric" nature may arise because the dimensionality of the decision space increases linearly with the sample size. Thus, asymptotically, the advantage of having a greater amount of data seems to only *just* overcome the disadvantage of having to make an increasing number of decisions. However, nonasymptotically, one cannot hope to drive both the FDR and the FNR to zero at any practical sample size in this general setting, at least when the mean signal lies below the maximum of the nulls (i.e., $r_n < 1$).

Intuition for the γ-distance. The distance D_γ plays a crucial role because of the scaling of order statistics under the tGG_γ model. If $W_1, \ldots, W_n$ are independent and identically distributed (i.i.d.) from a $\mathrm{tGG}_\gamma(0)$ model, then—ignoring constants inside the logarithm—we expect the ith-largest order statistic $W_{(i)}$ to be around $(\gamma i \log n)^{1/\gamma}$ if $i \ll n/2$, and around $-(\gamma i \log n)^{1/\gamma}$ if $n - i \ll n/2$. If an algorithm is to achieve an FNR on the order of $n^{-\kappa'}$, it must successfully identify all but the smallest $n^{-\kappa'}$ fraction of true signals. Thus, the algorithm's cutoff for rejection must exceed the $m - n^{-\kappa'}m$ order statistic of the signals, which is approximately

$$\mu - \left(\gamma \log \frac{m}{n^{-\kappa'}m}\right)^{1/\gamma} = (\gamma r \log n)^{1/\gamma} - (\gamma \kappa' \log n)^{1/\gamma}. \tag{3.18}$$

If we suppose that the FDR is also vanishing at a rate $n^{-\kappa}$, then first of all the algorithm must identify about $(1 \pm o(1))m$ indices as signals, because otherwise either the FDR or FNR would fail to vanish. Second, it must be that the $n^{-\kappa}m$th

or, equivalently, the $n^{1-\beta-\kappa}$th largest null is of the order of the quantity in (3.18). Combining these insights, we obtain the relation

$$(\gamma(\beta+\kappa) = (\gamma r \log n)^{1/\gamma} - (\gamma\kappa' \log n)^{1/\gamma},$$

which, after rearranging, yields the heuristic

$$\kappa' = \left(r^{1/\gamma} - (\beta+\kappa)^{1/\gamma}\right)^{1/\gamma}. \tag{3.19}$$

The theorems and corollaries in this paper together show that this intuition is exactly right.

Regime of linear sparsity. We turn to the regime of *linear sparsity*—that is, when the number of signals scales as $\pi_1 n$, for some scalar $\pi_1 \in (0,1)$. Recalling that we have parameterized the number of signals as $n^{1-\beta_n}$, some algebra leads to $\beta_n = (\log(1/\pi_1))/\log n$; thus, both Theorem 1 and Theorem 2 predict an upper and lower bound on the risk of the form

$$c_0 \exp\left(c_1 \left[\frac{\log n}{\log(1/\pi_1)}\right]^{(\gamma-1)/\gamma}\right) \cdot n^{-\kappa_*}. \tag{3.20}$$

Note that here we overload the exponent κ_* to the case when it is nonconstant. In order to interpret this result, observe that if $r_n = r$ is constant, then $\kappa_* = r/2^\gamma - o(1)$; thus, the rate is $n^{-r/2^\gamma}$ up to subpolynomial factors in n. On the other hand, if $r_n = 1/(\log^{(\gamma-1/2)/\gamma} n)$ is at the extreme lower limit permitted by the lower bound (ii) in (3.3b), then it is not difficult to see that $\kappa_* \approx \log^{-(\gamma-1/2)/\gamma} n$, which ensures that $n^{\kappa_*} \gg \exp\left(\log^{(\gamma-1)/\gamma} n\right)$, so that the risk (3.20) still approaches zero asymptotically, albeit subpolynomially, in n.

4. Proofs

We now turn to the proofs of our main results, namely Theorems 1 and 2. The proofs of the associated corollaries can be found in the Supplementary Material.

4.1. Proof of Theorem 1

The main idea of the proof is to reduce the problem of lower bounding the FNR_n of threshold-based procedures that use random, data-dependent thresholds T_n to the easier problem of lower bounding the FNR_n of threshold-based procedures that use a deterministic, data-independent threshold t_n. We refer to the latter class of procedures as *fixed-threshold procedures*, and we parameterize them by their target FDR $q_n = n^{-\kappa_n}$. Specifically, we define the *critical*

threshold, derived from the critical regime boundary $r_{\min}$ from equation (3.4), by

$$\tau_{\min}(\kappa_n) := \big(\gamma r_{\min}\big(\kappa_n\big)\log n\big)^{1/\gamma} \equiv \tau_{\min}(q_n) := \left(\gamma r_{\min}\left(\frac{\log(1/q_n)}{\log n}\right)\log n\right)^{1/\gamma}. \tag{4.1}$$

Here, and throughout the proof, we express $\tau_{\min}$ and $r_{\min}$ as functions of q_n rather than κ_n; this formulation makes certain calculations in the proof simpler to express.

