On false discovery rate thresholding for classification under sparsity

Pierre Neuvial, Etienne Roquain

Submitted to the Annals of Statistics
arXiv: arXiv:1106.6147

ON FALSE DISCOVERY RATE THRESHOLDING FOR CLASSIFICATION UNDER SPARSITY

By Pierre Neuvial and Etienne Roquain

We study the properties of false discovery rate (FDR) thresholding, viewed as a classification procedure. The "0"-class (null) is assumed to have a known density while the "1"-class (alternative) is obtained from the "0"-class either by translation or by scaling. Furthermore, the "1"-class is assumed to have a small number of elements w.r.t. the "0"-class (sparsity). We focus on densities of the Subbotin family, including Gaussian and Laplace models. Non-asymptotic oracle inequalities are derived for the excess risk of FDR thresholding. These inequalities lead to explicit rates of convergence of the excess risk to zero, as the number m of items to be classified tends to infinity and in a regime where the power of the Bayes rule is away from 0 and 1. Moreover, these theoretical investigations suggest an explicit choice for the target level α_m of FDR thresholding, as a function of m. Our oracle inequalities show theoretically that the resulting FDR thresholding adapts to the unknown sparsity regime contained in the data. This property is illustrated with numerical experiments.

1. Introduction.

1.1. *Background.* In many high-dimensional settings, such as microarray or neuro-imaging data analysis, we aim at detecting signal among several thousands of items (e.g., genes or voxels). For such problems, a standard error measure is the false discovery rate (FDR), which is defined as the expected proportion of errors among the items declared as significant.

Albeit motivated by pure testing considerations, the Benjamini Hochberg FDR controlling procedure proposed by Benjamini and Hochberg (1995) has recently been shown to enjoy remarkable properties as an estimation procedure Abramovich et al. (2006); Donoho and Jin (2006). More specifically, it turns out to be adaptive to the amount of signal contained in the data, which has been referred to as "adaptation to unknown sparsity".

In a classification framework, while Genovese and Wasserman (2002) contains what is to our knowledge the first analysis of FDR thresholding with respect to the mis-classification risk, an important theoretical breakthrough

Received June 2011.
AMS 2000 subject classifications: Primary 62H30; secondary 62H15.
Keywords and phrases: false discovery rate, sparsity, classification, multiple testing, Bayes' rule, adaptive procedure, oracle inequality.

has recently been made in Bogdan et al. (2011) (see also Bogdan et al. (2008)). The major contribution of Bogdan et al. (2011) is to create an asymptotic framework in which several multiple testing procedures can be compared in a sparse Gaussian scale mixture model. In particular, they proved that FDR thresholding is asymptotically optimal (as the number m of items goes to infinity) with respect to the mis-classification risk and thus adapts to unknown sparsity in that setting (for a suitable choice of the level parameter α_m). Also, they proposed an optimal choice for the rate of α_m as m grows to infinity.

The present paper can be seen as an extension of Bogdan et al. (2011). First, we prove that the property of adaptation to unknown sparsity also holds non-asymptotically, by using finite sample oracle inequalities. This leads to a more accurate asymptotic analysis, for which explicit convergence rates can be provided. Second, we show that these theoretical properties are not specific to the Gaussian scale model, but carry over to Subbotin location/scale models. They can also be extended to (fairly general) log-concave densities (as shown in the supplemental article), but we chose to focus on Subbotin densities in the main manuscript for simplicity. Finally, we additionally supply an explicit, finite sample, choice of the level α_m and provide an extensive numerical study that aims at illustrating graphically the property of adaptation to unknown sparsity.

1.2. *Initial setting.* Let us consider the following classification setting: let $(X_i, H_i) \in \mathbb{R} \times \{0,1\}$, $1 \le i \le m$, be m i.i.d. variables. Assume that the sample $X_1, \dots, X_m$ is observed without the labels $H_1, \dots, H_m$ and that the distribution of X_1 conditionally on $H_1 = 0$ is known a priori. We consider the following general classification problem: build a (measurable) classification rule $\hat{h}_m : \mathbb{R} \to \{0,1\}$, depending on $X_1, ..., X_m$, such that the (integrated) misclassification risk $R_m(\hat{h}_m)$ is as small as possible. We consider two possible choices for the risk $R_m(\cdot)$:

$$R_m^T(\hat{h}_m) = \mathbb{E}\left(m^{-1} \sum_{i=1}^{m} \mathbf{1}\{\hat{h}_m(X_i) \neq H_i\} \right); \tag{1}$$

$$R_m^I(\hat{h}_m) = \mathbb{P}(\hat{h}_m(X_{m+1}) \neq H_{m+1}), \tag{2}$$

where the expectation is taken with respect to $(X_i, H_i)_{1 \le i \le m}$ in (1) and to $(X_i, H_i)_{1 \le i \le m+1}$ in (2), for a new labeled data point $(X_{m+1}, H_{m+1}) \sim (X_1, H_1)$ independent of $(X_i, H_i)_{1 \le i \le m}$. The risks $R_m^T(\hat{h}_m)$ and $R_m^I(\hat{h}_m)$ are usually referred to as *transductive* and *inductive* risks, respectively (see Remark 1.1 for a short discussion on the choice of the risk). Note that these

two risks can be different in general because X_i appears "twice" in $\hat{h}_m(X_i)$. However, they coincide for procedures of the form $\hat{h}_m(\cdot) = h_m(\cdot)$, where $h_m : \mathbb{R} \to \{0,1\}$ is a deterministic function. The methodology investigated here can also be easily extended to a class of *weighted* mis-classification risks, as originally proposed by Bogdan et al. (2011) (in the case of the transductive risk) and further discussed in Section 6.2.

The distribution of (X_1, H_1) is assumed to belong to a specific parametric subset of distributions on $\mathbb{R} \times \{0,1\}$, which is defined as follows:

(i) the distribution of H_1 is such that the (unknown) mixture parameter $\tau_m = \pi_{0,m}/\pi_{1,m}$ satisfies $\tau_m > 1$, where $\pi_{0,m} = \mathbb{P}(H_1 = 0)$ and $\pi_{1,m} = \mathbb{P}(H_1 = 1) = 1 - \pi_{0,m}$.

(ii) the distribution of X_1 conditionally on $H_1 = 0$ has a density $d(\cdot)$ w.r.t. the Lebesgue measure on $\mathbb{R}$ that belongs to the family of so-called ζ-Subbotin densities, parametrized by $\zeta \geq 1$, and defined by

(3)
$$d(x) = (L_\zeta)^{-1} e^{-|x|^\zeta/\zeta}, \text{ with } L_\zeta = \int_{-\infty}^{+\infty} e^{-|x|^\zeta/\zeta} dx = 2\Gamma(1/\zeta)\zeta^{1/\zeta - 1}.$$

(iii) the distribution of X_1 conditionally on $H_1 = 1$ has a density $d_{1,m}(\cdot)$ w.r.t. the Lebesgue measure on $\mathbb{R}$ of either of the following two types:

- location: $d_{1,m}(x) = d(x - \mu_m)$, for an (unknown) location parameter $\mu_m > 0$;
- scale: $d_{1,m}(x) = d(x/\sigma_m)/\sigma_m$, for an (unknown) scale parameter $\sigma_m > 1$.

The density d is hence of the form $d(x) = e^{-\phi(|x|)}$ where $\phi(u) = u^\zeta/\zeta + \log(L_\zeta)$ is convex on $\mathbb{R}^+$ (log-concave density). This property is of primary interest when applying our methodology, see the supplemental article Neuvial and Roquain (2011). The particular values $\zeta = 1, 2$ give rise to the Laplace and Gaussian case, respectively. The classification problem under investigation is illustrated by Figure 1 (left panel), in the Gaussian location case. Moreover, let us note that we will exclude in our study the Laplace location model (that is, the location model using $\zeta = 1$). This particular model is not directly covered by our methodology and needs specific investigations, see Section S-3.3 in the supplemental article.

Our modeling is motivated by the following application: consider a microarray experiment for which measurements $Z_1, \ldots, Z_m$ for m genes are observed, each corresponding to a difference of expression levels between two experimental conditions (e.g., test versus reference sample). Let $H_1, \ldots, H_m$ be binary variables coded as 1 if the gene is differentially expressed and 0

if not. Assume that each Z_i is $\mathcal{N}(\delta_i, \sigma_\varepsilon^2)$ where δ_i is the (unknown) effect for gene i while σ_ε^2 quantifies the (known) measurement error. Next, assume the Bayesian paradigm that sets the following prior distribution for δ_i: the distribution of δ_i is $\mathcal{N}(0, \sigma_0^2)$ conditionally on $H_i = 0$ and $\mathcal{N}(\delta, \sigma_0^2 + \tau^2)$ conditionally on $H_i = 1$. Generally, σ_0^2 (≥ 0), the dispersion of the non-differentially expressed genes, is assumed to be known while δ (≥ 0) and τ^2 (≥ 0), the shift and additional dispersion of the non-differentially expressed genes, are unknown. Let $X_i = Z_i/\sigma$ for $\sigma^2 = \sigma_\varepsilon^2 + \sigma_0^2$ and consider the distribution unconditionally on the δ_i's. This corresponds to our model (in the Gaussian case) as follows:

- $\delta > 0$ and $\tau^2 = 0$: location model with $\mu_m = \delta/\sigma > 0$;
- $\delta = 0$ and $\tau^2 > 0$: scale model with $\sigma_m^2 = (\sigma^2 + \tau^2)/\sigma^2 > 1$.

The above convolution argument was originally proposed in Bogdan et al. (2011) for a Gaussian scale model: it explains how we can obtain test statistics that have the same distribution under the alternative even if the effects of the measurements are not equal.

Going back to our general setting, an important point is that the parameters — (τ_m, μ_m) in the location model, or (τ_m, σ_m) in the scale model — are assumed to *depend* on sample size m. The parameter τ_m, called the *sparsity* parameter, is assumed to tend to infinity as m tends to infinity, which means that the unlabeled sample only contains a small, vanishing proportion of label 1. This condition is denoted (Sp). As a counterpart, the other parameter — μ_m in the location model, or σ_m in the scale model — is assumed to tend to infinity fast enough to balance sparsity. This makes the problem "just solvable" under the sparsity constraint. More precisely, our setting corresponds to the case where the power of the Bayes procedure is bounded away from 0 and 1, and is denoted (BP). This is motivated by sparse high-dimensional problems for which the signal is strong but only carried by a small part of the data. For instance, in the above-mentioned application to microarray data, the two experimental conditions compared can be so close that only a very small proportion of genes are truly differentially expressed (e.g. two groups of patients having the same type of cancer but a different response to a cancer treatment Sawyers (2008)).

REMARK 1.1. *Our setting is close to the semi-supervised novelty detection (SSND) framework proposed in Blanchard et al. (2010), for which the knowledge of the distribution X_1 conditionally on $H_1 = 0$ is replaced by the observation of a finite i.i.d. sample with this distribution. In the latter work, the authors use the unlabeled data $X_1, \ldots, X_m$ to design a procedure $\widehat{h}_m$ that aims at classifying a new unlabeled data X_{m+1}. This approach is*

in accordance with the inductive risk defined by (2). *However, in other situations closer to standard multiple testing situations, one wants to classify* $X_1, \dots, X_m$ *meanwhile designing* $\widehat{h}_m$. *This gives rise to the transductive risk defined by* (1).

1.3. *Thresholding procedures.* Classically, the solution that minimizes the misclassification risks (1) and (2) is the so-called Bayes rule h_m^B that chooses the label 1 whenever $d_{1,m}(x)/d(x)$ is larger than a specific threshold. We easily check that the likelihood ratio $d_{1,m}(x)/d(x)$ is nondecreasing in x and $|x|$ for the location and the scale model, respectively. As a consequence, we can only focus on classification rules $\hat{h}_m(x)$ of the form $\mathbf{1}\{x \geq \hat{s}_m\}$, $\hat{s}_m \in \mathbb{R}$, for the location model, and $\mathbf{1}\{|x| \geq \hat{s}_m\}$, $\hat{s}_m \in \mathbb{R}^+$, for the scale model. Therefore, to minimize the mis-classification risks, thresholding procedures are classification rules of primary interest, and the main challenge consists in choosing the threshold $\hat{s}_m$ in function of $X_1, ..., X_m$.

