Electronic Journal of Statistics
Vol. 16 (2022) 4926–4981
ISSN: 1935-7524
https://doi.org/10.1214/22-EJS2050

Semi-supervised multiple testing*

David Mary

Université Côte d'Azur, Observatoire de la Côte d'Azur, CNRS, Laboratoire Lagrange, Bd de l'Observatoire, CS 34229, 06304, Nice cedex 4, France
e-mail: david.mary@oca.eu

Etienne Roquain

Laboratoire de Probabilités, Statistique et Modélisation, Sorbonne Université, Université de Paris & CNRS, 4, place Jussieu, 75005 Paris, France
e-mail: etienne.roquain@upmc.fr

Abstract: An important limitation of standard multiple testing procedures is that the null distribution should be known. Here, we consider a null distribution-free approach for multiple testing in the following semi-supervised setting: the user does not know the null distribution, but has at hand a sample drawn from this null distribution. In practical situations, this null training sample (NTS) can come from previous experiments, from a part of the data under test, from specific simulations, or from a sampling process. In this work, we present theoretical results that handle such a framework, with a focus on the false discovery rate (FDR) control and the Benjamini-Hochberg (BH) procedure. First, we provide upper and lower bounds for the FDR of the BH procedure based on empirical p-values, called here the semi-supervised BH procedure. These bounds match when $\alpha(n+1)/m$ is an integer, where n is the NTS sample size and m is the number of tests. Second, we give a power analysis for that procedure suggesting that it mimics an oracle power when n is sufficiently large in front of m; namely $n \gtrsim m/(\max(1,k))$, where k denotes the number of "detectable" alternatives. Third, to complete the picture, we also present a negative result that evidences an intrinsic transition phase to the general semi-supervised multiple testing problem and shows that the semi-supervised BH method is optimal in the sense that its performance boundary follows this transition phase. Our theoretical properties are supported by numerical experiments, which also show that the delineated boundary is of correct order without further tuning any constant. Finally, we demonstrate that our work provides a theoretical ground for standard practice in astronomical data analysis, and in particular for the procedure proposed in Mary et al. (2020) for galaxy detection.

Keywords and phrases: Multiple testing, BH procedure, empirical p-values, false discovery rate, knockoff, LASSO, phase transition, galaxy detection.

Received November 2021.

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

1. Introduction

1.1. Background and motivating examples

Multiple testing, with emphasis on large scale problems, is an important topic in modern statistics. Classical theory and performance guarantees heavily rely on the knowledge of the null distribution. However, in many practical situations, the null distribution is out of reach. A famous situation, described in a series of work by Efron (2004, 2007, 2008, 2009) and followed by, e.g., Schwartzman (2010), Azriel and Schwartzman (2015), Stephens (2017), Sun and Stephens (2018), Roquain and Verzelen (2020b) is the case where the null distribution is misspecified and is empirically adjusted from the data by fitting some parametric null model (typically Gaussian). In particular, it is well known that using an erroneous null can by disastrous in terms of false discovery rate (FDR), see, e.g., Roquain and Verzelen (2020a). Related works, relying on the famous two-group model (Efron et al., 2001), propose to estimate the null distribution together with the proportion of nulls and the alternative distribution, and to plug them into the so-called local FDR values, see Efron et al. (2001) and Padilla and Bickel (2012), Heller and Yekutieli (2014) among others. The latter can in turn be used into an FDR controlling procedure, see Sun and Cai (2007), Sun and Cai (2009), Cai and Sun (2009), Cai et al. (2019), Roquain and Verzelen (2020b), Abraham et al. (2021). The validity of such approaches, often given asymptotically in the number of tests, also requires strong model assumptions to ensure that these parameters can be correctly estimated.

Here, we consider a semi-supervised setting, with essentially no assumption on the null distribution. Instead, the user has at hand a sample, called the *null training sample* (NTS), of length $n \geq 1$, and generated according to this unknown null. This is motivated by the following generic situations:

- *Blackbox null sampling:* the exact expression of the null distribution is intractable, but a sampling machine is able to simulate according to the null distribution. In that case, the NTS is exogenous and its length n corresponds to the number of sampling, so can be chosen by the user. It is nevertheless typically limited in size by computation time constraints.
- *Null sample given:* the null distribution is unknown, but previous experiments or experts provide a fixed number n of examples under the null. The NTS is exogenous as in the above case, but n cannot be modified by the user.
- *Null sample learned from data:* the null distribution is unknown, but an independent part of the same data set provides an NTS for the user. In that case, the NTS is endogenous, of a given length n that cannot be modified by the user.

The case of "blackbox null sampling" is motivated by numerous situations. Two motivations come from Astrophysics; first when a code can be used to simulate images of astrophysical sources, see e.g. Bacon et al. (2021) (their Figure 15). Second, when the NTS comes from instrumental captures that are made

without the objects of interest, see e.g. Choquet et al. (2018) for the detection of exoplanetary debris disks (their Figure 5). In each of these situations, the null distribution is not accessible for the user, and only the NTS can be generated. More broadly, this case is motivated by recent advances in machine learning, especially implicit generative models, as generative adversarial networks (Goodfellow et al., 2014), or variational auto-encoders (Kingma and Welling, 2014), for which sampling is possible without knowing the underlying distribution. An illustration of the blackbox null sampling case is provided in Appendix A, on a toy example for which multiple likelihood ratio tests are simultaneously performed.

The case of "null sample given" is common in the machine learning context, where the learner is given a sample of "nominal patterns" but without labeled novelties. This is classically referred to as "one class classification" or "learning from positive and unlabeled examples" and we refer the reader to the work Blanchard et al. (2010) that pointed out many references in this abundant literature.

The case of "null sample learned from data" refers to the framework where it is possible to isolate part of the data to produce a sample that contains copies of the test statistics under the null, or approximately so. While it can be met in various datasets, it is motivated by a specific application in Astrophysics that is extensively developed in Section 7. It regards the detection of galaxies in the early Universe from image measurements in multiple wavelength channels. In this application, the distribution of the tests statistics under the null is unknown and it was proposed in Mary et al. (2020), Bacon et al. (2021) to estimate this distribution from a null training sample obtained from the data itself. The NTS is obtained as the population of the opposite of local minima and the whole NTS is used for testing each of the m local maxima.

In the three cases above, a crucial issue is to build a procedure for making discoveries while being fully interpretable, especially when the number of tests m is large. We thus focus on building a procedure that controls the false discovery rate (FDR), that is, the expected ratio of errors among the discoveries made by the procedure (Benjamini and Hochberg, 1995). Interestingly, controlling the FDR by using a simulated NTS has similarities with the recent "knockoff" method introduced in Barber and Candès (2015) which has been at the origin of an impressive scientific production over the last years, see, e.g., Weinstein et al. (2017), Katsevich and Sabatti (2019), Barber and Candès (2019), Bates et al. (2020). Further comparisons are given in Section 1.4.

When proper p-values can be built, the classical way to control the FDR at level α is to use the Benjamini Hochberg (BH) procedure (Benjamini and Hochberg, 1995). However, in the setting described above, the exact p-values are out of reach, so that the usual BH procedure cannot be used. In our context, we call it the *oracle* BH procedure, and denote it by BH^*_α, or BH^* for short. Instead, the NTS can be used to build empirical p-values, called $\hat{p}$-values for short. It is then natural to use the $\hat{p}$-values into the BH procedure, which is the procedure studied in this paper. We call it the *semi-supervised* BH procedure and denote it by $\widehat{\mathrm{BH}}_\alpha$, or $\widehat{\mathrm{BH}}$ for short.

Let us already note that plugging empirically-based p-values into the BH procedure is not new and has been widely explored in the literature, especially in a Monte Carlo framework, see, e.g., Guo and Peddada (2008), Sandve et al. (2011), Gandy and Hahn (2014), Zhang et al. (2019). However, while the same null sample is used to compute all p-values in our setting, most of the existing works focus on the case where m null samples are available, that is, each test uses a different sample, often generated via randomization process (e.g., permutations). In that case, the computational price is much higher and these works mostly aim at reducing this price. The case of *only one* null sample has been considered only recently to our knowledge, see Weinstein et al. (2017), Bates et al. (2021). The computational issue can be easily solved (see Algorithm 1), and our emphasis is rather on the theoretical guarantees of the resulting BH procedure ($\widehat{\mathrm{BH}}$). Further details and comparisons with existing literature are given in Section 1.3 and in Appendix F.1.

Finally, an important point of our work will be to determine how large n should be relatively to the number m of tests. Obviously, when n tends to infinity while the number m of tests is kept fixed, the situation becomes similar to the one where the null distribution is known (that is, when $n = \infty$). But the situation is more complex when both n and m gets large simultaneously, which is typical (e.g., in our galaxy detection example, we have $n \approx m = 3.3 \times 10^6$). As can be guessed, the full picture also depends on the sparsity of the signal. This will be adressed in our theory through a parameter called k, which is a proxy to the number of detectable alternatives.

1.2. Contributions

Main contributions First, we study the FDR of the procedure $\widehat{\mathrm{BH}}$, by providing upper and lower bounds (Theorem 3.1). These bounds hold in a strong sense, that is, for any couple (n, m) with $n, m \geq 1$, any number of true nulls m_0, any null distribution, and any marginal distribution of the alternatives. Moreover, these bounds match and equal $\alpha m_0/m$ when $\alpha(n+1)/m$ is an integer. In practice, this provides a first guideline for choosing n in order to avoid over-conservativeness of the procedure.

Second, we provide a power boundary for $\widehat{\mathrm{BH}}$, which puts forward the crucial role of n with respect to m: the power of $\widehat{\mathrm{BH}}$ is close to the one of the oracle BH^* if $n \gtrsim m/\alpha$ (Proposition 4.2), but is not when $n \lesssim m/\alpha$ (Proposition 4.3). This leads to the boundary $n \asymp m/\alpha$. In addition, we underline the role of the sparsity in the boundary with the following additional result. For distributions that are more favorable in the sense that the oracle BH^* is expected to make at least k true discoveries with high probability (a situation where we say that k alternatives are "detectable"), we show that the power boundary for $\widehat{\mathrm{BH}}$ occurs at $n \asymp m/(k\alpha)$. As an illustration, for $k = 1$, the boundary is $n \asymp m/\alpha$ and thus is the same as for general distributions. However, for the dense case $k = m/2$, the boundary reads $n \asymp 1/\alpha$. This indicates that an NTS of size $\gtrsim 1/\alpha$ is enough to recover the power of the oracle in this case. This is markedly different from the

case of general distributions. In particular, oracle performances can be achieved in the dense case for a constant value of n, regardless of m. Overall, this leads to a new "rule of thumb" with a transition at $n = m/(\alpha \max(1, k))$, which is implemented in the numerical experiments (Section 6) and in the astrophysical example (Section 7).

Third, we show that an intrinsic phase transition occurs in the general case at $n \asymp m$ (Corollary 5.3). The boundary $n \asymp m$ (α being fixed) can not be improved by another procedure: when $n \lesssim m$, no procedure (only based on the test sample and the NTS) can both control the FDR while having a power close to the one of BH* (Theorem 5.1). Since $\widehat{\text{BH}}$ does mimic the oracle when $n \gtrsim m$ (Proposition 4.2), this establishes a general minimax-type optimality property for $\widehat{\text{BH}}$. (Note that the test statistic is fixed in our setting so that BH* is an appropriate reference for power, see Remark 4.1 for a further discussion.)

Secondary contributions First, we show how $\widehat{\text{BH}}$ can be used in the "Blackbox null sampling" setting in Appendix A. We introduce the Blackbox BH procedure, which is defined as the semi-supervised BH procedure with a preliminary step where the NTS is properly generated, see Algorithm 2. While it can be used in a very broad context, we illustrate its use for likelihood ratio tests for which the oracle is accessible in Appendix A.2. A comparison with local FDR type approaches is also provided in that case.

Second, we put forward the following, perhaps seemingly paradoxal, fact for FDR control under negative dependence. Even in the classical setting where the true null is known, it is better not to use BH procedure, but to build instead artificially an NTS, and to use it along with the semi-supervised procedure $\widehat{\text{BH}}$. This approach is referred to as the randomized BH procedure, which is studied separately in Appendix B. While the superiority of the randomized BH procedure over the usual BH procedure in terms of FDR control is shown for an admittedly restrictive dependence structure, correcting the BH procedure to accommodate negative dependencies is known to be a challenging task (see, e.g., Fithian and Lei (2020) and references therein). We think that this intriguing side result is an important proof of concept for the randomized BH procedure.

Third, extensive numerical experiments are given in Section 6 that validate and illustrate our theoretical results. In particular, they corroborate the fact that the boundary where the power of $\widehat{\text{BH}}_\alpha$ gets of the order of the one of BH^*_α occurs around $n = m/(k\alpha)$ (without further tuning of the constant), where k is the number of "detectable" alternatives in the data. For instance, and perhaps counter-intuitively, it is shown that oracle performances can be achieved in a dense case for values of n as small as 5 or 10, regardless of m.

Fourth, a detailed application to galaxy detection is given in Section 7. Remarkably, the recent results of Bacon et al. (2021) suggest the likely discovery of an unexpected population of ultra-faint dwarf galaxies[1]. This discovery results from a two-stage detection process, whose first stage relies on a former version

[1] Also disseminated by the CNRS press release, see, https://www.cnrs.fr/en/first-images-cosmic-web-reveal-myriad-unsuspected-dwarf-galaxies

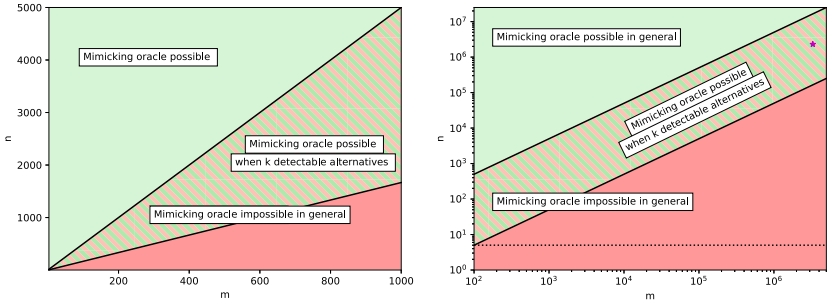


FIG 1. *Visualization of the general, distribution-free phase transition $n = m/\alpha$ for the semi-supervised multiple testing problem, as established in Section 5.2 (with $\alpha = 0.2$) and of the $\widehat{BH}$ boundary $n = m/(\alpha k)$, only valid for distributions with at least k detectable alternatives in the sense defined in Section 4.3. Left: $k = 3$. Right: $k = 100$; plot in* log-log *scale; the boundary $n = 1/\alpha$ ($k = m$) is added with a dotted line; the case of the MUSE data set (Section 7) is also added with a star symbol.*

of the semi-supervised Benjamini-Hochberg procedure developed in Mary et al. (2020), which also provides the same output as $\widehat{\mathrm{BH}}_\alpha$. Hence, the present paper provides a theoretical support to these findings, with guarantees both on the FDR and on the power.

Figure 1 summarizes the different power regimes put forward in our analysis. The transition phase $n = m/\alpha$ separates two regimes: the regime where oracle performances can be reached for any distribution ("mimicking the oracle possible in general", lime green) versus the regime where no procedure can reach the oracle performances ("mimicking the oracle impossible in general", tomato + red brick). The line $n = m/(\alpha k)$ is the performance boundary of $\widehat{\mathrm{BH}}$ for favorable distributions for which at least k alternatives are detectable (in which case oracle performances can be reached in the lime green + tomato area). Note that our theory proves that these boundaries hold only up to numerical constants, whereas the numerical experiments suggest that they hold with constant 1.

1.3. Related works

Permutation-based multiple testing A common way to generate a "null sample" from the data under test is to apply some randomization that preserves the null distribution, typically by performing permutations of individuals. While single testing using randomization is classical and can be traced back to Fisher (1935), several extensions have been proposed in the literature to accommodate multiple testing criteria, see Westfall and Young (1993), Lin (2005), Romano and Wolf (2005, 2007), Hemerik et al. (2019). In particular, an active line of research is dedicated to reduce the computation time of BH procedure with p-values obtained from permutation-based null samples: indeed, the usual permutation-based paradigm requires to generate a different "null sample" for each test, which

makes the use of such a BH procedure prohibitive in that framework. In Guo and Peddada (2008), they adapt the number of bootstrap samples sequentially to speed-up BH procedure by using bootstrap confidence intervals for p-values. This method is further refined in Gandy and Hahn (2014), where the procedure recovers with high probability the rejection set of the BH procedure using "ideal" p-values (exhausting all permutations). Another approach is used in Sandve et al. (2011) by allocating the Monte Carlo budget (total number of Monte Carlo samples) according to the significance of the test statistics, itself extending an idea of Besag and Clifford (1991) for single testing. More recently, Zhang et al. (2019) proposed to reduce the computation burden by following a bandit approach. While all these works are based on null training samples, the crucial difference is that our setting only relies on *one* null sample for all tests. The consequences are the following: first, the complexity of the procedure proposed here ($\widehat{\mathrm{BH}}$) is much smaller than that of the BH procedure with permutation-based p-values, the need for designing an efficient algorithmic strategy is far less critical than in the works mentioned above (note that our Algorithm 1 for $\widehat{\mathrm{BH}}$ is nevertheless efficient). Second, this advantage comes with a counterpart: in the case where the initial test statistics are independent, the permutation-based p-values are also independent, while our setting induces dependencies between the $\hat{p}$-values (the same NTS is used to build all $\hat{p}$-values). This makes the FDR control more difficult to obtain.

Finally, the above comparison has to be moderated by the fact that randomization testing and our semi-supervised setting each come with specific mathematical assumptions: randomization testing relies on a null distributional invariance which is very different from the assumption (Exch) below. Namely, the exchangeability property concerns the set of "variables" (nulls of the test sample plus the null training sample), whereas in permutation testing, the exchangeability concerns the set of individuals. As a consequence, mathematical results derived in each framework cannot be directly compared. In particular, it is important to note that we do not pretend to address the FDR controlling problem in the permutation-based framework. Our contribution lies in another framework, which thus departs from the Monte-Carlo literature mentioned above.

Multiple comparisons to control Multiple comparisons to control (MCC) is a long-established problem in multiple testing (Dunnett, 1955, Hsu, 1996, Finner and Strassburger, 2007, Fithian and Lei, 2020) where one typically aims at comparing several treatments to some common benchmark (control). In the MCC setting, one typical observes only one test statistic per treatments and one test statistics for the control. This would correspond to the case where the null training sample is of length $n = 1$, which is not the typical case considered here. Hence, to our knowledge, the connection to that part of the literature is only weak.

Other FDR controls Our work is closely related to the problem of semi-supervised novelty detections (Blanchard et al., 2010), developed in a machine

learning context, where the user has at hand both a null sample and an unlabeled sample and they aim at labeling the unlabeled sample. However, the procedures developed therein are significantly different from here: first, they adjust the test statistics by considering families of classifiers. Second, their FDR control is based on a concentration argument that adds an error term larger than $n^{-1/2}+m^{-1/2}$ (see Proposition 12 therein) and depending on the VC-dimension of the classifier class, while the FDR control in Theorem 3.1 is exact (no error term).

Finally, another closely related literature tackles the issue of learning the null distribution without null training sample (only using the original test statistics) but assuming that the null distribution belongs to a parametric model, typically Gaussian with unknown mean and variance. While the most classical line of research is the one following the "local FDR" methodology introduced by Efron, see, e.g., Efron (2008), theoretical results have been obtained by Carpentier et al. (2021), Roquain and Verzelen (2020b). The methodology developed here, and particularly the impossibility result (Section 4.2) and the boundary phenomenon (Section 5.2), are inspired from Roquain and Verzelen (2020b). However, the setting being markedly different, several substantial developments are required. Also, we underline that we derive here an FDR control without remainder terms, which was not the case in Carpentier et al. (2021), Roquain and Verzelen (2020b).

Naive solutions to our problem For completeness, let us discuss two naive solutions that can be straightforwardly used to derive a procedure with a proven FDR control in the present semi-supervised setting, and explain why they are not satisfactory. Recall that, even under independence of the test statistics, the $\hat{p}$-values are not independent, which is a problem to design an FDR controlling procedure that takes as input these $\hat{p}$-values.

First, one solution is to use the Benjamini-Yekutieli procedure or one of its extension Benjamini and Yekutieli (2001), Blanchard and Roquain (2008) that control the FDR under arbitrary dependence between the p-values, so also when used with $\hat{p}$-values. Namely, the semi-supervised Benjamini-Yekutieli procedure, denoted by $\widehat{\mathrm{BY}}_\alpha$ (or $\widehat{\mathrm{BY}}$ for short), considers $\widehat{\mathrm{BH}}_{\alpha/c_m}$ at level α/c_m where $c_m = 1+1/2+\cdots+1/m$. However, it is well known that the power loss is substantial with respect to BH procedure and this general fact also holds in our setting, as it will be shown in the numerical experiments, see Appendix F.1. In addition, Theorem 3.1 shows that under Assumption (Exch), the procedure $\widehat{\mathrm{BH}}$ already achieves the desired FDR control so there is no need to use the corrected procedure $\widehat{\mathrm{BY}}$.

A second naive solution, referred to as $\widehat{\mathrm{BH}}_{\mathrm{Split}}$, is to split the NTS of size n into m null samples $T^1,\dots,T^m$, each of size n/m (say that the latter ratio is an integer for simplicity) so that each $\hat{p}$-value uses a different part of the null sample, that is, each $\hat{p}_i$ is computed from the null training sample T^i. In that case, if the test statistics are independent, these modified $\hat{p}$-values are also independent, and the BH procedure using these modified $\hat{p}$-values does control the FDR by the original result of Benjamini and Hochberg (1995). However,

this reduces drastically the size of the (different) NTS used to calibrate each test (n/m instead of n), which leads again to a poor power, see Appendix F.1.

1.4. Link with previous literature on knockoff: Application to LASSO-based test statistics

The method investigated here has an important connection to the knockoff methodology (Candès et al., 2018, Weinstein et al., 2017, 2020). We make below a detailed description of these works to help distinguish the contribution of our work. Interestingly, this shows that the empirical BH procedure can be used with LASSO-based test statistics in a context of a Gaussian linear model with i.i.d. design matrix, see also Example 3.3 and Figure 3 below.

Setting Let us consider the (potentially high-dimensional) Gaussian linear model with m variables and N individuals where we observe

$$W = M\beta + \varepsilon, \ \ \beta \in \mathbb{R}^m, \ \ \varepsilon \sim \mathcal{N}(0, \sigma^2 I_N), \ \ \sigma > 0,$$

in which the $N \times m$ real design matrix M is random with a given known distribution. The aim is to test the null "$\beta_j = 0$" against the alternative "$\beta_j \neq 0$", simultaneously for all $1 \leq j \leq m$. Two kinds of $N \times n$ real "knockoff" matrix $\tilde{M}$ for M have been considered, with a corresponding specific choice of the test statistic:

- (A) the model-X knockoff (Candès et al., 2018): $\tilde{M}$ is obtained from M following a specific process, so that swapping the columns of the augmented design $\mathbb{M} = [\tilde{M}\ M]$ does not change its distribution (here $n = m$). In that case, the classical test statistic used is the LASSO coefficient difference (LCD), namely $X_j^{LCD} = |\hat{\beta}_{j+m}(\lambda)| - |\hat{\beta}_j(\lambda)|$, $1 \leq j \leq m$. In the latter, the LASSO solution $\hat{\beta}_j(\lambda)$ is computed in the linear model with the augmented $N \times (n+m)$ design matrix $\mathbb{M} = [\tilde{M}\ M]$, that is,

$$\hat{\beta}(\lambda) \in \underset{b \in \mathbb{R}^{n+m}}{\arg\min} \left\{ 0.5\|W - \mathbb{M}b\|^2 + \lambda\|b\|_1 \right\}, \quad \lambda \geq 0\,; \tag{1}$$

- (B) the counting knockoff (Weinstein et al., 2017, 2020): it is valid only when the entries of M are i.i.d. $\sim G$ (say). Then, the entries of $\tilde{M}$ are also i.i.d. $\sim G$ (hence G is known and n can be arbitrary). In that case, the test statistics Z_j, $1 \leq j \leq n+m$, are computed for each variable, typically as a function of the LASSO path $\{\hat{\beta}_j(\lambda), \lambda \geq 0\}$. Examples include the LASSO maximum (LM) statistic $Z_j^{LM} = \max\{\lambda \geq 0 \ : \ \hat{\beta}_j(\lambda) \neq 0\}$ (Weinstein et al., 2017) and the LASSO coefficient (LC) statistic $Z_j^{LC} = |\hat{\beta}_j(\lambda)|$, for a fixed $\lambda \geq 0$ (as in Weinstein et al., 2020). The LASSO solution $\hat{\beta}(\lambda)$ is still computed according to (1).

