ASYMPTOTIC BAYES OPTIMALITY UNDER SPARSITY FOR GENERALLY DISTRIBUTED EFFECT SIZES UNDER THE ALTERNATIVE

By Florian Frommlet*, Arijit Chakrabarti†, Magdalena Murawska‡ and Małgorzata Bogdan§

*Medical University of Vienna**, *Indian Statistical Institute*†,*Erasmus University*‡, *Wrocław University of Technology*§

Recent results concerning asymptotic Bayes-optimality under sparsity (ABOS) of multiple testing procedures are extended to fairly generally distributed effect sizes under the alternative. An asymptotic framework is considered where both the number of tests m and the sample size n go to infinity, while the fraction p of true alternatives converges to zero. It is shown that under mild restrictions on the loss function nontrivial asymptotic inference is possible only if n increases to infinity at least at the rate of $\log m$. Based on this assumption precise conditions are given under which the Bonferroni correction with nominal Family Wise Error Rate (FWER) level α and the Benjamini- Hochberg procedure (BH) at FDR level α are asymptotically optimal. When $n \propto \log m$ then α can remain fixed, whereas when n increases to infinity at a quicker rate, then α has to converge to zero roughly like $n^{-1/2}$. Under these conditions the Bonferroni correction is ABOS in case of extreme sparsity ($p \propto m^{-1}$), while BH adapts well to the unknown level of sparsity.

In the second part of this article these optimality results are carried over to model selection in the context of multiple regression with orthogonal regressors. Several modifications of Bayesian Information Criterion are considered, controlling either FWER or FDR, and conditions are provided under which these selection criteria are ABOS. Finally the performance of these criteria is examined in a brief simulation study.

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], [10], [16], [17], [28] or [32]). 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 Dicovery Rate or the expected number of false positives) at a certain, fixed level (see, [27], [31], [15], [34], [33], [22], [38] or [39]). A different notion of optimality is proposed in [36] and [8], which investigate multiple testing procedures in the context of minimizing the Bayes risk.

AMS 2000 subject classifications: Primary 62C25,62F05; secondary 62C10

Keywords and phrases: Multiple testing, Model selection, FDR, Bayes oracle, asymptotic optimality, two groups model

In many applications of high-dimensional multiple testing it is assumed that the proportion p of true alternative hypotheses among all tests is very small. In asymptotic analysis this is often expressed by the sparsity assumption, that p decreases to 0 as the total number of tests m increases to infinity. Recently, substantial efforts have been made to understand the asymptotic properties of multiple testing under sparsity (see, [16], [17], [1], [8]).

Bogdan et al. [8] consider the problem of testing hypotheses about means μ_i in normal populations $X_i \sim N(\mu_i, \sigma^2)$, $i = 1, \ldots, m$. Their analysis is based on a two-groups model, which assumes that the unknown means are generated by the scale mixture of two normal distributions: null and alternative. The classical case of testing $H_{0i} : \mu_i = 0$ corresponds to the situation when the variance of the null distribution is equal to 0. In [8] the ratio u of variances of the alternative distribution of μ_i and the null distribution of X_i, slowly increases to infinity as $p \to 0$, at a rate which guarantees that the limiting power of the Bayes classifier is larger than 0 and smaller than 1. Such sequences of alternative distributions are considered to be "on the verge of detectability". The Bayes risk is computed assuming that losses generated by the type I and type II errors are the same for all tests, and the total loss is the sum of losses for individual tests. In case of known p, σ^2 and u the risk is minimized by using Bayes classifiers for each individual test. This optimal rule, which is in practice unattainable, is referred to as the Bayes oracle.

Under the described asymptotic assumptions a multiple testing rule is classified as asymptotically Bayes optimal under sparsity (ABOS) if the ratio of the corresponding Bayes risk and the risk of the Bayes oracle converges to one. Bogdan et al. [8] characterize the class of multiple testing rules with fixed threshold which are ABOS, and they provide conditions under which the Bonferroni correction and the popular Benjamini–Hochberg multiple testing procedure (BH, [3]) are asymptotically optimal.

In the first part of this paper we extend the results of [8] concerned with testing $H_{0i} : \mu_i = 0$ to the case when the distribution of μ_i under the alternative $\nu(\mu)$ is fixed and not necessarily normal, while the number of individuals n used to calculate the test statistics $\bar{X}_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij}$ increases with m. It turns out that, given $p \propto m^{-\beta}$, signals are at the verge of detectability exactly when $n \propto \log m$. This situation is notably relevant in the context of bioinformatics data, where n is usually much smaller than m. We show that in this case BH and the Bonferroni correction are ABOS under the same assumptions as in [8]. In particular, we show that if $\nu(\mu)$ has a positive and bounded density on the real line then the Bonferroni correction at a fixed FWER $\alpha \in (0, 1)$ is ABOS if $p \propto m^{-1}$ and the ratio of losses for the false positive and the false negative δ decreases to 0 at such a rate that $\log \delta = o(\log m)$. In contrast BH at a fixed FDR level $\alpha \in (0, 1)$ adapts very well to the unknown level of sparsity and is ABOS whenever $p \propto m^{-\beta}$, $\beta \in (0, 1]$. As explained in [8] the assumption of decreasing δ is quite reasonable since the cost of missing a true signal usually increases when the total number of signal decreases. We also show that if $p \propto m^{-\beta}$ with $\beta \in (0, 1]$ then the step-down version of the FDR controlling procedure, SD, is ABOS under the same conditions as BH.

Unlike in [8] we also consider the case where the power of the Bayes oracle converges to 1. For $p \propto m^{-\beta}$ this relates to the case where n increases to infinity at a quicker rate than $\log m$. We show that in this case BH and SD are ABOS for any $\beta \in (0, 1]$ as long as FDR levels decrease to 0 approximately at the rate of $n^{-1/2}$, while δ is bounded from above and such that $\log \delta = o(\log m)$. Similarly, the Bonferroni correction is ABOS if its FWER converges to zero at the rate $n^{-1/2}$ and $p \propto \frac{1}{m}$. In this case the only assumption on ν is that it has a positive and bounded density in a neighborhood of 0. Extending the results of [8] to a more general class of distributions is based on techniques introduced by [29], where nontrivial modifications are required to deal with sparsity.

In the second part of the paper we use the results on multiple testing rules to prove asymptotic optimality of some model selection criteria for sparse least squares regression. Here we concentrate on the orthogonal design and study the two cases of known and unknown variance of the error term σ^2. As discussed in [6], in case of orthogonal design with known σ, penalized likelihood model selection criteria work analogously to multiple testing procedures which verify individually the significance of each regression coefficient. Based on this analogy it is very easy to prove that popular model selection criteria, like AIC [2] or BIC [35], are not consistent when $\frac{m}{\sqrt{n}}$ increases to infinity (see [6]). Specifically, under this scenario the expected number of false discoveries increases to infinity.

To solve this problem some modifications of AIC [12] and BIC (see, [5, 13]) were recently proposed in the literature. In this article we will concentrate on modifications of BIC, which is more appropriate to consider when one aims at minimizing the misclassification rate, or in our context the Bayes risk based on a generalized 0-1 loss. The first of the considered criteria, mBIC, was derived in [5] in a Bayesian setting using a prior on the model dimension which assumes that the expected number of true regressors does not depend on m. In case of orthogonality and known σ it was pointed out in [6] that mBIC is controlling the FWER. Optimality results at a sparsity level $p \propto m^{-1}$ follow immediately from the analysis for multiple testing.

In view of results on multiple testing it would actually be of great interest to study model selection criteria which control the FDR. In [1] penalized model selection schemes are discussed which have exactly this property. Quite similar penalties have been discussed in [23] and [26]. Starting from the penalty of [1] we will introduce several new modifications of BIC (mBIC1 - mBIC3), where mBIC2 has been shown already to perform very well in the application of genome wide association studies [24]. In case of known σ we prove that the FDR controlling criteria are ABOS for a wide range of sparsity levels, satisfying for example $p = m^{-\beta}$, with $\beta \in (0, 1]$.

In most applications it is much more realistic to assume that σ is not known. Under sparsity it is rather difficult to get reliable estimates on σ, and for that reason optimality results on the corresponding model selection criteria under sparsity are very rare in the literature. In a Bayesian approach with normally distributed error terms, σ is integrated out and in the corresponding version of BIC the residual sum of squares RSS is replaced by $\log RSS$. We will show that in this context mBIC is again ABOS in case of extreme sparsity. The

conditions we need for unknown σ are not much more restrictive than for known σ. Our proof is technically rather involved, and cannot be easily extended to prove ABOS for mBIC1 - mBIC3. However, in analogy to the case of known variance we conjecture that these criteria should be ABOS for a wide range of sparsity levels. This conjecture is underpinned by simulations, which show good properties of the new versions of mBIC both for known and unknown σ.

The rest of the paper is organized as follows. In Section 2 we present results for multiple testing, whereas Section 3 focuses on linear regression models under orthogonality. The main emphasis of Sections 2.1 and 2.2 is the generalization of results from [8] to the situation of general distributions under the alternative. Section 2.3 shows ABOS of Bonferroni correction in case of extreme sparsity. The most important theorems on multiple testing are given in Section 2.4, where ABOS of step-up and step-down FDR controlling procedures is proven. These results are needed in Section 3.2 to show ABOS of the FDR-controlling model selection criteria, after ABOS of mBIC for known variance was shown in Section 3.1. Optimality results of mBIC for unknown variance are proved in Section 3.3. Finally in Section 4 different model selection criteria are compared in a small simulation study. Most proofs of technical results can be found in the Appendix.

2. ABOS for multiple testing rules. Consider a set of m normal populations $\mathcal{N}(\mu_i, \sigma^2)$, $i = 1, \ldots, m$. 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 $X_i = (X_{1i}, \ldots, X_{ni})$ 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

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

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

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

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

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

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

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 [30]. The total loss is clearly minimized by applying the Bayes classifier to each individual test, the decision rule which was called Bayes oracle in [8].

Hence our first task is to determine the critical values a_n and b_n corresponding to the Bayes classifier for each individual test. As noted in [29], 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

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

Let $\delta = \delta_0/\delta_A$ denote the ratio of type I error and type II error losses, and let $f = (1-p)/p$ which serves as a measure of sparsity. In the forthcoming asymptotic analysis we will assume that $m \to \infty$ and that $n = n_m \to \infty$. Furthermore we will allow the parameters $\delta = \delta_m$ and $p = p_m$ to depend on m, whereas σ and ν are kept fixed. For simplicity of notation the index m will be omitted for n, δ and p. The most generic situation will be $p \to 0$, in which case $f \to \infty$. However, theorems are formulated in the more general setting under the following assumption:

Assumption (A): $n \to \infty$, $\delta f \to c \in (0, \infty]$, and $\frac{2\log(\delta f)}{n} \to C$, where $0 \leq C < \infty$.

REMARK 2.1. Under the model assumptions of [8] "signals on the verge of detectability" had to satisfy $\frac{2\log(\delta f)}{u} \to C \in (0,\infty)$, which yielded asymptotic power of the Bayes oracle within $(0,1)$. Here we are concerned with a different situation, where the alternative distribution for μ_i is not necessarily normal and does not depend on p, but the number n of individuals increases to infinity. In this setting the role of u is taken by n. Compared to the assumptions in [8] the major difference is that we additionally consider the case $\frac{2\log(\delta f)}{n} \to C = 0$, which means that the asymptotic power of the Bayes oracle is equal to 1. This additional case covers the interesting scenario where sparsity is of the form $p = m^{-\beta}$, $\beta > 0$, $\log\delta = o(\log m)$ and $n \in (m^{c_1}, m^{c_2})$, for any positive constants $c_1 < c_2$.

The generic situation will be concerned with sparsity and with the loss ratio δ having no dominating influence on the asymptotic results. We formalize this in

Assumption (B): $n \to \infty, p \to 0, \log\delta = o(\log p)$ and δ bounded from above.

If Assumption (B) holds, then $-\frac{2\log p}{n} \to C \geq 0$ is enough to guarantee that Assumption (A) is fulfilled. All theorems in Section 3 are formulated under Assumption (B).

The following assumption imposes a restriction on the measure ν, which will be used throughout this mansucript.

Assumption (C): Let $T := \sigma\sqrt{C}$. We assume that there exists $\epsilon > 0$ such that ν has a positive bounded density ρ with respect to 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.

The following Lemma provides the asymptotic critical points of the Bayes rule for distributions ν satisfying Assumption (C).

LEMMA 2.1. *Let Assumptions (A) and (C) hold. Then the critical values converge with limits*

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

The proof is given in Appendix 6.1.

Notation: Throughout the paper we will make use of the following notation: Let g_n and h_n be two sequences. Then $g_n \sim h_n$ indicates that $\frac{g_n}{h_n} \to 1$ as $n \to \infty$. If $g_n \to 0$ we write $g_n = o_n$.

The following Lemma 2.2 specifies the rate at which a_n and b_n converge to zero in case of $C = 0$.

LEMMA 2.2. *Let Assumptions (A) and (C) hold. If $C = 0$ then the critical values of the Bayes oracle fulfill*

$$\sqrt{n} e^{-\frac{n a_n^2}{2\sigma^2}} \sim \frac{\sqrt{2\pi}\sigma}{f\delta} \rho(0^-) \tag{2.5}$$

and

$$\sqrt{n} e^{-\frac{n b_n^2}{2\sigma^2}} \sim \frac{\sqrt{2\pi}\sigma}{f\delta} \rho(0^+) \ . \tag{2.6}$$

The proof is given in Appendix 6.2.

REMARK 2.2. As shown in the proof of Lemma 2.2, the accuracy of the approximations provided in (2.5) and (2.6) 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 (2.5) and (2.6) can be expressed as $1 + z_n$ with $z_n = o(n^{-1/2} \log n)$.

REMARK 2.3. The results of Lemmas 2.1 and 2.2 generalize the critical value of the Bayes rule specified in [7] and [8]. Note that for $\nu \sim \mathcal{N}(0, \tau^2)$ the "magnitude" of the true signal defined in [8] is given by $u = \frac{n\tau 2}{\sigma^2}$. Thus, according to Lemma 2.1, for $C > 0$ the Bayes classifier rejects the null hypothesis if

$$\frac{n\bar{X}_n^2}{\sigma^2} > \log(u f^2 \delta^2)(1 + o_n) \ \ ,$$

which agrees with the results of [8].

Next consider the case $C = 0$. For normal distribution $\mu_i \sim \mathcal{N}(0, \tau^2)$ it holds that $\rho(0^-) = \frac{1}{\sqrt{2\pi}\tau}$. Taking logarithms of (2.5) we obtain the accurate approximation

$$\frac{na_n^2}{\sigma^2} = 2\log\left(\frac{\delta f\sqrt{n}}{\sqrt{2\pi}\ \sigma\rho(0^-)}\right) + o_n = \log(uf^2\delta^2) + o_n$$

and because of $\rho(0^-) = \rho(0^+)$ the same relation holds for b_n.

