

Some optimality properties of FDR controlling rules under sparsity

Florian Frommlet

Department of Medical Statistics

Medical University of Vienna

Spitalgasse 23

A-1090 Vienna, Austria

e-mail: Florian.Frommlet@meduniwien.ac.at

and

Małgorzata Bogdan

Departments of Mathematics and Computer Science

Wroclaw University of Technology and

Jan Długosz University in Czestochowa

Department of Statistics

Stanford University

e-mail: Malgorzata.Bogdan@pwr.wroc.pl

Abstract: False Discovery Rate (FDR) and the Bayes risk are two different statistical measures, which can be used to evaluate and compare multiple testing procedures. Recent results show that under sparsity FDR controlling procedures, like the popular Benjamini-Hochberg (BH) procedure, perform also very well in terms of the Bayes risk. In particular asymptotic Bayes optimality under sparsity (ABOS) of BH was shown previously for location and scale models based on log-concave densities. This article extends previous work to a substantially larger set of distributions of effect sizes under the alternative, where the alternative distribution of true signals does not change with the number of tests m , while the sample size n slowly increases. ABOS of BH and the corresponding step-down procedure based on FDR levels proportional to $n^{-1/2}$ are proved. A simulation study shows that these asymptotic results are relevant already for relatively small values of m and n . Apart from showing asymptotic optimality of BH, our results on the optimal FDR level provide a natural extension of the well known results on the significance levels of Bayesian tests.

Keywords and phrases: Asymptotic optimality, Bayes risk, false discovery rate, multiple testing, two groups model.

Received December 2012.

1. Introduction

Driven by a vast number of applications, over the last few years multiple hypothesis testing with sparse alternatives has become a topic of intensive research (see [1, 9, 12, 13, 24] or [31]). As a result of this interest many new multiple testing procedures have been proposed, which can be compared according to several

different optimality criteria. In the classical context, a multiple testing procedure is considered to be *optimal* if it maximizes the number of true discoveries, while keeping one of the type I error measures (like Family Wise Error Rate, False Discovery Rate or the expected number of false positives) at a certain fixed level (see [10, 19, 23, 30, 34, 36, 41] or [42]). A different notion of optimality is proposed in [38] and [6], which investigate multiple testing procedures in the context of minimizing the Bayes risk.

The frequentist measures of accuracy for multiple testing procedures, like the False Discovery Rate (FDR) or the Family Wise Error Rate (FWER), are seemingly very different from the Bayes risk. Nevertheless several papers (see e.g. [4, 5] or [21]) reported that under sparsity (i.e. when the proportion p of alternatives among all tests is small) the popular Benjamini-Hochberg procedure (BH, [2]) has very good properties with respect to minimizing the Bayes risk under 0-1 loss. These empirical findings have been confirmed by the theoretical results reported in [6] and [33], where BH is proved to have some asymptotic optimality properties with respect to minimizing the Bayes risk under sparsity. These results complement findings of [1], where it is shown that the estimator of the vector of means based on the hard-thresholding rule, with cutoff values provided by BH, is asymptotically minimax under sparsity.

Asymptotic results of [6] are concerned with testing hypotheses about means of normal distributions, where the true means under the alternative have a normal distribution with a standard deviation slowly increasing with m . These results were further extended in [33] to a very general class of location and scale models. While the results of [33] are of substantial theoretical value, in practical applications it is perhaps too limiting to consider test statistics whose marginal distribution under the alternative differs from the null distribution only by a location or a scale parameter. In this article we present an extension of [6] which we believe to be of more practical relevance than the setting discussed in [33].

To motivate our proposed model we want to look at two examples from statistical genetics. First consider a rather simple two groups model for RNA micro-array data given in (3.11) of [16]. Each test statistic is assumed to have a normal distribution $N(\mu_i, \sigma^2)$, which typically can be justified by the Central Limit Theorem. Furthermore it is assumed that only a very small proportion p of the true means μ_1, \dots, μ_m are different from zero, and that their conditional distribution can be described by a normal distribution $N(\mu_0, \sigma_0^2)$. It follows that the marginal distribution of test statistics under the alternative are $N(\mu_0, \sigma^2 + \sigma_0^2)$, which differs from the null distribution $N(0, \sigma^2)$ both by a location and a scale parameter. Thus even this simple example goes beyond the framework of [33], where tests for the difference in only one of these parameters are considered.

A second example are genetic association studies, where it is quite common that effect sizes are modelled by exponential or Laplace distributions. For example in the model of [32] the effect sizes of causative SNPs are Laplace, while the environmental effects are normal, corresponding exactly to our setting. The marginal distribution under the alternative is thus the convolution of a Laplace and a normal distribution while the null distribution is again normal,

which again places this practically important example out of the class of models treated in [33].

In this article we will consider only testing the point null hypotheses that $\mu_i = 0$. This is less general than the normal scale mixture model from [6], where μ was allowed to be normally distributed with small variance under the null hypothesis. However, this restriction can be justified in many applications. For example whenever it is assumed in statistical genetics that disease characteristic phenotypes are influenced only by a small number of genes, then the restriction to point null hypotheses is legitimate. Considering point null hypotheses furthermore allows nontrivial asymptotic inference (which means positive asymptotic power) when keeping the distribution of true effects under the alternative fixed while increasing the number of tests. This assumption, natural in many practical applications, substitutes the assumptions of [6] and [33], where the magnitude of true effects increases with the number of tests.

We believe that both assumptions, increasing the effect size and increasing the sample size, are relevant in certain practical situations. For example in RNA micro-array experiments the sample size is often rather small, while effect sizes are believed to be quite large, which relates to the asymptotic analysis of [6] and [33]. On the other hand in association studies it is common to collect data from very large samples just to be able to detect relatively moderate effect sizes. For instance in genome wide association studies one typically has samples of size $n > 1000$, while the number of genetic markers is $m > 10^6$. On the other hand the number of markers k which are actually expected to be associated with the trait in question is usually considered to be rather moderate, like $k < 100$. Then for the proportion of markers under the alternative hypothesis it holds that $p < 10^{-4}$. These figures motivate an asymptotic scheme where the number of tests m goes to infinity, while the sample size $n = n_m$ increases at a slower rate, and $p = p_m$ converges to 0.

From a mathematical point of view, the major contribution of the present article is the extension of the results of [6] to cover a wide range of distributions of effect sizes μ_i under the alternative. The technical details on the class of feasible distributions are provided in Assumption (B), which includes for example all distributions with positive density over the real line. The main mathematical challenge is that one no longer obtains closed form expressions for the critical values of the Bayes rule. The rather elaborate asymptotic analysis which is then necessary to obtain approximate formulas of the Bayes risk uses some techniques introduced in [25].

We will develop the asymptotic theory under the general condition that $p \rightarrow 0$, but to explain the main message of this article we summarize here the results for the special case where $p \propto m^{-\beta}$, with $\beta \in (0, 1]$. We will see that in this scenario nontrivial asymptotic inference is possible only for n converging to infinity at least at a rate $n \propto \log m$. In terms of applications this allows the number of tests to be much larger than the sample size, like for example in genome wide association studies.

Asymptotic Bayes optimality under sparsity (ABOS) depends on the loss function under which the Bayes risk is computed. We consider a generalized 0-1

loss, and define δ as the ratio of losses between type I and type II errors. Our main result is concerned with BH which controls FDR at the nominal level α . When the loss ratio δ is kept fixed, and if α converges to 0 at a rate which is roughly proportional to $n^{-1/2}$, then BH is ABOS for any $\beta \in (0, 1]$. In that sense BH adapts well to any unknown level of sparsity.

Apart from BH we also consider the Bonferroni procedure which controls FWER at the nominal level α . Under the conditions on α which yield ABOS of BH, it turns out that the Bonferroni rule is ABOS only for the extremely sparse case $\beta = 1$ (which means that $p \propto 1/m$). We will furthermore illustrate some situations where the Bonferroni rule is optimal when $\beta < 1$, but then the nominal level α depends on β . Thus in contrast to BH the Bonferroni rule does not adapt well to an unknown level of sparsity.

Apart from these theoretical findings, we report the results of an extensive simulation study, comparing the performance of BH, Bonferroni correction and a multiple testing procedure based on the empirical Bayes estimate proposed in [26]. The study shows that our asymptotic results are relevant already for quite moderate values of m and n .

The rest of the paper is organized as follows. In Section 2 we present our statistical model, asymptotic assumptions and Theorem 2.2, which provides the conditions under which fixed threshold multiple testing rules are ABOS. The resulting Corollary 2.2 shows that the universal threshold $2 \log m$ of [14] is ABOS in the extremely sparse case $p \propto m^{-1}$ and for sample size n satisfying $\log n = o(\log m)$. The most important theorems on multiple testing are given in Section 3. Specifically, Theorem 3.2 specifies the conditions under which BH and its corresponding step down procedure are ABOS, while Theorem 3.3 gives respective conditions for the Bonferroni correction. The results of the simulation study are presented in Section 4. Finally, Section 5 summarizes our results and discusses the directions for future work. Most of the technical proofs have been put in the Appendix, which also includes a discussion on the relationship between rules controlling the Bayesian False Discovery Rate (BFDR) and the Bayes classifier.

2. Asymptotic Bayes optimality of multiple testing rules with fixed threshold

2.1. Statistical model

Consider a set of m normal populations $\mathcal{N}(\mu_i, \sigma^2)$, $i = 1, \dots, m$. We are interested in testing point null hypotheses $H_{0i} : \mu_i = 0$ against the alternatives $H_{Ai} : \mu_i \neq 0$, based on simple random samples $X_i = (X_{1i}, \dots, X_{ni})$ of size n from each of these populations. The effects under study μ_i are supposed to be independent and identically distributed according to a mixture distribution

$$\nu_{mix} = (1 - p)d_0 + p\nu, \quad (2.1)$$

where d_0 is the Dirac measure at 0, ν is a probability measure on the real line describing the distribution of μ_i under the alternative, and $p \in (0, 1)$ is

the proportion of alternatives among all tests. Since ν describes the alternative distribution of the different μ_i , we assume that $\nu(\{0\}) = 0$. Furthermore both positive and negative values of μ_i should be possible, that is

$$\nu(-\infty, 0) > 0 \quad \text{and} \quad \nu(0, \infty) > 0. \quad (2.2)$$

From (2.1) it easily follows that the marginal distribution of the sample mean $\bar{X}_i = \frac{1}{n} \sum_{j=1}^n X_{ji}$ is the mixture

$$\bar{X}_i \sim (1-p)\mathcal{N}(0, \sigma^2/n) + p (\nu * \mathcal{N}(0, \sigma^2/n)), \quad (2.3)$$

where the pdf of the second measure is computed by convolution of ν and $\mathcal{N}(0, \sigma^2/n)$.

Remark 2.1. To reduce the complexity of notation we assume that the population variances σ_i^2 are the same for every i . However, the results on asymptotic optimality presented in this article hold also if the sequence of possible different variances $\sigma_i^2 \in (0, \infty)$ satisfies $\liminf_{i \rightarrow \infty} \sigma_i^2 > 0$ and $\limsup_{i \rightarrow \infty} \sigma_i^2 < \infty$.

Remark 2.2. Note that for sufficiently large n the assumption that $\bar{X}_i|\mu_i \sim N(\mu_i, \sigma_i^2/n)$ can be justified by the Central Limit Theorem, even when the conditional distribution of $X_i|\mu_i$ is not normal. According to [17], under weak assumptions on the moments of the distribution of $X_i|\mu_i$, the level of the simultaneous tests based on the t-test statistics $\sqrt{n}(\bar{X}_i|\mu_i - \mu_i)/\hat{\sigma}_i$ is accurate provided $\log m$ increases at a strictly slower rate than $n^{1/3}$. Based on these findings we expect that under similar assumptions our asymptotic results can be generalized to the case where the distribution of $X_i|\mu_i$ is not normal and the variances σ_i^2 are estimated separately for every i . This is partially confirmed by our simulation study, which illustrates the performance of BH based on the t-statistics, with σ_i estimated separately for each i .

2.2. Asymptotic framework and the power of the Bayes oracle

Our decision theoretic framework for multiple testing is based on a generalization of the standard 0-1 loss. There are m decisions to be made. For each false rejection (type I error) we assign a loss of δ_0 , and for missing a true signal (type II error) a loss of δ_A . The total loss of a multiple testing procedure is then defined as the sum of losses for individual tests. The total loss is clearly minimized by applying the Bayes classifier to each individual test, the decision rule which was called Bayes oracle in [6].

As noted in [25], if $p \in (0, 1)$ then for any measure ν satisfying (2.2) and sufficiently large n , the Bayes classifier chooses H_{0i} if $\bar{X}_i \in (a_n, b_n)$, where the critical values a_n and b_n are uniquely defined by

$$\begin{aligned} a_n &< 0 < b_n \\ (1-p)\delta_0 &= p \delta_A \int_{\mathcal{R}} \exp\left(n\left(a_n \frac{\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right)\right) d\nu(\mu), \\ (1-p)\delta_0 &= p \delta_A \int_{\mathcal{R}} \exp\left(n\left(b_n \frac{\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right)\right) d\nu(\mu). \end{aligned} \quad (2.4)$$

Equivalently one can work with the scaled test statistics $Z_i = \frac{\sqrt{n}}{\sigma} \bar{X}_i$, for which the critical values $c_a := \frac{\sqrt{n}}{\sigma} a_n$ and $c_b := \frac{\sqrt{n}}{\sigma} b_n$ of the Bayes rule are given by the equations

$$\begin{aligned}
 c_a &< 0 < c_b \\
 (1-p)\delta_0 e^{-c_a^2/2} &= p \delta_A \int_{\mathcal{R}} \exp\left(-\frac{1}{2} \left[c_a - \frac{\sqrt{n}}{\sigma} \mu\right]^2\right) d\nu(\mu), \quad (2.5) \\
 (1-p)\delta_0 e^{-c_b^2/2} &= p \delta_A \int_{\mathcal{R}} \exp\left(-\frac{1}{2} \left[c_b - \frac{\sqrt{n}}{\sigma} \mu\right]^2\right) d\nu(\mu).
 \end{aligned}$$

All the results in the main text will be presented at the scale of the test statistics, that means in terms of c_a and c_b . However, some of the technical proofs in the appendix are more conveniently stated working with (2.4).

As explained in the Introduction, our asymptotic scheme which is motivated for example by genetic association studies assumes that the number of tests m goes to infinity, while the sample size $n = n_m$ increases at a slower rate, and $p = p_m$ converges to 0. From now on sequences will be indexed with m , but one has to be aware that $m \rightarrow \infty$ will also imply that $n \rightarrow \infty$ and $p \rightarrow 0$.

Also, in our analysis we are mainly interested in the situation where the loss ration $\delta = \delta_0/\delta_A$ is kept constant, which directly relates to the risk based on misclassification loss. However, we also consider the situation when $\delta = \delta_m$ slowly converges to zero, such that $\log \delta = o(\log p)$. This allows us in Theorem 3.2 to obtain optimality results for multiple testing rules at a constant nominal FDR level α .

