*The Annals of Statistics*
2022, Vol. 50, No. 2, 1095–1123
https://doi.org/10.1214/21-AOS2141


# FALSE DISCOVERY RATE CONTROL WITH UNKNOWN NULL DISTRIBUTION: IS IT POSSIBLE TO MIMIC THE ORACLE?

BY ETIENNE ROQUAIN[1,a] AND NICOLAS VERZELEN[2,b]

[1] *Université de Paris and Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation,* [a] *etienne.roquain@sorbonne-universite.fr*

[2] *INRAE, L'institut Agro, Université de Montpellier, MISTEA,* [b] *nicolas.verzelen@inrae.fr*

Classical multiple testing theory prescribes the null distribution, which is often too stringent an assumption for nowadays large scale experiments. This paper presents theoretical foundations to understand the limitations caused by ignoring the null distribution, and how it can be properly learned from the same data set, when possible. We explore this issue in the setting where the null distributions are Gaussian with unknown rescaling parameters (mean and variance) whereas the alternative distributions are let arbitrary. In that case, an oracle procedure is the Benjamini–Hochberg procedure applied with the true (unknown) null distribution and we aim at building a procedure that asymptotically mimics the performances of the oracle (AMO in short). Our main result establishes a phase transition at the sparsity boundary $n/\log(n)$: an AMO procedure exists if and only if the number of false nulls is of order less than $n/\log(n)$, where $n$ is the total number of tests. Further sparsity boundaries are derived for general location models where the shape of the null distribution is not necessarily Gaussian. In light of our impossibility results, we also pursue the less stringent aim of building a nonparametric confidence region for the null distribution. From a practical perspective, this provides goodness-of-fit tests for the null distribution and allows to assess the reliability of empirical null procedures via novel diagnostic graphs. Our results are illustrated on numerical experiments and real data sets, as detailed in a companion vignette (Roquain and Verzelen (2021)).

## 1. Introduction.

1.1. *Background*. In large-scale data analysis, the practitioner routinely faces the problem of simultaneously testing a large number $n$ of null hypotheses. In the last decades, an impressive amount of multiple testing procedures have been developed (see, e.g., Dickhaus (2014)). Theoretically-founded control of the amount of false rejections are provided notably by controlling the false discovery rate (FDR), that is, the average proportion of errors among the rejections, as done by the famous Benjamini–Hochberg procedure (BH), introduced by Benjamini and Hochberg (1995). However, most of the FDR controlling procedures developed in the multiple testing literature rely on the fact that the null distribution of the test statistics is known, either for finite $n$ or asymptotically. This is in plain contrast with common practice, where the null distribution is often *misspecified*:

- *The null distribution can be wrong*. This phenomenon, pointed out in a series of pioneering papers Efron (2004, 2007, 2008, 2009) and studied further in Schwartzman (2008, 2010), Azriel and Schwartzman (2015), Stephens (2017), Sun and Stephens (2018) is illustrated in Figure 1 for four classical datasets. As one can see, the theoretical null distribution

Received December 2020; revised July 2021.
*MSC2020 subject classifications.* Primary 62G10; secondary 62C20.
*Key words and phrases.* Benjamini–Hochberg procedure, false discovery rate, minimax, multiple testing, phase transition, sparsity, null distribution.

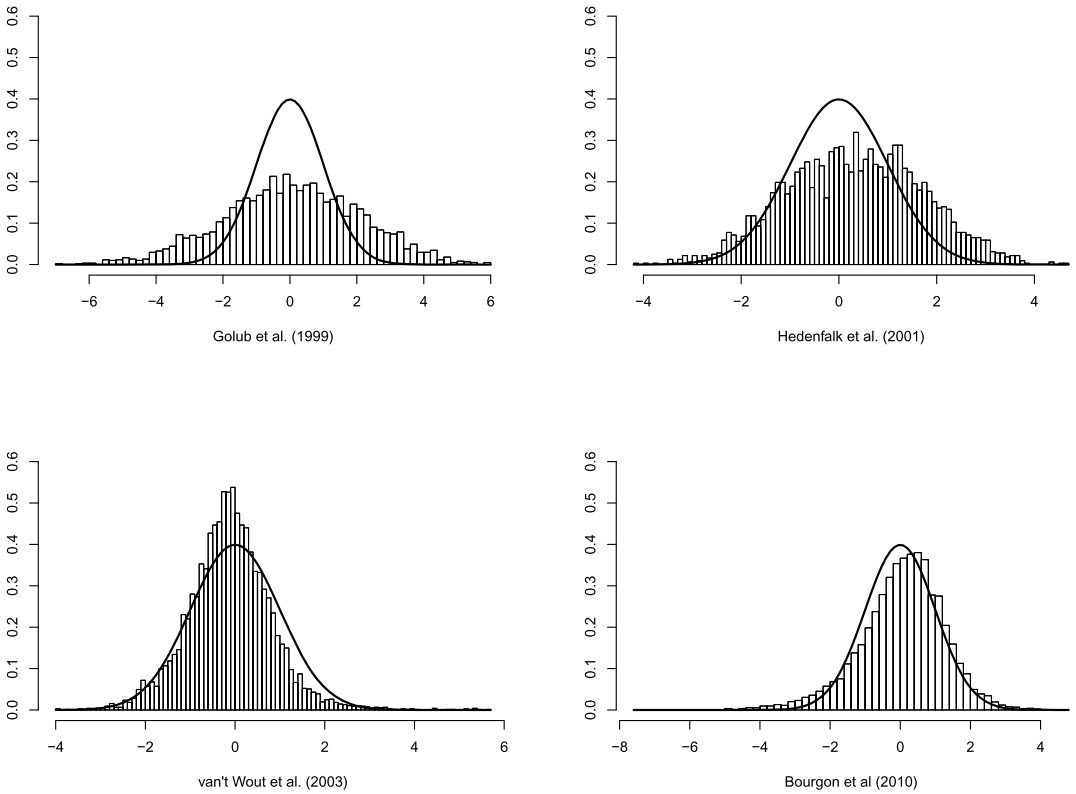


FIG. 1. *Histograms of the test statistics* (*rescaled to be all marginally standard Gaussian*), *for three data sets presented by Efron*: *Golub et al.* (1999) (*top left*); *Hedenfalk et al.* (2001) (*top right*); *van't Wout et al.* (2003) (*bottom left*); *and Bourgon, Gentleman and Huber* (2010) (*bottom right*). *The solid curve is the standard Gaussian density*. *Pictures reproducible from the vignette Roquain and Verzelen* (2021).

$\mathcal{N}(0, 1)$ does not faithfully describe the overall behavior of the measurements. As a result, using this theoretical null distribution into a standard multiple testing procedure (e.g., BH) can lead to an important resurgence of false discoveries. Markedly, this effect is sometimes more severe than simply ignoring the multiplicity of the tests (see Roquain and Verzelen (2021)), and thus the benefit of using a multiple testing correction can be lost.

- *The null distribution can be unknown*. Data often come from raw measurements that have been "cleaned" via many sophisticated normalization processes in which case the practitioner has no prior belief in the null distribution. Hence, the null distribution is implicitly defined as the "background noise" of the measurements and searching signal in the data boils down to make some assumption on this background (typically Gaussian) and find outliers, defined as items that significantly deviate from the background. This occurs for instance in astrophysics data sets (Miller et al. (2001), Sulis, Mary and Bigot (2017), Szalay, Connolly and Szokoly (1999)).

To address these issues, Efron popularized the concept of *empirical null distribution*, that is, of a null distribution estimated from the data, in the works Efron et al. (2001), Efron (2004, 2007, 2008, 2009) notably through the *two-group mixture model* and the *local FDR* method. Therein, an important message is that a significant improvement can already be obtained by replacing the theoretical null $\mathcal{N}(0, 1)$ by a Gaussian $\mathcal{N}(\theta, \sigma^2)$ with unspecified scaling parameters $\theta$ and $\sigma$. These works paved the way for many extensions (Cai and Jin (2010), Cai and Sun (2009), Cai, Sun and Wang (2019), Heller and Yekutieli (2014), Jin and Cai (2007), Muralidharan (2010), Nguyen and Matias (2014), Padilla and Bickel

(2012), Rebafka, Roquain and Villers (2019), Sun and Cai (2009)), which make this type of techniques widely used nowadays, mostly in genomics (Amar, Shamir and Yekutieli (2017), Consortium et al. (2007), Jiang and Yu (2016), Zablocki et al. (2014)) but also in other applied fields, such as neuroimaging; see, for example, Lee et al. (2016). However, when available, the theoretical FDR controlling properties often rely on stringent assumptions on the underlying mixture model (parameters fixed with $n$, specific alternatives and existence of suitable parameter estimators), which are not met in general.

1.2. *General aim.* In this work, we propose to fill this gap. We consider the issue of controlling the FDR when the null distribution is Gaussian $\mathcal{N}(\theta, \sigma^2)$, with unspecified scaling parameters $\theta \in \mathbb{R}$, $\sigma > 0$, whereas the alternative distributions are let arbitrary as in the original framework of Benjamini and Hochberg (1995). This provides a general and simple setting to address the following question:

> *When the null distribution is unknown, is it possible to build a procedure that both controls the FDR at the nominal level and has a power asymptotically mimicking the oracle?*

Here, the oracle procedure corresponds to the classical Benjamini–Hochberg (BH) procedure the statistician would have carried out if an oracle had given them the true values $\theta, \sigma^2$. When it exists, a procedure satisfying the two properties defined above is henceforth said to be *AMO* for *asymptotically mimicking the oracle*.

We refer the reader to Section 7.4 for a discussion about the choice of this oracle procedure.

1.3. *Our contributions.* We consider a setting where the statistician observes *independent* real random variables $Y_i$, $1 \le i \le n$. Among these $n$ random variables, at least $(n - k)$ follow the unknown null distribution $\mathcal{N}(\theta, \sigma^2)$ and the remaining variables follow arbitrary and unknown distributions. Hence, $k$ is an upper bound of the number of false nulls, which is referred henceforth as the sparsity parameter. The latter plays an important role in our results. A reason is that having few observations under the alternatives (i.e., $k$ small) facilitates *de facto* the problem of estimating the null distribution. Our model is formally introduced in Section 2 whereas its assumptions (normality, independence) are further discussed in Section 7.

In this manuscript, we first establish the following phase transition (illustrated in Figure 2): when $k$ is much larger than $n/\log(n)$, no AMO procedure exists. Hence, any multiple testing

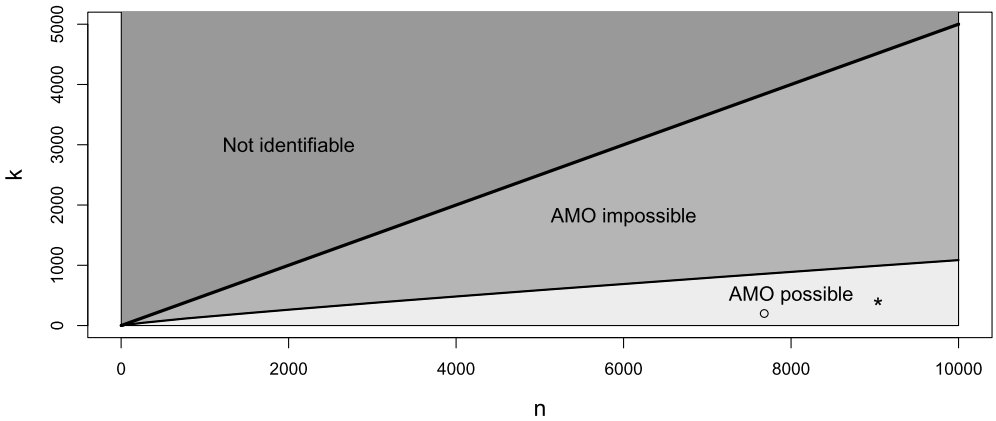


FIG. 2. *Phase transition $k \asymp n/\log n$ for $n \le 10^4$. Identifiability requires $k < n/2$. "AMO" stands for "asymptotically mimicking the performances of the oracle." For illustration, point* (7680, 200) (*circle*) *from the data set van't Wout et al.* (2003) *and point* (9038, 400) (*star*) *from the data set Bourgon, Gentleman and Huber* (2010) *have been added* (*assuming less than $k = 200$ and $k = 400$ true signals, resp.*). *Under these assumptions, both data sets fall into the AMO possible regime, below the sparsity boundary.*

procedure either violates the FDR control or is less powerful than the oracle procedure. On the other hand, when $k$ is much smaller than $n/\log(n)$, there exists a simple AMO procedure defined as follows: first compute the corrected $p$-values by plugging robust estimators of $\theta$ and $\sigma^2$ and then apply a BH procedure to these corrected $p$-values. This type of procedures is referred henceforth as a plug-in BH procedure. Figure 3 illustrates the behavior of such a plug-in procedure for different plugged values $(u, s^2)$ of the scaling parameters corresponding to the true, misspecified or estimated values of $(\theta, \sigma^2)$. This simple example shows that a wrong scaling of the null distribution can lead to poor performances, with either an uncontrolled increase of false discoveries (top-right panel), or an uncontrolled decrease of true discoveries (bottom-left panel). By contrast, fitting the null distribution with robust estimators of $\theta$ and $\sigma^2$ (bottom-right panel) seems to nearly mimic the oracle procedure (top-left panel), that is, BH procedure with the true scaling parameters $(\theta^2, \sigma^2)$.

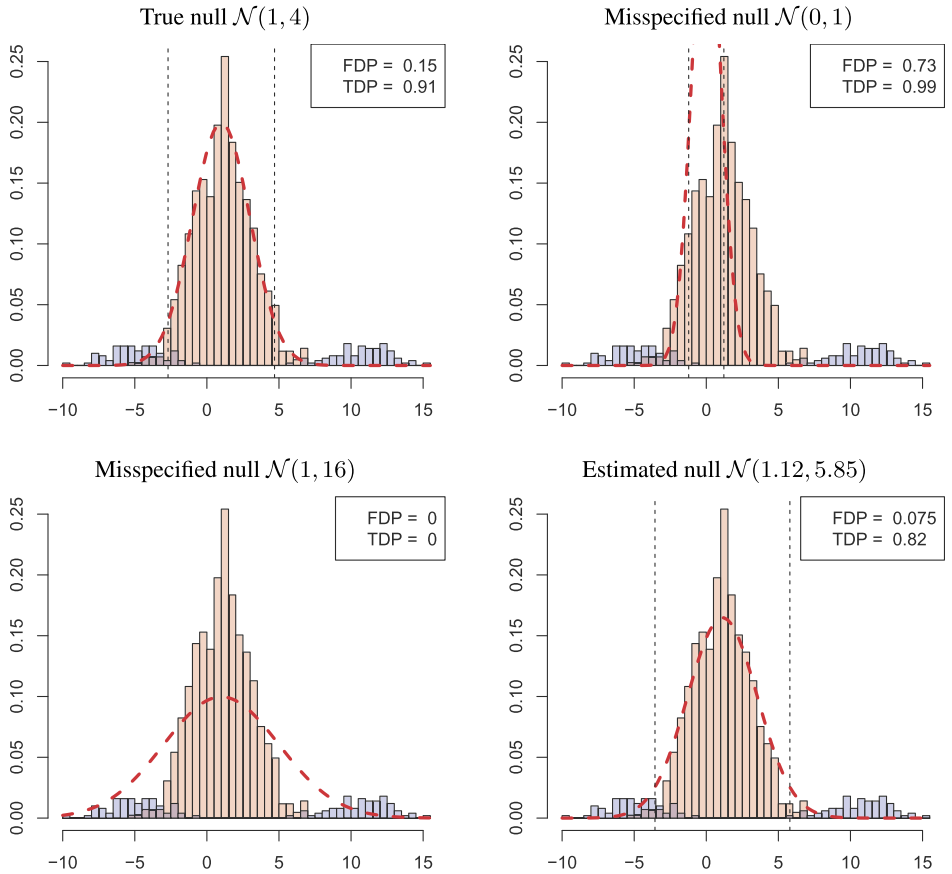


FIG. 3. *Illustration of the plug-in BH procedure with different plugged null distributions. The data have been generated as independent* $Y_i \sim \mathcal{N}(\theta + \mu_i, \sigma^2)$, *for* $\mu_i = 0$, $1 \le i \le n_0$, $\mu_i = 5$, $n_0 + 1 \le i \le n_0 + n_1/2$ *and* $\mu_i = -3$, $n_0 + n_1/2 + 1 \le i \le n$ *for* $n = 1000$, $n_0 = 850$, $n_1 = 150$, $\theta = 1$, $\sigma^2 = 4$, $\alpha = 0.2$. *Each panel displays the same overlap of the two following histograms of the data*: *colored in pink, the histogram of the* $Y_i$, $1 \le i \le n_0$, *generated under the null*; *colored in blue, the* (*rescaled*) *histogram of the* $Y_i$, $n_0 + 1 \le i \le n$, *generated under the alternative. The plug-in BH procedure is applied at level* $\alpha = 0.2$ *and its rejection threshold is displayed by the vertical dashed lines*: *the rejected null hypotheses correspond to the* $Y_i$*'s above the most-right vertical dashed line and below the most-left vertical dashed line. The FDP is the ratio of the false rejection number to the total rejection number*; *see* (2) *below. The TDP is the ratio of the true rejection number to the total number of false nulls* ($n_1$); *see* (3) *below. The plug-in BH procedure uses rescaled p-values* $p_i(u, s) = 2\overline{\Phi}(|Y_i - u|/s)$, $1 \le i \le n$, *where* $\overline{\Phi}$ *is the tail distribution of the standard normal distribution* (*see* (4) *below*) *using different values of* $u$, $s$. *Top left*: $u = 1$, $s^2 = 4$; *top right*: $u = 0$, $s^2 = 1$; *bottom left*: $u = 1$, $s^2 = 16$; *bottom right*: $u = \widetilde{\theta} \approx 1.12$, $s^2 = \widetilde{\sigma}^2 \approx 5.85$, *which are values derived from standard robust estimators*; *see* (14) *below.*