The FDR controlling method proposed by Benjamini and Hochberg (1995) (also called "Benjamini-Hochberg" thresholding) provides such a thresholding $\hat{s}_m$ in a very simple way once we can compute the quantile function $\overline{D}^{-1}(\cdot)$, where $\overline{D}(u) = (L_\varsigma)^{-1} \int_u^{+\infty} e^{-|x|^\varsigma/\varsigma} dx$ is the (known) upper-tail cumulative distribution function of X_1 conditionally on $H_1 = 0$. We recall below the algorithm for computing the FDR threshold in the location model (using test statistics rather than p-values).

ALGORITHM 1.2. *1. choose a nominal level* $\alpha_m \in (0, 1)$*;*
2. consider the order statistics of the X_k*'s:* $X_{(1)} \geq X_{(2)} \geq ... \geq X_{(m)}$*;*
3. take the integer $\hat{k} = \max\{1 \leq k \leq m : X_{(k)} \geq \overline{D}^{-1}(\alpha_m k/m)\}$ *when this set is non-empty and* $\hat{k} = 1$ *otherwise;*
4. use $\hat{h}_m^{FDR}(x) = \mathbf{1}\{x \geq \hat{s}_m^{FDR}\}$ *for* $\hat{s}_m^{FDR} = \overline{D}^{-1}(\alpha_m \hat{k}/m)$.

For the scale model, FDR thresholding has a similar form: $\hat{h}_m^{FDR}(x) = \mathbf{1}\{|x| \geq \hat{s}_m^{FDR}\}$ for $\hat{s}_m^{FDR} = \overline{D}^{-1}(\alpha_m \hat{k}/(2m))$, where $\hat{k} = \max\{1 \leq k \leq m : |X|_{(k)} \geq \overline{D}^{-1}(\alpha_m k/(2m))\}$ ($\hat{k} = 1$ if the set is empty) and $|X|_{(1)} \geq |X|_{(2)} \geq ... \geq |X|_{(m)}$. Algorithm 1.2 is illustrated in Figure 1 (right panel), in a Gaussian location setting. Since $\hat{s}_m^{FDR} = \overline{D}^{-1}(\alpha_m \hat{k}/m)$ takes its values in the range $[\overline{D}^{-1}(\alpha_m), \overline{D}^{-1}(\alpha_m/m)]$, it can be seen as a intermediate thresholding rule between the Bonferroni thresholding ($\overline{D}^{-1}(\alpha_m/m)$) and the uncorrected thresholding ($\overline{D}^{-1}(\alpha_m)$). Finally, an important feature of the FDR procedure is that it depends on a pre-specified level $\alpha_m \in (0, 1)$. In this work, the level α_m is simply used as a tuning parameter, chosen to make

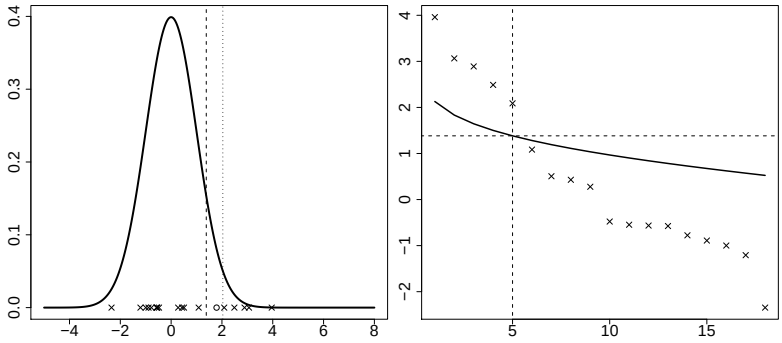

FIG 1. *Left: illustration of the considered classification problem for the Gaussian location model for the inductive risk (2); density of* $\mathcal{N}(0,1)$ *(solid line);* X_k, $k = 1, ..., m$ *(crosses); a new data point* X_{m+1} *to be classified (open circle); Bayes' rule (dotted line); FDR rule* $\hat{s}_m^{FDR}$ *for* $\alpha_m = 0.3$ *(dashed line). Right: illustration of the FDR algorithm for* $\alpha_m = 0.3$*;* $k \in \{1, ..., m\} \mapsto \overline{\Phi}^{-1}(\alpha_m k/m)$ *(solid line);* $X_{(k)}$*'s (crosses);* $\hat{s}_m^{FDR}$ *(dashed horizontal line);* $\hat{k} = 5$ *(dashed vertical line). Here,* $\overline{\Phi}(x) = \mathbb{P}(X \geq x)$ *for* $X \sim \mathcal{N}(0,1)$*.* $m = 18$*;* $\mu_m = 3$*;* $\tau_m = 5$*. For this realization,* 5 *labels "*1*" and* 13 *labels "*0*".*

the corresponding misclassification risk as small as possible. This contrasts with the standard philosophy of (multiple) testing for which α_m is meant to be a bound on the error rate and thus is fixed in the overall setting.

1.4. *Aim and scope of the paper.* Let $R_m(\cdot)$ being the risk defined either by (1) or (2). In this paper, we aim at studying the performance of FDR thresholding $\hat{h}_m = \hat{h}_m^{FDR}$ as a classification rule in terms of the excess risk $R_m(\hat{h}_m) - R_m(h_m^B)$ both in location and scale models. We investigate two types of theoretical results:

(i) Non-asymptotic oracle inequalities: prove for each (or some) $m \geq 2$, an inequality of the form

$$R_m(\hat{h}_m) - R_m(h_m^B) \leq b_m, \tag{4}$$

where b_m is an upper-bound which we aim to be "as small as possible". Typically, b_m depends on ζ, α_m and on the model parameters.

(ii) Convergence rates: find a sequence $(\alpha_m)_m$ for which there exists $D > 0$ such that for a large m,

$$R_m(\hat{h}_m) - R_m(h_m^B) \leq D \times R_m(h_m^B) \times \rho_m, \tag{5}$$

for a given rate $\rho_m = o(1)$.

Inequality (4) is meant to be non-asymptotic and involving quantities that can be explicitly derived (even the set of m for which (4) holds). It is of interest in its own right, but is also used to derive inequalities of the type (5), which are of asymptotic nature. The property (5) is called "asymptotic optimality at rate ρ_m". It implies that $R_m(\hat{h}_m) \sim R_m(h^B_m)$, that is, $\hat{h}_m$ is "asymptotically optimal", as defined in Bogdan et al. (2011). However, (5) is substantially more informative because it provides a rate of convergence.

It should be emphasized at this point that the trivial procedure $\hat{h}^0_m \equiv 0$ (which always chooses the label "0") satisfies (5) with $\rho_m = O(1)$ (under our setting (BP)). Therefore, proving (5) with $\rho_m = O(1)$ is not sufficient to get an interesting result and our goal is to obtain a rate ρ_m that tends to zero in (5). The reason for which $\hat{h}^0_m$ is already "competitive" is that we consider a sparse model in which the label "0" is generated with high probability.

1.5. *Overview of the paper.* First, Section 2 presents a more general setting than the one of Section 1.2. Namely, the location and scale models are particular cases of a general "p-value model" after a standardization of the original X_i's into p-values p_i's. While the "test statistic" formulation is often considered as more natural than the p-value one for many statisticians, the p-value formulation will be very convenient to provide a general answer to our problem. The so-obtained p-values are uniformly distributed on $(0, 1)$ under the label 0 while they follow a distribution with decreasing density f_m under the label 1. Hence, procedures of primary interest (including the Bayes rule) are p-value thresholding procedures, that choose label 1 for p-values smaller than some threshold $\hat{t}_m$. Throughout the paper, we focus on this type of procedures, and any procedure $\hat{h}_m$ is identified by its corresponding threshold $\hat{t}_m$ in the notation. Translated into this "p-value world", we describe in Section 2 the Bayes rule, the Bayes risk, condition (BP), BFDR and FDR thresholding.

The fundamental results are stated in Section 3 in the general p-value model. Following Abramovich et al. (2006); Donoho and Jin (2004, 2006); Bogdan et al. (2011), as BFDR thresholding is much easier to study than FDR thresholding from a mathematical point of view, the approach advocated here is as follows: first, we state an oracle inequality for BFDR, see Theorem 3.1. Second, we use a concentration argument of the FDR threshold around the BFDR threshold to obtain an oracle inequality of the form (4), see Theorem 3.2. At this point, the bounds involve quantities that are not written in an explicit form, and that depend on the density f_m of the p-values corresponding to the label 1.

The particular case where f_m comes either from a location or a scale model

is investigated in Section 4. An important property is that in these models, the upper-tail distribution function $\overline{D}(\cdot)$ and the quantile function $\overline{D}^{-1}(\cdot)$ can be suitably bounded , see Section S-5 in the supplemental article. By using this property, we derive from Theorems 3.1 and 3.2 several inequalities of the form (4) and (5). In particular, in the sparsity regime $\tau_m = m^\beta$, $0 < \beta \leq 1$, we derive that the FDR threshold $\hat{t}_m^{FDR}$ at level α_m is asymptotically optimal (under (BP) and (Sp)) in either of the following two cases:

- for the location model, $\zeta > 1$, if $\alpha_m \to 0$ and $\log \alpha_m = o\left((\log m)^{1-1/\zeta}\right)$;
- for the scale model, $\zeta \geq 1$, if $\alpha_m \to 0$ and $\log \alpha_m = o(\log m)$.

The latter is in accordance with the condition found in Bogdan et al. (2011) in the Gaussian scale model. Furthermore, choosing $\alpha_m \propto 1/(\log m)^{1-1/\zeta}$ (location) or $\alpha_m \propto 1/(\log m)$ (scale) provides a convergence rate $\rho_m = 1/(\log m)^{1-1/\zeta}$ (location) or $\rho_m = 1/(\log m)$ (scale), respectively.

At this point, one can argue that the latter convergence results are not fully satisfactory: first, these results do not provide an explicit choice for α_m for a given finite value of m. Second, the rate of convergence ρ_m being rather slow, we should check numerically that FDR thresholding has reasonably good performance for a moderately large m.

We investigate the choice of α_m by carefully studying Bayes' thresholding and how it is related to BFDR thresholding, see Sections 2.4 and 4.4. Next, for this choice of α_m, the performance of FDR thresholding is evaluated numerically in terms of (relative) excess risk, for several values of m, see Section 5. We show that the excess risk of FDR thresholding is small for a remarkably wide range of values for β, and increasingly so as m grows to infinity. This illustrates the adaptation of FDR thresholding to the unknown sparsity regime. Also, for comparison, we show that choosing α_m fixed with m (say, $\alpha_m \equiv 0.1$) can lead to higher FDR thresholding excess risk.

2. General setting.

2.1. *p-value model.* Let $(p_i, H_i) \in [0,1] \times \{0,1\}$, $1 \leq i \leq m$, be m i.i.d. variables. The distribution of (p_1, H_1) is assumed to belong to a specific subset of distributions on $[0,1] \times \{0,1\}$, which is defined as follows:

(i) same as (i) in Section 1.2;
(ii) the distribution of p_1 conditionally on $H_1 = 0$ is uniform on $(0,1)$;
(iii) the distribution of p_1 conditionally on $H_1 = 1$ has a c.d.f. F_m satisfying

($\mathrm{A}(F_m, \tau_m)$)
F_m is continuously increasing on $[0,1]$ and differentiable on $(0,1)$,
$f_m = F_m'$ is continuously decreasing with $f_m(0^+) > \tau_m > f_m(1^-)$.

This way, we obtain a family of i.i.d. p-values, where each p-value has a marginal distribution following the mixture model:

$$p_i \sim \pi_{0,m} U(0,1) + \pi_{1,m} F_m. \tag{6}$$

The model (6) is classical in the multiple testing literature and is usually called the "two-group mixture model". It has been widely used since its introduction by Efron et al. (2001) Efron et al. (2001), see for instance Storey (2003); Genovese and Wasserman (2004); Efron (2008); Bogdan et al. (2011).

The models presented in Section 1.2 are particular instances of this p-value model. In the scale model, we apply the standardization $p_i = 2\overline{D}(|X_i|)$, which yields $F_m(t) = 2\overline{D}(\overline{D}^{-1}(t/2)/\sigma_m)$. In the location model, we let $p_i = \overline{D}(X_i)$, which yields $F_m(t) = \overline{D}(\overline{D}^{-1}(t) - \mu_m)$. We can easily check that in both cases (A(F_m, τ_m)) is satisfied (additionally assuming $\zeta > 1$ for the location model), with $f_m(0^+) = +\infty$ and $f_m(1^-) < 1$ (scale) and $f_m(1^-) = 0$ (location), as proved in Section S-2.1 in the supplemental article.

