The Annals of Statistics
2020, Vol. 48, No. 3, 1281–1303
https://doi.org/10.1214/19-AOS1847

POST HOC CONFIDENCE BOUNDS ON FALSE POSITIVES USING REFERENCE FAMILIES

BY GILLES BLANCHARD[1], PIERRE NEUVIAL[2] AND ETIENNE ROQUAIN[3]

[1]*Institut für Mathematik, Universität Potsdam, gilles.blanchard@math.uni-potsdam.de*
[2]*Institut de Mathématiques de Toulouse, UMR 5219, Université de Toulouse, CNRS, pierre.neuvial@math.univ-toulouse.fr*
[3]*Sorbonne Université, Sorbonne Paris Cité, CNRS, Laboratoire de Probabilités Statistique et Modélisation, LPSM, etienne.roquain@upmc.fr*

We follow a post hoc, "user-agnostic" approach to false discovery control in a large-scale multiple testing framework, as introduced by Genovese and Wasserman [*J. Amer. Statist. Assoc.* **101** (2006) 1408–1417], Goeman and Solari [*Statist. Sci.* **26** (2011) 584–597]: the statistical guarantee on the number of correct rejections must hold for any set of candidate items, possibly selected by the user after having seen the data. To this end, we introduce a novel point of view based on a family of reference rejection sets and a suitable criterion, namely the joint familywise error rate over that family (JER for short). First, we establish how to derive post hoc bounds from a given JER control and analyze some general properties of this approach. We then develop procedures for controlling the JER in the case where reference regions are p-value level sets. These procedures adapt to dependencies and to the unknown quantity of signal (via a step-down principle). We also show interesting connections to confidence envelopes of Meinshausen [*Scand. J. Stat.* **33** (2006) 227–237]; Genovese and Wasserman [*J. Amer. Statist. Assoc.* **101** (2006) 1408–1417], the closed testing based approach of Goeman and Solari [*Statist. Sci.* **26** (2011) 584–597] and to the higher criticism of Donoho and Jin [*Ann. Statist.* **32** (2004) 962–994]. Our theoretical statements are supported by numerical experiments.

1. Introduction. Large-scale multiple inference with a rigorous statistical guarantee has become a topic of ever increasing relevance with the advent of very high-dimensional data in numerous application areas. Classical multiple testing procedures prescribe a rejection set based on the amount of false positives that the user might tolerate (e.g., false discovery rate control at level 5%). However, if the result does not correspond to what the user expected, they may tend to "snoop" in the data, possibly concentrating only on a set R of hypotheses that appear promising to them. Even when motivated by plausible justifications, any such approach will invalidate standard statistical guarantee because of the *selection effect*. This is illustrated on Figure 1, where only "noisy" measurements have been generated: within the selected set (in blue), 5 points stand out. However, this is only due to the selection effect: the blue data set comes from a larger data set (green) where these 5 measures are just the 5 maximum (noisy) measurements. As a consequence, while building a statistical guarantee on the selected set R, the overall size of the data set should be considered. This is the aim of the so-called "post-selection" (or post hoc) inference.

A particular case of post hoc inference is faced when the selected set R is obtained by a prespecified selection method, with a statistical guarantee holding either conditionally on the selection (Belloni, Chernozhukov and Hansen (2014), Fithian, Sun and Taylor (2014), Lee et al. (2016), Taylor and Tibshirani (2015)) or unconditionally (Benjamini and Yekutieli

Received March 2017; revised February 2019.
MSC2010 subject classifications. Primary 62G10; secondary 62H15.
Key words and phrases. Post hoc inference, multiple testing, Simes inequality, family-wise error rate, step-down algorithm, dependence, higher criticism.

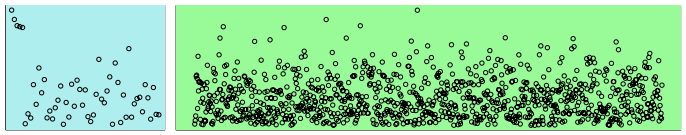

FIG. 1. *Illustration of the post hoc selection effect. Right*: *virtual data set with* 1000 *measurements. Left*: *data set of* 55 *measurements selected from the right dataset. Measures have been generated as i.i.d. absolute values of* $\mathcal{N}(0,1)$.

(2005)). Other approaches diminish the selection effect by using sample splitting; see, for example, Bühlmann and Mandozzi (2014), Dezeure et al. (2015) and references therein.

However, in these approaches, since the selection step is fixed, this does not allow for arbitrary "data snooping" or *ad hoc* selection rules often used in exploratory research. More generally, elaborate selection rules possibly consisting in several stages and involving user-fixed tuning constants are commonly used in a variety of contexts, for instance:

- in neural activity detection from brain imaging data, cluster-extent approaches (Woo, Krishnan and Wager (2014)) select voxels by a two-stage process, first building groups of contiguous voxels whose activity levels all pass a user-defined threshold, then performing a correction to select a subset of clusters. The second stage only ensures that each cluster contains at least one truly active voxel, but there is no additional statistical guarantee about the proportion of active voxels among the selected.
- in the context of gene or protein activity change detection, a two-sample rank test might be used to detect activity changes, while requiring that the log-ratio of average observed activities of the two samples ("fold change") is larger than a certain user-specified level; see Li (2012). In other words, for each hypothesis a statistic T_1 is used for constructing a standard test, but a different statistic T_2 is used for screening, with the two statistics not being independent.

A point of view argued in several papers in various statistical contexts (Bachoc, Preinerstorfer and Steinberger (2019), Berk et al. (2013), Goeman and Solari (2011)) is that in absence of precise information of the user's selection strategy, it is desirable to provide a statistical guarantee *simultaneously* for any possible selected set. In this paper, we adopt this view and focus on simultaneous upper bounds on the number of false positives on the selected set, as proposed in the seminal papers Genovese and Wasserman (2006) and Goeman and Solari (2011). More formally, our goal is to build a functional $V(\cdot)$ defined on all subsets of hypotheses, such that the following uniform guarantee holds:

$$\mathbb{P}\big(\forall R \subset \{1,\dots,m\} : |\mathcal{H}_0 \cap R| \le V(R)\big) \ge 1-\alpha, \tag{1}$$

where m is the number of null hypotheses to be tested (identified with their respective index) and $\mathcal{H}_0 \subset \{1,\dots,m\}$ corresponds to the (unknown) set of true null hypotheses. This general principle is "user-agnostic," in the sense that the provided inference is "ready for any selected set" (the "for all R" being inside the probability). Observe that a bound $V(\cdot)$ satisfying the above guarantee can also inform the choice of the final rejected set R; for example, the user is allowed to optimize some function of $V(R)$, possibly subject to geometrical or data-dependent constraints on R.

Note that providing such a bound $V(\cdot)$ is equivalent to build a uniform upper-bound on the false discovery proportion (FDP) $|\mathcal{H}_0 \cap R|/|R|$ by considering $V(R)/|R|$, which was the initial formulation of Genovese and Wasserman (2006). Such confidence envelopes for the FDP have also been considered in Genovese and Wasserman (2004), Section 6, as well as Meinshausen and Bühlmann (2005), Meinshausen (2006) when the coverage is restricted to

selection sets R that are p-value level sets, that is, of the form $R = \{i : p_i \le t\}$, for some $t \in [0, 1]$.

The main idea of our method is to build a reference family $(R_k)_{1\le k\le K}$ of rejection sets for which the guarantee (1) is ensured to hold (in restriction to that family) for some $\zeta_k = V(R_k)$. This will induce a post hoc bound, valid for any R, by an interpolation principle. Calibrating such a family brings new challenges, which can be formulated in terms of controlling a multiple testing criterion that we call "joint (familywise) error rate" (JER for short). While we formulate the latter in a very general way, let us first discuss as an introductive example the situation where the reference family consists of p-value level sets $R_k = \{i : p_i \le t_k\}$ and $\zeta_k = k - 1$. In that case, the JER of $\mathcal{T} = (t_k)_{1\le k\le K}$ is related to the distribution of $p_{(k:\mathcal{H}_0)}$, the kth smallest value in the set $\{p_i, i \in \mathcal{H}_0\}$ as follows:

$$\mathrm{JER}(\mathcal{T}) = \mathbb{P}\big(\exists k \in \{1, \dots, K \wedge m_0\} : p_{(k:\mathcal{H}_0)} < t_k\big), \tag{2}$$

where $m_0 = |\mathcal{H}_0|$ is the number of true null hypotheses. A general intuition is that the threshold t_k should be chosen as an appropriate quantile of the distribution of $p_{(k:\mathcal{H}_0)}$, with some extra slack to take into account for uniformity in k. We establish that if $\mathrm{JER}(\mathcal{T}) \le \alpha$ holds, then the functional

$$V(R) = \min_{k\in\{1,\dots,K\}} \Big\{ \sum_{i\in R} \mathbb{1}\big\{p_i(X) \ge t_k\big\} + k - 1 \Big\}, \quad R \subset \{1, \dots, m\} \tag{3}$$

is a valid post hoc bound.