It turns out that in case (B), for a statistic either computed with LM or LC, the observation vector $(Z_1, \ldots, Z_{n+m}) = (Y_1, \ldots, Y_n, X_1, \ldots, X_m)$ satisfies

that $(Y_1, \ldots, Y_n, X_j, j \in \mathcal{H}_0)$ is exchangeable conditionally on $(X_j, j \notin \mathcal{H}_0)$, where $\mathcal{H}_0 = \{j \in \{1, \ldots, m\} \ : \ \beta_j = 0\}$. While this exchangeability property is considered as obvious in Weinstein et al. (2017), we provide a formal result in Appendix E for completeness, see Lemma E.6. This assumption is the one considered in Theorem 3.1 of Weinstein et al. (2017), which is the same as our Assumption (Exch) below.

Procedures In Weinstein et al. (2017, 2020), the FDR controlling procedures are each time based on an estimation of the FDP:

- In case (A): the FDP estimator is based on a symmetrisation estimator of the FDP, that allows implicitly to compare the knockoffs test statistics to the original test statistics. The resulting procedure is called LCD-knockoff.
- In case (B): the FDP estimator is based on a direct comparison of knockoffs test statistics to the original test statistics. The resulting procedure is called LC-counting-knockoff and LM-counting-knockoff, if used with the LC test statistic and LM test statistic, respectively.

While the procedure LCD-knockoff is *not* the one of the present work, LM-counting-knockoff and LC-counting-knockoff are particular cases of the empirical BH procedure, applied with the LM test statistic (as in Weinstein et al., 2017) or the LC test statistic (as in Weinstein et al., 2020). This link between counting-knockoff and empirical BH has not been noticed before to our knowledge, and is stated in Lemma 2.2.

FDR control In Weinstein et al. (2017, 2020), it is proved that LCD-knockoff controls the FDR in case (A), while LM-counting-knockoff and LC-counting-knockoff controls the FDR in case (B). Theorem 3.1 in Weinstein et al. (2017) shows more generally the FDR control under Assumption (Exch). This is the same result has our upper bound in our Theorem 3.1. Our contribution w.r.t. that work is thus the lower bound in Theorem 3.1. It shows in particular that the upper bound is reached when $\alpha(n+1)/m$ is an integer.

Power results In Weinstein et al. (2017), they rely on the work by Bayati and Montanari (2011) to study the power of the LM-counting-knockoff in case (B) above. For this, they consider random effects on the true coefficient $\beta \in \mathbb{R}^m$ and use as an oracle the LASSO support (computed with the non-extended design matrix M). The comparison is done in terms of the asymptotical ROC curve. There is a tradeoff w.r.t. the parameter n: a large n provides a good approximation of the null distribution but deteriorates the quality of the LM test statistics. Next, the work Weinstein et al. (2020) further investigates the LC-counting-knockoff in case (B) and the LCD-knockoff in case (A) (thus covering the model-X knockoff framework), with an oracle given by the thresholded LASSO (computed with the non-extended design matrix M).

By contrast, the power results obtained in the present paper are much different: they are not developed in the particular LASSO-like cases (A)-(B) above, but in a setting where the test statistics are not varying with n and are mutually

independent, see (Indep) below. Hence, in our setting, there is no tradeoff in n: a larger n always leads to more power (see the phase transition in Figure 1). Another major difference is that our power study holds non asymptotically and in a minimax sense over different classes of alternatives, and in particular does not use random effects on the alternatives. In particular, our setting encompasses the sparse situation (proportion of alternatives tending to zero), which is excluded in the studies Weinstein et al. (2017, 2020). On the other hand, our power study is not able to deal with LASSO-type test statistics. In conclusion, the power study in Weinstein et al. (2017, 2020) and the one derived in the present work are of different nature (setting, statement) and they each have their own merit.

1.5. Organization of the paper

The paper is organized as follows: while the model, procedures and criteria are detailed in Section 2, FDR results are given in Section 3. Power properties of $\widehat{\mathrm{BH}}$ are then derived in Section 4 with upper and lower bounds, which delineate boundaries for $\widehat{\mathrm{BH}}$. Extending to any procedure the impossibility result below the boundary, the result of Section 5 delivers an optimality property of $\widehat{\mathrm{BH}}$ and a general phase transition for the semi-supervised multiple testing problem. We then illustrate our findings with numerical experiments in Section 6 and the motivating application to astrophysical data is investigated in Section 7. We conclude and discuss several open issues related to our work in Section 8. Two by-products of our theory are presented in Appendices A and B, with the blackbox BH procedure and the randomized BH procedure, respectively. The main proofs are given in Appendix C and Appendix D for FDR results and power results, respectively. Auxiliary results and proofs are postponed to Appendix E, while additional numerical experiments are given in Appendix F.

2. Preliminaries

2.1. Setting

For $n, m \geq 1$, let us observe a sample $Z = (Z_1, \ldots, Z_{n+m}) = (Y_1, \ldots, Y_n, X_1, \ldots, X_m) \in \mathbb{R}^{n+m}$, whose distribution is denoted by P, the model parameter, that belongs to some model $\mathcal{P}$. The sample $Y = (Y_1, \ldots, Y_n)$ is referred to as the null training sample (NTS), which is assumed to be identically distributed of marginal distribution $P_0 = P_0(P)$. We denote the upper-tail function of P_0 by $F_0(t) = \mathbb{P}(Y_i \geq t)$, $t \in \mathbb{R}$. P_0 is assumed to be unknown throughout the manuscript (except in Appendices A.2 and B). The only assumption made on P_0 (or equivalently F_0) throughout the manuscript is the following:

F_0 is assumed to be continuous and decreasing on the support of P_0. (Cont)

The sample $X = (X_1, \ldots, X_m)$ corresponds to the sample under test, referred to as the test sample. We consider the multiple testing problem where we would

like to test the i-th null hypothesis H_i: "$X_i \sim P_0$" (against the complementary alternative), simultaneously for $1 \leq i \leq m$. Note that while we allow for arbitrary alternatives here, this setting is typically suitable for alternatives that make X_i stochastically larger than under the null (decisions will be based upon large values of the X_i's). Classically, let us denote $\mathcal{H}_0(P) = \{i \in \{1, \dots, m\} : X_i \sim P_0\} \subseteq \{1, \dots, m\}$ the subset corresponding to true null hypotheses and $m_0(P) = |\mathcal{H}_0(P)|$. Let us denote $\mathcal{H}_1(P)$ the complement of $\mathcal{H}_0(P)$ in $\{1, \dots, m\}$ and $m_1(P) = m - m_0(P)$. Often, we omit the parameter P in the notation $P_0, \mathcal{H}_0, \mathcal{H}_1, m_0, m_1$ for simplicity.

Throughout the paper, we are going to consider various dependence assumptions between the Z_i's. The most simple assumption is

$$(Y_1, \dots, Y_n, X_i, i \in \mathcal{H}_0) \text{ are i.i.d. } \sim P_0 \text{ and independent of } (X_i, i \notin \mathcal{H}_0). \tag{Indep}$$

Note that (Indep) does not exclude dependencies between the elements of $(X_i, i \notin \mathcal{H}_0)$. We also use the following less restrictive condition:

$$(Y_1, \dots, Y_n, X_i, i \in \mathcal{H}_0) \text{ are exchangeable conditionally on } (X_i, i \notin \mathcal{H}_0). \tag{Exch}$$

Hence, under (Exch), there could be also some dependencies between the elements of $(Y_1, \dots, Y_n, X_i, i \in \mathcal{H}_0)$.

2.2. Procedures, criteria and p-values

A multiple testing procedure is a (measurable) function $R = R(Z)$ that returns a subset of $\{1, \dots, m\}$ corresponding to the indices i where H_i is rejected. For any such procedure R, the false discovery rate (FDR) of R is defined as the average of the false discovery proportion (FDP) of R under the model parameter $P \in \mathcal{P}$, that is,

$$\mathrm{FDR}(P, R) = \mathbb{E}_P[\mathrm{FDP}(P, R)], \quad \mathrm{FDP}(P, R) = \frac{\sum_{i \in \mathcal{H}_0} \mathbb{1}_{\{i \in R\}}}{1 \vee |R|}. \tag{2}$$

Similarly, the true discovery rate (TDR) is defined as the average of the true discovery proportion (TDP), that is,

$$\mathrm{TDR}(P, R) = \mathbb{E}_P[\mathrm{TDP}(P, R)], \quad \mathrm{TDP}(P, R) = \frac{\sum_{i \in \mathcal{H}_1} \mathbb{1}_{\{i \in R\}}}{1 \vee m_1(P)}. \tag{3}$$

Note that if $m_1(P) = 0$, $\mathrm{TDP}(P, R) = 0$ for all procedures R.

In the sequel, we will focus on p-value based procedures and we implicitly consider the situation where it is desirable to reject H_i for large values of X_i. If the null distribution P_0 is known, F_0 is known and we can consider $p_i(X) = F_0(X_i)$, $1 \leq i \leq m$. By definition, the p-value family $p_i = p_i(X)$, $1 \leq i \leq m$,

satisfies that for all $i \in \mathcal{H}_0(P)$, $p_i \sim U(0,1)$, and thus also the super-uniformity property

$$\forall i \in \mathcal{H}_0(P),\ \ \forall u \in [0,1], \mathbb{P}_{Z\sim P}(p_i \leq u) \leq u. \tag{4}$$

As it is required to obtain valid individual tests, condition (4) is generally considered as the definition of "valid" p-values.

Since in our framework P_0 is unknown, the above p-values are unknown oracle p-values and thus cannot be used in practice. Instead, the null sample $(Y_1, \ldots, Y_n)$ can be used to build the empirical p-values

$$\widetilde{p}_i(Z) = n^{-1} \sum_{j=1}^{n} \mathbb{1}_{\{Y_j \geq X_i\}},\ \ 1 \leq i \leq m. \tag{5}$$

However, the $\widetilde{p}_i$'s do not satisfies the necessary super-uniformity (4). For instance, for $u = 0$, the condition (4) is violated because the event $\widetilde{p}_i(Z) = 0$ can occur with a positive probability. Hence, using the $\widetilde{p}_i$'s as p-values is not appropriate, especially in a multiple testing context where under-estimating p-values can lead to an increased number of false discoveries. This phenomenon is well known and we refer the reader to the review of Phipson and Smyth (2010) for more details on this issue (see also the references therein). A common way to correct the $\widetilde{p}_i$'s is to make them slightly biased upward by considering instead the conservative version (see, e.g., Davison and Hinkley, 1997), given by

$$\widehat{p}_i(Z) = \widehat{F}_0(X_i) = (n+1)^{-1} \sum_{x \in \{X_i, Y_1, \ldots, Y_n\}} \mathbb{1}_{\{x \geq X_i\}},\ \ 1 \leq i \leq m, \tag{6}$$

where we let

$$\widehat{F}_0(x) = (n+1)^{-1} \left(1 + \sum_{j=1}^{n} \mathbb{1}_{\{Y_j \geq x\}}\right),\ \ x \in \mathbb{R}. \tag{7}$$

Under (Exch), since for any $i \in \mathcal{H}_0$, the variables $X_i, Y_1, \ldots, Y_n$ are exchangeable, the $\widehat{p}_i(Z)$'s do satisfy the super-uniformity (4), see, e.g., Lemma 5.2 in Arlot et al. (2010). Hence, the $\widehat{p}_i(Z)$ are "valid" p-values, that can in turn be plugged into multiple testing procedures.

2.3. BH procedures

In this work, an important class of multiple testing procedures is the BH-type procedures, which use as input different p-value families. The BH procedure is defined as follows: for some level $\alpha \in (0,1)$, order the p-values in increasing order $p_{(1)} \leq \cdots \leq p_{(m)}$ and then let

$$\mathrm{BH}_\alpha = \{i \in \{1, \ldots, m\} : p_i \leq \alpha \widehat{k}/m\}, \widehat{k} = \max\{k \in \{0, 1, \ldots, m\} : p_{(k)} \leq \alpha k/m\}, \tag{8}$$

where α is the nominal level of BH procedure and where we let $p_{(0)} = 0$ by convention.

Algorithm 1: $\widehat{\mathrm{BH}}_\alpha$, the semi-supervised BH procedure (case where (Cont) holds)

Data: $Z = (Z_1, \dots, Z_{n+m}) = (Y_1, \dots, Y_n, X_1, \dots, X_m) \in \mathbb{R}^{n+m}$ semi-supervised sample, α level

1. Order the Z_i's, that is, $Z_{\tau(1)} \geq \cdots \geq Z_{\tau(n+m)}$, for some permutation τ of $\{1, \dots, n+m\}$
2. Let $s_\ell = \mathbb{1}_{\{\tau(\ell) \leq n\}} \in \{0, 1\}$ which is 1 if and only if $Z_{\tau(\ell)}$ comes from sample $Y = (Y_1, \dots, Y_n)$
3. Let FDP $= 1$, $V = n$, $\ell = m + n$, $K = m$
4. While (FDP $> \alpha$ and $K \geq 1$) do $\ell = \ell - 1$
   - if $s_{\ell+1} = 1$, $V = V - 1$
   - else, $K = K - 1$

   FDP $= \frac{V+1}{n+1} \frac{m}{K}$ (or FDP $= 1$ if $K = 0$)

Result: $\widehat{\mathrm{BH}}_\alpha = \{i \in \{1, \dots, m\} \,:\, X_i \geq X_{(K)}\}$ (reject nothing if $K = 0$).

Definition 2.1. We consider the two following versions of BH procedure, depending on which p-value family is given as input:

- the *oracle BH procedure*, denoted by BH^*_α, is the BH procedure using the unknown p-values $p_i(X) = F_0(X_i)$, $1 \leq i \leq m$;
- the *semi-supervised BH procedure*, denoted by $\widehat{\mathrm{BH}}_\alpha$, is the BH procedure using the $\hat{p}$-values $\widehat{p}_i(Z)$, $1 \leq i \leq m$, given by (6).

Importantly, the output of $\widehat{\mathrm{BH}}_\alpha$ can be equivalently derived by using the following lemma (to be proved in Appendix C.1).

Lemma 2.2. *Letting $\mathcal{T} = \{X_i, 1 \leq i \leq m\}$, we have*

$$\widehat{BH}_\alpha = \{i \in \{1, \dots, m\} \,:\, X_i \geq \hat{t}\}$$
$$\hat{t} = \min\left\{ t \in \mathcal{T} \,:\, \frac{m}{n+1} \frac{1 + \sum_{i=1}^n \mathbb{1}_{\{Y_i \geq t\}}}{\sum_{i=1}^m \mathbb{1}_{\{X_i \geq t\}}} \leq \alpha \right\}.$$

Lemma 2.2 establishes that the semi-supervised BH procedure coincides with the "counting knockoff" procedure introduced in Weinstein et al. (2017).

Assuming moreover (Cont), the output of $\widehat{\mathrm{BH}}_\alpha$ can almost surely be derived by Algorithm 1. Figure 2 provides an illustration of Algorithm 1: it is a stepwise procedure that goes from the smallest values of the test statistics (right) to the largest values (left), and that stops the first time where the FDP falls below α. At each step, the FDP is estimated by the ratio of the number of null samples in the left part plus one $(V + 1)$, to the number of test statistics in the left part (K), this ratio being sample-sized corrected by the factor $m/(n + 1)$. Hence, at each step, the Y_i's are used as benchmarks to evaluate how many false discoveries are expected among the considered X_i's. Finally, while the above version of

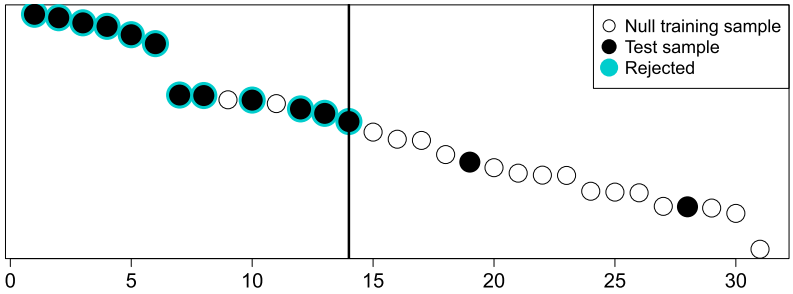


FIG 2. *Illustration of Algorithm 1 for $m = 14$, $n = 17$, $\alpha = 0.2$, and some realization of the ordered test statistics. At the point of rejections (vertical line) $\ell = 14$, we have $V = 2$, $K = 12$, $FDP = \frac{2+1}{n+1}\frac{m}{12} \leq 0.2$, while $FDP > 0.2$ in any further point $\ell > 14$. The algorithm makes $K = 12$ rejections in the test sample (depicted in blue). See text for further comments.*

Algorithm 1 was presented for simplicity, a shortcut (faster) version can obviously be obtained by iterating in the loop only over the indices ℓ corresponding to the X_i's (the FDP estimator is computed only at black points in Figure 2).

Remark 2.3. The equivalence between $\widehat{\mathrm{BH}}_\alpha$ and Algorithm 1 is not true if there are ties (this situation does not occur almost surely if (Cont) is satisfied): for instance, if $Z_i = 0$ for $1 \leq i \leq n+m$. Then all the $\widehat{p}$-values are equal to 1 and $\widehat{\mathrm{BH}}_\alpha$ rejects no null (at any level $\alpha \in (0,1)$). By contrast, it τ is such that $Z_{\tau(1)}, \ldots, Z_{\tau(m)}$ all come from the sample X (that is, $s_1 = \cdots = s_m = 0$), Algorithm 1 at a level $\alpha \geq 1/(n+1)$ rejects all nulls.

3. FDR control

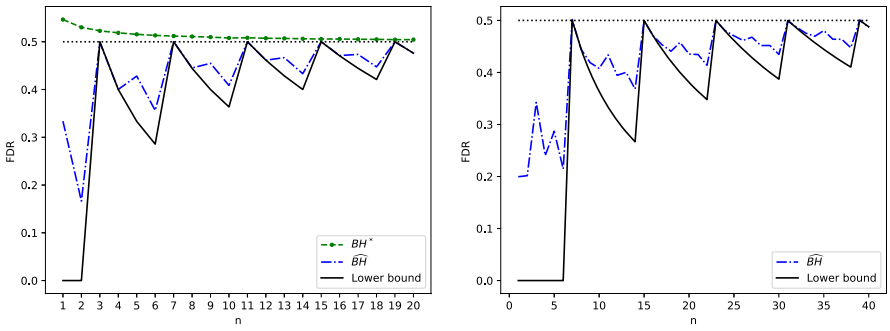
The FDR control result is as follows.

Theorem 3.1. *For all $n, m \geq 1$ and $\alpha \in (0,1)$, consider the semi-supervised BH procedure $\widehat{BH}_\alpha$ at level α as defined in Definition 2.1. Then, for any parameter P satisfying (Cont) and (Exch), the following holds:*

$$\frac{m_0}{m}\frac{m}{n+1}\left\lfloor \alpha \frac{n+1}{m} \right\rfloor \leq FDR(P, \widehat{BH}_\alpha) \leq \alpha m_0/m,$$

where $\lfloor x \rfloor$ denotes the largest integer smaller than or equal to x. In particular, when $\alpha(n+1)/m$ is an integer, the FDR bound is achieved, that is, $FDR(P, \widehat{BH}_\alpha) = \alpha m_0/m$.

The proof is given in Appendix C and is based on a super-martingale argument which is similar to that of Barber and Candès (2015). However, a major difference is that the underlying process is not an i.i.d. Bernoulli process, but is only exchangeable, see Lemma E.2 for more details. The lower bound part is obtained by looking carefully at the remainder term in the super-martingale



Equi-correlated LASSO

FIG 3. *FDR of* $\widehat{BH}$ *and* BH^* *as a function of* n*, the size of the null training sample* Y*. Left: maximal negative equicorrelated case as described in Example 3.2.* n *ranges from* 0 *to* 20*,* $m_0 = m = 2$*, and the FDR are evaluated with* 10^6 *simulations. Right: using LASSO coefficient statistics as described in Example 3.3.* n *ranges from* 0 *to* 40*,* $m_0 = m = 4$*,* $N = 4$*,* $G = \mathcal{N}(0,1)$*, and the FDR are evaluated with* 10^6 *simulations. To ensure* (Cont)*, the test statistics are disturbed in an i.i.d. manner by a* $\mathcal{N}(0,(0.001)^2)$*. LASSO implemented with module* `scikit-learn` *of python,* $\lambda = 0.1$ *and* $max_{iter} = 10^4$*. In both cases, the nominal level is* $\alpha = 0.5$ *(horizontal dashed line). The FDR lower bound delineated in Theorem 3.1 is also displayed.*

property. To our knowledge, this kind of refinement is new in the literature. This allows to evaluate the sharpness of the FDR bound.

In particular, Theorem 3.1 shows that under (Indep) (implying (Exch)) the semi-supervised BH procedure $\widehat{\mathrm{BH}}_\alpha$ has an FDR smaller than or equal to the one of BH^*_α. More precisely, since $\mathrm{FDR}(P, \mathrm{BH}^*_\alpha) = \alpha m_0/m$ under (Indep) (see Benjamini and Yekutieli (2001)), we have under (Indep),

$$\mathrm{FDR}(P, \widehat{\mathrm{BH}}_\alpha) \le \alpha m_0/m = \mathrm{FDR}(P, \mathrm{BH}^*_\alpha). \tag{9}$$

In addition, the FDR control of $\widehat{\mathrm{BH}}_\alpha$ holds under the more general condition (Exch). This is not the case for BH^*_α that can violate the FDR control under that condition. Hence, Theorem 3.1 puts forward an additional robustness of $\widehat{\mathrm{BH}}_\alpha$ w.r.t. the negative dependence, which is not satisfied by BH^*_α. We provide two examples below, see Figure 3 for an illustration.

Example 3.2 (Gaussian with maximal negative equicorrelation). Assume that $Z = (Y, X)$ is a centered Gaussian vector with equicorrelation $\rho < 0$ and variances equal to 1. Classically, since the length of Z is $n + m$, the condition $\rho \ge -1/(n+m-1)$ is necessary to provide that the $(n+m)\times(n+m)$ ρ-equicorrelated matrix (that is, with diagonal 1 and off-diagonal element ρ) is non-negative. For instance, the maximal negatively correlated case $\rho = -1/(n+m-1)$ can be easily realized as $Z = (1+1/(n+m-1))^{1/2}(W_i - \overline{W})_{1\le i\le n+m}$, with W_i, $1 \le i \le n+m$, i.i.d. $\mathcal{N}(0,1)$ and $\overline{W}$ denoting the sample mean of the W_i's, $1 \le i \le n+m$. For this specific distribution P of Z, we have $P_0 = P_0(P) = \mathcal{N}(0,1)$ and $\mathcal{H}_0 = \{1, \dots, m\}$. Also, Assumption (Exch) is satisfied so that $\widehat{\mathrm{BH}}_\alpha$ controls

the FDR at level α (with equality when $\alpha(n+1)/m$ is an integer). On the other hand, it is well known that BH^*_α has an FDR above α in that case (see also Figure 3). Additional illustrations are given in Section 6.2 in the numerical experiments. This example is also the starting point of the randomized BH procedure developed in Appendix B.