This impossibility result, together with the analysis of the plug-in procedure, is the central result of the paper. It is then extended in several directions. First, we show a stronger impossibility result that discards the existence of a procedure that always controls the FDR (even for nonsparse cases) while having a good power for simple cases ($k$ small). Indeed, any procedure whose FDR is controlled at the nominal level when $k$ is above the sparsity boundary cannot be as powerful as the oracle in simpler situations where $k$ is below the boundary. Second, we pinpoint the sparsity boundary when only the mean parameter $\theta$ is unknown or, more generally, when the density of the null distribution is arbitrary and known up to a location parameter (non-Gaussian case).

Finally, given the size of the impossible regime (dark gray areas in Figure 3), one can legitimately ask whether obtaining AMO procedures is not too demanding. Indeed, AMO procedures are required to perform as well as oracle BH procedures for all possible alternative distribution; see the definition in Section 2.6. Besides, their behavior depends on the true number of alternatives, which is in general unknown. This is why we also consider the looser but also more practical aim of assessing the validity of plugged BH procedures or more generally the validity of the empirical null approaches by a collection of tests and diagnostic graphs. This is achieved by building a nonparametric confidence region for the null distribution. On the one hand, this nonparametric confidence region leads to a goodness-of-fit test for the parametric assumption on the null distribution. On the other hand, we can plug this set of candidate null distributions into BH procedures to obtain a confidence region of rejection sets, among which the rejection set of the oracle procedure belongs to (with high probability). Hence, the stability of these rejection sets allows to assess whether one can be confident in the rejection sets of the plugged BH procedure. Diagnostic graphs are introduced to ease the interpretation. We illustrate the behavior of these diagnostic graphs on classical data sets; see Section 6 and the companion vignette (Roquain and Verzelen (2021)). More generally, these tests and diagnostic graphs easily adapt to any plug-in type empirical null method—not necessarily of the BH type—and to any candidate distribution family for the null, which makes it very versatile.

A more detailed and accurate description of our results will be given in Section 2.6 after the Introduction of required notation and definitions.

1.4. *Related works*. Many works are related to the theory developed here. First, as already mentioned, a wide literature has grown around the concept of "empirical null distribution," elaborating upon the work of Efron. While his proposal originally relies on the Gaussian null assumption, more sophisticated null distribution classes have been proposed later (Schwartzman (2008, 2010), Azriel and Schwartzman (2015), Muralidharan (2010), Sun and Stephens (2018)) to better model null coming from a multivariate correlated Gaussian vector. This results in a parametric family with much more parameters to fit. Related to this, estimating the parameters of the null has been considered in a more challenging multivariate factor model; see, for example, Fan and Han (2017), Fan, Han and Gu (2012), Friguet, Kloareg and Causeur (2009), Leek and Storey (2008). While the authors provide error bounds for the inferred factor models and consistency of the FDR estimates in some cases, none of these works establish FDR controls for the corresponding corrected BH procedure.

In fact, it turns out that few works have provided theoretical guarantees for multiple testing procedures using an empirical null distribution. Jin and Cai (2007) and Cai and Jin (2010) proposed a method to estimate the null in a particular context, but without evaluating the impact of such an operation when plugged into a multiple testing procedure. Such an attempt has been made by Ghosh (2012), who showed that the FDR control is maintained under the assumption that incorporating the empirical null distribution is an operation that can make the BH procedure only more conservative. Nevertheless, this assumption is admittedly difficult

to check. In Schwartzman (2008), the impact of the null estimation into the local/tailed area FDR estimator is measured when the null distribution belongs to an exponential family, but without further establishing FDR controlling results. Other studies have been developed in the one-sided context, for which contaminations (i.e., nonnull measurements) are assumed to arise on the upper tail (say) of the global distribution of the measurements ($z$-values). In that case, the left tail of the distribution can be used to learn the null. Such an idea has been exploited in Carpentier et al. (2021) to estimate the scaling parameters $\theta$ and $\sigma^2$ within the null $\mathcal{N}(\theta, \sigma^2)$ from the left-quantiles of the observed data. Doing so, they show that the plug-in BH procedure has performances close (asymptotically in $n$) to those of the BH procedure using the true unknown scaling. In addition, relaxing the Gaussian-null assumption, an FDR controlling procedure has been introduced in Arias-Castro and Chen (2017), Barber and Candès (2015), by only assuming that the measurements ($z$-values) are symmetric under the null. In that case, the null is implicitly learned by estimating the number of false discoveries occurring at the right-hand side of the null from its left-hand side. However, the one-sided contamination model is not the most common practical situation. The more general and realistic case of possibly two-sided alternatives will be considered throughout the paper.

Let us mention a few additional related studies with misspecified null: in Blanchard, Lee and Scott (2010), Mary and Roquain (2021), the null is unknown and estimated from an independent sample, so the setting is completely different. In Jing, Kong and Zhou (2014), the authors study the effect of nonnormality over the BH procedure using $p$-values calibrated with the Gaussian distribution. This is substantially different from our problem, where the null is assumed Gaussian with unknown parameters. Next, Pollard and van der Laan (2004) also discuss the choice of a null distribution, but the aim is to build a null that ensures a valid FWER-type error rate control, which is a goal markedly different from the one considered in this paper.

Besides, maybe on a more conceptual side, our work can be seen as a *frequentist minimax robust* analysis of empirical null distributions: first, we do assume that there exists a true null distribution and we utilize the estimated null parameters in subsequent inferences. Second, we let the alternative be arbitrary, which entails that the AMO properties should hold for any alternative and means that we consider worst-case type I/II error rates. For FDR control, this is classically referred to as a *strong* FDR control; see, for example, Dickhaus (2014). As for the notion of power, our analysis is formulated in terms of mimicking the power of the oracle in the worst case, which is new to our knowledge; see also Remark 2.3 below. Finally, the proofs of our impossibility results borrow some ideas from the literature on robust estimation and classical Huber contamination model (Huber (1964, 2011)).

### 1.5. *Notation and organization of the paper.*

*Notation.* For two sequences $u_n$ and $v_n$, $u_n \gg v_n$ means $v_n = o(u_n)$. Given a real number $x$, $\lfloor x \rfloor$ and $\lceil x \rceil$, respectively, denote the lower and upper integer parts of $x$. Given a finite set $A$, its cardinality is denoted by $|A|$. Given $x, y$, $x \wedge y$ (resp., $x \vee y$) stands for the minimum (resp., maximum) of $x$ and $y$. For $Y \sim P$, the corresponding probability is denoted $\mathbb{P}_{Y \sim P}$ or simply $\mathbb{P}$ when there is no confusion. The density of the standard normal distribution is denoted $\phi$ whereas $\overline{\Phi}$ stands for its right tail distribution, that is, $\overline{\Phi}(z) = \mathbb{P}(Z \geq z)$, $z \in \mathbb{R}$, $Z \sim \mathcal{N}(0, 1)$. Finally, given a vector $v \in \mathbb{R}^n$, we denote by $v_{(i)}$ the $i$th order statistic of $v$, that is, the $i$th smallest entry of $v$.

*Organization of the paper.* The setting and the main results are described in Section 2. While they are formulated in an asymptotical manner for simplicity, more accurate nonasymptotical counterparts are provided in Section 3: an impossibility result is given in

Section 3.1 (with a corollary given in Section 3.3) and a matching upper bound is provided in Section 3.2. Section 4 is devoted to the situation where the variance of the null is known, while Section 5 provides extensions to a general location model. The null confidence region as well the diagnostic graphs are presented in Section 6 together with some illustrations on synthetic and real data sets. Discussions are postponed to Section 7. Numerical experiments, proofs, lemmas and auxiliary results are deferred to the Supplementary Material (Roquain and Verzelen (2022)). An application of our approach on real data sets is developed carefully in a dedicated vignette (Roquain and Verzelen (2021)).

## 2. Setting and presentation of the main results.

2.1. *Framework for testing with an unknown null distribution.* To formalize our setting, we resort to a variation of Huber's model (Huber (1964)). We observe *independent* real random variables $Y_i$, $1 \le i \le n$. The distribution of the vector $Y = (Y_i)_{1\le i\le n}$ in $\mathbb{R}^n$ is denoted by $P = \bigotimes_{i=1}^n P_i$. We assume that most of the $P_i$'s follow the same (null) distribution while the others are "contaminated" and can be arbitrary. Also, following the setting introduced in Efron (2004), we shall assume in this manuscript that this null distribution is of the form $\mathcal{N}(\theta, \sigma^2)$ for some unknown scaling $(\theta, \sigma) \in \mathbb{R} \times (0, \infty)$ (except in Sections 5 and 6 where different or more general nulls are considered). Formally, this leads us to assuming that $P = \bigotimes_{i=1}^n P_i$ belongs to the collection $\mathcal{P}$ of all distributions satisfying

(1) there exists $(\theta, \sigma) \in \mathbb{R} \times (0, \infty)$ such that $\big|\{i \in \{1, \dots, n\} : P_i = \mathcal{N}(\theta, \sigma^2)\}\big| > n/2$.

In other words, (1) ensures that there exists a scaling $(\theta, \sigma)$ such that more than half of the $P_i$'s are $\mathcal{N}(\theta, \sigma^2)$. Condition (1) makes the problem identifiable with respect to the unknown null distribution: if $P$ is any distribution which satisfies the condition in (1) with two parameters $(\theta, \sigma)$ and $(\theta', \sigma')$, then we necessarily have $|\{i \in \{1, \dots, n\} : P_i = \mathcal{N}(\theta, \sigma^2) = \mathcal{N}(\theta', (\sigma')^2)\}| > 0$ and $\theta = \theta'$, $\sigma = \sigma'$. For $P \in \mathcal{P}$, we denote by $(\theta(P), \sigma(P))$ the unique couple satisfying (1). This allows us to formulate the multiple testing problem:

$$H_{0,i} : \text{“}P_i = \mathcal{N}\big(\theta(P), \sigma^2(P)\big)\text{”} \quad \text{against} \quad H_{1,i} : \text{“}P_i \ne \mathcal{N}\big(\theta(P), \sigma^2(P)\big)\text{”},$$

for all $1 \le i \le n$. We underline that $H_{0,i}$ is not a point mass null hypotheses, that is, "$P_i = P^0$", for some known distribution $P^0$, nor a composite null of the type "$P_i$ is a Gaussian distribution," but a point mass null hypothesis with value depending on all the marginals $(P_j, 1 \le j \le n)$.

Let us introduce some notation. We denote by $\mathcal{H}_0(P) = \{1 \le i \le n : \ H_{0,i} \text{ is true for } P\}$ the set of true null hypotheses, by $n_0(P) = |\mathcal{H}_0(P)|$ its cardinality and by $\mathcal{H}_1(P)$ its complement in $\{1, \dots, n\}$. We also let $n_1(P) = |\mathcal{H}_1(P)| = n - n_0(P)$, so that $n_1(P) < n/2$ by (1). As an illustration, if $P = \bigotimes_{i=1}^n P_i$ is given by

$$(P_i, 1 \le i \le n) = \big(P_1, P_2, \mathcal{N}(1,4), \mathcal{N}(1,4), P_5, \mathcal{N}(1,4), \mathcal{N}(1,4)\big)$$

for $n = 7$ and some distributions $P_1$, $P_2$, $P_5$ on $\mathbb{R}$ that are all different from $\mathcal{N}(1,4)$, we have $\theta(P) = 1$, $\sigma^2(P) = 4$, $\mathcal{H}_0(P) = \{3, 4, 6, 7\}$ and $n_1(P) = 3$.

We will sometimes consider an asymptotic situation where $n$ tends to infinity. In that case, the quantities $\mathcal{P}$, $P$, $Y$ (and those related) are all depending on $n$, but we remove such dependencies in the notation for the sake of brevity.

Our model assumptions are related to the classical Huber contamination models in robust statistics and to the two-group model in multiple testing theory. We postpone the corresponding discussion to Sections 7.2 and 7.3.

2.2. *Criteria.* A multiple testing procedure is defined as a measurable function $R$ taking as input the data $Y$ and returning a subset $R(Y) \subset \{1, \dots, n\}$ corresponding to the set of rejected null hypotheses among $(H_{0,i}, 1 \le i \le n)$. The amount of false positives of $R$ (type I errors) is classically measured by the false discovery proportion of $R$:

$$\mathrm{FDP}\big(P, R(Y)\big) = \frac{|R(Y) \cap \mathcal{H}_0(P)|}{|R(Y)| \vee 1}; \tag{2}$$

see Benjamini and Hochberg (1995). The expectation $\mathrm{FDR}(P, R) = \mathbb{E}_{Y\sim P}[\mathrm{FDP}(P, R(Y))]$ is the false discovery rate of the procedure $R$. The amount of true positives of $R$ is measured by

$$\mathrm{TDP}\big(P, R(Y)\big) = \frac{|R(Y) \cap \mathcal{H}_1(P)|}{n_1(P) \vee 1}, \tag{3}$$

and corresponds to the proportion of (correctly) rejected nulls among the set of false null hypotheses. It has been often used as a power metric for multiple testing procedures; see, for example, Arias-Castro and Chen (2017), Benjamini and Hochberg (1995), Rabinovich et al. (2020), Roquain and van de Wiel (2009).

2.3. *Plug-in BH procedures.* In our study, an important class of procedures are the BH procedures with rescaled $p$-values, that we call the plug-in BH procedures. This corresponds to first estimating the null distribution $(\theta(P), \sigma(P))$ and then plugging it into BH.