From data-dependent threshold to fixed threshold. Our first step is to reduce the analysis from data-dependent to fixed-threshold procedures. In particular, consider a threshold procedure, using a possibly random threshold T_n, that satisfies the FDR uppper bound $\mathrm{FDR}_n(T_n) \le q_n$. We claim that the FNR of any such procedure must be lower bounded as

$$\mathbb{E}[\mathrm{FNP}_n\big(T_n\big)] \;\ge\; \frac{\mathrm{FNR}_n\big(\tau_{\min}\big(4q_n\big)\big)}{16}. \tag{4.2}$$

This lower bound is crucial, because it reduces the study of random threshold procedures (LHS) to the study of fixed-threshold procedures (RHS). Its proof can be found in the Supplementary Material, Section S1.

Our next step is to lower bound the FNR for choices of the threshold $t \ge \tau_{\min}(q_n)$:

Lemma 1. *For any $t \ge \tau_{\min}(q_n)$, we have*

$$\mathrm{FNR}_n(t) \ge \begin{cases} \dfrac{\zeta^{2\beta_n^{(1-\gamma)/\gamma}}}{Z_\ell} \cdot n^{-D_\gamma(\beta_n+\kappa_n,r)} & \text{if } r > r_{\min}\big(\kappa_n(q_n)\big), \\ \dfrac{1}{2} & \text{otherwise,} \end{cases} \tag{4.3}$$

where ζ is defined as in (3.13).

The proof of this lemma can be found in the Supplementary Material, Section S2. Using Lemma 1 and the lower bound (4.2), we can now complete the proof of Theorem 1. We split the argument into two cases:

Case 1. First, suppose that $r \le r_{\min}(\kappa_n(4q_n))$. In this case, we have

$$\mathrm{FNR}_n(T_n) \overset{(i)}{\ge} \frac{\mathrm{FNR}_n\big(\tau_{\min}\big(4q_n\big)\big)}{16} \overset{(ii)}{\ge} \frac{1}{32},$$

where step (i) follows from the lower bound (4.2), and step (ii) follows by lower bounding the FNR by $1/2$, as is guaranteed by Lemma 1 in the regime $r \le r_{\min}(\kappa_n(4q_n))$.

Case 2. Otherwise, we may assume that $r > r_{\min}(4q_n)$. In this case, we have

$$\mathrm{FNR}_n(T_n) \overset{(i)}{\geq} \frac{\mathrm{FNR}_n\big(\tau_{\min}\big(4q_n\big)\big)}{16} \overset{(ii)}{\geq} \frac{\big(4\zeta\big)^{2\beta_n^{(1-\gamma)/\gamma}}}{Z_\ell} \cdot n^{-D_\gamma(\beta_n+\kappa_n,r)}.$$

Here, step (i) follows from the lower bound (4.2), whereas step (ii) follows from applying Lemma 1 in the regime $r > r_{\min}(\kappa_n(4q_n))$. With some further algebra, we find that

$$\begin{aligned}\mathrm{FNR}_n(T_n) &\geq \frac{1}{Z_\ell} \exp\big(2\log\big(4\zeta\big) \cdot \beta_n^{(1-\gamma)/\gamma}\big) n^{-D_\gamma(\beta+\kappa_n,r)} \\ &= c_0 \exp\big(c_1 \beta_n^{(1-\gamma)/\gamma}\big) n^{-D_\gamma(\beta+\kappa_n,r)},\end{aligned}$$

where $c_0 := 1/Z_\ell$ and $c_1 := 2\log(4\zeta)$. Note that because $Z_\ell > 0$ and $\zeta \geq 1$, the constants c_0 and c_1 are both positive, as claimed in the theorem statement.

4.2. Proof of Theorem 2

We now sketch the proof of Theorem 2, which states that the BH and BC algorithms achieve the minimax rate (3.9) when $r_n > r_{\min}(\kappa_n(c_{\mathrm{A}} q_n))$, where $\mathrm{A} \in \{\mathrm{BH},\ \mathrm{BC}\}$ and c_{A} is the algorithm-dependent constant defined in (3.13). For reasons of space, the details are relegated to the Supplementary Materials, Section S4.