The threshold family $t_k = \alpha k/m$, $1 \le k \le K = m$, is referred to as the Simes family throughout the paper. It satisfies $\mathrm{JER}(\mathcal{T}) \le \alpha$ when the family of p-value is positive regression dependent on each element of the subset $\mathcal{H}_0$ (in short, PRDS), as defined in Benjamini and Yekutieli (2001). The corresponding post hoc bound (3) is called the Simes post hoc bound, and will be a baseline for our work.

The bound $V(R)$ given by (3) has a simple graphical interpretation, based on the equivalent expression[1] $|R| - V(R) = \min\{u \in \{0, \dots, |R|\} : \forall v \in \{u + 1, \dots, \min(u + K, |R|)\} : p_{(v:R)} \ge t_{v-u}\}$. Two examples are displayed in Figure 2, for the Simes family, and another family based on the quantiles of the Beta distribution. The latter will be one of the new contributions of this paper; see Section 5.2. This already illustrates that an improvement is achievable when the sorted p-value curve has a specific shape.

REMARK 1.1. Note that the simple version (2) of the JER control was already implicitly defined by Meinshausen (2006). Also, the bound (3) can be seen as an extension of the "augmentation procedure" of van der Laan, Dudoit and Pollard (2004) and Genovese and Wasserman (2006); see Section 2 of the present paper for a proof in a more general context. Finally, the bound (3) and in particular the formula leading to the interpretation of Figure 2 turns out to coincide with the post hoc bound proposed in Goeman and Solari (2011) as a "shortcut" of the closed testing bound in the specific context of local tests; see Section S-1.4 for more details.

The main contributions of the present work are the following:

- We introduce a general and flexible framework to build post hoc bounds from reference rejection families. The confidence coverage of such a post hoc bound is ensured by showing that the reference family controls a JER criterion. We establish some fundamental properties of this method and of the resulting bounds (Section 2).

[1]The idea for this formulation and the graphical presentation used in Figure 2 is due to J. Goeman.

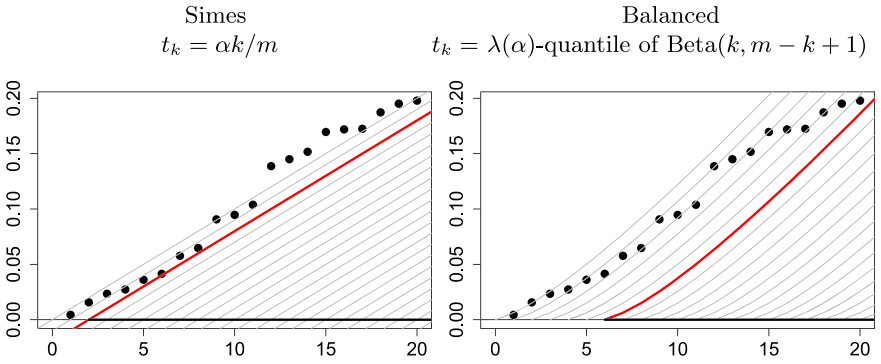

FIG. 2. *Sorted p-values of a subset R of $\{1,\dots,m\}$ (dots). Thresholds $v \in \{u+1,\dots,|R|\} \mapsto t_{v-u}$ for $u \in \{0,\dots,|R|\}$ (in red for $u = |R| - V(R)$, in light gray otherwise). The post hoc bound $V(R)$ (3) corresponds the length of the bold line on the X-axis. $K = m$, $|R| = 20$, $m = 50$, $\alpha = 0.5$. For the balanced threshold (right), the functional $\lambda(\alpha)$ is defined in Section 5.2 (single step, independence).*

- We develop JER controlling procedures of the more specific form given by (2), with adaptivity to known or unknown dependence and to the proportion of true null hypotheses (Sections 3 to 5).
- We explore connections of our work to confidence envelopes (Genovese and Wasserman (2004, 2006), Meinshausen (2006), Meinshausen and Bühlmann (2005)), closed testing (Goeman and Solari (2011)) and higher criticism (Donoho and Jin (2004)) (supplementary material).
- These procedures are implemented in an open-source R (R Core Team (2017)) package (Blanchard, Neuvial and Roquain (2019)). This package was used to perform numerical experiments (Section 6) to illustrate our theoretical statements.

The paper is organized as follows. In Section 2, we expose the general approach, with an emphasis on the computability of the obtained bound. We propose a low-complexity conservative proxy and analyze when it coincides with the optimal bound. In the following sections, we specifically focus on the JER control of the form (2), in some exemplary models under known or unknown dependence structure. The models are presented in Section 3. In Section 4, after briefly discussing the shortcomings of the basic JER control obtained using the classical Simes inequality, we present improvements to this basic case by considering more general threshold families called templates; incorporating adaptation to noise dependence structure, and to the proportion of null hypotheses using a step-down principle. Two specific examples of such templates combined with this improved methodology are developed in Section 5. In Section 6, we present the results of numerical simulations illustrating and comparing the developed methods. We conclude with a discussion of various points in Section 7. Due to space constraints, proofs as well as some additional results are postponed to the supplementary material Blanchard, Neuvial and Roquain (2019). The sections of this supplement are referred to with an additional symbol "S-" in the numbering.

2. JER control: Principle and properties. In this section, we introduce the framework (Section 2.1) for post hoc multiple testing inference, and propose a general approach to tackle this problem based on a reference family of rejection sets (Section 2.2). Proceeding from the general to the particular, we first study and discuss some generic properties of this approach (Section 2.3) before focusing on more specific choices for the reference family leading to (2) and (3) (Section 2.4). Formal proofs for theoretical claims in this section are found in Section S-7.1.

2.1. *Aim.* Formally, let X denote observed data generated from a statistical model $(\mathcal{X}, \mathfrak{X}, P)$, $P \in \mathcal{P}$, and assume we want to test a collection of null hypotheses $H_{0,i} \subset \mathcal{P}$ indexed by $i \in \mathbb{N}_m := \{1, \dots, m\}$. For any $P \in \mathcal{P}$, we denote by $\mathcal{H}_0(P)$ the set of (indices of) true null hypotheses satisfied by P, that is, $\mathcal{H}_0(P) = \{i \in \mathbb{N}_m : P \in H_{0,i}\}$, and by $m_0(P)$ its cardinality (or $\mathcal{H}_0$, m_0 for short). We denote by $\pi_0 = m_0/m$ the proportion of true nulls. We also let $\mathcal{H}_1(P) = \mathbb{N}_m \backslash \mathcal{H}_0(P)$ be the set of (indices of) false nulls and $m_1(P) = m - m_0(P)$ its cardinality (or $\mathcal{H}_1$, m_1 for short).