To emphasize similarity with the results for normal scale mixture models from [8] we introduce the notation

$$v := n\delta^2 f^2 \ .$$

Then according to Lemmas 2.1 and 2.2 the Bayes oracle threshold values satisfy

$$a_n = -\sigma\sqrt{\frac{\log v}{n}}(1 + o_n) \text{ and } b_n = \sigma\sqrt{\frac{\log v}{n}}(1 + o_n) \ . \tag{2.7}$$

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. Note that for the specified mixture model (2.3) type I error t_1 and type II error 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 compute the asymptotic risk R^B of the Bayes oracle.

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

$$R^B = mp\delta_A\sigma\sqrt{\frac{\log v}{n}}\left(\rho(0^-) + \rho(0^+)\right)(1 + o_n) \tag{2.9}$$

whereas for $0 < C < \infty$

$$R^B = mp\delta_A\ \nu(-T, T)(1 + o_n) \ . \tag{2.10}$$

The proof is given in Appendix 6.3.

Definition: A multiple testing rule is called asymptotically Bayes optimal under sparsity (ABOS) if its risk R satisfies $\frac{R}{R^B} \to 1$ under the conditions of Assumption (A).

2.1. *ABOS of fixed threshold rules.* The next theorem describes which multiple testing rules with fixed threshold are ABOS.

THEOREM 2.2. *Consider the testing rule which rejects H_{0i} if $\bar{X}_i$ falls out of the interval $(\tilde{a}_n, \tilde{b}_n)$, with $\tilde{a}_n < 0$ and $\tilde{b}_n > 0$. Under Assumptions (A) and (C) this rule is ABOS if and only if*

$$\frac{n\tilde{a}_n^2}{\sigma^2} = \log v + z_a \quad \textit{and} \quad \frac{n\tilde{b}_n^2}{\sigma^2} = \log v + z_b \tag{2.11}$$

where

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

and

$$\lim_{n\to\infty} z_a + 2\log\log v = \infty\ , \quad \lim_{n\to\infty} z_b + 2\log\log v = \infty\ . \tag{2.13}$$

The proof is given in Appendix 6.4.

As a simple consequence of Theorem 2.2 we have

COROLLARY 2.1. *Suppose that additional to the assumptions of Theorem 2.2 also Assumption (B) holds. If for $n = n_m$ the sparsity assumption*

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

is fulfilled, then thresholds of the form

$$c_a^2 \sim c_b^2 = \log(nm^2) + \xi, \quad \xi = o(\log(n/p^2)) \tag{2.15}$$

yield multiple testing rules which are ABOS, whenever $2\xi \geq -2\log(mp) + d$ for some arbitrary constant d. In particular this is the case when ξ is a constant.

Proof. Simply observe that $z = \log(nm^2) + \xi - \log(np^{-2}\delta^2)$ fulfills the requirements of Theorem 2.2 under the assumption of the corollary. □

REMARK 2.4. Corollary 2.1 addresses the situation of extreme sparsity, where the number m of tests increases to infinity, but the expected number of true signals remains constant or increases only very slowly with m. If additionally $\log n = o(\log m)$ then Corollary 2.1 implies that the universal threshold $2\log m$ of [18] is ABOS. This extends Remark 3.4 of [8] to the case where the distribution of μ_i under the alternative is not necessarily normal and does not change with m, while the number of individuals n slowly increases with m.

2.2. *BFDR controlling procedures.* One of our main goals is to study ABOS of FDR controlling procedures like the popular Benjamini–Hochberg procedure (BH,[3]). As in [8] the main technical tool to prove ABOS is to approximate the random threshold of BH by the threshold from a rule controlling the Bayesian false discovery rate (BFDR, see [19]). For that reason we will start our discussion here with results on the asymptotic properties of BFDR rules for general distributions of μ_i under the alternative. BFDR is defined as

$$BFDR = P(H_{0i} \text{ is true} | H_{0i} \text{ was rejected}) = \frac{(1-p)t_{1i}}{(1-p)t_{1i} + p\ (1-t_{2i})} \ , \tag{2.16}$$

where t_{1i} and t_{2i} are the probabilities of the corresponding type I and type II errors. Consider a fixed threshold rule based on $\bar{X}_i$ with threshold values $a < 0$ and $b > 0$. Then $t_{1i} = t_1$, $t_{2i} = t_2$, and under the mixture model (2.3)

$$t_1 = \Phi(\sqrt{n}a/\sigma) + 1 - \Phi(\sqrt{n}b/\sigma)$$

and

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

To obtain threshold values $a_n^B < 0$ and $b_n^B > 0$ with BFDR level α we have to solve $\frac{(1-p)t_1}{(1-p)t_1 + p(1-t_2)} = \alpha$, or equivalently

$$\frac{\alpha}{f(1-\alpha)} = \frac{\Phi(\sqrt{n}a_n^B/\sigma) + 1 - \Phi(\sqrt{n}b_n^B/\sigma)}{\int_{\mathbb{R}} (\Phi(\sqrt{n}(a_n^B-\mu)/\sigma) + 1 - \Phi(\sqrt{n}(b_n^B-\mu)/\sigma)) d\nu(\mu)} \ .$$

We will restrict our attention to rules based on symmetric thresholds, such that $a_n^B = -b_n^B$, and use

$$c_B^2 = c_B^2(n) := \frac{n\left(a_n^B\right)^2}{\sigma^2} = \frac{n\left(b_n^B\right)^2}{\sigma^2} \tag{2.17}$$

to denote the corresponding threshold for the scaled test statistics $Z_i = \frac{n\bar{X}_i^2}{\sigma^2}$. Then c_B satisfies the following equation

$$\frac{\alpha}{f(1-\alpha)} = \frac{2(1-\Phi(c_B))}{2 - \int_{\mathbb{R}} [\Phi(c_B + \sqrt{n}\mu/\sigma) + \Phi(c_B - \sqrt{n}\mu/\sigma)] d\nu(\mu)} \ . \tag{2.18}$$

As shown in Lemma 6.4 in Appendix 6.5, $\alpha \in (0, 1-p)$ guarantees existence and uniqueness of a solution c_B for (2.18). The following theorem provides conditions on α, for which the BFDR controlling rule specified in (2.18) is ABOS.

THEOREM 2.3. *Additional to Assumptions (A) and (C) suppose that* $\alpha \in (0, 1-p)$, $\alpha \to \alpha_\infty < 1$, *and*

$$f/\alpha \to \infty, \quad \frac{\log\left(\frac{f}{\alpha}\right)}{n} \to C_0 < \infty \ , \tag{2.19}$$

where C_0 is such that $\nu(-\sigma\sqrt{2C_0}, \sigma\sqrt{2C_0}) < 1$ and ν has no atoms at $\pm\sigma\sqrt{2C_0}$. The threshold value c_B of the rule controlling BFDR at level α is then given by

$$c_B^2 = 2\log\left(\frac{f}{\alpha}\right) - \log\left(2\log\left(\frac{f}{\alpha}\right)\right) + 2\log\left(\frac{\sqrt{2}\ (1-\alpha_\infty)}{\sqrt{\pi}\ C_1}\right) + o_n \ , \tag{2.20}$$

where

$$C_1 = 1 - \nu(-\sigma\sqrt{2C_0}, \sigma\sqrt{2C_0}) \ .$$

The BFDR controlling rule is ABOS if and only if

$$\frac{\log(f\delta\sqrt{n})}{\log(f/\alpha)} \to 1, \quad \textit{and} \quad 2\log(\alpha\delta\sqrt{n}) - \log\log(f/\alpha) \to -\infty \ . \tag{2.21}$$

In that case $C_0 = C/2$ and therefore $C_1 = 1 - \nu(-T, T)$.

The proof is given in Appendix 6.6.

COROLLARY 2.2. If in addition to the assumptions of Theorem 6.7 also Assumption (B) holds then the fixed threshold rule with BFDR at the level $\alpha \propto n^{-1/2}$ is ABOS.

COROLLARY 2.3. If in addition to the assumptions of Theorem 6.7 and Assumption (B) also $\delta \to 0$ and $n \propto -\log p$ then the fixed threshold rule with BFDR equal to $\alpha \in (0,1)$ is ABOS. It is not possible that a BFDR controlling rule is ABOS when both α and δ are constant.

REMARK 2.5. Based on (2.20) straight forward calculations yield the asymptotic type I error of the BFDR rule

$$t_1^B = \frac{C_1\alpha}{(1-\alpha_\infty)f}(1+o_n) \ . \tag{2.22}$$

The BFDR controlling rules discussed in this section require the knowledge of some of the parameters of the unknown mixture distribution and therefore they are not applicable in practice. However, the results on ABOS of the BFDR controlling rules can be used to prove ABOS of some popularly used multiple testing rules, like the Bonferroni correction or the Benjamini–Hochberg procedure. Asymptotic optimality results of these rules will be presented in the following sections.

2.3. *Bonferroni correction.* In applied sciences the most popular multiple testing procedure is still the fixed threshold rule of Bonferroni correction. In our setting its critical value c_{Bon} for the test statistic $\frac{\sqrt{n}|\bar{X}_i|}{\sigma}$ is defined by

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

The procedure controls the family wise error rate at level α. The following lemma specifies the conditions for α under which the Bonferroni procedure is ABOS.

LEMMA 2.3. *Suppose Assumptions (A), (C) and sparsity condition (2.14) hold. The Bonferroni procedure at FWER level α_n is ABOS if α_n satisfies the assumptions of Theorem 6.7.*

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

$$c^2_{Bon} = 2\log\left(\frac{m}{\alpha}\right) - \log\left(2\log\left(\frac{m}{\alpha}\right)\right) + \log(2/\pi) + o_n \ .$$

Comparison of this threshold with the asymptotic approximation to an optimal BFDR control rule (2.17) and (2.20) yields

$$c^2_{Bon} = c^2_B + 2\log mp + O_n(1) \ .$$

From (2.14) it follows easily that $c^2_{Bon} = c^2_B(1 + o_n)$. By assumption, the rule based on the threshold c^2_B is optimal, and hence c^2_{Bon} 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. □

2.4. *FDR controlling procedures.* The Benjamini–Hochberg rule [3], which we will also call step-up FDR controlling procedure, is defined as follows: For the square of the scaled test statistics $Z_i^2 = \frac{n\bar{X}_i^2}{\sigma^2}$ one computes two-sided p-values $p_i = 2(1 - \Phi(|Z_i|))$ which are then ordered $p_{[1]} \le p_{[2]} \le \cdots \le p_{[m]}$. For the step-up procedure at the FDR level α compute

$$k_F := \max\left\{i : \ p_{[i]} \le \frac{i\alpha}{m}\right\} \tag{2.23}$$

and reject the k_F hypothesis with p-values smaller or equal $p_{[k_F]}$. In view of the proof of ABOS for FDR controlling model selection criteria in Section 3.2 we will not only consider the step-up procedure, but also the corresponding step-down procedure at level α. For this compute

$$k_G := \min\left\{i : \ p_{[i]} > \frac{i\alpha}{m}\right\} \tag{2.24}$$

and reject the $k_G - 1$ hypotheses with p-values smaller than $p_{[k_G]}$. It is well known, that in practice both procedures behave very similar (see [1]).

Optimality results for the step-up FDR controlling rule were proven in [8] under the assumption of μ_i being normally distributed. A crucial step was the definition of a random threshold for the BH rule

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

with

$$\tilde{c}_{BH} = \inf\left\{y : \frac{2(1 - \Phi(y))}{1 - \check{F}_m(y)} \le \alpha\right\} \ . \tag{2.25}$$

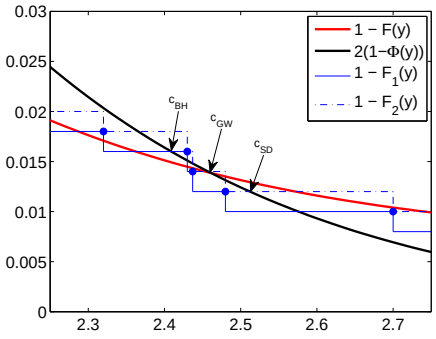


FIG 1. *Comparison of the random thresholds c_{BH} and c_{SD} with the nonrandom threshold c_{GW}. In the legend F_1 refers to $\check{F}_m$ and F_2 refers to $\hat{F}$.*

Here $1-\check{F}_m(y) = \#\{|Z_i| \geq y\}/m$. Alternatively let us denote $1-\hat{F}_m(y) = \#\{|Z_i| > y\}/m$. Similar as in case of BH it is easy to check that SD rejects the null hypothesis H_{0i} if and only if $Z_i^2 \geq c_{SD}^2$ where

$$c_{SD} = \sup\left\{ y : \frac{2(1-\Phi(y))}{1-\hat{F}_m(y)+1/m} > \alpha \right\} \quad . \tag{2.26}$$

It was proven by Genovese and Wassermann (GW) in [25] that for fixed p, as the number of tests increases, the random threshold c_{BH} can be approximated by the non-random threshold

$$c_{GW} : \frac{2(1-\Phi(c_{GW}))}{1-F(c_{GW})} = \alpha \quad , \tag{2.27}$$

where $F(y) = P(|Z_1| \leq y)$.

Figure 1 illustrates the thresholds c_{BH}, c_{SD} and c_{GW}. Comparing $\tilde{c}_{BH}$ and c_{SD} with c_{GW} the major change is in replacing the cumulative distribution function of $|Z_i|$ by the corresponding empirical distribution function. In [8] it was shown that also in case of sparsity c_{BH} can be well approximated by c_{GW}, and in Lemma 6.5 of Appendix 6.7 we will see that the same is true for c_{SD}. A much simpler result is that under sparsity the difference between c_{GW} and the corresponding BFDR controlling threshold c_B becomes asymptotically negligible.

THEOREM 2.4. *Suppose Assumptions (A) and (C) are true and that $p \to 0$. Consider the rule rejecting the null hypothesis H_{0i} if $\frac{n\bar{X}_i^2}{\sigma^2} \geq c_{GW}^2$. This rule is ABOS if and only if the corresponding BFDR controlling rule defined in (2.18) (for the same $\alpha = \alpha_n$) is ABOS. In this case we have*

$$c_{GW}^2 = c_B^2 + o_n \quad .$$

Proof. The proof of this statement follows exactly as the proof of Theorem 4.2 of [8]. □

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

THEOREM 2.5. *Apart from Assumptions (A) and (C) assume that*

$$mp \to s \in (0, \infty] \tag{2.28}$$

and

$$\alpha \text{ satisfies the conditions of Theorem 6.7,} \tag{2.29}$$

i.e the BFDR control rule at level α is asymptotically optimal.

For the denser case

$$p > \frac{\log^{\gamma_1} m}{m} \ , \text{ for some constant } \gamma_1 > 1 \tag{2.30}$$

the additional assumptions

$$n \leq m^{\gamma_2}, \quad \text{for some } \gamma_2 > 0 \text{ and } \ \frac{\log \log m}{\log(p\ \alpha)} \to 0 \tag{2.31}$$

should hold. Then both BH and SD are ABOS.