2.2. *Procedures, risks and the Bayes threshold.* A classification procedure is identified with a threshold $\hat{t}_m \in [0,1]$, that is, a measurable function of the p-value family $(p_i, i \in \{1,...,m\})$. The corresponding procedure chooses label 1 whenever the p-value is smaller than $\hat{t}_m$. In the p-value setting, the transductive and inductive misclassification risks of a threshold $\hat{t}_m$ can be written as follows:

$$R_m^T(\hat{t}_m) = m^{-1} \sum_{i=1}^{m} \mathbb{P}(p_i \le \hat{t}_m, H_i = 0) + m^{-1} \sum_{i=1}^{m} \mathbb{P}(p_i > \hat{t}_m, H_i = 1) \tag{7}$$

$$R_m^I(\hat{t}_m) = \mathbb{E}\big(\pi_{0,m}\hat{t}_m + \pi_{1,m}(1 - F_m(\hat{t}_m))\big). \tag{8}$$

In the particular case of a deterministic threshold $t_m \in [0,1]$, these two risks coincide and are equal to $R_m(t_m) = \pi_{0,m} t_m + \pi_{1,m}(1 - F_m(t_m))$. The following lemma identifies a solution minimizing both risks (7) and (8).

LEMMA 2.1. *Let $R_m(\cdot)$ being either $R_m^T(\cdot)$ or $R_m^I(\cdot)$. Under Assumption (A(F_m, τ_m)), the threshold*

$$t_m^B = f_m^{-1}(\tau_m) \in (0,1) \tag{9}$$

minimizes $R_m(\cdot)$, i.e., satisfies $R_m(t_m^B) = \min_{\hat{t}'_m} \{R_m(\hat{t}'_m)\}$, where the minimum is taken over all measurable functions from $[0,1]^m$ to $[0,1]$ that take as input the p-value family $(p_i, i \in \{1,...,m\})$.

The threshold t_m^B is called the *Bayes threshold* and $R_m(t_m^B)$ is called the *Bayes risk*. The Bayes threshold is unknown because it depends on τ_m and on the data distribution f_m.

Notation. In this paper, all the statements hold for both risks. Hence, throughout the paper, $R_m(\cdot)$ denotes either $R_m^T(\cdot)$ defined by (7) or $R_m^I(\cdot)$ defined by (8).

2.3. *Assumptions on the power of the Bayes rule and sparsity.* Under Assumption (A(F_m, τ_m)), let us denote the power of the Bayes procedure by

$$C_m = F_m(t_m^B) \in (0,1). \tag{10}$$

In our setting, we will typically assume that the signal is sparse while the power C_m of the Bayes procedure remains away from 0 or 1:

(BP) $\exists (C_-, C_+)$ s.t. $\forall m \geq 2, \;\; 0 < C_- \leq C_m \leq C_+ < 1$;

(Sp) $(\tau_m)_m$ is such that $\tau_m \to +\infty$ as $m \to +\infty$.

First note that Assumption (Sp) is very weak: it is required as soon as we assume some sparsity in the data. As a typical instance, $\tau_m = m^\beta$ satisfies (Sp), for any $\beta > 0$. Next, Assumption (BP) means that the best procedure is able to detect a "moderate" amount of signal. In Bogdan et al. (2011), a slightly stronger assumption has been introduced:

(VD) $\exists C \in (0,1)$ s.t. $C_m \to C$ as m tends to infinity,

which is referred to as "the verge of detectability". Condition (BP) encompasses (VD) and is more suitable to state explicit finite sample oracle inequalities (see, e.g., Remark 4.6 further on).

In the location (resp., scale) model, while the original parameters are (μ_m, τ_m) (resp., (σ_m, τ_m)), the model can be parametrized in function of (C_m, τ_m) by using (9) and (10). This way, F_m is uniquely determined from (C_m, τ_m) as follows: among the family of curves $\{\overline{D}(\overline{D}^{-1}(\cdot) - \mu)\}_{\mu \in \mathbb{R}}$ (resp., $\{2\overline{D}(\overline{D}^{-1}(\cdot/2)/\sigma)\}_{\sigma > 1}$), F_m is the unique curve such that the pre-image of C_m has a tangent of slope τ_m, that is, $f_m(F_m^{-1}(C_m)) = \tau_m$. This is illustrated in Figure 2 for the Laplace scale model. In this case, $\overline{D}(x) = d(x) = e^{-x}/2$ for $x \geq 0$ and thus $F_m(t) = t^{1/\sigma_m}$, so that the family of curves is simply $\{t \mapsto t^{1/\sigma}\}_{\sigma > 1}$.

REMARK 2.2. *Condition* (BP) *constrains the model parameters to be located in a very specific region. For instance, in the Gaussian location model with* $\tau_m = m^\beta$, *Condition* (BP) *implies that* $\mu_m \sim \sqrt{2\beta \log m}$ *(see Table S-3 in the supplemental article), which corresponds to choosing* (μ_m, β) *on the "estimation boundary", as displayed in Figure 1 of Donoho and Jin (2004).*

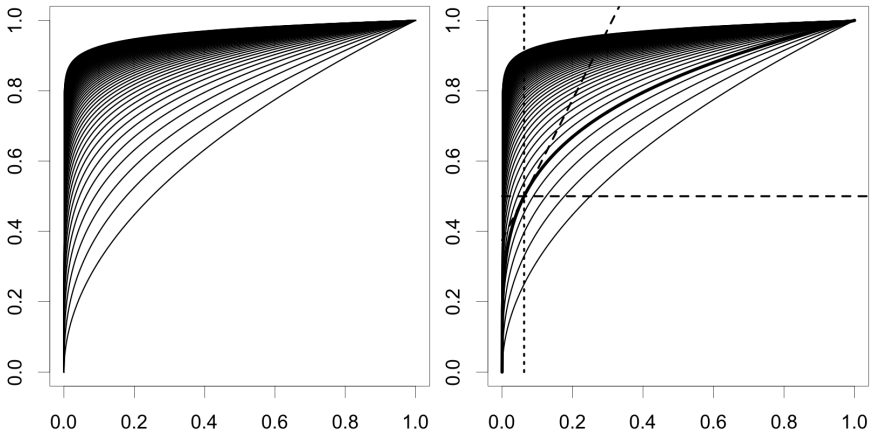

FIG 2. *Left: plot of the family of curves* $\{t \mapsto t^{1/(2+j/2)}\}_{j=0,\dots,56}$ *(thin solid curves). Right: choice (thick solid curve) within the family of curves* $\{t \mapsto t^{1/\sigma}\}_{\sigma>1}$ *that fulfills* (9) *and* (10) *for* $C_m = 1/2$ *(given by the dashed horizontal line) and* $\tau_m = 2$ *(slope of the dashed oblique line). This gives* $\sigma_m \simeq 4$*. The Bayes threshold* t_m^B *is given by the dotted vertical line.*

2.4. *BFDR thresholding.* Let us consider the following Bayesian quantity:

$$\text{BFDR}_m(t) = \mathbb{P}(H_i = 0 \mid p_i \le t) = \frac{\pi_{0,m} t}{G_m(t)} = \left(1 + \tau_m^{-1} F_m(t)/t\right)^{-1}, \tag{11}$$

for any $t \in (0,1)$ and where $G_m(t) = \pi_{0,m} t + \pi_{1,m} F_m(t)$. As introduced by Efron and Tibshirani (2002), the quantity defined by (11) is called "Bayesian FDR". It is not to be confounded with "Bayes FDR" defined by Sarkar et al. (2008). Also, under a two-class mixture model, $\text{BFDR}_m(t)$ coincides with the so-called "positive false discovery rate", itself connected to the original false discovery rate of Benjamini and Hochberg (1995), see Storey (2003) and Section 4 of Bogdan et al. (2011).

Under Assumption (A(F_m, τ_m)), the function $\Psi_m : t \in (0,1) \mapsto F_m(t)/t$ is decreasing from $f_m(0^+)$ to 1, with $f_m(0^+) \in (1, +\infty]$. Hence, the following result holds.

LEMMA 2.3. *Assume* (A(F_m, τ_m)) *and* $\alpha_m \in ((1 + f_m(0^+)/\tau_m)^{-1}, \pi_{0,m})$. *Then, equation* $BFDR_m(t) = \alpha_m$ *has a unique solution* $t = t_m^\star(\alpha_m) \in (0,1)$, *given by*

$$t_m^\star(\alpha_m) = \Psi_m^{-1}(q_m \tau_m), \tag{12}$$

for $q_m = \alpha_m^{-1} - 1 > 0$ *and* $\Psi_m(t) = F_m(t)/t$.

The threshold $t_m^\star(\alpha_m)$ is called the *BFDR threshold* at level α_m. Typically, it is well defined for any $\alpha_m \in (0, 1/2)$, because $\pi_{0,m} > 1/2$ and $f_m(0^+) = +\infty$[1] in the Subbotin location and scale models (additionally assuming $\zeta > 1$ for the location model). Obviously, the BFDR threshold is unknown because it depends on τ_m and on the distribution of the data. However, its interest lies in that it is close to the FDR threshold which is observable. For short, $t_m^\star(\alpha_m)$ will be denoted by $t_m^\star$ when not ambiguous.

Next, a quantity of interest in Lemma 2.3 is $q_m = \alpha_m^{-1} - 1 > 0$, called the *recovery parameter* (associated to α_m). As $\alpha_m = (1 + q_m)^{-1}$, considering α_m or q_m is equivalent. Since we would like to have $t_m^\star = \Psi_m^{-1}(q_m\tau_m)$ close to $t_m^B = f_m^{-1}(\tau_m)$, *the recovery parameter can be interpreted as a correction factor that cancels the difference between* $\Psi_m(t) = F_m(t)/t$ *and* $f_m(t) = F_m'(t)$. Clearly, the best choice for the recovery parameter is such that $t_m^\star = t_m^B$, that is,

$$q_m^{opt} = \tau_m^{-1}\Psi_m(f_m^{-1}(\tau_m)) = \frac{C_m}{\tau_m t_m^B}, \tag{13}$$

which is an unknown quantity, called the *optimal recovery parameter*. Note that from the concavity of F_m, we have $\Psi_m(t) \geq f_m(t)$ and thus $q_m^{opt} \geq 1$. As an illustration, for the Laplace scale model, we have $\sigma_m f_m(t) = \Psi_m(t)$ and thus the optimal recovery parameter is $q_m^{opt} = \sigma_m$.

The fact that $q_m^{opt} \geq 1$ suggests to always choose $q_m \geq 1$ (that is, $\alpha_m \leq 1/2$) into the BFDR threshold. A related result is that taking any sequence $(\alpha_m)_m$ such that $\alpha_m \geq \alpha_- > 1/2$ for all $m \geq 2$ never leads to an asymptotically optimal BFDR procedure, see Section S-6 in the supplemental article.

2.5. *FDR thresholding.* The first occurrence of FDR thresholding procedures seems to be in a series of papers by Eklund (1961-1963), see Seeger (1968). Much later, this procedure has been carefully studied by Benjamini and Hochberg (1995) by proving that it controls the FDR, see Benjamini and Hochberg (1995). As noted by many authors (see, e.g., (Sen, 1999; Efron and Tibshirani, 2002; Storey, 2002; Genovese and Wasserman, 2002; Abramovich et al., 2006)), this thresholding rule can be expressed as a function of the empirical c.d.f. $\widehat{\mathbb{G}}_m$ of the p-values in the following way: for any $\alpha_m \in (0, 1)$,

$$\hat{t}_m^{BH}(\alpha_m) = \max\{t \in [0, 1] \ : \ \widehat{\mathbb{G}}_m(t) \geq t/\alpha_m\}. \tag{14}$$

We simply denote $\hat{t}_m^{BH}(\alpha_m)$ by $\hat{t}_m^{BH}$ when not ambiguous. Classically, this implies that $t = \hat{t}_m^{BH}$ solves the equation $\widehat{\mathbb{G}}_m(t) = t/\alpha_m$ (this can be easily

[1]This condition implies that the setting is "non-critical", as defined in Chi (2007).

shown by using (14) together with the fact that $\widehat{\mathbb{G}}_m(\cdot)$ is a non-decreasing function). Hence, according to Lemma 2.3 and as already mentioned in the literature (see Bogdan et al. (2011)), $\hat{t}_m^{BH}$ can be seen as an empirical counterpart of the BFDR threshold at level $\alpha_m\pi_{0,m}$, in which the theoretical c.d.f. $G_m(t) = \pi_{0,m}t + \pi_{1,m}F_m(t)$ of the p-values has been replaced by the empirical c.d.f. $\widehat{\mathbb{G}}_m$ of the p-values. Next, once α_m has been chosen, (14) only involves observable quantities, so that the threshold $\hat{t}_m^{BH}$ only depends on the data. This is further illustrated on the left panel of Figure 3. Also, as already observed in Section 5.2 of Bogdan et al. (2011), since the BH procedure is never more conservative than the Bonferroni procedure, the following modification of $\hat{t}_m^{BH}$ can be proposed:

DEFINITION 2.4. *The FDR threshold at level α_m is defined by*

$$\hat{t}_m^{FDR}(\alpha_m) = \hat{t}_m^{BH}(\alpha_m) \vee (\alpha_m/m), \tag{15}$$

where $\hat{t}_m^{BH}(\alpha_m)$ is defined by (14).