Example 3.3 (LASSO coefficient correlations). In the Gaussian linear model described in Section 1.4 (case B), we can consider $Z_j = |\hat{\beta}_j(\lambda)| + \eta_j$, the LASSO coefficient test statistic disturbed by some i.i.d. random variables η_j that are $\mathcal{N}(0, \epsilon^2)$, with ϵ small enough. A well known fact is that some of the $|\hat{\beta}_j(\lambda)|$ will be equal to zero. Hence, the perturbation by η_j is necessary to make Assumption (Cont) hold true. While the Z_j's are not independent, Assumption (Exch) is also satisfied, as mentioned in Section 1.4. Hence, the upper and lower bounds of Theorem 3.1 both apply in that case. They are illustrated in Figure 3.

Remark 3.4. Since the first version of this work, earlier occurrences of the upper-bound proved in Theorem 3.1 have been reported to us (our work has been developed independently): first, it has been proved under assumption (Exch) in the work of Weinstein et al. (2017) by using the same martingale as ours (in a different context). Second, the upper-bound is a consequence of the work of Bates et al. (2021) who showed that the $\hat{p}$-values, despite their intricate structure, are positively regressively dependent on each one of the subset (PRDS). This is proved under the stronger assumption (Indep).

Remark 3.5. When $\alpha(n+1)/m$ is an integer, we can easily check that $\widehat{\mathrm{BH}}_\alpha$ coincide with $\widetilde{\mathrm{BH}}_\alpha$, the BH procedure applied to the naive, unbiased, $\tilde{p}$-values defined by (5). Hence, Theorem 3.1 implies that $\mathrm{FDR}(P, \widetilde{\mathrm{BH}}_\alpha) = \alpha m_0/m$ in that case (under (Exch)). This shows that, perhaps surprisingly, the naive way to build empirical p-values eventually leads to a correct FDR control for such values of n. Simulations will show that this is not the necessarily the case for other values of n, see Section 6.

4. Power result

Section 3 showed that $\widehat{\mathrm{BH}}_\alpha$ has an FDR smaller than or equal to the one of the oracle BH^*_α under (Indep), see (9). Now, an important concern is to check whether the *power* of $\widehat{\mathrm{BH}}_\alpha$ is comparable to the one of BH^*_α. In this section, we explore this issue under Assumption (Indep) and the power comparison is established by comparing the true discovery proportions (3) of $\widehat{\mathrm{BH}}_\alpha$ and $\mathrm{BH}^*_{\alpha'}$, for α' slightly below α. In a nutshell, we establish that the TDP of $\widehat{\mathrm{BH}}_\alpha$ is larger than the one of $\mathrm{BH}^*_{\alpha'}$ with a probability tending to 1, for any model parameter, when n/m is large (Section 4.1), while we show that it is not true when n/m is small (Section 4.2). Together, this means that the boundary achieved by the procedure $\widehat{\mathrm{BH}}_\alpha$ is $n \asymp m/\alpha$. We then present the case of particular, more favorable distributions, for which at least k alternatives are "detectable" (Section 4.3). In that case, the boundary achieved by $\widehat{\mathrm{BH}}_\alpha$ is shown to be $nk \asymp m/\alpha$.

To state our results, let us finally introduce an additional notation: let

$$\mathcal{P}_{n,m} = \left\{ P = P_0^{\otimes n} \otimes \bigotimes_{i=1}^{m} P_i \ : \ P_i \text{ continuous distribution on } \mathbb{R},\ 0 \le i \le m \right\}. \tag{10}$$

The distribution P belongs to $\mathcal{P}_{n,m}$ in the semi-supervised setting presented in Section 2.1 and under (Indep). Hence, P can be considered as the parameter set of the model under that assumption. In addition, since we look at power results, we are going to focus on distributions in $\mathcal{P}_{n,m}$ with at least one true alternative. We denote by

$$\mathcal{A}_{n,m} = \{P \in \mathcal{P}_{n,m} \ : \ m_1(P) \ge 1\} \tag{11}$$

the corresponding set.

Remark 4.1 (Optimality of BH*). In the past literature, numerous works proposed approaches that improve, sometimes substantially, the baseline BH procedure (as local FDR methods listed in introduction). Hence, a common belief is that the BH procedure is well known to be conservative and suboptimal when controlling the FDR. This belief makes the aim of mimicking the performance of the oracle BH procedure somewhat questionable. However, we argue that this belief is not justified when the test statistic used before applying BH algorithm is suitably chosen, typically using a likelihood ratio or a local FDR transformation. This is shown in particular with simulations in the setting of Appendix A, where BH* (= BH used with the oracle p-values) improves over a local FDR method, itself well known to enjoy optimality properties (Cai et al., 2019). In a nutshell, the possible conservativeness of BH procedure when controlling the FDR is not due to the BH algorithm *per se* but rather to the test statistic used as entries of this algorithm. To come back to our framework, the test statistic is assumed to be fixed once for all in our work. Hence, *given the chosen test statistic*, BH* is close to be optimal and the aim considered in this section perfectly makes sense.

4.1. Upper bound

The following result shows that, under (Indep), when $n \ge \gamma m$ with γ large enough, the semi-supervised BH procedure at level α rejects at least all null hypotheses rejected by the oracle BH procedure at level $\alpha' = \alpha(1-\eta)$, with high probability and with η small.

Proposition 4.2. *Recall* $\mathcal{A}_{n,m}$ (11). *Let* $\alpha, \gamma, \eta \in (0,1)$ *and let*

$$\gamma^*(\alpha,\eta) = \alpha^{-1}\eta^{-2}\frac{(3\log 2)(2+(\log 2)/3)(1+\eta)}{(1-\eta/(2\log 2))^2} > 0. \tag{12}$$

Then, for all $n, m \ge 1$ *with* $n \ge \gamma m$, *for all* $P \in \mathcal{A}_{n,m}$, *we have*

$$\mathbb{P}_{Z\sim P}(BH^*_{\alpha(1-\eta)} \subseteq \widehat{BH}_\alpha) \ge 1 - (1/2)^{3\gamma/\gamma^*(\alpha,\eta)-1}. \tag{13}$$

In particular, for all $n, m \geq 1$ with $n \geq \gamma m$,

$$\sup_{P \in \mathcal{A}_{n,m}} \{\mathbb{P}_{Z\sim P}(TDP(P, BH^*_{\alpha(1-\eta)}) > TDP(P, \widehat{BH}_\alpha))\} \leq (1/2)^{3\gamma/\gamma^*(\alpha,\eta)-1}$$

Proposition 4.2 is proved in Appendix D.1. It is based on a concentration argument of the empirical c.d.f. of the Y_i's, which relies on the independence assumption between the Y_i's. Note that taking γ much larger than $\gamma^*(\alpha, \eta)$ makes the probability in (13) arbitrarily close to 1. For illustration, if $\alpha = \eta = 0.1$, taking $\gamma = \gamma^*(\alpha, \eta) \approx 5928$ provides a probability in (13) at least 3/4. Note that, as is often the case in non asymptotical results, the value of this constant is an artefact of the proof. It has not to be interpreted has a meaningful value that separates feasible and unfeasible regimes.

4.2. Lower bound

The previous section shows that the power of $\widehat{\mathrm{BH}}$ is close to the one of the oracle BH procedure provided that n/m is sufficiently large. We can legitimately ask whether this condition is necessary. The following result addresses this point.

Proposition 4.3. *Recall $\mathcal{A}_{n,m}$ (11). Let $\alpha \in (0, 1)$ and $\eta \in [0, 1)$. Consider $n, m \geq 1$ with $n/m \leq 1/(4\alpha)$. Then*

$$\sup_{P \in \mathcal{A}_{n,m}} \{\mathbb{P}_{Z\sim P}(TDP(P, BH^*_{\alpha(1-\eta)}) > TDP(P, \widehat{BH}_\alpha))\} \geq 1 - 2\alpha. \tag{14}$$

Proposition 4.3 is proved in Appendix D.2. It is a consequence of the fact that all $\hat{p}$-values are larger than $1/(n+1)$ (see (6)), while $\widehat{\mathrm{BH}}$ controls the FDR (Theorem 3.1).

Propositions 4.2 and 4.3 are matching upper and lower bounds, up to constants. Put together, these results establish that the semi-supervised BH procedure achieves the boundary $n \asymp m/\alpha$: for $n \leq m/(4\alpha)$, there exists a configuration $P \in \mathcal{A}_{n,m}$ such that the power of $\widehat{\mathrm{BH}}_\alpha$ is less than the one of the oracle $\mathrm{BH}^*_{\alpha(1-\eta)}$ (with probability at least $1 - 2\alpha$), while for $n \gg m/\alpha$ all configurations $P \in \mathcal{A}_{n,m}$ are such that the power of $\widehat{\mathrm{BH}}_\alpha$ is larger than the one of the oracle (with probability arbitrarily close to 1).

4.3. Refinement to more favorable distributions

If there are enough alternatives, with enough signal strength, we show here that the boundary achieved by $\widehat{\mathrm{BH}}$ can be much better than $n \asymp m/\alpha$. We extend for this Proposition 4.2 and Proposition 4.3 to a specific set of "more favorable" distributions.

For $\alpha \in (0, 1)$, $n, m \geq 1$ and $1 \leq k \leq m$, consider the subset of $\mathcal{A}_{n,m}$ given by

$$\mathcal{A}_{n,m,k,\alpha,\beta} = \left\{P \in \mathcal{A}_{n,m} \ : \ m_1(P) \geq k,\ \mathbb{P}_{Z\sim P}\left(|\mathcal{H}_1(P) \cap \mathrm{BH}^*_{\alpha/2}| \leq k-1\right) \leq \beta\right\}.$$

In words, $\mathcal{A}_{n,m,k,\alpha,\beta}$ is the set of distributions such that at least k null hypotheses are false while the probability that the procedure $\mathrm{BH}^*_{\alpha/2}$ makes at most $k-1$ number of true discoveries is smaller than β. From an intuitive point of view, this means that the distribution contains at least k "detectable" alternatives, in the sense that they are detectable with large probability by the oracle itself (at level $\alpha/2$).

Now, the idea is that for a distribution $P \in \mathcal{A}_{n,m,k,\alpha,\beta}$, the threshold of the oracle procedure is at least $\alpha k/m$ with large probability, so that the precision $1/(n+1)$ of the $\hat{p}$-values is enough to mimic the power of the oracle BH if and only if $1/n \ll \alpha k/m$, that is, $nk \gg m$. The following result proves that this informal argument is correct.

Proposition 4.4. *Let $\alpha \in (0,1)$, $\eta \in (0,1/2)$ and $\beta \in (0,1)$. Then the following holds for $n, m \geq 1$ and $1 \leq k \leq m$:*

(i) if $nk/m \geq \gamma$, for some $\gamma > 0$,

$$\sup_{P \in \mathcal{A}_{n,m,k,\alpha,\beta}} \{\mathbb{P}_{Z\sim P}(\mathit{TDP}(P, \mathit{BH}^*_{\alpha(1-\eta)}) > \mathit{TDP}(P, \widehat{\mathit{BH}}_\alpha))\} \leq \beta + (1/2)^{3\gamma/\gamma^*(\alpha,\eta)-1},$$

where $\gamma^*(\alpha,\eta)$ is given by (12).

(ii) if $nk/m \leq 1/(4\alpha)$,

$$\sup_{P \in \mathcal{A}_{n,m,k,\alpha,\beta}} \{\mathbb{P}_{Z\sim P}(\mathit{TDP}(P, \mathit{BH}^*_{\alpha(1-\eta)}) > \mathit{TDP}(P, \widehat{\mathit{BH}}_\alpha))\} \geq 1 - \beta - 2\alpha.$$

Proposition 4.4 is proved in Appendix D.3. Point (i) above is an upper-bound: in particular, it shows that having $nk/m \geq \gamma^*(\alpha,\eta)$ is enough for $\widehat{\mathrm{BH}}_\alpha$ to mimic the power of the oracle $\mathrm{BH}^*_{\alpha(1-\eta)}$ with probability at least $1-\beta-1/4$ when the underlying distribution belongs to the set $\mathcal{A}_{n,m,k,\alpha,\beta}$. Interestingly, the condition $nk/m \geq \gamma^*(\alpha,\eta)$ is much weaker than the previous condition $n/m \geq \gamma^*(\alpha,\eta)$ when k gets large.

Point (ii) is a lower-bound showing that the order given in the upper-bound is correct. Together, (i) and (ii) ensure that the boundary achieved by $\widehat{\mathrm{BH}}_\alpha$ is $nk \asymp m$ on the distribution set $\mathcal{A}_{n,m,k,\alpha,\beta}$. In addition, when α gets small and $1/\alpha$ cannot be considered as a constant, our result is able to track the dependence in α; since $\gamma^*(\alpha,\eta)$ is of order $1/\alpha$ (see (12)), the boundary reads $nk \asymp m/\alpha$. In addition, the constant in $\gamma^*(\alpha,\eta)$ turns out to be largely overestimated. The effective transition for the feasible regime when k rejections are expected seems to be exactly at $nk = m/\alpha$ in the numerical experiments, see Section 6. This provides a useful "rule of thumb" for a practical use.

5. Optimality

For a fixed level α, the previous results show that the semi-supervised BH procedure $\widehat{\mathrm{BH}}_\alpha$ mimics the oracle BH procedure when $n \gg m$ both in terms of FDR

(Theorem 3.1) and power (Proposition 4.2). However, when $n \ll m$, while $\widehat{\mathrm{BH}}_\alpha$ still controls the FDR, it looses the power property (Proposition 4.3). Hence, it does not mimic the oracle in that regime. However, this does not exclude that a different procedure, that would use the data Z more cleverly, might be able to mimic the oracle when $n \ll m$. In this section, we show that *no procedure* can mimic the oracle in that regime (Theorem 5.1). This shows a general phase transition to the problem of mimicking the oracle (Corollary 5.3) and establishes that $\widehat{\mathrm{BH}}_\alpha$ achieves this transition, which thus delineates a kind of optimality satisfied by the semi-supervised BH procedure.

5.1. General lower bound

Recall $\mathcal{P}_{n,m}$ (10) and $\mathcal{A}_{n,m}$ (11). Taken together, Theorem 3.1 and Proposition 4.2 show that for any $\alpha, \eta, \gamma \in (0,1)$ the procedure $R = \widehat{\mathrm{BH}}_\alpha$ (as a sequence in $n, m \geq 1$) satisfies simultaneously the two following properties:

$$\sup_{\substack{n,m\geq 1\\ n\geq m\gamma}} \sup_{P\in\mathcal{P}_{n,m}} \{\mathrm{FDR}(P,R) - \mathrm{FDR}(P,\mathrm{BH}^*_\alpha)\} \leq \delta_1; \tag{15}$$

$$\sup_{\substack{n,m\geq 1\\ n\geq m\gamma}} \sup_{P\in\mathcal{A}_{n,m}} \mathbb{P}(\mathrm{TDP}(P,\mathrm{BH}^*_{\alpha(1-\eta)}) > \mathrm{TDP}(P,R)) \leq \delta_2. \tag{16}$$

for $\delta_1 = 0$ and $\delta_2 = (1/2)^{3\gamma/\gamma^*(\alpha,\eta)-1} > 0$. This quantifies how $\widehat{\mathrm{BH}}_\alpha$ mimics the oracle $(\mathrm{BH}^*_\alpha)_{\alpha\in(0,1)}$ both in terms of FDR and power when $\gamma/\gamma^*(\alpha,\eta)$ grows.

In the regime where γ is too small, the following result shows that achieving simultaneously (15) and (16) is not possible.

Theorem 5.1. *Recall $\mathcal{P}_{n,m}$ (10) and $\mathcal{A}_{n,m}$ (11). Let $\alpha \in (0,1/4)$ and $\gamma, \eta \in (0,1)$ and let*

$$\gamma_*(\alpha,\eta) = (1 + (\alpha(1-\eta))^{-1/2})^{-3}/64, \tag{17}$$

Consider $n, m \geq 1$ with $n \leq \gamma m$. Then for any procedure R (based only on Z), one has either

$$\sup_{P\in\mathcal{P}_{n,m}} \{FDR(P,R) - FDR(BH^*_\alpha, R)\} \geq 1/2 - \alpha \tag{18}$$

or

$$\sup_{P\in\mathcal{A}_{n,m}} \{\mathbb{P}_{Z\sim P}(TDP(P,R) < TDP(P,BH^*_{\alpha(1-\eta)}))\} \geq 1/2 - (1/4)(\gamma/\gamma_*(\alpha,\eta))^{1/3}. \tag{19}$$

As a result, no procedure R (as a sequence in $n, m \geq 1$) can satisfy simultaneously (15) and (16) for $\delta_1 < 1/2 - \alpha$ and $\delta_2 < 1/2 - (1/4)(\gamma/\gamma_*(\alpha,\eta))^{1/3}$.

The proof of Theorem 5.1 is given in Appendix D.4. It relies on building two nearly indistinguishable configurations $Q_1, Q_2 \in \mathcal{P}_{n,m}$ such that: either $\mathrm{FDR}(Q_1, R)$ is large, or with large probability under Q_2, R makes no discovery while the oracle makes at least one correct discovery. Note that the result in

Theorem 5.1 is silent if $1/2 - (1/4)(\gamma/\gamma_*(\alpha,\eta))^{1/3} \leq 0$, that is, $\gamma \geq 8\gamma_*(\alpha,\eta)$. Hence, Theorem 5.1 is only informative whenever $\gamma < 8\gamma_*(\alpha,\eta)$. When $\gamma < \gamma_*(\alpha,\eta)$, the RHS of (19) is in addition strictly larger than 1/4.

5.2. Phase transition

Let us elaborate further on the phase transition that we have put forward. To this end, we introduce the following definition.

Definition 5.2. For $\delta_1 \in [0,1)$, $\delta_2 \in [0,1)$, $\gamma > 0$ and $\alpha, \eta \in (0,1)$, a procedure R is said to be (δ_1,δ_2)-mimicking the oracle $(\mathrm{BH}^*_\alpha)_{\alpha\in(0,1)}$, for a training-to-test sample size at least γ, and a level α with relaxation η, in short R is $MO(\gamma,\alpha,\eta,\delta_1,\delta_2)$, when (15) and (16) simultaneously hold for these values of $\delta_1,\delta_2,\alpha,\gamma,\eta$.

According to this definition, Theorem 3.1, Proposition 4.2 and Theorem 5.1 can be combined as follows:

Corollary 5.3. *Let $\alpha \in (0,1/4)$, $\eta \in (0,1)$, and consider $\gamma^*(\alpha,\eta)$ defined by (12) and $\gamma_*(\alpha,\eta)$ defined by (17). Then for any $\gamma > 0$:*

(i) If $\gamma < \gamma_*(\alpha,\eta)$, then there exists no $MO(\gamma,\alpha,\eta,\delta_1,\delta_2)$ procedure for any possible value of $\delta_1,\delta_2 \in (0,1/4]$. This is even true for $\delta_1 < 1/2-\alpha$ and $\delta_2 < 1/2 - (1/4)(\gamma/(\gamma_*(\alpha,\eta)))^{1/3}$);

(ii) If $\gamma \geq \gamma^*(\alpha,\eta)$ then there exists an $MO(\gamma,\alpha,\eta,\delta_1,\delta_2)$ procedure for some values of $\delta_1,\delta_2 \in [0,1/4]$. This is achieved by $\widehat{BH}_\alpha$, even with $\delta_1 = 0$ and $\delta_2 = (1/2)^{3\gamma/\gamma^*(\alpha,\eta)-1}$.

This phase transition is illustrated in Figure 1 in the introduction of the paper. The transition is provided under the simpler form $n = m/\alpha$ for comparison with Section 4.

Note that the impossibility result above is a "worst case" analysis over the distributions $P \in \mathcal{P}_{n,m}$ (FDR) and $P \in \mathcal{A}_{n,m}$ (power), that is, suprema are taken in (18) and (19). In particular, under the more stringent assumption $P \in \mathcal{A}_{n,m,k,\alpha,\beta}$, mimicking the oracle becomes already possible whenever $n \gtrsim m/(\alpha k)$ (as reported in Figure 1 for $k = 3$ or 100).

This general phase transition is in line with the recent results by Roquain and Verzelen (2020b). Nevertheless, their setting is markedly different: it is unsupervised (no NTS) and the null distribution is assumed to belong to the Gaussian distribution family with unknown mean and variance. The phase transition found there was $\ell \asymp m/\log(m)$ (with our notation) where ℓ is a lower bound on the number of alternatives $m_1(P)$ (with no signal strength assumption). Here, the situation is notably different, with a boundary function of the length n of the NTS. The situation is also very different in terms of FDR control: the mimicking procedure $\widehat{\mathrm{BH}}$ provides here an FDR control both above and below the transition boundary, while such property is not possible in the setting of Roquain and Verzelen (2020b) (as proved in Corollary 3.3 therein).

6. Numerical illustrations

This section provides several numerical illustrations for the theoretical findings derived in Sections 3 and 4.

6.1. Simulation setting

While our experiments mostly focus on the two BH-type procedures $\widehat{\mathrm{BH}}$ and BH^*, we will also consider other competitors: $\widetilde{\mathrm{BH}}$, which is the BH procedure applied to the unbiased $\tilde{p}$-values defined by (5) (Section 2.2) and the "naive" procedures $\widehat{\mathrm{BY}}$ and $\widehat{\mathrm{BH}}_{\mathrm{Split}}$ described in Section 1.3. Also, for simplicity, the way to evaluate how the power of $\widehat{\mathrm{BH}}$ mimics the one of BH^* slightly departs from our theoretical study: first, we compare $\widehat{\mathrm{BH}}$ to the oracle BH^* taken at the same level α (say, $\eta = 0$ with the notation of Section 4). This makes the power mimicking more challenging. Second, to stick with the standard way of comparing procedures (for BH^*, $\widehat{\mathrm{BH}}$ or their competitors), the considered power criterion is simply the TDR (3) (average of the TDP). Unless specified, the setting is Gaussian with a null distribution $P_0 = \mathcal{N}(0,1)$ and an alternative $\mathcal{N}(\mu,1)$, for a given value of $\mu > 0$. Across the sections below, we made various choices of n, m and of the sparsity m_1 (number of alternatives). We sometimes fix the level α to the (unusual large) value 0.5 for better visibility of the curves and faster computation time, but the results scale accordingly for smaller values of α (the interested reader can refer to Figures 9 and 11 for situations where $\alpha = 0.2$). Finally, the FDR (resp. TDR) curves are here estimated by Monte-Carlo simulations. The plots show the estimates $\widehat{\mathrm{FDR}}$ and $\widehat{\mathrm{TDR}}$ with two error bars: one estimating the standard deviation of $\widehat{\mathrm{FDR}}$ (resp. $\widehat{\mathrm{TDR}}$) and one estimating the standard variation of FDP (resp. TDP) (these two deviations being proportional).

6.2. FDR control under the full null

The first experiment concerns the case where $m_0 = m$, which corresponds to the so-called "full null" configuration where there is no alternative. We consider two dependence framework: the independent case (all Z_i's independent) and the negatively equicorrelated case described in Example 3.2. Recall that $\widehat{\mathrm{BH}}$ is proved to control the FDR at level α in both cases (Theorem 3.1), while BH^* is only proved to control the FDR at level $\alpha m_0/m = \alpha$ in the independent case. Also, $\widetilde{\mathrm{BH}}$ (BH procedure applied to the unbiased $\tilde{p}$-values defined in Section 2.2) is not proved to control the FDR since the $\tilde{p}$-values do not satisfy the super-uniformity property (4).