Proof. 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 (2.31) the proof for Type 1 error of BH follows along the same lines as the proof of Lemma 5.4 in [8]. Also, under the condition of extreme sparsity (2.14) according to Lemma 2.3 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. Finally, the necessary bound of the type II error component of the risk of SD for the denser case (2.30) is provided in Appendix 6.7. This proof substantially relies on the second condition in (2.31). □

REMARK 2.6. The upper bound on m provided in the second condition of (2.31) 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

$$n \geq m^{\gamma_3} \quad \text{for some } \ \gamma_3 > 0 \ . \tag{2.32}$$

(It is easy to show that (2.32) implies the upper bound on m in (2.31) given the other assumptions of Theorem 2.5) .

The following Corollaries are easy consequences of Theorem 2.5.

COROLLARY 2.4. *Suppose Assumptions (A) and (C) hold. If $p = m^{-\beta}$ with $\beta \in (0, 1]$, $n \leq m^{\gamma_2}$ for some $\gamma_2 > 0$ and δ is bounded from above such that $\log \delta = o(\log m)$ then BH and SD at FDR level $\alpha \propto n^{-1/2}$ are ABOS.*

COROLLARY 2.5. *Suppose Assumptions (A) and (C) hold. If* $p = m^{-\beta}$ *with* $\beta \in (0,1]$, $n \propto \log m$ *and* δ *converges to zero such that* $\log \delta = o(\log m)$ *then BH and SD at a fixed FDR level* $\alpha \in (0,1)$ *are ABOS.*

REMARK 2.7. Corollary (2.4) states that under some mild restrictions on δ BH and SD at the FDR level $\alpha \propto n^{-1/2}$ are ABOS. Corollary (2.5) says that in case when $n \propto \log m$ then under the additional requirement that $\delta \to 0$, BH and SD at the fixed FDR level $\alpha \in (0,1)$ are also ABOS. This result substantially extends the results of [8] to the case where the prior on μ_i is fixed and not necessarily 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 [8] the condition $\delta \to 0$ is quite reasonable in this context, since the cost of missing a true positive is usually large if p is very small.

3. ABOS in the context of multiple regression. It is well known [8, 23] that there is a strong connection between model selection for multiple regression and multiple testing rules. Under the simplified assumption of an orthogonal design matrix and known variance of the error term the two problems actually become identical. Consider a multiple linear regression model

$$Y_{n\times 1} = X_{n\times(m+1)}\beta_{(m+1)\times 1} + \epsilon_{n\times 1} \ ,$$

where the first column in the design matrix consists of ones and $\epsilon \sim N\left(0, \sigma^2 I_{n\times n}\right)$. Let us additionally assume that

$$X'X = nI_{(m+1)\times(m+1)} \ , \tag{3.33}$$

and that the regression coefficients $\beta_1, \ldots, \beta_m$ can be modelled as independent random variables from the following mixture distribution

$$(1-p)d_0 + p\nu \ . \tag{3.34}$$

Under the assumptions (3.33) and (3.34) least squares estimates $\hat{\beta}_i$, $1 \le i \le m$, are independent random variables from the mixture distribution

$$(1-p)N\left(0, \frac{\sigma^2}{n}\right) + p\left(\nu * N\left(0, \frac{\sigma^2}{n}\right)\right) \ . \tag{3.35}$$

This is identical with (2.3) and thus the problem of detecting true regressors is equivalent to the multiple testing problem. Therefore, in case when each false positive (falsely detected regressor) induces the cost δ_0 and each false negative induces the cost δ_A, thresholds of the Bayes rule and the optimal Bayes risk are obtained just like in Lemma 2.2 and in Theorem 2.1.

As mentioned in the introduction we will focus here on the case of Assumption (B), where the loss ratio has no particular influence on the asymptotic results. In this case $\frac{2\log f\delta}{n} = -\frac{2\log p}{n}(1+o_n)$. We also consider only sparsity parameters $p \to 0$ satisfying assumption (2.28). Since under orthogonal designs $m < n$ one has $-\log p = O(\log m) = O(\log n)$, and finally $-\frac{2\log p(1+o_n)}{n} \to 0$. Thus,

under orthogonal designs assumptions (B) and (2.28) imply Assumption (A) with $C = 0$. Therefore we will refrain from referring to Assumption (A) in this section.

We will first discuss a model selection criterion which is ABOS in case of extreme sparsity (2.14), as in Corollary 2.1. However, it is easy to see that for $m \leq n$ that sparsity assumption reduces to

$$mp \to s \in (0, \infty], \quad \frac{\log(mp)}{\log n} \to 0 \; . \tag{3.36}$$

3.1. *ABOS of mBIC when σ is known.* It was shown in [5] in the context of QTL mapping that for large m classical model selection criteria like AIC or BIC tend to select too large models. Based on Bayesian ideas a modified version of BIC (mBIC) was proposed to take into account the number of available regressors. When σ is known the mBIC criterion suggests choosing the model M for which

$$\frac{RSS_M}{\sigma^2} + k(\log n + 2 \log m + d) \tag{3.37}$$

obtains a minimum, where RSS_M refers to the residual sum of squares for model M, $k = k(M)$ is the number of regressors in the model and d is a certain constant. A comprehensive introduction into the ideas leading to mBIC is given in [9].

REMARK 3.1. It follows from the derivation of mBIC that from a Bayesian perspective $\exp(-d/2)$ is the a priori expected number of regressors. If there is no prior knowledge on the model size the recommended standard choice is $d = -2\log(4)$, which guarantees control of FWER at level 0.1 for $n \geq 200$ and $m \geq 10$. For further details see [9].

Apart from ABOS we want to show consistency of mBIC.

Definition. A model selection rule is said to be consistent if the probability of selecting the true model converges to 1 as $m \to \infty$.

THEOREM 3.1. *Consider the orthogonal regression model specified by the conditions (3.33) and (3.34) and let assumptions (B) and (C) (with C=0) hold. Under (3.36) mBIC is ABOS, while under the considerably weaker assumption*

$$mp \to s \in (0, \infty], \quad mp\sqrt{\frac{\log n}{n}} \to 0 \tag{3.38}$$

mBIC is consistent.

Proof. It is easy to check that under assumption (3.33) mBIC suggests choosing those regressors for which

$$\frac{n\hat{\beta}_j^2}{\sigma^2} > \log n + 2\log m + d \; .$$

From Corollary 2.1 one immediately concludes that under the sparsity assumption (3.36) this selection rule is ABOS.

To prove consistency of mBIC let the random variable M_j be Bernoulli distributed where a misclassification of predictor X_j denotes a success. If t_1 and t_2 denote the probability of type I and type II error of mBIC, respectively, then for sufficiently large n

$$P(M_j = 1) = (1-p)t_1 + pt_2 \leq Kp\sqrt{\frac{\log n}{n}}$$

for some constant K, where the last inequality is shown in Appendix 6.8. Using Markov's inequality the probability of picking the wrong model (which is the probability of at least one wrong misclassification) can thus be bounded like

$$P(\sum_{j=1}^{m} M_j \geq 1) \leq E\left(\sum_{j=1}^{m} M_j\right) \leq Kmp\sqrt{\frac{\log n}{n}}\,, \tag{3.39}$$

which according to (3.38) converges to 0. □

REMARK 3.2. Theorem 3.1 addresses the situation of sparsity, where the expected number of true signals remains constant or slowly increases with m. The assumption $mp \to s < \infty$ was used when deriving the mBIC penalty in [5]. Theorem 3.1 actually tells us that mBIC remains optimal when the number of true signals is mildly growing, for example $mp = \log m$ is still conceivable. This scenario might be more realistic in many applications, where one would hope that by increasing the number of markers one could actually be able to detect more true signals. However, the situation described is still very sparse, which is one motivation to introduce in Section 3.2 criteria which are slightly less restrictive.

REMARK 3.3. Note that under the assumption $mp \to s < \infty$ the expected value of the number of false positives EP produced by the standard BIC is equal to $EP = m(1-p)t_1 = \frac{m}{\sqrt{n \log n}}(1 + o_{n,m})$. Thus BIC is not consistent when $\lim_{n\to\infty} \frac{m}{\sqrt{n \log n}} > 0$.

REMARK 3.4. Another interesting situation arises for distributions ν for which there exists an open interval including 0 such that $\nu(-l, r) = 0$ (cf. [29]). It can be shown that in this situation the mBIC rule is not optimal anymore, although its risk still converges to 0.

3.2. *Modifications of BIC controlling FDR.* As shown in [9] there exists a close connection between mBIC penalty and the Bonferroni correction for multiple testing. In a recent paper [1] Abramovich et al. have been discussing extensively penalized model selection schemes which control the false discovery rate. Their starting point is the close relationship between step-up and step-down

FDR controlling procedures at level α and the following penalizing scheme: For models of size k define the selection criterion

$$\frac{RSS_M}{\sigma^2} + \sum_{l=1}^{k} q_N^2(\alpha l/2m) \ , \tag{3.40}$$

where $q_N(\eta)$ is the $(1-\eta)$ - quantile of the standard normal distribution. It can be shown quite easily that the size of models selected by this procedure is larger or equal k_G and smaller or equal k_F (see [1]). The procedure is therefore nested between BH and SD, and from Theorem 2.5 it immediately follows that it is also ABOS.

We will adopt approximations of the FDR penalization (3.40) to amend BIC. A simple argument involving the normal tail approximation shows that

$$q_N^2(\alpha l/2m) \sim 2\log(m/l) - \log[2\log(m/\alpha l)] + \log(2/\pi) - 2\log\alpha \ .$$

In view of Corollary 2.2 we are mainly interested in the case where $\alpha \propto n^{-1/2}$ which leads to the criterion

$$\text{mBIC1:} \quad \frac{RSS_M}{\sigma^2} + k(\log(nm^2) + d_1) - 2\log(k!) - \sum_{i=1}^{k} \log\log(nm^2/i^2) \ . \tag{3.41}$$

Here the constant d_1 can be chosen appropriately to control FDR at a given level. Neglecting the last term of the mBIC1 penalty, which is of a lower order than the two preceding terms, leads to the following simplified form of (3.41),

$$\text{mBIC2:} \quad \frac{RSS_M}{\sigma^2} + k(\log(nm^2) + d_2) - 2\log(k!) \ . \tag{3.42}$$

This might be thought of as a first order approximation of the FDR penalization, whereas mBIC1 is a second order approximation. Interestingly, the penalty in mBIC2 is very similar to a modification of RIC introduced in [26], with additional penalty term

$$2\sum_{i=1}^{k} \log(m/i) = k\log(m^2) - 2\log(k!) \ ,$$

which was motivated by an empirical Bayes approach.

Abramovich et al. consider in [1] the approximation $\sum_{l=1}^{k} q_N^2(\alpha l/2m) \sim k q_N^2(\alpha k/2m)$, which can be justified by using the Sterling approximation for $k!$. The resulting first order criterion has the form

$$\text{mBIC3:} \quad \frac{RSS_M}{\sigma^2} + k(\log(nm^2) + d_3) - 2k\log(k) \ . \tag{3.43}$$

Compared with (3.42) this means essentially that $\log(k!)$ is substituted by $\log(k^k)$.

REMARK 3.5. In the simulation study of Section 4 the constant of mBIC1 is chosen as $d_1 = 0$, which guarantees control of FDR at a level below 0.06 for sample size n larger than 200. For mBIC2 the constant $d_2 = -2\log(4)$ is used, which coincides with the recommended standard choice of d for mBIC. For moderate m and n as in the simulation study mBIC1 with $d_1 = 0$ and mBIC2 with $d_2 = -2\log(4)$ have rather similar penalties for small k. In case of mBIC3 the Sterling approximation leads to $d_3 = d_2 + 2$.

THEOREM 3.2. *Consider the orthogonal regression model specified by the conditions (3.33) and (3.34). Let Assumptions (B) and (C) as well as (2.28) be true. For the denser case (2.30) the additional condition (2.31) is assumed to hold. Then the rules mBIC1, mBIC2 and mBIC3 are ABOS. The rules are consistent under the additional assumption (3.38).*

The proof is given in Appendix 6.9.

REMARK 3.6. The FDR controlling selection rules mBIC1 - mBIC3 are ABOS under much less restrictions on the sparsity levels than mBIC. However, conditions for consistency are exactly the same. Actually given the other assumptions of Theorem 3.2 it follows that (3.38) is also necessary for the Bayes rule to be consistent.

3.3. *ABOS of mBIC when σ is unknown.* We have seen that for known σ and under the simplifying assumption of an orthogonal design matrix, the problem of model selection using mBIC in multiple regression is equivalent to multiple testing, in the sense that a regressor is included in the model chosen by mBIC if and only if the corresponding square of the sample regression coefficient is larger than a fixed threshold. In case of unknown σ the situation gets much more complicated and no such direct connection with multiple testing can be established. We are only interested in the comparison of models which include the intercept. In this case the Bayesian Information Criterion chooses that model which minimizes $BIC = n\log RSS_M + k\log n$. The corresponding version of mBIC becomes

$$mBIC = n\log RSS_M + k(\log n + 2\log m + d) \quad . \tag{3.44}$$

Our main goal is to show that also in case of unknown σ mBIC is asymptotically optimal.

Some problem occurs when (3.44) is used as a selection criterion for very large models. To be able to estimate the parameters of a model M we need the restriction that $k \leq n-2$. But if k is getting close to n then overfitting will lead to extremely small $\log RSS_M$, and the global minimum of (3.44) is likely to be attained by models of maximum size $k = n - 2$ (if that many regressors are available). It can be ruled out that such models are correct under the assumption of sparsity. To cope with this pathology we will restrict L, the maximal number of regressors to be allowed in addition to the common intercept term, by

$$L = o\left(\frac{n}{(\log n + 2\log m)^2 \log m}\right) \text{ as } n \to \infty \quad . \tag{3.45}$$

On the other hand to bound the type II error it is necessary to search among sufficiently large models, and we require the lower bound

$$L \geq mp(\log n)^{1+\eta} \text{ for some } \eta > 0 \text{ and all sufficiently large } n \,. \tag{3.46}$$

THEOREM 3.3. *Suppose as in Theorem 3.1 that Assumptions (B) and (C), (3.33), (3.34) hold. Furthermore assume that (3.45) and (3.46) are true. Then the mBIC criterion (3.44) is ABOS under (3.36), and consistent under (3.38).*

The somewhat lengthy proof of this theorem is provided in Appendix 6.10.

REMARK 3.7. Note that except for the conditions (3.45) and (3.46) on the potential model size L the assumptions for ABOS of mBIC in case of unknown σ are exactly the same as in Theorem 3.1 for known σ. We conjecture that similarly the results of Theorem 3.2 concerning ABOS of the FDR controlling modifications of BIC should also hold in case of unknown σ. However, the techniques used for the proof of Theorem 3.3 cannot easily be extended to mBIC1 - mBIC3. We will come back to this point in the simulation study in the next section.

4. Simulation results. We employ computer simulations to investigate the performance of the proposed model selection rules for multiple regression. For the sake of simple notation in this section m denotes the number of regressors plus intercept. We use orthogonal designs with $n = m$, where the design matrices $X_{m\times m}$ are chosen as Hadamard matrices, whose elements are equal to 1 or -1. For each of the simulation runs the number of nonzero regression coefficients k^* was simulated from a binomial distribution $B(m, p)$. Then the values of nonzero coefficients $\beta_1, \ldots, \beta_{k^*}$ were simulated from a normal distribution $N(0, \tau^2)$, with $\tau^2 = 0.9$. Finally the values of the response variable were simulated according to the multiple regression model

$$Y_i = \sum_{j=1}^{k^*} \beta_j X_{ij} + \epsilon_j \;\;,$$

where $\epsilon_j \sim N(0, 1)$. The specific value of the variance of regression coefficients $\tau^2 = 0.9$ is selected in such a way that the power of the Bayes oracle for $m = 256$ is in the range between 50% and 60%. This choice allows to assess differences in performance of the considered model selection rules.