We simply denote $\hat{t}_m^{FDR}(\alpha_m)$ by $\hat{t}_m^{FDR}$ when not ambiguous. The threshold $\hat{t}_m^{FDR}$ is the one that we use throughout this paper. This modification does not change the risk $R_m^T(\cdot)$, that is, $R_m^T(\hat{t}_m^{BH}) = R_m^T(\hat{t}_m^{FDR})$, but can affect the risk $R_m^I(\cdot)$, that is, $R_m^I(\hat{t}_m^{BH}) \neq R_m^I(\hat{t}_m^{FDR})$ in general.

Finally, while relation (15) uses p-values whereas the algorithms defined in Section 1.3 use test statistics, it is easy to check that the resulting procedures are the same.

REMARK 2.5 (Adaptive FDR procedures under sparsity). *To get a better FDR controlling procedure, one classical approach is to modify (15) by dividing α_m by a (more or less explicit) estimator of $\pi_{0,m}$ and by possibly using a step-up-down algorithm, see e.g., Tamhane et al. (1998); Sarkar (2002); Benjamini et al. (2006); Sarkar (2008); Blanchard and Roquain (2009); Finner et al. (2009); Gavrilov et al. (2009). However, this method seems not helpful in our sparse setting because $\pi_{0,m}$ is very close to 1. As a result, we focus in this paper only on the original (non-adaptive) version of FDR thresholding (15).*

3. Results in the general model. This section presents relations of the form (4) and (5) for the BFDR and FDR thresholds. Our first main result deals with the BFDR threshold.

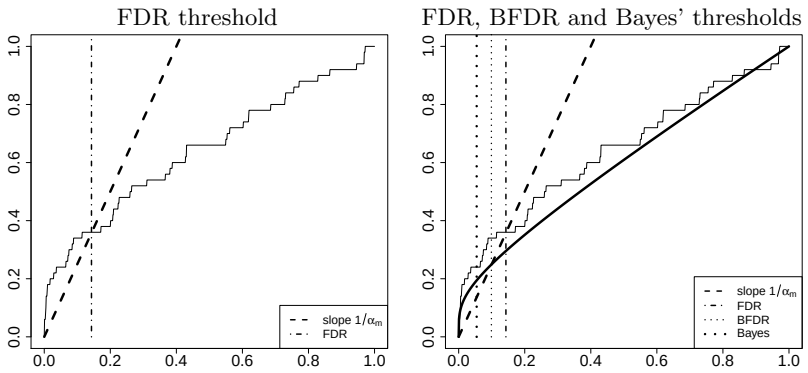

FIG 3. *Left: illustration of the FDR threshold* (15)*: e.c.d.f. of the p-value (solid line), line of slope* $1/\alpha_m$ *(dotted line), the FDR threshold at level* α_m *(X-coordinate of the vertical dashed dotted line). Right: illustration of the FDR threshold as an empirical surrogate for the BFDR threshold; compared to the left picture, we added the c.d.f. of the p-values (thick solid line), the BFDR threshold at level* $\alpha_m \pi_{0,m}$ *(dotted vertical line) and the Bayes threshold (dashed vertical line). In both panels, we consider the Laplace scale model with* $C_m = 0.5$*;* $m = 50$*;* $\beta = 0.2$*;* $\tau_m = m^\beta$*;* $\sigma_m \simeq 4.2$*;* $\alpha_m = 0.4$.

THEOREM 3.1. *Assume* (A(F_m, τ_m)) *and consider the BFDR threshold* $t^\star_m$ *at a level* $\alpha_m \in ((1 + f_m(0^+)/\tau_m)^{-1}, \pi_{0,m})$ *corresponding to a recovery parameter* $q_m = \alpha_m^{-1} - 1$*. Consider* $q_m^{opt} \geq 1$ *the optimal recovery parameter given by* (13)*. Then the following holds:*

(i) if $\alpha_m \leq 1/2$*, we have for any* $m \geq 2$*,*

$$R_m(t^\star_m) - R_m(t^B_m) \leq \pi_{1,m}\{(C_m/q_m - C_m/q_m^{opt}) \vee \gamma_m\}, \tag{16}$$

where we let $\gamma_m = (C_m - F_m(\Psi_m^{-1}(q_m\tau_m)))_+$.
In particular, under (BP)*, if* $\alpha_m \to 0$ *and* $\gamma_m \to 0$*, the BFDR threshold* $t^\star_m$ *is asymptotically optimal at rate* $\rho_m = \alpha_m + \gamma_m$.

(ii) we have for any $m \geq 2$*,*

$$\frac{R_m(t^\star_m)}{R_m(t^B_m)} \geq \frac{\pi_{1,m}}{R_m(t^B_m)} \left(1 - (1 - q_m^{-1})_+ F_m(q_m^{-1}\tau_m^{-1})\right). \tag{17}$$

In particular, under (BP)*, if* $R_m(t^B_m) \sim \pi_{1,m}(1 - C_m)$ *and if* q_m *is chosen such that*

$$\liminf_m \left\{ \frac{1 - (1 - q_m^{-1})_+ F_m(q_m^{-1}\tau_m^{-1})}{1 - C_m} \right\} > 1, \tag{18}$$

$t^\star_m$ *is not asymptotically optimal.*

Theorem 3.1 is proved in Section 7. Theorem 3.1 (i) presents an upper-bound for the excess risk when choosing q_m instead of q_m^{opt} in BFDR thresholding. First, both sides of (16) are equal to zero when $q_m = q_m^{opt}$. Hence, this bound is sharp in that case. Second, Assumption "$\alpha_m \leq 1/2$" in Theorem 3.1 (i) is only a technical detail that allows to get C_m/q_m instead of $1/q_m$ in the RHS of (16) (moreover, it is quite natural, see the end of Section 2.4). Third, γ_m has a simple interpretation as the difference between the power of Bayes' thresholding and BFDR thresholding. Fourth, the bound (16) induces the following trade-off for choosing α_m: on the one hand, α_m has to be chosen small enough to make C_m/q_m small; on the other hand, γ_m increases as α_m decreases to zero. Finally note that, in Theorem 3.1 (i), the second statement is a consequence of the first one because $R_m(t_m^B) \geq \pi_{1,m}(1-C_m)$. Theorem 3.1 (ii) states lower bounds which are useful to identify regimes of α_m that do not lead to an asymptotically optimal BFDR thresholding (see Corollary 4.4 (i) further on).

Our second main result deals with FDR thresholding.

THEOREM 3.2. *Let $\varepsilon \in (0,1)$, assume (A(F_m, τ_m)) and consider the FDR threshold $\hat{t}_m^{FDR}$ at level $\alpha_m > (1-\varepsilon)^{-1}(\pi_{0,m} + \pi_{1,m} f_m(0^+))^{-1}$. Then the following holds: for any $m \geq 2$,*

$$R_m(\hat{t}_m^{FDR}) - R_m(t_m^B) \leq \pi_{1,m} \frac{\alpha_m}{1-\alpha_m} + m^{-1} \frac{\alpha_m}{(1-\alpha_m)^2} + \pi_{1,m} \left\{ \gamma'_m \wedge \left(\gamma_m^{\varepsilon} + e^{-m\varepsilon^2 (\tau_m+1)^{-1}(C_m - \gamma_m^{\varepsilon})/4} \right) \right\}, \tag{19}$$

for $\gamma_m^{\varepsilon} = (C_m - F_m(\Psi_m^{-1}(q_m^{\varepsilon}\tau_m)))_+$ with $q_m^{\varepsilon} = (\alpha_m \pi_{0,m}(1-\varepsilon))^{-1} - 1$ and $\gamma'_m = (C_m - F_m(\alpha_m/m))_+$. In particular, under (BP) and assuming $\alpha_m \to 0$,

(i) if $m/\tau_m \to +\infty$, $\gamma_m^{\varepsilon} \to 0$ and additionally $\forall \kappa > 0$, $e^{-\kappa m/\tau_m} = o(\gamma_m^{\varepsilon})$, the FDR threshold $\hat{t}_m^{FDR}$ is asymptotically optimal at rate $\rho_m = \alpha_m + \gamma_m^{\varepsilon}$.

(ii) if $m/\tau_m \to l \in (0, +\infty)$ with $\gamma'_m \to 0$, the FDR threshold $\hat{t}_m^{FDR}$ is asymptotically optimal at rate $\rho_m = \alpha_m + \gamma'_m$.

Theorem 3.2 is proved in Section 7. The proof mainly follows the methodology of Bogdan et al. (2011), but is more general and concise. It relies on tools developed in Finner and Roters (2002); Genovese and Wasserman (2002); Ferreira and Zwinderman (2006); Finner et al. (2009); Roquain and Villers (2011); Roquain (2011). The main argument for the proof is that the FDR threshold $\hat{t}_m^{FDR}(\alpha_m)$ is either well concentrated around the BFDR threshold $t_m^{\star}(\alpha_m \pi_{0,m})$ (as illustrated in the right panel of Figure 3) or close

to the Bonferroni threshold α_m/m. This argument was already used in Bogdan et al. (2011).

Let us comment briefly on Theorem 3.2: first, as in the BFDR case, choosing α_m such that the bound in (19) is minimal involves a trade-off because γ_m^{ε} and γ_m' are quantities that increase when α_m decreases to zero. Second, let us note that cases (i) and (ii) in Theorem 3.2 are intended to cover regimes where the FDR is close to BFDR (moderately sparse) and where the FDR threshold is close to the Bonferroni threshold (extremely sparse), respectively. In particular, these two regimes cover the case where $\tau_m = m^{\beta}$ with $\beta \in (0, 1]$. Finally, the bounds and convergence rates derived in Theorems 3.1 and 3.2 strongly depend on the nature of F_m. We derive a more explicit expression of the latter in the next section, in the particular cases of location and scale models coming from a Subbotin density.

REMARK 3.3 (Conservative upper-bound for γ_m). *By the concavity of F_m, we have $q_m\tau_m = \Psi_m(t_m^\star) \geq f_m(t_m^\star)$, which yields*

$$\gamma_m \leq C_m - F_m(f_m^{-1}(q_m\tau_m)) \in [0, 1). \tag{20}$$

When f_m^{-1} is easier to use than Ψ_m^{-1}, it is tempting to use (20) to upper bound the excess risk in Theorems 3.1 and 3.2. However, this can inflate the resulting upper-bound too much. This point is discussed in Section S-3.4 in the supplemental article for the case of a Gaussian density (for which this results in an additional $\log\log\tau_m$ factor in the bound).

4. Application to location and scale models.

4.1. *The Bayes risk and optimal recovery parameter.* A preliminary task is to study the behavior of t_m^B, $R_m(t_m^B)$ and $q_m^{opt} = C_m/(\tau_m t_m^B)$ both in location and scale models. While finite sample inequalities are given in Section S-2.2 in the supplemental article, we only report in this subsection some resulting asymptotic relations for short. Let us define the following rates, which will be useful throughout the paper:

$$r_m^{loc} = (\zeta \log \tau_m + |\overline{D}^{-1}(C_m)|^{\zeta})^{1-1/\zeta}; \tag{21}$$

$$r_m^{sc} = \zeta \log \tau_m + (\overline{D}^{-1}(C_m/2))^{\zeta}. \tag{22}$$

Under (Sp), note that the rates r_m^{loc} (resp., r_m^{sc}) tend to infinity. Furthermore, by using Section S-2.2 in the supplemental article, we have $\mu_m = (r_m^{loc})^{1/(\zeta-1)} - \overline{D}^{-1}(C_m)$ in the location model and $\sigma_m \geq (r_m^{sc})^{1/\zeta}/(\overline{D}^{-1}(C_m/2))$ in the scale model.