Figure 4 displays the obtained FDR curves for BH^* (green), $\widehat{\mathrm{BH}}$ (blue) and $\widetilde{\mathrm{BH}}$ (red). The obtained results are consistent with our theoretical findings: negative correlations induce an FDR of BH^* slightly above the targeted level, although this effect tends to reduce when n gets larger. This is because the

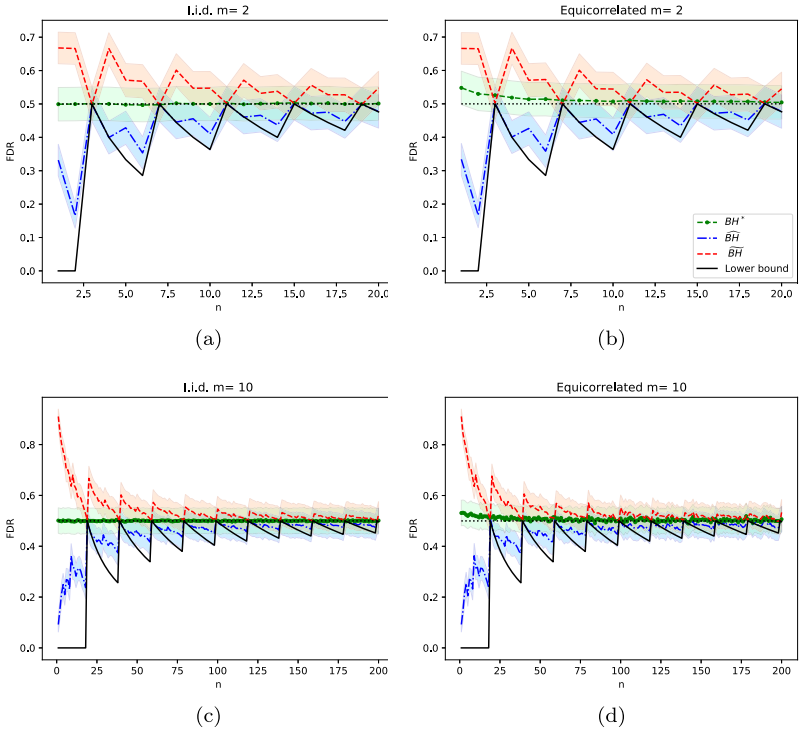


FIG 4. *FDR result in the case of i.i.d. samples (left column) and Gaussian negative equicorrelation (right). The cases* $m = 2$ *(top row),* $m = 10$ *(bottom row) have been investigated with respectively* 10^5 *and* 10^4 *Monte Carlo simulations. The* 2σ *confidence intervals on the estimated FDR are not visible. The standard deviation (divided by a factor* 10*) of the FDP is shown by shaded areas. The figure shows the results for BH* (green),* $\widehat{BH}$ *(blue),* $\widetilde{BH}$ *(red, see text), and the lower bound of Theorem 3.1 (black).*

negative correlation $\rho = -(m+n-1)^{-1}$ decreases (in absolute value) when n grows. As expected, $\widehat{\mathrm{BH}}$ maintains the FDR control is any case. Meanwhile, $\widetilde{\mathrm{BH}}$ fails to control the FDR in any case, except for some values of n where it has the same FDR value as $\widehat{\mathrm{BH}}$ (see also Remark 3.5). Hence, we shall discard $\widetilde{\mathrm{BH}}$ from our plots in the sequel. Interestingly, we also displayed the lower bound of Theorem 3.1 in Figure 4: while it correctly lower bounds the estimated FDR of $\widehat{\mathrm{BH}}$ for any n, it illustrates that the FDR is exactly α for $n \in \{3, 7, 11, 15, \dots\}$ as the theory establishes (the curves might also suggest that the lower bound is sharp for $n \in \{4, 8, 12, 16, \dots\}$, which is not covered by our theory). Finally, note that these results are in expectation: as shown by the shaded areas, there can be large variations for particular samples. This is inherent to the BH procedure when the number of discoveries is not large.

6.3. Power study

Figure 5 compares the performances of the procedures BH^* (dark green and khaki) and $\widehat{\mathrm{BH}}$ (dark blue and cyan) in terms of FDR (dark colors) and TDP (light colors) in the dense case where $m_1 = m_0 = \frac{m}{2}$, with $\mu = 1$ (left column) and $\mu = 2$ (right column). Regarding the FDR first, the plots show that the FDR of $\widehat{\mathrm{BH}}$ tends to the oracle FDR (which is $0.25 = \alpha\frac{m_0}{m} = \frac{\alpha}{2}$ here). For a fixed value of n, the convergence is faster for smaller values of m. This is coherent with Theorem 3.1, ensuring that the FDR of $\widehat{\mathrm{BH}}$ is equal to $\alpha/2$ for $n = m/\alpha - 1$. On the other hand, the variance in the FDR (blue shaded area) is smaller at fixed n when m increases, because the larger sample size tends to stabilize the result.

Turning to the power results, the plots show that the power of $\widehat{\mathrm{BH}}$ also tends to that of BH^* in this sparsity regime, with also faster convergence for smaller values of m (at fixed n). This is well expected from the "rule of thumb" delineated in Section 4.3 and ensuring that the transition occurs for $n \approx m/(\max(1,k)\alpha)$ where k is a lower bound on the typical true discovery number of the oracle. Given the displayed results, the value of k could be chosen around $(2, 4, 20, 40)$, so that this rule would predict a transition for n occurring around $(10, 5, 10, 5)$ (top-left, top-right, bottom-left, bottom-right). Strikingly enough, the transitions indeed occur at these points in the different TDR curves.

The sparse case where $m_1 = 1$ is considered in Figure 6, with $\mu = 1$ (left column) and $\mu = 3$ (right column) and a slightly increased range for n. Here, the oracle FDR is 0.45 for $m = 10$ and 0.495 for $m = 100$. The observations made regarding the FDR and TDR in Figure 5 are qualitatively the same. Moreover, in the sparse case, the convergence to the asymptotic regime is slower than in the dense case, while increasing m for fixed n slows down more significantly the convergence than in the dense case. This is coherent with the rule of thumb $n \approx m/(\max(1,k)\alpha)$, predicting that the transition n occurs around $m/\alpha = 2m$ here (only one alternative here). In addition, it is apparent on the plots that the value of the transition n predicted by this rule turns out to be particularly well adjusted, at least in this simulation setup.

Finally, Figure 7 compares the FDR and TDR of the procedures BH^* and $\widehat{\mathrm{BH}}$ for larger values of m and n and $\alpha = 0.2$. We fix $m = 10^3$ and the size of the NTS ranges from $n = 1$ to 5×10^4. In each plot, we see that the performances of $\widehat{\mathrm{BH}}$ indeed increase with n. Despite the increased signal amplitude in the sparse case, the situation is more difficult both in terms of convergence (which is slower) and of variance in the FDP and TDP (which are larger). Interestingly, this corroborates again the rule of thumb predicting a transition n around 20 and 625 (for the choices $k \approx 250$ and $k \approx 8$) for the dense and sparse situations, respectively.

6.4. Additional experiments

Appendix F presents the following additional experiments: first, Appendix F.1

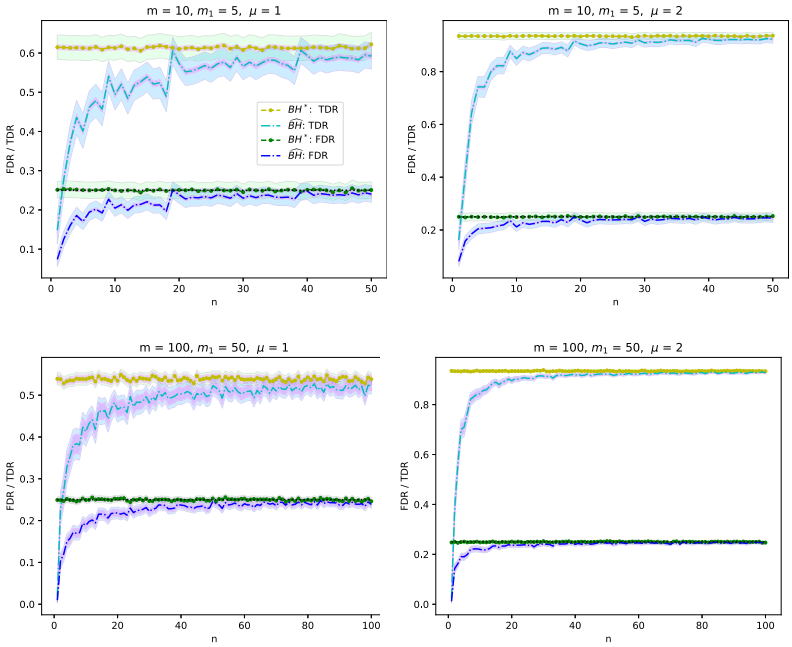


FIG 5. *FDR and TDR results for the dense case:* $m_1 = \frac{m}{2}$*, with* $\mu = 1$ *(left column) and* $\mu = 2$ *(right column). The number of tests* m *equals* 10 *in the top row and* 100 *in the bottom row. The number of Monte Carlo simulations used for estimating the FDR and TDR is* 10^4 *(top row) and* 10^3 *(bottom row). The* 2σ *confidence interval on the estimated FDR and TDR is plotted in magenta. In all plots the standard deviation (divided by* 10*) of the FDP and TDP are shown in shaded green for* BH^* *and shaded blue for* $\widehat{BH}$*.*

presents a comparison with the naive procedures $\widehat{\mathrm{BY}}$ and $\widehat{\mathrm{BH}}_{\mathrm{Split}}$. They are both shown to be over-conservative and much less powerful than $\widehat{\mathrm{BH}}$. Second, a case study with a Student distribution, leading to similar conclusions, is presented in Appendix F.2. Third, Appendix F.3 is devoted to simulations for very small values of n ($n = 5$ or 10) with increasing values of m: it shows that $\widehat{\mathrm{BH}}$ can achieve oracle performances in that dense case, regardless of m.

7. Application

One of the major scientific goals of the MUSE integral field spectrograph, which is installed on one of the 8 m telescopes at the Very Large Telescope in Chile, is the detection of distant and consequently ultra faint galaxies in the early Universe. MUSE delivers 3-dimensional datacubes (two spatial dimensions and one spectral dimension) composed of images taken in different wavelengths channels of the visible spectrum. The values of the data samples correspond to light fluxes. Ordinary datacubes are composed with a pile of 300×300 pixels im-

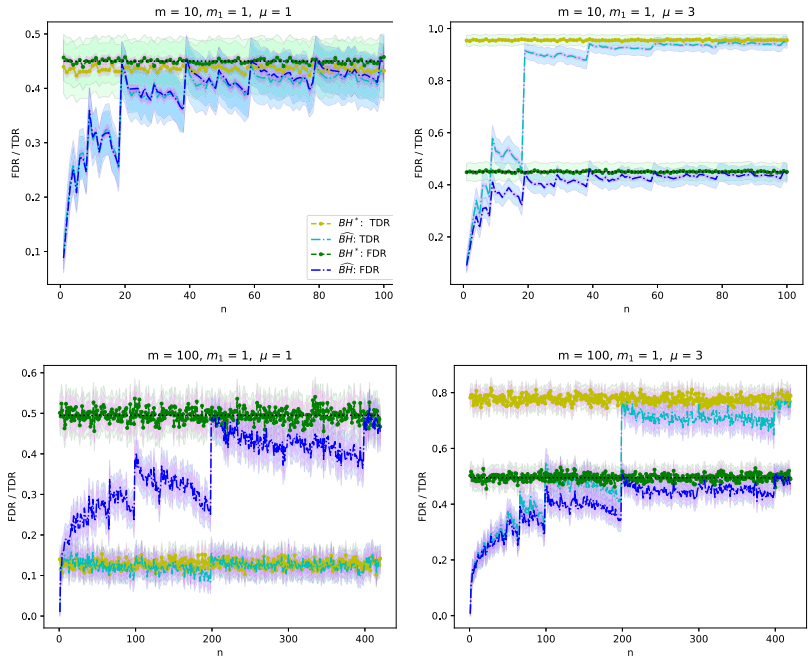


FIG 6. *FDR and TDR results for the sparse case:* $m_1 = 1$, *with* $\mu = 1$ *(left column) and* $\mu = 3$ *(right column). The number of tests* m *equals* 10 *for the top row and* 100 *for the bottom row. The number of Monte Carlo simulations used for estimating the FDR and TDR is* 10^4 *(top row) and* 10^3 *(bottom row). The* 2σ *confidence interval on the estimated FDR and TDR is plotted in magenta. In all plots the standard deviation (divided by* 10*) of the FDP and TDP are shown in shaded green for* BH^* *and shaded blue for* $\widehat{BH}$.

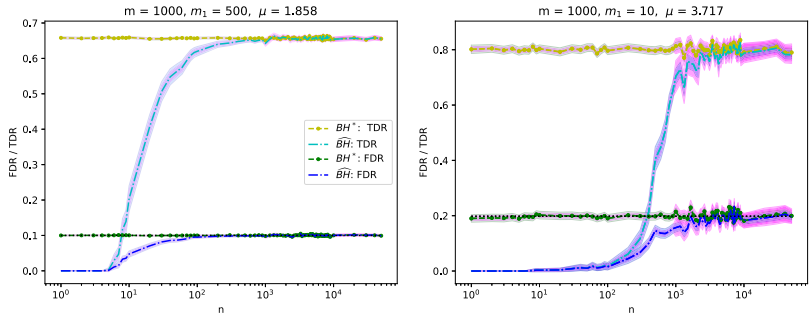


FIG 7. *FDR and TDR results for* $m = 10^3$ *as a function of* n*, for* $\alpha = 0.2$*. Left: dense case:* $m_1 = 500$ *and* $\mu = \frac{1}{2}\sqrt{2\log m} \approx 1.86$*. Right:* $m_1 = 10$ *and* $\mu = \sqrt{2\log m} \approx 3.72$ *(right). The number of Monte Carlo simulations used for estimating the FDR and TDR is* 10^3 *for* $n < 10^3$ *and* 10^2 *otherwise. The* 2σ *confidence interval on the estimated FDR and TDR is plotted in magenta. In all plots the standard deviation (divided by* 10*) of the FDP and TDP are shown in shaded green for* BH^* *and shaded blue for* $\widehat{BH}$.

ages in 3700 consecutive visible wavelengths, leading to more than 300 millions voxels.

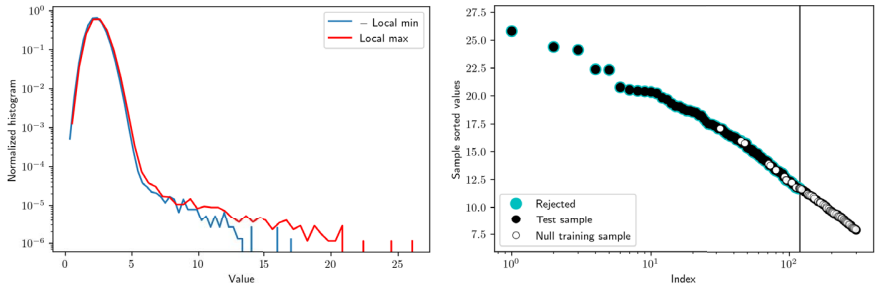
After multiple calibration and preprocessing stages, the problem of detecting faint galaxies boils down to a typical needle in a haystack problem. The haystack is the datacube, which can be considered as a discrete-valued 3-dimensional random process. This process is generated by various noise sources and by the residual perturbations of numerous bright sources. Consequently, the statistics of the random process are poorly constrained. In this haystack, each needle (there are hundreds of them) is a small group of connected voxels, centered on the galaxy's position, in which the flux locally increases.

A dedicated detection strategy, proposed by Mary et al. (2020) and further exploited by Bacon et al. (2021), consists in considering as final test statistics the 3-dimensional local maxima of the processed datacube. In the resulting testing problem, there is one null hypothesis linked to each of the m local maxima, with m typically in the range $[10^5,\ 10^6]$. If we denote by x, y, z the position of a particular local maximum, we test $H_{0,x,y,z}$: "There is no galaxy centred at position (x, y, z)", against $H_{1,x,y,z}$: "There is one galaxy centred at this position" and the considered error criterion is the FDR.

As evoked above, the distribution of the local maxima under the null hypothesis is fairly unknown. To circumvent this difficulty, Mary et al. (2020) proposed to use the population of the opposite values of the local minima (say, Y_i, in number n) as an independent "proxy" (a NTS) for the local maxima (say, X_i, in number m). They reported numerical simulations suggesting that a procedure close to the Benjamini-Hochberg procedure using p-values computed from this NTS controls the FDR. This astrophysical application involves a common but unknown distribution P_0 under the null hypothesis and the possibility of using a NTS to improve the control of the FDR. The sample sizes considered are $n = 2.3 \times 10^6$ and $m = 3.3 \times 10^6$ so $n < m$ and both are large. It is thus interesting to see which light the present study sheds on this initial approach.

Let us check that the setting described in Section 2.1 is reasonable for this empirical study. First, the distribution of the NTS correctly matches the null distribution of the test sample. Indeed, the empirical distribution of the values of the NTS and of the test sample are shown in Figure 8, left panel. The similarity of the two distributions in the left and central parts suggests that the NTS (blue) can serve as a useful proxy for the test sample (red). The right tail of the test sample is logically heavier owing to the presence of galaxies, which tend to shift the values of the local maxima upwards. Second, the measurements at the local extrema are plausibly independent, because the typical distance between local extrema is larger than the size of the filter used in the preprocessing. This justifies that Assumption (Indep) is reasonably true here.

While the procedure proposed by Mary et al. (2020) is very close to the $\widehat{\text{BH}}$ procedure with Algorithm 1, it differs in the following point. Instead of using $\text{FDP} = \frac{V+1}{K}\frac{m}{n+1}$ (see step 4 of Algorithm 1), Mary et al. (2020) use $\text{FDP} = \frac{V}{K}\frac{N_1}{N_0}$, where N_1 (resp. N_0) are the number of voxels of the region where the local extrema of the test sample (resp., the NTS) are computed. Because



(a) (b)

FIG 8. *MUSE application example. (a) Empirical distributions of the values of the NTS (the Y_i, computed as the opposite of the local minima, in blue) and of the test sample (the X_i, local maxima, in red). (b) Results of Algorithm 1 (same color code as in Figure 2). Only the* 700 *largest sample values are shown. The black vertical line indicates the rejection threshold* $K = 105$ *and* 14 *samples of the NTS are above this threshold.*

n is large, V is large as well and using V instead of $V + 1$ has no numerical impact in this regime. The normalization factors are in fact very similar as well, with $\frac{m}{n+1} \approx 1.440$ and $\frac{N_1}{N_0} \approx 1.442$. In effect, it turns out that there is no numerical difference in running these two versions of Algorithm 1: in both cases, the procedure rejects exactly 105 local maxima at target FDR $\alpha = 0.2$, a situation shown in the right panel of Figure 8. The rejected local maxima in Mary et al. (2020) being the same as those rejected by $\widehat{\text{BH}}$, the discovery set inherits the properties delineated in the present work: first, the FDR control is established from Theorem 3.1. Second, we have $n \approx m$ while both are large. This means that we are just at the border of the boundary identified in Section 5.2, so the theory is silent for this case. Nevertheless, the distribution of the data exhibits some minimum amount of signal, perhaps $k \gtrsim 50$ fairly detectable alternatives. Hence, the refined upper-bound given in Proposition 4.4 can also be applied: since the training-to-test ratio n/m is above the boundary, that is, $n/m \approx 1$ is much larger than $1/(k\alpha) \lesssim 0.1$, the power of $\widehat{\text{BH}}$ should be close to the one of the oracle for this data set.

To conclude, the present paper illustrates that $\widehat{\text{BH}}$, together with our theoretical findings, delivers interpretable and useful results for common practice. Meanwhile, it validates the use of the procedure proposed in Mary et al. (2020) on this particular data set.

8. Conclusion and discussion

8.1. Summary

In a nutshell, this paper evaluated how classical multiple testing methodology can generalize when replacing the knowledge of the null distribution P_0 by examples $Y_1, \ldots, Y_n$ following this null. While this situation is very frequent in

practice, it has only been scarcely studied so far and this paper contributed to fill this gap. The FDR control guarantee holds whatever n, m, with no assumption on P_0 and for any marginal alternative, with a bound $\alpha m_0/m$ (achieved when $\alpha(n+1)/m$ is an integer), which is similar to the result obtained in the original work of Benjamini and Hochberg (1995) in case where P_0 is known. In addition, the power is comparable to the one of the oracle when $n \gtrsim m/(\alpha \max(1,k))$, where k is a confidence lower bound on the number of true discoveries made by the oracle. This "rule of thumb" has been both validated by theory and numerical experiments. Finally, we demonstrated that our work brought a theoretical support and thus more interpretability in a worked-out application to recent breakthrough findings in astrophysics. In practice, our "rule of thumb" $n \gtrsim m/(\alpha \max(1,k))$ can be used as follows: if the user has no strong prior belief in a minimum number k of discoveries, choosing $k = 0$ might be safer, which leads to the condition $n \gtrsim m/\alpha$. By contrast, if k can be accurately guessed a priori, the less demanding condition $n \gtrsim m/\alpha k$ can be opted for.

This work also completed the picture by exhibiting a theoretical intrinsic limitation of the semi-supervised multiple testing setting when the null training sample is not populated enough. It is impossible to control the FDR while mimicking the oracle power for $n \lesssim m$ when letting the sparsity and the distribution of the alternative arbitrary. This delineates a setting-intrinsic phase transition at $n \asymp m$.

8.2. Future work

Given that semi-supervised multiple testing setting is versatile, our work raises a number of new perspectives. For instance, in recent machine learning, this setting conveniently bypasses model assumptions on P_0 and only needs a number of null examples, that can be generated by a suitable "blackbox". Nevertheless, in order to avoid potential bias in the null training sample, this blackbox should be properly calibrated with significant prior calibrations and preprocessing steps. While building such an approach deserves an entire devoted study, we anticipate that studying the robustness of the procedure $\widehat{\mathrm{BH}}$ with respect to the NTS is a key point: what about the case where $Y_1, \ldots, Y_n$ are i.i.d. $\sim P_0'$ with $P_0' \approx P_0$?

Another avenue for future work is to decline recent advances in multiple testing into this semi-supervised setting. For instance, while $\widehat{\mathrm{BH}}$ is devoted to the FDR criterion, an interesting and challenging issue is to design semi-supervised counterparts suitable for other criteria, as FDX Genovese and Wasserman (2004), online FDR Foster and Stine (2008), Xu and Ramdas (2021) or post hoc bounds Genovese and Wasserman (2006), Goeman and Solari (2011). In particular, since the variability of the FDP of $\widehat{\mathrm{BH}}$ is increased by the NTS, considering criteria accounting for this effect seems particularly interesting. Since various dependence assumptions are used in such studies, we also expect that our main assumption (Exch) can be relaxed in some of these frameworks.

Finally, proper calibrations of the individual tests sometimes require to consider hypothesis-dependent null distributions, that is, null distributions $P_{0,i}$ that

depend on $i \in \{1, \dots, m\}$ (see, e.g., Sulis et al., 2017, 2020 for a concrete example). Since m null training samples should be considered in that case, it poses a complexity issue and generalizing our result to this setting is both theoretically challenging and useful to support or improve procedures used in common practice.

Appendix A: By-product 1: Blackbox BH procedure

A.1. Setting and procedure

In this section, we consider the same formal setting and notation as in Section 2.1, except that the test statistics $X_1, \dots, X_m$ are given along with a "blackbox sampler" able to produce i.i.d. realizations of the null P_0, even if P_0 is not known. As in the motivations described in Section 1.1, such a blackbox can come from an external code implemented by an expert of the application domain, or from a machine learning program that has been sufficiently trained. Our work easily allows to design a multiple testing inference in that situation. Namely, Algorithm 2 below can be used to produce a sampled BH procedure, that we call the Blackbox BH procedure (bbBH). By Theorem 3.1, the bbBH procedure achieves an FDR equal to $\alpha m_0/m$ (when α is a rational number), provided that $(X_i, i \in \mathcal{H}_0)$ are i.i.d. $\sim P_0$ and independent of $(X_i, i \notin \mathcal{H}_0)$. Also, since n is chosen so that $(n+1)\alpha/m \geq 1$, it is just above the boundary put forward in Section 4, which might indicate that the power of bbBH should be comparable to that of the oracle.