In the first part of the simulation study we consider sparsity parameters $p \in \{0.001, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2\}$ and simulate for $m = 256$ as well as $m = 1024$. In the second part we will look at a wider range of sample sizes $n = m \in \{128, 256, 512, 1024, 2048, 4096\}$, while the sparsity parameters are computed according to $p \propto m^{-\beta}$ for four different levels $\beta \in \{1, 1/2, 1/4, 1/8\}$.

We compared the following model selection criteria:

1. The Bayes Oracle (2.4) with $\delta_0 = \delta_A$. This oracle is aimed at minimizing the expected number of wrongly classified regressors and in our setting

includes those explanatory variables for which

$$n\hat{\beta}_i^2 > \frac{n\tau^2+1}{n\tau^2}\left(\log\left(n\tau^2+1\right)+2\log\left(\frac{1-p}{p}\right)\right) \quad . \tag{4.47}$$

2. Modified versions of Bayesian information criterion:
   (a) mBIC: (3.37) with $d = -2\log 4$
   (b) mBIC1: (3.41) with $d_1 = 0$
   (c) mBIC2: (3.42) with $d_2 = -2\log 4$
   (d) mBIC3: (3.43) with $d_3 = -2\log 4 + 2$

   The values of the constants are chosen according to Remark 3.1 and Remark 3.5.
3. Step up and step down FDR controlling procedures, (2.23) and (2.24) at FDR levels $\alpha = 0.05$. These procedures test individually each of the regression coefficients based on simple regression models.

Modified versions of BIC and FDR controlling procedures are investigated under two scenarios: when σ is known and when it is unknown. In case when σ is unknown modified versions of BIC are based on $n\log RSS_M$ instead of $\frac{RSS_M}{\sigma^2}$ (see (3.37) and (3.44)). For unknown σ the FDR controlling procedures are based on t-tests instead of z-tests.

To identify the regression models, which are "best" with respect to our model selection criteria, we start with ordering explanatory variables based on the results of simple regression t-tests. This procedure gives us the proper sequence of nested models, since under the orthogonal design the estimate of a regression coefficient for a given explanatory variable does not depend on the other regressors included in the model. Then we compare values of model selection criteria for these nested models, starting from the null model, with no explanatory variables, and ending with a model of dimension $k_{max} = 0.3m$. The need for using the bound on the maximal number of components in the considered models results from the fact that under our design the residual sum of squares for the full model is equal to 0. Therefore, in case of unknown σ, all modified versions of BIC are optimized by the full model (see the discussion before introducing assumption (3.45)). Despite of this, according to Theorem 3.3 and the results of [14] on the consistency of similar model selection rules, we expect that our model selection criteria are consistent if the true design is sparse and k_{max} goes to infinity at a slower rate than m. The choice $k_{max} = 0.3m$ corresponds to the expected upper bound of model sizes for the sparsity level $p = 0.2$.

For all considered procedures we report several characteristics, which are calculated based on 10000 replicates. For each of these replications we compute the number of chosen variables that do not appear in the true model (false positives, FP) and the number of true regressors which were not detected (false negatives, FN). These values are used to calculate the following statistics:

1. Misclassification probability: $\mathrm{MP} = (\mathrm{FP} + \mathrm{FN})/(m-1)$.
2. False discovery rate: $\mathrm{FDR} = \frac{\mathrm{FP}}{\mathrm{FP}+\ k^*-\mathrm{FN}}$, or 0 in case of no discoveries.
3. Power $= \frac{k^*-\mathrm{FN}}{k^*}$ (cases for which $k^* = 0$ are excluded from this analysis).

For each scenario the values of MP, FDR and Power are averaged over all 10000 simulations.

4.1. *First part of Simulation.* The results of this part of the simulation study are illustrated in Figure 3 and Figure 4 in Appendix 6.11. Figure 3 presents the graphs of our computed characteristics as functions of the sparsity parameter p in case of known σ. The two plots (a) and (b) of the first line show that, as expected, the Bayes oracle has the lowest misclassification probability MP. However, the differences in MP between the Bayes oracle and FDR controlling procedures, as well as mBIC1-mBIC3, are hardly observable. For $p < 0.05$ also MP of mBIC is comparable to and sometimes even better than MP of other criteria. However, for $p = 0.05$ differences become observable, and for $p > 0.05$ MP of mBIC substantially exceeds the values reported for other methods. Qualitatively there is no different behavior in the plots for $m = 256$ and $m = 1024$, though it is clear that MP gets smaller for larger sample size. These observations agree well with our results on the asymptotic optimality of mBIC in case of extreme sparsity, and of the FDR controlling procedures and mBIC1-mBIC3 in a wider range of sparsity levels. Apparently our asymptotic analysis describes the situation already quite well for $m = 256$.

Plots (c) and (d) of Figure 3 show the FDR of different procedures. FDR of the Bayes oracle increases from 0 for $p = 0$ to 0.08 for $p = 0.2$ in case of $m = 256$, and to 0.03 for $m = 1024$. As expected, FDR of both step up and step down multiple testing procedures slowly decreases from approximately 0.05 for $p = 0$ to 0.04 for $p = 0.2$ independently of the sample size. The same pattern is observed for the first modified version of BIC aimed at controlling FDR, mBIC1. For $m = 256$ its FDR behaves almost identical to BH, whereas for $m = 1024$ FDR starts at 0.03 and decreases to 0.02. FDR of mBIC2 and mBIC3 behave quite differently in case of extreme sparsity. Due to the choice of constants d_1 and d_2, FDR of mBIC2 is close to FDR of mBIC1 for small p. In contrast mBIC3 has extremely small FDR for p close to 0, which is due to the fact that for small k Sterling's approximation is not valid. For larger p (resulting in the choice of larger models) mBIC2 and mBIC3 behave more and more similar, and their FDR stabilizes at a level of approximately 0.05 for $m = 256$ and at 0.025 for $m = 1024$, being thus slightly larger than FDR of mBIC1. Finally FDR of the modified version of BIC aimed at controlling the Family Wise Error Rate, mBIC, quickly decreases; for $m = 256$ from approximately 0.043 for $p = 0$ to 0.0015 for $p = 0.2$, and for $m = 1024$ from approximately 0.015 down to 0.001.

The pattern of the graphs (e) and (f) for power corresponds to the behavior of FDR. At $p = 0.001$ clearly the Bayes oracle has smallest power. In case of $m = 256$ for $p \geq 0.01$ the power of the Bayes oracle exceeds the power of other model selection criteria, whereas for $m = 1024$ BH and SD have largest power. However, the differences of power between all criteria apart from mBIC are very small and for $p > 0.001$ do not exceed 4%. Also, it is interesting to observe that the power of these criteria slowly increases with p. mBIC performs substantially different than the other methods. Its power is significantly smaller and remains constant as a function of p. Graphs (e) and (f) illustrate also that as expected power increases with sample size.

In Figure 4 the results for unknown σ are reported. The most obvious difference between the case of known and unknown σ is observed for the multiple testing procedures based on simple regression tests. FDR of these procedures is close to the nominal level of 0.05 only when p is very close to 0. For larger values of p other important regressors inflate the residual error in simple regression tests, which leads to a very low power, low FDR and large misclassification rate. As a consequence, when σ is unknown simple regression tests perform substantially worse than other methods based on model selection strategies. This finding has been discussed extensively in [24] in the context of genome wide association studies.

Concerning modified versions of BIC the performance of mBIC1-mBIC3 is only slightly affected by the fact that σ is unknown when $p \leq 0.1$. However, for $p = 0.2$ and $m = 256$ one observes a significant increase of FDR and MP when compared to the known σ case. In particular mBIC2 and mBIC3 have a sudden increase of FDR which results also in a significantly larger MP than that of the Bayes rule. mBIC1 suffers from the same problem, though to a lesser extent. Thus for larger p the second order approximation in mBIC1 proves beneficial.

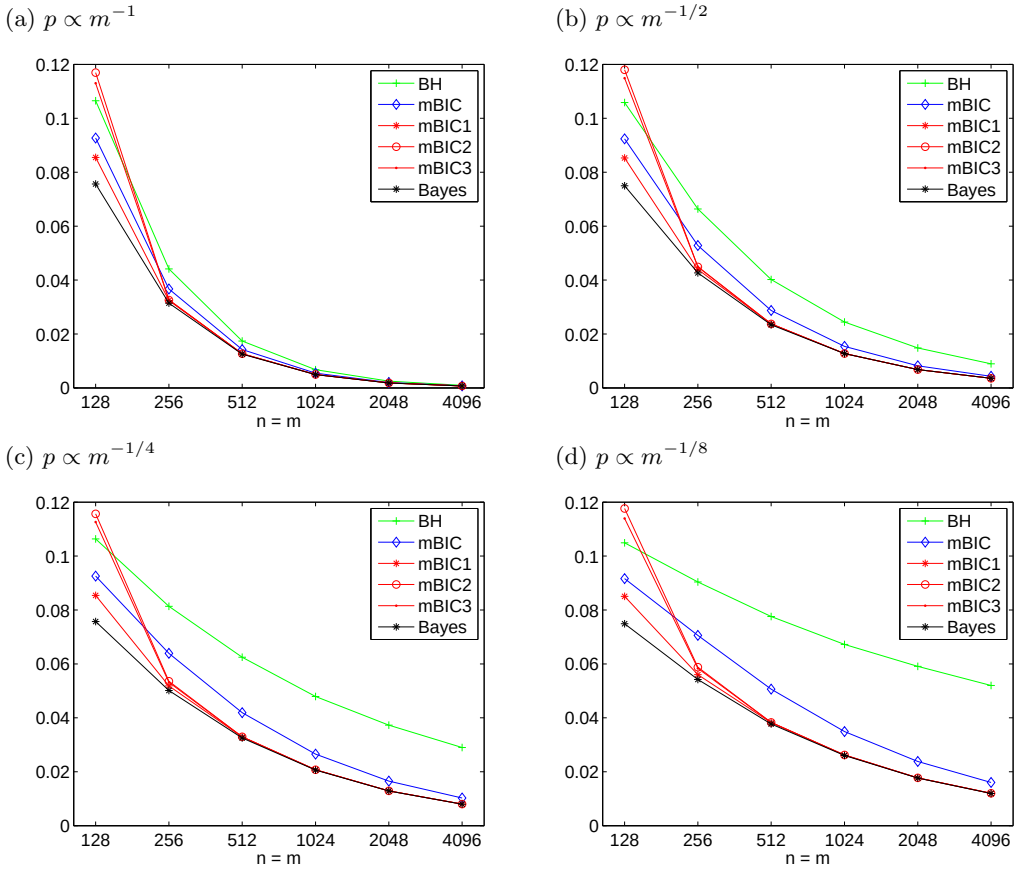
While mBIC2 and mBIC3 are getting for larger p too liberal, mBIC has the opposite tendency. Especially for $m = 256$ the fact that σ is unknown leads to a substantial decrease of power and FDR for large values of p. For $m = 1024$ the relative performance of mBIC substantially improves and is only slightly worse than for known σ. However, both in terms of power and MP mBIC is still performing much worse than mBIC1 - mBIC3.

4.2. *Second part of Simulation.* Here we want to assess numerically the asymptotic behavior which was analyzed theoretically in Section 3. To this end we will perform similar computations as above, but consider the wider range of sample sizes $n = m \in \{128, 256, 512, 1024, 2048, 4096\}$. The sparsity parameter is computed as $p = c_\beta m^{-\beta}$, where we analyze the extremely sparse case $\beta = 1$ as well as $\beta \in \{1/2, 1/4, 1/8\}$. For each scenario the factor c_β is chosen such that for $m = 128$ we always have $p = 0.125$. The misclassification probability for the four different scenarios and for the various methods are provided in Figure 2. We no longer consider SD, as it has been seen before to behave more or less identical with BH. We also present here only the case of unknown σ, which is of particular interest in view of the unproven conjecture that mBIC1 - mBIC3 will be ABOS for a wider range of sparsity levels than mBIC.

For $m = 128$ (and $p = 0.125$) mBIC1 has lower misclassification rate than all other criteria. mBIC2 and mBIC3 have relatively large misclassification rate, and are performing worse than mBIC. We had seen this behavior before already for $m = 256$ and $p = 0.2$. If there are relatively many true signals and m is small then mBIC2 and mBIC3 tend to be slightly too liberal.

For $\beta = 1$ the misclassification rate of all procedures converges towards that of the optimal Bayes rule. In particular it is confirmed that mBIC is ABOS in case of extreme sparsity, although mBIC1 - mBIC3 perform even better. In case of extreme sparsity it seems that even BH behaves relatively well. For smaller β a multiple testing approach is not suitable in case of unknown σ as we discussed already above.

FIG 2. *Asymptotic behavior of the misclassification rate MP at sparsity $p \propto m^{-\beta}$ for different values of β.*



The smaller β, the poorer becomes the performance of mBIC. Although it seems that its misclassification rate still converges towards that of the Bayes rule, this is only true in absolute terms. Already for $\beta = 1/2$ the ratio of the misclassification rates between BH and the Bayes rule remains more or less constant at 1.2. For $\beta = 1/8$ this ratio is actually growing, and mBIC is certainly not optimal. On the other hand MP of mBIC1 - mBIC3 rapidly converges towards MP of the Bayes rule in all four scenarios, which supports our conjecture that an analogue of Theorem 3.2 should also hold in case of unknown σ. Finally Figure 2 suggests that regardless of the sparsity level β all modifications of BIC are consistent selection rules in the asymptotic framework of Assumption B.

5. Discussion. The first part of this paper generalizes optimality results of [8] for multiple testing procedures. Instead of scaled normal distributions we consider models of a larger class of distributions under the alternative. Only point null hypotheses are considered and the measure under the alternative is kept fixed. The asymptotics is thus not driven by a scaling parameter which determines the effect size, but rather by the sample size n which is assumed to

become large. In that context we study two situations: The "verge of detectability" case as in [8], where the power of the Bayes oracle is positive but less than 1. In this article the notion of "the verge of detectability" is extended to the practically important case where the distribution of the effect size is fixed and the sample size n slowly increases with the number of tests m. When sparsity is of the form $p \propto m^{-\beta}$ and the ratio of losses δ is bounded from above, then the "verge of detectability" is obtained when n grows proportionally to $\log m$. The second analyzed case is concerned with asymptotic power equal to 1, which is naturally associated with the situation where n grows faster than $\log m$.

In both cases all optimality results of [8] could be proved for the considered general class of distributions, where in the second case the analysis is slightly more involved and some additional mild restrictions on the asymptotic behavior of the loss ratio δ are necessary. In particular it was shown that the Bonferroni selection rule is ABOS in case of extreme sparsity, whereas the Benjamini–Hochberg rule is ABOS under a much wider range of sparsity levels. Thus results of [8] have been extended to many practically important cases, where the distribution of the true effects is not symmetric. A new result is that the step down version of the FDR - controlling procedure is ABOS under almost the same conditions as BH.