Our main objective in this paper is to find a function $V(X, R)$ (denoted by $V(R)$ for short) satisfying

$$\text{for all } P \in \mathcal{P}, \quad \mathbb{P}_{X\sim P}\big(\forall R \subset \mathbb{N}_m, \big|R \cap \mathcal{H}_0(P)\big| \le V(R)\big) \ge 1 - \alpha. \tag{PH_α}$$

If the above is satisfied, $V(R)$ gives a level $1-\alpha$ confidence bound for the number of false rejections in a set R of (indices of) rejected hypotheses that is *uniformly valid* over all possible choices of R. Letting $S(R) = |R| - V(R)$, the property (PH_α) equivalently provides the following simultaneous lower bound on $|R \cap \mathcal{H}_1(P)|$, that is, evidence of signal in R:

$$\text{for all } P \in \mathcal{P}, \quad \mathbb{P}_{X\sim P}\big(\forall R \subset \mathbb{N}_m, \big|R \cap \mathcal{H}_1(P)\big| \ge S(R)\big) \ge 1 - \alpha.$$

As the the above bounds are uniformly valid over all possible choice of R, they will apply (with probability at least $1-\alpha$) to any arbitrary data-dependent choice of R made by the user, including choices made after looking at the value of the bound itself for different candidates for R. For instance, R can be chosen as maximizing $|\hat{R}|$ among those $\hat{R}$ satisfying $S(\hat{R})/|\hat{R}| \ge 0.5$ (more than half of signal in $\hat{R}$ with high probability). Obviously, the theoretical guarantees for $\hat{R}$ also hold because the bounds are uniform in $R \subset \mathbb{N}_m$.

2.2. *General principle.* The question of how to obtain a control of the general form (PH_α) is statistical as well as computational in nature, since it is not practically feasible to consider individually all 2^m possibilities for candidate rejection sets R as soon as m exceeds a couple of dozens. Provided that the statistical guarantee holds, we would ideally wish that the bound $V(R)$ is computable efficiently for any candidate R (or family thereof) suggested by the user.

In this section, we consider a general approach to the problem based on a reference family with a controlled Joint familywise Error Rate (JER). The basic argument is illustrated by Figure 3. Imagine that a subset A of hypotheses is guaranteed to contain less than 5 true nulls, that is, $|A \cap \mathcal{H}_0(P)| \le 5$. Then this also provides information on other subsets $R \subset \mathbb{N}_m$ with $R \ne A$. Namely, for any $R \subset \mathbb{N}_m$, $|R \cap \mathcal{H}_1(P)| \ge |R \cap A| - 5$. Of course, while this information is useful for R if $|R \cap A| \ge 6$, it is not if $|R \cap A| \le 5$ (nonpositive bound), as in Figure 3. Next, if we want to improve the bound, we can consider another set B (here including A) with the property $|B \cap \mathcal{H}_0(P)| \le 7$ (say). In the situation pictured in Figure 3, this ensures that R contains at least one element which is in $\mathcal{H}_1(P)$. Similarly, adding another set C (here disjoint from A and B) with the property $|C \cap \mathcal{H}_0(P)| \le 1$ (say), ensures that R contains at least two elements which are in $\mathcal{H}_1(P)$.

More generally, let us assume that we have at hand $\mathfrak{R} = ((R_1(X), \zeta_1(X)), \dots, (R_K(X), \zeta_K(X)))$ a data-dependent collection of subsets R_k of $\mathbb{N}_m$ and integer numbers ζ_k (we will often omit the dependence in X to ease notation), such that, with probability larger than $1-\alpha$, the set $R_k(X)$ does not contain more than $\zeta_k(X)$ elements of $\mathcal{H}_0(P)$, uniformly over k, that is,

$$\text{For all } P \in \mathcal{P}, \quad \mathrm{JER}(\mathfrak{R}, P) \le \alpha, \tag{4}$$

where we have denoted

$$\mathrm{JER}(\mathfrak{R}, P) := \mathbb{P}_{X\sim P}\big(\exists k \in \mathbb{N}_K : \big|R_k(X) \cap \mathcal{H}_0\big| > \zeta_k(X)\big). \tag{5}$$

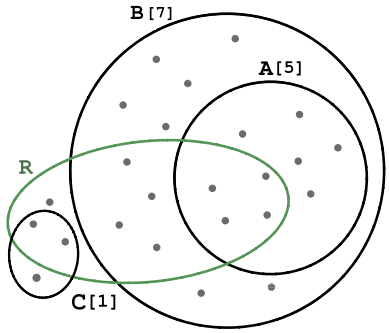

FIG. 3. *Toy example*: *use of a reference family with three subsets* A, B *and* C *to build a post hoc bound on the number of true positives in an arbitrary candidate rejection set* R. *In brackets*, *a known bound on the number of false positives in each set*.

We see $\mathfrak{R}$ as a *reference family* of rejection sets for which a statistical guarantee on the number of false rejections is ensured, and based on which we will build a post hoc bound. The cardinality (or size) K of the reference family is also allowed to be data-dependent in the most general form, although this dependence is not acknowledged for in our notation for simplicity.

How can we "interpolate" from the control on a reference family (4) to a control on all possible rejection sets (PH_α)? On the event where $\forall k \in \mathbb{N}_K$, $|R_k(X) \cap \mathcal{H}_0| \le \zeta_k(X)$, the only available information on the unknown subset $\mathcal{H}_0$ is that it is an element of the collection of subsets

$$\mathcal{A}(\mathfrak{R}) = \big\{A \subset \mathbb{N}_m : \forall k \in \mathbb{N}_K, |R_k \cap A| \le \zeta_k\big\}. \tag{6}$$

As a result, the best we can do to bound $|R \cap \mathcal{H}_0|$ for any proposed rejection set R is a worst-case bound under this constraint:

$$V^*_{\mathfrak{R}}(R) := \max_{A \in \mathcal{A}(\mathfrak{R})} |R \cap A|, \quad R \subset \mathbb{N}_m. \tag{7}$$

The next result formalizes the link between JER control and the associated post hoc bound. It is a purely deterministic result, analyzing the information available under JER control.

PROPOSITION 2.1. *Let* $\mathfrak{R} = (R_k, \zeta_k)_{k \in \mathbb{N}_K}$ *be a collection of subsets* $R_k \subset \mathbb{N}_m$ *and of integers* ζ_k. *Then for any* $A \subset \mathbb{N}_m$, *the following statements are equivalent*:

1. $\forall k \in \mathbb{N}_K : |R_k \cap A| \le \zeta_k$;
2. $A \in \mathcal{A}(\mathfrak{R})$;
3. $\forall R \subset \mathbb{N}_m, \quad |R \cap A| \le V^*_{\mathfrak{R}}(R)$;
4. $|A| \le V^*_{\mathfrak{R}}(A)$.

Furthermore, *if a function* $V : \mathcal{P}(\mathbb{N}_m) \to \mathbb{N}$ *satisfies that for any* $A \subset \mathbb{N}_m$, *point* 1 *implies point* 3 (*wherein* $V^*_{\mathfrak{R}}$ *is replaced by* V), *then for all* $R \subset \mathbb{N}_m$, *it holds* $V^*_{\mathfrak{R}}(R) \le V(R)$.

Note that point 1 of the above proposition is the complement of the event appearing in (5) ($(R_k, \zeta_k)_{k \in \mathbb{N}_K}$ is a reference family and A is taken equal to $\mathcal{H}_0$) while point 2 is the event appearing in (1) ($V^*_{\mathfrak{R}}$ is a post hoc bound). The last part of the proposition establishes the optimality of $V^*_{\mathfrak{R}}$ in this context.

An important problem is that $V^*(R)$ (we will sometimes drop the index $\mathfrak{R}$ for simplicity) can be hard to compute. In fact, the next proposition shows that it is, in full generality, an NP-hard problem.

PROPOSITION 2.2. *The problem of computing* $V^*_{\mathfrak{R}}(R)$ *given any arbitrary reference family* $\mathfrak{R} = (R_k, \zeta_k)_{1 \le k \le K}$ (*with* $R_k \subset \mathbb{N}_m$, $\zeta_k \in \mathbb{N}$), *and* $R \subset \mathbb{N}_m$, *is NP-hard*.

Naturally, Proposition 2.2 does not imply that computing the optimal bound $V^*(R)$ is always infeasible: depending on the choice of the reference family, we might be in a particular case where this can be done efficiently We will discuss precisely such a situation below when the regions are nested.

2.3. *A computable upper bound for V^* and its properties.* We introduce the following coarser but simpler bound:

$$\overline{V}_{\mathfrak{R}}(R) := \min_{k\in\mathbb{N}_K} \big(|R \setminus R_k| + \zeta_k\big) \wedge |R|, \quad R \subset \mathbb{N}_m. \tag{8}$$

Given the reference family and R, the bound $\overline{V}$ is computable in time $\mathcal{O}(mK)$. The next proposition is a counterpart of Proposition 2.1 for $\overline{V}$.