Algorithm 2: Blackbox BH procedure

Data: $X = (X_1, \dots, X_m) \in \mathbb{R}^m$, a nominal level $\alpha \in (0, 1)$ (assumed to be a rational number) and a blackbox sampler of the null distribution P_0

1. Choose $n \geq 1$ the smallest integer such that $(n+1)\alpha/m$ is an integer
2. Sample $(Y_1, \dots, Y_n)$ i.i.d. according to the null distribution P_0
3. Apply the semi-supervised BH procedure $\widehat{\mathrm{BH}}_\alpha$ to $Z = (Y, X)$, see Algorithm 1

Result: Reject the nulls in the set $\widehat{\mathrm{BH}}_\alpha$.

A.2. Illustration with simultaneous likelihood ratio tests

To illustrate further the interest of the bbBH procedure, we consider in this section the problem of controlling the FDR while choosing the best individual test statistics. To this end, let us consider the common setup where we observe m independent measurements $T_1, \dots, T_m \in \mathbb{R}$, with T_i either distributed as a null distribution G_0 or as an alternative distribution G_1, where G_0 and G_1 are *known* distribution with densities g_0 and g_1, respectively. For each $i \in \{1, \dots, m\}$, we consider a likelihood ratio test of the null hypothesis $H_{0,i}$: "$T_i \sim G_0$" against

the alternative $H_{1,i}$: "$T_i \sim G_1$". It rejects the null whenever $\{X_i \geq c\}$ for $X_i = g_1(T_i)/g_0(T_i)$ (with the convention $X_i = +\infty$ if $g_0(T_i) = 0$) and some constant $c > 0$ such that $\bar{F}_0(c) = \alpha$ with

$$\bar{F}_0(t) = \int_{\mathbb{R}} \mathbb{1}_{\{g_1(u) > t g_0(u)\}} g_0(u) du, \ \ t \geq 0. \tag{20}$$

Denote P_0 the distribution of X_i under G_0 and assume that $\bar{F}_0$ is continuous and decreasing on the support of P_0. The oracle BH procedure BH*, which is not accessible in general, can be nevertheless approximated in this setting via a numerical approximation of the function $\bar{F}_0$. By contrast, we can also build a "blackbox" that generates realizations of P_0 by simulating $T_1, \ldots, T_n$ i.i.d. $\sim G_0$ and then letting $Y_i = g_1(T_i)/g_0(T_i)$, $1 \leq i \leq n$. Hence, we can apply the bbBH procedure (Algorithm 2) to control the FDR at level α in this model (and even have an FDR equal to $\alpha m_0/m$), while having a power close to the one of the oracle.

For comparisons, we also introduce two other procedures: first, the BH procedure directly applied to the original test statistics T_i's (with respect to the known null G_0), which is referred to as BH0 below. Compared to bbBH, BH0 has the advantage to be not-random. However, since the individual tests based on the T_i are less powerful than those based on the likelihood ratio X_i, bbBH is in general more powerful than BH0. The second procedure is the classical FDR controlling method based on the local FDR values Efron et al. (2001), Sun and Cai (2007), denoted by "locfdr", which can be used specifically for this example, see Section F.4 for more details.

The performances of these procedures are illustrated by a numerical experiment in Appendix F.4. The conclusions of this experiment are as follows:

- All procedures correctly control the FDR;
- As expected bbBH, BH* and locfdr have better power than BH0;
- The two procedures locfdr and bbBH both mimic the power of the oracle BH*, although bbBH is better adjusted for m small, while locfdr is slightly less variant.

Overall, this section validates the use of bbBH in a "toy blackbox setting" where alternative procedures can be employed. This suggests that bbBH will perform favorably in general blackbox settings for which no such alternative exists.

Appendix B: By-product 2: The randomized BH procedure

Let us consider, only for the present section, the usual framework where the null distribution is known and no NTS is given. In particular, BH* boils down to the usual BH procedure.

Recall that an important part of multiple testing literature is devoted to find procedures that control rigorously the FDR at level α while maximizing the power. We emphasize that, in this framework, having an FDR equal to $\alpha + 10^{-10}$ (say) is not allowed: the inequality FDR $\leq \alpha$ must hold under any configuration

of the model, which is particularly challenging when negative dependences are possible. For instance, we refer to the very recent work of Fithian and Lei (2020) (see also references therein), that aims at modifying the BH procedure in order to control the FDR under negative dependence. The point of this section is to show that Theorem 3.1 and Example 3.2 allow to solve this problem in a simple way for some (admittedly specific) dependence structure.

Assume that X is an m-dimensional Gaussian equi-correlated vector with individual variances equal to 1 and with *known* covariance $\rho \in [-1/m, 0)$. Consider $n = n(\rho, m) \geq 1$ the largest integer so that $\rho \geq -1/(n+m-1)$, that is, $n = \lfloor -\rho^{-1} - m + 1 \rfloor$ and generate a n-sample $Y_1, \ldots, Y_n$ such that (Y, X) is Gaussian equi-correlated ρ. This can be done easily via Proposition E.3. Then Theorem 3.1 provides that the procedure $\widehat{\mathrm{BH}}_\alpha$ controls the FDR at level α. Here, since the NTS is generated by the user, this procedure can be seen as a *randomized BH procedure*, (randBH in short). Algorithm 3 gives the full steps to implement randBH.

Algorithm 3: Randomized BH procedure

Data: $X = (X_1, \ldots, X_m) \in \mathbb{R}^m$, $\rho \in [-1/m, 0)$, α

1. Compute $n \geq 1$ the largest integer so that $\rho \geq -1/(n+m-1)$, that is, $n = \lfloor -\rho^{-1} - m + 1 \rfloor$
2. Let $T = X$
3. For k from m to $n+m-1$:
   - draw $U \sim \mathcal{N}(0,1)$ independently of the rest
   - let $T_{k+1} = \frac{\rho}{1+(k-1)\rho}(T_1 + \cdots + T_k) + \left(1 - k\frac{\rho^2}{1+(k-1)\rho}\right)^{1/2} U$
   - let $T = (T_1, \ldots, T_{k+1})$
4. Let $Y = (T_{m+1}, \ldots, T_{n+m})$
5. Apply the semi-supervised BH procedure $\widehat{\mathrm{BH}}_\alpha$ to $Z = (Y, X)$, see Algorithm 1

Result: Reject the nulls in the set $\widehat{\mathrm{BH}}_\alpha$.

Also, we would like to make a disclaimer: we do not pretend that RandBH is applicable in general practice, at least under the current form, because it is linked to a too specific dependence structure. Rather, the message is that randomization (plus using p-values biased upwards) can help the BH procedure to be more robust with respect to negative dependencies. We think that this intriguing side result is an important proof of concept.

Finally, this phenomenon can be derived for other negative dependence structures: however, the reachable distributions of $(X_i, i \in \mathcal{H}_0)$ should necessarily be expressible as a marginal of a larger vector $(Y_1, \ldots, Y_n, X_i, i \in \mathcal{H}_0)$ that is exchangeable in order to satisfy (Exch).

Appendix C: Proof of Theorem 3.1

We assume throughout the proof that $\mathcal{H}_0 = \{1, \dots, m_0\}$ without loss of generality.

C.1. Proof of Lemma 2.2

Let us recall that the $\hat{p}$-values are given by $\widehat{p}_i = \widehat{F}_0(X_i)$, $1 \le i \le m$, with $\widehat{F}_0$ given by (6). Hence, they can be ordered as $\widehat{p}_{(1)} \le \cdots \le \widehat{p}_{(m)}$ with a permutation also ensuring $X_{(1)} \ge \cdots \ge X_{(m)}$ and $\widehat{p}_{(k)} = \widehat{F}_0(X_{(k)})$, $1 \le k \le m$. Now, by definition

$$\widehat{\mathrm{BH}}_\alpha = \{i \in \{1, \dots, m\} : \widehat{p}_i \le \widehat{p}_{(\widehat{k})}\}, \widehat{k} = \max\{k \in \{0, 1, \dots, m\} : \widehat{p}_{(k)} \le \alpha k/m\}.$$

Since $\frac{m\widehat{p}_{(k)}}{k} = \frac{m}{n+1} \frac{1+\sum_{j=1}^n \mathbb{1}_{\{Y_j \ge X_{(k)}\}}}{\sum_{i=1}^m \mathbb{1}_{\{X_i \ge X_{(k)}\}}}$, we have by definition of $\hat{t}$ that $\hat{t} = X_{(\hat{k})}$. To conclude, we only have to prove the equality

$$\{i \in \{1, \dots, m\} \,:\, \widehat{F}_0(X_i) \le \widehat{F}_0(X_{(\hat{k})})\} = \{i \in \{1, \dots, m\} \,:\, X_i \ge X_{(\hat{k})}\}.$$

The inclusion $\supseteq$ is clear by the monotonicity of $\widehat{F}_0$. In addition, by definition of $\hat{k}$, we have $X_{(\hat{k}+1)} < X_{(\hat{k})}$, so that the set on the right-hand-side is of cardinal $\hat{k}$. Since, again by definition of $\hat{k}$, the set on the left-hand-side is also of cardinal $\hat{k}$, the two sets are equal.

C.2. Reformulation of $\widehat{BH}$

Let us reformulate $\widehat{\mathrm{BH}}$ according to Algorithm 1, which will be useful for the proof. Recall that this is possible because (Cont) holds. For this, let us order the Z_i's, that is, $Z_{\tau(1)} \ge \cdots \ge Z_{\tau(n+m)}$ and consider $s_\ell = \mathbb{1}_{\{\tau(\ell) \le n\}} \in \{0, 1\}$, $1 \le \ell \le n+m$, which is 1 if and only if $Z_{\tau(\ell)}$ comes from sample $Y = (Y_1, \dots, Y_n)$. Then, we easily see that $\widehat{\mathrm{BH}}_\alpha$ reject H_i if $X_i \ge Z_{\tau(\hat{\ell})}$ where

$$\hat{\ell} = \max\left\{\ell \in \{1, \dots, n+m\} \,:\, \widehat{\mathrm{FDP}}_\ell \le \alpha\right\}, \widehat{\mathrm{FDP}}_\ell = \frac{m}{n+1} \frac{1 + \sum_{\ell'=1}^{\ell} s_{\ell'}}{1 \vee \sum_{\ell'=1}^{\ell} (1 - s_{\ell'})}, \tag{21}$$

with no rejection if this set is empty.

C.3. Randomization lemma

First, let us provide two lemmas that will be useful for the proof. Consider the sample $W = (Y_1, \dots, Y_n, X_1, \dots, X_{m_0})$ of size $n + m_0$. Consider π the permutation of $\{1, \dots, n+m_0\}$ that orders the W_i's in decreasing order, that is, $W_{\pi(1)} \ge \cdots \ge W_{\pi(n+m_0)}$ and let $s_{0,\ell} = \mathbb{1}_{\{\pi(\ell) \le n\}} \in \{0, 1\}$ for any $\ell \in \{1, \dots, n + m_0\}$

which equals 1 if and only if $W_{\pi(\ell)}$ comes from the sample Y. Under (Cont) and (Exch), since the W_i's are exchangeable and F_0 is continuous, there is almost surely no tie in the sample W and π is uniformly distributed among all permutations of $\{1,\dots,n+m_0\}$, conditionally on $(X_i, m_0+1 \le i \le m)$. Hence, the following lemma holds.

Lemma C.1. *Under (Cont) and (Exch), the set $S_0 = \{\ell \in \{1,\dots,n+m_0\} : s_{0,\ell} = 1\}$ is uniformly distributed among all subset of $\{1,\dots,n+m_0\}$ of cardinality n and this, independently from the order statistics $(W_{\pi(1)},\dots,W_{\pi(n+m_0)})$, and conditionally on $(X_i, m_0+1 \le i \le m)$.*

To study the new procedure, we should now make the link between $s_{0,\ell}$ and s_ℓ. Denote

$$L = \{\ell \in \{1,\dots,n+m\} : \tau(\ell) \le n+m_0\}. \tag{22}$$

The integer L corresponds to the ordered indices of the Z_i's coming from the Y_i's. Then we map $\{1,\dots,n+m_0\}$ to L by using a bijection only depending on the order statistics $(W_{\pi(1)},\dots,W_{\pi(n+m_0)})$ and $(X_i, m_0+1 \le i \le m)$, and thus a bijection independent of S_0. Hence, the above lemma entails the following result.

Lemma C.2. *Under (Exch), the set $S = \{\ell \in L : s_\ell = 1\}$ is uniformly distributed among all subsets of L of cardinality n and this, independently from the order statistics $(W_{\pi(1)},\dots,W_{\pi(n+m_0)})$ and conditionally on $(X_i, m_0+1 \le i \le m)$.*

Also note that $s_\ell = 0$ when $\ell \notin L$, and we introduce the following notation:

$$V_\ell = \sum_{1\le \ell' \le \ell, \ell' \in L} s_{\ell'} = \sum_{1 \le \ell' \le \ell} s_{\ell'}, \text{ for all } \ell \in \{1,\dots,n+m\}. \tag{23}$$

C.4. Core argument for the proof

When $\widehat{\mathrm{BH}}_\alpha$ makes at least one rejection, $\hat\ell \in \{1,\dots,n+m\}$ exists. Let in addition $\hat\ell = 0$ when $\widehat{\mathrm{BH}}_\alpha$ makes no rejection. When $\hat\ell > 0$, we also denote $\hat t = Z_{\tau(\hat\ell)}$. Now, by definition,

$$\begin{aligned}\mathrm{FDR}(P, \widehat{\mathrm{BH}}_\alpha) &= \mathbb{E}[\mathrm{FDP}(P, \widehat{\mathrm{BH}}_\alpha)] = \mathbb{E}[\mathrm{FDP}(P, \widehat{\mathrm{BH}}_\alpha)\mathbb{1}_{\{\hat\ell>0\}}] \\ &= \mathbb{E}\left[\frac{\sum_{i\in\mathcal{H}_0}\mathbb{1}_{\{X_i\ge\hat t\}}}{1\vee\sum_{i=1}^m \mathbb{1}_{\{X_i\ge\hat t\}}}\mathbb{1}_{\{\hat\ell>0\}}\right]. \end{aligned}\tag{24}$$

Now, relying on (21), we have almost surely

$$\begin{aligned}&\frac{\sum_{i\in\mathcal{H}_0}\mathbb{1}_{\{X_i\ge\hat t\}}}{1\vee\sum_{i=1}^m \mathbb{1}_{\{X_i\ge\hat t\}}}\mathbb{1}_{\{\hat\ell>0\}} \\ &= \frac{n+1}{m}\frac{\sum_{i\in\mathcal{H}_0}\mathbb{1}_{\{X_i\ge\hat t\}}}{\sum_{i=1}^n \mathbb{1}_{\{Y_i\ge\hat t\}}+1}\frac{m}{n+1}\frac{\sum_{i=1}^n \mathbb{1}_{\{Y_i\ge\hat t\}}+1}{1\vee\sum_{i=1}^m \mathbb{1}_{\{X_i\ge\hat t\}}}\mathbb{1}_{\{\hat\ell>0\}}\end{aligned}$$

$$= \frac{m_0}{m}\ \widehat{\mathrm{FDP}}_{\hat\ell}\ \times \frac{n+1}{m_0}\,\frac{\sum_{i\in\mathcal{H}_0}\mathbb{1}_{\{X_i\geq\hat t\}}}{\sum_{i=1}^n \mathbb{1}_{\{Y_i\geq\hat t\}}+1}\mathbb{1}_{\{\hat\ell>0\}}, \tag{25}$$

In the next section, we will prove the following equality:

$$\mathbb{E}\left[\frac{n+1}{m_0}\,\frac{\sum_{i\in\mathcal{H}_0}\mathbb{1}_{\{X_i\geq\widehat{t}\}}}{\sum_{i=1}^n \mathbb{1}_{\{Y_i\geq\widehat{t}\}}+1}\mathbb{1}_{\{\hat\ell>0\}}\right] = 1. \tag{26}$$

Let us check that this implies the statements of Theorem 3.1: first, since by definition $\widehat{\mathrm{FDP}}_{\hat\ell}\leq\alpha$ when $\hat\ell>0$, relations (24)-(25)-(26) imply $\mathrm{FDR}(P,\widehat{\mathrm{BH}}_\alpha)\leq \frac{m_0}{m}\alpha$. Second, if $\alpha(n+1)/m$ is an integer, we have $\widehat{\mathrm{FDP}}_{\hat\ell}=\alpha$ when $\hat\ell>0$ by Lemma E.4, hence relations (24)-(25)-(26) implies $\mathrm{FDR}(P,\widehat{\mathrm{BH}}_\alpha)=\frac{m_0}{m}\alpha$. Finally, Lemma E.5 gives $\widehat{\mathrm{FDP}}_{\hat\ell}\geq\frac{m}{n+1}\lfloor\alpha\frac{n+1}{m}\rfloor$ when $\hat\ell>0$, which gives that $\mathrm{FDR}(P,\widehat{\mathrm{BH}}_\alpha)\geq\frac{m_0}{m}\frac{m}{n+1}\lfloor\alpha\frac{n+1}{m}\rfloor$.

C.5. Super-martingale argument

Let $\xi=((W_{\pi(1)},\dots,W_{\pi(n+m_0)}),(X_i,m_0+1\leq i\leq m))$ for short. The proof is based on a super-martingale argument. Recall the equivalent definition (21), so that (26) is proved if

$$\mathbb{E}\left[M_{\hat\ell}\mathbb{1}_{\{\hat\ell\geq1\}}\mid\xi\right]=\frac{m_0}{n+1},\ \ M_\ell=\frac{\sum_{\ell\in L,1\leq\ell'\leq\ell}(1-s_{\ell'})}{\sum_{\ell'\in L,1\leq\ell'\leq\ell}s_{\ell'}+1}=\frac{m_{0,\ell}-V_\ell}{V_\ell+1}, \tag{27}$$

for $1\leq\ell\leq m+n$, where V_ℓ is given by (23) and $m_{0,\ell}$ denotes the cardinal of $\{1\leq\ell'\leq\ell\ :\ \ell'\in L\}$ for $1\leq\ell\leq m+n$. By Lemma C.2, the randomness in the above expectation is only carried by the binary variable $(s_\ell,\ell\in L)$ for which $S=\{\ell\in L\ :\ s_\ell=1\}$ is uniformly distributed among all subsets of L of cardinality n, conditionally on ξ (L is fixed in particular, conditionally on ξ).

Let us define the σ-fields

$$\mathcal{F}_\ell=\sigma\left((V_{\ell'},\ \ell\leq\ell'\leq m+n),\ \xi\right),\ \ 1\leq\ell\leq m+n, \tag{28}$$

where $\sigma(\cdot)$ denotes the σ-field operator. The latter form a filtration $\mathcal{F}_{m+n}\subseteq\mathcal{F}_{m+n-1}\subseteq\cdots\subseteq\mathcal{F}_1$. Note also that

$$\mathcal{F}_\ell=\sigma\left((s_{\ell'})_{\ell'\in L,\ell+1\leq\ell'\leq m+n},V_\ell,\xi\right)=\sigma\left((s_{\ell'})_{\ell+1\leq\ell'\leq m+n},V_\ell,\xi\right).$$

A first key point is that $\{\hat\ell\leq\ell-1\}\in\mathcal{F}_\ell$ for all $\ell\in\{1,\dots,m+n\}$, which means that $\hat\ell$ is a stopping time with respect to the filtration $(\mathcal{F}_\ell)_\ell$. Indeed, by (21), we have

$$\{\hat\ell\leq\ell-1\}=\left\{\forall\ell'\in\{\ell,\dots,m+n\},\frac{m}{n+1}\,\frac{1+V_{\ell'}}{1\vee(\ell'-V_{\ell'})}>\alpha\right\}.$$

Hence, $\{\hat\ell\leq\ell-1\}$ is an event measurable in $V_{\ell'}$, $\ell'\geq\ell$.

A second key point is the following lemma:

Lemma C.3. *Consider the process* $(M_\ell)_{1\le \ell\le m+n}$ *defined by* (27) *and the filtration* (28). *Then* $(M_{m+n}, M_{m-1}, \dots, M_1)$ *is a super-martingale with respect to the filtration* $(\mathcal{F}_{m+n}, \mathcal{F}_{m+n-1}, \dots, \mathcal{F}_1)$ *(note that time is running backwards) that is,* $M_\ell \in \mathcal{F}_\ell$ *for all* $\ell \in \{1, \dots, m+n\}$ *and*

$$\mathbb{E}(M_\ell \mid \mathcal{F}_{\ell+1}) = M_{\ell+1} - \mathbb{1}_{\{V_{\ell+1}=0, \ell+1\in L\}} \le M_{\ell+1}, \quad 1 \le \ell \le m+n-1, \quad (29)$$

where V_ℓ *is defined by* (23) *and* L *is given by* (22).

Applying this lemma, we obtain

$$\begin{aligned}
&\mathbb{E}[M_{\hat\ell}\mathbb{1}_{\{\hat\ell\ge 1\}} \mid \xi] \\
&= \sum_{\ell=1}^{m+n} \mathbb{E}[M_\ell\, \mathbb{1}_{\{\hat\ell=\ell\}} \mid \xi] \\
&= \sum_{\ell=1}^{m+n} \mathbb{E}[M_\ell\, (\mathbb{1}_{\{1\le\hat\ell\le\ell\}} - \mathbb{1}_{\{1\le\hat\ell\le\ell-1\}}) \mid \xi] \\
&= \mathbb{E}[M_{m+n} \mid \xi] + \sum_{\ell=1}^{m+n-1} \mathbb{E}[(M_\ell \mathbb{1}_{\{1\le\hat\ell\le\ell\}} \mid \xi] - \sum_{\ell=1}^{m+n} \mathbb{E}[M_\ell \mathbb{1}_{\{1\le\hat\ell\le\ell-1\}} \mid \xi].
\end{aligned}$$

Hence, we obtain

$$\begin{aligned}
\mathbb{E}[M_{\hat\ell}\mathbb{1}_{\{\hat\ell\ge 1\}} \mid \xi] &= \mathbb{E}[M_{m+n} \mid \xi] + \sum_{\ell=1}^{m+n-1} \mathbb{E}[(M_\ell - M_{\ell+1})\mathbb{1}_{\{1\le\hat\ell\le\ell\}} \mid \xi] \\
&= \mathbb{E}[M_{m+n} \mid \xi] + \sum_{\ell=1}^{m+n-1} \mathbb{E}[\mathbb{1}_{\{1\le\hat\ell\le\ell\}} \mathbb{E}[(M_\ell - M_{\ell+1}) \mid \mathcal{F}_{\ell+1}] \mid \xi] \\
&= \mathbb{E}[M_{m+n} \mid \xi] + \sum_{\ell=1}^{m+n-1} \mathbb{E}[\mathbb{1}_{\{1\le\hat\ell\le\ell\}} (\mathbb{E}[M_\ell \mid \mathcal{F}_{\ell+1}] - M_{\ell+1}) \mid \xi] \\
&= \mathbb{E}[M_{m+n} \mid \xi] - \sum_{\ell=1}^{m+n-1} \mathbb{P}[1 \le \hat\ell \le \ell, V_{\ell+1} = 0, \ell+1 \in L] \\
&= \mathbb{E}[M_{m+n} \mid \xi],
\end{aligned}$$

by using successively (29), the fact that $\{1 \le \hat\ell \le \ell\} \in \mathcal{F}_{\ell+1}$ and Lemma C.4 (below). Now, we conclude because

$$\mathbb{E}[M_{m+n} \mid \xi] = \frac{\sum_{\ell\in L}(1-s_\ell)}{\sum_{\ell\in L} s_\ell + 1} = \frac{m_0}{n+1}.$$

Lemma C.4. *For all* $\ell \in \{2, \dots, m+n\}$, *consider* V_ℓ *defined by* (23), $\hat\ell$ *and* $\widehat{FDP}_\ell$ *defined by* (21). *If* $V_\ell = 0$, *then* $\hat\ell \ge \ell$.