Optimality results were then transferred into the context of linear regression. The simplest situation is concerned with orthogonal regressors and known error variance σ^2, where optimality results from multiple testing can be directly applied. We analyzed the performance of mBIC, a modification of BIC which was previously introduced for model selection in high dimensional data [5], and which is known to control the family wise error rate under the given conditions [9]. It turns out that mBIC is ABOS in case of extreme sparsity, namely under the same conditions as the Bonferroni selection rule for multiple testing. Additionally three different FDR-controlling modifications of BIC were introduced. Optimality results for these selection rules, mBIC1 - mBIC3, entirely correspond to results for the step up and step down FDR controlling procedures in multiple testing. Thus mBIC1 - mBIC3 are ABOS under a much wider range of sparsity levels than mBIC. All modified versions of BIC (including mBIC) are consistent under the same assumption on sparsity levels which guarantee consistency of the Bayes oracle.

Next we showed ABOS of mBIC under extreme sparsity in case of unknown σ, a situation which is technically much more demanding than the previous case of known σ. We conjecture that in analogy to the known σ case, mBIC1 - mBIC3 should be ABOS when removing the extreme sparsity restriction. While we were not able to give a formal proof, simulation results strongly support this conjecture. Furthermore mBIC in case of unknown σ is consistent under the same conditions on sparsity levels under which the Bayes oracle is consistent. The same is expected to hold for mBIC1 - mBIC3. Apart from our simulation study, consistency of the modified versions of BIC for unknown σ can also be conjectured based on recent consistency results for the extended version of Bayesian Information Criterion, EBIC, reported in [13] and [14]. As discussed in [40], if the dimension of the maximal allowable model k_{max} satisfies $k_{max}/m \to \infty$ then mBIC2 is asymptotically equivalent to the standard version

of EBIC, based on a uniform prior on the model dimension. It follows that mBIC2 can be interpreted as an approximation of the Bayesian rule, in which the prior on the true number of regressors is uniform over the set $\{0, \ldots, k_{max}\}$, with $k_{max} = o(m)$.

The results presented in this article are important to understand optimality of model selection criteria under sparsity. However, they are somewhat preliminary as they are only considering the case of orthogonal regressors. In most applications where sparsity is an issue one is also dealing with $m > n$, that is the number of regressors exceeds the sample size. Our current analysis is explicitly not applicable to this situation. However, we believe that the majority of results can be extended to the case $m > n$ if the design matrix satisfies certain conditions for identifiability of small models, which are discussed for example in [11], [4] or [14]. These expectations are confirmed by the successful application of mBIC2 to analyze genome wide association study data, as reported in [24]. Theoretical analysis of asymptotic optimality properties of modifications of BIC under non-orthogonal designs is the topic of further research.

Acknowledgment. We want to thank Professor Jayanta K. Ghosh for many discussions and guidance.

This work is partially funded by by the WWTF grant MA09-007a for F. Frommlet and by grant 1 P03A 01430 of the Polish Ministry of Science and Higher Education for M. Bogdan.

6. Appendix.

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

LEMMA 6.1. *Let $a_n \to a$ be any convergent sequence. Define $h_n(\mu) := \exp\left(a_n \frac{\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right)$ and $h(\mu) := \exp\left(a \frac{\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right)$. Then*

$$\lim_{n\to\infty} \|h_n\|_{L^n(\nu)} = \|h\|_{L^\infty(\nu)} \ . \tag{6.48}$$

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 (6.48) easily follows by the triangle inequality and the fact that $\lim_{n\to\infty} \|h\|_{L^n(\nu)} = \|h\|_{L^\infty(\nu)}$. □

Now we are ready to prove Lemma 2.1.

Proof.

Let $h_n(\mu) = \exp\left(a_n \frac{\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right)$. 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 (6.1) for any subsequence $a_j \to a$ it holds

$$\lim_j \|h_j\|_{L^j(\nu)} = \left\| \exp\left(a \frac{\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right) \right\|_{L^\infty(\nu)} \ . \tag{6.49}$$

If $a \in S$ then $\|h\|_{L^\infty(\nu)} = \exp\left(\frac{a^2}{2\sigma^2}\right)$ and taking logarithms yields $a = -\sqrt{C}\sigma$. Thus the only potential accumulation point of a_n within S is $-T$. To complete the proof of Lemma 2.1 we will show that $a \notin S$ leads to a contradiction with Assumption (C).

If $a \notin S$ then $a \in (l_a, r_a)$ where $l_a < r_a$ are the boundaries of S, closest to a. It is immediately clear that either

$$\|h\|_{L^\infty(\nu)} = h(l_a) \tag{6.50}$$

or

$$\|h\|_{L^\infty(\nu)} = h(r_a) \ . \tag{6.51}$$

The maximum is taken on the right boundary (6.51) when $a \in (\frac{l_a+r_a}{2}, r_a)$ and for $r_a \neq 0$ we obtain that $a = \frac{1}{2}\left(\frac{T^2}{r_a} + r_a\right)$. Now, since $a \leq 0$ these conditions imply

$$r_a < 0 \ , \ T^2 > r_a^2 \ \text{ and } \ T^2 < l_a r_a \ ,$$

and we conclude that $-T \in (l_a, r_a)$. But according to Assumption (C) we have $-T \in S$, which contradicts $(l_a, r_a) \notin S$.

Similarly, one can show that for any value $a \in (l_a, \frac{l_a+r_a}{2})$ the case (6.50) also leads to a contradiction with Assumption (C) .

Now consider the remaining case (6.51) and $r_a = 0$. Then (6.49) implies that $T = 0$. However, due to Assumption (C) μ has a positive density in some neighborhood of 0, in contradiction with r_a lying on the boundary of the support of μ.

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

□

6.2. *Proof of Lemma 2.2.*

Proof. By Lemma 2.1 a_n converges to 0. Also, by Assumption (C) there exists $\epsilon > 0$ such that $\nu(\mu)$ has a density $\rho(\mu)$ on the interval $(-\epsilon, \epsilon)$. It is immediately clear that

$$\int\limits_{(\epsilon,\infty)} h_n^n(\mu)\ d\nu(\mu) \leq e^{-n\frac{\epsilon^2}{2\sigma^2}}\nu(\epsilon,\infty) \ . \tag{6.52}$$

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

$$\int\limits_{(-\infty,-\epsilon)} h_n^n(\mu)\ d\nu(\mu) \leq e^{-n\frac{\epsilon^2}{4\sigma^2}}\nu(-\infty,-\epsilon) \ . \tag{6.53}$$

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

$$\int\limits_{-\epsilon}^{\epsilon} h_n^n(\mu)\ \rho(\mu)d\mu = \rho_n \exp\left(\frac{na_n^2}{2\sigma^2}\right) \int\limits_{-\epsilon}^{\epsilon} \exp\left(-n\frac{(\mu-a_n)^2}{2\sigma^2}\right)\ d\mu$$

$$= \rho_n e^{\frac{na_n^2}{2\sigma^2}} \frac{\sqrt{2\pi}\sigma}{\sqrt{n}} \left[\Phi(\sqrt{n}(\epsilon - a_n)/\sigma) - \Phi(\sqrt{n}(-\epsilon - a_n)/\sigma)\right] \ , \tag{6.54}$$

where $\rho_n \in [\inf_{\mu\in(-\epsilon,\epsilon)} \rho(\mu), \sup_{\mu\in(-\epsilon,\epsilon)} \rho(\mu)]$, and $0 < \inf_{\mu\in(-\epsilon,\epsilon)} \rho(\mu) \leq \sup_{\mu\in(-\epsilon,\epsilon)} \rho(\mu) < \infty$ according to Assumption (C).

Note that $\Phi(\sqrt{n}(\epsilon - a_n)/\sigma) \to 1$ as well as $\Phi(\sqrt{n}(-\epsilon - a_n)/\sigma) \to 0$ (because $a_n \to 0$). Comparing (6.52), (6.53) and (6.54) we observe that the integral over $(-\epsilon, \epsilon)$ dominates the two remaining terms and from (2.4) it follows that

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

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

$$g_n \to 0 \quad \text{with} \quad \frac{\log n}{ng_n^2} \to 0, \quad \frac{\log(\delta f)}{ng_n^2} \to 0 \ . \tag{6.55}$$

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 (6.54) with the corresponding boundaries, and finally obtain

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

which yields (2.5). The proof for b_n is exactly the same. □

Remark 6.1. The proof of Lemma 2.2 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 [29]. 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.

6.3. *Proof of Theorem 2.1.*

Proof.

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 $a_n \to 0$, 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

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

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 $-\frac{\sqrt{n} b_n}{\sigma} \to -\infty$ we further obtain

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

where the last equality holds due to (2.7).

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

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

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

$$t_2 = \sigma \sqrt{\frac{\log v}{n}} \left(\rho(0^-) + \rho(0^+) \right) (1 + o_n(1)) \, . \tag{6.56}$$

In case of $0 < C < \infty$ we know from Lemma 2.1 that $a_n \to -T$ and $b_n \to T$, where $T = \sigma \sqrt{C} > 0$. For $\mu \in (-T, T)$, $\Psi_n(\mu) \to 1$, while for $\mu \in (-\infty, T) \cup (T, \infty)$, $\Psi_n(\mu) \to 0$. Then by the dominated convergence theorem,

$$t_2 = \int_{-\infty}^{\infty} \Psi_n(\mu) d\nu(\mu) = \nu(-T, T) \, (1 + o_n) \, , \tag{6.57}$$

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 (6.56) and (6.57) to complete the proof of Theorem 2.1 it is enough to show that

$$f\delta t_1/t_2 \to 0 \quad . \tag{6.58}$$

In case of $C = 0$, (2.7) and the normal tail approximation yield $t_1 \propto (v \log v)^{-1/2}$. Thus from (6.56) we easily obtain

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

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

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

Because $a_n \to -T$ similar considerations as in (6.52) show that the integral vanishes rapidly for $\mu \notin (-T-\epsilon, -T+\epsilon)$. Now observe that

$$\frac{1}{\sqrt{2\pi}} \int_{-T-\epsilon}^{-T+\epsilon} \exp\left(-\frac{n}{2\sigma^2}(a_n - \mu)^2\right) \rho(\mu) \, d\mu \le M_\rho \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \exp\left(-\frac{n}{2\sigma^2}(a_n - \mu)^2\right) \; d\mu \quad ,$$

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 $\delta f t_{1a} = O\left(\frac{1}{n}\right)$. Analogous considerations for t_{1b} finish the proof.

□

6.4. *Proof of Theorem 2.2.*

Proof. 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 (6.56), but using $\tilde{a}_n$, and $\tilde{b}_n$ instead of a_n and b_n. Taking into account (2.11) and (2.12) one thus obtains

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

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

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

Thus the type II error component of the risk $R_2 = mp\delta_A t_2$ satisfies $R_2 = R^B(1 + o_n)$.

Now, using (6.56) and the tail approximation for the type I error we obtain

$$R_1/R^B = \frac{m(1-p)\delta_0 t_1}{R^B} = C_{\sigma\rho}\frac{\exp(-z_a/2) + \exp(-z_b/2)}{\log v}(1 + o_n) \ , \tag{6.59}$$

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

In case of $C > 0$ due to (2.12) it holds that $\tilde{a}_n \to -T$ and $\tilde{b}_n \to T$, and hence thresholds specified by (2.11) also have type II error of the form (6.57). For sufficiency it remains to establish (6.58). 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

$$R_1/R^B = C_\nu\frac{\exp(-z_a/2) + \exp(-z_b/2)}{\log v}(1 + o_n) , \tag{6.60}$$

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

Concerning necessity, similar arguments as in the proof of Theorem 3.2 of [8] show that (2.12) is necessary for ABOS. In that case the computations leading to (6.59) and (6.60) are still valid and imply the necessity of (2.13).

□

6.5. *Lemma on the existence of the exact BFDR controlling rule.*

We first prove the following result

LEMMA 6.2. *For any fixed $s \neq 0$ the function*

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

satisfies

a) $f(0) = 1$,
b) $\lim_{c\to\infty} f(c) = \infty$,
c) $f(c)$ *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

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

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.

□

The following Lemma 6.7 easily follows from Lemma 6.2.

LEMMA 6.3. *Let* $\nu(\cdot)$ *be any probability measure such that* $\nu(0) < 1$. *Let us define*

$$H(c) := \frac{\int_R (2 - \Phi(c - \sqrt{n}\mu/\sigma) - \Phi(c + \sqrt{n}\mu/\sigma))d\nu(\mu)}{2(1 - \Phi(c))} \ . \tag{6.61}$$

Then it holds

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

LEMMA 6.4. *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 continously decreasing from $1-p$ for $c=0$ to 0 for $c \to \infty$.

Proof. Observe that

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

with $H(c)$ as in (6.61). Thus Lemma 6.4 is a direct consequence of Lemma 6.7. □

6.6. *Proof of Theorem 6.7.*

Proof.

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 $u_j^B \to \infty$. 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 $u_j^B \to \infty$ we can apply the tail approximation for the normal distribution and obtain from (2.18)

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

But on the other hand the second assumption of (2.19) yields $\left(\frac{\alpha_j}{f}\right)^{1/j} \to \exp(-C_0)$ which contradicts $u_j^B \to \infty$.

If $u_j := u_j^B \to u < \infty$ then the denominator of (2.18) converges to a constant $C_{\nu,u} = 1-\nu(-u,u)$. Under the first assumption of (2.19) equation (2.18) can only hold if $\sqrt{j}u_j \to \infty$. Thus we can apply again the tail approximation to obtain

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

Taking logarithms and some simple calculations yield

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

Now, the second condition in (2.19) implies that $u = \sigma\sqrt{2C_0}$, which completes the proof of (2.20).

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 [8]. Define $s_n := \frac{\log(f\delta\sqrt{n})}{\log(f/\alpha)} - 1$, then necessary and sufficient conditions for optimality are $s_n \to 0$ and $2s_n \log(f/n) - \log\log(f/\alpha) \to -\infty$ which immediately provides (2.21). From the first equation in (2.21) it follows that in case of ABOS $C_0 = C/2$, where C is the constant from Assumption (A). □

6.7. *Lemmas needed for Theorem 2.5.* 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 6.5. *Let c_{SD} be the random threshold SD threshold at the level α_n and let $\tilde{c}_1 = \tilde{c}_{1n}$ be the GW threshold (2.27) at the level $\alpha_{1n} = \alpha_n\xi_m$, where $\xi_m = (\log m)^{-s}$ with $s > 1$. Suppose that Assumptions (A) and (C), (2.28), (2.29) and (2.31) hold with $\alpha = \alpha_n$. Then $\tilde{c}_1$ is ABOS. Moreover, for every $\gamma_u > 0$ it holds for sufficiently large $m = m_n$ that*

$$P(c_{SD} \geq \tilde{c}_1) \leq m^{-\gamma_u} \quad . \tag{6.62}$$

PROOF. Based on the second condition in (2.31) it is easy to show that α_{1n} satisfies the asymptotic optimality assumptions provided in Theorem . Thus, Theorem 2.4 immediately yields that $\tilde{c}_1$ is ABOS.