We will focus on an asymptotic scenario under which nontrivial asymptotic inference is possible. Specifically, we concentrate on the situation where the optimal Bayes rule has non-vanishing asymptotic power. As shown in the proof of Theorem 2.1 the power of the Bayes rule diminishes when p is getting smaller, which will be balanced by increasing the sample size. Lemma 2.1 shows that nontrivial asymptotic inference is possible only if the sequence $\frac{\log p}{n}$ remains bounded.

Lemma 2.1. *Assume that the distribution ν of true signals under the alternative is such that $\nu(0, \infty) > 0$ and $\nu(-\infty, 0) > 0$. Moreover, assume that $p = p_m$ converges to zero and that the loss function satisfies $\log \delta = o(\log p)$. Then the asymptotic power of the Bayes classifier is positive only if the sequence $\frac{\log p}{n}$ remains bounded.*

Proof. The proof is given in Section A of the Appendix. □

Remark 2.3. We will see later that under the conditions of Theorem 2.1 it actually holds that the asymptotic power of the Bayes classifier is positive if and only if the sequence $\frac{\log p}{n}$ remains bounded. However, at this point the Lemma mainly serves as a motivation for the asymptotic framework adopted in Assumption (A).

Corollary 2.1. *Assume that the distribution ν of true signals under the alternative is such that $\nu(0, \infty) > 0$ and $\nu(-\infty, 0) > 0$. Moreover, assume that $p \propto m^{-\beta}$ for some $\beta > 0$ and that the loss function satisfies $\log \delta = o(\log m)$. Then the asymptotic power of the Bayes classifier is positive only if the sequence $\frac{\log m}{n}$ remains bounded.*

Remark 2.4. Corollary 2.1 specifies the relationship between the number of tests and the sample size under which the asymptotic power of the Bayes classifier does not vanish in the limit. Similar conditions appear quite often in the literature on high dimensional inference. For an example in the context of multiple testing see [17] as discussed in Remark 2.2. Similarly, in the context of model selection the popular Lasso [43] has asymptotically optimal tuning parameter $\lambda \propto \sqrt{\log m/n}$, where m is the total number of regressors. As discussed in [8] Lasso is consistent if the vector of true regression coefficients satisfies $\|\beta\|_1 = o(\sqrt{n/\log m})$. So, if $\|\beta\|_1$ stays constant then for consistency one needs $\frac{\log m}{n} \rightarrow 0$.

Motivated by the above considerations we will investigate the performance of the multiple testing rules under the following asymptotic assumption.

Assumption (A): $p \rightarrow 0$, δ is bounded from above and such that $\log \delta = o(\log p)$ and $\frac{-2 \log p}{n} \rightarrow C$, where $0 \leq C < \infty$.

The second set of assumptions used throughout this paper imposes a mild restriction on the measure ν .

Assumption (B): Let $T := \sigma\sqrt{C}$, where C is the constant defined in Assumption (A). Assume that there exists $\epsilon > 0$ such that ν has a positive bounded density ρ with respect to the Lebesgue measure on $[-T - \epsilon, -T + \epsilon]$ and $[T - \epsilon, T + \epsilon]$. In case of $C = 0$ it is further assumed that $\rho(0^-) := \lim_{\mu \uparrow 0} \rho(\mu)$ and $\rho(0^+) := \lim_{\mu \downarrow 0} \rho(\mu)$ both exist and are finite and positive.

Remark 2.5. In principle this assumption requires only that the measure ν has a positive density in the neighborhood of one or two points, depending on the constant C specified in Assumption (A). Thus it is for example satisfied whenever ν has a positive, bounded density on the real line. Note that in case when $C = 0$ the Assumption (B) is satisfied whenever ν has a positive density at $\mu = 0$ (i.e. very small signals are probable).

Using the techniques from [25] it can be shown that under Assumptions (A) and (B) the threshold values of the Bayes oracle satisfy

$$c_a = -\sqrt{\log v} (1 + o_m) \text{ and } c_b = \sqrt{\log v} (1 + o_m), \quad (2.6)$$

where $v := n\delta^2/p^2$ (the formula is derived in Section B of the Appendix).

The risk for a multiple testing rule is computed under the additive loss of individual tests simply as the sum of the risks of individual tests. A multiple testing rule using for given m the same critical values for all m test statistics will be called henceforth a fixed threshold rule. Typically the acceptance region

of such a rule will be an interval of the form

$$(\tilde{c}_a, \tilde{c}_b) = \left(\frac{\sqrt{n}\tilde{a}_m}{\sigma}, \frac{\sqrt{n}\tilde{b}_m}{\sigma} \right), \tag{2.7}$$

where \tilde{c}_a, \tilde{c}_b are the critical values for the scaled test statistics Z_i , and \tilde{a}_m, \tilde{b}_m are the critical values for the empirical means \bar{X}_i .

Note that for the specified mixture model (2.3) type I error rates t_1 and type II error rates t_2 of fixed threshold rules are identical for each individual test. The corresponding risk is therefore defined as

$$R = R_1 + R_2 = m(1 - p)t_1\delta_0 + mpt_2\delta_A. \tag{2.8}$$

In the following theorem we provide the asymptotic risk R_{opt} of the Bayes oracle.

Theorem 2.1. *Under Assumptions (A) and (B) the risk obtained by the Bayes rule (2.4) takes for $C = 0$ the form*

$$R_{opt} = mp\delta_A\sigma\sqrt{\frac{\log v}{n}} (\rho(0^-) + \rho(0^+)) (1 + o_m) \tag{2.9}$$

whereas for $0 < C < \infty$

$$R_{opt} = mp\delta_A \nu(-T, T)(1 + o_m). \tag{2.10}$$

Proof. The proof combines techniques of [25] and [6], and is given in Section C of the Appendix. □

Remark 2.6. Similarly as in [6], the asymptotics of the optimal Bayes risk is determined by the cost associated with the Type II error. From the proof of Theorem 2.1 it immediately follows that the Bayes rule has asymptotic power $1 - \nu(-T, T)$.

Remark 2.7. In a setting of sparsity one might consider the question how the trivial rule of not rejecting any hypothesis performs. This procedure has the risk $R_0 = mp\delta_A$, which in case of $C > 0$ can be rewritten as $R_0 = \frac{R_{opt}}{\nu(-T, T)} (1 + o_m)$. Therefore when the asymptotic power of the Bayes rule is close to zero the trivial rule is doing almost as well as the Bayes rule, and it will be difficult for any other statistical procedure to perform reasonably well. From a practical perspective the following results for $C > 0$ are therefore more meaningful when the asymptotic power of the Bayes rule is sufficiently larger than zero.

2.3. Asymptotic Bayes optimality under sparsity

The Bayes classifier defined in (2.4) and in (2.5) requires the full knowledge of the mixture distribution. Since this information is usually not available, in practice one needs to consider other multiple testing rules. This section provides

a theorem characterizing the set of fixed threshold rules, which asymptotically attain the optimal risk.

Definition. A multiple testing rule is called asymptotically Bayes optimal under sparsity (ABOS) if its risk R satisfies $\frac{R}{R_{opt}} \rightarrow 1$ under Assumption (A).

Theorem 2.2. Consider the fixed threshold testing rule which accepts H_{0i} if and only if $Z_i \in (\tilde{c}_a, \tilde{c}_b)$, with $\tilde{c}_a < 0$ and $\tilde{c}_b > 0$. Under Assumptions (A) and (B) this rule is ABOS if and only if

$$\tilde{c}_a = -\sqrt{\log v + z_a} \quad \text{and} \quad \tilde{c}_b = \sqrt{\log v + z_b} \quad (2.11)$$

where

$$z_a = o(\log v), \quad z_b = o(\log v), \quad (2.12)$$

and

$$\lim_{m \rightarrow \infty} z_a + 2 \log \log v = \infty, \quad \lim_{m \rightarrow \infty} z_b + 2 \log \log v = \infty. \quad (2.13)$$

The proof is provided in Section D of the Appendix.

The following corollary gives a simple multiple testing rule, which is asymptotically optimal in case of extreme sparsity, where the number m of tests increases to infinity, but the expected number of true signals mp remains constant or increases only very slowly with m .

Corollary 2.2. If for $n = n_m$ the extreme sparsity assumption

$$mp \rightarrow s \in (0, \infty], \quad \frac{\log(mp)}{\log(n/p^2)} \rightarrow 0, \quad (2.14)$$

is fulfilled, then thresholds of the form

$$\tilde{c}_a = -\sqrt{\log(nm^2) + d_a}, \quad \tilde{c}_b = \sqrt{\log(nm^2) + d_b}, \quad d_a, d_b \in R, \quad (2.15)$$

yield multiple testing rules which are ABOS.

If additionally $\log n = o(\log m)$ then the universal threshold of [14]

$$\tilde{c}_a^2 = \tilde{c}_b^2 = 2 \log m + d, \quad d \in R \quad (2.16)$$

is ABOS.

Proof. Simply observe that $z_1 = \log(nm^2) + d$ and $z_2 = 2 \log m + d$ fulfill the requirements of Theorem 2.2 under the assumption of the corollary. \square

Remark 2.8. Corollary 2.2 states that under extreme sparsity and when $\log n = o(\log m)$ the universal threshold $2 \log m$ of [14] is ABOS. However, when $\log n$ increases to infinity at least at the rate $\log m$ then the universal threshold needs to be supplemented by the $\log n$ term to preserve its ABOS property. Similar criteria in the context of model selection in multiple regression have been proposed and discussed e.g. in [22, 3] and [7].

3. ABOS of some popular multiple testing rules

In this Section we will present the main results of this article: First we give conditions under which the FDR controlling procedure BH and its corresponding step down procedure SD at nominal FDR level α are ABOS. Then we will consider the Bonferroni correction at nominal FWER level α . Specifically ABOS of BH and SD is derived under the following Condition (Op) on α :

Condition (Op): $\alpha \rightarrow \alpha_\infty < 1$, $\frac{\log \alpha}{n} \rightarrow 0$ and

$$\alpha = \frac{v^{1/(2(z_m-1))}}{p} \quad \text{with} \quad z_m \rightarrow 0 \quad \text{and} \quad z_m \log v + \log \log v \rightarrow \infty . \quad (3.1)$$

This condition is necessary and sufficient for the asymptotic optimality of rules based on the Genovese-Wasserman approximation to BH (see Theorem 3.1), and of fixed threshold rules controlling the Bayesian False Discovery Rate, which are discussed in Section E of the Appendix. In Theorem 3.2 Condition (Op) is sufficient for ABOS of BH and SD.

Condition (Op) is rather technical and not so easy to interpret. However, we decided to leave it in this form since it encompasses several important and simple examples, formulated in the corollaries after Theorem 3.2. To get a better understanding of the meaning of Condition (Op) consider the following consequences which are easy to check (it is always assumed that Assumption (A) holds):

- a) Condition (Op) is satisfied if $\alpha \propto n^{-1/2}$. Simply set $z_m = 0$.
- b) Condition (Op) cannot be fulfilled when both $\alpha = \text{const}$ and $\delta = \text{const}$.
- c) If $n \propto -\log p$ and $\delta \rightarrow 0$ under the restrictions of Assumption (A), then Condition (Op) is satisfied for $\alpha = \text{const}$.
- d) If $n \propto -\log p$ and $\delta = \text{const}$, then Condition (Op) is satisfied for $\alpha \rightarrow 0$ such that $\log \alpha = o(\log p)$.

Assertion a) delivers perhaps the most important consequence, that our optimality results will hold for nominal levels $\alpha \propto n^{-1/2}$. Assertions c) and d) are concerned with signals being at the verge of detectability, where it is possible to get optimality for constant α , but only when letting δ getting smaller with growing m . From a practical perspective this is quite reasonable. It means that the relative cost of missing a true signal is increased when the signals become more sparse. If one wants to consider only $\delta = \text{const}$, which includes the simple misclassification rate, then the last assertion states that at the verge of detectability it is possible to attain optimality for $\alpha \rightarrow 0$ at a rate which is substantially slower than $\alpha \propto n^{-1/2}$.

3.1. FDR controlling procedures

The Benjamini–Hochberg rule [2], which we will also call step-up FDR controlling procedure, is defined as follows: Denote the ordered p-values p_i of the scaled

test statistics Z_i as $p_{[1]} \leq p_{[2]} \leq \dots \leq p_{[m]}$. For the step-up procedure at the FDR level α compute

$$k_F := \max \left\{ i : p_{[i]} \leq \frac{i\alpha}{m} \right\} \quad (3.2)$$

and reject the k_F hypotheses with p-values smaller or equal $p_{[k_F]}$.

In our asymptotic analysis we will also consider the corresponding step-down procedure (SD) at level α . For this one needs to compute

$$k_G := \min \left\{ i : p_{[i]} > \frac{i\alpha}{m} \right\} \quad (3.3)$$

and reject the $k_G - 1$ hypotheses with p-values smaller than $p_{[k_G]}$. It is known that under sparsity both procedures behave very similar (see [1]). However, the proof of ABOS for SD turns out to be slightly more challenging than for BH.

The proof of optimality results for the step-up FDR controlling rule in [6] and [33] relied upon the definition of a random threshold for the BH rule

$$c_{BH} = \min\{c_{Bon}, \tilde{c}_{BH}\} \text{ ,}$$

with

$$\tilde{c}_{BH} = \inf \left\{ y : \frac{2(1 - \Phi(y))}{1 - F_m^{(BH)}(y)} \leq \alpha \right\} \text{ .} \quad (3.4)$$

Here $1 - F_m^{(BH)}(y) = \#\{|Z_i| \geq y\}/m$. A detailed motivation of (3.4) is given in [15].

Alternatively let us denote $1 - F_m^{(SD)}(y) = \#\{|Z_i| > y\}/m$. Similarly as in case of BH it is easy to check that SD rejects the null hypothesis H_{0i} if and only if $|Z_i| \geq c_{SD}$, where

$$c_{SD} = \sup \left\{ y : \frac{2(1 - \Phi(y))}{1 - F_m^{(SD)}(y) + 1/m} > \alpha \right\} \text{ .} \quad (3.5)$$

It was proved by Genovese and Wasserman (GW) in [21] that for fixed p , as the number of tests increases, the random threshold c_{BH} can be approximated by the non-random threshold

$$c_{GW} : \frac{2(1 - \Phi(c_{GW}))}{1 - F(c_{GW})} = \alpha \text{ ,} \quad (3.6)$$

where $F(y) = P(|Z_1| \leq y)$. From Corollary F.1 of Appendix F it immediately follows that for any $\alpha \in (0, 1)$ there exists a unique positive solution of (3.6).