Proof. Recall (21) and let $\ell \in \{2, \dots, m+n\}$. If $V_\ell = 0$, this implies that for any $\ell' \in \{1, \dots, \ell\}$, we have $V_{\ell'} \le V_\ell = 0$ and thus,

$$\widehat{\mathrm{FDP}}_{\ell'} = \frac{m}{n+1}\frac{1}{\ell'}$$

because $\sum_{\ell''=1}^{\ell'}(1-s_{\ell''}) = \ell' - V_{\ell'} = \ell'$. As a result, the function $\ell' \in \{1,\dots,\ell\} \mapsto \widehat{\mathrm{FDP}}_{\ell'}$ is decreasing. This implies $\hat{\ell} \geq \ell$ by definition of $\hat{\ell}$. □

C.6. Proof of Lemma C.3

Recall

$$M_\ell = \frac{\sum_{\ell' \in L, 1 \leq \ell' \leq \ell}(1-s_{\ell'})}{\sum_{\ell' \in L, 1 \leq \ell' \leq \ell} s_{\ell'} + 1} = \frac{m_{0,\ell} - V_\ell}{V_\ell + 1}, 1 \leq \ell \leq m+n,$$

and let us prove (29). Let $1 \leq \ell \leq m+n-1$. For $\ell+1 \notin L$, we have $M_\ell = M_{\ell+1}$ so (29) holds true. Assume thus $\ell+1 \in L$. We have in that case

$$\begin{aligned} M_\ell &= \frac{m_{0,\ell} - V_\ell}{V_\ell + 1} \\ &= \frac{m_{0,\ell+1} + s_{\ell+1} - 1 - V_{\ell+1}}{V_{\ell+1} - s_{\ell+1} + 1}. \end{aligned}$$

because $m_{0,\ell+1} = m_{0,\ell} + 1$. Remember $\mathcal{F}_\ell = \sigma\left((s_{\ell'})_{\ell+1 \leq \ell' \leq m+n, \ell' \in L}, V_\ell, \xi\right)$. We should now study the distribution of $s_{\ell+1}$ conditionally on $(s_{\ell'})_{\ell+2 \leq \ell' \leq m+n}, V_\ell, \xi$. Remember that $S = \{\ell \in L : s_\ell = 1\}$ is uniformly distributed among all subsets of L of cardinality n, conditionally on ξ. Hence, by applying Lemma E.2 below (with $q = n + m_0$, $u = m_{0,\ell}$ and $\{1,\dots,q\}$ in place of L), we obtain

$$\mathbb{P}\left(s_{\ell+1} = 1 \mid (s_{\ell'})_{\ell+2 \leq \ell' \leq m+n, \ell' \in L}, V_{\ell+1}, \xi\right) = V_{\ell+1}/m_{0,\ell+1}.$$

This gives for $V_{\ell+1} \geq 1$,

$$\begin{aligned} \mathbb{E}(M_\ell \mid \mathcal{F}_{\ell+1}) &= \frac{V_{\ell+1}}{m_{0,\ell+1}} \frac{m_{0,\ell+1} - V_{\ell+1}}{V_{\ell+1}} + \frac{m_{0,\ell+1} - V_{\ell+1}}{m_{0,\ell+1}} \frac{m_{0,\ell+1} - 1 - V_{\ell+1}}{V_{\ell+1} + 1} \\ &= \frac{m_{0,\ell+1} - V_{\ell+1}}{m_{0,\ell+1}} \left(1 + \frac{m_{0,\ell+1} - 1 - V_{\ell+1}}{V_{\ell+1} + 1}\right) \\ &= \frac{m_{0,\ell+1} - V_{\ell+1}}{V_{\ell+1} + 1} = M_{\ell+1}. \end{aligned}$$

The last thing to check is that if $V_{\ell+1} = 0$ then $M_\ell = m_{0,\ell+1} - 1$ and $M_{\ell+1} = m_{0,\ell+1}$, so that $\mathbb{E}(M_\ell \mid \mathcal{F}_{\ell+1}) = M_{\ell+1} - 1$. This gives (29) for any possible value of $V_{\ell+1}$.

Appendix D: Proofs for power results

D.1. Proof of Proposition 4.2

Recall that for the oracle p-values $p_i = F_0(X_i)$, $1 \leq i \leq m$, sorted as $p_{(0)} = 0 \leq p_{(1)} \leq \cdots \leq p_{(m)}$, the oracle BH procedure at level α is defined by

$$\mathrm{BH}^*_\alpha = \{i \in \{1,\dots,m\} : p_i \leq p_{(k^*)}\},$$

for $k^* = \max\{k \in \{0,1,\dots,m\} \,:\, p_{(k)} \le \alpha k/m\}$, or, equivalently,

$$\mathrm{BH}^*_\alpha = \{i \in \{1,\dots,m\} \,:\, X_i \ge X_{(k^*)}\},$$

for $k^* = \max\{k \in \{0,1,\dots,m\} \,:\, X_{(k)} \ge F_0^{-1}(\alpha k/m)\}$. The $\hat{p}$-values are $\widehat{p}_i = \widehat{F}_0(X_i)$, $1 \le i \le m$, where $\widehat{F}_0$ is defined by (7), and the semi-supervised BH procedure at level $\alpha(1+\eta)$ is given by

$$\widehat{\mathrm{BH}}_{\alpha(1+\eta)} = \{i \in \{1,\dots,m\} \,:\, \widehat{F}_0(X_i) \le \widehat{F}_0(X_{(\widehat{k})})\},$$

with $\widehat{k} = \max\{k \in \{0,1,\dots,m\} \,:\, \widehat{F}_0(X_{(k)}) \le \alpha(1+\eta)k/m\}$. Hence, since $\widehat{F}_0$ is non-increasing, we have that $X_i \ge X_{(k^*)}$ implies $\hat{F}_0(X_i) \le \hat{F}_0(X_{(k^*)})$, hence

$$\mathrm{BH}^*_\alpha \subseteq \{i \in \{1,\dots,m\} \,:\, \widehat{F}_0(X_i) \le \widehat{F}_0(X_{(k^*)})\}$$

which is itself contained in $\widehat{\mathrm{BH}}_{\alpha(1+\eta)}$ provided that $k^* \le \widehat{k}$. By definition of $\widehat{k}$, the latter holds true whenever

$$\widehat{F}_0(X_{(k^*)}) \le \alpha(1+\eta)k^*/m. \tag{30}$$

Since $X_{(k^*)} \ge F_0^{-1}(\alpha k^*/m)$, we have $\widehat{F}_0(X_{(k^*)}) \le \widehat{F}_0(F_0^{-1}(\alpha k^*/m))$ and (30) holds if $\widehat{F}_0(F_0^{-1}(\alpha k^*/m)) \le \alpha(1+\eta)k^*/m$. To sum up, we obtained that

$$\mathbb{P}(\mathrm{BH}^*_\alpha \subseteq \widehat{\mathrm{BH}}_{\alpha(1+\eta)}) \ge \mathbb{P}(\Omega)$$

with $\Omega = \{\widehat{F}_0(F_0^{-1}(\alpha k^*/m)) \le \alpha(1+\eta)k^*/m\}$.

Now, let us upper bound the probability of Ω^c as follows. Letting $u_k = F_0^{-1}(\alpha k/m)$, and $\widetilde{F}_0(x) = n^{-1}\sum_{j=1}^n \mathbb{1}_{\{Y_j \ge x\}} = \frac{n+1}{n}\widehat{F}_0(x) - 1/n$, see (7), (so that $\mathbb{E}\widetilde{F}_0(x) = F_0(x)$), we have $\mathbb{P}(\Omega^c) = \mathbb{P}(\Omega^c, k^* > 0)$ with

$$\begin{aligned}
\mathbb{P}(\Omega^c, k^* > 0) &\le \sum_{k=1}^m \mathbb{P}(\widehat{F}_0(u_k) > \alpha(1+\eta)k/m) \\
&= \sum_{k=1}^m \mathbb{P}\left(\widetilde{F}_0(u_k) > \frac{(n+1)\alpha(1+\eta)k/m - 1}{n}\right) \\
&\le \sum_{k=1}^m \mathbb{P}\left(\widetilde{F}_0(u_k) > \alpha(1+\eta)k/m - 1/n\right) \\
&= \sum_{k=1}^m \mathbb{P}(n(\widetilde{F}_0 - F_0)(u_k) > \alpha\eta kn/m - 1) \\
&\le \sum_{k=1}^m \mathbb{P}(n(\widetilde{F}_0 - F_0)(u_k) > c\alpha\eta kn/m),
\end{aligned}$$

for some constant $c \in (0,1)$ such that $\eta\alpha n/m \ge 1/(1-c)$ (to be chosen later) and by noting that $F_0(u_k) = \alpha k/m$ and $a = \eta\alpha kn/m$ satisfies $a - 1 \ge ca$ by

the imposed condition on c. Applying Bernstein's inequality (Lemma E.1 with $W_i = \mathbb{1}_{\{Y_i \geq u_k\}}$, $\mathcal{M} = 1$, $V = \alpha k n/m$, $A = c\alpha\eta k n/m$), we obtain

$$\begin{aligned}\mathbb{P}(\Omega^c) &\leq \sum_{k=1}^{m} \exp\left\{-\frac{(c\alpha\eta kn/m)^2}{2\alpha kn/m + 2(c\alpha\eta kn/m)/3)}\right\} \\ &\leq \sum_{k=1}^{m} \exp\left\{-\frac{c^2\eta^2\alpha kn/m}{2+2c\eta/3}\right\} \leq \frac{e^{-z}}{1-e^{-z}} \leq 2e^{-z},\end{aligned}$$

for $z = \frac{c^2\eta^2\alpha n/m}{2+2c\eta/3}$ that is assumed to be such that $z \geq \log 2$. Now, we only have to choose c such that the two following conditions are satisfied:

$$\eta^2\alpha n/m \geq \frac{\eta}{1-c} \vee \frac{(\log 2)(2+2c\eta/3)}{c^2}.$$

To make the tradeoff, we choose $c = 1-\eta/(2\log 2)$. Since $2(1-\eta/(2\log 2))\eta/3 \leq (\log 2)/3$, this gives rise to the sufficient condition

$$\eta^2\alpha\gamma \geq (2\log 2) \vee \frac{\kappa\log(2)}{(1-\eta/(2\log 2))^2} = \frac{\kappa\log(2)}{(1-\eta/(2\log 2))^2} \tag{31}$$

where $\kappa = 2 + (\log 2)/3 \geq 2$. Noting that $z \geq D(\eta)\eta^2\alpha\gamma$ for $D(\eta) = (1-\eta/(2\log 2))^2/\kappa$ (z is above $\log 2$ by construction), we have proved that for any $\alpha, \eta \in (0,1)$ satisfying (31),

$$\mathbb{P}_{Z\sim P}(\mathrm{BH}^*_\alpha \subseteq \widehat{\mathrm{BH}}_{\alpha(1+\eta)}) \geq 1 - 2\exp(-D(\eta)\alpha\eta^2\gamma).$$

Applying this for $\alpha' = \alpha/(1+\eta) \in (0,1)$ in place of α, we obtain that for all $\alpha, \eta \in (0,1)$ with $\eta^2\alpha\gamma \geq (1+\eta)\frac{\kappa\log 2}{(1-\eta/(2\log 2))^2} = \frac{(1+\eta)\log 2}{D(\eta)}$,

$$\mathbb{P}_{Z\sim P}(\mathrm{BH}^*_{\alpha'} \subseteq \widehat{\mathrm{BH}}_\alpha) \geq 1 - 2\exp(-D(\eta)\alpha\gamma\eta^2/(1+\eta)).$$

Since $\mathrm{BH}^*_{\alpha(1-\eta)} \subseteq \mathrm{BH}^*_{\alpha'}$ because $1-\eta \leq 1/(1+\eta)$, we obtain

$$\mathbb{P}_{Z\sim P}(\mathrm{BH}^*_{\alpha(1-\eta)} \subseteq \widehat{\mathrm{BH}}_\alpha) \geq 1 - 2\exp(-D(\eta)\alpha\gamma\eta^2/(1+\eta)).$$

Now, we note that $D(\eta)\alpha\gamma\eta^2/(1+\eta)) = (3\log 2)\gamma/\gamma^*(\alpha,\eta)$ so that the latter bound is equal to $1-(1/2)^{3\gamma/\gamma^*(\alpha,\eta)-1}$. In addition, the condition $\eta^2\alpha\gamma \geq \frac{(1+\eta)\log 2}{D(\eta)}$ is equivalent to $\gamma \geq \gamma^*(\alpha,\eta)/3$. Also, the bound trivially holds if $\gamma \leq \gamma^*(\alpha,\eta)/3$. This shows the result.

D.2. Proof of Proposition 4.3

We start by proving the following result.

Lemma D.1. *Assume* (Indep) *and let* $\alpha \in (0,1)$ *and* $n, m \geq 1$. *Then for any* $P \in \mathcal{A}_{n,m}$, *we have that* $2\alpha m_1(P) \leq m/(n+1)$ *implies*

$$\mathbb{P}_{Z\sim P}(TDP(P, \widehat{BH}_\alpha) = 0) \geq 1 - 2\alpha. \tag{32}$$

In particular, if $(n+1)/m \leq 1/(2\alpha)$, *inequality* (32) *holds for all* $P \in \mathcal{A}_{n,m}$ *with* $m_1(P) = 1$.

Before proving Lemma D.1, let us show that it implies Proposition 4.3. For this, let us consider $\alpha \in (0, 1/4)$, $\eta \in [0,1)$ and $n, m \geq 1$ with $n/m \leq 1/(4\alpha)$. Then we have $(n+1)/m \leq 1/(4\alpha) + 1/m \leq 1/(2\alpha)$. Applying (32) for $P_a = \mathcal{N}(0,1)^{\otimes(n+m-1)} \otimes \mathcal{N}(a,1)$, $a > 0$ (note that $m_1(P_a) = 1$), we have for all $a > 0$,

$$\mathbb{P}_{Z\sim P_a}(\mathrm{TDP}(P, \widehat{\mathrm{BH}}_\alpha) = 0) \geq 1 - 2\alpha.$$

Now, we have that $\mathrm{BH}^*_{\alpha(1-\eta)}$ rejects the only null hypotheses that are false for P_a provided that $\bar{\Phi}(X_m) \leq \alpha(1-\eta)/m$, where $\bar{\Phi}$ denotes the standard Gaussian upper tail function. This occurs with probability $\bar{\Phi}(\bar{\Phi}^{-1}(\alpha(1-\eta)/m) - a)$. Therefore, for all $a > 0$,

$$\begin{aligned}&\mathbb{P}_{Z\sim P_a}(\mathrm{TDP}(P, \widehat{\mathrm{BH}}_\alpha) = 0, \mathrm{TDP}(P, \mathrm{BH}^*_{\alpha(1-\eta)}) > 0)\\ &\geq \bar{\Phi}(\bar{\Phi}^{-1}(\alpha(1-\eta)/m) - a) - 2\alpha.\end{aligned}$$

This entails for all $a > 0$,

$$\begin{aligned}&\sup_{P\in\mathcal{A}_{n,m}} \{\mathbb{P}_{Z\sim P}(\mathrm{TDP}(P, \widehat{\mathrm{BH}}_\alpha) < \mathrm{TDP}(P, \mathrm{BH}^*_{\alpha(1-\eta)}))\}\\ &\geq \bar{\Phi}(\bar{\Phi}^{-1}(\alpha(1-\eta)/m) - a) - 2\alpha.\end{aligned}$$

Now making a tending to infinity gives (14).

Let us now prove Lemma D.1. Consider $P \in \mathcal{A}_{n,m}$. Assume $2\alpha m_1(P) \leq m/(n+1)$. Denote by $\hat{k} \geq 0$ the number of rejections of $\widehat{\mathrm{BH}}_\alpha$. First observe that $\hat{k} \geq 2m_1(P)$ implies $\mathrm{FDP}(P, \widehat{\mathrm{BH}}_\alpha) \geq (\hat{k} - m_1(P))/\hat{k} \geq 1/2$. Applying the Markov inequality, we thus derive

$$\mathbb{P}_{Z\sim P}(\hat{k} \geq 2m_1(P)) \leq \mathbb{P}_{Z\sim P}(\mathrm{FDP}(P, \widehat{\mathrm{BH}}_\alpha) \geq 1/2) \leq 2\,\mathrm{FDR}(P, \widehat{\mathrm{BH}}_\alpha) \leq 2\alpha,$$

because $\mathrm{FDR}(P, \widehat{\mathrm{BH}}_\alpha) \leq \alpha$ by Theorem 3.1. On the other hand, if $\hat{k} < 2m_1(P)$ then because $2\alpha m_1(P) \leq m/(n+1)$, we have that all $\hat{p}$-values are larger than or equal to (see (6))

$$1/(n+1) \geq 2\alpha m_1(P)/m > \alpha\hat{k}/m.$$

Hence $\hat{k} = 0$ by definition of the BH procedure (8). This entails $\mathrm{TDP}(P, \widehat{\mathrm{BH}}_\alpha) = 0$. Putting the above relations together, we obtain

$$\mathbb{P}_{Z\sim P}(\mathrm{TDP}(P, \widehat{\mathrm{BH}}_\alpha) > 0) \leq \mathbb{P}_{Z\sim P}(\hat{k} \geq 2m_1(P)) \leq 2\alpha,$$

which concludes the proof.

D.3. Proof of Proposition 4.4

Point (i) is similar to the proof of Proposition 4.2, see Appendix D.1. The only difference is that we can use that the number of correct rejections of the oracle procedure at a level $\alpha' \in (\alpha/2, 1)$ is larger or equal to k, with large probability, because $P \in \mathcal{A}_{n,m,k,\alpha,\beta}$.

Consider $nk/m \geq \gamma$ for some $\gamma > 0$. We first prove that for $\alpha' \in (\alpha/2, 1)$, $\eta \in (0, 1/2)$, if $\eta^2 \alpha' \gamma \geq \frac{\kappa \log(2)}{(1-\eta/(2\log 2))^2}$,

$$\mathbb{P}_{Z\sim P}(\mathrm{BH}^*_{\alpha'} \subseteq \widehat{\mathrm{BH}}_{\alpha'(1+\eta)}) \geq 1 - \beta - 2\exp(-D(\eta)\alpha'\eta^2\gamma).$$

For this, we use exactly the same proof as in Appendix D.1, except that we use $k^* \geq k$ when $k^* > 0$ on an event of probability larger than $1 - \beta$ (see notation therein). Hence, we obtain

$$\mathbb{P}(\Omega^c, k^* > 0) \leq \beta + \mathbb{P}(\Omega^c, k^* \geq k) \leq \sum_{k'=k}^{m} e^{-k'z} \leq 2e^{-kz},$$

where $z = \frac{c^2\eta^2\alpha' n/m}{2+2c\eta/3}$, which proves the intermediate result above. Now, Point (i) comes by applying this with $\alpha' = \alpha/(1+\eta) \geq \alpha(1-\eta) \geq \alpha/2$ because $\eta < 1/2$.

Point (ii) is similar to the proof of Proposition 4.3, see Appendix D.2. Consider any distribution $P \in \mathcal{A}_{n,m,k,\alpha,\beta}$ with $m_1(P) = k$ and

$$\mathbb{P}_{Z\sim P}(\mathrm{TDP}(P, \mathrm{BH}^*_{\alpha(1-\eta)}) = 1) \geq 1 - \beta$$

(we easily check that such a distribution exists for Gaussian alternatives with a common alternative mean large enough). Applying Lemma D.1 with this distribution P, we obtain that if $2\alpha k \leq m/(n+1)$, that is, $(n+1)k/m \leq 1/(2\alpha)$,

$$\mathbb{P}_{Z\sim P}(\mathrm{TDP}(P, \widehat{\mathrm{BH}}_{\alpha}) = 0) \geq 1 - 2\alpha.$$

This gives Point (ii) by noting that $(n+1)k/m \leq 1/(4\alpha) + k/m \leq 1/(2\alpha)$ because $k/m \leq 1 \leq 1/(4\alpha)$.

D.4. Proof of Theorem 5.1

The proof relies on the construction of particular distributions for the Y_i's and the X_i's. Also remember that the BH procedure rejects for large values of X_i's. For instance, if $X_i \sim U(0,1)$ under the null, any $X_i \in [1 - \alpha/m, 1]$ will be in the rejection set of the oracle BH procedure at level α.

Let $\alpha' = \alpha(1-\eta)$. For some constant $\kappa > 0$ with $\kappa/n < 1$ (to be chosen later on), let us consider the two following distribution on $\mathbb{R}$

$$\mu = (1 - \kappa/n)\, U(0,1) + (\kappa/n)\, U(1 - \alpha'/m, 1) \tag{33}$$

and the following distributions on $\mathbb{R}^{n+m}$

$$Q_1 = \mu^{\otimes(n+m)}, \;\; Q_{2,u} = U(0,1)^{\otimes n} \otimes \bigotimes_{i=1}^{m} \left((1-u_i)U(0,1) + u_i U(1-\alpha'/m, 1)\right), \tag{34}$$

for $u \in \mathbb{R}^m$. Observe that $\mathcal{H}_0(Q_1) = \{1,\dots,m\}$, $m_0(Q_1) = m$, $\mathcal{H}_0(Q_{2,u}) = \{i \in \{1,\dots,m\} \,:\, u_i = 0\}$, $m_0(Q_{2,u}) = \sum_{i=1}^m (1-u_i)$, for all $u \in \mathbb{R}^m$. Now consider U_i, $1 \le i \le m$, that are i.i.d. $\mathcal{B}(\kappa/n)$ and $U = (U_i)_{1 \le i \le n}$. Then any $Z \sim Q_{2,U}$ is distributed as $Q_2 = U(0,1)^{\otimes n} \otimes \mu^{\otimes m}$ unconditionally on U.