Since Benjamini–Hochberg (BH) procedure is defined through the $p$-value family, we first define, for $u \in \mathbb{R}$ and $s > 0$, the rescaled $p$-values

$$p_i(u, s) = 2\overline{\Phi}\left(\frac{|Y_i - u|}{s}\right), \quad u \in \mathbb{R}, s > 0, 1 \le i \le n, \tag{4}$$

which corresponds to the situation where $\theta(P)$, $\sigma(P)$ have been estimated by $u$, $s$, respectively. By convention, the value $s = +\infty$ is allowed here, which gives a rescaled $p$-value always equal to 1. The oracle $p$-values are then given by

$$p_i^* = p_i\big(\theta(P), \sigma(P)\big), \quad 1 \le i \le n. \tag{5}$$

DEFINITION 2.1. Let $\alpha \in (0, 1)$, $u \in \mathbb{R}$, $s > 0$ and $P \in \mathcal{P}$. The plug-in BH procedure of level $\alpha$ with scaling $u$ and $s$ is given by

$$\mathrm{BH}_\alpha(Y; u, s) = \big\{1 \le i \le n : p_i(u, s) \le T_\alpha(Y; u, s)\big\} \tag{6}$$

$$= \big\{1 \le i \le n : p_i(u, s) \le T_\alpha(Y; u, s) \vee (\alpha/n)\big\}; \tag{7}$$

$$T_\alpha(Y; u, s) = \max\left\{t \in [0, 1] : \sum_{i=1}^{n} \mathbb{1}\big\{p_i(u, s) \le t\big\} \ge nt/\alpha\right\}. \tag{8}$$

In particular, the oracle BH procedure (of level $\alpha$) is defined as the plug-in BH procedure (of level $\alpha$) with scaling $\theta(P)$ and $\sigma(P)$, that is, is defined by $\mathrm{BH}_\alpha^*(Y) = \mathrm{BH}_\alpha(Y; \theta(P), \sigma(P))$.

Note that the equivalence between (6) and (7) comes from the fact that if $T_\alpha(Y; u, s) > 0$, then $T_\alpha(Y; u, s) \ge \alpha/n$ and if $T_\alpha(Y; u, s) = 0$ then all $p$-values are larger than $\alpha/n$ and the sets in (6) and (7) are both empty.

When not ambiguous, we will sometimes drop $Y$ in the notation $\mathrm{BH}_\alpha(Y; u, s)$, $T_\alpha(Y; u, s)$, and $\mathrm{BH}_\alpha^*(Y)$ for short. The oracle procedure $\mathrm{BH}_\alpha^*$ corresponds to the situation where the true scaling $(\theta(P), \sigma(P))$ is directly plugged into the BH procedure. It is the oracle procedure in our study. In our framework, the $p$-values $p_i^*$ are all independent, with the property $p_i^* \sim$

$U(0,1)$ whenever $i \in \mathcal{H}_0(P)$. Hence, it is well known (Benjamini and Hochberg (1995), Benjamini and Yekutieli (2001)) that its FDR satisfies the following:

$$\forall P \in \mathcal{P}, \quad \mathrm{FDR}\big(P, \mathrm{BH}_\alpha^*\big) = \alpha n_0(P)/n. \tag{9}$$

To mimic $\mathrm{BH}_\alpha^*$, natural candidates are the plug-in BH procedures $\mathrm{BH}_\alpha(\widehat{\theta}, \widehat{\sigma})$, for some suitable estimators $\widehat{\theta}, \widehat{\sigma}$ of $\theta(P), \sigma(P)$ (by convention, the value $\widehat{\sigma} = \infty$ is allowed here). In the sequel, $(\widehat{\theta}, \widehat{\sigma})$ is called a rescaling.

2.4. *AMO procedures*. To evaluate how a procedure is mimicking $\mathrm{BH}_\alpha^*$ on some sparsity range, let us define the following notation: for any procedure $R(Y) \subset \{1, \dots, n\}$, any sparsity parameter $k \in [1, n/2]$ and any level $\alpha \in (0,1)$, we let

$$\mathbf{I}(R,k) = \sup_{\substack{P \in \mathcal{P} \\ n_1(P) \le k}} \big\{\mathrm{FDR}(P,R)\big\}; \tag{10}$$

$$\mathbf{II}(R,k,\alpha) = \sup_{\substack{P \in \mathcal{P} \\ n_1(P) \le k}} \big\{\mathbb{P}_{Y\sim P}\big(\mathrm{TDP}(P,R) < \mathrm{TDP}\big(P, \mathrm{BH}_\alpha^*\big)\big)\big\}. \tag{11}$$

The latter can be seen as maximum type I/II risks, with a supremum taken on distributions $P \in \mathcal{P}$ with $n_1 = n_1(P) \le k$, $\theta = \theta(P)$ and $\sigma = \sigma(P)$ varying accordingly. In particular, $\mathbf{I}(R,k) \le \alpha$ corresponds to a strong FDR control on the range of distributions with at most $k$ false nulls, and we have $\mathbf{I}(\mathrm{BH}_\alpha^*, k) = \alpha$ for any $k$ when using the particular oracle BH procedure; see (9). The criterion $\mathbf{II}(R,k,\alpha)$ is a (maximum) type II risk defined relatively to $\mathrm{BH}_\alpha^*$: it is small when the TDP of $R$ is at least as large as the one of $\mathrm{BH}_\alpha^*$, with a large probability. In particular, the map $\alpha \mapsto \mathbf{II}(R,k,\alpha)$ is nondecreasing. Then a procedure is said to mimic the oracle if it maintains the strong FDR control while having a small (maximum) relative type II risk.

DEFINITION 2.2. Let $R = (R_\alpha)_{\alpha\in(0,1)}$ be a sequence of multiple testing procedures, both depending on the nominal level $\alpha$ and of the number $n$ of tests. For a given sparsity sequence $k_n \in [1, n/2)$, the procedure sequence $R$ is asymptotically mimicking the oracle BH procedure, AMO in short, whenever the two following properties hold: there exists a positive sequence $\eta_n \to 0$ such that

$$\limsup_n \sup_{\alpha\in(1/n,1/2)} \big\{\mathbf{I}(R_\alpha, k_n) - \alpha\big\} \le 0; \tag{12}$$

$$\lim_n \sup_{\alpha\in(1/n,1/2)} \big\{\mathbf{II}\big(R_\alpha, k_n, \alpha(1-\eta_n)\big)\big\} = 0. \tag{13}$$

Furthermore, if $\widehat{\theta}$ and $\widehat{\sigma}$ are two (sequences of) estimators of $\theta(P)$ and $\sigma(P)$, respectively, the rescaling $(\widehat{\theta}, \widehat{\sigma})$ is said to be AMO if the sequence of plug-in BH procedures $(\mathrm{BH}_\alpha(\widehat{\theta}, \widehat{\sigma}))_{\alpha\in(0,1)}$ is AMO.

In this definition, the performances of the oracle BH procedure are mimicked both in terms of FDR and TDP. Note that the power statement is made slightly weaker than one could expect at first sight, with a slight decrease of the level in $\mathrm{BH}^*_{\alpha(1-\eta_n)}$. Since $\eta_n$ converges to 0, this modification is very light. However, taking $\eta_n = 0$ would certainly be too demanding for a plug-in procedure $\mathrm{BH}_\alpha(\widehat{\theta}, \widehat{\sigma})$ as its TDP could fluctuate on both sides of $\mathrm{TDP}(P, \mathrm{BH}_\alpha^*)$. Alternatively, if one wants a comparison with the oracle procedure $\mathrm{BH}_\alpha^*$ (without modification of the level), the convergence (13) can be equivalently replaced by $\lim_n \sup_{\alpha\in(1/n,1/2)}\{\mathbf{II}(R_{\alpha(1+\eta_n)}, k_n, \alpha))\} = 0$. This would not change our results. Also, we underline that, while the statements (12) and (13) are formulated in an asymptotic manner for compactness, all our results can be formulated nonasymptotically.

REMARK 2.3. Instead of stochastically comparing the true discovery proportions in (11), an alternative could have been to compare their expectations. The expectation of the TDP, called the true discovery rate (TDR), is the standard notion of power in the literature (see, for instance, Arias-Castro and Chen (2017), Rabinovich et al. (2020), Roquain and van de Wiel (2009)) where specific classes of alternative distributions are considered. Here, the TDR is not a suitable measure of power, because the alternative distribution is let completely free in (11). As a result, in some cases, the TDR is maximized by trivial procedures that typically reject no null hypothesis with probability $1-\alpha$ and reject all null hypotheses with probability $\alpha$. As such procedures are obviously undesirable, we focus on the stronger TDP stochastic domination property required in (13).

2.5. *Robust estimation of* $(\theta(P), \sigma(P))$. Since our framework allows arbitrary alternative distributions, we consider simple robust estimators of $(\theta(P), \sigma(P))$ defined by

$$\widetilde{\theta} = Y_{(\lceil n/2 \rceil)}; \qquad \widetilde{\sigma} = U_{(\lceil n/2 \rceil)} / \overline{\Phi}^{-1}(1/4), \tag{14}$$

where $U_i = |Y_i - Y_{(\lceil n/2 \rceil)}|$, $1 \le i \le n$, and $\overline{\Phi}^{-1}(1/4) \approx 0.674$. While $\widetilde{\theta}$ is the sample median, $\widetilde{\sigma}$ corresponds to a suitable rescaling of $U_{(\lceil n/2 \rceil)}$, the median absolute deviation (MAD) of the sample. Under the null, the variables $|Y_i - \theta|/\sigma$ are i.i.d. and distributed as the absolute value of a standard Gaussian variable. Hence, taking the median of the $|Y_i - \theta|$ should be a robust estimator of $\sigma$ times the median of the absolute value of a standard Gaussian variable, that is, of $\sigma \overline{\Phi}^{-1}(1/4)$. Rescaling suitably this quantity and replacing $\theta$ by $\widetilde{\theta}$ leads to the definition of $\widetilde{\sigma}$. The two estimators defined by (14) are minimax optimal; see, for example, Chen, Gao and Ren (2018) for a result in a slightly different mixture model. We will use here specific properties of these estimators to be found in Section S-6.1.

2.6. *Presentation of the results*.

2.6.1. *Main result*. We now state the main result of the paper.

THEOREM 2.4. *In the setting of Section 2 and according to Definition 2.2, the following hold*:

(i) *for a sparsity parameter* $k_n \gg n/\log(n)$, *there exists no sequence of procedures that is AMO*;

(ii) *for a sparsity parameter* $k_n \ll n/\log(n)$, *the sequence of plug-in BH procedures* $(BH_\alpha(\widetilde{\theta}, \widetilde{\sigma}))_{\alpha \in (0,1)}$ *is AMO, for the scaling* $(\widetilde{\theta}, \widetilde{\sigma})$ *given by standard robust estimators* (14).

This result delineates a phase transition for the problem of multiple testing with an unknown Gaussian null distribution. The transition is expressed in function of the sparsity parameter $k$ (see Figure 2 for a graphical illustration).

In more detail, part (i) of Theorem 2.4 (lower bound) means that when the proportion of true alternative hypotheses is much larger than $1/\log(n)$, it is not possible to perform as well as an oracle that knows the null distribution in advance. Obtaining impossibility results on FDR control has recently received some attention in multiple testing literature (Arias-Castro and Chen (2017), Castillo and Roquain (2020), Rabinovich et al. (2020)), but those are restricted to the class of thresholding-based procedures. Here, our impossibility holds for *any* multiple testing procedure. The proof of our lower bound relies on a Le Cam's two-point reduction scheme. Namely, it is derived by identifying two mixture distributions on $\mathbb{R}^n$ that are indistinguishable while corresponding to distant null distributions (see Figure 4) and by studying the impact of such a fuzzy configuration on the FDR and TDP metrics. While this

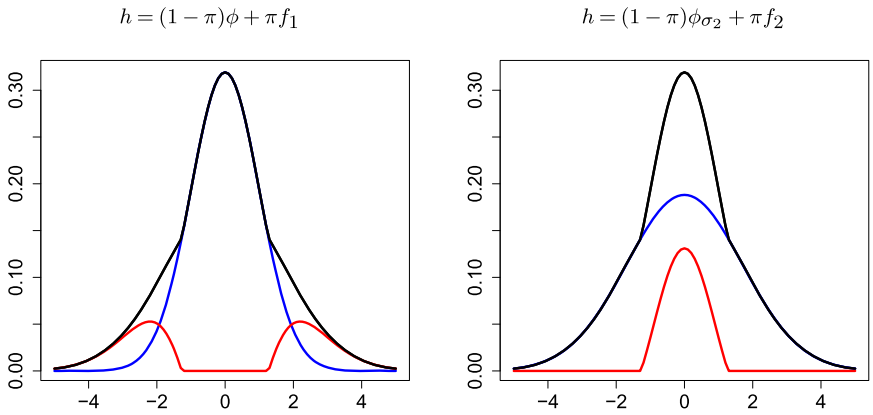


FIG. 4. *Left*: *the density h given by* (17) (*black*), *interpreted as a mixture between the null* $\mathcal{N}(0,1)$ ($(1-\pi)\phi$ *in blue*) *and the alternative* $f_1$ ($\pi f_1$ *in red*). *Right*: *the same h interpreted as a mixture between the null* $\mathcal{N}(0,\sigma_2^2)$ ($(1-\pi)\phi_{\sigma_2}$ *in blue*) *and the alternative* $f_2$ ($\pi f_2$ *in red*). $\pi = 1/5$, $\sigma_2^2 \approx 2.88$.

argument is classical in the estimation or (single) testing literature (see, e.g., Tsybakov (2009) and Donoho and Jin (2006)), it is to our knowledge new in the multiple testing context.

Part (ii) of Theorem 2.4 (upper bound) is proved in Section 3.2. For this, we extend the ideas used in Carpentier et al. (2021) to accommodate the new two-sided geometry of the test statistics. In particular, correcting the $Y_i$'s by $\widehat{\theta}$ changes the order of the $p$-values, which was not the case in the one-sided situation. Our proof relies on the symmetry of the Gaussian distribution and on special properties of the BH procedure rejection set when removing one element of the $p$-value family; see, for example, Ferreira and Zwinderman (2006). Also note that the estimators $(\widetilde{\theta}, \widetilde{\sigma})$ do not use the knowledge of $k_n$, which implies that the procedure is adaptive with respect to the sparsity $k_n$ on the range $k_n \ll n/\log(n)$.

2.6.2. *Extending the scope of the main result.* We provide three complementary results. First, in the testing literature, controlling the inflation of type I error rates is generally considered as more crucial than enhancing the power. In our framework, we can always design a plug-in BH procedure that controls the FDR by simply setting $\widehat{\sigma} = \infty$, which is equivalent to taking $R(Y) = \varnothing$ (no rejection). In view of this remark, the impossibility result above shows that mimicking the power of the oracle is impossible above the boundary. Hence, we can reinterpret the statement of Theorem 2.4 as follows:

(i) in the dense regime ($k_n \gg n/\log(n)$), it is possible to achieve (12) but not with (13);
(ii) in the sparse regime ($k_n \ll n/\log(n)$), it is possible to achieve both (12) and (13).

A natural question is then: can we achieve the best of the two worlds? Is it possible to find a rescaling satisfying (12) in the dense regime and both (12) and (13) in the sparse regime? We establish in Section 3.3 that such a procedure does not exist; see Corollary 3.3. As a consequence, any procedure controlling the FDR in the dense regime is not AMO in the sparse regime. Conversely, any AMO procedure in the sparse regime is not able to control the FDR in the dense regime. This is the case in particular for the plug-in procedure $\mathrm{BH}_\alpha(\widetilde{\theta}, \widetilde{\sigma})$ considered in Theorem 2.4(ii). More formally, combining Corollary 3.3 ($\alpha = c_3/2$) and Theorem 3.2 below leads the following result.

COROLLARY 2.5. *There exist numerical constants* $\alpha_0 \in (0, 1/2)$ *and* $c > 0$ *such that for any sequence* $u_n \to \infty$,

$$\liminf_n \{\mathbf{I}(BH_{\alpha_0}(\widetilde{\theta}, \widetilde{\sigma}), u_n n/\log(n)) - \alpha_0\} > c.$$

Second, in Section 4, we show an analogue of Theorem 2.4 when $\sigma = \sigma(P)$ is known. Hence, the only unknown null parameter is $\theta$ and the class of rescaling is restricted to those of the form $(\widehat{\theta}, \sigma)$, where $\widehat{\theta}$ is an estimator of $\theta$. We establish that the sparsity boundary is slightly modified in this case: impossibility is shown for $k_n \gg n/\log^{1/2}(n)$, while $(\widetilde{\theta}, \sigma)$ is AMO for $k_n \ll n/\log^{1/2}(n)$ (Theorem 4.1). While the upper-bound part is similar to the upper-bound part of Theorem 2.4 above, the lower bound arguments have to be adapted to the case where only the location parameter is unknown. More precisely, we establish two types of lower bounds. We first develop a lower bound valid for any multiple testing procedure (Theorem 4.2), which follows the same philosophy as the lower bound developed in Theorem 2.4 (via Theorem 3.1). Next, we provide a refined lower bound specifically tailored to plug-in BH type procedures. Contrary to the previous lower bounds, it does not state type I error/type II error trade-offs but it establishes that uniform control of the FDR alone is already out of reach. Namely, this result shows that, on the sparsity range $k_n \gg n/\log^{1/2}(n)$, any plug-in procedure exhibits a FDP close to $1/2$ and makes around $n^{3/4}$ false discoveries, this on an event of probability close to $1/2$ (see Theorem 4.4). Intuitively, this comes from the fact that $\widehat{\sigma} = \sigma$ is fixed to the true value, and thus cannot compensate the estimation error of $\widehat{\theta}$, which irremediably leads to many false discoveries in that regime.