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

$$\{c_{SD} \geq \tilde{c}_1\} = \left\{\tilde{H}(c_{SD}) \leq \alpha_{1n}\right\} \quad . \tag{6.63}$$

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

$$\frac{2(1-\Phi(c_{SD}))}{1-\hat{F}_m(c_{SD}) + 1/m} = \alpha_n$$

and thus

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

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

$$P(c_{SD} \geq \tilde{c}_1) \leq P\left(\inf_{c \geq \tilde{c}_1} \frac{1-\hat{F}_m(c) + 1/m}{1-F(c)} \leq \xi_m\right) \quad . \tag{6.64}$$

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 $U_1, \ldots, U_m$. 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 [37] can be applied. Define the event

$B_i = \{\hat{G}_m(u)$ intersects the line $y = (u-a)/(bm)$ at height i/m 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

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

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

Now, similar to Lemma 10.3.1 of [37] (page 414) we can apply Stirling's formula, which for $i < m\xi_m - 1$ yields

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

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 (6.62). □

The next lemma discusses the type II error component of the risk of SD.

LEMMA 6.6. *Under the assumptions of Theorem 2.5 the type II error component of the risk of SD satisfies*

$$R_A \leq R_B(1 + o_m) \ . \tag{6.65}$$

PROOF. For the extremely sparse case (2.14) we have seen already that the result follows by comparing with the Bonferroni rule which is ABOS according to Lemma 2.3. It remains to show the result for the denser case (2.30) and to note that both cases overlap.

Denote by L_A the number of false negatives under the SD rule and let $\tilde{c}_1$ be defined as in Lemma 6.5. Clearly

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

and furthermore

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

where L_1 is the number of false negatives produced by the rule based on the threshold $\tilde{c}_1$. Since by Lemma 6.5 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 $P(c_{SD} > \tilde{c}_1) \leq m^{-\gamma_u}$ 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 (2.30) and (2.31), 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. □

6.8. *Lemma needed for Theorem 3.1.*

LEMMA 6.7. *Assume that (3.33), (3.34), (3.38) as well as assumptions (B) and (C) hold. Then the following bounds are valid for the type I and type II error rates of mBIC:*

$$\frac{t_1}{p} = O\left((n \log n)^{-1/2}\right), \quad t_2 = O\left(\sqrt{\frac{\log n}{n}}\right) \ . \tag{6.66}$$

PROOF. Let $h_{n,m} := \log n + 2 \log m + d$. From the tail approximation of the standard normal distribution we immediately obtain

$$t_1 \sim \sqrt{\frac{2}{\pi h_{n,m}}} e^{-h_{n,m}/2} \leq \frac{c}{m} (n \log n)^{-1/2}$$

for some constant c. Using the fact that $mp \to s > 0$ from assumption (3.38) gives the first bound of (6.66).

To bound type II error we proceed similarly as in the proof of Theorem 2.1. We have $t_2 = \int \Psi_n(\mu)\, d\nu(\mu)$ with

$$\Psi_n(\mu) = \Phi\left(\sqrt{h_{n,m}} - \frac{\sqrt{n}\mu}{\sigma}\right) - \Phi\left(-\sqrt{h_{n,m}} - \frac{\sqrt{n}\mu}{\sigma}\right) .$$

The asymptotic behavior of this integral is obtained by similar analysis like that leading to (6.56), resulting in

$$t_2 = \sigma(\rho(0^-) + \rho(0^+))\frac{\sqrt{\log n + 2\log m}}{\sqrt{n}}(1 + o_{n,m}), \tag{6.67}$$

which completes the proof of Lemma 6.7 (since $m \leq n$).

□

6.9. *Proof of Theorem 3.2.*

Proof. In [1] it was shown that the step-up procedure BH corresponds to the smallest local minimum of the selection criterion (3.40), whereas SD corresponds to the largest local minimum of (3.40). Now mBIC1 is searching for the global minimum of (3.41), but we can again consider the smallest local minimum as well as the largest local minimum of (3.41). These will correspond to step-up and step-down procedures based on the comparison

$$\frac{n\hat{\beta}^2_{[k]}}{\sigma^2} \geq \log nm^2 + d - 2\log(k) - \log\log(nm^2/k^2) .$$

Translating this comparison to the level of p-values when applying the usual tail approximation for the standard normal distribution yields

$$p_{[k]} \leq \frac{Ak}{m} \quad \text{with} \quad A^2 = \frac{2e^{-d}}{\pi n z(k,m,n)}, \tag{6.68}$$

where $z(k,m,n) = 1 + \frac{d - \log\log(nm^2/k^2)}{\log(nm^2/k^2)}$. Since for sufficiently large n

$$1 - \frac{2\log\log n}{\log n} = z_1(n) < z(k,m,n) < z_2(n) = 1 - \frac{\log\log n}{6\log n} \quad ,$$

it holds that mBIC1 can be sandwiched between the step-up and step-down BH procedures with the FDR levels $\alpha_i = \sqrt{\frac{2e^{-d}}{\pi n z_i(n)}},\ i = 1, 2$, correspondingly. Since both α_i satisfy (2.21) the conditions of Theorem 2.5 are fulfilled and mBIC1 is itself ABOS.

Similar considerations give the result for mBIC2, for which we obtain $z(k,m,n) = \log(nm^2/k^2 + d)$ in (6.68). Using the inequalities

$$\log n < z(k,m,n) < 3\log n$$

for n large enough to sandwich mBIC2 between step-up and step-down procedures, ABOS of mBIC2 follows immediately from the fact that $\alpha \propto \frac{1}{\sqrt{n \log n}}$ fulfills (2.21).

Finally for mBIC3 we get

$$z(k,m,n) = e^2(1-1/k)^{2(k-1)} \left(\log(nm^2/k^2) + d + 2 + 2(k-1)\log(1-1/k)\right) ,$$

and we find again $\log n < z(k,m,n) < 3e^2 \log n$ which yields ABOS as above.

The consistency result is obtained the following way. From ABOS and the Markov inequality in (3.39) one easily concludes that all three criteria are consistent exactly when the Bayes oracle is consistent. Consider the asymptotic formulas concerning type II error (6.56) and type I error (6.58) for the special case $\delta = 1$. Then it immediately follows that the Bayes oracle is consistent under assumption (3.38).

□

6.10. *Proof of Theorem 3.3.*

Recall that in our setting (two groups model, orthogonality) it is reasonable to think in terms of type I error (misclassification of a regressor under H_0) and type II error (misclassification of a true signal) for model selection procedures. To prove Theorem 3.3 we first bound the type I and the type II errors in Lemma 6.8 and Lemma 6.9 respectively. Both these results will be proved assuming minimal conditions under which the individual bounds on type I and type II errors hold. The conditions in Theorem 3.3 ensure that both lemmas hold and additionally that the overall upper bound on the total risk of mBIC is asymptotically equivalent to that of the Bayes Oracle.

To bound the type I error we will make use of the following corollary given after Theorem 2 of Section 16.7, Vol.2 of [21].

COROLLARY 6.1. *Let F be the common distribution function of i.i.d. random variables $X_1, \ldots, X_n$ with $E(X_i) = 0$, $Var\,(X_i) = \sigma^2$ and let F_n be the distribution function of the normalized sum $(X_1 + \cdots + X_n)/(\sqrt{n}\sigma)$. If $1 < x = o(\sqrt{n})$, then for any $\epsilon > 0$, for all sufficiently large n,*

$$\frac{\exp(-(1+\epsilon)x^2/2)}{\sqrt{x}} < 1 - F_n(x) < \frac{\exp(-(1-\epsilon)x^2/2)}{\sqrt{x}} \tag{6.69}$$

LEMMA 6.8. *Assume $n \to \infty$, $m = m(n) \to \infty$ and that (3.33), (3.34) and (3.45) hold. Then the type I error probability of the decision rule based on mBIC criterion (3.44) is bounded by*

$$t_1 \leq \frac{C_1}{\sqrt{n}m}(1 + o_{n,m}) \ , \tag{6.70}$$

with $C_1 = \frac{\sqrt{2}}{\sqrt{\pi}} \exp(-d/2)$.

Proof. Let $1 \leq i \leq m$. Then the probability of type I error correspoding to β_i is given by

$$t_{1i} = P(A_i|B_i),$$

where B_i denotes the event that $\beta_i = 0$ and A_i denotes the event that the corresponding regressor is included in the model chosen by mBIC. Through exchangeability, it follows that $t_{1i} = t_1$, for each $1 \leq i \leq m$. Let us compare two models M and $M \bigcup X_i$ where $X_i \notin M$. mBIC considers supplementing model M with the variable X_i only in the case

$$\log \frac{RSS_M}{RSS_M - n\hat{\beta}_i^2} \geq \frac{\log n + 2\log m + d}{n} \tag{6.71}$$

where $RSS_M = Y'Y - \sum_{j \in M} n\hat{\beta}_j^2$ denotes the residual sum of squares of the model M (we have used the orthogonality assumption (3.33) here). Henceforth we will use the abbreviations $Z_{i,M} := \log \frac{RSS_M}{RSS_M - n\hat{\beta}_i^2}$ and $u_{n,m} := \frac{\log n + 2\log m + d}{n}$. Note that if $M_1 \subset M_2$ then $RSS_{M_1} \geq RSS_{M_2}$. Therefore $RSS_{M_1}/(RSS_{M_1} - n\hat{\beta}_i^2) \leq RSS_{M_2}/(RSS_{M_2} - n\hat{\beta}_i^2)$ and the event that a given false positive is added to the model M_1 is contained in the event that it is added to the model M_2. According to these considerations we obtain an upper bound for the type I error

$$t_1 \leq P(\bigcup_{M \in \Omega_L} \{Z_{i,M} \geq u_{n,m}\}|B_i),$$

where Ω_L is the set of all models with $L-1$ regressors in addition to the the common intercept term, such that $X_i \notin \Omega_L$. We bound the probability of the right hand side above in three intermediate steps.

Step 1: Let $T_i = \frac{\hat{\beta}_i^2}{\sigma^2}$ and $\epsilon_{n,m} = \frac{\log(\log n + 2\log m)}{n}$. Then

$$\{Z_{i,M} \geq u_{n,m}\} \subset \{Z_{i,M} - T_i \geq \epsilon_{n,m}\} \cup \{T_i \geq u_{n,m} - \epsilon_{n,m}\} \ ,$$

and therefore

$$\begin{aligned} &P(\bigcup_{M \in \Omega_L} \{Z_{i,M} \geq u_{n,m}\}|B_i) \leq \\ &P(\bigcup_{M \in \Omega_L} \{Z_{i,M} - T_i \geq \epsilon_{n,m}\}|B_i) + P(T_i \geq u_{n,m} - \epsilon_{n,m}|B_i) \ . \end{aligned}$$

The second term on the right hand side of the above inequality can be expressed as

$$P(T_i \geq u_{n,m} - \epsilon_{n,m}|B_i) = P\left(\frac{n\hat{\beta}_i^2}{\sigma^2} \geq \log n + 2\log m + d - \log(\log n + 2\log m)|B_i\right) \ ,$$

and from the normal tail approximation we obtain

$$P(T_i \geq u_{n,m} - \epsilon_{n,m}|B_i) = C_1 \frac{1}{\sqrt{n}m}(1 + o_{n,m}) \ ,$$

where $C_1 = \sqrt{\frac{2}{\pi}}\exp(-d/2)$. Now to establish inequality (6.70) it remains to be shown that $P(\bigcup_{M\in\Omega_L}\{Z_{i,M} - T_i \geq \epsilon_{n,m}\}|B_i)$ is of lower order.

Step 2: Let $\delta_n = \frac{1}{n}$. Similar arguments as in Step 1 yield

$$\{Z_{i,M}-T_i \geq \epsilon_{n,m}\} \subset \{Z_{i,M} > -\log(1-(T_i+\delta_n))\}\cup\{-T_i-\log(1-(T_i+\delta_n)) \geq \epsilon_{n,m}\} \ .$$

The first set on the right hand side can be rewritten as $\{\frac{n\hat{\beta}_i^2}{RSS_M} > T_i + \delta_n\}$ and therefore

$$\begin{aligned} &P(\bigcup_{M\in\Omega_L}\{Z_{i,M} - T_i \geq \epsilon_{n,m}\}|B_i) \\ &\leq P(\bigcup_{M\in\Omega_L}\{\frac{n\hat{\beta}_i^2}{RSS_M} > T_i + \delta_n\}|B_i) + P(-T_i - \log(1-(T_i+\delta_n)) \geq \epsilon_{n,m}|B_i) \ . \end{aligned}$$

To bound the second term note that $-\log(1-x) \leq x + 2x^2$ for $0 \leq x \leq 1/2$. Hence

$$P(-T_i-\log(1-(T_i+\delta_n)|) \geq \epsilon_{n,m}|B_i) \leq P\left((T_i+\delta_n)^2 > \frac{\epsilon_{n,m}-\delta_n}{2}|B_i\right)+P(T_i+\delta_n \geq 1/2|B_i) \ .$$

First note that

$$P\left((T_i+\delta_n)^2 > \frac{\epsilon_{n,m}-\delta_n}{2}|B_i\right) = P\left(nT_i > \frac{\sqrt{n\log(\log n + 2\log m)}}{2}(1+o_{n,m})|B_i\right)$$

and for sufficiently large n the normal tail approximation yields

$$P\left((T_i+\delta_n)^2 > \frac{\epsilon_{n,m}-\delta_n}{4}|B_i\right) < \exp\left(-\frac{\sqrt{n}}{2}\right) \ .$$

Second we have

$$P(T_i+\delta_n \geq 1/2|B_i) = P(nT_i \geq n/2-1|B_i) \leq \sqrt{\frac{2}{\pi}}\exp(-(n/4-1/2)).$$

Combining the two bounds obtained above, it follows that

$$P(-T_i - \log(1-(T_i+\delta_n)) \geq \epsilon_{n,m}|B_i) = o\left(\frac{1}{\sqrt{n}m}\right),$$

since $m < n$.