For any procedure $R = R(Z)$, we have $\mathrm{FDR}(Q_1, R) = \mathbb{P}_{Z\sim Q_1}(|R| > 0)$ (remember $\mathcal{H}_0(Q_1) = \{1,\dots,m\}$). Since $\mathbb{P}_{Z\sim Q_1}(|R| > 0) + \mathbb{P}_{Z\sim Q_1}(|R| = 0) = 1$. Either $\mathrm{FDR}(Q_1, R) = \mathbb{P}_{Z\sim Q_1}(|R| > 0) \ge 1/2$, or $\mathbb{P}_{Z\sim Q_1}(|R| = 0) \ge 1/2$, in which case $\mathbb{P}_{Z\sim Q_2}(|R| = 0) \ge 1/2 - d_{tv}(Q_1, Q_2)$, where $d_{tv}(Q_1, Q_2) = \sup_{\mathcal{A}} |Q_1(A) - Q_2(A)|$ denotes the total variation distance between the distributions Q_1 and Q_2. Hence, in the latter case, we obtain (recall the definition of the U_i's above)

$$\begin{aligned}
&1/2 - d_{tv}(Q_1, Q_2) \\
&\le \mathbb{E}_U \mathbb{P}_{Z\sim Q_{2,U}}(|R(Z)| = 0) \\
&\le \mathbb{E}_U \left(\mathbb{1}_{\{\sum_{i=1}^m U_i \ge 1\}} \mathbb{P}_{Z\sim Q_{2,U}}(|R(Z)| = 0) \right) + \mathbb{P}_U \left(\sum_{i=1}^m U_i = 0 \right) \\
&\le \mathbb{E}_U \left(\mathbb{1}_{\{\sum_{i=1}^m U_i \ge 1\}} \mathbb{P}_{Z\sim Q_{2,U}}(|R(Z)| = 0) \right) + (1-\kappa/n)^m \\
&\le \mathbb{E}_U \left(\mathbb{1}_{\{\sum_{i=1}^m U_i \ge 1\}} \mathbb{P}_{Z\sim Q_{2,U}}(|R(Z)| = 0, |\mathrm{BH}^*_{\alpha'} \cap \mathcal{H}_1(Q_{2,U})| \ge 1) \right) + e^{-\kappa m/n},
\end{aligned}$$

because by definition of $Q_{2,U}$ the null hypothesis corresponding to any index i with $U_i = 1$ corresponds to a X_i larger than $1 - \alpha'/m$ and thus is rejected by $\mathrm{BH}^*_{\alpha'}$. Note that we also used $(1-\kappa/n)^m \le e^{-\kappa m/n}$ because for all $u \in [0,1)$, $\log(1-u) \le -u$. The last display entails that

$$\begin{aligned}
&\sup_{u \in \mathbb{R}^m} \left\{ \mathbb{P}_{Z\sim Q_{2,u}}(|R(Z)| = 0, |\mathrm{BH}^*_{\alpha'} \cap \mathcal{H}_1(Q_{2,u})| \ge 1) \right\} \\
&\ge 1/2 - e^{-\kappa m/n} - d_{tv}(Q_1, Q_2).
\end{aligned}$$

Summing up, we obtained that for any procedure R, either $\mathrm{FDR}(Q_1, R) \ge 1/2$, or there exists a distribution $Q_{2,u}$, $u \in \mathbb{R}^m$, with $m_1(Q_{2,u}) \ge 1$ and

$$\mathbb{P}_{Z\sim Q_{2,u}} \left(\mathrm{FDP}(R, Q_{2,u}) < \mathrm{FDP}(\mathrm{BH}^*_{\alpha'/2}, Q_{2,u}) \right) \ge 1/2 - e^{-\kappa m/n} - d_{tv}(Q_1, Q_2).$$

It only remains to upper bound the total variation distance $d_{tv}(Q_1, Q_2)$. From Le Cam's inequalities and tensorization identities for Hellinger distances, see, e.g., Tsybakov (2009) Section 2.4, we have that

$$d_{tv}(Q_1, Q_2)^2$$

$$\leq \int_{\mathbb{R}^n}\int_{\mathbb{R}^m}\left(\prod_{i=1}^n f_\mu^{1/2}(y_i)\prod_{i=1}^m f_\mu^{1/2}(x_i)-\prod_{i=1}^n g^{1/2}(y_i)\prod_{i=1}^m f_\mu^{1/2}(x_i)\right)^2 dxdy$$
$$=\int_{\mathbb{R}^n}\left(\prod_{i=1}^n f_\mu^{1/2}(y_i)-\prod_{i=1}^n g^{1/2}(y_i)\right)^2 dy \leq n\int_{\mathbb{R}}\left(f_\mu^{1/2}(y)-g^{1/2}(y)\right)^2 dy,$$

where $f_\mu(y)=(1-\kappa/n)\mathbb{1}_{\{y\in[0,1]\}}+\frac{\kappa m}{n\alpha'}\mathbb{1}_{\{y\in[1-\alpha'/m,1]\}}$, $y\in\mathbb{R}$, denotes the density of μ, while $g(y)=\mathbb{1}_{\{y\in[0,1]\}}$, $y\in\mathbb{R}$, denotes the density of $U(0,1)$. Now, we have

$$\begin{aligned}&d_{tv}(Q_1,Q_2)^2\\&\leq n\left[\int_{1-\frac{\alpha'}{m}}^1\left(\left(\frac{\kappa m}{n\alpha'}+1-\kappa/n\right)^{1/2}-1\right)^2 dy+\int_0^{1-\frac{\alpha'}{m}}\left(1-(1-\kappa/n)^{1/2}\right)^2 dy\right]\\&=\frac{n\alpha'}{m}\left(\left(\frac{\kappa m}{n\alpha'}+1-\kappa/n\right)^{1/2}-1\right)^2+n\left(1-(1-\kappa/n)^{1/2}\right)^2.\end{aligned}$$

Now note that $1\leq\frac{\kappa m}{n\alpha'}+1-\kappa/n\leq\frac{\kappa m}{n\alpha'}+1$, which entails

$$\left(\left(\frac{\kappa m}{n\alpha'}+1-\kappa/n\right)^{1/2}-1\right)^2\leq\left(\left(1+\frac{\kappa m}{n\alpha'}\right)^{1/2}-1\right)^2\leq\left(\frac{\kappa m}{2n\alpha'}\right)^2$$

where we used that for all $u\geq 0$, $(1+u)^{1/2}-1\leq u/2$. Furthermore, for all $u\in[0,1]$, $1-(1-u)^{1/2}\leq u$ and thus $\left(1-(1-\kappa/n)^{1/2}\right)^2\leq(\kappa/n)^2$. Hence, we obtain (since $m\geq 4\alpha'$),

$$d_{tv}(Q_1,Q_2)^2\leq\frac{n\alpha'}{m}\left(\frac{\kappa m}{2n\alpha'}\right)^2+n(\kappa/n)^2=(\kappa^2/n)\left(\frac{m}{4\alpha'}+1\right)\leq\kappa^2\frac{m}{2n\alpha'}.$$

Now, to make $e^{-\kappa m/n}+\kappa\sqrt{\frac{m}{2n\alpha'}}$ small, we can choose $\kappa=(n/m)\log(1+m/n)$, to get the bound $\frac{n}{m}+\sqrt{\frac{n\log(1+m/n)}{2m\alpha'}}\leq\gamma+\sqrt{\frac{\gamma\log(1+\gamma^{-1})}{2\alpha(1-\eta)}}$, because $\log(1+m/n)\geq\log(2)\geq 1\ n/m\geq\gamma$ and $h(u)=u\log(1+1/u)$ is increasing on $\mathbb{R}_+$ (for instance, we have $h''(u)=-1/(u(u+1)^2)$ and $h'(10)>0$). We also check that $\kappa/n<1$, which holds because $\gamma\log(1+\gamma^{-1})\leq\log(2)<1\leq n$.

Finally, we obtained that for any procedure R, either (18) holds or

$$\sup_{P\in\mathcal{A}_{n,m}}\{\mathbb{P}_{Z\sim P}(\mathrm{TDP}(P,R)<\mathrm{TDP}(P,\mathrm{BH}^*_{\alpha(1-\eta)}))\}\geq 1/2-\gamma-\sqrt{\frac{\gamma\log(1+\gamma^{-1})}{2\alpha(1-\eta)}}\tag{35}$$

holds. Since for all $x>1$, we have $x^{-1/3}\log(1+x)\leq 2$, this entails

$$\gamma+\sqrt{\frac{\gamma\log(1+\gamma^{-1})}{2\alpha(1-\eta)}}\leq\gamma^{1/3}(1+(\alpha(1-\eta))^{-1/2})=(\gamma/(64\gamma_*(\alpha,\eta)))^{1/3},$$

by definition of $\gamma_*(\alpha,\eta)$. This gives the main statement. Let us now prove the additional statement. Choose any sequence $\gamma_k\in\mathbb{Q}$ with $0<\gamma_k\leq\gamma$ and $\gamma_k\to\gamma$

when $k \to \infty$. Since γ_k is of the form n/m with $n, m \geq 1$ being two integers, the previous statement applied with $n = m\gamma_k$ shows that

$$\max\left(\sup_{\substack{n,m\geq 1\\ n\geq m\gamma_k}} \sup_{P\in\mathcal{P}_{n,m}} \{\mathrm{FDR}(P,R) - \mathrm{FDR}(\mathrm{BH}^*_\alpha, R)\} - (1/2-\alpha),\right.$$

$$\left.\sup_{\substack{n,m\geq 1\\ n\geq m\gamma_k}} \sup_{P\in\mathcal{A}_{n,m}} \{\mathbb{P}(\mathrm{TDP}(P,\mathrm{BH}^*_{\alpha(1-\eta)}) > \mathrm{TDP}(P,R)\} - 1/2 - \left(\frac{\gamma_k}{8\gamma_*(\alpha,\eta)}\right)^{1/3}\right)$$

is nonnegative. Hence, this also holds if we take the supremum over the indices $n, m \geq 1, n \geq m\gamma$. Making k tending to infinity, we get that

$$\max\left(\sup_{\substack{n,m\geq 1\\ n\geq m\gamma}} \sup_{P\in\mathcal{P}_{n,m}} \{\mathrm{FDR}(P,R) - \mathrm{FDR}(\mathrm{BH}^*_\alpha, R)\} - (1/2-\alpha),\right.$$

$$\left.\sup_{\substack{n,m\geq 1\\ n\geq m\gamma}} \sup_{P\in\mathcal{A}_{n,m}} \{\mathbb{P}(\mathrm{TDP}(P,\mathrm{BH}^*_{\alpha(1-\eta)}) > \mathrm{TDP}(P,R)\} - 1/2 - \left(\frac{\gamma}{8\gamma_*(\alpha,\eta)}\right)^{1/3}\right)$$

is nonnegative. This excludes that (15) and (16) simultaneously holds for $\delta_1 < 1/2 - \alpha$ and $\delta_2 < 1/2 - (\gamma/(8\gamma_*(\alpha,\eta)))^{1/3}$.

Appendix E: Auxiliary results

Lemma E.1. *[Bernstein's inequality] Let W_i, $1 \leq i \leq n$ centered independent variables with $|W_i| \leq \mathcal{M}$ and $\sum_{i=1}^n \mathrm{Var}(W_i) \leq V$, then for any $A > 0$,*

$$P\left[\sum_{i=1}^n W_i > A\right] \leq \exp\left\{-\frac{1}{2}A^2/(V + \mathcal{M}A/3)\right\}.$$

Lemma E.2. *Let $\varepsilon_1, \ldots, \varepsilon_q \in \{0,1\}$ be exchangeable binary random variables, $1 \leq u \leq q$, and $V = \sum_{i=1}^u \varepsilon_i$, then*

$$\mathbb{P}(\varepsilon_u = 1 \mid V, \varepsilon_q, \ldots, \varepsilon_{u+1}) = V/u.$$

In particular, this holds if the set $\{1 \leq i \leq q : \varepsilon_i = 1\}$ is uniformly distributed among the subsets of $\{1, \ldots, q\}$ of size n, for some $1 \leq n \leq q$.

Note that the following stronger result holds: conditionally on the variables V and $\varepsilon_q, \ldots, \varepsilon_{u+1}$, the set $\{1 \leq i \leq u : \varepsilon_i = 1\}$ is uniformly distributed among the subsets of $\{1, \ldots, u\}$ of size $n - V$.

Proof. Let us first observe that $(\varepsilon_u, \ldots, \varepsilon_1)$ are exchangeable conditionally on $\sum_{i=1}^u \varepsilon_i, \varepsilon_q, \ldots, \varepsilon_{u+1}$. Indeed, for any permutation g of $\{1, \ldots, u\}$, we have that

$$(\varepsilon_{g(u)}, \ldots, \varepsilon_{g(1)}, \varepsilon_q, \ldots, \varepsilon_{u+1}) \sim (\varepsilon_u, \ldots, \varepsilon_1, \varepsilon_q, \ldots, \varepsilon_{u+1})$$

and since $\sum_{i=1}^{u} \varepsilon_i = \sum_{i=1}^{u} \varepsilon_{g(i)}$, we have thus

$$(\varepsilon_{g(u)}, \dots, \varepsilon_{g(1)}, \sum_{i=1}^{u} \varepsilon_i, \varepsilon_q, \dots, \varepsilon_{u+1}) \sim (\varepsilon_u, \dots, \varepsilon_1, \sum_{i=1}^{u} \varepsilon_i, \varepsilon_q, \dots, \varepsilon_{u+1}),$$

which entails the first observation.

Hence, we have

$$V = \sum_{i=1}^{u} \mathbb{P}(\varepsilon_i = 1 \mid V, \varepsilon_q, \dots, \varepsilon_{u+1}) = u\, \mathbb{P}(\varepsilon_u = 1 \mid V, \varepsilon_q, \dots, \varepsilon_{u+1}),$$

which gives the result. □

Proposition E.3. *Let $k \geq 1$ and consider $(X_1, \dots, X_k)$ a k-dimensional centered Gaussian vector with individual variance 1 and equi-correlation equal to $\rho \in [-1/k, 1]$. Let*

$$X_{k+1} = a(X_1 + \cdots + X_k) + bU, \;\; a = \frac{\rho}{1 + (k-1)\rho}, \;\; b = (1 - ak\rho)^{1/2},$$

where $U \sim \mathcal{N}(0,1)$ is independent of $X_1, \dots, X_k$. Then the random vector $(X_1, \dots, X_k, X_{k+1})$ is a $(k+1)$-dimensional centered Gaussian vector with individual variance 1 and equi-correlation ρ.

As an illustration, in the extremal case $\rho = -1/k$, we have $a = -1$, $b = 0$ and $X_{k+1} = -(X_1 + \cdots + X_k)$. The opposite extremal case is $\rho = 1$, for which $a = 1/k$, $b = 0$. More generally, when $\rho \in [-1/k, 1]$, a is increasing in ρ from -1 ($\rho = -1/k$) to $1/k$ ($\rho = 1$) and we can check that b is well defined because $|ak\rho| \leq 1$: when $\rho \leq 0$, $|ak\rho| \leq |a| \leq 1$ and when $\rho \geq 0$, $|ak\rho| \leq |ak| \leq 1$.

Proof. Since the vector $(X_1, \dots, X_k, U)$ is Gaussian, so is $(X_1, \dots, X_k, X_{k+1})$ and we just have to check that $\mathrm{Var}(X_{k+1}) = 1$ and for all $i \in \{1, \dots, k\}$, $\mathrm{Cov}(X_i, X_{k+1}) = \rho$. This comes from

$$\begin{aligned}
\mathrm{Var}(X_{k+1}) &= a^2 \,\mathrm{Var}(X_1 + \cdots + X_k) + b^2 \\
&= a^2(k + k(k-1)\rho) + b^2 \\
&= a\frac{\rho}{1 + (k-1)\rho} k(1 + (k-1)\rho) + 1 - ak\rho = 1,
\end{aligned}$$

and, for all $i \in \{1, \dots, k\}$,

$$\mathrm{Cov}(X_i, X_{k+1}) = a\left(1 + \sum_{1 \leq j \leq k, j \neq i} \mathrm{Cov}(X_i, X_j)\right) = a(1 + (k-1)\rho) = \rho. \quad \square$$

Proposition E.4. *Let $\alpha \in (0,1)$ and $n, m \geq 1$ such that $\alpha(n+1)/m$ is an integer. Then if $\hat{\ell}$ given by (21) exists, we have $\widehat{FDP}_{\hat{\ell}} = \alpha$*

Proof. Recall

$$\widehat{\mathrm{FDP}}_\ell = \frac{m}{n+1}\frac{1+\sum_{\ell'=1}^{\ell} s_{\ell'}}{1 \vee \sum_{\ell'=1}^{\ell}(1-s_{\ell'})}, \quad 1 \le \ell \le n+m.$$

Let us introduce the sets $\mathcal{L} = \left\{\ell \in \{1,\dots,n+m\} : \widehat{\mathrm{FDP}}_\ell \le \alpha\right\}$ and $\mathcal{L}' = \left\{\ell \in \{1,\dots,n+m\} : \widehat{\mathrm{FDP}}_\ell < \alpha\right\}$. By assumption, $\mathcal{L} \neq \emptyset$ and $\hat{\ell} = \max \mathcal{L}$. If $\mathcal{L}' = \emptyset$, then necessarily $\hat{\ell} \notin \mathcal{L}'$ which means $\widehat{\mathrm{FDP}}_{\hat{\ell}} = \alpha$. So we assume in the sequel $\mathcal{L}' \neq \emptyset$ and consider $\tilde{\ell} = \max \mathcal{L}' \le \hat{\ell}$.

Let us prove $\widehat{\mathrm{FDP}}_{\tilde{\ell}+1} = \alpha$. First note that $\tilde{\ell} \le n+m-1$ because $\widehat{\mathrm{FDP}}_{n+m-1} = 1$. In addition, we have by definition $\widehat{\mathrm{FDP}}_{\tilde{\ell}+1} \ge \alpha$, which means that $s_{\tilde{\ell}+1} = 1$. Let $v = 1 + \sum_{\ell'=1}^{\tilde{\ell}+1} s_{\ell'}$, $k = 1 \vee \sum_{\ell'=1}^{\tilde{\ell}+1}(1-s_{\ell'})$ and $a = \alpha(n+1)/m$, so that $\frac{n+1}{m}\widehat{\mathrm{FDP}}_{\tilde{\ell}+1} = v/k \ge a$ and $\frac{n+1}{m}\widehat{\mathrm{FDP}}_{\tilde{\ell}} = (v-1)/k < a$. But since v, k and a (by assumption) are integers, we have that $v - 1 < ak$ implies $v \le ak$ and thus we obtain $v = ak$. This gives $\widehat{\mathrm{FDP}}_{\tilde{\ell}+1} = \frac{m}{n+1} v/k = \alpha$.

Now, since $\widehat{\mathrm{FDP}}_{\tilde{\ell}+1} \le \alpha$ we have $\hat{\ell} \ge \tilde{\ell}+1$. But by definition of $\tilde{\ell}$, this implies $\widehat{\mathrm{FDP}}_{\hat{\ell}} \ge \alpha$. Since $\widehat{\mathrm{FDP}}_{\hat{\ell}} \le \alpha$ also holds, this gives $\widehat{\mathrm{FDP}}_{\hat{\ell}} = \alpha$. □

Proposition E.5. *Let $\alpha \in (0,1)$ and $n, m \ge 1$. Then if $\hat{\ell}$ given by (21) exists, we have $\widehat{FDP}_{\hat{\ell}} \ge \frac{m}{n+1}\lfloor \alpha \frac{n+1}{m} \rfloor$.*

Proof. Let $\alpha' = \frac{m}{n+1}\lfloor \alpha \frac{n+1}{m} \rfloor \le \alpha$ and $\mathcal{L}' = \left\{\ell \in \{1,\dots,n+m\} : \widehat{\mathrm{FDP}}_\ell \le \alpha'\right\}$. If this set is empty, this means $\widehat{\mathrm{FDP}}_{\hat{\ell}} > \alpha'$ and the conclusion holds. If this set is not empty, we can consider its maximum $\widetilde{\ell} = \max \mathcal{L}'$. Obviously, we have $\widetilde{\ell} \le \hat{\ell}$. If $\widetilde{\ell} = \hat{\ell}$, then by Lemma E.4 (since $\alpha'(n+1)/m$ is an integer), we have $\widehat{\mathrm{FDP}}_{\widetilde{\ell}} = \alpha'$, and thus also $\widehat{\mathrm{FDP}}_{\hat{\ell}} = \alpha'$ and the conclusion holds. If $\widetilde{\ell} < \hat{\ell}$, we have $\widehat{\mathrm{FDP}}_{\hat{\ell}} > \alpha'$ because $\widetilde{\ell}$ is a maximum. This shows $\widehat{\mathrm{FDP}}_{\hat{\ell}} \ge \alpha'$ in any case and proves the result. □

Lemma E.6. *In the Gaussian linear model of Section 1.4, in case of counting knockoff (case (B), with the notation therein), let us consider the LASSO solution (1) (the solution always exists and we assume that the map giving the solution is fixed and measurable). Then, for any permutation $\sigma : \{1,\dots,n+m\} \to \{1,\dots,n+m\}$ of the columns of $\mathbb{X} = [\tilde{X}\ X]$ that leaves invariant all elements of the set $\{n+j, j : \beta_j \neq 0, 1 \le j \le m\}$ (alternatives), we have*

$$(\{\hat{\beta}_j(\lambda), \lambda \ge 0\})_{1 \le j \le n+m} \sim (\{\hat{\beta}_{\sigma(j)}(\lambda), \lambda \ge 0\})_{1 \le j \le n+m},$$

that is, the set of LASSO paths is invariant under the permutation σ. In particular, when considering the LASSO maximum statistics $Z_j^{LM} = \max\{\lambda \ge 0 : \hat{\beta}_j(\lambda) \neq 0\}$, $1 \le j \le n+m$, or the LASSO coefficient statistics $Z_j^{LC} = |\hat{\beta}_j(\lambda)|$, $1 \le j \le n+m$, the vector $(Z_1,\dots,Z_{n+m}) = (Y_1,\dots,Y_n,X_1,\dots,X_m)$ satisfies Assumption (Exch).

Proof. Since $\|Y - \mathbb{X}b\|^2 = (Y - \mathbb{X}b)^T(Y - \mathbb{X}b) = Y^TY - 2Y^T\mathbb{X}b + b^T\mathbb{X}^T\mathbb{X}b$, we can write the LASSO solution path as $\{\hat{\beta}_j(\lambda), \lambda \geq 0\} = f(\mathbb{X}^TY, \mathbb{X}^T\mathbb{X})$, for some measurable function f. Moreover, by denoting $\mathbb{X}^\sigma$ the matrix $\mathbb{X}$ where the columns have been permuted according the permutation $\sigma : \{1, \dots, n + m\} \to \{1, \dots, n + m\}$, we have by the i.i.d. property of $\mathbb{X}$, that $\mathbb{X}^\sigma \sim \mathbb{X}$ for any permutation σ. Since $\mathbb{X}$ is also independent of ε, we have

$$\begin{aligned}(\mathbb{X}^TY, \mathbb{X}^T\mathbb{X}) &= (\mathbb{X}^T(\mathbb{X}[0\ \beta]^T + \varepsilon), \mathbb{X}^T\mathbb{X}) \\ &\sim ((\mathbb{X}^\sigma)^T(\mathbb{X}^\sigma[0\ \beta]^T + \varepsilon), (\mathbb{X}^\sigma)^T\mathbb{X}^\sigma).\end{aligned}$$

If in addition σ leaves invariant the set $\{n + j, j : \beta_j \neq 0, 1 \leq j \leq m\}$, we have $\mathbb{X}^\sigma[0\,\beta]^T = \sum_{j=1}^m \beta_j \mathbb{X}_{\sigma(n+j)} = \sum_{j:\beta_j\neq 0} \beta_j \mathbb{X}_{\sigma(n+j)} = \sum_{j:\beta_j\neq 0} \beta_j \mathbb{X}_{n+j} = \mathbb{X}[0\,\beta]^T$, the last display is equal to $((\mathbb{X}^\sigma)^T(\mathbb{X}[0\beta]^T + \varepsilon), (\mathbb{X}^\sigma)^T\mathbb{X}^\sigma) = ((\mathbb{X}^\sigma)^TY, (\mathbb{X}^\sigma)^T\mathbb{X}^\sigma)$. This shows

$$\begin{aligned}(\{\hat{\beta}_j(\lambda), \lambda \geq 0\})_{1\leq j\leq n+m} &= f(\mathbb{X}^TY, \mathbb{X}^T\mathbb{X}) \\ &\sim f((\mathbb{X}^\sigma)^TY, (\mathbb{X}^\sigma)^T\mathbb{X}^\sigma) \\ &= (\{\hat{\beta}_{\sigma(j)}(\lambda), \lambda \geq 0\})_{1\leq j\leq n+m},\end{aligned}$$

and concludes the proof. □

Appendix F: Additional numerical experiments

F.1. Comparison to naive procedures

Figure 9 shows a comparison with the two naive procedures defined in Section 1.3 for $\alpha = 0.2$ and $m = 10$ as a function of n. The plots display the FDR and TDR results for BH* (dark green and khaki), $\widehat{\text{BH}}$ (dark blue and cyan), $\widehat{\text{BY}}$ (red and magenta) and $\widehat{\text{BH}}_{\text{Split}}$ (gray-blue and black). The left column corresponds to i.i.d. samples and the right column to equicorrelated Gaussian samples (see Section 6.2). We note immediately very similar performances in the i.i.d. and correlated cases because the considered values of n are large. Under the "full null' (top row), the dense (middle row) and sparse (bottom row) cases, the convergence of the FDR is clearly much slower for the naive approaches than for BH* and $\widehat{\text{BH}}$. The same effect is true for the TDR in the signal-present cases. As expected, the $\widehat{\text{BY}}$ approach is overly conservative to force the FDR control, which leads to a substantial loss in power. The $\widehat{\text{BH}}_{\text{Split}}$ approach indeed controls the FDR but also suffers from power loss at fixed n with respect to the proposed $\widehat{\text{BH}}$ procedure. Note that for $\widehat{\text{BH}}_{\text{Split}}$, the rule of thumb $nk \asymp m/\alpha$ found in Section 4 is expected to become $nk \asymp m^2/\alpha$, because the splitting process uses a training sample of length n/m instead of m. This is what we observe in the middle and lower panels, where values of $k = 1$ and $k \approx 3$ respectively lead to $n = 500$ and $n \approx 170$. In contrast, the corresponding values of n are 10 times smaller for $\widehat{\text{BH}}$, which is consequently much more powerful at fixed n than the naive approaches considered here.