PROPOSITION 4.1. *Consider a ζ-Subbotin density (3) with $\zeta \geq 1$ for a scale model and $\zeta > 1$ for a location model. Let $(\tau_m, C_m) \in (1,\infty) \times (0,1)$ be the parameters of the model. Let r_m be equal to r_m^{loc} defined by (21) in the location model or to r_m^{sc} defined by (22) in the scale model. Then, under (BP) and (Sp), we have $\mu_m \sim r_m^{loc} \sim (\zeta \log \tau_m)^{1/\zeta}$ and $\sigma_m \sim (r_m^{sc})^{1/\zeta}/(\overline{D}^{-1}(C_m/2)) \sim (\zeta \log \tau_m)^{1/\zeta}/(\overline{D}^{-1}(C_m/2))$ and*

$$R_m(t_m^B) \sim \pi_{1,m}(1 - C_m) \tag{23}$$

$$t_m^B = O\left(R_m(t_m^B)/r_m\right) \tag{24}$$

$$q_m^{opt} \sim \begin{cases} \frac{C_m}{d(\overline{D}^{-1}(C_m))}(\zeta \log \tau_m)^{1-1/\zeta} & \textit{(location)} \\ \frac{C_m/2}{\overline{D}^{-1}(C_m/2)d(\overline{D}^{-1}(C_m/2))}\,\zeta \log \tau_m & \textit{(scale).} \end{cases} \tag{25}$$

From (23) and (24), by assuming (BP) and (Sp), the probability of a type I error $(\pi_{0,m} t_m^B)$ is always of smaller order than the probability of a type II error $(\pi_{1,m}(1 - C_m))$. The latter had already been observed in Bogdan et al. (2011) in the particular case of a Gaussian scale model.

REMARK 4.2. *From (23) and since the risk of null thresholding is $R_m(0) = \pi_{1,m}$, a substantial improvement over the null threshold can only be expected in the regime where $C_m \geq C_-$, where C_- is "far" from 0.*

4.2. *Finite sample oracle inequalities.* The following result can be derived from Theorem 3.1 (i) and Theorem 3.2. It is proved in Section S-2.3 in the supplemental article.

COROLLARY 4.3. *Consider a ζ-Subbotin density (3) with $\zeta > 1$ for a location model and $\zeta \geq 1$ for a scale model, and let $(\tau_m, C_m) \in (1,\infty) \times (0,1)$ be the parameters of the model. Let $r_m = r_m^{loc}$ (defined by (21)) and $K_m = d(0)$ in the location model or $r_m = r_m^{sc}$ (defined by (22)) and $K_m = 2\overline{D}^{-1}(C_m/2)d(\overline{D}^{-1}(C_m/2))$ in the scale model. Let $\alpha_m \in (0, 1/2)$ and denote the corresponding recovery parameter by $q_m = \alpha_m^{-1} - 1$. Consider $q_m^{opt} \geq 1$ the optimal recovery parameter given by (13). Let $\nu \in (0,1)$. Then:*

(i) The BFDR threshold $t_m^\star$ at level α_m defined by (12) satisfies that for any $m \geq 2$ such that $r_m \geq \frac{K_m}{C_m(1-\nu)}(\log(q_m/q_m^{opt}) - \log \nu)$,

$$\begin{aligned} &R_m(t_m^\star) - R_m(t_m^B) \\ &\leq \pi_{1,m}\left\{\left(\frac{C_m}{q_m} - \frac{C_m}{q_m^{opt}}\right) \vee \left(K_m \frac{\log(q_m/q_m^{opt}) - \log \nu}{r_m}\right)\right\}; \end{aligned} \tag{26}$$

(ii) Letting $\varepsilon \in (0,1)$, $D_{1,m} = -\log(\nu\pi_{0,m}(1-\varepsilon))$ *and* $D_{2,m} = \log(\nu^{-1}C_m\tau_m^{-1}m)$, *the FDR threshold* $\hat{t}_m^{FDR}$ *at level* α_m *defined by* (15) *satisfies that, for any* $a \in \{1,2\}$, *for any* $m \geq 2$ *such that* $r_m \geq \frac{K_m}{C_m(1-\nu)}(\log(\alpha_m^{-1}/q_m^{opt}) + D_{a,m})$,

$$R_m(\hat{t}_m^{FDR}) - R_m(t_m^B) \leq \pi_{1,m}\left(\frac{\alpha_m}{1-\alpha_m} + K_m\frac{(\log(\alpha_m^{-1}/q_m^{opt}) + D_{a,m})_+}{r_m}\right)$$
$$+\frac{\alpha_m/m}{(1-\alpha_m)^2} + \pi_{1,m}\mathbf{1}\{a=1\}e^{-m(\tau_m+1)^{-1}\nu\varepsilon^2C_m/4}. \tag{27}$$

Corollary 4.3 (ii) contains two distinct cases. The case $a = 1$ should be used when m/τ_m is large, because the remainder term containing the exponential becomes small (whereas $D_{1,m}$ is approximately constant). The case $a = 2$ is intended to deal with the regime where m/τ_m is not large, because $D_{2,m}$ is of the order of a constant in that case. The finite sample oracle inequalities (26) and (27) are useful to derive explicit rates of convergence, as we will see in the next section. Let us also mention that an exact computation of the excess risk of BFDR thresholding can be derived in the Laplace case, see Section S-3.2 in the supplemental article.

4.3. *Asymptotic optimality with rates.* In this section, we provide a sufficient condition on α_m such that, under (BP) and (Sp), BFDR/FDR thresholding is asymptotically optimal (according to (5)) and we provide an explicit rate ρ_m. Furthermore, we establish that this condition is necessary for the optimality of BFDR thresholding.

COROLLARY 4.4. *Take* $\zeta > 1$, $\gamma = 1 - \zeta^{-1}$ *for the location case and* $\zeta \geq 1$, $\gamma = 1$ *for the scale case. Consider a* ζ*-Subbotin density* (3) *in the sparsity regime* $\tau_m = m^\beta$, $0 < \beta \leq 1$ *and under* (BP). *Then the following holds:*

(i) The BFDR threshold $t_m^\star$ *is asymptotically optimal if and only if*

$$\alpha_m \to 0 \text{ and } \log\alpha_m = o\left((\log m)^\gamma\right), \tag{28}$$

in which case it is asymptotically optimal at rate $\rho_m = \alpha_m + \frac{(\log(\alpha_m^{-1}/(\log m)^\gamma))_+}{(\log m)^\gamma}$.

(ii) The FDR threshold $\hat{t}_m^{FDR}$ *at a level* α_m *satisfying* (28) *is asymptotically optimal at rate* $\rho_m = \alpha_m + \frac{(\log(\alpha_m^{-1}/(\log m)^\gamma))_+}{(\log m)^\gamma}$.

(iii) Choosing $\alpha_m \propto 1/(\log m)^\gamma$, *BFDR and FDR thresholding are both asymptotically optimal at rate* $\rho_m = 1/(\log m)^\gamma$.

In the particular case of a Gaussian scale model ($\zeta = 2$), Corollary 4.4 recovers Corollary 4.2 and Corollary 5.1 of Bogdan et al. (2011). Corollary 4.4 additionally provides a rate, and encompasses the location case and other values of ζ.

REMARK 4.5 (Lower bound for the Laplace scale model). *We can legitimately ask whether the rate $\rho_m = (\log m)^{-\gamma}$ can be improved. We show that this rate is the smallest that one can obtain over a sparsity class $\beta \in [\beta_-, 1]$ for some $\beta_- \in (0,1)$, in the particular case of BFDR thresholding and in the Laplace scale model, see Corollary S-3.2 in the supplemental article. While the calculations become significantly more difficult in the other models, we believe that the minimal rate for the relative excess risk of the BFDR is still $(\log m)^{-\gamma}$ in a Subbotin location and scale models. Also, since the FDR can be seen as a stochastic variation around the BFDR, we may conjecture that this rate is also minimal for FDR thresholding.*

4.4. *Choosing α_m.* Let us consider the sparsity regime $\tau_m = m^\beta$, $\beta \in (0,1)$. Corollary 4.4 suggests to choose α_m such that $\alpha_m \propto (\log m)^{-\gamma}$. This is in accordance with the recommendation of Bogdan et al. (2011) in the Gaussian scale model, see Remark 5.3 therein. In this section, we propose an explicit choice of α_m from an *a priori* value (β_0, C_0) of the unknown parameter (β, C_m).

Let us choose a value (β_0, C_0) *a priori* for (β, C_m). A natural choice for α_m is the value which would be optimal if the parameters of the model were $(\beta, C_m) = (\beta_0, C_0)$. Namely, by using (13) in Section 2.4, we choose $\alpha_m = \alpha_m^{opt}(\beta_0, C_0)$, where

(29)
$$\alpha_m^{opt}(\beta_0, C_0) = \big(1 + q_m^{opt}(\beta_0, C_0)\big)^{-1} \text{ with } q_m^{opt}(\beta_0, C_0) = m^{-\beta_0} C_0 / F_{m,0}^{-1}(C_0),$$

by denoting $F_{m,0}$ the c.d.f. of the p-values following the alternative for the model parameters (β_0, C_0). For instance,

- Gaussian location: $F_{m,0}^{-1}(C_0) = \overline{\Phi}\left(\left\{\overline{\Phi}^{-1}(C_0)^2 + 2\beta_0 \log m\right\}^{1/2}\right)$;
- Gaussian scale: $F_{m,0}^{-1}(C_0) = 2\overline{\Phi}\left(\overline{\Phi}^{-1}(C_0/2)x\right)$, where $x > 1$ is the solution of $2\beta_0 \log m + 2\log x = (\overline{\Phi}^{-1}(C_0/2))^2(x^2 - 1)$;
- Laplace scale: $q_m^{opt}(\beta_0, C_0) = y$, where $y > 1$ is the solution of $\beta_0 \log m + \log y = (y-1)\log(1/C_0)$,

where $\overline{\Phi}(z)$ denotes $\mathbb{P}(Z \geq z)$ for $Z \sim \mathcal{N}(0,1)$.

The above choice of α_m does depend on (β_0, C_0), which can be interpreted as a "guess" on the value of the unknown parameter (β, C_m). Hence, when

no prior information on (β, C_m) is available from the data, the above choice of α_m can appear of limited interest in practice. However, we would like to make the following two points:

- asymptotically, choosing $\alpha_m = \alpha_m^{opt}(\beta_0, C_0)$ always yields an optimal (B)FDR thresholding (under (BP)), even if $(\beta_0, C_0) \neq (\beta, C_m)$: by Proposition 4.1, we get $\alpha_m^{opt}(\beta_0, C_0) \propto (\log m)^{-\gamma}$ and thus the asymptotic optimality is a direct consequence of Corollary 4.4 (iii);
- non-asymptotically, our numerical experiments suggest that $\alpha_m = \alpha_m^{opt}(\beta_0, C_0)$ performs fairly well when we have at hand an *a priori* on the location of the model parameters: if (β, C_m) is supposed to be in some specific (but possibly large) region of the "sparsity×power" square, choosing any (β_0, C_0) in that region yields a thresholding procedure with a reasonably small risk, see Section 5 and Section S-7 in the supplemental article.

Finally, let us note that the choice $\alpha_m = \alpha_m^{opt}(\beta_0, C_0)$ is motivated by the analysis of the BFDR risk, not that of the FDR risk. Hence, it might be possible to choose a better α_m for FDR thresholding, especially for small values of m for which BFDR and FDR are different. Because obtaining such a refinement appeared quite challenging, and as our proposed choice already performed well, we decided not to investigate this question further.

REMARK 4.6. *By choosing $\alpha_m = \alpha_m^{opt}(\beta_0, C_0)$ as in (29), we can legitimately ask how large the constants are in the finite sample inequalities coming from Corollary 4.3 in standard cases. To simplify the problem, let us focus on the BFDR threshold and consider a ζ-Subbotin location model with $\zeta > 1$. Taking $\tau_m = m^\beta$, the parameters of the model are $(\beta, C_m) \in (0, 1] \times (0, 1)$. Assume that the parameter sequence $(C_m)_m$ satisfies (BP) for some $0 < C_- \leq C_+ < 1$. Then Corollary S-2.4 in the supplemental article provides explicit constants $D = D(\beta, C_-, C_+, \beta_0, C_0, \nu)$ and $M = M(\beta, C_-, C_+, \beta_0, C_0, \nu)$ such that the following inequality holds*

$$\left(R_m(t_m^\star) - R_m(t_m^B)\right) / R_m(t_m^B) \leq D/(\log m)^{1-1/\zeta}, \text{ for any } m \geq M. \tag{30}$$

As an illustration, in the Gaussian case ($\zeta = 2$), for $\beta = 0.7$, $C_- = 0.5$, $C_+ = 0.7$, $\beta_0 = C_0 = 0.5$ and $\nu = 0.25$, we have $M \simeq 61.6$ and $D \simeq 2.66$. As expected, these constants are over-estimated: for instance, by taking $m = 1000$, the LHS of (30) is smaller than 0.1 (see Figure 4 in the next section) while the RHS of (30) is $D/\sqrt{\log(1000)} \simeq 1.01$. Finally, we can check that D becomes large when β is close to 0 or C_+ is close to 1. These configurations correspond to the cases where the data are almost non-sparse and where the

Bayes rule can have almost full power, respectively. They can be seen as limit cases for our methodology.