Step 3: We will now bound the remaining term

$$P\left(\bigcup_{M\in\Omega_L}\left\{\frac{n\hat{\beta}_i^2}{RSS_M} > T_i + \frac{1}{n}\right\}|B_i\right) \ .$$

Observing that

$$\left\{\frac{n\hat{\beta}_i^2}{\sigma^2}\left(\frac{n\sigma^2}{RSS_M}-1\right) > 1\right\} \subset \left\{\frac{n\hat{\beta}_i^2}{\sigma^2} > c_{n,m}\right\} \cup \left\{\frac{n\sigma^2}{RSS_M} - 1 > \frac{1}{c_{n,m}}\right\} \ ,$$

where we choose $c_{n,m} = \log n + 2\log m$, we conclude that

$$P\left(\bigcup_{M\in\Omega_L}\left\{\frac{n\hat{\beta}_i^2}{RSS_M} > T_i + \frac{1}{n}\right\}|B_i\right) \leq P\left(\frac{n\hat{\beta}_i^2}{\sigma^2} \geq c_{n,m}|B_i\right) + \sum_{M\in\Omega_L} P\left(\frac{n\sigma^2}{RSS_M} - 1 \geq \frac{1}{c_{n,m}}|B_i\right) \ .$$

By the tail approximation of the standard normal distribution we obtain that for sufficiently large n

$$P\left(\frac{n\hat{\beta}_i^2}{\sigma^2} \geq c_{n,m}|B_i\right) = \frac{1}{\sqrt{n}m}o_{n,m} \ .$$

We now have to bound the remaining series. Fix any model $M \in \Omega_L$ and assume that k true signals are not included in M. Under assumptions (3.33) and (3.34) it immediately follows that $RSS_M = W_k + Z_k$, where $Z_k = n\sum_{r=1}^{k}\hat{\beta}_{j_r}^2$ refers to the k true signals which were not detected, and $W_k \sim \sigma^2\chi^2_{(n-L-k)}$ is the remainder term. Z_k and W_k are independent and Z_k is stochastically larger than a $\sigma^2\chi_k^2$ distributed random variable. Therefore RSS_M is stochastically larger than $\sigma^2\chi^2_{(n-L)}$. But this argument holds for any k, and we conclude that

$$P\left(\frac{n\sigma^2}{RSS_M} - 1 \geq \frac{1}{c_{n,m}}|B_i\right) \leq P\left(\frac{n}{\chi^2_{n-L}} - 1 \geq \frac{1}{c_{n,m}}\right) \ .$$

Now

$$P\left(\frac{n}{\chi^2_{n-L}} - 1 \geq \frac{1}{c_{n,m}}\right) = P\left(\frac{\chi^2_{n-L} - (n-L)}{\sqrt{2(n-L)}} \leq \frac{L - \frac{n}{1+c_{n,m}}}{\sqrt{2(n-L)}}\right)$$

will be bounded using a normal tail approximation argument. ¿From assumption (3.45) it follows that $L = o(n/c_{n,m})$ and therefore

$$\frac{L - \frac{n}{1+c_{n,m}}}{\sqrt{2(n-L)}} = -\frac{\sqrt{n}}{\sqrt{2}(\log n + 2\log m)}(1 + o_{n,m}) \ .$$

Applying Corollary 6.1 with $x = \frac{\sqrt{n}}{\sqrt{2}(\log n + 2\log m)} = o(\sqrt{n-L})$ yields that for every $\epsilon > 0$ and n large enough (dependent on ϵ) it holds

$$P\left(\frac{\chi^2_{n-L} - (n-L)}{\sqrt{2(n-L)}} \leq \frac{L - \frac{n}{1+c_{n,m}}}{\sqrt{2(n-L)}}\right) \leq \exp\left(-\frac{(1-\epsilon)n^2}{2(\log n + 2\log m)^2}\right) \ .$$

The number of models with $L-1$ regressors is $\binom{m}{L-1} < m^L$. Thus, for sufficiently large n

$$\sum_{M\in\Omega_L} P\left(\frac{n\sigma^2}{RSS_M} - 1 \geq \frac{1}{c_{n,m}}\right) \leq m^L \exp\left(-\frac{(1-\epsilon)n}{4(\log n + 2\log m)^2}\right) = o\left(\frac{1}{\sqrt{n}m}\right) \ ,$$

which finishes the proof. □

Next we compute a bound for the type II error:

LEMMA 6.9. *Assume $n \to \infty$, $m = m(n) \to \infty$ and that (3.33), (3.34), (3.45), (3.46), (3.38) and Assumption (C) hold. Then the type II error of the decision rule based on mBIC criterion (3.44) is bounded by*

$$t_2 \leq \sigma(\rho(0^-) + \rho(0^+)) \frac{\sqrt{\log n + 2 \log m}}{\sqrt{n}} (1 + o_{n,m}) \tag{6.72}$$

Proof. Let $1 \leq i \leq m$ and suppose $\tilde{B}_i$ is the event that $\beta_i \neq 0$ and let $\tilde{A}_i$ be the event that the corresponding regressor is not detected. Then we have type II error $t_{2i} = P(\tilde{A}_i | \tilde{B}_i)$, and by exchangeability it follows that $t_{2i} = t_2$ is independent of i, for each $1 \leq i \leq m$.

Let us introduce the symbol D to denote the event that none of the X_j's corresponding to the null hypothesis are included in the model chosen by mBIC. Similar to the proof of Lemma 6.8 one can show that for every $i \neq j$, $P(A_j | B_j, \tilde{B}_i) = O\left(\frac{1}{\sqrt{n}m}\right)$. Then $P(D^c | \tilde{B}_i) = O(\frac{1}{\sqrt{n}})$ and thus

$$t_2 = P(\tilde{A}_i \cap D | \tilde{B}_i) + O(n^{-1/2}) \, .$$

To shorten the notation we now define $\tilde{A}_i^D = \tilde{A}_i \cap D$.

Note that

$$t_2 = \sum_{k=1}^{m} P(\tilde{A}_i^D | K = k, \tilde{B}_i) P(K = k | \tilde{B}_i) \ \ ,$$

where K is the number of nonzero β's among $\beta_1, \ldots \beta_m$. Under assumption (3.34), given $\tilde{B}_i$, $K - 1$ has a binomial distribution $B(m - 1, p_m)$. Define $L' = \lfloor m p_m (\log n)^{1+\eta} \rfloor$, where $\lfloor z \rfloor$ denotes the largest integer less than or equal to z. Using the assumption (3.38) and Bennett's inequality, it is easy to show that

$$P(K > L' | \tilde{B}_i) = o(n^{-1/2}). \tag{6.73}$$

Thus

$$t_2 \leq \sum_{k=1}^{L'} P(\tilde{A}_i^D | K = k, \tilde{B}_i) P(K = k | \tilde{B}_i) + O(n^{-1/2}) \ \ . \tag{6.74}$$

Note that here we made use of assumption (3.46).

Given $K = k$, let the ordered values of the squares of the estimates of the regression coefficients corresponding to the true regressors among $X_1, \ldots, X_m$ be denoted by $\hat{\beta}^2_{(1)} \leq \hat{\beta}^2_{(2)} \leq \ldots \leq \hat{\beta}^2_{(k)}$. Clearly

$$P(\tilde{A}_i^D | K = k, \tilde{B}_i) = \sum_{r=1}^{k} P(\tilde{A}_i^D | \hat{\beta}_i = \hat{\beta}_{(r)}, K = k, \tilde{B}_i) P(\hat{\beta}_i = \hat{\beta}_{(r)} | K = k, \tilde{B}_i) \ \ ,$$

and using the fact that $\hat{\beta}_i$'s corresponding to true signals are i.i.d. continuous random variables, the above equation can be rewritten as

$$P(\tilde{A}_i^D | K = k, \tilde{B}_i) = \frac{1}{k} \sum_{r=1}^{k} P(\tilde{A}_{(r)}^D | K = k) \ \ ,$$

where $\tilde{A}^D_{(r)}$ is the generic event that neither the regressor corresponding to $\hat{\beta}_{(r)}$ nor any false positives are detected by mBIC. Note that the events $\tilde{A}^D_{(r)}$'s are nested, i.e, $\tilde{A}^D_{(r+1)} \subset \tilde{A}^D_{(r)}$, and thus

$$\frac{1}{k}\sum_{r=1}^{k} P(\tilde{A}^D_{(r)}|K=k) = \sum_{r=1}^{k} \frac{r}{k} P(\tilde{A}^D_{(r)} \cap (\tilde{A}^D_{(r+1)})^c|K=k) \ , \tag{6.75}$$

where we define $\{\tilde{A}^D_{(k+1)}|K=k\} = \emptyset$. Thus we can write

$$t_2 \leq \sum_{k=1}^{L'}\sum_{r=1}^{k} \frac{r}{k} P(\tilde{A}^D_{(r)} \cap (\tilde{A}^D_{(r+1)})^c|K=k) P(K=k|\tilde{B}_i) + O(n^{-1/2}).$$

The event $\tilde{A}^D_{(r)} \cap \tilde{(}A^D_{(r+1)})^c$ implies that the model chosen by mBIC includes the $(k-r)$ true regressors having the largest absolute values of estimated regression coefficients, denoted by $X_{(r+1)}, \ldots, X_{(k)}$, while $X_{(j)}$ for $1 \leq j \leq r$ are not included. This event also corresponds to the situation when no false positives are included. Hence the model includes $k - r < L' \leq L$ regressors, and in any case we have not yet exhausted our maximum model size. So, since $X_{(r)}$ is not included in the model we can infer that mBIC criterion is getting larger by adding $X_{(r)}$. Denoting by RSS_{k-r} the residual sum of squares of the model including $X_{(r+1)}, \ldots, X_{(k)}$ (or only the intercept in case $r = k$), we have

$$P(\tilde{A}^D_{(r)} \cap (\tilde{A}^D_{(r+1)})^c|K=k) \leq P\left(\log\left(\frac{RSS_{k-r}}{RSS_{k-r} - n\hat{\beta}^2_{(r)}}\right) \leq u_{n,m}|K=k\right) \ .$$

Since for $x \in (0,1)$, $\log(1/(1-x)) \geq x$,

$$P(\tilde{A}^D_{(r)} \cap (\tilde{A}^D_{(r+1)})^c|K=k) \leq P\left(\frac{n\hat{\beta}^2_{(r)}}{RSS_{k-r}} \leq u_{n,m}|K=k\right) \ . \tag{6.76}$$

Under $K = k$, RSS_{k-r} is the sum of two independent random variables, the first being a $\sigma^2\chi^2_{n-k-1}$ random variable (χ^2_{n-k-1} being a central chi-square with $(n-k-1)$ degrees of freedom), while the second is $\sum_{j=1}^{r} n\hat{\beta}^2_{(j)}$. Therefore

$$P\left(\frac{n\hat{\beta}^2_{(r)}}{RSS_{k-r}} \leq u_{n,m}|K=k\right) = P\left(\frac{n\hat{\beta}^2_{(r)}}{\sigma^2\chi^2_{(n-k-1)} + \sum_{j=1}^{r} n\hat{\beta}^2_{(j)}} \leq u_{n,m}\right) \ ,$$

and because of $\hat{\beta}^2_{(r)} > \hat{\beta}^2_{(j)}$, for $r > j$ one obtains

$$P\left(\frac{n\hat{\beta}^2_{(r)}}{RSS_{k-r}} \leq u_{n,m}|K=k\right) \leq P(n\hat{\beta}^2_{(r)} \leq \sigma^2\chi^2_{n-k-1}u_{n,m}(1+o_{n,m})) \ ,$$

where $(1+o_{n,m}) = \frac{1}{1-ru_{n,m}}$. (Note that $r \leq k \leq L' = mp_m(\log n)^{1+\eta}$, and under assumption 3.45 $ru_{n,m} \to 0$ as $n \to \infty$).

Define $b_n = \frac{\log(\log n)}{4 \log n}$. Then

$$\left(n-k-1+\sqrt{2}(n-k-1)^{1-b_n}\right)/n = 1+o_{n,m},$$

and therefore

$$\begin{aligned} P(n\hat{\beta}^2_{(r)} \leq \sigma^2 \chi^2_{n-k-1} u_{n,m}(1+o_{n,m})) \;\leq\; & P(\hat{\beta}^2_{(r)} \leq \sigma^2 u_{n,m}(1+o_{n,m})) \\ & + P\left(\chi^2_{n-k-1} > n-k-1+\sqrt{2}(n-k-1)^{1-b_n}\right) \quad . \end{aligned}$$

A simple application of Chebyshev's inequality yields

$$P\left(\chi^2_{n-k-1} > n-k-1+\sqrt{2}(n-k-1)^{1-b_n}\right) \leq (n-k-1)^{-1+2b_n} \; .$$

Now, observe that

$$\begin{aligned} & \sum_{k=1}^{L'} \sum_{r=1}^{k} \tfrac{r}{k} \; P\left(\chi^2_{n-k-1} > n-k-1+\sqrt{2}(n-k-1)^{1-b_n}\right) \; P(K=k|B_i) \\ & \leq (n-k-1)^{-1+2b_n} E((K+1)/2) \leq mp \; (n-k-1)^{-1+2b_n} = o(n^{-1/2}) \quad , \end{aligned}$$

where the last equality follows after some calculations from (3.38). Therefore

$$t_2 \leq \sum_{k=1}^{L'} \sum_{r=1}^{k} \frac{r}{k} \; P(\hat{\beta}^2_{(r)} \leq \sigma^2 u_{n,m}(1+o_{n,m})) \; P(K=k|\tilde{B}_i) + O(n^{-1/2}) \quad . \tag{6.77}$$

Let us define

$$q = q_{n,m} := P\left(\hat{\beta}^2_i \leq \sigma^2 u_{n,m}(1+o_{n,m})\right) \quad . \tag{6.78}$$

Given that $\hat{\beta}_i \sim \nu * \mathcal{N}(0, \sigma^2/n)$ straight forward computations yield

$$q = \int\limits_{\mu \in I\!R} \left[\Phi\left(\sqrt{nu_{n,m}}(1+o_{n,m}) - \frac{\sqrt{n}\mu}{\sigma}\right) - \Phi\left(-\sqrt{nu_{n,m}}(1+o_{n,m}) - \frac{\sqrt{n}\mu}{\sigma}\right)\right] d\nu(\mu).$$

The asymptotic behavior of this integral is obtained by similar analysis like that leading to (6.56), resulting in

$$q = \sigma(\rho(0^-) + \rho(0^+)) \frac{\sqrt{\log n + 2\log m}}{\sqrt{n}} (1+o_{n,m}).$$

Define the contribution for fixed r in the sum on the right hand side of (6.77) as

$$\Psi_r := \sum_{k=r}^{L'} \frac{r}{k} P(\hat{\beta}^2_{(r)} \leq \sigma^2 u_{n,m}(1+o_{n,m})) \; P(K=k|\tilde{B}_i).$$

For $r = 1$ we observe that

$$P(\hat{\beta}^2_{(1)} \leq \sigma^2 u_{n,m}(1 + o_{n,m})) = 1 - (1-q)^k = kq - \sum_{j=2}^{k} \binom{k}{j} (-q)^j \ ,$$

with the convention that $\binom{1}{2}$=0.