PROPOSITION 2.3. *Let $\mathfrak{R} = (R_k, \zeta_k)_{k\in\mathbb{N}_K}$ be a collection of subsets $R_k \subset \mathbb{N}_m$ and of integers ζ_k. Then for any $A \subset \mathbb{N}_m$, the following statements are equivalent*:

1. $\forall k \in \mathbb{N}_K : |R_k \cap A| \le \zeta_k$;
2. $\forall R \subset \mathbb{N}_m, \quad |R \cap A| \le \overline{V}_{\mathfrak{R}}(R)$;
3. $|A| \le \overline{V}_{\mathfrak{R}}(A)$.

For all $R \subset \mathbb{N}_m$, it holds $V^*_{\mathfrak{R}}(R) \le \overline{V}_{\mathfrak{R}}(R)$.

Observe that $\overline{V}(R)$ is also nondecreasing in the sense that $R \subset R'$ implies $\overline{V}(R) \le \overline{V}(R')$. We turn to studying further properties.

Self-consistency. Given some reference family $\mathfrak{R} = (R_k, \zeta_k)_{1\le k\le K}$, on the large probability event for which the control $|R_k \cap \mathcal{H}_0(P)| \le \zeta_k$, $1 \le k \le K$ holds, $\overline{V}_{\mathfrak{R}}$ provides a bound for $|R_k \cap \mathcal{H}_0(P)|$ itself, namely

$$\tilde{\zeta}_k := \overline{V}_{\mathfrak{R}}(R_k) = \min_{j\in\mathbb{N}_K} \big(|R_k \setminus R_j| + \zeta_j\big) \wedge |R_k|, \quad 1 \le k \le K. \tag{9}$$

Obviously, $\tilde{\zeta}_k \le \zeta_k$, with a possible strict inequality. Nevertheless, the next proposition shows that there is no advantage in "iterating" the post hoc bound $\overline{V}$ with ζ replaced by $\tilde{\zeta}$, thus showing a form of self-consistency of the bound $\overline{V}_{\mathfrak{R}}$.

PROPOSITION 2.4. *For any reference family $\mathfrak{R} = (R_k, \zeta_k)_{1\le k\le K}$, define $(\tilde{\zeta}_k)_{1\le k\le K}$ by* (9). *Denoting $\widetilde{\mathfrak{R}} = (R_k, \tilde{\zeta}_k)_{1\le k\le K}$, we have*

$$\overline{V}_{\mathfrak{R}}(R) = \min_{k\in\mathbb{N}_K} \big(|R \setminus R_k| + \tilde{\zeta}_k\big) \wedge |R| = \overline{V}_{\widetilde{\mathfrak{R}}}(R), \quad R \subset \mathbb{N}_m. \tag{10}$$

Optimality under nestedness assumption. In the situation where the sets $(R_k)_{1\le k\le K}$ are nested, it holds that $\overline{V} = V^*$, that is, the formula for $\overline{V}$ provides a computationally efficient way to compute the optimal bound in this case.

PROPOSITION 2.5. *For any reference family $\mathfrak{R} = (R_k, \zeta_k)_{1\le k\le K}$ such that $R_k \subset R_{k'}$ whenever $k \le k'$, we have $\overline{V}_{\mathfrak{R}}(R) = V^*_{\mathfrak{R}}(R)$.*

REMARK 2.6. The bound $\overline{V}(R)$ was introduced in Genovese and Wasserman (2006) in the particular case $K = 1$, $\zeta_1 = 0$, with a reference to the augmentation procedure of van der Laan, Dudoit and Pollard (2004). The latter builds a k-FWER controlling procedure by adding $k-1$ arbitrary hypotheses to the rejection set of a given FWER controlling procedure. Genovese and Wasserman (2006), noting that fixing any single value of k is suboptimal in terms of power, also put forward the principle of taking the minimum obtained for several

values of k via a union bound principle. See an extended discussion on this point in Section S-1.1. Also, since any one-element family is nested, Proposition 2.5 encompasses Theorem 5 of Genovese and Wasserman (2006) ($K = 1$, $\zeta_1 = 0$), extending it to the case of a whole nested reference family.

REMARK 2.7. The results of the paper can equivalently be stated in terms of false positives using V, V^* and $\overline{V}$ or in terms of true positives S, S^* and $\overline{S}$, where for any $R \in \mathbb{N}_m$ $S^*(R) := |R| - V^*(R)$ and $\overline{S}(R) := |R| - \overline{S}(R)$. For simplicity, we have chosen to focus on V.

2.4. *From general reference families to specific instances.*

Specific instances. We have developed post hoc bounds for reference families $\mathfrak{R}$ in a very general form. Specific cases can be considered, recovering in particular previous literature:

(A) $\zeta_k = k - 1$ for all k: in this case, each individual rejection region R_k has controlled k-FWER, and the control is uniform over the regions. In the standard case discussed in the Introduction where these regions are p-value level sets $R_k = \{i : p_i \le t_k\}$, nestedness holds, and thus the bound $\overline{V}$ given by (8) is optimal by Proposition 2.5.

(B) $\zeta_k = |R_k| - 1$ for all k: adopting a different point of view, let us associate to each $R \subset \mathbb{N}_m$ the *intersection hypothesis* $H_{0,R} := \bigcap_{i \in R} H_{0,i}$. In this view, each R corresponds to a hypothesis rather than a collection of hypotheses. The statement (4) is interpreted as saying that with high probability, each individual rejection region R_k contains at least one true rejection. Consequently, rejecting all intersection hypotheses H_{0,R_k}, $k = 1, \dots, K$ can be done without committing any error. This corresponds to an overall FWER control over this family of hypotheses.

From Section 3 onwards, we will focus on case (A) ($\zeta_k = k - 1$ and nestedness) and on how to obtain JER control then. In situation (B), JER control can in particular be obtained by defining a test for each local hypothesis $H_{0,R}$, thus recovering the setting of Genovese and Wasserman (2006), Goeman and Solari (2011); see Section S-1 for a more detailed discussion.

References families of different types can be considered and be useful in other situations as well. For instance, consider the setting where the reference regions R_k have little or no overlap to each other. In such cases, the bound $\overline{V}_{\mathfrak{R}}$ is a poor proxy for $V^*_{\mathfrak{R}}$ and other approximations should be considered, as for example,

$$\widetilde{V}_{\mathfrak{R}}(R) := \left(\sum_{k \in \{1,\dots,K\}} |R \cap R_k| \wedge \zeta_k + \left| R \setminus \bigcup_{k=1}^{K} R_k \right| \right) \wedge |R|, \quad R \subset \mathbb{N}_m. \tag{11}$$

It is not difficult to see that $\widetilde{V}_{\mathfrak{R}}(\cdot) = V^*_{\mathfrak{R}}(\cdot)$ when the reference sets R_k are disjoint. This setting is in particular useful if the signal is spatially structured; see Durand et al. (2018) for a detailed analysis of the JER approach in this case, and further corresponding developments.

Focus of the next sections. For the remainder of this paper, we focus on the common situation where a test statistic $T_i(X)$ is available for each null hypothesis $H_{0,i}$, which in turn is transformed into a p-value $p_i(X)$, for all $i \in \mathbb{N}_m$, and we choose a reference family by p-value thresholding:

$$R_k(X) = \{i \in \mathbb{N}_m : p_i(X) < t_k\}, \quad k \in \{1, \dots, K\}, \tag{12}$$

where the $t_k \in \mathbb{R}$, $1 \le k \le K$, are associated thresholds, possibly depending on X (K being deterministic). We easily check that the simpler expressions (2) and (3) announced in the Introduction hold in that context.

3. Model assumptions. Properties of the p-value process $(p_i(X), i \in \mathbb{N}_m)$ depend on the underlying model assumptions. In this paper, we distinguish between two general situations, depending on whether the dependence structure is known or not.