Figure 1 illustrates the thresholds c_{BH} , c_{SD} and c_{GW} . Comparing \tilde{c}_{BH} and c_{SD} with c_{GW} the only change is in replacing the cumulative distribution function of $|Z_i|$ by the corresponding empirical distribution function. Thus, for fixed p and m converging to infinity, c_{SD} will also converge to c_{GW} . Approximations

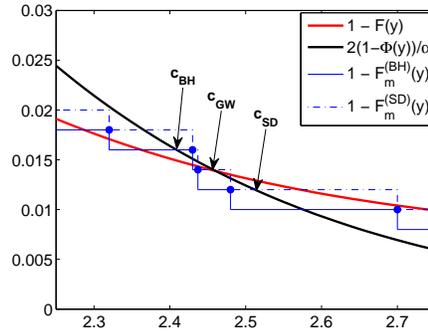


FIG 1. Comparison of the random thresholds c_{BH} and c_{SD} with the nonrandom threshold c_{GW} .

of c_{BH} by c_{GW} in case of sparsity were given in [6]. In Section G of the Appendix we consider the approximation of c_{SD} with c_{GW} under our asymptotic assumptions. A much simpler result is that under sparsity the multiple testing rule based on the Genovese-Wasserman threshold c_{GW} is ABOS.

Theorem 3.1. *Suppose Assumptions (A) and (B) are true. Consider the rule rejecting the null hypothesis H_{0i} if $\frac{\sqrt{n}|\bar{X}_i|}{\sigma} \geq |c_{GW}|$. This rule is ABOS if and only if the Condition(Op) holds.*

Proof. The proof of Theorem 3.1 appears in Section F of the Appendix. □

The next theorem provides the optimality result of BH and SD for generally distributed effect sizes under the alternative.

Theorem 3.2. *In addition to Assumptions (A),(B) assume that*

$$mp \rightarrow s \in (0, \infty]. \tag{3.7}$$

For the denser case

$$p > \frac{(\log m)^{\gamma_1}}{m}, \text{ for some constant } \gamma_1 > 1, \tag{3.8}$$

the additional assumptions

$$n \leq m^{\gamma_2}, \text{ for some } \gamma_2 > 0, \text{ and } \frac{\log \log m}{\log(p \alpha)} \rightarrow 0 \tag{3.9}$$

should hold. Then both BH and SD are ABOS if their nominal FDR levels α satisfy the Condition (Op) of Theorem 3.3.

Proof. The proof is given in Section G of the Appendix. □

Remark 3.1. The upper bound on m provided in the second condition of (3.9) is not very restrictive. Specifically, it is satisfied whenever $p \propto m^{-\beta}$ with $\beta \in (0, 1]$. For p decreasing to 0 at a slower rate (for example like $(\log m)^{-1}$) one can replace this bound with the condition

$$n \geq m^{\gamma_3} \quad \text{for some } \gamma_3 > 0 . \quad (3.10)$$

The following Corollaries are easy consequences of Theorem 3.2.

Corollary 3.1. *Suppose Assumptions (A) and (B) hold. If $p = m^{-\beta}$ with $\beta \in (0, 1]$ and $n \leq m^{\gamma_2}$, for some $\gamma_2 > 0$, then BH and SD at FDR level $\alpha \propto n^{-1/2}$ are ABOS. Note that in this case the condition $\log \delta = o(\log p)$ in Assumption (A) reduces to $\log \delta = o(\log m)$.*

Corollary 3.2. *Suppose Assumptions (A) and (B) hold. If $p = m^{-\beta}$ with $\beta \in (0, 1]$, $n \propto \log m$ and δ converges to zero then BH and SD at a fixed FDR level $\alpha \in (0, 1)$ are ABOS. If $\delta = \text{const}$ then BH and SD are ABOS if α converges to 0 such that $\log \alpha = o(\log m)$.*

Remark 3.2. Corollary 3.1 states that under some mild restrictions on δ BH and SD at the FDR level $\alpha \propto n^{-1/2}$ are ABOS for the whole range of sparsity parameters $p = m^{-\beta}$ with $\beta \in (0, 1]$. This illustrates that these rules adapt to the unknown sparsity parameter. Corollary 3.2 says that in case when $n \propto \log m$ and under the additional requirement that δ converges slowly towards zero, BH and SD at the fixed FDR level $\alpha \in (0, 1)$ are also ABOS. This result substantially extends the results of [6] to the case where the prior on μ_i is fixed and not normal, while the sample size n slowly increases to infinity. This additionally justifies the use of the fixed FDR level for BH in many applications, like e.g. in bioinformatics, where n is much smaller than m . As discussed in [6] the condition $\delta \rightarrow 0$ is quite reasonable in this context, since the cost of missing a true positive is usually large if p is very small. Note that Condition (Op) does not allow to keep both δ and α fixed at the same time.

Remark 3.3. Our results on the optimal FDR level are closely related to the well known fact that for fixed p the significance level of the Bayes test is roughly proportional to $n^{-1/2}$ (see e.g. [11]). Interestingly, in the context of multiple testing under moderate sparsity (i.e. when $\beta \in (0, 1)$) the Bayesian classifier controls rather FDR than FWER.

3.2. Bonferroni correction

In applied sciences the most popular multiple testing procedure is still the Bonferroni correction. For the scaled test statistics $Z_i = \frac{\sqrt{n}\bar{X}_i}{\sigma}$ one computes p-values $p_i = 2(1 - \Phi(|Z_i|))$. The test decision is based on comparing these p-values with the corrected significance level α/m , where in case of $p_i < \alpha/m$ one rejects the null hypothesis.

Equivalently this procedure can be viewed as a fixed threshold rule, where the corresponding critical value c_{Bon} for the test statistic Z_i is defined by

$$1 - \Phi(c_{Bon}) = \frac{\alpha}{2m}.$$

The procedure controls the family wise error rate at level α .

Note that the threshold of the Bonferroni correction is determined by the choice of the nominal FWER α . So, one can expect that by using an appropriate choice of α one can obtain an optimal rule. Unfortunately, the threshold of the optimal Bayes classifier depends on n , δ and the unknown sparsity level p . As a consequence the optimal FWER α will also depend on these parameters. The following Theorem 3.3 states that under the conditions on α for which BH is optimal the Bonferroni correction is ABOS only in the extremely sparse case, where p is roughly proportional to $\frac{1}{m}$.

Theorem 3.3. *Suppose Assumptions (A), (B) and the Condition (Op) for $\alpha = \alpha_m$ are satisfied. The Bonferroni procedure at FWER level α is ABOS if and only if sparsity condition (2.14) holds.*

Proof. The proof of Theorem 3.3 is provided in Section H of the Appendix. The proof relies upon technical results concerning the relationship between BFDR controlling rules and the Bayes classifier, which are discussed in Appendix Section E. □

Remark 3.4. Our results show that under Condition (Op) the Bonferroni correction is ABOS only under extreme sparsity, while BH is ABOS under a much wider range of sparsity. This is due to the fact that the Bonferroni procedure has larger Type II error than BH. The Bonferroni-Holm procedure serves as a popular alternative, which is more powerful than the Bonferroni procedure while still controlling FWER. The Bonferroni-Holm procedure works like the SD procedure (3.3), but the ordered p-values are compared with $\alpha/(m - i + 1)$ instead of $i\alpha/m$. Thus the Holm procedure is sandwiched between Bonferroni and SD, and one might ask if it is ABOS for a wider range of sparsity parameters than the Bonferroni correction. Unfortunately this is not the case. The gain in power from the Bonferroni-Holm procedure comes from an implicit estimation of $1 - p$, but this does not bring any advantage when $p \rightarrow 0$.

The following proposition gives conditions on α under which the Bonferroni rule is ABOS also for the denser case.

Proposition 3.1. *Suppose $p \propto m^{-\beta}$, $\delta = \text{const}$ and $\frac{2\beta \log m}{n} \rightarrow C \in [0, \infty)$. Then for any $\beta \in (0, 1]$ the Bonferroni rule at nominal level α is ABOS under the condition*

$$\alpha \propto \frac{m^{1-\beta}}{\sqrt{n}}. \tag{3.11}$$

The proof is given in Appendix H. For the ease of presentation the proposition is presented with the simple sufficient condition (3.11). However, the condition indicates that the significance level α of the Bonferroni rule depends on the

unknown sparsity parameter β . In that sense, contrary to the FDR controlling procedures, the Bonferroni rule does not adapt to the unknown level of sparsity.

4. Simulations

The purpose of this simulation study is to illustrate the theoretical results of Section 3. We focus particularly on the implications of Corollary 3.1 and 3.2. For various scenarios we generated independent test statistics according to the mixture model (2.3). The performance of different multiple testing procedures under sparsity is compared with the Bayes oracle, where the risk ratio is estimated for a given loss function based on 2000 simulation runs for each scenario.

The following features are analyzed systematically:

1. *Multiple testing rules*: We consider the Bonferroni rule and the Benjamini-Hochberg procedure. Initially we also compared two further procedures, namely the step-down procedure (3.3) as well as the asymptotically optimal rejection curve from [19]. However, under sparsity the performance of both procedures is almost indistinguishable from BH, and therefore these results are not shown. In our simulation study we also included the empirical Bayes multiple testing procedure from Johnstone and Silverman [26, 27]. As shown in [26] this procedure adapts to the unknown sparsity and provides an asymptotically minimax thresholding rule for the estimation of unknown means.

The effect of different *significance levels* for Bonferroni (*FDR levels* for BH) was studied. In particular we considered fixed nominal levels $\alpha = 0.05$ and $\alpha = 0.1$, as well as $\alpha \propto n^{-1/2}$ (where $\alpha = 0.1$ for $n = 5$). In the manuscript we only present results for fixed $\alpha = 0.1$, and in case of BH for $\alpha \propto n^{-1/2}$.

The empirical Bayes approach of [26] was based on a Laplace prior, with density

$$f_L(\mu) = \frac{1}{2\tau} \exp(-|\mu|/\tau), \quad (4.1)$$

which performed best in the simulations of Johnstone and Silverman. Note that the Laplace prior is heavy tailed and spiked at zero, which makes it very well suited for detection of signals under the assumption of sparsity. This prior is often used in the context of sparse model selection and is directly related to the L_1 penalty applied in Lasso [43]. In accordance with the presentation in [26] the procedure EB uses the scaling parameter $\tau = 2$, for which the variance of the Laplace distribution is equal to 8. The procedure EB2 estimates the scaling parameter τ (see [27] for details).

2. *Known and unknown error variance σ^2* : In all simulations the error variance term was $\sigma^2 = 1$. We performed analysis both for known σ (z-tests) as well as under the assumption that σ is not known (t-tests). To apply the empirical Bayes approach of [26] in case of unknown σ we transformed t-test statistics into z-scores, yielding test statistics which are standard normal under the null hypothesis.

3. For the sample size n and the number of tests m two prototypic scenarios are considered.

- *Polynomial case:* $n \in \{5, 10 \dots, 200\}$, and $m = 0.8 \cdot n^2$:
 $m \in \{20, 80, 180, 320, 500, 720, 980, 1280, \dots, 32000\}$
- *Exponential case:* $n \in \{5, 10 \dots, 40\}$, and $m = \text{round}(5 \cdot (1.3)^n)$:
 $m \in \{19, 69, 256, 950, 3528, 13100, 48639, 180594\}$

For $n \in \{5, 10\}$ the difference between the polynomial and the exponential scenario is quite small. But for larger n the number of tests in the exponential case gets much larger, and leads to the situation referred to as ‘being on the verge of detectability’. Results for the polynomial case are presented in Figure 2 and Figure 3, results for the exponential case in Figure 4 and Figure 5.

4. Simulations were run for two different *sparsity levels*:

$$p \propto m^{-\beta} \text{ for } \beta = 1, \text{ and } \beta = 1/4 .$$

For the case of extreme sparsity $\beta = 1$ we used the proportionality constant 8, which yields for $n = 5$ the sparsity parameter $p = 0.4$ in the polynomial case, and $p = 0.4211$ in the exponential case. For the denser case $\beta = 1/4$ we wanted sparsity for $n = 5$ to be the same as for the case of extreme sparsity, which is achieved using proportionality constants 0.8459 for the polynomial case, and 0.8791 for the exponential case.

5. Concerning the *loss function* mainly the usual 0 - 1 loss is studied, which corresponds to $\delta = 1$ and results in the Bayes risk being equal to the misclassification rate. In view of Corollary 3.2 we analyze for the exponential case also the loss $\delta = 1/n$, which corresponds to a loss ratio $\delta \propto (\log(m/5))^{-1}$ (with proportionality constant $\log(1.3)$).

6. *Distribution of effect sizes under the alternative:* Apart from the standard normal distribution we considered an asymmetric double exponential (ADE) distribution with density

$$f_{ADE}(\mu) = \begin{cases} \lambda_1/2 \exp(\lambda_1\mu) & \mu \leq 0 \\ \lambda_2/2 \exp(-\lambda_2\mu) & \mu > 0 \end{cases} \quad \lambda_1 > 0, \lambda_2 > 0 . \quad (4.2)$$

The particular choice of ADE is motivated two-fold. In [35] it was suggested to make use of the asymmetric Laplace distribution to model microarray data. While [35] models the data directly with ADE it appears to be reasonable to consider a model where the effect sizes under the alternative are stemming from ADE. In statistical genetics there is a long history to model effect sizes using Laplace distributions, and ADE is a natural asymmetric generalization. On the other hand ADE is simple enough to allow to compute the thresholds of the Bayes rule (details about this are given below). Results are presented for the specific choice $\lambda_1 = 1.5$ and $\lambda_2 = 3$.

Remark 4.1. The scenario with normally distributed mean under the alternative $\mu \sim \mathcal{N}(0, \tau^2)$ coincides with the scale mixture model discussed in [6].

The ADE as a generalization of the Laplace distribution gives an example of a distribution with heavier tails. If we make the specific choice $\lambda_1 = c_1/\tau$ and $\lambda_2 = c_2/\tau$, then ADE can be viewed as a scaling family with scaling parameter τ . While in [6] the focus was on normal scaling mixture models and asymptotic analysis was performed in terms of the parameter τ^2/σ^2 , this article studies the more general class of distributions characterized by Assumption (B), and the asymptotic analysis is driven by the sample size n . However, for scaling families it is easy to take into account both the influence of effect size τ^2/σ^2 and the influence of sample size n , by working with the parameter $u = n\tau^2/\sigma^2$. The following simulations are performed with $\sigma = 1$, and for the normal mixture model τ is kept fixed at $\tau = 1$, a value which guarantees that the Bayes oracle has reasonable power to detect signals within the considered range of n and m . The presented results are representative for any other choice of n, σ and τ yielding the same parameter $u = n\tau^2/\sigma^2$.