Thus

$$\begin{aligned} \Psi_1 &\leq q + \sum_{k=2}^{L'} \frac{1}{k} \sum_{j=2}^{k} \binom{k}{j} q^j P(K = k|\tilde{B}_i) \\ &= q + \sum_{j=2}^{L'} \frac{q^j}{j} \sum_{k=j}^{L'} \binom{k-1}{j-1} P(K = k|\tilde{B}_i) \\ &\leq q + \sum_{j=2}^{L'} \frac{q^j}{j\ (j-1)!} (mp)^{j-1} \\ &\leq q + q^2 mp e^{mpq} = q(1 + o(q)) \ . \end{aligned}$$

as long as $mpq \to 0$ which is guaranteed by (3.38). The first inequality follows from the fact that under $\tilde{B}_i$, $K \sim \text{Bin}(m-1, p)$ and thus

(6.79)

$$E((K-1)(K-2)\cdots(K-j+1)) = (m-1)(m-2)\ldots(m-j+1)p^{j-1} \leq (mp)^{j-1}.$$

Finally we have to bound the contribution Ψ_r in the sum on the the right hand side of (6.77) stemming from $r > 1$. Note that

$$P(\hat{\beta}^2_{(r)} \leq \sigma^2 u_{n,m}(1 + o_{n,m})) = \sum_{j=r}^{k} \binom{k}{j} q^j (1-q)^{k-j} \ .$$

Similar computations as above using (6.79) yield

$$\Psi_r = \sum_{k=r}^{L'} \frac{r}{k} \sum_{j=r}^{k} \binom{k}{j} q^j (1-q)^{k-j} P(K = k|\tilde{B}_i) \leq \sum_{j=r}^{L'} \frac{q^j}{(j-1)!} s^{j-1} \leq q^r s^{r-1} e^{qs}.$$

Summing over all possible values of $r > 1$ finally gives

$$\sum_{r=2}^{L'} \Psi_r \leq \frac{q^2 s}{1 - qs} e^{qs} = o(q).$$

Thus we have shown that

$$t_2 \leq \sum_{r=1}^{L'} \Psi_r + O(n^{-1/2}) \leq q(1 + o_{n,m}),$$

since $O(n^{-1/2}) = o(q)$. This completes the proof of the lemma. □

Proof of Theorem 3.3

Proof. First note that the assumption on mp in (3.36) is stronger than that in assumption (3.38). Given (3.36) it is easy to see that the type II error estimate in Lemma 6.9 is asymptotically of the same form as the type II error of the Bayes oracle for $C = 0$ in (6.56). To show ABOS it is therefore sufficient that the risk component of the type I error is of smaller order than the Bayes risk. From Lemma 6.8 we conclude

$$\frac{R_1}{R_{BO}} = O\left(\frac{\delta}{mp\sqrt{\log v}}\right) .$$

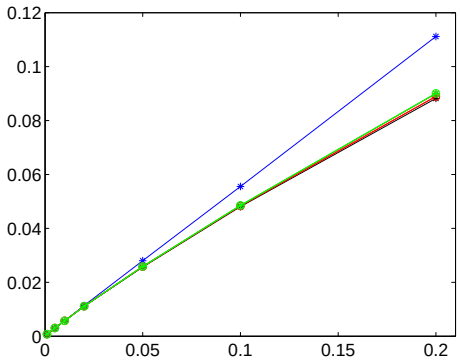
Under Assumption (B) δ is bounded from above and ABOS follows.

Consistency follows exactly the same way as in Theorem 3.1. □

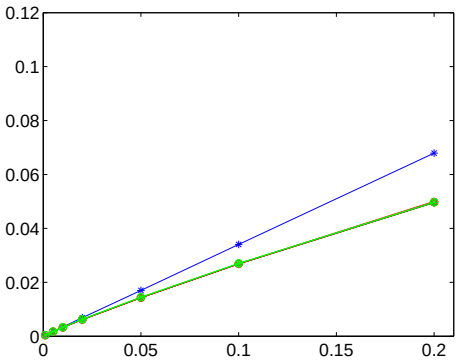
6.11. *Figures of the first part of the simulation study.*

FIG 3. *Simulation runs for known σ. Misclassification probability (MP), False Discovery Rate (FDR) and Power for different selection rules and sparsity parameter p at values of $p \in \{0.001, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2\}$.*

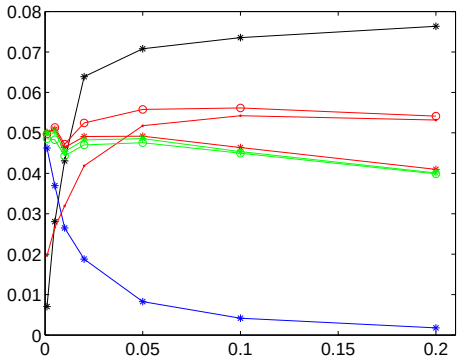
(a) MP , $n = m = 256$



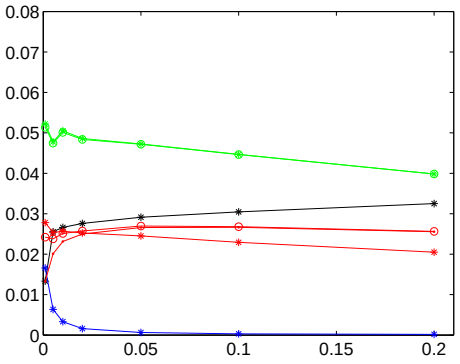
(b) MP , $n = m = 1024$



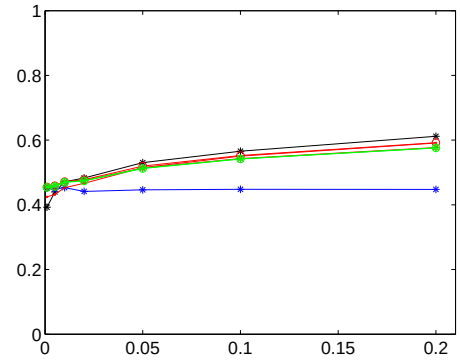
(c) FDR , $n = m = 256$



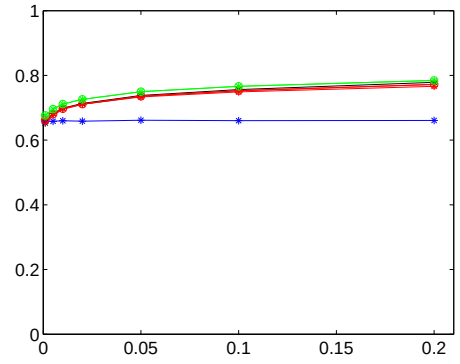
(d) FDR , $n = m = 1024$



(e) Power , $n = m = 256$



(f) Power , $n = m = 1024$



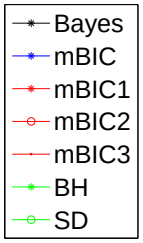
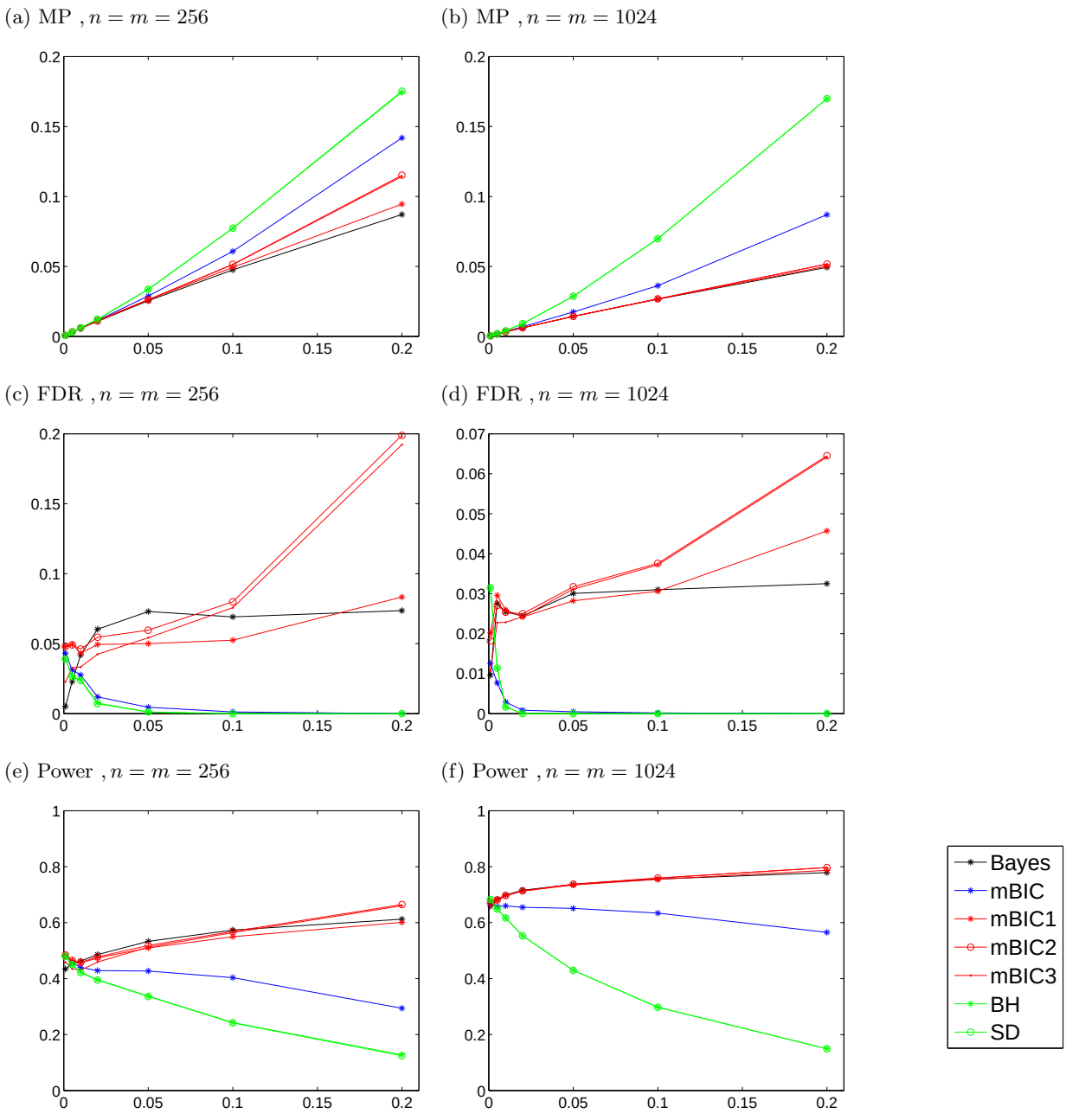


FIG 4. *Simulation runs for unknown σ. Misclassification probability (MP), False Discovery Rate (FDR) and Power for different selection rules and sparsity parameter p at values of $p \in \{0.001, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2\}$.*



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

[2] AKAIKE H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19** (6), 716–723.

[3] 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

[4] Bickel, P.J., Ritov, Y., and Tsybakov, A.B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* **37**, 1705–1732.

[5] Bogdan, M., Ghosh, J.K., and Doerge, R.W. (2004). Modifying the Schwarz Bayesian Information Criterion to locate multiple interacting quantitive trait loci. *Genetics* **167**, 989–999.

[6] 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.

[7] Bogdan, M., Chakrabarti, A., Frommlet, F. and Ghosh, J. K. (2010). The Bayes oracle and asymptotic optimality of multiple testing procedures under sparsity, *arXiv:1002.3501*.

[8] Bogdan, M., Chakrabati, A., Frommlet F. and Ghosh, J.K. (2011) Asymptotic Bayes Optimality under sparsity of some multiple testing procedures. *Ann. Statist.*, *To appear*.

[9] Bogdan, M., Ghosh, J.K. and Żak-Szatkowska, M. (2008) Selecting explanatory variables with the modified version of Bayesian Information Criterion, *Quality and Reliability Engineering International* **24**, 627-641.

[10] 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.

[11] Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35**, 2313–2351.

[12] Chakrabarti, A. and Ghosh, J.K. (2006). Some aspects of Bayesian model selection for prediction. *Bayesian Statistics* **8**, 51–90, Oxford University Press.

[13] Chen, J. and Chen, Z. (2008). Extended Bayesian Information criteria for model selection with large model spaces. *Biometrika* **95**(3), 759–771.

[14] Chen, Z. and Luo, S. (2010). Extended BIC for linear regression models with diverging number of parameters and high or ultra-high feature spaces. *Preprint*.

[15] Chi, Z. (2008). False discovery rate control with multivariate p-values. *Electronic Journal of Statistics* **2**, 368–411.

[16] Donoho, D.L. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32**, 962–994.

[17] Donoho, D.L. and Jin, J. (2006). Asymptotic minimaxity of false discovery rate thresholding for sparse exponential data. *Ann. Statist.* **34**, 2980–3018.

[18] 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.

[19] Efron, B. and Tibshirani, R. (2002). Empirical bayes methods and false discovery rates for microarrays. *Genetic Epidemiology*, **23**, 70–86.

[20] Efron, B. (2008). Microarrays, Empirical Bayes and the two-group model. *Stat. Sci.*, **23**(1), 1–22.

[21] Feller, W. (1966). An introduction to probability theory and its applications. Vol. 2: Wiley, New York.

[22] Finner H., Dickhaus, T. and Roters, M. (2009). On the false discovery rate and an asymptotically optimal rejection curve. *Ann. Statist.* **37**, 596–618.

[23] Foster, D.P., and George, E.I. (1994) The risk inflation criterion for multiple regression. *Ann. Statist.* **22**, 1947–1975.

[24] Frommlet, F., Ruhaltinger, F., Twarog, P. and Bogdan, M. (2011). Modified versions of Bayesian Information Criterion for genome-wide association studies. CSDA, doi:10.1016/j.csda.2011.05.005

[25] 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.

[26] George, E.I. and Foster, D.P. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87**, 731–747.

[27] 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.

[28] 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.

[29] Johnson, B.R. and Truax, D.R. (1973). Asymptotic behavior of Bayes tests and Bayes risk. *Ann. Statist.* **2**, 278–294.

[30] Lehmann, E. L. 1957. A theory of some multiple decision problems, I. *Ann. Math. Stat.* **28**, 1–25.

[31] 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.

[32] 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

[33] 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.

[34] Roquain, E., and van de Wiel, M. A. (2009). Optimal weighting for false discovery rate control. *Electronic Journal of Statistics* **3**, 678–711.

[35] Schwarz, G. (1978). Estimating the Dimension of a Model. *Ann. Statist.* **6**(2), 461-464.

[36] Scott, J.G. and Berger, J.O. (2006). An exploration of aspects of Bayesian multiple testing. *J. Statist. Plann. Inference* **136**(7), 2144–2162.

[37] Shorack, G. R. and Wellner, J. A. (1986). *Empirical processes with applications to Statistics*, Wiley Series in Probability and Mathematical Statistics.

[38] Storey, J.D. (2007). The optimal discovery procedure: a new approach to simultaneous significance testing. *J. R. Statist. Soc. B* **69**, 347-368.

[39] 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.
[40] Żak-Szatkowska, M. and Bogdan, M. (2011). Modified versions of Bayesian Information Criterion for sparse Generalized Linear Models, **CSDA**, in revision, available at *www.im.pwr.wroc.pl/∼mbogdan/Preprints*

Department of Medical Statistics
Medical University of Vienna
Spitalgasse 23
A-1090 Vienna, Austria
Florian.Frommlet@meduniwien.ac.at

203 B.T.Road
Bayesian and Interdisciplinary Research Unit
Indian Statistical Institute
Kolkata 700108, West Bengal,India
arc@isical.ac.in

Department of Biostatistics
Erasmus University Medical Center
PO Box 2040, 3000 CA Rotterdam,the Netherlands
m.murawska@erasmusmc.nl

Department of Mathematics and Computer Science
Wrocław University of Technology
ul. Janiszewskiego 14a
50-370 Wroclaw, Poland
E-mail: Malgorzata.Bogdan@pwr.wroc.pl