The proof strategy for both algorithms is essentially the same. Given a target FDR rate q_n, we apply each algorithm with q_n as the target FDR level, and then prove that the resulting threshold satisfies $t_{\mathrm{A}} \leq \tau_{\min}(c_{\mathrm{A}} q_n)$ with high probability. Letting $\tau_{\min,\mathrm{A}} = \tau_{\min}\big(c_{\mathrm{A}} q_n\big)$, we can formulate the specific claims we seek as:

$$\mathbb{P}\big(t_{\mathrm{BH}} > \tau_{\min,\mathrm{BH}}\big) \leq \exp\left(-\frac{n^{1-r_{\max}}}{24}\right) \tag{4.4}$$

and

$$\mathbb{P}\left(t_{\mathrm{BC}} > \tau_{\min,\mathrm{BC}}\right) \leq q_n + \exp\left(-\frac{n^{1-r_{\max}}}{24}\right). \tag{4.5}$$

The known properties of the algorithms guarantee the required FDR bounds (as studied by Arias-Castro and Chen (2017); Foygel Barber and Candès (2015); Benjamini and Hochberg (1995)). The following converse to Lemma 1, coupled with the probabilistic upper bounds (4.4) and (4.5), provides the requisite upper bounds on the FNR.

Lemma 2. *If $r_n > r_{\min}(cq_n)$ and $t \leq \tau_{\min}(cq_n)$, for some $c > 0$, then we have*

$$\mathrm{FNR}_n\big(t\big) \leq \frac{\big(\max\big\{c,\ 1/c\big\} \cdot \zeta\big)^{2\beta_n^{(1-\gamma)/\gamma}}}{Z_u} \cdot n^{-D_\gamma(\beta_n+\kappa_n,r)},$$

where constant ζ *is defined in* (3.13).

5. Discussion

Despite considerable interest in multiple testing with FDR control, we have relatively little understanding of the nonasymptotic trade-off between controlling FDR and the analogous measure of power known as the FNR. In this study, we explored this issue in the context of the sparse generalized Gaussian model, deriving the first nonasymptotic lower bounds on the sum of the FDR and FNR. We complemented these lower bounds by establishing the nonasymptotic minimaxity of both the BH and the BC procedures for FDR control. The theoretical predictions are validated using simple simulations, and our results include recent asymptotic results (Arias-Castro and Chen (2017)) as special cases. Our work introduces a simple proof strategy based on a reduction to deterministic and data-oblivious procedures. We suspect this core idea may apply to other multiple testing settings. In particular, because our arguments do not depend on the CDF asymptotics in the way that many classical analyses of both global null testing and FDR control procedures do, we hope they can be adapted to other problems as well, as described below.

As mentioned after the statement of Theorem 2, the practical implications of our results are somewhat pessimistic. Even for rather simple problems with $r - \beta$ of constant order, the resulting rate at which the risk tends to zero can be far slower than $n^{-1/2}$. (Indeed, it seems such a parametric rate is only achievable when $\gamma = 1, r_n \to 1, \beta_n \to 0$.) Hence, in practice, one must carefully consider whether a good FDR or a good FNR is more important, because achieving both may not be possible, unless most of the signals to be identified are rather large.

Future directions

We have focused on establishing a nonasymptotic trade-off between the FDR and FNR in what is arguably the simplest interesting model of the problem. By way of contrast, much of the recent multiple testing literature focuses on developing valid FDR control procedures that can gain power or precision by explicitly using prior knowledge and structure in various ways, including using null-proportion adaptivity (Storey (2002); Storey, Taylor and Siegmund (2004)), grouping of hypotheses (Foygel Barber and Ramdas (2016); Hu, Zhao and Zhou (2010)), prior or penalty weights (Benjamini and Hochberg (1997); Genovese, Roeder and Wasserman (2006)), or other forms of structure (Li and Foygel Barber

(2016); Ramdas et al. (2017)).

Similarly, the issue of dependence—either positive or arbitrary—between test statistics has been an area of focus (Benjamini and Yekutieli (2001); Blanchard and Roquain (2008); Ramdas et al. (2017)). (Dependence has already been explored for the higher criticism statistic applied to the detection problem (Hall and Jin (2008); Jin and Ke (2016); Hall and Jin (2010))). Still others have studied the nonexchangeability of hypotheses, either in the context of multiple scales of signal strength, or in the context of online FDR procedures (Foster and Stine (2008); Javanmard and Montanari (2015)).

Owing to the increasing importance of structured, dependent, and nonexchangeable settings, developing analogues of our results for such settings is a worthwhile direction for future work. Furthermore, it is far from clear that known procedures are optimal under assumptions of structure, dependence, or various kinds of non-exchangeability, so that an improved understanding of the fundamental difficulty of the multiple testing problem under such assumptions may yield improved algorithms. Chen and Arias-Castro (2017) have made progress in this direction by providing *upper* bounds for existing procedures for the online FDR problem (Javanmard and Montanari (2015)), but much still remains unknown.