Third, we extend our results to the case where the null distribution has a known symmetric density $g$ with an unknown location parameter; see Section 5. Therein, we derive lower bounds in two different regimes, when $k_n/n$ tends to zero (Theorem 5.1) and when $k_n/n$ is of order constant (Theorem 5.2). Also, we provide a general upper bound matching the lower bounds under assumptions on $g$ (Theorem 5.3). As expected, the sparsity boundary depends on $g$. For instance, for $\zeta$-Subbotin null $g(x) = L_\zeta^{-1} e^{-|x|^\zeta/\zeta}$, $\zeta > 1$, the boundary is proved to be $k_n \asymp n/(\log(n))^{1-1/\zeta}$ (Corollary 5.4), which recovers the Gaussian case for $\zeta = 2$. For the Laplace distribution $g(x) = e^{-|x|}/2$, AMO scaling is possible as long as $k_n \ll n$ (Corollary 5.5). Finally, we further explore the behavior of any procedure for the Laplace distribution on the boundary when $k_n$ is of the same order as $n$ (Proposition 5.6).

2.6.3. *Confidence region for the null and applications.* Our previous analysis shows that, when the sparsity is not strong enough, we cannot hope to build a procedure that mimics the properties of the oracle BH procedure. This holds in the minimax sense, that is, this impossibility is shown to be met under a least favorable configuration (see Figure 4 below). However, if the underlying distribution $P$ is reasonably far from these worst-case configurations, it is not necessarily impossible to perform as best as the oracle BH procedure under the distribution $P$. Hence, for some specific data sets that are not sparse, one can possibly reliably estimate the null distribution and plug a BH procedure. This raises the issue of deriving data-driven and distribution-dependent measures of the reliability of plug-in null estimation methods. This is the topic of Section 6. In that section, we state a general, nonasymptotic, confidence region for the null distribution (Theorem 6.1). The latter holds without any assumption on the sparsity and on the null distribution. It only requires an upper-bound $k$ on the number of false nulls.

We then derive from this result several corollaries. First, this rejection set also induces a goodness-of-fit test for any given null distribution (Corollary 6.2) or for any family of null distributions (Corollary 6.3). As shown in the companion vignette (Roquain and Verzelen (2021)) for several data sets, the theoretical null $\mathcal{N}(0, 1)$ is rejected while the family of Gaussian null is accepted. Again, this reinforces the interest in using Gaussian empirical nulls, as Efron suggested in the first place.

Then, plugging this confidence region into a Benjamini–Hochberg procedure, we obtain a confidence region for the rejection set—or the size of the rejection set—of the oracle BH procedure; see Corollary 6.4. This complements our theoretical results as this provides a

purely data-driven diagnostic for assessing the possible validity of a plug-in procedure: if the rejection set (or number) is fairly stable over the confidence region, then the statistician can confidently use the plug-in approach. By contrast, if the rejection set varies a lot in the confidence region and, in particular, if the empty set belongs to the confidence region of rejection sets, then the user cannot confidently reject hypotheses and should certainly make no rejection. To ease the interpretation, we introduce a visualization method in that section via "diagnostic graphs"; see Figure 5. Finally, we underline that this approach of building a confidence region for the rejection set can be applied to any type of null distribution, not necessarily Gaussian.

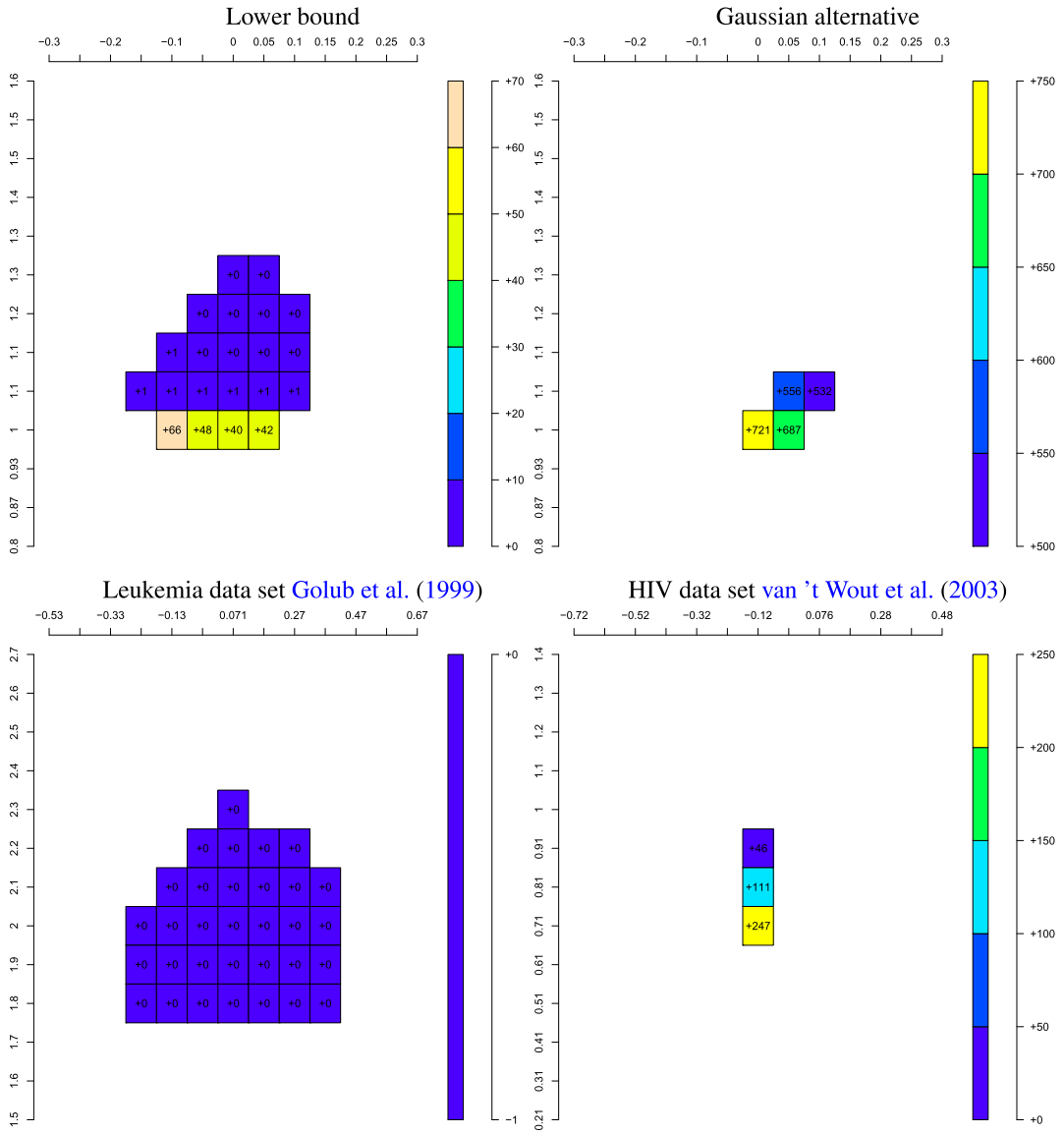


FIG. 5. *Plot of the confidence region* $\mathcal{S}_{k,\alpha}$ (40) *for* $\alpha = 0.1$, $k/n = 0.1$, *in the Gaussian Huber model* ($X$*-axis*: $\theta$. $Y$*-axis*: $\sigma$). *Top*: *two simulated data sets with* $n = 10{,}000$. *Bottom*: *the two real data sets Golub et al.* (1999) (*bottom left*) *and van't Wout et al.* (2003) (*bottom right*) *for which* $n = 3051$ *and* $n = 7680$, *respectively. In each pixel* $(\theta, \sigma)$ *of the region, the depicted number is the rejection number of the plug-in BH procedure at level* $\alpha = 0.1$ *using the corresponding scaling. More details are given in the text. The bottom panels are reproducible from the vignette* (*Roquain and Verzelen* (2021)).

## 3. Nonasymptotical bounds.

3.1. *Lower bound.* To prove part (i) of Theorem 2.4, we establish a more general, nonasymptotic, impossibility result.

THEOREM 3.1. *There exist numerical positive constants $c_1$–$c_5$ such that the following holds for all $n \geq c_1$ and any $\alpha \in (0,1)$. Consider any two positive numbers $k_1 \leq k_2$ satisfying*

$$c_2 \frac{n \log(2/\alpha)}{\log(n)} \left[1 + \log\left(\frac{k_2}{k_1}\right)\right] \leq k_2 < n/2. \tag{15}$$

*Then, for any multiple testing procedure $R$ such that*

$$FDR(P,R) \leq c_3, \quad \text{for any } P \in \mathcal{P} \text{ with } n_1(P) \leq k_2\ ,$$

*there exists some $P \in \mathcal{P}$ with $n_1(P) \leq k_1$ such that we have*

$$\mathbb{P}_{Y\sim P}\big(|R(Y) \cap \mathcal{H}_1(P)| = 0\big) \geq 2/5;$$

$$\mathbb{P}_{Y\sim P}\left[|BH^*_{\alpha/2} \cap \mathcal{H}_1(P)| \geq c_4 \alpha^{-1} \left\{\frac{n}{\log n}\right\}^{1/2}\right] \geq 1 - e^{-c_5 \alpha^{-1}\{n/\log(n)\}^{1/2}} \geq 4/5. \tag{16}$$

*In particular, we have that $\mathbf{I}(R, k_2) \leq c_3$ implies $\mathbf{II}(R, k_1, \alpha/2) \geq 1/5$.*

Theorem 3.1 states that, for any procedure $R$, either the FDR is not controlled at the nominal level $\alpha \leq c_3$ for all $P$ with $n_1(P) \leq k_2$ or there exists a distribution $P$ with $n_1(P) \leq k_1$ such that $R$ does not make any correct rejection with positive probability while the oracle procedure $\mathrm{BH}^*_{\alpha/2}$ make at least (of the order of) $\{n/\log(n)\}^{1/2}$ correct rejections with probability close to one.

Now, let us show that Theorem 3.1 implies part (i) of Theorem 2.4. Consider any sequence $k_n$ with $n/2 > k_n \gg n/(\log n)$, any sequence $\eta_n \to 0$, an arbitrary sequence of procedures $(R_\alpha)_{\alpha\in(0,1)}$, and choose $\alpha = (c_3 \wedge 1)/2$. Clearly, for $n$ large enough, the sparsity parameters $k_1 = k_2 = k_n$ satisfy the requirements of Theorem 3.1, and thus, for $n$ large, either $\mathbf{I}(R_\alpha, k_n) - \alpha > (c_3 \wedge 1)/2$ or $\mathbf{II}(R_\alpha, k_n, \alpha/2) \geq 1/5$. This entails that (12) and (13) cannot hold simultaneously.

Let us provide some high-level ideas of the proof of Theorem 3.1; we refer to Section S-2 for the details. As explained later in Section 7, two-group models can be viewed as random instances of our setting and it suffices to prove that no AMO procedure exists in this setting. Let us assume that $Y_i$, $1 \leq i \leq n$, are i.i.d. and have a common distribution given by the mixture density

$$h = (1-\pi)\phi + \pi f_1, \tag{17}$$

where $\phi$ is the density of the standard Gaussian distribution, $f_1$ is the density of the alternative and $\pi \in (0,1)$ is a prescribed proportion of signal. This density is depicted in the left panel of Figure 4, for some specific choice of $\pi$ and $f_1$. Here, the alternative density looks nicely separated from the null $\phi$, which indicates that the oracle procedure should typically make some rejections. By contrast, consider the situation depicted in the right panel of Figure 4, where the null density is given by $\phi_{\sigma_2}(\cdot) = \sigma_2^{-1}\phi(\cdot/\sigma_2)$ and the alternative is given by a density $f_2$, concentrated near 0. In that situation, the alternative density is not well distinguishable from the null so that the oracle procedure, which "knows" what is the null distribution, makes no rejection with high probability to ensure a correct FDR control.

The point is that $f_1$ and $f_2$ are chosen so that the two mixture densities in the left and right panel coincide. Hence, when the data are generated by this mixture, *any data-driven*

*procedure* cannot decipher whether the data arise from $(\phi, f_1)$ or $(\phi_{\sigma_2}, f_2)$. As a result, a data-driven procedure is not able to mimic the behavior of the oracle as soon as the distribution of the rejection number of the oracle highly differs in the two situations. Quantifying precisely the latter provides a condition on the sparsity parameter $\pi = \pi_n$, namely $\pi_n \gg 1/\log(n)$, under which no AMO procedure exists.

3.2. *Upper bound.* In this section, we prove Part (ii) of Theorem 2.4. The following result states FDR and power oracle inequalities.

THEOREM 3.2. *In the setting of Section 2, there exist universal constants $c_1, c_2 > 0$ such that the following holds for all $n \geq c_1$ and $\alpha \in (0, 0.5)$. Consider any number $k \leq 0.1n$ such that $\eta = c_2 \log(n/\alpha)((k/n) \vee n^{-1/6}) \leq 0.05$. Then we have*

$$\mathbf{I}\big(BH_\alpha(\widetilde{\theta}, \widetilde{\sigma}), k\big) \leq \alpha(1+\eta) + e^{-n^{1/2}}; \tag{18}$$

$$\mathbf{II}\big(BH_\alpha(\widetilde{\theta}, \widetilde{\sigma}), k, \alpha(1-\eta)\big) \leq e^{-n^{1/2}}. \tag{19}$$

Let us check that Theorem 3.2 implies (ii) of Theorem 2.4. If $\log(n)k_n/n$ tends to zero and $\alpha \in (1/n, 1/2)$, we have $\eta \leq 2c_2 \log(n)(\frac{k_n}{n} \vee n^{-1/6})$, which is smaller than 0.05 for $n$ large enough, and by (18) above and (9),

$$\sup_{\alpha \in (1/n, 1/2)} \big\{\mathbf{I}\big(\mathrm{BH}_\alpha(\widetilde{\theta}, \widetilde{\sigma}), k\big) - \alpha\big\} \leq \eta + e^{-n^{1/2}},$$

which converges to 0 as $n$ grows to infinity. This gives (12) for $(\widehat{\theta}, \widehat{\sigma}) = (\widetilde{\theta}, \widetilde{\sigma})$. Similarly,

$$\sup_{\alpha \in (1/n, 1/2)} \big\{\mathbf{II}\big(\mathrm{BH}_\alpha(\widetilde{\theta}, \widetilde{\sigma}), k, \alpha(1-\eta)\big)\big\} \leq e^{-n^{1/2}} \to 0,$$

which gives (13) for $(\widehat{\theta}, \widehat{\sigma}) = (\widetilde{\theta}, \widetilde{\sigma})$ and $\eta_n = \eta$.