REMARK 4.7. *By using Proposition 4.1, as* $m \to +\infty$, $\alpha_m^{opt}(\beta_0, C_0) \sim \alpha_m^{\infty}(\beta_0, C_0)$, *for an equivalent* $\alpha_m^{\infty}(\beta_0, C_0)$ *having a very simple form, see Section S-3.1 in the supplemental article. Therefore, we could use* $\alpha_m^{\infty}(\beta_0, C_0)$ *instead of* $\alpha_m^{opt}(\beta_0, C_0)$. *Numerical comparisons between the (B)FDR risk obtained according to* $\alpha_m^{opt}(\beta_0, C_0)$ *and* $\alpha_m^{\infty}(\beta_0, C_0)$ *are provided in Section S-7 in the supplemental article. While* $\alpha_m^{\infty}(\beta_0, C_0)$ *qualitatively leads to the same results when* m *is large (say,* $m \geq 1,000$*), the use of* $\alpha_m^{opt}(\beta_0, C_0)$ *is more accurate for a small* m.

5. Numerical experiments. In order to complement the convergence results stated above, it is of interest to study the behavior of FDR and BFDR thresholding for a small or moderate m in numerical experiments. These experiments have been performed for the inductive risk $R_m(\cdot) = R_m^I(\cdot)$ defined by (8).

5.1. *Exact formula for the FDR risk.* The BFDR threshold $t_m^\star$ can be approximated numerically, which allows us to compute $R_m(t_m^\star)$. Computing $R_m(\hat{t}_m^{FDR})$ is more complicated, because the FDR threshold $\hat{t}_m^{FDR}$ is not deterministic. However, we can avoid performing cumbersome and somewhat imprecise simulations to compute $R_m(\hat{t}_m^{FDR})$, by using the approach proposed in Finner and Roters (2002) and Roquain and Villers (2011). Using this methodology, the full distribution of $\hat{t}_m^{FDR}$ may be written as a function of the joint c.d.f. of the order statistics of i.i.d. uniform variables. Let for any $k \geq 0$ and for any $(t_1, ..., t_k) \in [0,1]^k$, $\Psi_k(t_1, ..., t_k) = \mathbb{P}(U_{(1)} \leq t_1, ..., U_{(k)} \leq t_k)$, where $(U_i)_{1 \leq i \leq k}$ is a sequence of i.i.d. uniform variables on $(0,1)$ and with the convention $\Psi_0(\cdot) = 1$. The Ψ_k's can be evaluated e.g. by using Steck's recursion (see (Shorack and Wellner, 1986), pages 366-369). Then, relation (10) in Roquain and Villers (2011) entails

$$R_m(\hat{t}_m^{FDR}) = \sum_{k=0}^{m} \binom{m}{k} R_m\left(\frac{\alpha(k \vee 1)}{m}\right) G_m(\alpha k/m)^k \times \Psi_{m-k}\big(1 - G_m(\alpha m/m), ..., 1 - G_m(\alpha(k+1)/m)\big), \tag{31}$$

where $G_m(t) = \pi_{0,m} t + \pi_{1,m} F_m(t)$. For reasonably large m ($m \leq 10,000$ in what follows), expression (31) can be used for computing the *exact* risk of FDR thresholding $\hat{t}_m^{FDR}$ in our experiment.

5.2. *Adaptation to unknown sparsity.* We quantify the quality of a thresholding procedure using the relative excess risk

$$\mathcal{E}_m(\hat{t}_m) = (R_m(\hat{t}_m) - R_m(t_m^B))/R_m(t_m^B).$$

The closer the relative excess risk $\mathcal{E}_m(\hat{t}_m)$ is to 0, the better the corresponding classification procedure is.

Figure 4 compares relative excess risks of different procedures in the Gaussian location model (results for the Gaussian scale and the Laplace scale models are qualitatively similar, see Figures S-1 and S-2 in the supplemental article). Each row of plots corresponds to a particular procedure, and each column to a particular value of $m \in \{25, 10^2, 10^3, 10^4, 10^5, 10^6\}$. The first row corresponds to the Bayes procedure defined by (9), where the model parameters are taken as $(\beta, C_m) = (\beta_0, C_0)$. It is denoted by *Bayes0*. Next, we consider BFDR (rows 2 to 5) and FDR (rows 6 to 9) thresholding at level α_m, for $\alpha_m \in \{0.1, 0.2, 0.25\}$ (independent of m) and for the choice $\alpha_m = \alpha_m^{opt}(\beta_0, C_0)$ defined in Section 4.4. For each procedure and each value of m, the behavior of the relative excess risk is studied as the (unknown) true model parameters (β, C_m) vary in $[0,1] \times [0,1]$, and we arbitrarily choose β_0 and C_0 as the midpoints of the corresponding intervals, i.e. $(\beta_0, C_0) = (1/2, 1/2)$ (similar results are obtained for other values of (β_0, C_0), see Figures S-4, S-5, and S-6 in the supplemental article). Colors reflect the value of the relative excess risk. They range from white ($R_m = R_m(t_m^B)$) to dark red ($R_m \geq 2R_m(t_m^B)$). Black lines represent the level set $\mathcal{E}_m = 0.1$, that is, they delineate a region of the (β, C_m) plane in which the excess risk of the procedure under study is ten times less than the Bayes risk. The number at the bottom left of each plot gives the fraction of configurations (β, C_m) for which $\mathcal{E}_m \leq 0.1$. This evaluates the quality of a procedure uniformly across all the (β, C_m) values.

For $m = 10^6$, we did not undertake exact FDR risk calculations: they were too computationally intensive, as the complexity of the calculation of function Ψ_k used in (31) is quadratic in m. However, FDR risk is expected to be well approximated by BFDR risk for such a large value of m, as confirmed by the fact that FDR and BFDR plots at a given level α are increasingly similar as m increases.

Bayes0 performs well when the sparsity parameter β is correctly specified, and its performance is fairly robust to C_m. However, it performs poorly when β is misspecified, and increasingly so as m increases. The results are markedly different for the other thresholding methods. BFDR thresholding and FDR thresholding are less adaptive to C_m than Bayes0, but much more adaptive to the sparsity parameter β, as illustrated by the fact that the

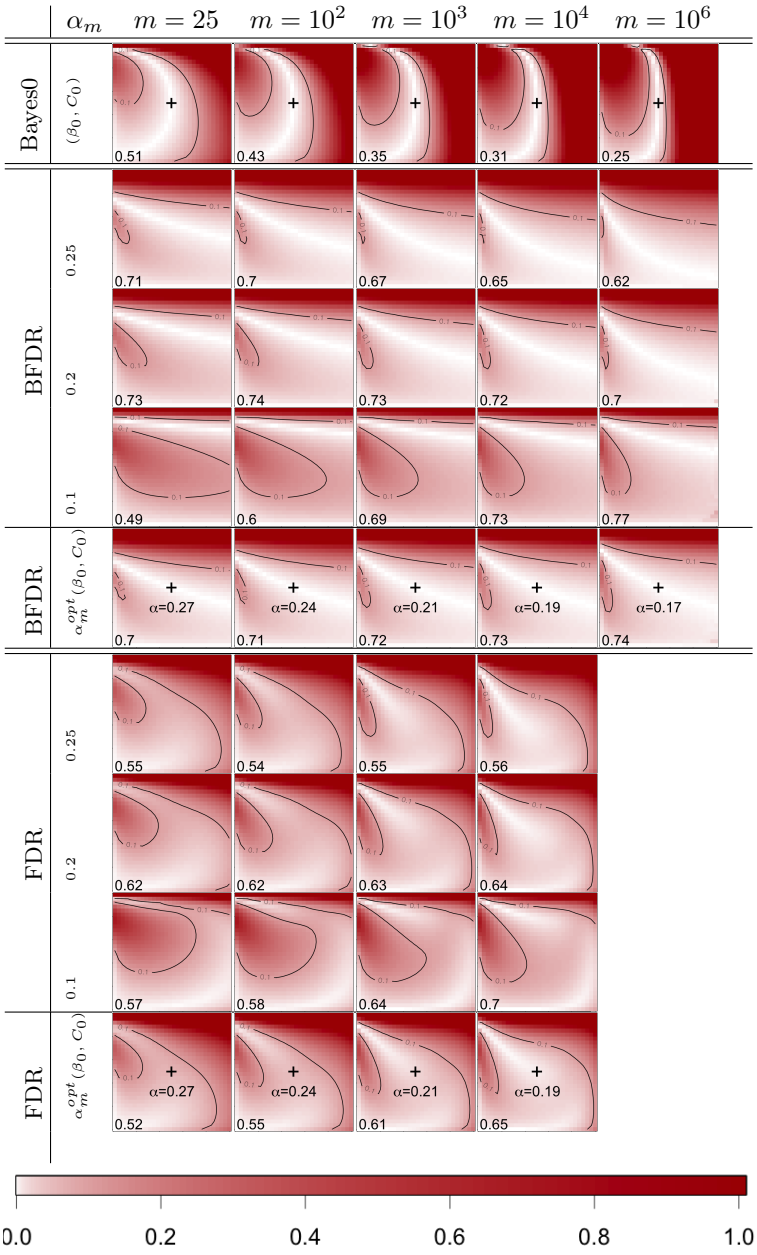

FIG 4. *Adaptation to sparsity by (B)FDR thresholding in the Gaussian location model. relative excess risks $\mathcal{E}_m$ for various thresholding procedures (rows) and different values of m (columns). In each panel, the corresponding risk is plotted as a function of $\beta \in [0,1]$ (horizontal axis) and $C_m \in [0,1]$ (vertical axis). Colors range from white (low risk) to dark red (high risk), as indicated by the color bar at the bottom. Black lines represent the level set $\mathcal{E}_m = 0.1$. The point $(\beta, C_m) = (\beta_0, C_0)$ is marked by "+". We chose $\beta_0 = 1/2$ and $C_0 = 1/2$. See main text for details.*

configurations with low relative excess risk span the whole range of β.

For $\alpha_m = \alpha_m^{opt}(\beta_0, C_0)$, the fraction of configurations (β, C_m) for which $\mathcal{E}_m \leq 0.1$ increases as m increases. This illustrates the asymptotic optimality of (B)FDR thresholding, as stated in Corollary 4.4 (iii), because $\alpha_m^{opt}(\beta_0, C_0) \propto (\log m)^{-1/2}$. Additionally, observe that the (β, C_m)-region around (β_0, C_0) contains only very small values of $\mathcal{E}_m$, even for moderate m. This suggests that, non-asymptotically, $\alpha_m^{opt}(\beta_0, C_0)$ is a reasonable choice for α_m, when we know *a priori* that the parameters lie in some specific region of the (β, C_m)-square.

Next, let us consider the case of (B)FDR thresholding using a fixed value of $\alpha_m = \alpha$. While our theoretical results show that choosing α_m fixed with m (and in particular not tending to zero) is always asymptotically sub-optimal, the results shown by Figure 4 are less clear-cut. An explanation is that $(\log m)^{-1/2}$ decreases only slowly to zero (e.g., $\alpha_m^{opt}(\beta_0, C_0) \simeq 0.17$ for $m = 10^6$), hence the asymptotic is quite "far" and not fully attained by our experiments.

Hence, from a more practical point of view, in a classical situation where m does not exceed, say, 10^6, a practitioner willing to use the (B)FDR can consider two different approaches to calibrate α_m: the first one is to take some arbitrary value, e.g., 0.05, 0.1, or 0.2. The overall excess risk might be small, but the location of the region of smallest excess risk (pictured in white in our figures) is unknown, and depends strongly on α and m (and even ζ). In contrast, the second method $\alpha_m^{opt}(\beta_0, C_0)$ "stabilizes" the region of the (β, C_m)-square where the (B)FDR has good performance across all the values of m (and ζ). Thus, while the first method has a clear interpretation in terms of FDR, the second approach is more interpretable w.r.t. the sparsity and power parameters and is recommended when these parameters are felt to correctly parametrize the model.