Finally, a general proof technique for establishing nonasymptotic lower bounds in multiple testing remains an important direction for future work. In this study, we pursued an approach based on a reduction to a class of nonadaptive procedures, a principle that could perhaps be applied to other multiple testing problems. However, our arguments are, however, based on analytical calculations and, therefore, are sensitive to the specific observation model under consideration. Thus, an especially pressing problem is that of developing approaches that depend on the intrinsic structural properties of the test statistic distributions, and that are less brittle when it becomes inconvenient to reason about the analytical forms.

Supplementary Material

The online Supplementary Material contains proofs that—for space reasons—we could not accommodate in the main body of the paper. These include parts of the proofs of the theorems, as well as proofs of the corollaries and technical lemmas.

Acknowledgements

This work was partially supported by Office of Naval Research Grant DOD-ONR-N00014, Air Force Office of Scientific Research Grant AFOSR-FA9550-14-1-0016, and Army Research Office grant W911NF-16-1-0368. In addition, MR was supported by an NSF Graduate Research Fellowship and a Fannie and John Hertz Foundation Google Fellowship.

References

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

Arias-Castro, E. and Chen, S. (2017). Distribution-free multiple testing. *Electronic Journal of Statistics* **11**, 1983–2001.

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

Benjamini, Y. and Hochberg, Y. (1997). Multiple hypotheses testing with weights. *Scandinavian Journal of Statistics* **24**, 407–418.

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

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

Bogdan, M., Chakrabarti, A., Frommlet, F. and Ghosh, J. (2011). Asymptotic Bayes-optimality under sparsity of some multiple testing procedures. *The Annals of Statistics* **39**, 1551–1579.

Chen, S. and Arias-Castro, E. (2017). Sequential multiple testing. *arXiv preprint arXiv:1705.10190*.

Chi, Z. (2007). On the performance of FDR control: Constraints and a partial solution. *The Annals of Statistics* **35**, 1409–1431.

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

Donoho, D. and Jin, J. (2006). Asymptotic minimaxity of false discovery rate thresholding for sparse exponential data. *The Annals of Statistics* **34**, 2980–3018.

Donoho, D. and Jin, J. (2015). Higher criticism for large-scale inference, especially for rare and weak effects. *Statistical Science* **30**, 1–25.

Foster, D. and Stine, R. (2008). α-investing: A procedure for sequential control of expected false discoveries. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**, 429–444.

Foygel Barber, R. and Candès, E. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics* **43**, 2055–2085.

Foygel Barber, R. and Ramdas, A. (2016). The p-filter: Multi-layer FDR control for grouped hypotheses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

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

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

Hall, P. and Jin, J. (2008). Properties of higher criticism under strong dependence. *The Annals of Statistics* **36**, 381–402.

Hall, P. and Jin, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *The Annals of Statistics* **38**, 1686–1732.

Hu, J., Zhao, H. and Zhou, H. (2010). False discovery rate control with groups. *Journal of the American Statistical Association* **105**, 1215–1227.

Javanmard, A. and Montanari, A. (2015). On online control of false discovery rate. *arXiv preprint arXiv:1502.06197*.

Ji, P. and Jin, J. (2012). UPS delivers optimal phase diagram in high-dimensional variable selection. *The Annals of Statistics* **40**, 73–103.

Ji, P. and Zhao, Z. (2014). Rate optimal multiple testing procedure in high-dimensional regression. *arXiv preprint arXiv:1404.2961*.

Jin, J. and Ke, T. (2016). Rare and weak effects in large-scale inference: Methods and phase diagrams. *Statistica Sinica* **26**, 1–34.

Li, A. and Foygel Barber, R. (2016). Multiple testing with the structure adaptive Benjamini-Hochberg algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **81**, 45–74.

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

Ramdas, A., Foygel Barber, R., Wainwright, M. and Jordan, M. (2017). A unified treatment of multiple testing with prior knowledge. *The Annals of Statistics* **47**, 2790–2821.

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

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

Tukey, J. (1953). *The Problem of Multiple Comparisons: Introduction and Parts A, B, and C*. Princeton University.

Department of EECS, UC Berkeley, Berkeley, CA 94720, USA.

E-mail: rabinovich@berkeley.edu

Department of Statistics and EECS, UC Berkeley, Berkeley, CA 94720, USA.

E-mail: ramdas@berkeley.edu

Department of Statistics and EECS, UC Berkeley, Berkeley, CA 94720, USA.

E-mail: jordan@berkeley.edu

Department of Statistics and EECS, UC Berkeley, Berkeley, CA 94720, USA.

E-mail: wainwrig@berkeley.edu

(Received November 2017; accepted June 2018)