The proof of Theorem 3.2 is given in Section S-3. The general argument can be summarized as follows. Observing that the estimators $\widetilde{\theta}$, $\widetilde{\sigma}$ converge at the rate $n_1(P)/n + n^{-1/2}$ (Lemma S-6.1), we mainly have to quantify the impact of these errors on the FDR/TDP metrics. To show (18), we establish that the FDR metric is at worst perturbed by the estimation rate multiplied by $\log(n/\alpha)$. Here, $\alpha/n$ corresponds to the smallest $p$-value threshold of the BH procedure. This can be shown by studying how the $p$-value process is affected by misspecifying the scaling parameters (Lemma S-3.2). A difficulty stems from the fact that the FDR metric is not monotonic in the rejection set, so that specific properties of BH procedure and of the estimators $\widetilde{\theta}$, $\widetilde{\sigma}$ are required (Lemmas S-3.1 and S-3.3). The second result (19) is proved similarly, the main difference being that we need a slight decrease in the level $\alpha$ (Lemma S-3.2) of the oracle procedure $\mathrm{BH}^*_\alpha$ to compare the BH thresholds $T_\alpha(\theta(P), \sigma(P))$ and $T_\alpha(\widehat{\theta}, \widehat{\sigma})$. This results in a level $\alpha(1-\eta)$ instead of $\alpha$ in (19).

3.3. *Relation between FDR and power across the boundary.* Theorem 2.4 establishes that it is impossible to perform as well as the oracle BH procedure when $k_n \gg n/\log(n)$. As simultaneously controlling the FDR and power mimicking is out of reach, one may require that, at least, the FDR is controlled. Theorem 3.1, applied with $k_1 < k_2$, shows that controlling the FDR in the dense case has consequences on the relative type II risk in the sparse case. More precisely, for some $\epsilon > 0$, Condition (15) and $k_1 \leq k_2$ is satisfied for $k_2 = \log(1/\epsilon)n/\log(n)$ and $k_1 = \epsilon^{(c_2 \log(2/\alpha))^{-1}} \log(1/\epsilon) e \frac{n}{\log n}$ (for $\epsilon$ in a specific range), which entails the following result.

COROLLARY 3.3. *Consider the same numerical constants $c_1$–$c_3$ as in Theorem 3.1 above. Take any $\alpha \in (0, c_3)$, any $n \geq c_1$ and fix any $\epsilon \in (n^{-1/2}; (\alpha/2)^{c_2})$. Then for any procedure $R$ with $\mathbf{I}(R, k_2) \leq c_3$ for a sparsity $k_2 = \log(1/\epsilon)n/\log(n)$, we have $\mathbf{II}(R, k_1, \alpha/2) \geq 1/5$ for a sparsity $k_1 = \epsilon^{(c_2 \log(2/\alpha))^{-1}} \log(1/\epsilon)en/\log n$. In particular, if $n^{-1/4} < (\alpha/2)^{c_2}$, we have for any procedure $R$:*

- *if $\mathbf{I}(R, n/4) \leq \alpha$, then $\mathbf{II}(R, n^{1-\delta}e/4, \alpha/2) \geq 1/5$;*
- *if $\mathbf{II}(R, n^{1-\delta}e/4, \alpha/2) < 1/5$, then $\mathbf{I}(R, n/4) > c_3$,*

*where we let $\delta = 1/(4c_2 \log(2/\alpha)) > 0$.*

In plain words, the above corollary entails that a procedure $R$ controlling the FDR up to a sparsity $\log(1/\epsilon)\frac{n}{\log n}$ (that is of order larger than or equal to the boundary $n/\log(n)$ of Theorem 2.4), suffers from a power loss in a sparse setting where $n_1(P)$ is of order $\epsilon^{(c_2 \log(2/\alpha))^{-1}} \log(1/\epsilon)\frac{n}{\log n}$, for which AMO is theoretically possible (as stated in Theorem 3.2). As $\epsilon$ decreases, $R$ is assumed to control the FDR in denser settings and becomes overconservative in sparser settings. The case $\epsilon = n^{-1/4}$, requiring that the FDR is controlled at the nominal level up to a sparsity $n/4$ enforces a power loss in some "easy" settings where $n_1(P)/n$ is polynomially small. In other words, if we require FDR control in the dense regime, we will pay a high power price in the "easy" regime where AMO is achievable. Conversely, any AMO procedure in sparse regime violates the FDR control in the dense regime. Corollary 2.5 formalizes this fact with the plug-in BH procedure of Theorem 3.2.

**4. Known variance.** This section is dedicated to the simpler case where $\sigma(P)$ is known to the statistician, so that only the mean $\theta(P)$ has to be estimated. In this setting, it turns out that the boundary for AMO is $n/\log^{1/2}(n)$ instead of $n/\log(n)$.

THEOREM 4.1. *In the setting of Section 2 and according to Definition 2.2, the following hold*:

(i) *for a sparsity parameter $k_n \gg n/\log^{1/2}(n)$, there exists no (sequence of) procedure that is AMO*;

(ii) *for a sparsity parameter $k_n \ll n/\log^{1/2}(n)$, the scaling $(\widetilde{\theta}, \sigma(P))$ given by (14) is AMO.*

The upper bound (ii) is proved similar to the upper bound of Theorem 2.4, but with the weaker condition $k_n \log^{1/2}(n)/n = o(1)$. For this, one readily checks that Theorem 3.2 extends to the case where $\widehat{\sigma} = \sigma(P)$ up to replacing $\eta$ by $\eta = c_2 \log^{1/2}(n/\alpha)(\frac{n_1(P)+1}{n} + n^{-1/6})$ (and possibly modifying the constants $c_1$ and $c_2$). The proofs are exactly the same, except that Lemma S-3.2 has to be replaced by Lemma S-4.3. See Section S-4.3 for details. Let us additionally provide a heuristic explanation of this boundary. Roughly, the oracle BH procedure is equivalent to the plug-in BH procedure if the corrected observations $Y_i - \widehat{\theta}$ can be compared to the Gaussian quantiles $\overline{\Phi}^{-1}(\alpha\ell/(2n))$ in the same way as the variables $Y_i - \theta$ do. Hence, the plug-in operation will mimic the oracle if

$$|\hat{\theta} - \theta| \ll \min_{\ell} \{\overline{\Phi}^{-1}(\alpha\ell/(2n)) - \overline{\Phi}^{-1}(\alpha(\ell-1)/(2n))\} \asymp \frac{\alpha/n}{\phi(\overline{\Phi}^{-1}(\alpha/n))},$$

which leads to the condition $k/n \ll 1/\log^{1/2} n$, by using the standard properties on the Gaussian tail distribution (Section S-7) and the estimation rate of $\widetilde{\theta}$ (Section S-6.1).

In the remainder of this section, we focus on the impossibility results. We first establish in Theorem 4.2 the counterpart of Theorem 3.1, which will also prove Part (i) of Theorem 4.1. This lower bound is valid nonasymptotically and for arbitrary testing procedures. Next, we provide a sharper lower bound for plug-in procedures.

4.1. *Lower bound for a general procedure.*

THEOREM 4.2. *There exist numerical positive constants $c_1$–$c_5$ such that the following holds for all $n \geq c_1$ and any $\alpha \in (0,1)$. Consider two positive numbers $k_1 \leq k_2$ satisfying*

$$c_2 \frac{n \log(2/\alpha)}{\log^{1/2}(n)} \left\{ 1 + \log\left(\frac{k_2}{k_1}\right) \right\}^{1/2} \leq k_2 < n/2. \tag{20}$$

*Then for any multiple testing procedure $R$ satisfying*

$$FDR(P,R) \leq c_3, \quad \text{for any } P \in \mathcal{P} \text{ with } n_1(P) \leq k_2 \ ,$$

*there exists some $P \in \mathcal{P}$ with $n_1(P) \leq k_1$ such that we have*

$$\mathbb{P}_{Y\sim P}\big(|R(Y) \cap \mathcal{H}_1(P)| = 0\big) \geq 2/5;$$

$$\mathbb{P}_{Y\sim P}\left[ |BH^*_{\alpha/2} \cap \mathcal{H}_1(P)| \geq c_4 \alpha^{-1} \left\{ \frac{n}{\log n} \right\}^{1/2} \right] \geq 1 - e^{-c_5 \alpha^{-1} \{n/\log(n)\}^{1/2}} \geq 4/5. \tag{21}$$

*In particular, we have that $\mathbf{I}(R,k_2) \leq c_3$ implies $\mathbf{II}(R,k_1,\alpha/2) \geq 1/5$.*

This result is qualitatively similar to Theorem 3.1, up to the change the boundary condition (15) into (20). Taking $k_1 = k_2 = k_n \gg n/\log^{1/2}(n)$, we deduce part (i) of Theorem 4.1.

As in Section 3.3, we also deduce from Theorem 4.2 that no procedure $R$ can simultaneously control the FDR at the nominal level up to some $k_n \gg n/\log^{1/2}(n)$ while being also AMO for all sequences $k_n \ll n/\log^{1/2}(n)$.

COROLLARY 4.3. *Consider the same numerical constants $c_1$–$c_5$ as in Theorem 4.2 above. Take any $\alpha \in (0,c_3)$, any $n \geq c_1$ and fix any $\epsilon \in (n^{-1/4}; (\alpha/2)^{c_2} e^{-(c_2 \log(2/\alpha))^2})$. Then for any procedure $R$ with $\mathbf{I}(R,k_2) \leq c_3$ for a sparsity $k_2 = \log^{1/2}(1/\epsilon) \frac{n}{\log^{1/2} n}$, we have $\mathbf{II}(R,k_1,\alpha/2) \geq 1/5$ for a sparsity $k_1 = \epsilon^{(c_2 \log(2/\alpha))^{-2}} \log^{1/2}(1/\epsilon) e \frac{n}{\log^{1/2} n}$. In particular, if $n^{-1/16} < (\alpha/2)^{c_2} e^{-(c_2 \log(2/\alpha))^2}$, we have for any procedure $R$:*

- *if $\mathbf{I}(R,n/4) \leq \alpha$, then $\mathbf{II}(R, n^{1-\delta} e/4, \alpha/2) \geq 1/5$;*
- *if $\mathbf{II}(R, n^{1-\delta} e/4, \alpha/2) < 1/5$, then $\mathbf{I}(R,n/4) > c_3$,*

*where we let $\delta = 1/(16 c_2^2 \log^2(2/\alpha)) > 0$.*

4.2. *Lower bound for plug-in procedures.* In the previous section, we established an impossibility result for all multiple testing procedures $R$. In this section, we turn our attention to the special case of plug-in procedures $\mathrm{BH}_\alpha(\widehat{\theta}, \sigma(P))$ where $\widehat{\theta}$ is any estimator of $\theta(P)$.

THEOREM 4.4. *There exist positive numerical constants $c_1$–$c_3$ such that the following holds for all $\alpha \in (0,1)$, all $n \geq N(\alpha)$, any estimator $\widehat{\theta}$, and all $k$ satisfying*

$$c_1 \frac{n \log(2/\alpha)}{\log^{1/2}(n)} \leq k < \frac{n}{2}. \tag{22}$$

*There exists $P \in \mathcal{P}$ with $n_1(P) \leq k$ and an event $\Omega$ of probability higher than $1/2 - c_2/n$ such that, on $\Omega$, the plug-in procedure $BH_\alpha(Y; \widehat{\theta}, \sigma(P))$ satisfies both*

$$\begin{aligned} &\big|BH_\alpha\big(Y; \widehat{\theta}, \sigma(P)\big) \cap \mathcal{H}_0(P)\big| \geq 0.5 n^{3/4}; \\ &FDP\big(P, BH_\alpha\big(Y; \widehat{\theta}, \sigma(P)\big)\big) \geq \frac{1}{2 + c_3 n^{-1/5}}. \end{aligned} \tag{23}$$

This theorem enforces that no plug-in procedure $\mathrm{BH}_\alpha(Y;\widehat{\theta},\sigma(P))$ is able to control the FDR at the nominal level in dense settings ($k_n \gg n/\log^{1/2}(n)$). In fact, the FDP of plug-in procedure $\mathrm{BH}_\alpha(Y;\widehat{\theta},\sigma(P))$ is even shown to be at least of the order of $1/2$ with probability close to $1/2$. On the same event, the plug-in procedure $\mathrm{BH}_\alpha(Y;\widehat{\theta},\sigma(P))$ makes many false rejections. This statement is much stronger than the one of Theorem 4.2 (in the case $k_1 = k_2$).

In contrast to the previous lower bounds, the proof of Theorem 4.4 relies on a tighter control of the shifted $p$-value process and quantifies its impact on the BH threshold.

**5. Extension to general location models.** In this section, we generalize our approach to the case where the null distribution is not necessarily Gaussian. For simplicity, we focus here on the location model. Let $\mathcal{G}$ denote the collection of densities on $\mathbb{R}$ that are symmetric, continuous and nonincreasing on $\mathbb{R}_+$. Given any $g \in \mathcal{G}$, we extend the setting of Section 2, by now assuming that $P = \bigotimes_{i=1}^n P_i$ belongs to the collection $\mathcal{P}_g$ of all distributions on $\mathbb{R}^n$ satisfying

$$\text{there exists } \theta \in \mathbb{R} \text{ such that } \big|\{i \in \{1,\dots,n\} : P_i \text{ has density } g(\cdot-\theta)\}\big| > n/2. \tag{24}$$

In other words, we assume that there exists $\theta$ such that at least half of the $P_i$'s have for density $g(\cdot-\theta)$. Such $\theta$ is therefore uniquely defined from $P$, and we denote it again by $\theta(P)$. The testing problem becomes

$$H_{0,i} : \text{“}P_i \sim g\big(\cdot-\theta(P)\big)\text{”} \quad \text{against} \quad H_{1,i} : \text{“}P_i \nsim g\big(\cdot-\theta(P)\big)\text{”}, \quad \text{for all } 1 \le i \le n.$$

The rescaled $p$-values are now defined by

$$p_i(u) = 2\overline{G}\big(|Y_i - u|\big), \quad u \in \mathbb{R}, 1 \le i \le n, \tag{25}$$

where $\overline{G}(y) = \int_y^{+\infty} g(x)dx$, $y \in \mathbb{R}$. The oracle $p$-values are given by $p_i^* = 2\overline{G}(|Y_i - \theta(P)|)$, $1 \le i \le n$. The BH procedure at level $\alpha$ using $p$-values $p_i(u)$, $1 \le i \le m$, is denoted by $\mathrm{BH}_\alpha(u)$, whereas the oracle version is still denoted by $\mathrm{BH}_\alpha^*$.

For a given sparsity sequence $k_n \in [1, n/2)$, the sequence of procedures $R = (R_\alpha)_{\alpha\in(0,1)}$ is said to be AMO if there exists a positive sequence $\eta_n \to 0$ such that

$$\limsup_n \sup_{\alpha\in(1/n,1/2)} \big\{\mathbf{I}_g(R_\alpha, k_n) - \alpha\big\} \le 0; \tag{26}$$

$$\lim_n \sup_{\alpha\in(1/n,1/2)} \big\{\mathbf{II}_g\big(R_\alpha, k_n, \alpha(1-\eta_n)\big)\big\} = 0, \tag{27}$$

where $\mathbf{I}_g(\cdot)$ and $\mathbf{II}_g(\cdot)$ are respectively defined as (10) and (11), except that $\mathcal{P}$ is replaced by $\mathcal{P}_g$ therein. Similarly, for any sequence of estimators $\widehat{\theta}$ of $\theta(P)$, the rescaling $\widehat{\theta}$ is said to be AMO if $(R_\alpha)_{\alpha\in(0,1)} = (\mathrm{BH}_\alpha(\widehat{\theta}))_{\alpha\in(0,1)}$ is AMO.

5.1. *Lower bounds.* We first state two conditions under which (26) and (27) cannot hold together.