3.1. *Location model.* To give some intuition behind the general assumptions of the next section, we start by considering a specific location model

$$X_i = \mu_i + \varepsilon_i, \quad i \in \mathbb{N}_m, \tag{13}$$

where the ε_i are identically distributed with a common known marginal distribution which is assumed to be continuous, integrable and symmetric. We denote $\overline{F}(x) = \mathbb{P}(\varepsilon_1 \geq x)$, $x \in \mathbb{R}$. We consider the one-sided (resp., two-sided) testing problem with null hypotheses $H_{0,i}$: "$\mu_i \leq 0$" (resp., $H_{0,i}$:"$\mu_i = 0$") versus the alternative hypotheses $H_{1,i}$:"$\mu_i > 0$" (resp., $H_{1,i}$:"$\mu_i \neq 0$") for all $i \in \mathbb{N}_m$. Classical p-values are then given by $p_i(X) = \overline{F}(X_i)$ (resp., $p_i(X) = 2\overline{F}(|X_i|)$). As many procedures of multiple testing theory, our results will rely on the (joint) distribution of $(p_i(X))_{i \in \mathcal{H}_0(P)}$ or some approximation/bound of it.

Known dependence. In the case where the (joint) distribution of ε is known, we can consider "least favorable" p-values $q_i(X) = \overline{F}(X_i - \mu_i)$ ($q_i = 2\overline{F}(|X_i - \mu_i|)$). While the $q_i(X)$'s are not observed, they can be used purely as a technical device. Interestingly, these variables satisfy the following pointwise property: for all $i \in \mathcal{H}_0$, $p_i(X) \geq q_i(X)$, both in the one-sided and two-sided case. In addition, their joint distribution, that is, $\nu_m = \mathcal{D}((q_i(X))_{1 \leq i \leq m})$, is assumed to be known. For instance, under independence of the ε_i's, $\nu_m = U(0,1)^{\otimes m}$.

Unknown dependence. In the case where the (joint) distribution of ε is unknown, so is ν_m and the above least favorable p-values cannot be generated. In this situation, we focus on the two-sided situation, and assume that we have at hand n i.i.d. copies $(X_{i,j})_{i \in \mathbb{N}_m} \in \mathbb{R}^m$, $j \in \mathbb{N}_n$, where each $(X_{i,j})_{i \in \mathbb{N}_m}$ follows the location model (13). The p-values are assumed to be given by $p_i(X) = \overline{G}(|T(X_{i,j}, 1 \leq j \leq n)|)$, where $T(X_{i,j}, 1 \leq j \leq n)$ is some statistic, and the (known) function $\overline{G}$ is given by $\overline{G}(x) = \mathbb{P}(|T(\varepsilon_j, 1 \leq j \leq n)| \geq x)$, $x \geq 0$, for n i.i.d. copies ε_j, $1 \leq j \leq n$ of ε_1. Then, by a standard argument (see, e.g., Arlot, Blanchard and Roquain (2010)), the joint distribution of $(p_i(X))_{i \in \mathcal{H}_0(P)}$ can be approximated by random sign-flipping: let $\mathcal{G} = \{-1, 1\}^n$ denote the group of signs $s \in \{-1, 1\}^n$ that acts on the observed X in the following way:

$$(s.X)_{i,j} = s_j X_{i,j}, \quad i \in \mathbb{N}_m, j \in \mathbb{N}_n.$$

Then, if $i \in \mathcal{H}_0$, by symmetry, the distribution of $p_i(X)$ is equal to the one of $p_i(s.X)$, for some random sign s uniformly generated in $\mathcal{G}$. As a consequence, the distribution of $(p_i(s.X))_{i \in \mathcal{H}_0(P)}$ conditionally on X can act as proxy for the distribution of $(p_i(X))_{i \in \mathcal{H}_0(P)}$. This "randomization property" will be formalized in detail in the next section.

Both known and unknown situations can be met in the simple Gaussian location model for which $\varepsilon \sim \mathcal{N}(0, \Sigma)$ with some covariance matrix Σ (assuming $\Sigma_{i,i} = 1$ for $i \in \mathbb{N}_m$ for simplicity). On the one hand, the known dependence case corresponds to the case where Σ is known (with $\nu_m = \mathcal{N}(0, \Sigma)$). It can be met in practice in a standard Gaussian linear model or in marginal regression; see Fan, Han and Gu (2012). On the other hand, the unknown dependence case corresponds to the general situation where we have no information on Σ. A suitable statistic is then $T(X_{i,j}, 1 \leq j \leq n) = n^{-1/2} \sum_{j=1}^n X_{i,j}$, for which $\overline{G}(x) = 2\mathbb{P}(Z \geq x)$, $x \geq 0$, $Z \sim \mathcal{N}(0,1)$.

Also, mainly for illustrative purposes, we will use throughout the paper the ρ-equicorrelated covariance matrix for which $\Sigma_{i,j} = \rho$ for $1 \leq i \neq j \leq m$, for some $\rho \in [0,1]$ (either known or not).

3.2. *General framework and assumptions.* Now that we have a concrete example in mind, we go beyond the location model by presenting general assumptions on the p-value family $(p_i(X), i \in \mathcal{H}_0)$.

Known dependence. We assume that there exists a family of "least favorable" variables $(q_i(X))_{1\le i\le m}$ such that for all $P \in \mathcal{P}$,

$$\text{(LeastFavor)} \qquad \begin{cases} \forall i \in \mathcal{H}_0(P),\ p_i(X) \ge q_i(X) & P\text{-a.s.} \\ \nu_m = \mathcal{D}\big((q_i(X))_{1\le i\le m}\big) & \text{does not depend on } P. \end{cases}$$

While (LeastFavor) is satisfied in particular in the location model (with known dependence), it encompasses some other models (e.g., scaling model).

Unknown dependence. A classical way to adapt to unknown dependence in a multiple testing setting is to use resampling-based procedures, as introduced in Westfall and Young (1993) and reviewed in Dudoit and van der Laan (2008) for instance. However, establishing a rigorous nonasymptotic control is challenging and the seminal work of Romano and Wolf (2005) has paved the way for this by using randomization strategies. We follow this approach by assuming the existence of a finite transformation group $\mathcal{G}$ acting onto the observation set $\mathcal{X}$. Next, by denoting $p_{\mathcal{H}_0}(x)$ the null p-value vector $(p_i(x))_{i\in\mathcal{H}_0(P)}$ for $x \in \mathcal{X}$, we assume that the joint distribution of the transformed null p-values is invariant under the action of any $g \in \mathcal{G}$, that is,

$$\text{(Rand)} \qquad \forall P \in \mathcal{P}, \forall g \in \mathcal{G}, \quad \big(p_{\mathcal{H}_0}(g'.X)\big)_{g'\in\mathcal{G}} \sim \big(p_{\mathcal{H}_0}(g'.g.X)\big)_{g'\in\mathcal{G}},$$

where $g.X$ denotes X that has been transformed by g. This assumption has been introduced in Hemerik and Goeman (2018) and is slightly weaker than the so-called randomization hypothesis of Romano and Wolf (2005). It is easy to check that (Rand) is satisfied in the location model (with unknown dependence) for the above mentioned sign-flipping group $\mathcal{G} = \{-1, 1\}^n$, by using the symmetry of the noise. Assumption (Rand) is also met in permutation-based two-sample multiple testing problems, as described in Section S-5.

4. Methodology for adaptive JER control.

4.1. *Limitations of JER control based on Simes inequalities.* A particular form of JER control (2) may be obtained directly from the Simes inequality (Simes (1986)): denoting by $p_{(k:m)}$ the kth smallest p-value,

$$\mathbb{P}_{X\sim P}\left(\exists k \in \{1, \dots, m\} : p_{(k:m)} < \frac{\alpha k}{m}\right) \le \alpha, \tag{14}$$

provided that the p-value family is PRDS. A straightforward consequence is that (2) is satisfied for the choice $t_k = \alpha k/m, k \in \mathbb{N}_K$, for any choice of K. This is described in more detail in Section S-2. However, the corresponding reference family, called *Simes reference family* in the sequel, suffers from several limitations, which are briefly described in the next two paragraphs.