For the normal scaling mixture model one readily computes that threshold values of the Bayes oracle for \bar{X}_i are given by $b_l = -\sqrt{c_N}$ and $b_r = \sqrt{c_N}$ with

$$c_N^2 = \frac{1+u}{u}(\log(1+u) + 2\log(\delta(1-p)/p)). \quad (4.3)$$

For ADE threshold values of the Bayes oracle have no closed form expression. However, for (4.2) the optimal threshold values $b_l < 0$ and $b_r > 0$ of \bar{X}_i are quite easily obtained as numerical solutions of

$$\begin{aligned} & \lambda_1/2 \exp(\lambda_1^2/2n + \lambda_1 x) \Phi(-\sqrt{n}(x + \lambda_1/n)) \\ + & \lambda_2/2 \exp(\lambda_2^2/2n - \lambda_2 x) \Phi(\sqrt{n}(x - \lambda_2/n)) \\ = & \frac{\sqrt{n}}{\sqrt{2\pi}} \exp(-nx^2/2) \delta(1-p)/p. \end{aligned}$$

Figure 2 illustrates the performance of multiple testing procedures for polynomial growth rates of m with respect to n , in case of known error variance σ^2 . In the upper two panels (a) and (b) the mean under the alternative is simulated from a standard normal distribution $\mu \sim \mathcal{N}(0, 1)$, whereas the lower two panels (c) and (d) relate to the asymmetric double exponential (4.2). Panels of the left column (a) and (c) show the extremely sparse case $\beta = 1$, whereas (b) and (d) of the right column show the denser case $\beta = 1/4$.

The first observation is that the estimates of the risk ratio for the extremely sparse case in Figure 2 (a) and (c) show much more fluctuation due to simulation error than estimates in panels (b) and (d). This is not surprising because under extreme sparsity the expected number of true signals does not increase with m , which makes it difficult to accurately estimate the mixture density. In accordance with our theoretical results Bonferroni correction is doing pretty well in case of extreme sparsity, where the risk ratio rapidly converges towards the optimal value 1. In contrast for the denser case the Bonferroni rule is way too conservative and its risk ratio is an increasing function of n .

The BH procedure with fixed FDR level $\alpha = 0.1$ (abbreviated as BH01) is getting too liberal for large n , and its risk ratio starts growing. However, in

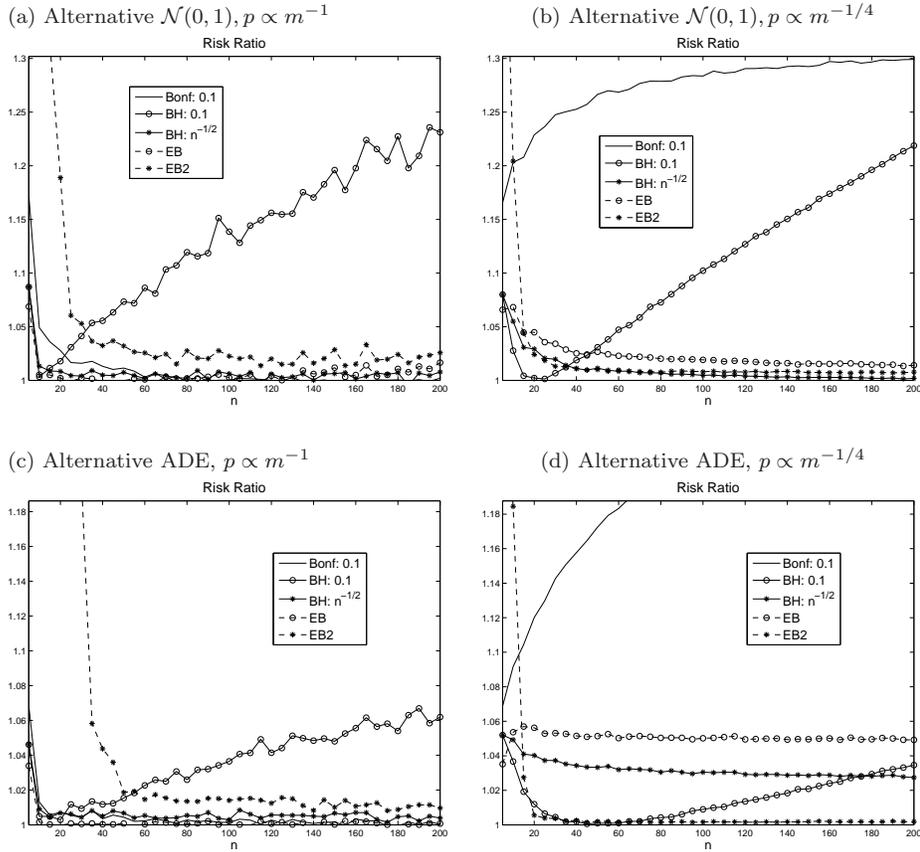


FIG 2. Risk ratio in case of known σ for the polynomial growth rate of m with respect to n as described in the main text. The loss ratio $\delta = 1$ is kept fixed.

Figure 2 (b) BH01 seems to attain optimal risk for $n = 20$. What actually happens here is that FDR of the Bayes oracle is a decreasing function in n , with $\text{FDR} \simeq 0.2$ for $n = 5$ and $\text{FDR} \simeq 0.03$ for $n = 200$. Roughly at $n = 20$ FDR of the Bayes oracle is close to the nominal FDR level 0.1 of BH01, resulting in the minimum of the risk ratio. In Figure 2 (d) the same observation can be made at $n = 50$. Much better performance for large n has the BH procedure at FDR level proportional to $n^{-1/2}$ (abbreviated as BHn). Both in case of extreme sparsity as well as for the denser case its risk ratio converges towards 1, which confirms the theoretical results from Corollary 3.1.

The observed behavior of risk ratios for Bonferroni, BH01 and BHn is qualitatively quite similar for normal alternatives (panels (a) and (b)) and ADE (panels (c) and (d)), though there are obviously big quantitative differences. In particular in Figure 2 (d) the risk ratio of BHn with decreasing FDR level con-

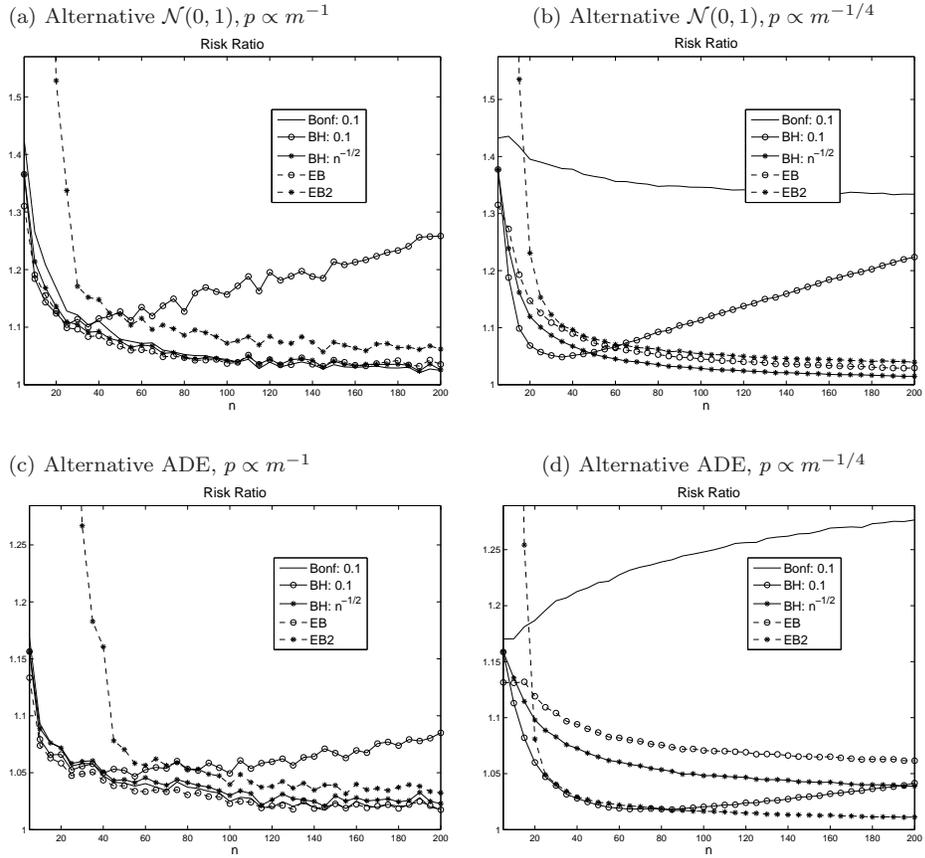


FIG 3. Risk ratio in case of *unknown* σ for the *polynomial growth rate* of m with respect to n as described in the main text. The loss ratio $\delta = 1$ is kept fixed.

verges rather slowly towards 1. This is mainly because in this scenario the Bayes oracle is much more liberal than BHn. We observed also in other simulation scenarios (not presented here), that in case of very large effect sizes (large variance of the alternative distribution) BHn is more conservative than the Bayes oracle, and then only for very large values of m and n the risk ratio gets close to 1. However, in such an “easy” situation the misclassification rate of BHn will be rather small, only the Bayes oracle will still perform that much better.

Concerning the empirical Bayes approach our simulation study confirms very good results reported in [26]. Both EB and EB2 perform quite well under all scenarios presented in Figure 2. Qualitatively EB performs similarly to BHn. For the extremely sparse case the risks of these procedures are very similar. In the denser case EB has a substantially larger risk than BHn, but still seems to be ABOS. In the denser case (panels (b) and (d)) EB2 performs substantially

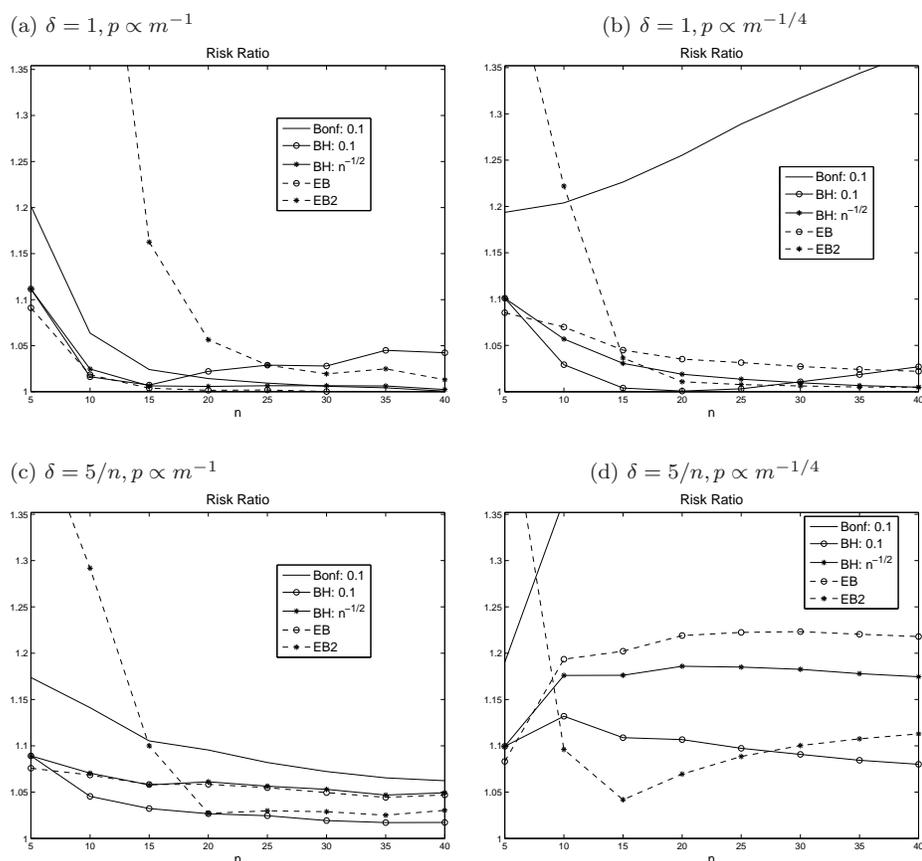


FIG 4. Risk ratio in case of *known* σ for the *exponential growth rate* of m with respect to n as described in the main text. Distribution of μ under the alternative is in all scenarios $\mathcal{N}(0,1)$. Asymptotic power of the Bayes rule is 0.47 for $\beta = 1$ and 0.72 for $\beta = 1/4$.

better than EB as long as $m > 300$. In particular for ADE (or in general for alternative distributions with large variance) EB2 is performing extremely well in the denser case. However, for smaller m EB2 has difficulties to estimate p correctly, leading to a rather large misclassification rate. Also, in the extremely sparse case EB2 needs as least $m > 1000$ tests to be able to estimate both the mixture rate p as well as the scaling parameter τ simultaneously, and even then it tends to have larger risk ratio than BHn.

Figure 3 presents the same simulation results for polynomial growth rate as Figure 2, but now in case of unknown σ . Qualitatively we obtain very similar results as for known σ , with the most striking difference that for small n the risk ratio is generally much larger in case of unknown σ . This is obviously a consequence of the fact that the Bayes oracle has the advantage of knowing σ ,

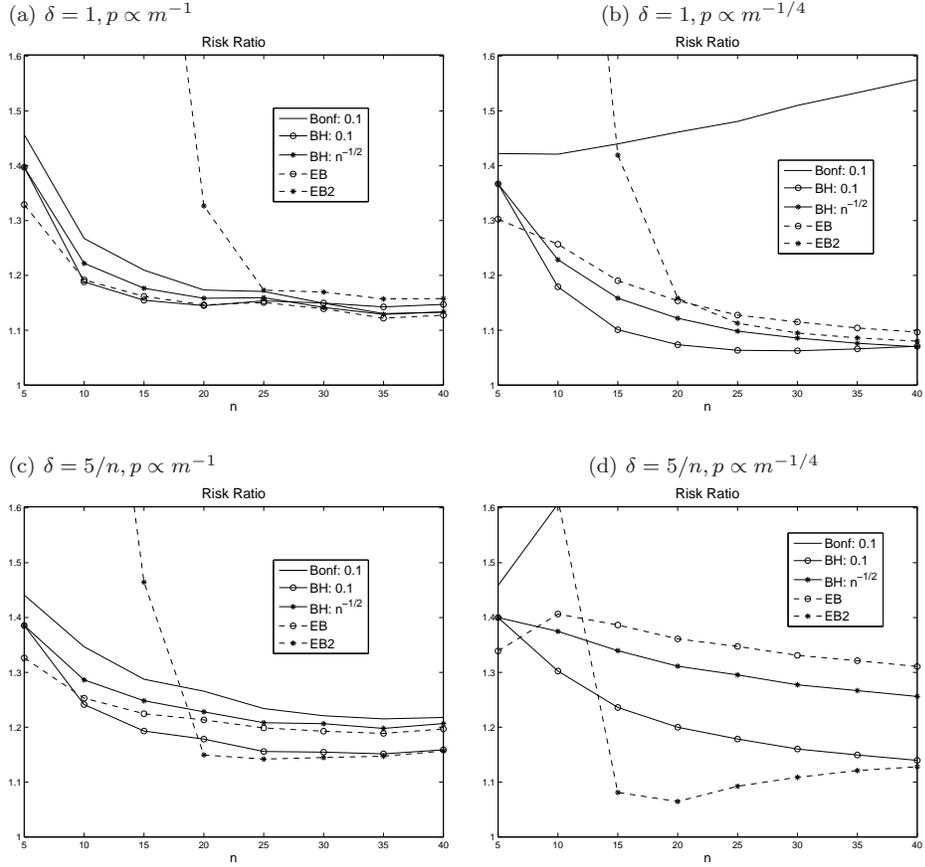


FIG 5. Risk ratio in case of *unknown* σ for the *exponential growth rate* of m with respect to n as described in the main text. Distribution of μ under the alternative is in all scenarios $\mathcal{N}(0, 1)$. Asymptotic power of the Bayes rule is 0.47 for $\beta = 1$ and 0.72 for $\beta = 1/4$.

and for small n it is difficult to estimate σ from the data. The risk ratio of the Bonferroni procedure converges towards 1 only in case of extreme sparsity (Figures 3 (a) and (c)), whereas for BHn convergence towards 1 occurs in all four scenarios.