THEOREM 5.1. *Consider any $g \in \mathcal{G}$. There exist numerical positive constants $c_1$ and $c_2$ and a constant $c_g$ (only depending on $g$) such that the following holds for all $n > 2k \ge c_1$ and any $\alpha \in (0, 1/2)$. Assume that*

$$\frac{k}{nc_g} \ge \min_{t\in[\frac{\alpha}{2n};\frac{\alpha}{12}]}\left[\overline{G}^{-1}\left(\frac{t}{2}\right) - \overline{G}^{-1}\left(\frac{12t}{\alpha}\right)\right], \tag{28}$$

*and consider*

$$t_0 = \max\left\{t \in \left[\frac{\alpha}{2n};\frac{\alpha}{12}\right] \text{ s.t. } \overline{G}^{-1}\left(\frac{t}{2}\right) - \overline{G}^{-1}\left(\frac{12t}{\alpha}\right) \le \frac{k}{nc_g}\right\}.$$

*Then for any multiple testing procedure $R$ satisfying*

$$FDR(P, R) \le \frac{1}{5}, \quad \text{for all } P \in \mathcal{P}_g \text{ with } n_1(P) \le k \ ,$$

*there exists some $P \in \mathcal{P}_g$ with $n_1(P) \le k$ such that we have*

$$\begin{aligned} &\mathbb{P}_{Y\sim P}\big(|R(Y) \cap \mathcal{H}_1(P)| = 0\big) \ge 2/5; \\ &\mathbb{P}_{Y\sim P}\left[|BH^*_{\alpha/2} \cap \mathcal{H}_1(P)| \ge \frac{2nt_0}{\alpha}\right] \ge 1 - e^{-c_2\alpha^{-1}nt_0}. \end{aligned} \tag{29}$$

*In particular, $\mathbf{I}_g(R, k) \le 1/5$ implies $\mathbf{II}_g(R, k, \alpha/2) \ge 2/5 - e^{-c_2\alpha^{-1}nt_0}$.*

A consequence of Theorem 5.1 is that, for some sparsity sequence $k_n$ satisfying $n > 2k_n \ge c_1$ and (28) with $e^{-c_2nt_0} \le 1/5$, it is not possible to achieve any AMO procedure in the sense defined above. Interestingly, Condition (28) depends on the variations of $\overline{G}^{-1}(t)$ for small $t > 0$. Taking $g = \phi$ and $t = 1/\{n\log(n)\}^{1/2}$ and using the relations stated in Lemma S-7.2, we recover Theorem 4.2 (case $k_1 = k_2$) obtained in the Gaussian location model and the corresponding sharp condition $k \gg n/\{\log(n)\}^{1/2}$.

While Theorem 5.1 goes beyond the Gaussian case, it does not cover all possible distributions $G$. Consider the Laplace function $g(x) = e^{-|x|}/2$, so that $\overline{G}^{-1}(t) = \log(1/(2t))$. Then Condition (28) cannot be guaranteed even when $k/n$ is of the order of a constant. More generally, Theorem 5.1 is silent for any $g$ such that $\min_{t\in[\frac{\alpha}{2n};\frac{\alpha}{12}]}[\overline{G}^{-1}(\frac{t}{2}) - \overline{G}^{-1}(\frac{12t}{\alpha})]$ is of the order of a constant.

The next result is dedicated to this case. Remember that, when $k/n \ge 1/2$, $\theta(P)$ is not identifiable. We show that there exists a threshold $\pi_\alpha < 1/2$, such that deriving a AMO scaling is impossible when $k/n$ belongs to the region $(\pi_\alpha, 1/2)$. Markedly, $\pi_\alpha$ does not depend on $g$. For $\alpha \in (0, 1)$, it is defined by

$$\pi_\alpha = \frac{\sqrt{(1-\alpha)} - (1-\alpha)}{\alpha} \in (0, 1/2). \tag{30}$$

THEOREM 5.2. *Consider any $\alpha \in (0, 1)$ and $\pi_\alpha$ given by (30). There exist a positive constant $c_\alpha$ (only depending on $\alpha$) such that following holds for any $\overline{\pi} \in (\pi_\alpha, 1/2)$, any $g \in \mathcal{G}$ and $n$ larger than a constant depending on $\alpha$ and $\overline{\pi}$. For any multiple testing procedure $R$ satisfying*

$$FDR(P, R) \le 1/4, \quad \text{for all } P \in \mathcal{P}_g \text{ and } n_1(P) \le \overline{\pi}n \ ,$$

*there exists $P \in \mathcal{P}_g$ with $n_1(P) \le \overline{\pi}n$ such that we have*

$$\begin{aligned} &\mathbb{P}_{Y\sim P}\big(|R(Y)| = 0\big) \ge 1/3; \\ &\mathbb{P}_{Y\sim P}\left[|BH^*_{\alpha} \cap \mathcal{H}_1(P)| \ge n\frac{\overline{\pi}}{4}\right] \ge 1 - 10e^{-c_\alpha n(\overline{\pi}-\pi_\alpha)^2} \ge 3/4. \end{aligned} \tag{31}$$

*In particular, $\mathbf{I}_g(R, \overline{\pi}n) \le 1/4$ implies $\mathbf{II}_g(R, \overline{\pi}n, \alpha) \ge 1/12$.*

To illustrate the above result, take $\alpha \in (0, 1/4]$ and $\overline{\pi} \in (\pi_\alpha, 1/2)$. Applying the above result for $\alpha' < \alpha$ with $\pi_{\alpha'} < \overline{\pi}$, we obtain that, for any procedure $R$ with $\mathbf{I}_g(R, \overline{\pi}n) \le \alpha$, we have $\mathbf{II}_g(R, \overline{\pi}n, \alpha') \ge 1/12$. In particular, this shows that there exists no AMO scaling in the regime $k_n = n\overline{\pi}$, for $\overline{\pi} \in (\pi_\alpha, 1/2)$. In addition, this holds uniformly over all $g$ in the class $\mathcal{G}$.

5.2. *Upper bound.* Since any $g \in \mathcal{G}$ is symmetric and puts a mass around "0," $\theta(P)$ also corresponds to the median of the null distribution. We consider, again, $\widetilde{\theta} = Y_{\lceil n/2 \rceil}$ as the estimator of $\theta(P)$ and plug it into BH to build $\mathrm{BH}_\alpha(\widetilde{\theta})$. The following result holds.

THEOREM 5.3. *Consider any $g \in \mathcal{G}$. There exist constants $c_1(g), c_2(g) > 0$ only depending on $g$ such that the following holds for all $n \geq c_1(g)$ and $\alpha \in (0, 0.5)$. Consider an integer $k \leq 0.1n$ such that*

$$\eta = c_2(g)\big((k/n) \vee n^{-1/6}\big) \max_{t \in [0.95\alpha/n, \alpha]} \left\{ \frac{1}{\overline{G}^{-1}(t/2) - \overline{G}^{-1}(t)} \frac{g(\overline{G}^{-1}(t))}{g(\overline{G}^{-1}(t/2))} \right\} \leq 0.05. \tag{32}$$

*Then we have*

$$\mathbf{I}_g\big(BH_\alpha(\widetilde{\theta}), k\big) \leq \alpha(1+\eta) + e^{-n^{1/2}}; \tag{33}$$

$$\mathbf{II}_g\big(BH_\alpha(\widetilde{\theta}), k, \alpha(1-\eta)\big) \leq e^{-n^{1/2}}. \tag{34}$$

If we consider any asymptotic setting where $\eta$ in (32) converges to 0, then it follows from the above theorem that $\widetilde{\theta}$ is a AMO scaling.

Comparing the lower bound condition (28) to the upper bound condition (32), we observe that those are matching up to the term

$$\max_{t \in [0.95\alpha/n, \alpha]} \left\{ \frac{g(\overline{G}^{-1}(t))}{g(\overline{G}^{-1}(t/2))} \right\}.$$

Interestingly, the latter is of the order of a constant for the Subbotin–Laplace cases, which provides new sparsity boundaries, as we present in the next section.

5.3. *Application to Subbotin distributions.* We now apply our general results to the class of Subbotin distributions.

COROLLARY 5.4. *Consider the location Subbotin null model for which $g(x) = L_\zeta^{-1} e^{-|x|^\zeta/\zeta}$, for some fixed $\zeta > 1$ and the normalization constant $L_\zeta = 2\Gamma(1/\zeta)\zeta^{1/\zeta - 1}$. Then*:

(i) *for a sparsity parameter $k_n \gg n/(\log(n))^{1-1/\zeta}$, there exists no sequence of procedures that is AMO.*

(ii) *for a sparsity parameter $k_n \ll n/(\log(n))^{1-1/\zeta}$, the scaling $\widetilde{\theta} = Y_{\lceil n/2 \rceil}$ is AMO.*

COROLLARY 5.5. *Let us consider the Laplace density $g(x) = 0.5e^{-|x|}$. Then for a sparsity parameter $k_n \ll n$, the scaling $\widetilde{\theta} = Y_{\lceil n/2 \rceil}$ is AMO.*

5.4. *An additional result for the Laplace location model.* Our general theory implies that, in the Laplace location model, an AMO scaling is possible when $k_n \ll n$ (Corollary 5.5) and is impossible if $\liminf k_n/n > \pi_\alpha$ (Theorem 5.2). However, it is silent when $k_n/n$ converges to a small constant $\pi \in (0, 1)$. In this section, we investigate this constant proportion regime. We establish that AMO scaling is impossible and that one needs to incur a small but yet nonnegligible loss. Define, for any $\alpha \in (0, 1)$,

$$\pi_\alpha^* = \frac{1 - \sqrt{\alpha}}{2 - \sqrt{\alpha}} \in (\pi_\alpha, 1/2). \tag{35}$$

PROPOSITION 5.6 (Lower bound for the Laplace distribution). *There exists a positive and increasing function* $\zeta : (0, 1/2) \mapsto \mathbb{R}_+$ *with* $\lim_{1/2} \zeta = +\infty$ *such that the following holds for any* $\alpha \in (0, 1)$, *any* $\overline{\pi} < \pi_\alpha^*$ *and for any* $n$ *larger than a constant depending only on* $\alpha$ *and* $\overline{\pi}$. *For any procedure* $R$ *satisfying*

$$FDR[P, R] \le \alpha n_0(P)/n, \quad \text{for all } P \in \mathcal{P}_g \text{ with } n_1(P) \le \overline{\pi} n \ ,$$

*there exists a distribution* $P \in \mathcal{P}_g$ *with* $n_1(P) \le \overline{\pi} n$ *such that*

$$\mathbb{P}_{Y\sim P}\big[|BH_\alpha^*| > 0\big] - \mathbb{P}_{Y\sim P}\big[|R(Y)| > 0\big] \ge \alpha\zeta(\overline{\pi}) - c_{\overline{\pi}} n^{-1/3},$$

*where* $c_{\overline{\pi}}$ *only depends on* $\overline{\pi}$.

Recall that, for any distribution $P$, the FDR of $\mathrm{BH}_\alpha^*$ is equal to $\alpha n_0(P)/n$; see (9). Hence, the above proposition states that any procedure achieving the same FDR bound as the oracle procedure is strictly more conservative than the oracle, in the sense that $\mathbb{P}_{Y\sim P}[|\mathrm{BH}_\alpha^*| > 0, |R(Y)| = 0] \ge \alpha\zeta(\overline{\pi}) + o(1) > 0$. In addition, the amplitude of $\alpha\zeta(\overline{\pi})$ is increasing with $\overline{\pi}$, which is expected. Also, the assumption $\overline{\pi} < \pi_\alpha^*$ is technical. In particular, we can easily prove that, for larger $\overline{\pi}$, the result remains true by replacing $\zeta(\overline{\pi})$ by $\zeta(\overline{\pi} \wedge \pi_\alpha^*)$.

REMARK 5.7. On the feasibility side, we can show that in the regime where $n_1(P)/n$ converges to a small constant, the plug-in BH procedure at level $\alpha$ is yet not AMO, but is comparable to oracle BH procedures with modified nominal levels $\alpha' \ne \alpha$. Recall that $p_i(u) = 2\overline{G}(|Y_i - u|) = e^{-|Y_i-u|}$ and $p_i^* = e^{-|Y_i-\theta(P)|}$. As a consequence, given an estimator $\widehat{\theta}$, the ratio $p_i(\widehat{\theta})/p_i^*$ belongs to $[e^{-|\widehat{\theta}-\theta(P)|}; e^{|\widehat{\theta}-\theta(P)|}]$. Assuming that $\alpha e^{|\widehat{\theta}-\theta(P)|} < 1$, it follows from the definition of $\mathrm{BH}_\alpha(u)$ that

$$\mathrm{BH}^*_{\alpha e^{-|\widehat{\theta}-\theta(P)|}} \subset \mathrm{BH}_\alpha(\widehat{\theta}) \subset \mathrm{BH}^*_{\alpha e^{|\widehat{\theta}-\theta(P)|}}.$$

As a consequence, as long as $|\widehat{\theta} - \theta(P)| \le \log(1/\alpha)$, $\mathrm{BH}_\alpha(\widehat{\theta})$ is sandwiched between two oracle BH procedures with modified nominal levels. As an example, the median estimator $\widehat{\theta} = Y_{\lceil n/2 \rceil}$ satisfies $|\widehat{\theta} - \theta| \le c n_1(P)/n$ with high probability when $n_1(P)/n$ is small enough (see the proof of Theorem 5.3). As a consequence, with high probability, we have

$$\mathrm{BH}^*_{\alpha e^{-cn_1(P)/n}} \subset \mathrm{BH}_\alpha(\widehat{\theta}) \subset \mathrm{BH}^*_{\alpha e^{cn_1(P)/n}}.$$

Conversely, Proposition 5.6 entails that no multiple testing procedure can be sandwiched by oracle procedures with level $\alpha(1 - o(1))$ and $\alpha(1 + o(1))$.

**6. Confidence region for the null and applications.** In this section, we tackle the issue of building a confidence superset on the possible null distributions for $P$. This has not been considered yet in the literature to the best of our knowledge.

6.1. *A confidence region for the null.* We come back to the general Huber model described in Section 2, although we do not assume that the null distribution is necessarily Gaussian. That is, the observations $Y_i$, $1 \le i \le n$ are only assumed to be independent. Their respective c.d.f.'s are denoted by $F_i$, $1 \le i \le n$, and we let

$$\mathcal{F}_{0,k}(P) = \big\{F_0 \text{ c.d.f. } : \big|\{i \in \{1, \dots, n\} : F_i = F_0\}\big| \ge n - k\big\} \tag{36}$$

the set of all possible null c.d.f.'s for $P$, for some prescribed, known, maximum amount of contaminated marginals $k \in [0, n - 1]$. As before, if $k < n/2$, we have $n - k > n/2$ and the null distribution is an identifiable parameter of the model. Indeed, in that case, $F_0$ can be identified as the predominant element of the vector $(F_i)_{1\le i\le n}$. Equivalently, the set $\mathcal{F}_{0,k}(P)$ has cardinality at most 1. By contrast, if $k \ge n/2$, we have $n - k \le n/2$ and several nulls

may be possible for $P$, which entails that the null distribution is not necessary identifiable. However, this is not an obstacle to building a confidence region of level $(1-\alpha)$ for $F_0$; the region simply needs to contain all possible nulls $F_0$ of the set $\mathcal{F}_{0,k}(P)$.

Our inference is based on the empirical c.d.f. $\widehat{F}_n$ of the sample $Y$: the idea is that for any $F_0 \in \mathcal{F}_{0,k}(P)$, the function $(\widehat{F}_n - (1-k/n)F_0)/(k/n)$ should be close to be a c.d.f., which induces some constraints. This idea bears similarities with the existing literature, in particular with Genovese and Wasserman (2004), that derived confidence interval for the proportion of signal when the true null is known and uniform.

For some $\alpha \in (0,1)$, let us denote

$$\mathcal{F}_{1-\alpha}(Y) = \big\{F_0 \text{ c.d.f. } : \forall j \in \{0,\dots,n\}, \widehat{a}_n(j;F_0) \le \widehat{b}_n(j;F_0)\big\}; \tag{37}$$

$$\widehat{a}_n(j;F_0) = 0 \vee \frac{\max_{0\le \ell \le j}\{\ell/n - (1-k/n)F_0(Y_{(\ell)})\} - c_{n,\alpha}}{k/n}; \tag{38}$$

$$\widehat{b}_n(j;F_0) = 1 \wedge \frac{\min_{j\le \ell \le n}\{\ell/n - (1-k/n)F_0(Y^-_{(\ell+1)})\} + c_{n,\alpha}}{k/n}, \tag{39}$$

where $c_{n,\alpha} = \{-(1-k/n)\log(\alpha/2)/(2n)\}^{1/2}$, $F_0(y^-) = \lim_{x\to y^-} F_0(x)$ and $Y_{(1)} \le \cdots \le Y_{(n)}$ denote the order statistics ($Y_{(0)} = -\infty$, $Y_{(n+1)} = +\infty$) of the observed sample ($Y_i, 1 \le i \le n$). Note that all these quantities depend on $k$, but we have omitted it in the notation for short. The following result holds.