Sharpness and conservativeness. We carried out a simulation study in the Gaussian equicorrelated model where the one-sided test statistics follow the distribution $\mathcal{N}(0, \Sigma)$ with $\Sigma_{ii} = 1$ and $\Sigma_{ij} = \rho$ for $i \ne j$, for some $\rho \ge 0$. This p-value family is PRDS. We consider a white setting ($m_0 = m = 1000$). In Table 1, we quantify the conservativeness of JER control in this model as the ratio of the JER actually achieved (estimated from 1000 simulations) to the target JER level α (for $\alpha = 0.2$). While the JER actually achieved by the Simes reference family is α for $\rho = 0$ (a consequence of the sharpness of the Simes inequality under independence), it is less than $\alpha/2$ for $\rho = 0.4$.

TABLE 1
Conservativeness of JER control based on Simes inequality in the Gaussian equi-correlated model. Here, $m_0 = m = 1000$ *and* $\alpha = 0.2$

Equi-correlation level: ρ	0	0.1	0.2	0.4	0.8
Achieved JER $\times \alpha^{-1}$	1.00	0.89	0.73	0.46	0.39

Unbalancedness. Let us consider a "favorable" case for the Simes procedure, for which the p-values are i.i.d. uniform on $(0, 1)$. In this case, the Simes inequality (14) is an equality. However, we argue that the errors in the event described in (14) are *not balanced* w.r.t. the parameter k. As an illustration, $\mathbb{P}(p_{(1:m)} < \alpha/m) = 1 - (1 - \frac{\alpha}{m})^m = \alpha + o(\alpha)$, hence the probability of the event in (14) is already almost exhausted for $k = 1$. More generally, some values of the function $k \mapsto \mathbb{P}(p_{(k:m)} < \alpha k/m)$ are given in Table 2 for $m = 1000$, where $p_{(k:m)} \sim \text{Beta}(k, m + 1 - k)$. As a consequence, the Simes family seems to favor some of the k's when controlling the JER. In addition, the structure of this unbalancedness is somewhat arbitrary, and imposed to the user of the procedure, which may be undesirable. This phenomenon is quantified more formally in Section S-3.3; see (S-14).

In order to address these limitations, we aim in the rest of Section 4 at building a thresholding-based reference family $\mathfrak{R}$ for which the quantity $\text{JER}(\mathfrak{R}, P)$ is as close as possible to α, for a wide spectrum of distributions P. To this end, we combine two approaches:

- incorporating the dependence structure of the noise (either known or unknown);
- using a step-down algorithm to adapt to the unknown set $\mathcal{H}_0$.

4.2. *Threshold templates*. We start with considering a reference family $\mathfrak{R}_\lambda$ of the form (12), parametrized by $\lambda \in [0, 1]$ and itself based on a parametrized family of thresholds $t_k(\lambda)$ which we call *template*. The second step will be to to choose $\lambda = \lambda(\alpha)$ so that the JER control (4) is satisfied, which we call λ-*calibration.*

DEFINITION 4.1. A *one-parameter threshold template* (simply referred to as *template* in the sequel for short) is a family of functions $t_k(\lambda)$, $\lambda \in [0, 1]$, $1 \le k \le K$, such that $K \in \{1, \dots, m\}$ and for all $k \in \{1, \dots, K\}$, $t_k(0) = 0$ and $t_k(\cdot)$ is nondecreasing and left continuous on $[0, 1]$. The parameter K is called the *size* of the template.

In general, a template is allowed to depend on the observation X. For a given template and fixed λ, we refer to $t_k(\lambda)$, $1 \le k \le K$, as thresholds and denote by $\mathfrak{R}_\lambda$ the associated reference family given by (12). Several choices of template are possible as we will see in Section 5. Here, we work with a generic, fixed template $t_k(\lambda)$, $\lambda \in [0, 1]$, $1 \le k \le K$. We denote the generalized inverse of $t_k(\cdot)$ by $t_k^{-1}(y) = \max\{x \in [0, 1] : t_k(x) \le y\}$, for any $y \in \mathbb{R} \cup \{-\infty, +\infty\}$.

TABLE 2
$\mathbb{P}(p_{(k:m)} < \alpha k/m)$ *when* $p_{(k:m)} \sim \text{Beta}(k, m + 1 - k)$, $m = 1000$ *and* $\alpha = 0.05$

k	1	2	5	10	100
$\mathbb{P}(p_{(k:m)} \le \alpha k/m)$	4.9×10^{-2}	4.7×10^{-3}	6.6×10^{-6}	1.6×10^{-10}	5.8×10^{-93}

Since $t_k(\cdot)$ is monotonic, for any p-value family $\{p_i, i \in \mathbb{N}_m\}$, we have $t_k(\lambda) > p_{(k:\mathcal{H}_0)}$ if and only if $\lambda > t_k^{-1}(p_{(k:\mathcal{H}_0)})$. Hence, in view of (2), we obtain

$$\begin{aligned}\mathrm{JER}(\mathfrak{R}_\lambda, P) &= \mathbb{P}_{X\sim P}\big(\exists k \in \{1, \dots, K \wedge m_0\} : p_{(k:\mathcal{H}_0)} < t_k(\lambda)\big)\\ &= \mathbb{P}_{X\sim P}\big(\exists k \in \{1, \dots, K \wedge m_0\} : t_k^{-1}(p_{(k:\mathcal{H}_0)}) < \lambda\big).\end{aligned}$$

This proves the following result.

LEMMA 4.2. *Consider a general p-value model and any (possibly data-dependent) template $t_k(\lambda)$, $\lambda \in [0,1]$, $1 \le k \le K$. Then, for any $\lambda \in [0,1]$, the error rate (5) of the reference family $\mathfrak{R}_\lambda$ given by (12) can be written as follows: for any $P \in \mathcal{P}$,*

$$\mathrm{JER}(\mathfrak{R}_\lambda, P) = \mathbb{P}_{X\sim P}\Big(\min_{1\le k\le K\wedge m_0}\big\{t_k^{-1}\big(p_{(k:\mathcal{H}_0)}(X)\big)\big\} < \lambda\Big). \tag{15}$$

4.3. *Single-step and step-down procedures by λ-calibration.* The JER control (4) can now be achieved by choosing λ in an appropriate way.

DEFINITION 4.3. Given a threshold template $t_k(\lambda)$, $\lambda \in [0,1]$, $1 \le k \le K$, a (possibly data-dependent) functional $\lambda(\alpha, A)$, $\alpha \in (0,1)$, $A \subset \mathbb{N}_m$, is called a λ-calibration if it is nonincreasing in A, that is,

$$\forall \alpha \in (0,1), \forall A, A' \subset \{1, \dots, m\} \quad \text{with } A \subset A', \lambda\big(\alpha, A'\big) \le \lambda(\alpha, A), \tag{16}$$

and satisfies $\forall \alpha \in (0,1)$, $\forall P \in \mathcal{P}$,

$$\mathbb{P}_{X\sim P}\Big(\min_{1\le k\le K\wedge m_0}\big\{t_k^{-1}\big(p_{(k:\mathcal{H}_0(P))}(X)\big)\big\} < \lambda\big(\alpha, \mathcal{H}_0(P)\big)\Big) \le \alpha. \tag{17}$$

Two examples of possible λ-calibrations will be provided in Sections 4.4 and 4.5. In the remainder of Section 4.2, we consider that some λ-calibration is given.

The dependence of the calibration on the set A adds extra flexibility which will allow us to apply a step-down principle and get a more accurate procedure. A consequence of Lemma 4.2 is that the procedure $\mathfrak{R}_{\lambda(\alpha,\mathcal{H}_0)}$ has a controlled JER (given a template and a calibration), in other words taking $A = \mathcal{H}_0$ provides an "oracle" calibration, but since $\mathcal{H}_0$ is unknown, $\lambda(\alpha, \mathcal{H}_0)$ cannot be used. However, a consequence of (16) is that $\lambda(\alpha, \mathbb{N}_m) \le \lambda(\alpha, \mathcal{H}_0)$, so that $\lambda(\alpha, \mathbb{N}_m)$ can be used as a (single-step) conservative substitute for $\lambda(\alpha, \mathcal{H}_0)$. This provides the following result.