Figure 4 and Figure 5 refer to simulations at exponential growth rates, which relates to the situation of ‘being on the verge of detectability’. In case of known σ Figures 4 (a) and (b) illustrate again the implications of Corollary 3.1, which states that for fixed δ the BH procedure with $\alpha \propto n^{-1/2}$ is ABOS. Just like in the polynomial case BH01 is for larger n too liberal, and its risk ratio is increasing with n . In contrast the risk ratio of BHn converges to 1 both in panel (a) and (b), resembling very much the results of Figure 2. In Figure 4 (c) and (d) risk ratios are shown at a loss ratio $\delta = n^{-1}$. In that case BH01 with fixed

FDR level performs better than BHn with decreasing α level. Here the risk ratio of BH01 actually appears to be decreasing for large n which corresponds with Corollary 3.2, although convergence towards 1 seems to be rather slow.

Finally in Figure 5 simulation results at the verge of detectability in case of unknown σ are reported. While in the denser case $\beta = 1/4$ (panels (b) and (d)) the qualitative behavior of all risk ratios is fairly similar to the case of known σ in Figure 4, this seems to be no longer true for extreme sparsity (panels (a) and (c)). To a certain extent these findings are in accordance with results of [17], which require $\log m = o(n^{1/3})$ for the accurate uniform approximation of the distribution of t-statistics by a normal distribution.

5. Discussion

In this paper we extend the asymptotic optimality results on FWER and FDR controlling rules from [6] to the practically relevant situation, where the distribution of effect sizes under the alternative does not change with the number of tests, and has an arbitrary positive density over the entire real line. This allows to describe realistic scenarios, where the distribution of effect sizes is not symmetric.

Our results are easiest to describe for a sparsity parameter p of the form $p \propto m^{-\beta}$, with $\beta \in (0, 1]$. When furthermore the ratio of losses for type I and type II error fulfills $\log \delta = o(\log m)$, then nontrivial inference is possible only when the sample size n increases to infinity at least at the rate of $\log m$. We show that in general if the FDR level $\alpha \propto n^{-1/2}$ then BH and the corresponding step-down procedure are optimal for all sparsity levels $\beta \in (0, 1]$. For the limiting case of $n \propto \log m$ the FDR controlling procedures at a fixed level $\alpha \in (0, 1)$ are asymptotically optimal when δ slowly converges to 0. This condition on δ seems justifiable, since typically the cost of missing the true signal gets larger if the proportion of true signals decreases.

We also show that for $\alpha \propto n^{-1/2}$ the Bonferroni correction and the universal threshold of [14] are ABOS in the extremely sparse case of $p \propto 1/m$. Interestingly, the universal threshold remains ABOS also when $\delta = \text{const}$. For the denser case $\beta < 1$ the Bonferroni rule is ABOS when the nominal significance level depends on the sparsity parameter β roughly like $\alpha \propto m^{1-\beta}n^{-1/2}$. Thus contrary to the FDR controlling rules at level $\alpha \propto n^{-1/2}$ the Bonferroni rule does not adapt well to the unknown level of sparsity.

Our results on the optimal FWER and FDR level can be seen as an extension of the well known results that the significance levels of Bayesian tests is roughly proportional to $n^{-1/2}$ (see e.g. [11]). They suggest that in the context of Bayesian multiple testing under sparsity FDR is a better analogue of the type I error than FWER. Specifically, in section E of the Appendix we observe that under our sparsity assumptions the Bayesian classifier controls FDR at the level $(n(\log n + \log m))^{-1/2}$. Additionally, our asymptotic results provide some margin for the FDR level, under which FDR controlling rules are still ABOS.

Our model assumes that $\bar{X}_i|\mu_i \sim N(\mu_i, \sigma^2/n)$. Asymptotic results can also be applied directly to the case when the variances of $X_i|\mu_i$ are known, but

not necessarily equal to each other (assuming that the sequence of variances is bounded and bounded away from zero). Also, as shown in [17], even when the distribution of $X_i|\mu_i$ is not normal and variances are estimated separately for each i , the level of the simultaneous tests based on the t-test statistics $\sqrt{n}(\bar{X}_i|\mu_i - \mu_i)/\hat{\sigma}_i$ is accurate provided $\log m$ increases at a strictly slower rate than $n^{1/3}$. Based on these results we expect that under similar assumptions our asymptotic results can be generalized to the case where the distribution of $X_i|\mu_i$ is not normal and the variances σ_i^2 are estimated separately for every i . We consider this as an interesting topic for further research.

Motivated by the genetic applications from the introduction (for example [16] and [32]) we have focussed essentially on continuous distributions of μ_i (or at least on distributions which are continuous in an environment of $\pm T$). One referee brought up the topic of discretely distributed μ_i , which indeed would need a rather different analysis. Discrete distributions fulfill Condition A_1 of [25], thus when the asymptotic power of the Bayes oracle equals one (that is when $C = 0$) the techniques of [25] can be applied to obtain formulas for the optimal risk analogously to (2.9). From a statistical point of view it is easier to detect signals when the smallest possible effect size is bounded away from 0. Accordingly for discrete distributions the optimal risk decays exponentially with n . It is open to study which effect this has on the conditions under which FDR-controlling rules are ABOS. For signals at the verge of detectability (that is when $C > 0$) we would expect that results for discrete distributions are qualitatively rather similar to the results presented in this paper. However, there remain certain technical difficulties to be solved when the distribution is such that it has no probability mass at $\pm T$.

Another interesting topic for further research is the analysis of the asymptotic optimality properties of the plug-in version of the Bayes oracle, based on the empirical bayes estimates (EB) of the mixture distribution. Encouraging theoretical results from [26] and [9] make us believe that EB procedures are ABOS when $mp \rightarrow \infty$, i.e. when the expected number of true signals increases with m . However, our simulation study suggests that EB procedures might be ABOS even in the extremely sparse case $p \propto 1/m$. Also, based on our simulation study and the results of [5], empirical Bayes or full Bayes multiple testing procedures perform better than BH for somewhat denser scenarios. To distinguish between these rules in our asymptotic context one could consider investigating the rate of convergence of the risk ratio to 1, as it was proposed in [33] in the context of testing within families parametrized by the location or the scale parameter.

Our theory is based on the assumption that the test statistics $\bar{X}_1, \dots, \bar{X}_m$ are independent. However, it can be rather directly applied to the case where the correlation matrix between the test statistics has a diagonal block structure (e.g. correlated SNPs, gene pathways), when one tests the significance of the entire block based on some aggregated test statistic. Then our results hold as long as the distribution of the aggregated test statistics can be approximated by a normal (or chi-square) distribution.

If the correlation structure is more involved, then the Bayes classifier for a given hypothesis will use the information from other test statistics. Depending

on the relationship between m and n such a Bayes classifier can be approximated by full or empirical Bayes methods, which may estimate the dependence structure using some parametric assumptions or one of the (sparse) high-dimensional methods to estimate covariance matrices. We believe that such methods will in general work much better than BH based on individual p-values. It remains to verify under which forms of dependence BH is still ABOS, and to find the conditions under which empirical or full Bayes methods are ABOS. Due to a more complicated form of the Bayes classifier the extension of our results to the case of dependent test statistics will require further development of the proof technique and remains an interesting topic for a further research.

Acknowledgment

We want to thank Arijit Chakrabarti for a careful reading of our proofs, and Jayanta K. Ghosh for discussions and guidance. Furthermore we want to thank two anonymous reviewers whose comments greatly helped to improve the presentation of this article.

This research has been funded by the Vienna Science and Technology Fund (WWTF) through project MA09-007a, and by grant N N201 414139 of the Polish Ministry of Science and Higher Education.

Appendix A: Asymptotic power of the Bayes classifier

Throughout the appendix we will make use of the notation $f = \frac{1-p}{p}$. In this section we present the proof of Lemma 2.1.

Proof. Let $h_n(\mu) = \exp(a_n \frac{\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2})$. Then $(\delta f)^{1/n} = \|h_n\|_{L^n(\nu)}$.

Now, assume that the sequence $\frac{\log p}{n}$ is unbounded. This implies that the sequence $(\delta f)^{1/n} \rightarrow \infty$, and it easily follows that $a_n \rightarrow -\infty$. Similarly, $(\delta f)^{1/n} \rightarrow \infty$ implies that $b_n \rightarrow \infty$.

Now, observe that the type II error of the Bayes oracle is given by $t_2 = \int \Psi_n(\mu) d\nu(\mu)$ with

$$\Psi_n(\mu) = \Phi\left(\frac{\sqrt{n}(b_n - \mu)}{\sigma}\right) - \Phi\left(\frac{\sqrt{n}(a_n - \mu)}{\sigma}\right).$$

It is easy to check that if $a_n \rightarrow -\infty$ and $b_n \rightarrow \infty$ then for every $\mu \in R$ it holds that $\Psi_n(\mu) \rightarrow 1$ and, by the dominated convergence theorem, $t_2 \rightarrow 1$. □

Appendix B: Threshold values of the Bayes oracle, the derivation of equation (2.6)

In this section we will derive equation (2.6). The first Lemma provides asymptotic formulas for the critical points of the Bayes rule for distributions ν satisfying Assumption (B). Lemma B.3 refines these results for the case $C = 0$.

Lemma B.1. *Let Assumptions (A) and (B) hold. Then the critical values converge with limits*

$$a_n \rightarrow -T \text{ and } b_n \rightarrow T .$$

The proof of Lemma 2.1 relies on the following technical result.

Lemma B.2. *Let $a_n \rightarrow a$ be any convergent sequence. Define $h_n(\mu) := \exp(a_n \frac{\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2})$ and $h(\mu) := \exp(a \frac{\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2})$. Then*

$$\lim_{n \rightarrow \infty} \|h_n\|_{L^n(\nu)} = \|h\|_{L^\infty(\nu)} . \quad (\text{B.1})$$

Proof. First note that for all n it holds that $h_n \in L^\infty(\nu)$, and therefore also $h_n \in L^m(\nu), \forall m > 0$. It is easy to check that $\lim_n \|h_n - h\|_{L^\infty(\nu)} = 0$. Thus for any $\epsilon > 0$ and sufficiently large n we have $\|h_n - h\|_{L^n(\nu)} \leq \|h_n - h\|_{L^\infty(\nu)} < \epsilon$. Now (B.1) easily follows by the triangle inequality and the fact that $\lim_{n \rightarrow \infty} \|h\|_{L^n(\nu)} = \|h\|_{L^\infty(\nu)}$. \square

Now we are ready to prove Lemma B.1.

Proof. Let, as before denote $h_n(\mu) = \exp(a_n \frac{\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2})$, where a_n are now the solutions of (2.4). Then $(\delta f)^{1/n} = \|h_n\|_{L^n(\nu)}$ and due to Assumption (A) $\lim_n (\delta f)^{1/n} = e^{C/2}$. Note that a_n has to be bounded, otherwise the sequence $\|h_n\|_{L^n(\nu)}$ could not be bounded. Let a be an accumulation point of a_n . By Lemma (B.2) for any subsequence $a_j \rightarrow a$ it holds

$$\lim_j \|h_j\|_{L^j(\nu)} = \left\| \exp\left(a \frac{\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right) \right\|_{L^\infty(\nu)} . \quad (\text{B.2})$$

Let S denote the support of ν . By (B.2) it holds

$$\begin{aligned} C/2 &= \sup_{\mu \in S} \left(a \frac{\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} \right) \\ &\leq \sup_{\mu \in \mathbb{R}} \left(-\frac{(\mu - a)^2}{2\sigma^2} + \frac{a^2}{2\sigma^2} \right) = \frac{a^2}{2\sigma^2} . \end{aligned} \quad (\text{B.3})$$

Since $a \leq 0$ (B.3) implies that $a \leq -T$.

On the other hand, using the assumption that $-T \in S$ we have

$$C/2 = \frac{T^2}{2\sigma^2} \geq -a \frac{T}{\sigma^2} - \frac{T^2}{2\sigma^2} , \quad (\text{B.4})$$

which implies that $a \geq -T$. Thus (B.3) and (B.4) lead to the conclusion that $a = -T$.

The proof that $b_n \rightarrow T$ goes exactly along the same lines. \square

The following Lemma B.3 specifies the rate at which a_n and b_n converge to zero in case of $C = 0$.

Lemma B.3. *Throughout this proof and the rest of the appendix we will make use of the notation $o_m = o_m(1)$.*

Let Assumptions (A) and (B) hold. If $C = 0$ then the critical values of the Bayes oracle fulfill

$$\sqrt{n}e^{-\frac{na_n^2}{2\sigma^2}} \sim \frac{\sqrt{2\pi}\sigma}{f\delta} \rho(0^-) \tag{B.5}$$

and

$$\sqrt{n}e^{-\frac{nb_n^2}{2\sigma^2}} \sim \frac{\sqrt{2\pi}\sigma}{f\delta} \rho(0^+) . \tag{B.6}$$

Proof. By Lemma B.1 a_n converges to 0. Also, by Assumption (B) there exists $\epsilon > 0$ such that $\nu(\mu)$ has a density $\rho(\mu)$ on the interval $(-\epsilon, \epsilon)$. It is immediately clear that

$$\int_{(\epsilon, \infty)} h_n^n(\mu) d\nu(\mu) \leq e^{-n\frac{\epsilon^2}{2\sigma^2}} \nu(\epsilon, \infty) . \tag{B.7}$$

Also, there exists n_0 such that for every $\mu < -\epsilon$ and $n > n_0$ it holds $a_n\mu < \mu^2/4$ (because $a_n \rightarrow 0$). Thus for $n > n_0$

$$\int_{(-\infty, -\epsilon)} h_n^n(\mu) d\nu(\mu) \leq e^{-n\frac{\epsilon^2}{4\sigma^2}} \nu(-\infty, -\epsilon) . \tag{B.8}$$

Concerning the integral over the interval $(-\epsilon, \epsilon)$, by completion of squares one derives

$$\begin{aligned} \int_{-\epsilon}^{\epsilon} h_n^n(\mu) \rho(\mu) d\mu &= \rho_n \exp\left(\frac{na_n^2}{2\sigma^2}\right) \int_{-\epsilon}^{\epsilon} \exp\left(-n\frac{(\mu-a_n)^2}{2\sigma^2}\right) d\mu \\ &= \rho_n e^{\frac{na_n^2}{2\sigma^2}} \frac{\sqrt{2\pi}\sigma}{\sqrt{n}} [\Phi(\sqrt{n}(\epsilon - a_n)/\sigma) - \Phi(\sqrt{n}(-\epsilon - a_n)/\sigma)] . \end{aligned} \tag{B.9}$$

Here ρ_n is a real number, with $\rho_n \in [\inf_{\mu \in (-\epsilon, \epsilon)} \rho(\mu), \sup_{\mu \in (-\epsilon, \epsilon)} \rho(\mu)]$, where according to Assumption (B) $0 < \inf_{\mu \in (-\epsilon, \epsilon)} \rho(\mu) \leq \sup_{\mu \in (-\epsilon, \epsilon)} \rho(\mu) < \infty$.