THEOREM 6.1. *For a given sparsity parameter $k \in [0, n-1]$, the region $\mathcal{F}_{1-\alpha}(Y)$ defined by* (37) *is a $(1-\alpha)$-confidence superset of the set $\mathcal{F}_{0,k}(P)$* (36) *of possible nulls for $P$ with at most $k$ contaminations, in the following sense*:

$$\mathbb{P}_{Y\sim P}\big(\mathcal{F}_{0,k}(P) \subset \mathcal{F}_{1-\alpha}(Y)\big) \ge 1-\alpha.$$

Compared to our previous findings, this result is less demanding on the sparsity parameter: it only assumes $n_1(P) \le k$ and not that $n_1(P)/n$ tends to zero at some rate. Also, it is nonparametric, and usable in combination with any possible modeling for the null.

6.2. *Application* 1: *A goodness-of-fit test for a given null distribution*. Considering any known c.d.f. $F_0$, the confidence region derived in Theorem 6.1 provides a way to test the null hypothesis $H_0'$: "$F_0 \in \mathcal{F}_{0,k}(P)$," that is, "$F_0$ is a possible null c.d.f. for $P$ with at most $k$ contaminations."

COROLLARY 6.2. *Consider $k \in [0, n-1]$ and the sets $\mathcal{F}_{1-\alpha}(Y)$* (37) *and $\mathcal{F}_{0,k}(P)$* (36). *Consider any c.d.f. $F_0$. Then the test rejecting the null hypothesis $H_0'$: "$F_0 \in \mathcal{F}_{0,k}(P)$" whenever $F_0 \notin \mathcal{F}_{1-\alpha}(Y)$, that is, if there exists $j \in \{0,\dots,n\}$, such that $\widehat{a}_n(j;F_0) > \widehat{b}_n(j;F_0)$, is of level $\alpha$.*

Since for any $F_0 \in \mathcal{F}_{0,k}(P)$, we have $\mathbb{P}(F_0 \notin \mathcal{F}_{1-\alpha}(Y)) \le 1 - \mathbb{P}(\mathcal{F}_{0,k}(P) \subset \mathcal{F}_{1-\alpha}(Y))$, the proof is straightforward from Theorem 6.1. In particular, this result can be used to test whether the *theoretical* null distribution $\mathcal{N}(0,1)$ is suitable for some data set, given some maximum proportion of contaminations, say $k/n = 10\%$. As shown in the vignette (Roquain and Verzelen (2021)), this test rejects $H_0'$ for many data sets.

Next, we can also build a goodness-of-fit test of level $\alpha$ for a family $(F_{0,\vartheta})_{\vartheta\in\Theta}$ of null c.d.f.'s. This corresponds to consider the null hypothesis $H_0''$: "$\exists \vartheta \in \Theta : F_{0,\vartheta} \in \mathcal{F}_{0,k}(P)$", that is, "in the family $(F_{0,\vartheta})_{\vartheta\in\Theta}$ there is at least a possible null c.d.f. for $P$ with at most $k$ contaminations."

COROLLARY 6.3. *Consider $k \in [0, n-1]$ and the sets $\mathcal{F}_{1-\alpha}(Y)$ (37) and $\mathcal{F}_{0,k}(P)$ (36). Consider any family of c.d.f.'s $(F_{0,\vartheta})_{\vartheta\in\Theta}$. Then the test rejecting the null hypothesis $H_0''$: "$\exists\vartheta \in \Theta : F_{0,\vartheta} \in \mathcal{F}_{0,k}(P)$" whenever $\forall\vartheta \in \Theta$, $F_{0,\vartheta} \notin \mathcal{F}_{1-\alpha}(Y)$, that is, when the confidence region does not contain any null distribution of the family, is of level $\alpha$.*

Since for any $\vartheta_0 \in \Theta$ with $F_{0,\vartheta_0} \in \mathcal{F}_{0,k}(P)$, we have $\mathbb{P}(\forall\vartheta \in \Theta, F_{0,\vartheta} \notin \mathcal{F}_{1-\alpha}(Y)) \le \mathbb{P}(F_{0,\vartheta_0} \notin \mathcal{F}_{1-\alpha}(Y)) \le 1 - \mathbb{P}(\mathcal{F}_{0,k}(P) \subset \mathcal{F}_{1-\alpha}(Y))$, the proof is straightforward from Theorem 6.1. As a typical instance, this can be used to build a goodness-of-fit test for the family of Gaussian null distribution with arbitrary scaling. Interestingly, this test never rejects this null hypothesis for the classical data sets used in the vignette, which shows that considering a Gaussian null with unknown scaling can be suitable for these data.

The two aforementioned tests provide a way to validate Efron's paradigm who discarded the theoretical null $\mathcal{N}(0, 1)$, while still using empirical Gaussian nulls.

6.3. *Application* 2: *A reliability indicator for empirical null procedures*. For simplicity, let us focus on the case of Gaussian null distributions as in the setting of Section 2, for which $k < n/2$. The confidence region derived in Theorem 6.1 induces a confidence region for the true scaling $(\theta(P), \sigma(P))$ given by

$$\begin{aligned}\mathcal{S}_{k,\alpha} = \big\{&(\theta,\sigma) \in \mathbb{R} \times (0,\infty) : \\ &\forall j \in \{0,\dots,n\}, \widehat{a}_n\big(j; \Phi\big((\cdot-\theta)/\sigma\big)\big) \le \widehat{b}_n\big(j; \Phi\big((\cdot-\theta)/\sigma\big)\big)\big\}.\end{aligned} \tag{40}$$

COROLLARY 6.4. *Consider the setting of Section 2. Provided that $n_1(P) \le k$, the set $\mathcal{S}_{k,\alpha}$ is a $(1-\alpha)$-confidence region for the true scaling $(\theta(P), \sigma(P))$. In particular, with probability at least $1-\alpha$, the oracle BH procedure is one of the procedures $BH_\alpha(u, s)$, $(u, s) \in \mathcal{S}_{k,\alpha}$.*

Corollary 6.4 can be used to build a "diagnostic graph" for the oracle BH procedure, as depicted in Figure 5 for different data sets, $\alpha = 0.1$, $k/n = 0.1$ and $n = 10{,}000$. The colored areas correspond to points $(\theta, \sigma)$ that belong to the region $\mathcal{S}_{k,\alpha}$ (while other points are left in white color). In addition, in each of these points $(\theta, \sigma)$, we have depicted the number of rejections of the plug-in BH procedure using the corresponding null $\mathcal{N}(\theta, \sigma^2)$. The rejection number is used for the sake of readability, but we could have reported the rejection set for further details. From Corollary 6.4, we know that, with probability at least 90%, the oracle procedure has a rejection number equals to one of these numbers. As a result, we propose the following practical interpretation for these diagnostic graphs: when all these rejection numbers are of the same order and larger than $r \ge 1$, this means that the BH oracle procedure should make at least $r$ rejections (the user could also look more closely to the corresponding rejection sets to validate these rejections). On the other hand, if one of these rejection numbers is 0, it means that the BH oracle procedure possibly makes no rejection. In this case, our recommendation is to return no discovery.

Let us apply this general diagnostic to the various examples depicted in Figure 5. The two top data sets correspond to simulated observations (only 1 run each time), for which the true underlying distribution is known. First, in the "lower bound" setting, the null is $\mathcal{N}(0, \sigma^2)$ and the alternative density is $\frac{0.9}{0.1}[\phi - \phi(\cdot/\sigma)/\sigma]_+$ ($\sigma \approx 1.26$). This distribution arises in the proof of our impossibility result; see Section 3.1. The region is wide in this setting and contains the true null $\mathcal{N}(0, \sigma^2)$ (with no rejection) but also the erroneous null $\mathcal{N}(0, 1)$ (with some rejections). Since the confidence set contains the empty rejection set, our recommendation leads to no discovery. This is indeed correct because the oracle procedure makes no rejection in that case. Second, in the "Gaussian alternative" setting, the null is $\mathcal{N}(0, 1)$ and the alternative

is $\mathcal{N}(3,1)$. The region looks much more narrowed and contains only scaling for which the corresponding plug-in BH procedures make many findings (at least 532). Our recommendation is thus to declare these rejections as real discoveries (possibly after checking that they also belong to the other rejected sets with 556, 687 and 721 rejections). Doing so, we are able to almost mimic the rejection set of the oracle (here, the one with 721 rejections), even if we are in the impossibility regime where $k \gg n/\log(n)$. Here, the confidence region is able to "guess" that the underlying distribution is favorable enough to make the problem of estimating the null possible. This illustrates that the confidence region is "distribution-dependent."

Next, the bottom panels in Figure 5 display the region for two classical real data sets. For the leukemia data set, the region contains scaling for which the plug-in BH procedure makes no rejection (this turns out to be true for all of them). Hence, we recommend no rejection for this data set. By contrast, for the HIV data set, there is evidence that the oracle BH procedure is able to reject at least 46 variables.

Finally, while our diagnostic graph is used here with plug-in BH procedures, it could be used in combination with any other procedures based on a prescribed null as, for example, local FDR methods. We also underline that the above analysis easily adapts to any parametric null family $(F_{0,\vartheta})_{\vartheta\in\Theta}$, not necessarily Gaussian. The $(1-\alpha)$-confidence region for the true null parameter(s) $\vartheta(P)$ is then simply replaced by

$$\Theta_{k,\alpha} = \big\{\vartheta \in \Theta : \forall j \in \{0,\dots,n\}, \widehat{a}_n(j; F_{0,\vartheta}) \le \widehat{b}_n(j; F_{0,\vartheta})\big\},$$

the rest of the analysis being unchanged.

## 7. Discussion.

7.1. *Conclusion.* Elaborating upon Efron's problem, we have presented a general theory to assess whether one can estimate the null and plug-it into a BH procedure, while keeping the FDP and TDP similar to the oracle BH procedure. As expected, the sparsity parameter $k$, that is, the upper bound on the number of alternatives, plays a central role. The obtained sparsity boundaries were shown to depend (i) on the fact that the null variance is known or not and (ii) on the variations of the quantile function of the null distribution. We eventually went beyond the worst case analysis by designing a confidence region for the null distribution, which only requires the knowledge of an upper bound on the signal proportion. This allowed us to define goodness-of-fit tests for null distributions and diagnostic graphs for assessing the reliability of empirical null procedures. This is illustrated in detail in the vignette (Roquain and Verzelen (2021)). We now further discuss several aspects of our work in light of the related literature.

7.2. *Assumptions on the model.* Our model relies on several assumptions. First, the null distribution is assumed to belong to a Gaussian family (or to a location family in Section 5). If this assumption does not hold, and the null distribution is indeed far from being Gaussian, then our analysis is not correct and the FDR control is likely to fail. Nevertheless, the new null distribution goodness-of-fit test described next to Corollary 6.3 can be used to validate if the Gaussian null family is plausible. Before plugging the estimated null into some FDR controlling procedure, we therefore recommend to use our test as a "sanity check" (see the companion vignette for a concrete example).

Second, we assume that the observations $Y_i$, $1 \le i \le n$ are independent. What happens if they are some dependencies between the $Y_i$'s? While this question is challenging in general, we discuss here two possible answers: first, our theoretical results being extensively based on concentration inequalities (Bernstein, DKW), a generalization would need to use analogues in the dependent case, for instance using some kind of weak dependence between the $Y_i$'s; see Merlevède, Peligrad and Rio (2009), Naaman (2021). While our numerical experiments suggest that the results are robust against weak dependences (block-correlation structure, Section S-1), supporting this fact with a theoretical statement is an

interesting avenue for future research. Second, our framework naturally handles the following specific strongly dependent case: consider the Gaussian equicorrelated structure $Y_i = \mu_i + \rho^{1/2}U + (1-\rho)^{1/2}\xi_i$, $1 \le i \le n$, where $U, \xi_1, \dots, \xi_n$ are all i.i.d. $\mathcal{N}(0,1)$, $\rho \in (0,1)$ and $\mu_i$ is the signal (null if and only if we are under the null). Then, *conditionally on the factor* $U$, we have $Y \sim N(\mu + \rho^{1/2}U\mathbf{1}, (1-\rho)I_n)$, with $\mathbf{1} = (1, \dots, 1) \in \mathbb{R}^n$. Hence, letting $\theta = \rho^{1/2}U$ and $\sigma = (1-\rho)^{1/2}$, we fall into the situation described by our model, conditionally on $U$. This is in line with the well-known fact that unobserved covariates can strongly affect the null distribution, which was one important original motivation for estimating the *conditional null*; see Efron (2008). While our theory does not cover the case of general covariates, it encompasses this particular case, which is an important proof of concept in our eyes.

Finally, let us note that when each $Y_i$ comes from a (Gaussian-normalized) test statistic computed over several individuals, a classical alternative approach is to learn the (conditional) null distribution by permuting the individuals. However, this method is not able to properly learn the (conditional) null, as originally discussed in Efron (2008) and further illustrated in the vignette (Roquain and Verzelen (2021)), so does not correctly account for the dependence in factor models as the one presented above.

7.3. *Connections with Huber's contamination model and two-group models.* Originally, Huber's contamination model was introduced in robust statistics as a mixture model with density $h = (1-\pi)\phi_{\theta,\sigma} + \pi f$ where $\pi \in [0, 1/2)$, $\phi_{\theta,\sigma}$ stands for the density of a $\mathcal{N}(\theta, \sigma^2)$ and $f$ is an arbitrary density. When sampling according to this model, one observes a proportion close to $(1-\pi)$ of data sampled from the normal distribution and a proportion close to $\pi$ of contaminated data. In our framework, the contaminated data account for the false null hypotheses and noncontaminated data for the true null hypotheses. Hence, Huber's contamination model interprets as a specific random instance of our model (1). Indeed, one can sample $n$ observations according to $h$ by first generating $n$ Bernoulli random variables $Z_i \in \{0,1\}$ with parameter $\pi$. Next, if $Z_i = 0$, then $Y_i$ is sampled according to $\phi_{\theta,\sigma}$, whereas if, $Z_i = 1$, $Y_i$ is sampled according to $f$. As a consequence, conditionally to the $Z_i$'s, the distribution $P$ of $Y$ satisfies $n_1(P) = \sum_{i=1}^n Z_i$ and $(\theta(P), \sigma(P)) = (\theta, \sigma)$, at least when $\sum_{i=1}^n Z_i < n/2$ and all false null distributions $P_i$ are identical and have density $f$. This random instance of our model (1) is central for proving impossibility results in Section 3.1.

In the multiple testing terminology, Huber's contamination model can be interpreted as a two-group model where the null distribution is a normal distribution with unknown parameters whereas the alternative distribution is let completely arbitrary. Many recent contributions have been devoted to specific forms of the two-group model, but the authors generally assume that the data are sampled as a mixture of chosen parametric families (Cai and Jin (2010), Jin and Cai (2007), Sun and Stephens (2018)). Hence, our model includes a semiparametric version of the two-group model by letting the alternative distribution to be arbitrary. It has the virtue of being in line with the original empirical null framework of Efron (2008) while being rich enough to ensure strong FDR controls as usually required in the multiple testing field (Dickhaus (2014)).

7.4. *Choice of BH as an oracle procedure—comparison to local FDR-type procedure.* In this work, we have chosen the oracle BH as the reference procedure. When the null and alternative distributions are known, more powerful (and optimal) procedures have been developed. For instance, approaches based on the local FDR (Sun and Cai (2007)) and the maximin approach (Rosset et al. (2020)) outperform oracle BH while controlling the marginal FDR and FDR, respectively. However, estimating such oracle procedures is only possible in specific mixture model configurations. To quote (Rosset et al. (2020)), "care has to be taken

in proper estimation of the mixture parameters in order to avoid an unacceptable inflation of the FDR level." In contrast, BH enjoys a strong control of the FDR for arbitrary alternative distributions. Besides, BH procedure is known to enjoy optimality power properties for specific alternatives including sparse one-sided Gaussian alternatives; see Arias-Castro and Chen (2017), Rabinovich et al. (2020).