PROPOSITION 4.4. *In the framework of Lemma 4.2, consider $\lambda(\alpha) = \lambda(\alpha, \mathbb{N}_m)$ for some λ-calibration as in Definition 4.3. Then the procedure $\mathfrak{R}_{\lambda(\alpha)}$ controls the JER criterion at level α in the sense of (4).*

Above, the fact that $\lambda(\alpha, \mathbb{N}_m)$ is smaller than $\lambda(\alpha, \mathcal{H}_0)$ induces a loss in the JER control. This loss can sometimes be substantial, as illustrated with numerical experiments in Section 6; this effect is further studied theoretically in Section S-3.2. This loss can be reduced by using $\lambda(\alpha, \widehat{A})$, where $\widehat{A}$ is the output of the the following step-down algorithm.

While the update of $A^{(j)}$ only depends on $t_1(\cdot)$ in Algorithm 1, $\widehat{A}$ may depend on all the t_k's through the functional $\lambda(\alpha, \cdot)$. The following result is proved in Section S-7.2.

PROPOSITION 4.5. *In the framework of Lemma 4.2, consider any λ-calibration as in Definition 4.3 and compute $\widehat{A}$ by Algorithm 1. Then the procedure $\mathfrak{R}_{\lambda(\alpha,\widehat{A})}$ controls the JER at level α in the sense of (4).*

REMARK 4.6. When we choose $K = 1$, Algorithm 1 reduces to the usual FWER controlling step-down algorithm (see, e.g., Romano and Wolf (2005)).

Algorithm 1: General step-down algorithm

$j \leftarrow 0$;
$A^{(0)} \leftarrow \mathbb{N}_m$;
repeat
  $j \leftarrow j + 1$;
  $\lambda_j \leftarrow \lambda(\alpha, A^{(j-1)})$;
  $A^{(j)} \leftarrow \{i \in \mathbb{N}_m \,:\, p_i(X) \geq t_1(\lambda_j)\}$;
until $A^{(j)} = A^{(j-1)}$;
return $A^{(j)}$;

4.4. *Valid λ-calibration for known dependence.* Let us focus on the situation where the dependence is known; see Section 3.2. The template is assumed to be deterministic in this section. Assumption (LeastFavor) (with ν_m defined therein) and Lemma 4.2 thus give

$$\mathrm{JER}(\mathfrak{R}_\lambda, P) \leq \mathbb{P}_{q\sim\nu_m}\Big(\min_{1\leq k\leq K\wedge m_0}\{t_k^{-1}(q_{(k:\mathcal{H}_0)})\} < \lambda\Big), \tag{18}$$

which provides the following valid λ-calibration: for all $A \subset \{1, \ldots, m\}$,

$$\lambda(\alpha, A) = \max\Big\{\lambda \geq 0 : \mathbb{P}_{q\sim\nu_m}\Big(\min_{1\leq k\leq K\wedge|A|}\{t_k^{-1}(q_{(k:A)})\} < \lambda\Big) \leq \alpha\Big\}. \tag{19}$$

Property (16) can be easily checked. Note that $\lambda(\alpha, \cdot)$ depends on ν_m and on the template, although it is not explicit from the notation for short. We have proved the following result.

THEOREM 4.7 (λ-calibration for known dependence). *Consider any p-value family satisfying* (LeastFavor), *a deterministic template and the associated reference family $\mathfrak{R}_\lambda$. Then the* (*deterministic*) *functional $\lambda(\cdot, \cdot)$ defined by* (19) *is a λ-calibration in the sense of Definition* 4.3, *and thus $\mathfrak{R}_{\lambda(\alpha,\mathbb{N}_m)}$ and $\mathfrak{R}_{\lambda(\alpha,\widehat{A})}$ both control the JER at level α.*

4.5. *Valid λ-calibration for unknown dependence.* Let us consider now the case where the dependence is unknown; see Section 3.2. The template is still assumed to be deterministic in this section. We use the notation defined therein and in particular assumption (Rand). Let us consider a (random) B-tuple $(g_1, g_2, \ldots, g_B)$ of $\mathcal{G}$ (for some $B \geq 2$), where g_1 is the identity element of $\mathcal{G}$ and $g_2, \ldots, g_B$ have been drawn (independently of the other variables) as i.i.d. variables, each being uniformly distributed on $\mathcal{G}$.

Let us consider some template $t_k(\cdot)$, $1 \leq k \leq K$, and, for short, denote for all $A \subset \mathbb{N}_m$, $\Psi(X, A) = \min_{1\leq k\leq K\wedge|A|}\{t_k^{-1}(p_{(k:A)}(X))\}$. Now introduce the (data-dependent) λ-calibration

$$\lambda(\alpha, A) = \max\left\{\lambda \geq 0 : B^{-1}\sum_{j=1}^{B}\mathbb{1}\{\Psi(g_j.X, A) < \lambda\} \leq \alpha\right\}. \tag{20}$$

In practice, we can compute this functional easily as $\lambda(\alpha, A) = \Psi_{(\lfloor\alpha B\rfloor+1)}$ where $\Psi_{(1)} \leq \Psi_{(2)} \leq \cdots \leq \Psi_{(B)}$ denote the ordered sample $(\Psi(g_j.X, A), 1 \leq j \leq B)$. Then the following result holds and is proved in Section S-7.2.

THEOREM 4.8 (λ-calibration for unknown dependence). *Consider any p-value family satisfying* (Rand), *a deterministic template and the associated reference family $\mathfrak{R}_\lambda$. Then the* (*data-dependent*) *functional $\lambda(\cdot, \cdot)$ defined by* (20) *is a λ-calibration in the sense of Definition* 4.3 *and $\mathfrak{R}_{\lambda(\alpha,\mathbb{N}_m)}$ and $\mathfrak{R}_{\lambda(\alpha,\widehat{A})}$ both control the JER at level α.*

A related idea has been proposed independently by Hemerik, Solari and Goeman (2019) to build confidence envelopes for the false discovery proportion.

5. Application: Two examples of template-based reference families. In this section, we apply the methodology presented in the previous section for two particular instances of templates. Throughout this section, the λ-calibration functional $\lambda(\alpha, A)$ is either given by (19) (known dependence) or by (20) (unknown dependence).

5.1. *Linear template.* Motivated by the Simes inequality (see (14)), we define the *linear template* (of size K) by

$$t_k^L(\lambda) = \lambda k/m, \quad \lambda \in [0,1], 1 \le k \le K. \tag{21}$$

Hence we have $(t_k^L)^{-1}(u) = 1 \wedge (\frac{m}{k}u)$ which corresponds to a specific λ-calibration denoted by $\lambda^L(\alpha, A)$. For each K, this gives rise to two new reference families:

- The *single-step linear reference family* (of size K), denoted $\mathfrak{R}^L$, is given by $\mathfrak{R}^L = (R_1^L(X), \ldots, R_K^L(X))$, where

$$R_k^L(X) = \left\{ i \in \mathbb{N}_m : p_i < \lambda^L(\alpha, \mathbb{N}_m)\frac{k}{m} \right\}, \quad 1 \le k \le K. \tag{22}$$

- The *step-down linear reference family* (of size K), denoted $\mathfrak{R}^{L,\mathrm{sd}}$, is given by $\mathfrak{R}^{L,\mathrm{sd}} = (R_1^{L,\mathrm{sd}}(X), \ldots, R_K^{L,\mathrm{sd}}(X))$, where

$$R_k^{L,\mathrm{sd}}(X) = \left\{ i \in \mathbb{N}_m : p_i < \lambda^L(\alpha, \hat{A})\frac{k}{m} \right\}, \quad 1 \le k \le K, \tag{23}$$

  where $\hat{A}$ is derived from Algorithm 1, used with $\lambda(\cdot) = \lambda^L(\cdot)$ and $t_1(\cdot) = t_1^L(\cdot)$.

Theorems 4.7 and 4.8 ensure that the reference families $\mathfrak{R}^L$ and $\mathfrak{R}^{L,\mathrm{sd}}$ control the JER at level α both in the known and unknown dependence case.

There exists also distribution free calibrations of the type $\lambda(\alpha) = \alpha/c_m$ that are valid under arbitrary dependence. First, the Hommel bound corresponding to $c_m = \sum_{i=1}^m 1/i$ (see Section S-2.1). Second, a union bound argument can be used to give $c_m = K$, as suggested in Genovese and Wasserman (2006) (see the sentence before equation (24) therein). By contrast, the advantage of our proposed λ-calibrations is their adaptivity to the dependence structure. The magnitude of $\lambda^L(\alpha, \mathbb{N}_m)$ is studied numerically in Section S-3.1 in the case of known dependence, while the numerical experiments in Section 6 illustrate the sharpness of the associated JER control.