Finally note that, when considering the weighted mis-classification risk (as formally defined in (33) and studied in Section S-4 in the supplemental article), there exists a particular choice of the weight (as a function of m) such that the optimal (B)FDR level $\alpha_m^{opt}(\beta_0, C_0)$ does not depend on m, making (B)FDR thresholding with fixed values of α_m asymptotically optimal, as noted by Bogdan et al. (2011). This point is discussed in Section 6.2.

6. Discussion.

6.1. *Asymptotic minimaxity over a sparsity class.* Let us consider the sparsity range $\tau_m = m^\beta$, with $\beta_- \leq \beta \leq 1$, for some given $\beta_- \in (0, 1)$. Assume (BP) with C_- and C_+ defined therein. Denote the set $[\beta_-, 1] \times$

$[C_-, C_+]$ by Θ for short. The minimax risk is defined by

$$R_m^\star = \inf_{\hat{t}_m} \left\{ \sup_{(\beta, C_m)\in\Theta} \{R_m(\hat{t}_m)\} \right\},$$

where the infimum is taken over the set of thresholds that can be written as measurable functions of the p-values. Obviously, $R_m^\star \geq \sup_{(\beta,C_m)\in\Theta}\{R_m(t_m^B)\}$, where t_m^B is the Bayes threshold. Hence, by taking the supremum w.r.t. (β, C_m) in our excess risk inequalities, we are able to derive minimax results. However, this requires a precise formulation of (5) where the dependence in β of the constant D is explicit. For simplicity, let us consider the Laplace scale model. By using (S-24) and (S-29) in the supplemental article, and by taking $\alpha_m \propto (\log m)^{-1}$, we can derive that there exists a constant $D' > 0$ (independent of β_-, C_- and C_+) such that for a large m,

$$\begin{aligned} \sup_{(\beta,C_m)\in\Theta} \{R_m(\hat{t}_m^{FDR})\} &\leq \sup_{(\beta,C_m)\in\Theta} \{R_m(t_m^B)\} \left(1 + \frac{-\log(\beta_-/2)}{\beta_-(1-C_+)} \frac{D'}{\log m}\right) \\ &\leq R_m^\star \left(1 + \frac{-\log(\beta_-/2)}{\beta_-(1-C_+)} \frac{D'}{\log m}\right). \end{aligned} \tag{32}$$

This entails that $\hat{t}_m^{FDR}$ is asymptotically minimax, that is,

$$\sup_{(\beta,C_m)\in\Theta} \{R_m(\hat{t}_m^{FDR})\} \sim R_m^\star.$$

This property can be seen as an analogue to the asymptotically minimaxity stated in Theorem 1.1 in Abramovich et al. (2006) and Theorem 1.3 in Donoho and Jin (2006), in an estimation context.

Finally, regarding (32), an interesting avenue for future research would be to establish whether there are asymptotically minimax rules $\hat{t}_m$ such that $\sup_{(\beta,C_m)\in\Theta} \{R_m(\hat{t}_m)\} = R_m^\star(1 + o(\rho_m))$ for a rate ρ_m smaller than $(\log m)^{-1}$.

6.2. *Extension to weighted mis-classification risk.* In our sparse setting, where we assume that there are many more labels "0" than labels "1", one could consider that mis-classifying a "0" is less important than mis-classifying a "1". This suggests to consider the following weighted risk:

$$R_{m,\lambda_m}(\hat{t}_m) = \mathbb{E}(\pi_{0,m}\hat{t}_m + \lambda_m \pi_{1,m}(1 - F_m(\hat{t}_m))), \tag{33}$$

for a known factor $\lambda_m \in (1, \tau_m)$. This weighted risk was extensively used in Bogdan et al. (2011). In Section S-4 in the supplemental article, we show

that all our results can be adapted to this risk. Essentially, when considering R_{m,λ_m} instead of R_m, our results hold after replacing τ_m by τ_m/λ_m and q_m by $q_m\lambda_m$.

As an illustration, let us consider here the case of a ζ-Subbotin density, $\tau_m = m^\beta$, $\beta \in (0,1]$, $\log\lambda_m = o((\log m)^\gamma)$, where $\gamma = 1-\zeta^{-1}$ and $\gamma = 1$ for the location and scale cases, respectively. As displayed in Table S-4 in the supplemental article, under the (corresponding) assumptions (BP) and (Sp), we show that a sufficient condition for FDR thresholding to be asymptotically optimal for the risk R_{m,λ_m} is to take $q_m^{-1} = O(1)$, $q_m\lambda_m \to \infty$ and $\log q_m = o\left((\log m)^\gamma\right)$. This recovers Theorem 5.3 of Bogdan et al. (2011) when applied to the particular case of a Gaussian scale model (for which $\gamma = 1$). Furthermore, we show that taking $q_m \propto q_m^{opt}$, that is $q_m \propto \lambda_m^{-1}(\log m)^\gamma$, leads to the optimality rate $\rho_m = (\log m)^{-\gamma}$ for the relative excess risk based on R_{m,λ_m}. While the order of q_m^{opt} is not modified when $\lambda_m \propto 1$, it may be substantially different when $\lambda_m \to \infty$. Typically, $\lambda_m \propto (\log m)^\gamma$ leads to $q_m^{opt} \propto 1$. Hence, when considering R_{m,λ_m} instead of R_m, the value of λ_m should be carefully taken into account when choosing α_m to obtain a small excess risk.

Conversely, our result states that FDR thresholding with a pre-specified value of $\alpha_m = \alpha$, (say, $\alpha = 0.05$), is optimal over the range of weighted misclassification risks using a λ_m satisfying $\lambda_m \to \infty$ and $\log\lambda_m = o\left((\log m)^\gamma\right)$, and that choosing $\lambda_m \propto (\log m)^\gamma$ leads to the optimality rate $\rho_m = (\log m)^{-\gamma}$.

7. Proofs of Theorem 3.1 and Theorem 3.2. The proofs are first established for the misclassification risk $R_m = R_m^I$ defined by (8). The case of the misclassification risk R_m^T, defined by (7) is examined in Section 7.4.

7.1. *Relations for BFDR.* Let us first state the following result.

PROPOSITION 7.1. *Consider the setting and the notation of Theorem 3.1. Then we have for any $m \geq 2$,*

$$R_m(t_m^\star) - R_m(t_m^B) = \pi_{1,m}C_m/q_m - \pi_{0,m}t_m^B + \pi_{1,m}(1-q_m^{-1})(C_m - F_m(t_m^\star)). \tag{34}$$

Furthermore, if $\alpha_m \leq 1/2$, we have for any $m \geq 2$,

$$R_m(t_m^\star) - R_m(t_m^B)) \leq \pi_{1,m}C_m/q_m - \pi_{0,m}t_m^B + \pi_{1,m}(1-q_m^{-1})\gamma_m; \tag{35}$$

$$R_m(t_m^\star) - R_m(t_m^B) \leq \pi_{1,m}(C_m/q_m - \tau_m t_m^B) \vee \gamma_m. \tag{36}$$

PROOF. To prove (34), we use $F_m(t_m^\star) = t_m^\star q_m \tau_m$ and $\tau_m = \pi_{0,m}/\pi_{1,m}$, to write

$$\begin{aligned} &R_m(t_m^\star) - R_m(t_m^B) \\ &= \pi_{0,m} t_m^\star - \pi_{0,m} t_m^B + \pi_{1,m}(C_m - F_m(t_m^\star)) \\ &= \pi_{1,m} F_m(t_m^\star)/q_m - \pi_{0,m} t_m^B + \pi_{1,m}(C_m - F_m(t_m^\star)). \end{aligned} \tag{37}$$

Expression (35) is an easy consequence of (34). Finally, (37) and (34) entail

$$R_m(t_m^\star) - R_m(t_m^B) \le \begin{cases} \pi_{1,m} C_m/q_m - \pi_{0,m} t_m^B & \text{if } t_m^B \le t_m^\star \\ \pi_{1,m}(C_m - F_m(\Psi_m^{-1}(q_m\tau_m))) & \text{if } t_m^B \ge t_m^\star \end{cases},$$

which yields (36).

□

7.2. *Proof of Theorem 3.1.* Theorem 3.1 (i) follows from (36) because $\pi_{0,m} t_m^B = \pi_{1,m} C_m/q_m^{opt}$ by definition. Let us now prove (ii). First note that

$$R_m(t_m^\star) = \pi_{1,m} - \pi_{1,m} F_m(t_m^\star)(1 - q_m^{-1}). \tag{38}$$

Using (38) and the upper bound $t_m^\star = F_m(t_m^\star)(q_m\tau_m)^{-1} \le (q_m\tau_m)^{-1}$, we obtain $R_m(t_m^\star) \ge \pi_{1,m}(1-(1-q_m^{-1})_+ F_m(t_m^\star)) \ge \pi_{1,m}(1-(1-q_m^{-1})_+ F_m(q_m^{-1}\tau_m^{-1}))$. This entails (17) and (18).

7.3. *Proof of Theorem 3.2.* Write $\hat{t}_m$ instead of $\hat{t}_m^{FDR}$ for short. To establish (19), let us first write the risk of FDR thresholding as $R_m(\hat{t}_m) = T_{1,m} + T_{2,m}$, with $T_{1,m} = \pi_{0,m}\mathbb{E}(\hat{t}_m)$ and $T_{2,m} = \pi_{1,m}(1 - \mathbb{E}(F_m(\hat{t}_m)))$. In the sequel, $T_{1,m}$ and $T_{2,m}$ are examined separately.

7.3.1. *Bounding $T_{1,m}$.* The next result is a variation of Lemma 7.1 and Lemma 7.2 in Bogdan et al. (2011).

PROPOSITION 7.2. *The following bound holds:*

$$T_{1,m} \le \pi_{1,m} \frac{\alpha_m}{1-\alpha_m} + m^{-1} \frac{\alpha_m}{(1-\alpha_m)^2}. \tag{39}$$

PROOF. To prove Proposition 7.2, we follow the proof of Lemma 7.1 in Bogdan et al. (2011) with slight simplifications. Recall that we have by definition $\hat{t}_m = \hat{t}_m^{BH} \vee (\alpha_m/m)$. Hence, we have $\mathbb{E}(\hat{t}_m|H) \le \alpha_m/m + \mathbb{E}(\hat{t}_m^{BH}|H)$. By integrating w.r.t. the label vector H, it is thus sufficient to prove

$$E(\hat{t}_m^{BH}|H) \le \pi_{1,m} \frac{\alpha_m}{1-\alpha_m} + m^{-1} \frac{\alpha_m^2}{(1-\alpha_m)^2}. \tag{40}$$

Let $m_1(H) = \sum_{i=1}^m H_i$ and $m_0(H) = m - m_1(H)$. By exchangeability of $(p_i, H_i)_i$, we can assume without loss of generality that the p-values corresponding to a label $H_i = 0$ are $p_1, ..., p_{m_0(H)}$ for simplicity. Let us denote $\hat{t}_{m,0}$ the thresholding $\hat{t}_m^{BH}$ defined by (14), applied to the p-value family $p_i, 1 \leq i \leq m$, in which each of the p-value $p_{m_0(H)+1}, ..., p_m$ has been replaced by 0. Classically, we have

$$\hat{t}_{m,0} = \alpha_m(m_1(H) + \hat{k}_{m,0})/m,$$

where $\hat{k}_{m,0} = \max\{k \in \{0, 1, ..., m_0(H)\} \ : \ q_{(k)} \leq \alpha_m(m_1(H) + k)/m\}$, where $(q_1, \ldots, q_{m_0(H)}) = (p_1, ..., p_{m_0(H)})$ is the set of p-values corresponding to zero labels (see, e.g., Lemma 7.1 in Roquain and Villers (2011)). Since $\hat{t}_m^{BH}$ is non-increasing in each p-value, setting some p-values equal to 0 can only increase $\hat{t}_m^{BH}$. This entails

$$\mathbb{E}(\hat{t}_m^{BH} \mid H) \leq \mathbb{E}(\hat{t}_{m,0} \mid H) = \alpha_m(m_1(H) + \mathbb{E}(\hat{k}_{m,0} \mid H))/m. \tag{41}$$

Next, we use Lemma 4.2 in Finner and Roters (2002) (by taking "$n = m_0(H)$, $\beta = \alpha_m$, $\tau = \alpha_m/m$" with their notation), to derive that for any $H \in \{0,1\}^m$,

$$\begin{aligned}
\mathbb{E}(\hat{k}_{m,0} \mid H) &= \alpha_m \frac{m_0(H)}{m} \sum_{i=0}^{m_0(H)-1} \binom{m_0(H)-1}{i} (m_1(H) + i + 1) i! \left(\frac{\alpha_m}{m}\right)^i \\
&\leq \alpha_m \sum_{i \geq 0} (m_1(H) + i + 1)\alpha_m^i \\
&= \alpha_m(m_1(H)/(1-\alpha_m) + 1/(1-\alpha_m)^2). \qquad (42)
\end{aligned}$$

The bound (40) thus follows from (41).