As an aside, our impossibility results apply to any empirical null testing procedure. In particular, they entail impossibility results for local FDR-type ones: if the sparsity parameter is above the boundary, any local FDR-type procedure either violates the FDR control, or is less powerful than the oracle BH procedure. This is also illustrated empirically in our numerical experiments; see Section S-1.

7.5. *Choosing the empirical null method in practice.* Our analysis entails that a BH procedure with plugged robust estimators of the scaling parameters achieves the phase transition and is therefore "optimal" with that respect. Nevertheless, this phase transition is of a "worst-case" nature in accordance with our model that lets the alternative distribution to be arbitrary. In some situations, the distribution of the alternatives may be more favorable so that the null distribution can be better learned than suggested by the worst-case rates. If it turns out that all the alternatives follow an identical Gaussian distribution as prescribed by the Gaussian two-group model (Sun and Cai (2007)), then local FDR-type procedures are certainly more suited. Hence, we underline that our plug-in procedure with robust estimators of the parameters does not always outperform other procedures. Choosing among these methods depends on some unknown characteristic of the alternative distributions and exceeds the scope of this paper.

Still, regardless of the empirical null method chosen by the statistician, we recommend, in practice, to rely on the diagnostic graphs introduced in Section 6 to validate the inference. Indeed, as long as the empirical null method is of plug-in type (which is typically the case for the local FDR methods), these graphs can assess the validity of the plug-in method.

7.6. *Open problems.* This work paves the way for several extensions. First, one direction is to investigate the sparsity boundary when the model is reduced, for example, by considering more constrained alternatives. A first hint has been given for one-sided alternatives in Carpentier et al. (2021), where both a uniform FDR control and power results can be achieved in dense settings, for example, $k = n/2$ (say), which is markedly different from what we obtained here. In future work, many more structured setting can be considered, for example, decreasing alternative densities, temporal/spatial structure on the signal, and so on. Second, the problem could also be made more difficult by considering a more complex model for the null, for instance, by dropping the assumption that $g$ is known, but assuming instead that it belongs to some parametric or nonparametric class. Each of these settings should come with a new phase transition that is worth investigating. Finally, the new proposed confidence region is based on the DKW inequality, which is not always the most accurate tool. Reducing the size of the confidence region is an interesting avenue for future investigations.

**Acknowledgments.** We are grateful to Ery Arias-Castro, David Mary, to anonymous referees and to the editorial team for insightful comments that helped us to improve the presentation of the manuscript.

**Funding.** This work has been supported by ANR-16-CE40-0019 (SansSouci), ANR-17-CE40-0001 (BASICS), ANR-21-CE23-0035 (ASCAI) and by the GDR ISIS through the "projets exploratoires" program (project TASTY).

## SUPPLEMENTARY MATERIAL

**Supplement to "False discovery rate control with unknown null distribution: Is it possible to mimic the oracle?"** (DOI: 10.1214/21-AOS2141SUPP; .pdf). This supplement contains numerical experiments, proofs of our results, technical lemmas and auxiliary results for the main paper.

## REFERENCES


AMAR, D., SHAMIR, R. and YEKUTIELI, D. (2017). Extracting replicable associations across multiple studies: Empirical Bayes algorithms for controlling the false discovery rate. *PLoS Comput. Biol.* **13** e1005700. https://doi.org/10.1371/journal.pcbi.1005700

ARIAS-CASTRO, E. and CHEN, S. (2017). Distribution-free multiple testing. *Electron. J. Stat.* **11** 1983–2001. MR3651021 https://doi.org/10.1214/17-EJS1277

AZRIEL, D. and SCHWARTZMAN, A. (2015). The empirical distribution of a large number of correlated normal variables. *J. Amer. Statist. Assoc.* **110** 1217–1228. MR3420696 https://doi.org/10.1080/01621459.2014.958156

BARBER, R. F. and CANDÈS, E. J. (2015). Controlling the false discovery rate via knockoffs. *Ann. Statist.* **43** 2055–2085. MR3375876 https://doi.org/10.1214/15-AOS1337

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

BENJAMINI, Y. and YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29** 1165–1188. MR1869245 https://doi.org/10.1214/aos/1013699998

BLANCHARD, G., LEE, G. and SCOTT, C. (2010). Semi-supervised novelty detection. *J. Mach. Learn. Res.* **11** 2973–3009. MR2746544

BOURGON, R., GENTLEMAN, R. and HUBER, W. (2010). Independent filtering increases detection power for high-throughput experiments. *Proc. Natl. Acad. Sci. USA* **107** 9546–9551. https://doi.org/10.1073/pnas.0914005107

CAI, T. T. and JIN, J. (2010). Optimal rates of convergence for estimating the null density and proportion of nonnull effects in large-scale multiple testing. *Ann. Statist.* **38** 100–145. MR2589318 https://doi.org/10.1214/09-AOS696

CAI, T. T. and SUN, W. (2009). Simultaneous testing of grouped hypotheses: Finding needles in multiple haystacks. *J. Amer. Statist. Assoc.* **104** 1467–1481. MR2597000 https://doi.org/10.1198/jasa.2009.tm08415

CAI, T. T., SUN, W. and WANG, W. (2019). Covariate-assisted ranking and screening for large-scale two-sample inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 187–234. MR3928141

CARPENTIER, A., DELATTRE, S., ROQUAIN, E. and VERZELEN, N. (2021). Estimating minimum effect with outlier selection. *Ann. Statist.* **49** 272–294. MR4206678 https://doi.org/10.1214/20-AOS1956

CASTILLO, I. and ROQUAIN, É. (2020). On spike and slab empirical Bayes multiple testing. *Ann. Statist.* **48** 2548–2574. MR4152112 https://doi.org/10.1214/19-AOS1897

CHEN, M., GAO, C. and REN, Z. (2018). Robust covariance and scatter matrix estimation under Huber's contamination model. *Ann. Statist.* **46** 1932–1960. MR3845006 https://doi.org/10.1214/17-AOS1607

CONSORTIUM, E. P. et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the encode pilot project. *Nature* **447** 799.

DICKHAUS, T. (2014). *Simultaneous Statistical Inference*. Springer, Heidelberg. With applications in the life sciences. MR3184277 https://doi.org/10.1007/978-3-642-45182-9

DONOHO, D. and JIN, J. (2006). Asymptotic minimaxity of false discovery rate thresholding for sparse exponential data. *Ann. Statist.* **34** 2980–3018. MR2329475 https://doi.org/10.1214/009053606000000920

EFRON, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *J. Amer. Statist. Assoc.* **99** 96–104. MR2054289 https://doi.org/10.1198/016214504000000089

EFRON, B. (2007). Doing thousands of hypothesis tests at the same time. *Metron* **LXV** 3–21.

EFRON, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.* **23** 1–22. MR2431866 https://doi.org/10.1214/07-STS236

EFRON, B. (2009). Empirical Bayes estimates for large-scale prediction problems. *J. Amer. Statist. Assoc.* **104** 1015–1028. MR2562003 https://doi.org/10.1198/jasa.2009.tm08523

EFRON, B., TIBSHIRANI, R., STOREY, J. D. and TUSHER, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.* **96** 1151–1160. MR1946571 https://doi.org/10.1198/016214501753382129

FAN, J. and HAN, X. (2017). Estimation of the false discovery proportion with unknown dependence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1143–1164. MR3689312 https://doi.org/10.1111/rssb.12204

FAN, J., HAN, X. and GU, W. (2012). Estimating false discovery proportion under arbitrary covariance dependence. *J. Amer. Statist. Assoc.* **107** 1019–1035. MR3010887 https://doi.org/10.1080/01621459.2012.720478

FERREIRA, J. A. and ZWINDERMAN, A. H. (2006). On the Benjamini–Hochberg method. *Ann. Statist.* **34** 1827–1849. MR2283719 https://doi.org/10.1214/009053606000000425

FRIGUET, C., KLOAREG, M. and CAUSEUR, D. (2009). A factor model approach to multiple testing under dependence. *J. Amer. Statist. Assoc.* **104** 1406–1415. MR2750571 https://doi.org/10.1198/jasa.2009.tm08332

GENOVESE, C. and WASSERMAN, L. (2004). A stochastic process approach to false discovery control. *Ann. Statist.* **32** 1035–1061. MR2065197 https://doi.org/10.1214/009053604000000283

GHOSH, D. (2012). Incorporating the empirical null hypothesis into the Benjamini–Hochberg procedure. *Stat. Appl. Genet. Mol. Biol.* **11** 11. MR2958610 https://doi.org/10.1515/1544-6115.1735

GOLUB, T. R., SLONIM, D. K., TAMAYO, P., HUARD, C., GAASENBEEK, M., MESIROV, J. P., COLLER, H., LOH, M. L., DOWNING, J. R. et al. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286** 531–537.

HEDENFALK, I., DUGGAN, D., CHEN, Y., RADMACHER, M., BITTNER, M., SIMON, R., MELTZER, P., GUSTERSON, B., ESTELLER, M. et al. (2001). Gene-expression profiles in hereditary breast cancer. *N. Engl. J. Med.* **344** 539–548.

HELLER, R. and YEKUTIELI, D. (2014). Replicability analysis for genome-wide association studies. *Ann. Appl. Stat.* **8** 481–498. MR3191999 https://doi.org/10.1214/13-AOAS697

HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Stat.* **35** 73–101. MR0161415 https://doi.org/10.1214/aoms/1177703732

HUBER, P. J. (2011). Robust statistics. In *International Encyclopedia of Statistical Science* 1248–1251. Springer, Berlin.

JIANG, W. and YU, W. (2016). Controlling the joint local false discovery rate is more powerful than meta-analysis methods in joint analysis of summary statistics from multiple genome-wide association studies. *Bioinformatics* **33** 500–507.

JIN, J. and CAI, T. T. (2007). Estimating the null and the proportional of nonnull effects in large-scale multiple comparisons. *J. Amer. Statist. Assoc.* **102** 495–506. MR2325113 https://doi.org/10.1198/016214507000000167

JING, B.-Y., KONG, X.-B. and ZHOU, W. (2014). FDR control in multiple testing under non-normality. *Statist. Sinica* **24** 1879–1899. MR3308667

LEE, N., KIM, A.-Y., PARK, C.-H. and KIM, S.-H. (2016). An improvement on local fdr analysis applied to functional mri data. *J. Neurosci. Methods* **267** 115–125.

LEEK, J. T. and STOREY, J. D. (2008). A general framework for multiple testing dependence. *Proc. Natl. Acad. Sci. USA* **105** 18718–18723.

MARY, D. and ROQUAIN, E. (2021). Semi-supervised multiple testing. arXiv preprint. Available at arXiv:2106.13501.

MERLEVÈDE, F., PELIGRAD, M. and RIO, E. (2009). Bernstein inequality and moderate deviations under strong mixing conditions. In *High Dimensional Probability V*: *The Luminy Volume*. *Inst. Math. Stat.* (*IMS*) *Collect.* **5** 273–292. IMS, Beachwood, OH. MR2797953 https://doi.org/10.1214/09-IMSCOLL518

MILLER, C. J., GENOVESE, C., NICHOL, R. C., WASSERMAN, L., CONNOLLY, A., REICHART, D., HOPKINS, A., SCHNEIDER, J. and MOORE, A. (2001). Controlling the false-discovery rate in astrophysical data analysis. *Astron. J.* **122** 3492–3505.

MURALIDHARAN, O. (2010). An empirical Bayes mixture method for effect size and false discovery rate estimation. *Ann. Appl. Stat.* **4** 422–438. MR2758178 https://doi.org/10.1214/09-AOAS276

NAAMAN, M. (2021). On the tight constant in the multivariate Dvoretzky–Kiefer–Wolfowitz inequality. *Statist. Probab. Lett.* **173** 109088. MR4233456 https://doi.org/10.1016/j.spl.2021.109088

NGUYEN, V. H. and MATIAS, C. (2014). On efficient estimators of the proportion of true null hypotheses in a multiple testing setup. *Scand. J. Stat.* **41** 1167–1194. MR3277044 https://doi.org/10.1111/sjos.12091

PADILLA, M. and BICKEL, D. R. (2012). Estimators of the local false discovery rate designed for small numbers of tests. *Stat. Appl. Genet. Mol. Biol.* **11** 4. MR2990984 https://doi.org/10.1515/1544-6115.1807

POLLARD, K. S. and VAN DER LAAN, M. J. (2004). Choice of a null distribution in resampling-based multiple testing. *J. Statist. Plann. Inference* **125** 85–100. MR2086890 https://doi.org/10.1016/j.jspi.2003.07.019

RABINOVICH, M., RAMDAS, A., JORDAN, M. I. and WAINWRIGHT, M. J. (2020). Optimal rates and trade-offs in multiple testing. *Statist. Sinica* **30** 741–762. MR4214160

REBAFKA, T., ROQUAIN, E. and VILLERS, F. (2019). Graph inference with clustering and false discovery rate control. arXiv preprint. Available at arXiv:1907.10176.

ROQUAIN, E. and VAN DE WIEL, M. A. (2009). Optimal weighting for false discovery rate control. *Electron. J. Stat.* **3** 678–711. MR2521216 https://doi.org/10.1214/09-EJS430

ROQUAIN, E. and VERZELEN, N. (2021). False discovery rate control with unknown null distribution: Illustrations on real data sets. Available at https://github.com/eroquain/empiricalnull/blob/main/vignette.pdf.

ROQUAIN, E. and VERZELEN, N. (2022). Supplement to "False discovery rate control with unknown null distribution: is it possible to mimic the oracle?." https://doi.org/10.1214/21-AOS2141SUPP

ROSSET, S., HELLER, R., PAINSKY, A. and AHARONI, E. (2020). Optimal and maximin procedures for multiple testing problems.

SCHWARTZMAN, A. (2008). Empirical null and false discovery rate inference for exponential families. *Ann. Appl. Stat.* **2** 1332–1359. MR2655662 https://doi.org/10.1214/08-AOAS184

SCHWARTZMAN, A. (2010). Comment: "Correlated $z$-values and the accuracy of large-scale statistical estimates". *J. Amer. Statist. Assoc.* **105** 1059–1063. MR2752600 https://doi.org/10.1198/jasa.2010.tm10237

STEPHENS, M. (2017). False discovery rates: A new deal. *Biostatistics* **18** 275–294. MR3824755 https://doi.org/10.1093/biostatistics/kxw041

SULIS, S., MARY, D. and BIGOT, L. (2017). A study of periodograms standardized using training datasets and application to exoplanet detection. *IEEE Trans. Signal Process.* **65** 2136–2150. MR3608115 https://doi.org/10.1109/TSP.2017.2652391

SUN, W. and CAI, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *J. Amer. Statist. Assoc.* **102** 901–912. MR2411657 https://doi.org/10.1198/016214507000000545

SUN, W. and CAI, T. T. (2009). Large-scale multiple testing under dependence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 393–424. MR2649603 https://doi.org/10.1111/j.1467-9868.2008.00694.x

SUN, L. and STEPHENS, M. (2018). Solving the empirical bayes normal means problem with correlated noise. arXiv preprint. Available at arXiv:1812.07488.

SZALAY, A. S., CONNOLLY, A. J. and SZOKOLY, G. P. (1999). Simultaneous multicolor detection of faint galaxies in the hubble deep field. *Astron. J.* **117** 68–74.

TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation. Springer Series in Statistics*. Springer, New York. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats. MR2724359 https://doi.org/10.1007/b13794

VAN'T WOUT, A. B., LEHRMAN, G. K., MIKHEEVA, S. A., O'KEEFFE, G. C., KATZE, M. G., BUMGARNER, R. E., GEISS, G. K. and MULLINS, J. I. (2003). Cellular gene expression upon human immunodeficiency virus type 1 infection of cd4+-t-cell lines. *J. Virol.* **77** 1392–1402.

ZABLOCKI, R. W., SCHORK, A. J., LEVINE, R. A., ANDREASSEN, O. A., DALE, A. M. and THOMPSON, W. K. (2014). Covariate-modulated local false discovery rate for genome-wide association studies. *Bioinformatics* **30** 2098–2104.