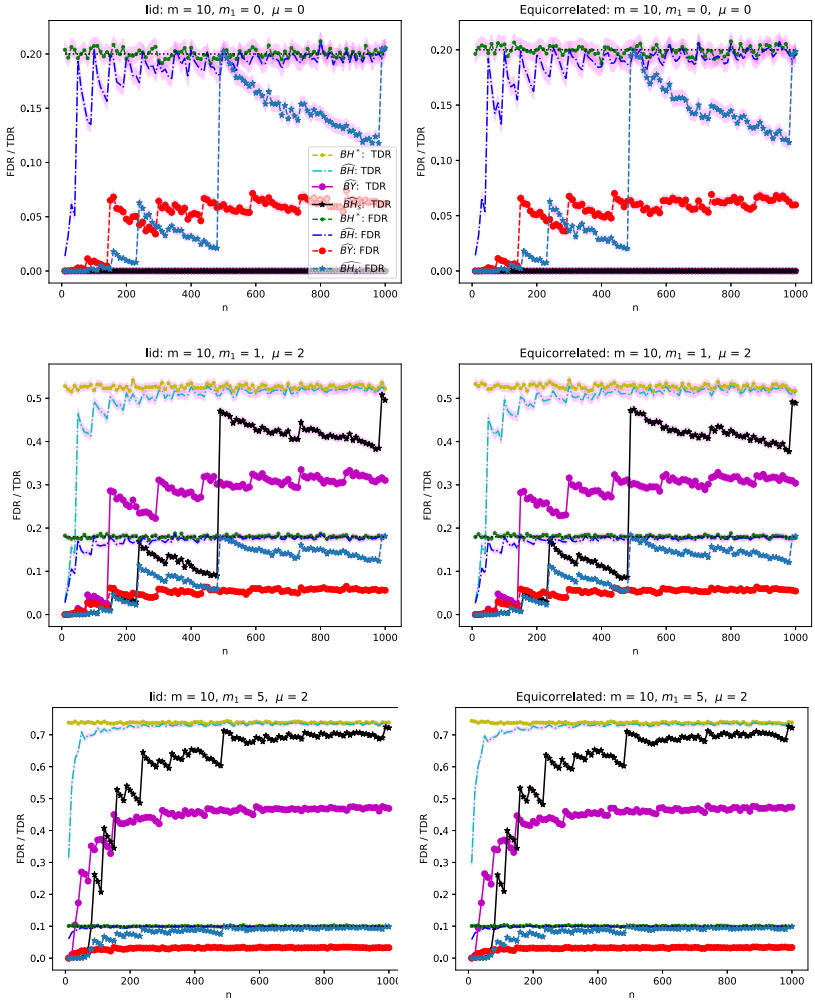


FIG 9. *Comparison with naive procedures for $m = 10$ as a function of n, for $\alpha = 0.2$. The plots display the FDR and TDR results for BH^* (dark green and khaki), $\widehat{BH}$ (dark blue and cyan), $\widehat{BY}$ (red and magenta) and $\widehat{BH}_{Split}$ (gray-blue and black). Left column: i.i.d. samples. Right column: equicorrelated Gaussian samples (see Section 6.2). Top row: full null configuration, middle row: sparse case ($m_1 = 1$), bottom row: dense case ($m_1 = 5$). The number of Monte Carlo simulations used for estimating the FDR and TDR is 10^4 for all plots. The 2σ confidence interval on the estimated FDR and TDR, when visible, is plotted in magenta.*

F.2. Results in a non Gaussian case

Figure 10 illustrates the FDR and TDR in the case where the null distribution P_0 is a Student distribution with zero mean and $\nu = 3$ degrees of freedom,

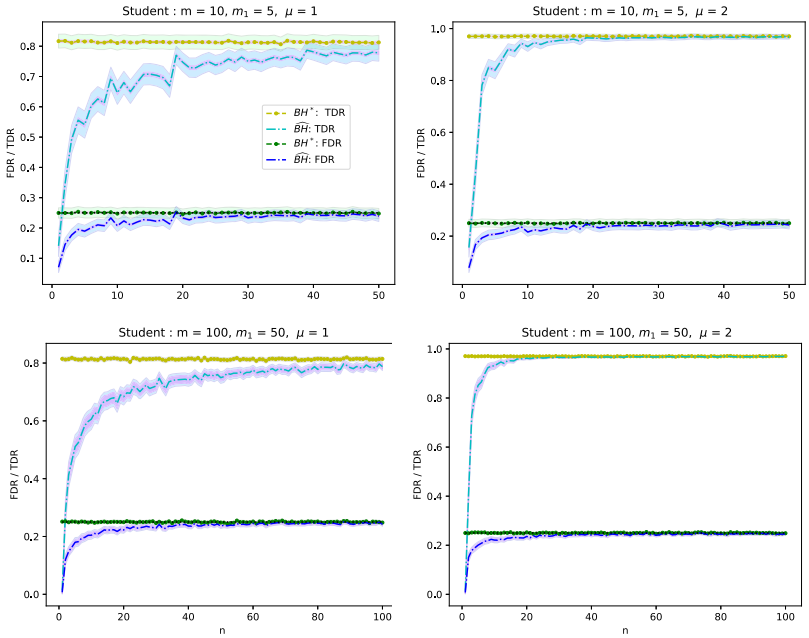


FIG 10. *FDR and TDR results for a Student distribution with three degrees of freedom, in the dense case (compare to Figure 5) :* $m_1 = \frac{m}{2}$*, with* $\mu = 1$ *(left column) and* $\mu = 2$ *(right column). The number of tests* m *equals* 10 *in the top row and* 100 *in the bottom row. The number of Monte Carlo simulations used for estimating the FDR and TDR is* 10^4 *(top row) and* 10^3 *(bottom row). The* 2σ *confidence interval on the estimated FDR and TDR is plotted in magenta. In all plots the standard deviation (divided by* 10*) of the FDP and TDP are shown in shaded green for* BH^* *and shaded blue for* $\widehat{BH}$*.*

rescaled to have unit variance. Comparing with the Gaussian case of (Figure 5), where all other simulation parameters are the same, very similar conclusions can be drawn. In particular, the FDR control of $\widehat{\text{BH}}$ indeed holds also in the case of a heavy tailed distribution since our results are distribution free. Note finally that while the power is larger in the Student case than in the Gaussian case in this setting, the situation can be reversed for other couples (α, μ). As ν increases however, the powers in the Gaussian and standardized Student cases become indeed similar, and hardly distinguishable when ν reaches ≈ 20.

F.3. Results for small values of n

Figure 11 displays the TDR of $\widehat{\text{BH}}$ and BH^* for $n \in \{5, 10\}$ and $\mu \in \{1, 3, 4\}$ in the fully dense case where $m_1 = m$. In this case, when μ is large, there are $k = m$ detectable alternatives, so that the rule of thumb $n = m/(\alpha k)$ reads $n = 1/\alpha = 5$ here. We indeed observe that $\widehat{\text{BH}}$ as a power close to the one of the oracle for $n = 5$ (even more for $n = 10$) when $\mu \in \{3, 4\}$, regardless of m. This once again suggests that the rule of thumb is valid without tuning any constant.

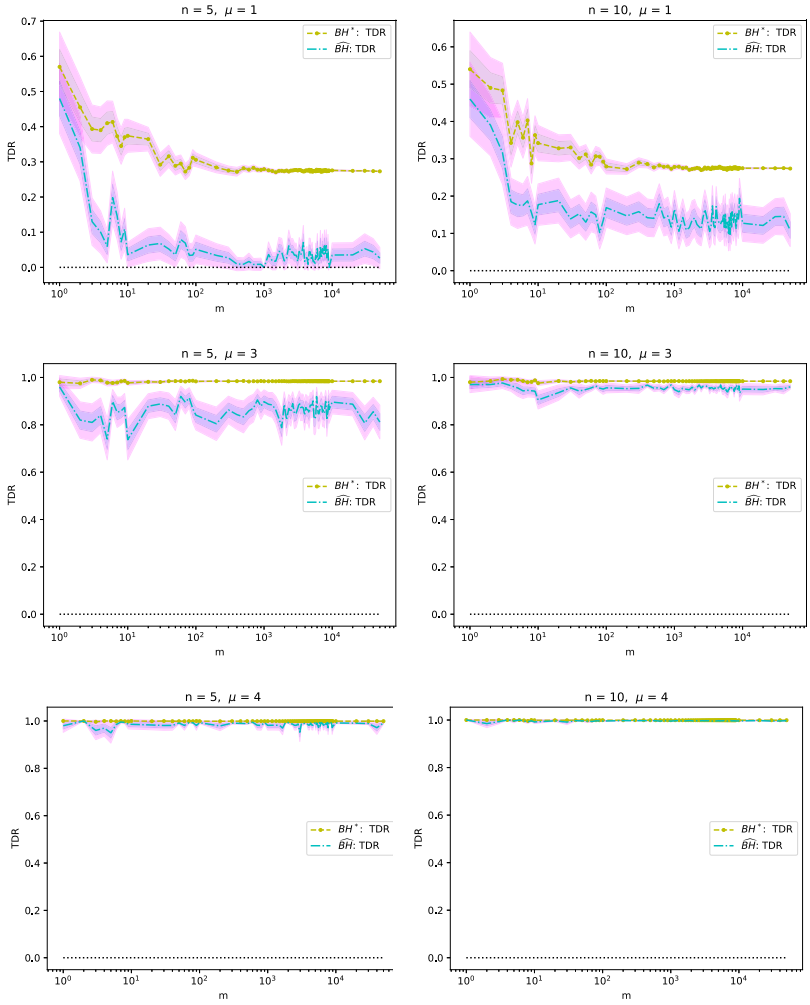


FIG 11. *TDR of* $\widehat{BH}$ *(*$\alpha = 0.2$*) in the fully dense case* $m_1 = m$ *at fixed* $n = 5$ *(left column) and* $n = 10$ *(right column) for* m *varying in the range* $[1\ 5 \times 10^4]$*. The signal amplitude* μ *increases from* $\mu = 1$ *(top row) to* $\mu = 3$ *(bottom row). The plots display the TDR results for* BH^* *(khaki) and* $\widehat{BH}$ *(cyan). The number of Monte Carlo simulations used for estimating TDR is* 10^2 *for all plots. The standard deviation (divided by* 10*) of the FDP is shown in shaded blue and the* 2σ *confidence interval on the estimated TDR is plotted in magenta.*

F.4. Results for bbBH procedure

This section describes the numerical experiment discussed in Appendix A.

First, we describe in detail the third procedure (locfdr): it is based on ℓ-values $\ell_i = \pi_0 g_0(T_i)/g(T_i)$, $1 \le i \le m$, where π_0 is the probability that a null hypothesis is true and using the notation of Appendix A. Note that the latter

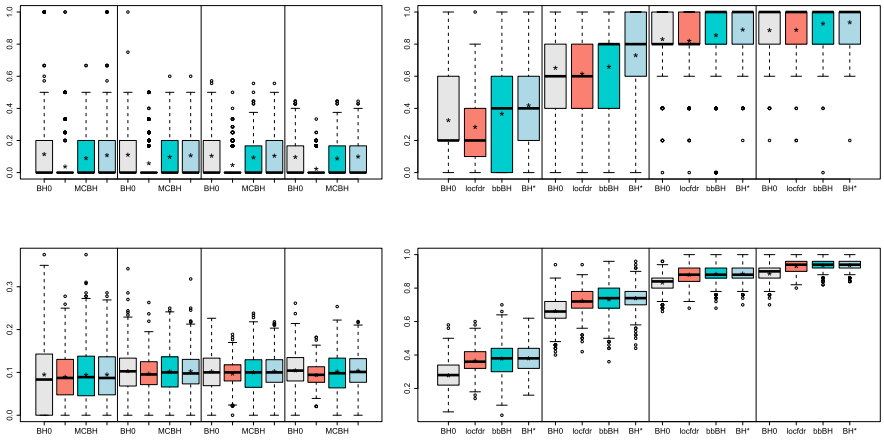


FIG 12. *Boxplots of the FDP (left) and TDP (right) for the procedures BH*0*, bbBH, Locfdr, BH***, see text. For each boxplot, the FDR (left) and TDR (right) are depicted with the symbol "*$\star$*". Each picture is composed of* 4 *panels, one for each value of the alternative mean* $\mu \in \{1, 2, 3, 4\}$*. Top* $m = 10$*, Bottom* $m = 100$*.* $\alpha = 0.2$*,* $m_0 = m/2$*,* 500 *replications.*

is not well defined in our setting since the null hypothesis are not random. Indeed, ℓ-values are generally defined in the so-called "two group model" (Efron et al., 2001) that uses an additional mixture effect for the configuration vector $\theta = (\theta_i)_{1 \le i \le m} \in \{0, 1\}^m$ with $\theta_i = 0$ if and only if i-th null hypothesis is true. Nevertheless, we can fix π_0 to the value m_0/m and compute the ℓ-values accordingly. Then, the version of the local FDR procedure controlling the FDR (introduced in Sun and Cai (2007)) reads as follows:

- first order the ℓ-values $\ell_{(1)} \le \cdots \le \ell_{(m)}$;
- reject the null corresponding to the $\hat{k}$ smallest ℓ-values where

$$\hat{k} = \max \left\{ k \in \{0, \dots, m\} \ : \ k^{-1} \sum_{i=1}^{k} \ell_{(i)} \le \alpha \right\}.$$

Since the value of m_0/m is used in this locfdr procedure and to make the comparison fair with BH0, bbBH and BH*, we apply the locfdr at level $\alpha/(m_0/m)$. This way, all procedures uses the same parameter informations and target the same FDR level $\alpha m_0/m$.

Figure 12 displays the FDP/TDP achieved by each procedure, in a Gaussian setting where g_0 is the density of the $\mathcal{N}(0, 1)$ and g_1 is the density of the Cauchy distribution with mean μ, taken in the range $\{1, 2, 3, 4\}$.

Acknowledgments

We are grateful to Lihua Lei for very interesting discussions, to Sabine Houssaye for her help when proving Lemma E.4 and to Guillaume Lecué for helpful comments.

References

[1] Abraham, K., Castillo, I., and Gassiat, E. (2021). Multiple testing in non-parametric hidden markov models: An empirical bayes approach. *arXiv preprint arXiv:2101.03838*.

[2] Arlot, S., Blanchard, G., and Roquain, E. (2010). Some nonasymptotic results on resampling in high dimension. I. Confidence regions. *Ann. Statist.*, 38(1):51–82. MR2589316

[3] Azriel, D. and Schwartzman, A. (2015). The empirical distribution of a large number of correlated normal variables. *Journal of the American Statistical Association*, 110(511):1217–1228. MR3420696

[4] Bacon, R., Mary, D., Garel, T., Blaizot, J., Maseda, M., Schaye, J., Wisotzki, L., Conseil, S., Brinchmann, J., Leclercq, F., Abril-Melgarejo, V., Boogaard, L., Bouché, N. F., Contini, T., Feltre, A., Guiderdoni, B., Herenz, C., Kollatschny, W., Kusakabe, H., Matthee, J., Michel-Dansac, L., Nanayakkara, T., Richard, J., Roth, M., Schmidt, K. B., Steinmetz, M., Tresse, L., Urrutia, T., Verhamme, A., Weilbacher, P. M., Zabl, J., and Zoutendijk, S. L. (2021). The muse extremely deep field: The cosmic web in emission at high redshift. *A&A*, 647:A107.

[5] Barber, R. F. and Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *Ann. Statist.*, 43(5):2055–2085. MR3375876

[6] Barber, R. F. and Candès, E. J. (2019). A knockoff filter for high-dimensional selective inference. *Ann. Stat.*, 47(5):2504–2537. MR3988764

[7] Bates, S., Candès, E., Janson, L., and Wang, W. (2020). Metropolized knockoff sampling. *Journal of the American Statistical Association*, pages 1–15. MR4309282

[8] Bates, S., Candès, E., Lei, L., Romano, Y., and Sesia, M. (2021). Testing for outliers with conformal p-values.

[9] Bayati, M. and Montanari, A. (2011). The lasso risk for gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017. MR2951312

[10] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, 57(1):289–300. MR1325392

[11] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, 29(4):1165–1188. MR1869245

[12] Besag, J. and Clifford, P. (1991). Sequential monte carlo p-values. *Biometrika*, 78(2):301–304. MR1131163

[13] Blanchard, G., Lee, G., and Scott, C. (2010). Semi-supervised novelty detection. *J. Mach. Learn. Res.*, 11:2973–3009. MR2746544

[14] Blanchard, G. and Roquain, E. (2008). Two simple sufficient conditions for FDR control. *Electron. J. Stat.*, 2:963–992. MR2448601

[15] Cai, T. T. and Sun, W. (2009). Simultaneous testing of grouped hypotheses: finding needles in multiple haystacks. *J. Amer. Statist. Assoc.*, 104(488):1467–1481. MR2597000

[16] Cai, T. T., Sun, W., and Wang, W. (2019). Covariate-assisted ranking and

screening for large-scale two-sample inference. In *Royal Statistical Society*, volume 81. MR3928141

[17] Candès, E., Fan, Y., Janson, L., and Lv, J. (2018). Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *J. R. Stat. Soc., Ser. B, Stat. Methodol.*, 80(3):551–577. MR3798878

[18] Carpentier, A., Delattre, S., Roquain, E., and Verzelen, N. (2021). Estimating minimum effect with outlier selection. *Annals of Statistics*, 49(1):272–294. MR4206678

[19] Choquet, É., Bryden, G., Perrin, M. D., Soummer, R., Augereau, J.-C., Chen, C. H., Debes, J. H., Gofas-Salas, E., Hagan, J. B., Hines, D. C., Mawet, D., Morales, F., Pueyo, L., Rajan, A., Ren, B., Schneider, G., Stark, C. C., and Wolff, S. (2018). HD 104860 and HD 192758: Two debris disks newly imaged in scattered light with the Hubble space telescope. *The Astrophysical Journal*, 854(1):53.

[20] Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap methods and their application*. Number 1. Cambridge University Press. MR1478673

[21] Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, 50(272):1096–1121.

[22] Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Am. Stat. Assoc.*, 99(465):96–104. MR2054289

[23] Efron, B. (2007). Doing thousands of hypothesis tests at the same time. *Metron - International Journal of Statistics*, LXV(1):3–21.

[24] Efron, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.*, 23(1):1–22. MR2431866

[25] Efron, B. (2009). Empirical Bayes estimates for large-scale prediction problems. *J. Am. Stat. Assoc.*, 104(487):1015–1028. MR2562003

[26] Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.*, 96(456):1151–1160. MR1946571

[27] Finner, H. and Strassburger, K. (2007). Step-up related simultaneous confidence intervals for mcc and mcb. *Biometrical Journal*, 49(1):40–51. MR2339215

[28] Fisher, R. A. (1935). *The Design of Experiments*. Oliver and Boyd, Edinburgh.

[29] Fithian, W. and Lei, L. (2020). Conditional calibration for false discovery rate control under dependence.

[30] Foster, D. P. and Stine, R. A. (2008). α-investing: a procedure for sequential control of expected false discoveries. *J. R. Stat. Soc., Ser. B, Stat. Methodol.*, 70(2):429–444. MR2424761

[31] Gandy, A. and Hahn, G. (2014). MMCTest – a safe algorithm for implementing multiple Monte Carlo tests. *Scand. J. Stat.*, 41(4):1083–1101. MR3277039

[32] Genovese, C. and Wasserman, L. (2004). A stochastic process approach to false discovery control. *Ann. Statist.*, 32(3):1035–1061. MR2065197

[33] Genovese, C. R. and Wasserman, L. (2006). Exceedance control of the

false discovery proportion. *J. Amer. Statist. Assoc.*, 101(476):1408–1417. MR2279468

[34] Goeman, J. J. and Solari, A. (2011). Multiple testing for exploratory research. *Statist. Sci.*, 26(4):584–597. MR2951390

[35] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

[36] Guo, W. and Peddada, S. (2008). Adaptive choice of the number of bootstrap samples in large scale multiple testing. *Stat. Appl. Genet. Mol. Biol.*, 7(1):19. Id/No 13. MR2386329

[37] Heller, R. and Yekutieli, D. (2014). Replicability analysis for genome-wide association studies. *Ann. Appl. Stat.*, 8(1):481–498. MR3191999

[38] Hemerik, J., Solari, A., and Goeman, J. J. (2019). Permutation-based simultaneous confidence bounds for the false discovery proportion. *Biometrika*, 106(3):635–649. MR3992394

[39] Hsu, J. (1996). *Multiple comparisons: theory and methods*. CRC Press. MR1629127

[40] Katsevich, E. and Sabatti, C. (2019). Multilayer knockoff filter: Controlled variable selection at multiple resolutions. *The Annals of Applied Statistics*, 13(1):1. MR3937419

[41] Kingma, D. P. and Welling, M. (2014). Auto-encoding variational bayes. In Bengio, Y. and LeCun, Y., editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

[42] Lin, D. (2005). An efficient monte carlo approach to assessing statistical significance in genomic studies. *Bioinformatics*, 21(6):781–787.

[43] Mary, D., Bacon, R., Conseil, S., Piqueras, L., and Schutz, A. (2020). ORIGIN: Blind detection of faint emission line galaxies in muse datacubes. *A&A*, 635:A194.

[44] 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(5):Art. 4, front matter+39. MR2990984

[45] Phipson, B. and Smyth, G. K. (2010). Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1). MR2746025

[46] Romano, J. P. and Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *J. Amer. Statist. Assoc.*, 100(469):94–108. MR2156821

[47] Romano, J. P. and Wolf, M. (2007). Control of generalized error rates in multiple testing. *Ann. Statist.*, 35(4):1378–1408. MR2351090

[48] Roquain, E. and Verzelen, N. (2020a). False discovery rate control with unknown null distribution: illustrations on real data sets. `https://github.com/eroquain/empiricalnull/blob/main/vignette.pdf`.

[49] Roquain, E. and Verzelen, N. (2020b). False discovery rate control with unknown null distribution: is it possible to mimic the oracle?

[50] Sandve, G. K., Ferkingstad, E., and Nygård, S. (2011). Sequential monte carlo multiple testing. *Bioinformatics*, 27(23):3235–3241.

[51] Schwartzman, A. (2010). Comment: "Correlated z-values and the accuracy of large-scale statistical estimates". *J. Amer. Statist. Assoc.*, 105(491):1059–1063. MR2752600

[52] Stephens, M. (2017). False discovery rates: a new deal. *Biostatistics*, 18(2):275–294. MR3824755

[53] Sulis, S., Mary, D., and Bigot, L. (2017). A study of periodograms standardized using training datasets and application to exoplanet detection. *IEEE Transactions on Signal Processing*, 65(8):2136–2150. MR3608115

[54] Sulis, S., Mary, D., and Bigot, L. (2020). 3D magneto-hydrodynamical simulations of stellar convective noise for improved exoplanet detection - I. Case of regularly sampled radial velocity observations. *A&A*, 635:A146.

[55] Sun, L. and Stephens, M. (2018). Solving the empirical bayes normal means problem with correlated noise.

[56] Sun, W. and Cai, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *J. Am. Stat. Assoc.*, 102(479):901–912. MR2411657

[57] Sun, W. and Cai, T. T. (2009). Large-scale multiple testing under dependence. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 71(2):393–424. MR2649603

[58] 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

[59] Weinstein, A., Barber, R., and Candès, E. (2017). A power and prediction analysis for knockoffs with lasso statistics.

[60] Weinstein, A., Su, W. J., Bogdan, M., Barber, R. F., and Candes, E. J. (2020). A power analysis for knockoffs with the lasso coefficient-difference statistic. *arXiv preprint arXiv:2007.15346*.

[61] Westfall, P. H. and Young, S. S. (1993). *Resampling-Based Multiple Testing*. Wiley. Examples and Methods for P-Value Adjustment.

[62] Xu, Z. and Ramdas, A. (2021). Dynamic algorithms for online multiple testing.

[63] Zhang, M. J., Zou, J., and Tse, D. (2019). Adaptive Monte Carlo Multiple Testing via Multi-Armed Bandits. *arXiv:1902.00197 [cs, math, q-bio, stat]*.