Note that $\Phi(\sqrt{n}(\epsilon - a_n)/\sigma) \rightarrow 1$ as well as $\Phi(\sqrt{n}(-\epsilon - a_n)/\sigma) \rightarrow 0$ (because $a_n \rightarrow 0$). Comparing (B.7), (B.8) and (B.9) we observe that the integral over $(-\epsilon, \epsilon)$ dominates the two remaining terms and from (2.4) it follows that

$$1 = \sqrt{\frac{2\pi}{n}} \sigma (f\delta)^{-1} \rho_n \exp\left(\frac{na_n^2}{2\sigma^2}\right) (1 + o_m) .$$

Thus we may conclude that the sequence

$$S_n := (f\delta)^{-1} n^{-1/2} \exp\left(\frac{na_n^2}{2\sigma^2}\right)$$

is bounded and therefore for any convergent subsequence it holds that

$$a_n \sim -\frac{\sigma \sqrt{\log n + 2 \log(\delta f)}}{\sqrt{n}} .$$

To get the exact behavior we further split the domain of the integral in $(-\epsilon, -g_n)$, $(-g_n, 0)$ and $(0, \epsilon)$, where g_n is a positive sequence such that $a_n = o(g_n)$, or more specifically

$$g_n \rightarrow 0 \quad \text{with} \quad \frac{\log n}{ng_n^2} \rightarrow 0, \quad \frac{\log(\delta f)}{ng_n^2} \rightarrow 0. \quad (\text{B.10})$$

For the first interval we get a bound by evaluating the integrand at $-g_n$, for the second and third interval we repeat the computations leading to (B.9) with the corresponding boundaries, and finally obtain

$$\delta f = \int_{-g_n}^0 h_n^n(\mu) \rho(\mu) d\mu (1 + o_m) = \frac{\rho(0^-) \sqrt{2\pi} \sigma}{\sqrt{n}} \exp\left(\frac{na_n^2}{2\sigma^2}\right) (1 + o_m)$$

which yields (B.5). The proof for b_n is exactly the same. \square

Remark B.1. The proof of Lemma B.3 relies upon choosing a suitable sequence g_n . The choice of the sequence g_n strongly depends on the asymptotic behavior of δf . If for example for sufficiently large n , $\delta f \leq n^\alpha$, with $\alpha > 0$, one might use $g_n = \frac{\log n}{\sqrt{n}}$, the choice of Johnson and Truax (1973). Another situation occurs if $\delta f \sim e^{n^{1-\gamma}}$ with $0 < \gamma < 1$, where $g_n = n^{-\gamma/3}$ is a suitable choice.

Remark B.2. As shown in the proof of Lemma B.3, the accuracy of the approximations provided in (B.5) and (B.6) depends on the asymptotic behavior of δf and on the regularity of ρ in a neighborhood of 0. Assuming for example that ρ is one-sided Lipschitz (on both sides of 0) and that δf is polynomially bounded one obtains that the ratio of the right and left-hand sides of (B.5) and (B.6) can be expressed as $1 + z_n$ with $z_n = o(n^{-1/2} \log n)$.

In case of $C = 0$ equation (2.6) then immediately follows from Lemma B.3 where we use the notation $v = n\delta^2/p^2$ and remember that $f = 1/p (1 + o_m)$. Otherwise if $C > 0$ Lemma B.1 provides $a_n \rightarrow -T = -\sigma\sqrt{C}$ and equation (2.6) follows from Assumption (A).

Appendix C: Proof of Theorem 2.1

Notice, that the type II error of the Bayes oracle is given by $t_2 = \int \Psi_n(\mu) d\nu(\mu)$ with

$$\Psi_n(\mu) = \Phi\left(\frac{\sqrt{n}(b_n - \mu)}{\sigma}\right) - \Phi\left(\frac{\sqrt{n}(a_n - \mu)}{\sigma}\right).$$

We will now calculate the asymptotic formula for the type II error in case when $C = 0$. Consider first the integral over $\mu \in (-\infty, 0)$. Remember that $a_n \rightarrow 0$, thus for n sufficiently large ν has a density $\rho(\mu)$ on $(2a_n, 0)$ and it holds that

$$\int_{2a_n}^0 \Psi_n d\nu = \int_{2a_n}^0 \left[\Phi\left(\frac{\sqrt{n}(b_n - \mu)}{\sigma}\right) - \Phi\left(\frac{\sqrt{n}(a_n - \mu)}{\sigma}\right) \right] \rho(\mu) d\mu.$$

Applying the mean value theorem and substitution yields

$$\int_{2a_n}^0 \Psi_n d\nu = \rho_n \frac{\sigma}{\sqrt{n}} \int_{\frac{\sqrt{n}}{\sigma} a_n}^{-\frac{\sqrt{n}}{\sigma} a_n} \left[\Phi \left(\frac{\sqrt{n}(b_n - a_n)}{\sigma} - z \right) - \Phi(-z) \right] dz$$

for some $\rho_n \in [\inf_{\mu \in (2a_n, 0)} \rho(\mu), \sup_{\mu \in (2a_n, 0)} \rho(\mu)]$. Using the facts that $\int_{-x}^x \Phi(z) dz = x$ and $-\frac{\sqrt{n}b_n}{\sigma} \rightarrow -\infty$ we further obtain

$$\begin{aligned} \int_{2a_n}^0 \Psi_n d\nu &= \rho_n \frac{\sigma}{\sqrt{n}} \left[\int_{\frac{\sqrt{n}}{\sigma} a_n}^{-\frac{\sqrt{n}}{\sigma} a_n} [1 - \Phi(-z)] dz - \int_{\frac{\sqrt{n}}{\sigma} a_n}^{-\frac{\sqrt{n}}{\sigma} a_n} \Phi \left(z + \frac{\sqrt{n}}{\sigma}(a_n - b_n) \right) dz \right] \\ &= -\rho(0^-) a_n (1 + o_m) = \sigma \rho(0^-) \sqrt{\frac{\log v}{n}} (1 + o_m) . \end{aligned}$$

where the last equality holds due to equation (2.6) from the manuscript.

It remains to show that the integral over $(-\infty, 2a_n)$ is of lower order. It holds that

$$\begin{aligned} \int_{-\infty}^{2a_n} \Psi_n d\nu &\leq \int_{-\infty}^{2a_n} (1 - \Phi(\sqrt{n}(a_n - \mu)/\sigma)) d\nu \\ &\leq 1 - \Phi(-a_n \sqrt{n}/\sigma) = O\left((v \log v)^{-1/2}\right) . \end{aligned}$$

Assumption (A) yields $f\delta \log v \rightarrow \infty$, and hence $(v \log v)^{-1/2} = o\left(\sqrt{\frac{\log v}{n}}\right)$. Similar computations for the interval $(0, \infty)$ lead to

$$t_2 = \sigma \sqrt{\frac{\log v}{n}} (\rho(0^-) + \rho(0^+)) (1 + o_m(1)) . \tag{C.1}$$

In case of $0 < C < \infty$ we know from Lemma B.1 that $a_n \rightarrow -T$ and $b_n \rightarrow T$, where $T = \sigma\sqrt{C} > 0$. For $\mu \in (-T, T)$, $\Psi_n(\mu) \rightarrow 1$, while for $\mu \in (-\infty, T) \cup (T, \infty)$, $\Psi_n(\mu) \rightarrow 0$. Then by the dominated convergence theorem,

$$t_2 = \int_{-\infty}^{\infty} \Psi_n(\mu) d\nu(\mu) = \nu(-T, T) (1 + o_m) , \tag{C.2}$$

and $\nu(-T, T) > 0$, since the distribution has a positive density in neighborhoods of $-T$ and T .

The Bayes risk can be written as

$$R = mp\delta_A t_2 (1 + f\delta t_1/t_2) .$$

Thus by (C.1) and (C.2) to complete the proof of Theorem 2.1 it is enough to show that

$$f\delta t_1/t_2 \rightarrow 0 . \tag{C.3}$$

In case of $C = 0$, equation (2.6) and the normal tail approximation yield $t_1 \propto (v \log v)^{-1/2}$. Thus from (C.1) we easily obtain

$$f \delta \frac{t_1}{t_2} \propto \frac{f \delta \sqrt{n}}{\sqrt{v} \log v} = \frac{1}{\log v} \rightarrow 0 .$$

In case of $C > 0$ we write $t_1 = t_{1a} + t_{1b}$, where $t_{1a} = \Phi(\sqrt{n}a_n/\sigma)$ and $t_{1b} = 1 - \Phi(\sqrt{n}b_n/\sigma)$. Using the fundamental equality (2.4) for a_n yields

$$\delta f t_{1a} \sim \frac{\sigma}{T \sqrt{n}} \frac{1}{\sqrt{2\pi}} \int_{\mathcal{R}} \exp\left(-\frac{n}{2\sigma^2}(a_n - \mu)^2\right) d\nu(\mu) .$$

Because $a_n \rightarrow -T$ similar considerations as in (B.7) show that the integral vanishes rapidly for $\mu \notin (-T - \epsilon, -T + \epsilon)$. Now observe that

$$\begin{aligned} & \frac{1}{\sqrt{2\pi}} \int_{-T-\epsilon}^{-T+\epsilon} \exp\left(-\frac{n}{2\sigma^2}(a_n - \mu)^2\right) \rho(\mu) d\mu \\ & \leq M_\rho \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \exp\left(-\frac{n}{2\sigma^2}(a_n - \mu)^2\right) d\mu , \end{aligned}$$

where $M_\rho = \sup_{\mu \in (-T-\epsilon, -T+\epsilon)} \rho(\mu) < \infty$. Moreover,

$$\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \exp\left(-\frac{n}{2\sigma^2}(a_n - \mu)^2\right) d\mu = \frac{\sigma}{\sqrt{n}} .$$

Thus we finally obtain $\delta f t_{1a} = O(\frac{1}{n})$. Analogous considerations for t_{1b} finish the proof.

Appendix D: Proof of Theorem 2.2

First consider the case $C = 0$. To prove sufficiency of (2.12) and (2.13) for ABOS of a fixed threshold rule note that computing type II error for rules of the form (2.11) involves similar computations to those leading to (C.1), but using \tilde{a}_n , and \tilde{b}_n defined in (2.7) instead of a_n and b_n . Taking into account (2.11) and (2.12) one thus obtains

$$\int_{2\tilde{a}_n}^0 \Psi_n d\nu = -\rho(0^-)\tilde{a}_n(1 + o_m) = \rho(0^-)\sigma\sqrt{\frac{\log v}{n}}(1 + o_m) ,$$

which is asymptotically equivalent to the first contribution of the type II error of the Bayes Oracle. On the other hand

$$\int_{-\infty}^{2\tilde{a}_n} \Psi_n d\nu \leq 1 - \Phi(-\tilde{a}_n\sqrt{n}/\sigma) \sim \frac{\exp(-z_a/2)}{\sqrt{2\pi v[\log(v) + z_a]}} = o\left(\sqrt{\frac{\log v}{n}}\right) ,$$

where the last equality follows from the first part of Assumption (A) and (2.13). Similar calculations on the interval $[0, \infty]$ yield

$$\int_0^\infty \Psi_n \, d\nu = \rho(0^+) \sigma \sqrt{\frac{\log v}{n}} (1 + o_m) .$$

Thus the type II error component of the risk $R_2 = mp\delta_A t_2$ satisfies $R_2 = R_{opt}(1 + o_m)$.

Now, using (C.1) and the tail approximation for the type I error we obtain

$$R_1/R_{opt} = \frac{m(1-p)\delta_0 t_1}{R_{opt}} = C_{\sigma\rho} \frac{\exp(-z_a/2) + \exp(-z_b/2)}{\log v} (1 + o_m) , \quad (D.1)$$

where $C_{\sigma\rho} = \frac{1}{\sqrt{2\pi}\sigma(\rho(0^-) + \rho(0^+))}$. Thus under assumption (2.13) $R_1 = o(R_{opt})$, which completes the proof of sufficiency for $C = 0$.

In case of $C > 0$ due to (2.12) it holds that $\tilde{a}_n \rightarrow -T$ and $\tilde{b}_n \rightarrow T$, and hence thresholds specified by (2.11) also have type II error of the form (C.2). For sufficiency it remains to establish (C.3). To this end note that the type I error can be written approximately as

$$t_1 \sim \frac{1}{\sqrt{2\pi}} \frac{\exp(-z_a/2) + \exp(-z_b/2)}{\sqrt{v \log v}} .$$

Hence

$$R_1/R_{opt} = C_\nu \frac{\exp(-z_a/2) + \exp(-z_b/2)}{\log v} (1 + o_m) , \quad (D.2)$$

where $C_\nu = \frac{\sqrt{C}}{\nu(-T, T)}$. Thus, under assumption (2.13) again $R_1 = o(R_{opt})$, and the proof of sufficiency is completed.

Concerning necessity, similar arguments as in the proof of Theorem 3.2 of Bogdan et al. (2011) show that (2.12) is necessary for ABOS. In that case the computations leading to (D.1) and (D.2) are still valid and imply the necessity of (2.13).

Appendix E: Bayesian False Discovery Rate vs the Bayes risk

To derive results on classical multiple testing procedures controlling FWER or FDR we first consider rules controlling the Bayesian False Discovery Rate (BFDR), a concept which was introduced in [15]:

$$BFDR = P(H_{0i} \text{ is true} | H_{0i} \text{ was rejected}) = \frac{(1-p)t_{1i}}{(1-p)t_{1i} + p(1-t_{2i})} , \quad (E.1)$$

where t_{1i} and t_{2i} are the probabilities of the corresponding type I and type II errors. As discussed in [40], under the mixture model (2.3) the BFDR of a fixed threshold multiple testing rule is related to FDR according to the formula

$$FDR = BFDR \times P(R > 0) ,$$

where R is the total number of rejections.