□

7.3.2. *Bounding $T_{2,m}$.* Let us consider t_m^ε the BFDR threshold associated to level $\alpha_m \pi_{0,m}(1-\varepsilon)$. Note that by definition of t_m^ε we have $\pi_{0,m}(1-\varepsilon)G_m(t_m^\varepsilon) = t_m^\varepsilon/\alpha_m$. Here, we state the following inequalities, which, combined with Proposition 7.2 establishes Theorem 3.2:

$$T_{2,m} \leq \pi_{1,m}(1 - F_m(\alpha_m/m)); \tag{43}$$

$$T_{2,m} \leq \pi_{1,m}(1 - F_m(t_m^\varepsilon)) + \pi_{1,m} \exp\{-m(\tau_m + 1)^{-1}(C_m - \gamma_m^\varepsilon)\varepsilon^2/4\}. \tag{44}$$

First, (43) is an easy consequence of $\hat{t}_m \geq \alpha_m/m$. Second, expression (44) derives from (45) of Lemma 7.3 because

$$\begin{aligned}
\mathbb{E}(1 - F_m(\hat{t}_m)) &= \mathbb{E}((1 - F_m(\hat{t}_m))\mathbf{1}\{\hat{t}_m < t_m^\varepsilon\}) + \mathbb{E}((1 - F_m(\hat{t}_m))\mathbf{1}\{\hat{t}_m \geq t_m^\varepsilon\}) \\
&\leq \mathbb{P}(\hat{t}_m^{BH} < t_m^\varepsilon) + 1 - F_m(t_m^\varepsilon),
\end{aligned}$$

(by using $\hat{t}_m \geq \hat{t}_m^{BH}$) and because $G_m(t_m^\varepsilon) \geq \pi_{1,m} F_m(t_m^\varepsilon) \geq (\tau_m+1)^{-1}(C_m - \gamma_m^\varepsilon)$.

LEMMA 7.3. *The following bound holds:*

$$\mathbb{P}(\hat{t}_m^{BH} < t_m^\varepsilon) \leq \exp\{-mG_m(t_m^\varepsilon)\varepsilon^2/4\}. \tag{45}$$

We prove Lemma 7.3 by using a variation of the method described in the proof of Theorem 1 in Genovese and Wasserman (2002) (we use Bennett's inequality instead of Hoeffding's inequality). For any $t_0 \in (0,1)$ such that $t_0/\alpha_m - G_m(t_0) < 0$, we have $\mathbb{P}(\hat{t}_m^{BH} < t_0) \leq \mathbb{P}(\widehat{\mathbb{G}}_m(t_0) < t_0/\alpha_m) \leq \mathbb{P}(\widehat{\mathbb{G}}_m(t_0) - G_m(t_0) < t_0/\alpha_m - G_m(t_0))$. Next, by using Bennett's inequality (see, e.g., Proposition 2.8 in Massart (2007)) and by letting $h(u) = (1+u)\log(1+u) - u$, for any $u > 0$, we obtain

$$\mathbb{P}(\hat{t}_m^{BH} < t_0) \leq \exp\left\{-mG_m(t_0)h\left(\frac{G_m(t_0) - t_0/\alpha_m}{G_m(t_0)}\right)\right\}.$$

Finally, for $t_0 = t_m^\varepsilon$, since we have $G_m(t_m^\varepsilon) - t_m^\varepsilon/\alpha_m = (1 - \pi_{0,m}(1-\varepsilon))G_m(t_m^\varepsilon) \geq \varepsilon G_m(t_m^\varepsilon)$, we obtain (45) by using that $h(u) \geq u^2/4$ for any $u > 0$.

7.4. *Proofs for the risk R_m^T.* Let us recall that R_m^T and R_m^I are equal for a deterministic threshold and thus also for the BFDR threshold. Hence, Theorem 3.1 also holds for the risk R_m^T and we only have to prove Theorem 3.2.

First note that since $R_m^T(\hat{t}_m^{FDR}) = R_m^T(\hat{t}_m^{BH})$, we can work directly with $\hat{t}_m^{BH}$. Proving the type I error bound (39) can be done similarly: with the same notation, the type I error can be written conditionally on H as

$$\begin{aligned}\mathbb{E}\left(m^{-1}\sum_{i=1}^{m_0(H)} \mathbf{1}\{p_i \leq \hat{t}_m^{BH}\} \,\middle|\, H\right) &\leq \mathbb{E}\left(m^{-1}\sum_{i=1}^{m_0(H)} \mathbf{1}\{p_i \leq \hat{t}_{m,0}\} \,\middle|\, H\right) \\ &= m^{-1}\mathbb{E}(\hat{k}_{m,0} \mid H) \\ &\leq \pi_{1,m}\frac{\alpha_m}{1-\alpha_m} + m^{-1}\frac{\alpha_m}{(1-\alpha_m)^2},\end{aligned}$$

by using (42). Hence, (39) is proved for the risk R_m^T.

Next, the proof for bounding the type II error derives essentially from the following argument, which is quite standard in the multiple testing methodology, see e.g. Ferreira and Zwinderman (2006); Finner et al. (2009); Roquain and Villers (2011); Roquain (2011). Let us denote

$$\tilde{t}_m = \max\{t \in [0,1] \ : \ \alpha_m \widetilde{\mathbb{G}}_m(t) \geq t\},$$

where $\widetilde{\mathbb{G}}_m(t) = m^{-1}(1 + \sum_{i=2}^m \mathbf{1}\{p_i \leq t\})$ denotes the empirical c.d.f. of the p-values where p_1 has been replaced by 0. Then, for any realization of the p-value family, $p_1 \leq \hat{t}_m^{BH}$ is equivalent to $p_1 \leq \tilde{t}_m$ (see, e.g., Proof of Theorem 2.1 in Ferreira and Zwinderman (2006) and Section 3.2 of Roquain (2011)). This entails that the type II error is equal to $\pi_{1,m}(1 - \mathbb{E}(F_m(\tilde{t}_m)))$ (by using the exchangeability of $(H_i, p_i)_{1 \leq i \leq m}$). Finally, since $\tilde{t}_m \geq \hat{t}_m^{BH}$ and $\tilde{t}_m \geq \alpha_m/m$, we have $\tilde{t}_m \geq \hat{t}_m^{FDR}$. Hence $\pi_{1,m}(1 - \mathbb{E}(F_m(\tilde{t}_m))) \leq \pi_{1,m}(1 - \mathbb{E}(F_m(\hat{t}_m^{FDR})))$ and the bounds (43) and (44) also hold for the risk R_m^T.

Acknowledgements

We would like to thank Guillaume Lecué and Nicolas Verzelen for interesting discussions. We are also grateful to anonymous referees, an associated editor, and an editor for their very helpful comments and suggestions. The second author was supported by the French Agence Nationale de la Recherche (ANR grant references: ANR-09-JCJC-0027-01, ANR-PARCIMONIE, ANR-09-JCJC-0101-01) and by the French ministry of foreign and european affairs (EGIDE - PROCOPE project number 21887 NJ).

SUPPLEMENTARY MATERIAL

Supplement to: on false discovery rate thresholding for classification under sparsity

(). Proofs, additional experiments and supplementary notes for the present paper can be bound in Neuvial and Roquain (2011) on AoS website.

References.

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(2):584–653.

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(1):289–300.

Benjamini, Y., Krieger, A. M., and Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3):491–507.

Blanchard, G., Lee, G., and Scott, C. (2010). Semi-supervised novelty detection. *J. Mach. Learn. Res.*, 11:2973–3009.

Blanchard, G. and Roquain, E. (2009). Adaptive false discovery rate control under independence and dependence. *J. Mach. Learn. Res.*, 10:2837–2871.

Bogdan, M., Chakrabarti, A., Frommlet, F., and Ghosh, J. K. (2011). Asymptotic bayes-optimality under sparsity of some multiple testing procedures. *Ann. Statist.*, 39(3):1551–1579.

Bogdan, M., Ghosh, J. K., and Tokdar, S. T. (2008). A comparison of the Benjamini-Hochberg procedure with some Bayesian rules for multiple testing. In *Beyond parametrics in interdisciplinary research: Festschrift in honor of Professor Pranab K. Sen*, volume 1 of *Inst. Math. Stat. Collect.*, pages 211–230. Inst. Math. Statist., Beachwood, OH.

Chi, Z. (2007). On the performance of FDR control: constraints and a partial solution. *Ann. Statist.*, 35(4):1409–1431.

Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.*, 32(3):962–994.

Donoho, D. and Jin, J. (2006). Asymptotic minimaxity of false discovery rate thresholding for sparse exponential data. *Ann. Statist.*, 34(6):2980–3018.

Efron, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.*, 23(1):1–22.

Efron, B. and Tibshirani, R. (2002). Empirical bayes methods and false discovery rates for microarrays. *Genet. Epidemiol.*, 23(1):70–86.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.*, 96(456):1151–1160.

Ferreira, J. A. and Zwinderman, A. H. (2006). On the Benjamini-Hochberg method. *Ann. Statist.*, 34(4):1827–1849.

Finner, H., Dickhaus, T., and Roters, M. (2009). On the false discovery rate and an asymptotically optimal rejection curve. *Ann. Statist.*, 37(2):596–618.

Finner, H. and Roters, M. (2002). Multiple hypotheses testing and expected number of type I errors. *Ann. Statist.*, 30(1):220–238.

Gavrilov, Y., Benjamini, Y., and Sarkar, S. K. (2009). An adaptive step-down procedure with proven FDR control under independence. *Ann. Statist.*, 37(2):619–629.

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.

Genovese, C. and Wasserman, L. (2004). A stochastic process approach to false discovery control. *Ann. Statist.*, 32(3):1035–1061.

Massart, P. (2007). *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

Neuvial, P. and Roquain, E. (2011). Supplement to: on false discovery rate thresholding for classification under sparsity. Submitted.

Roquain, E. (2011). Type I error rate control for testing many hypotheses: a survey with proofs. *J. Soc. Fr. Stat.*, 152(2):3–38.

Roquain, E. and Villers, F. (2011). Exact calculations for false discovery proportion with application to least favorable configurations. *Ann. Statist.*, 39(1):584–612.

Sarkar, S. K. (2002). Some results on false discovery rate in stepwise multiple testing procedures. *Ann. Statist.*, 30(1):239–257.

Sarkar, S. K. (2008). On methods controlling the false discovery rate. *Sankhya, Ser. A*, 70:135–168.

Sarkar, S. K., Zhou, T., and Ghosh, D. (2008). A general decision theoretic formulation of procedures controlling FDR and FNR from a Bayesian perspective. *Statist. Sinica*, 18(3):925–945.

Sawyers, C. L. (2008). The cancer biomarker problem. *Nature*, 452(7187):548–552.

Seeger, P. (1968). A note on a method for the analysis of significances en masse. *Technometrics*, 10(3):586–593.

Sen, P. K. (1999). Some remarks on Simes-type multiple tests of significance. *J. Statist. Plann. Inference*, 82(1-2):139–145. Multiple comparisons (Tel Aviv, 1996).

Shorack, G. R. and Wellner, J. A. (1986). *Empirical processes with applications to statistics*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York.

Storey, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 64(3):479–498.

Storey, J. D. (2003). The positive false discovery rate: a Bayesian interpretation and the

q-value. *Ann. Statist.*, 31(6):2013–2035.
Tamhane, A. C., Liu, W., and Dunnett, C. W. (1998). A generalized step-up-down multiple test procedure. *Canad. J. Statist.*, 26(2):353–363.

Laboratoire Statistique et Génome
Université d'Évry Val d'Essonne
UMR CNRS 8071 – USC INRA
23 boulevard de France, 91 037 Évry, France
E-mail: pierre.neuvial@genopole.cnrs.fr

UPMC Univ Paris 6, LPMA,
4, Place Jussieu, 75252 Paris cedex 05, France
E-mail: etienne.roquain@upmc.fr