Based on the formulas for the type I and type II errors of the Bayes classifier presented in Section C of the Appendix, it is easy to show that under assumptions (A) and (B) both BFDR and FDR of the Bayes classifier are proportional to $(n(\log n - \log p))^{-1/2}$. It follows that for sparsity levels of the form $p \propto m^{-\beta}$ both are proportional to $(n(\log n + \log m))^{-1/2}$. Now Theorem 2.2 provides some margin for thresholds which yield asymptotically optimal rules. Below we will translate this margin of “asymptotically optimal” thresholds into the margin of “asymptotically optimal” BFDR levels.

Here we will restrict our attention to the BFDR controlling rules based on symmetric thresholds, such that $a_m^B = -b_m^B$, and use

$$c_B^2 = c_B^2(m) := \frac{n (a_m^B)^2}{\sigma^2} \quad (\text{E.2})$$

to denote the corresponding threshold for the scaled test statistics $Z_i = \frac{\sqrt{n}\bar{X}_i}{\sigma}$.

As shown in Lemma E.3, under the mixture model (2.3) BFDR decreases continuously from $1 - p$ for $c_B = 0$ to 0 as $c_B \rightarrow \infty$, implying that each $\alpha \in (0, 1 - p]$ corresponds uniquely to a threshold value c_B . To obtain the threshold value c_B with BFDR level α is equivalent to solving

$$\frac{t_1(c_B)}{1 - t_2(c_B)} = \frac{p}{1 - p} \frac{\alpha}{1 - \alpha}, \quad (\text{E.3})$$

where

$$t_1(c_B) = 2(1 - \Phi(c_B))$$

and

$$1 - t_2(c_B) = 2 - \int_{\mathbb{R}} (\Phi(c_B + \sqrt{n}\mu/\sigma) + \Phi(c_B - \sqrt{n}\mu/\sigma)) d\nu(\mu).$$

This is the key relation for the proof of the following theorem, which provides conditions on α for which the BFDR controlling rule is ABOS and gives the asymptotic approximation of the corresponding threshold c_B .

Theorem E.1. *Assume that Assumptions (A) and (B) hold. Additionally assume that $\alpha \rightarrow \alpha_\infty < 1$ and $\frac{\log \alpha}{n} \rightarrow 0$. As before let $v = \frac{n\delta^2}{p^2}$. Then a rule with $\text{BFDR} = \alpha$ is ABOS if and only if*

$$\alpha = \frac{v^{1/(2(z_m-1))}}{p} \quad \text{with} \quad z_m \rightarrow 0 \quad \text{and} \quad z_m \log v + \log \log v \rightarrow \infty. \quad (\text{E.4})$$

The threshold value c_B of a rule with BFDR equal to α is then given by

$$c_B^2 = -2 \log(p\alpha) - \log(2(-\log(p\alpha))) + 2 \log\left(\frac{\sqrt{2}(1 - \alpha_\infty)}{\sqrt{\pi} C_1}\right) + o_m, \quad (\text{E.5})$$

where

$$C_1 = 1 - \nu(-T, T).$$

Proof. According to (E.3) the threshold value c_B with BFDR level α fulfills

$$\frac{\alpha}{f(1-\alpha)} = \frac{2(1-\Phi(c_B))}{2-\int_{\mathbb{R}}[\Phi(c_B+\sqrt{n}\mu/\sigma)+\Phi(c_B-\sqrt{n}\mu/\sigma)]d\nu(\mu)}. \tag{E.6}$$

Let us define $u_n^B = c_B\sigma/\sqrt{n}$. First we want to show that u_n^B is bounded. Assume on the contrary that for some subsequence $u_j^B \rightarrow \infty$. It holds for any constant $K > 0$, that

$$\begin{aligned} & \int_{\mathbb{R}} \Phi(\sqrt{j}(-u_j^B - \mu)/\sigma)d\nu + 1 - \int_{\mathbb{R}} \Phi(\sqrt{j}(u_j^B - \mu)/\sigma)d\nu \\ & \geq (\nu(-\infty, -K) + \nu(K, \infty))(1 - \Phi(\sqrt{j}(u_j^B - K)/\sigma)). \end{aligned}$$

If $u_j^B \rightarrow \infty$ we can apply the tail approximation for the normal distribution and obtain from (E.6)

$$\frac{\alpha_j}{f} \leq \frac{(1-\alpha_j)(u_j^B - K)}{(\nu(-\infty, -K) + \nu(K, \infty))u_j^B} \exp\left(-\frac{j(u_j^B K - K^2)}{2\sigma^2}\right) (1 + o_j).$$

But on the other hand Assumption (A) together with the assumption that $n^{-1} \log \alpha \rightarrow 0$ yields $(\frac{\alpha_j}{f})^{1/j} \rightarrow \exp(-C/2)$, which contradicts $u_j^B \rightarrow \infty$.

If $u_j := u_j^B \rightarrow u < \infty$ then the denominator of (E.6) converges to a constant $C_{\nu,u} = 1 - \nu(-u, u)$. Because of $p \rightarrow 0$ and $\alpha \rightarrow \alpha_\infty < 1$ equation (E.6) can only hold if $\sqrt{j}u_j \rightarrow \infty$. Thus we can apply again the normal tail approximation to obtain

$$\frac{\alpha_j}{f} = \sqrt{\frac{2}{\pi}} \frac{1-\alpha_j}{c_B C_{\nu,u}} \exp(-\frac{c_B^2}{2})(1 + o_j).$$

Taking logarithms and some simple calculations yield

$$c_B^2 = 2 \log\left(\frac{f}{\alpha}\right) - \log\left(2 \log\left(\frac{f}{\alpha}\right)\right) + \log\left(\frac{2}{\pi}\right) + 2 \log\left(\frac{1-\alpha_\infty}{C_{\nu,u}}\right) + o_j.$$

Multiplying by σ^2/n and using again Assumption (A) together with $n^{-1} \log \alpha \rightarrow 0$ implies that $u = \sigma\sqrt{C}$, and thus $C_{\nu,u} = 1 - \nu(-T, T)$ which completes the proof of (E.5).

The critical value has exactly the same form as in the case of normal distributions and the result on ABOS follows exactly the same way as in Bogdan et al (2011). Define $s_m := \frac{\log(f\delta\sqrt{n})}{\log(f/\alpha)} - 1$, then necessary and sufficient conditions for optimality are $s_m \rightarrow 0$ and $2s_m \log(f/\alpha) - \log \log(f/\alpha) \rightarrow -\infty$. Elementary computations then yield the reformulated conditions (E.4), where we write $z_m = -s_m$. □

Lemmas showing the existence of the exact BFDR controlling rule
 We first prove the following result

Lemma E.1. For any fixed $s \neq 0$ the function

$$Q(c) := \frac{2 - \Phi(c - s) - \Phi(c + s)}{2(1 - \Phi(c))}$$

satisfies

- a) $Q(0) = 1$,
- b) $\lim_{c \rightarrow \infty} Q(c) = \infty$,
- c) $Q(c)$ is increasing in c for $c \geq 0$.

Proof. Points a) and b) easily follow by elementary algebra. To prove point c) let us define

$$g(c) := (1 - \Phi(c))\phi(c - s) - (1 - \Phi(c - s))\phi(c).$$

Then straight forward calculations yield

$$Q'(c) > 0 \quad \Leftrightarrow \quad g(c + s) > g(c)$$

Let us consider at first the case of $s > 0$. In this situation it is enough to show that $g(c)$ is increasing. We find

$$g'(c) = (1 - \Phi(c))\phi'(c - s) - (1 - \Phi(c - s))\phi'(c)$$

and define $h(c) = \frac{\phi'(c)}{1 - \Phi(c)}$. Then clearly

$$g'(c) > 0 \quad \Leftrightarrow \quad h(c - s) > h(c).$$

To show that $h(c)$ is a decreasing function observe that

$$h'(c) = \frac{1}{2\pi(1 - \Phi(c))^2} e^{-c^2/2} \left(\sqrt{2\pi}(c^2 - 1)(1 - \Phi(c)) - ce^{-c^2/2} \right).$$

Now, the standard bound on the tail of the normal distribution yields

$$\sqrt{2\pi}c^2(1 - \Phi(c)) < ce^{-c^2/2},$$

which implies that $h'(c) < 0$.

The proof for $s < 0$ goes analogously. In that case $g(c)$ has to be decreasing, which yields $h(c) > h(c - s)$, and again $h(c)$ is a decreasing function. \square

The following Lemma E.2 easily follows from Lemma E.1.

Lemma E.2. Let $\nu(\cdot)$ be any probability measure such that $\nu(0) < 1$. Let us define

$$H(c) := \frac{\int_{\mathbb{R}} (2 - \Phi(c - \sqrt{n}\mu/\sigma) - \Phi(c + \sqrt{n}\mu/\sigma)) d\nu(\mu)}{2(1 - \Phi(c))}. \quad (\text{E.7})$$

Then it holds that

- a) $H(0) = 1$,
- b) $\lim_{c \rightarrow \infty} H(c) = \infty$,
- c) $H(c)$ is increasing on $[0, \infty]$.

Lemma E.3. Let $\nu(\cdot)$ be any probability measure such that $\nu(0) < 1$. Let

$$BFDR(c) = \frac{(1-p)t_1(c)}{(1-p)t_1(c) + p(1-t_2(c))} ,$$

where

$$t_1(c) = 2(1 - \Phi(c))$$

and

$$t_2(c) = 1 - \int_{\mathbb{R}} (\Phi(-c - \sqrt{n}\mu/\sigma) + 1 - \Phi(c - \sqrt{n}\mu/\sigma)) d\nu .$$

Then $BFDR(c)$ is continuously decreasing from $1 - p$ for $c = 0$ to 0 for $c \rightarrow \infty$.

Proof. Observe that

$$BFDR(c) = \frac{1}{1 + \frac{p}{1-p}H(c)} ,$$

with $H(c)$ as in (E.7). Thus Lemma E.3 is a direct consequence of Lemma E.2. □

Appendix F: Existence of c_{GW} and Proof of 3.1

Corollary F.1. For any p there is a unique positive solution c_{GW} of (3.6) for each $\alpha \in (0, 1)$.

Proof. Note that

$$\frac{2(1 - \Phi(c))}{1 - F(c)} = \frac{t_1(c)}{(1-p)t_1(c) + p(1-t_2(c))} = BFDR(c)/(1-p) ,$$

and thus from Lemma E.3 it follows that the expression on the left decreases from 1 for $c = 0$ to 0 for $c \rightarrow \infty$. □

Proof of Theorem 3.1:

Proof. First observe that the threshold c_{GW} (3.6) at the level α coincides with the threshold of the BFDR controlling rule at the level $\alpha' = \alpha(1 - p)$. The rest of the proof follows by observing that when $p \rightarrow 0$ then α' satisfies Condition (Op) if and only if α satisfies Condition (Op). □

Appendix G: Proof of Theorem 3.2

To prove optimality of the type II risk component of SD in the denser case we first show that with large probability the random threshold of SD is bounded from above by the asymptotically optimal threshold \tilde{c}_{1n} .

Lemma G.1. *Let c_{SD} be the random SD threshold at the level α_m and let $\tilde{c}_1 = \tilde{c}_{1m}$ be the GW threshold (3.6) at the level $\alpha_{1m} = \alpha_m \xi_m$, where $\xi_m = (\log m)^{-s}$ with $s > 1$. Suppose that Assumptions (A), (B), Condition (Op), (3.7) and (3.9) hold with $\alpha = \alpha_m$. Then \tilde{c}_1 is ABOS. Moreover, for every $\gamma_u > 0$ it holds for sufficiently large m that*

$$P(c_{SD} \geq \tilde{c}_1) \leq m^{-\gamma_u} . \quad (\text{G.1})$$

Proof. Based on the second condition in (3.9) it is easy to show that α_{1m} satisfies the asymptotic optimality assumptions provided in Theorem E.1. Thus, Theorem 3.1 immediately yields that \tilde{c}_1 is ABOS.

To prove the second assertion of the Lemma we first note that by Lemma E.2 the function $\tilde{H}(c) := \frac{2(1-\Phi(c))}{1-F(c)}$ is decreasing. Therefore according to the definition of \tilde{c}_1 ,

$$\{c_{SD} \geq \tilde{c}_1\} = \left\{ \tilde{H}(c_{SD}) \leq \alpha_{1m} \right\} . \quad (\text{G.2})$$

On the other hand the definition of c_{SD} actually gives

$$\frac{2(1-\Phi(c_{SD}))}{1-F_m^{(SD)}(c_{SD}) + 1/m} = \alpha_m$$

and thus

$$\{c_{SD} \geq \tilde{c}_1\} = \left\{ \frac{1 - F_m^{(SD)}(c_{SD}) + 1/m}{1 - F(c_{SD})} \leq \xi_m \right\} .$$

Taking another intersection of the right hand side with $\{c_{SD} \geq \tilde{c}_1\}$ we can conclude that

$$P(c_{SD} \geq \tilde{c}_1) \leq P\left(\inf_{c \geq \tilde{c}_1} \frac{1 - F_m^{(SD)}(c) + 1/m}{1 - F(c)} \leq \xi_m \right) . \quad (\text{G.3})$$

Using the standard transformation $U_i = F(|Z_i|)$ one obtains

$$P(c_{SD} \geq \tilde{c}_1) \leq P\left(\inf_{t \in [z_{1m}, 1]} \frac{1 - \hat{G}_m(t) + 1/m}{1 - t} \leq \xi_m \right)$$

where $z_{1m} = F(\tilde{c}_1)$, and $\hat{G}_m(t)$ is the empirical cdf of U_1, \dots, U_m . Now, using the transformation $u = 1 - t$ and observing that $V_i = 1 - U_i$ also has a uniform distribution we obtain

$$P(c_{SD} \geq \tilde{c}_1) \leq P\left(\inf_{u \in [0, 1 - z_{1m}]} \frac{\hat{G}_m(u) + 1/m}{u} \leq \xi_m \right)$$

This is equivalent to computing the probability that the empirical process $\hat{G}_m(u)$ intersects the line $L = -\frac{1}{m} + u\xi_m$ within the interval $[\frac{1}{m\xi_m}, 1 - z_{1m}]$. For this

type of problem Proposition 9.1.1 of Shorack and Wellner (Beginning of chapter 9, page 344) can be applied. Define the event

$$B_i = \{\hat{G}_m(u) \text{ intersects the line } y = (u-a)/(bm) \text{ at height } i/m \text{ but not below}\}$$

Then

$$P(B_i) = \binom{m}{i} a(a+ib)^{i-1} (1-a-ib)^{m-i}.$$

In our case $a = b = \frac{1}{m\xi_m}$ and thus

$$P(B_i) = \binom{m}{i} \frac{1}{m\xi_m} \left(\frac{1+i}{m\xi_m}\right)^{i-1} \left(1 - \frac{1+i}{m\xi_m}\right)^{m-i} \quad \text{for } i < m\xi_m - 1$$

and $P(B_i) = 0$ for $i \geq m\xi_m - 1$.

Now, similar to Lemma 10.3.1 of Shorack and Wellner (page 414) we can apply Stirling's formula, which for $i < m\xi_m - 1$ yields

$$\begin{aligned} P(B_i) &< \frac{m!}{(i+1)!(m-i)!} \left(\frac{1+i}{m\xi_m}\right)^i \left(1 - \frac{1+i}{m\xi_m}\right)^{m-i} \\ &< \frac{m^{m+1/2} e^{-m} \sqrt{2\pi} \exp(1/12m)}{(i+1)^{i+3/2} e^{-(i+1)} \sqrt{2\pi} (m-i-1)^{m-i+1/2} e^{-(m-i)} \sqrt{2\pi}} \left(\frac{1+i}{m\xi_m}\right)^i \\ &\quad \times \left(1 - \frac{1+i}{m\xi_m}\right)^{m-i} \\ &< \frac{\exp(1/12m+1)}{\sqrt{2\pi}} \frac{1}{(i+1)^{3/2} \sqrt{1-i/m}} \xi_m^{-i} \left(\frac{m-(1+i)/\xi_m}{m-i}\right)^{m-i} \\ &< \frac{\exp(1/12m+1)}{\sqrt{2\pi}} \frac{1}{(i+1)^{3/2} \sqrt{1-i/m}} \xi_m^{-i} \exp\left(-i\left(\frac{1+i}{i\xi_m} - 1\right)\right) \end{aligned}$$

In the last step we adapted the inequality $\left(1 - \frac{i(\lambda-1)}{n-i}\right)^{n-i} < e^{-i(\lambda-1)}$ used by Shorack and Wellner in the proof of Lemma 10.3.1. In summary we find that

$$P(B_i) < K \xi_m^{-i} \exp(-(i+1)/\xi_m).$$

for some constant K which can be chosen such that it does not depend on m or i . As long as $\frac{1}{\xi_m} \exp(-1/\xi_m) < 1$ we then have

$$P(c_{SD} \geq \tilde{c}_1) \leq K \sum_{i=0}^{\infty} \xi_m^{-i} \exp(-(i+1)/\xi_m) = K \frac{\exp(-1/\xi_m)}{1 - \frac{1}{\xi_m} \exp(-1/\xi_m)}.$$

Remembering that $\xi_m = (\log m)^{-s}$ with $s > 1$ finally yields (G.1). □

With this Lemma we are ready to prove Theorem 3.2 itself. First note that BH is more liberal than SD, thus it is enough to control the risk contribution of Type 1 error for BH, as well as the risk contribution of Type 2 error for SD.

Under the first condition in (3.9) the proof for Type 1 error of BH follows along the same lines as the proof of Lemma 5.4 in [6]. Also, under the condition of extreme sparsity (2.14) according to Theorem 3.3 the Bonferroni procedure is ABOS. Therefore the optimality of the type II error component of the risk of SD in the extremely sparse case follows directly from a comparison with the more conservative Bonferroni correction. There remains to bound the type II error component of the risk of SD for the denser case (3.8), which overlaps with the extremely sparse case.

Denote by L_A the number of false negatives under the SD rule, and by R_A the corresponding type II error component of the risk. Furthermore let \tilde{c}_1 be the GW threshold defined in Lemma G.1. Clearly

$$E(L_A) \leq E(L_A | c_{SD} \leq \tilde{c}_1) P(c_{SD} \leq \tilde{c}_1) + m P(c_{SD} > \tilde{c}_1) ,$$

and also

$$E(L_A | c_{SD} \leq \tilde{c}_1) P(c_{SD} \leq \tilde{c}_1) \leq EL_1 ,$$

where L_1 is the number of false negatives produced by the rule based on the threshold \tilde{c}_1 . Since by Lemma G.1 the rule based on \tilde{c}_1 is asymptotically optimal, it follows that $\delta_A EL_1 = R_{opt}(1 + o_m)$. On the other hand by Lemma G.1 it holds that $P(c_{SD} > \tilde{c}_1) \leq m^{-\gamma_u}$ for any $\gamma_u > 0$ if only m is sufficiently large, and therefore

$$R_A = \delta_A EL_A \leq R^B(1 + o_m) + \delta_A m^{1-\gamma_u} .$$

Now by using assumptions (3.8) and (3.9), and choosing e. g. $\gamma_u = \gamma_2/2 + 1$, we conclude that $\delta_A m^{1-\gamma_u} = o(R_{opt})$, and the proof is thus complete.

Appendix H: Proofs of Theorems 3.3 and Proposition 3.1

Proof of Theorem 3.3:

Proof. If $m \rightarrow \infty$ then the threshold for the Bonferroni correction can be written as

$$c_{Bon}^2 = 2 \log \left(\frac{m}{\alpha} \right) - \log \left(2 \log \left(\frac{m}{\alpha} \right) \right) + \log(2/\pi) + o_m . \quad (\text{H.1})$$

Comparison of this threshold with the asymptotic approximation to an optimal BFDR control rule (E.5) yields

$$c_{Bon}^2 = c_B^2 + 2 \log mp + O_m(1) .$$

From (2.14) it follows easily that $c_{Bon}^2 = c_B^2(1 + o_m)$. By assumption, the rule based on the threshold c_B^2 is optimal, and hence c_{Bon}^2 satisfies condition (2.12) of Theorem 2.2. Condition (2.13) is satisfied, since by assumption $\log mp$ is bounded from below and thus ABOS of the Bonferroni correction follows.

Otherwise assume that (2.14) does not hold, which means that there is a subsequence with $\log(mp)/\log(n/p^2) \rightarrow D \in (0, \infty]$. But then

$$c_{Bon}^2 = c_B^2 + 2D \log(n/p^2) + O_m(1)$$

and $\log(n/p^2) \neq o(\log v)$. Specifically remember that $v = n\delta^2/p^2$, and because of Assumption (A) δ is bounded from above and $\log \delta = o(\log p)$. Thus it is not possible that $\log(n/p^2)/\log(\delta^2) \rightarrow 0$. It follows that condition (2.12) cannot be fulfilled, and the Bonferroni rule is not ABOS. \square

Proof of Proposition 3.1

Proof. For fixed δ the Bonferroni threshold (H.1) can be rewritten as

$$c_{Bonf}^2 = 2 \log(m/\alpha) - \log 2 \log(m/\alpha) + O(1) = \log v + z$$

with

$$z = 2 \log mp - \log(n\alpha^2) - \log 2 \log(m/\alpha) + O(1).$$

Under assumption (3.11) it is clear that $\log 2 \log(m/\alpha) = o(v)$, and thus the first condition for a fixed threshold rule to be optimal becomes

$$\frac{2(1-\beta) \log m - \log(n\alpha^2)}{\log n - 2 \log p} \rightarrow 0.$$

This is trivially fulfilled because (3.11) is equivalent to $2(1-\beta) \log m = \log(n\alpha^2) + \text{const}$. It is easy to check that this choice also fulfills the second condition for optimality. \square

References

- [1] ABRAMOVICH F., BENJAMINI Y., DONOHO D. L. and JOHNSTONE I. M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.* **34**, 584–653. MR2281879 [MR2281879](#)
- [2] BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B.* **57**, 289–300. MR1325392 [MR1325392](#)
- [3] BOGDAN, M., GHOSH, J. K., and DOERGE, R. W. (2004). Modifying the Schwarz Bayesian Information Criterion to locate multiple interacting quantitative trait loci. *Genetics* **167**, 989–999.
- [4] BOGDAN, M., GHOSH, J. K., OCHMAN, A. and TOKDAR, S. T. (2007) On the Empirical Bayes approach to the problem of multiple testing. *Quality and Reliability Engineering International* **23**, 727–739.
- [5] BOGDAN, M., GHOSH, J. K. and TOKDAR S. T. (2008). A comparison of the Simes-Benjamini-Hochberg procedure with some Bayesian rules for multiple testing. *IMS Collections, Vol.1, Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen, edited by N. Balakrishnan, Edsel Peña and Mervyn J. Silvapulle* 211–230. Beachwood Ohio. [MR2462208](#)
- [6] BOGDAN, M., CHAKRABATI, A., FROMMLET F. and GHOSH, J. K. (2011) Asymptotic Bayes Optimality under sparsity of some multiple testing procedures. *Ann. Statist.*, **39**, 1551–1579. [MR2850212](#)

- [7] BOGDAN, M., GHOSH, J. K. and ŽAK-SZATKOWSKA, M. (2008) Selecting explanatory variables with the modified version of Bayesian Information Criterion, *Quality and Reliability Engineering International* **24**, 627–641.
- [8] BÜHLMANN, P. and VAN DE GEER, S. (2011) *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer. [MR2807761](#)
- [9] CAI, T. and JIN, J. (2010). Optimal rates of convergence for estimating the null and proportion of non-null effects in large-scale multiple testing. *Ann. Statist.* **38**, 100–145. [MR2589318](#)
- [10] CHI, Z. (2008). False discovery rate control with multivariate p -values. *Electronic Journal of Statistics* **2**, 368–411. [MR2411440](#)
- [11] COX, D. R. and HINKLEY, D. V. (1974). *Theoretical Statistics*. Chapman and Hall. [MR0370837](#)
- [12] DONOHO, D. L. and JIN, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32**, 962–994. [MR2065195](#)
- [13] DONOHO, D. L. and JIN, J. (2006). Asymptotic minimaxity of false discovery rate thresholding for sparse exponential data. *Ann. Statist.* **34**, 2980–3018. [MR2329475](#)
- [14] DONOHO, D. L. and JOHNSTONE, I. M. (1994). Minimax risk over l_p -balls for l_q -error. *Probab. Theory Related Fields* **99**, 277–303. [MR1278886](#)
- [15] EFRON, B. and TIBSHIRANI, R. (2002). Empirical bayes methods and false discovery rates for microarrays. *Genetic Epidemiology*, **23**, 70–86.
- [16] EFRON, B. (2008). Microarrays, Empirical Bayes and the two-group model. *Stat. Sci.*, **23**(1), 1–22. [MR2431866](#)
- [17] FAN, J., HALL, P. and YAO, Q. (2007). To how many simultaneous hypothesis tests can normal, Student’s t or bootstrap calibration be applied? *J. Amer. Statist. Assoc.*, **102**, 1282–1288. [MR2372536](#)
- [18] FELLER, W. (1966). *An introduction to probability theory and its applications*. Vol. 2: Wiley, New York. [MR0210154](#)
- [19] FINNER, H., DICKHAUS, T. and ROTERS, M. (2009). On the false discovery rate and an asymptotically optimal rejection curve. *Ann. Statist.* **37**, 596–618. [MR2502644](#)
- [20] FROMMLET, F., RUHALTINGER, F., TWAROG, P. and BOGDAN, M. (2012). Modified versions of Bayesian Information Criterion for genome-wide association studies. *Comput. Stat. Data An.*, **56**, 1038–1051. [MR2897552](#)
- [21] GENOVESE, C. and WASSERMAN, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64**(3), 499–517. [MR1924303](#)
- [22] GEORGE, E. I. and FOSTER, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87**, 731–747. [MR1813972](#)
- [23] GUO, W. and RAO, M. B. (2008). On optimality of the Benjamini-Hochberg procedure for the false discovery rate. *Statistics and Probability Letters* **78**, 2024–2030. [MR2528629](#)

- [24] JIN, J. and CAI, T. C. (2007). Estimating the null and the proportion of non-null effects in large-scale multiple comparisons. *J. Amer. Statist. Assoc.* **102**, 495-506. [MR2325113](#)
- [25] JOHNSON, B. R. and TRUAX, D. R. (1973). Asymptotic behavior of Bayes tests and Bayes risk. *Ann. Statist.* **2**, 278-294. [MR0362588](#)
- [26] JOHNSTONE, I. M. and SILVERMAN, B. W. (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.* **32**, 1594-1649. [MR2089135](#)
- [27] JOHNSTONE, I. M. and SILVERMAN, B. W. (2005). EbayesThresh: R programs for empirical Bayes thresholding. *J. Stat. Software* **12**, Issue 8.
- [28] LEHMANN, E. L. 1957. A theory of some multiple decision problems, I. *Ann. Math. Stat.* **28**, 1-25. [MR0084952](#)
- [29] LEHMANN, E. L. and ROMANO, J. P. (2005). Generalizations of the familywise error rate. *Ann. Statist.* **33**, 1138-1154. [MR2195631](#)
- [30] LEHMANN, E. L., ROMANO, J. P. and POPPER SHAFFER, J. (2005). On optimality of stepdown and stepup multiple test procedures. *Ann. Statist.* **33**, 1084-1108. [MR2195629](#)
- [31] MEINSHAUSEN, N. and RICE, J. (2006). Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Ann. Statist.* **34**, 373-393. [MR2275246](#) [MR2275246](#)
- [32] MEUWISSEN, T. and GODDARD, M. (2010). Accurate Prediction of Genetic Values for Complex Traits by Whole-Genome Resequencing. *Genetics* **185** (2), 623-631
- [33] NEUVIAL, P. and ROQUAIN, E. (2011). On false discovery rate thresholding for classification under sparsity. *arXiv:1106.6147*
- [34] PEÑA, E. A., HABIGER, J. D., and WU, W. (2011). Power-enhanced multiple decision functions controlling family-wise error and false discovery rates. *Ann. Statist.* **39**(1), 556-583. [MR2797856](#)
- [35] PURDOM, E. and HOLMES, S. P. (2005) Error Distribution for Gene Expression Data. *SAGMB* **4**(1), Article 16 [MR2170432](#)
- [36] ROQUAIN, E., and VAN DE WIEL, M. A. (2009). Optimal weighting for false discovery rate control. *Electronic Journal of Statistics* **3**, 678-711. [MR2521216](#)
- [37] SCHWARZ, G. (1978). Estimating the Dimension of a Model. *Ann. Statist.* **6**(2), 461-464. [MR0468014](#)
- [38] SCOTT, J. G. and BERGER, J. O. (2006). An exploration of aspects of Bayesian multiple testing. *J. Statist. Plann. Inference* **136**(7), 2144-2162. [MR2235051](#)
- [39] SHORACK, G. R. and WELLNER, J. A. (1986). *Empirical processes with applications to Statistics*, Wiley Series in Probability and Mathematical Statistics. [MR0838963](#)
- [40] STOREY, J. D. (2003). The positive false discovery rate: a Bayesian interpretation and the q -value. *Ann. Statist.*, **31**(6), 2013-2035. [MR2036398](#)
- [41] STOREY, J. D. (2007). The optimal discovery procedure: a new approach to simultaneous significance testing. *J. R. Statist. Soc. B* **69**, 347-368. [MR2323757](#)

- [42] SUN, W. and CAI, T. C. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *J. Amer. Statist. Assoc.* **102**, 901–912. [MR2411657](#)
- [43] TIBSHIRANI, R. (1996). Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B*, **58**, 267–288. [MR1379242](#)
- [44] ŹAK-SZATKOWSKA, M. and BOGDAN, M. (2011). Modified versions of Bayesian Information Criterion for sparse Generalized Linear Models, *Comput. Stat. Data An.* **55**, 2908–2924. [MR2813055](#)