5.2. *Balanced template.* Considering a linear template is not always appropriate: as mentioned in Section 4.1, under independence and $K = m$, the Simes reference family suffers from a kind of unbalancedness. Ideally, a *balanced* reference family R_k would have the property that $\mathbb{P}(|R_k \cap \mathcal{H}_0| \ge k)$ is a constant not depending on $k = 1, \ldots, K$. While strict balancedness seems out of reach, since these probabilities depend on $\mathcal{H}_0$, we can ensure balancedness under the full null configuration ($\mathbb{N}_m = \mathcal{H}_0$) by calibrating the template as a quantile at a common level for all k, as follows. For each $k \in \mathbb{N}_m$, let us define

$$\begin{cases} F_k(x) = \mathbb{P}_{q\sim\nu_m}(q_{(k:m)} \le x) & \text{(known dep.)}, \\ F_k(x) = B^{-1}\displaystyle\sum_{j=1}^{B} \mathbb{1}\{p_{(k:m)}(g_j.X) \le x\} & \text{(unknown dep.)}, \end{cases} \quad x \in [0,1].$$

The *balanced template* (of size K) is then given by

$$t_k^B(\lambda) = F_k^{-1}(\lambda) = \min\{x \in [0,1] : F_k(x) \geq \lambda\} \quad \text{with } k \in \{1,\dots,K\}. \tag{24}$$

From an intuitive point of view, for each k, the threshold $t_k^B(\lambda)$ corresponds to a procedure controlling the k-FWER at level λ. It is straightforward to check that $t_k^B(\cdot)$ fulfills the requirements of Definition 4.1 while $(t_k^B)^{-1}(x) = F_k(x)$ for all $x \in [0,1]$. This corresponds to a specific λ-calibration denoted by $\lambda^B(\alpha, A)$. For each K, this gives rise to two new reference families:

- The *single-step balanced reference family* (of size K), denoted $\mathfrak{R}^B$, is given by $\mathfrak{R}^B = (R_1^B(X),\dots,R_K^B(X))$, where

$$R_k^B(X) = \{i \in \mathbb{N}_m : p_i < t_k^B(\lambda^B(\alpha, \mathbb{N}_m))\}, \quad 1 \leq k \leq K. \tag{25}$$

- The *step-down balanced reference family* (of size K), denoted $\mathfrak{R}^{B,\mathrm{sd}}$, is given by $\mathfrak{R}^{B,\mathrm{sd}} = (R_1^{B,\mathrm{sd}}(X),\dots,R_K^{B,\mathrm{sd}}(X))$, where

$$R_k^{B,\mathrm{sd}}(X) = \{i \in \mathbb{N}_m : p_i < t_k^B(\lambda^B(\alpha, \hat{A}))\}, \quad 1 \leq k \leq K, \tag{26}$$

  where $\hat{A}$ is derived from Algorithm 1, used with $\lambda(\cdot) = \lambda^B(\cdot)$ and $t_1(\cdot) = t_1^B(\cdot)$.

We give in section Section S-6 a detailed construction of the reference families $\mathfrak{R}^B$ and $\mathfrak{R}^{B,\mathrm{sd}}$. Theorem 4.7 ensures that both of these reference families control the JER at level α in the case of a known dependence.

However, for unknown dependence, Theorem 4.8 cannot be directly applied to the balanced template. Indeed, although this is not acknowledged by the notation for simplicity, F_k, and thus $t_k^B(\lambda)$ depends on the observation X. Our proof does not generalize easily to such a data-dependent rejection template, although the numerical experiments of Section 6 suggest that the JER control is also valid in that situation.

REMARK 5.1. The step-down refinement can be substantial for a balanced template, as illustrated in the numerical experiments of Section 6, and further discussed in Section §-3.2.

REMARK 5.2. Under independence, the balanced template $t_k^B(\cdot)$ corresponds to using quantiles of a Beta distribution, which was proposed in Genovese and Wasserman (2006). However, these authors address uniformity with respect to k through a union bound argument, which corresponds to divide the confidence level by the family cardinal K, while our λ-calibration method divides the level by a factor at most $(\log m)^{1/4}$; see Lemma S-3.1.

REMARK 5.3. By considering the two-sample setting with unknown dependency structure (see Section S-5) our balanced procedure is related to the work of Meinshausen (2006), where permutations are used to build FDP confidence envelopes. However, there appears to be a gap in the theoretical analysis justifying the validity of such an approach (Theorem 1 of Meinshausen (2006), more specifically equation (12) therein), which seems to have been overlooked so far. The reason is similar to the one making our proof not cover the case of a data-dependent template $t_k(X,\lambda)$: the fact that for all λ and $g \in \mathcal{G}$, $(t_k(g.X,\lambda))_{1\leq k\leq K} = (t_k(X,\lambda))_{1\leq k\leq K}$ and $(p_i(g.X))_{i\in\mathcal{H}_0} \sim (p_i(X))_{i\in\mathcal{H}_0}$, does not imply (in general) equality of the joint distributions $((t_k(X,\lambda))_{1\leq k\leq K}, (p_i(X))_{i\in\mathcal{H}_0})$ and $((t_k(g.X,\lambda))_{1\leq k\leq K}, (p_i(g.X))_{i\in\mathcal{H}_0})$.

6. Numerical experiments. We report numerical experiments performed in the two-sided location model (13) described in Section 3.1 in the case of an *unknown dependence*. The observations $(X_{i,j})_{i\in\mathbb{N}_m} \in \mathbb{R}^m$, $j \in \mathbb{N}_n$ are distributed as ρ-equi-correlated, and the test statistics for $i \in \mathbb{N}_m$ is $T(X_{i,j}, 1 \le j \le n) = n^{-1/2}\sum_{j=1}^n X_{i,j}$. We use sign-flipping (as described in that section) to approximate the joint distribution of the test statistics under the null. The location parameter is set to $\mu_i = n^{-1/2}\overline{\mu}\mathbb{1}\{i \in \mathcal{H}_1\}$, where $\overline{\mu} > 0$ quantifies the signal-to-noise ratio (SNR). We have also performed experiments in the same model but assuming *known dependence*, in order to illustrate Theorem 4.7. The results of these experiments are quite similar to those reported here for unknown dependence.

6.1. *JER control.* The target JER level is set to $\alpha = 0.25$, and the simulation parameters are: $m = n = 1000$, $\rho \in \{0, 0.2, 0.4\}$, $\pi_0 \in \{0.8, 0.9, 0.99\}$ (corresponding to $m_1 \in \{200, 100, 10\}$) and $\overline{\mu} \in \{0, 1, 2, 3, 4, 5\}$. For each setting, we report the empirical JER achieved, that is, the proportion of simulation runs (out of a total of 10,000 runs) for which $|R_k(X) \cap \mathcal{H}_0(P)| > k$ for at least one $k \in \{1, \dots, K\}$. The results are summarized by Figure 4 for the linear template, and by Figure 5 for the balanced template. Each figure is a matrix of panels, where each row corresponds to one value of the sparsity parameter π_0, and each column corresponds to one value of the equi-correlation parameter ρ. In each panel, the empirical JER achieved by several procedures is displayed as a function of the signal-to-noise ratio parameter $\overline{\mu}$. The target JER level α is represented by a horizontal dashed line, and for the linear template, the level $\pi_0\alpha$ is represented by a horizontal dotted line. In both figures, each color corresponds to a different λ-calibration:

single-step	Step down	Oracle
$\lambda(\alpha, \mathbb{N}_m)$	$\lambda(\alpha, \widehat{A})$	$\lambda(\alpha, \mathcal{H}_0)$

Additionally, for the linear template, "Simes" corresponds to $\lambda = \alpha$ (no λ-calibration). Figure 4 illustrates that the JER is controlled at the target level α in all situations for the linear template, which is expected according to Proposition 4.8. Oracle calibration yields exact JER control, up to sampling fluctuations. As discussed in Section 4.1, the Simes reference family with parameter α yields JER equal to $\pi_0\alpha$ under independence ($\rho = 0$), while it is more conservative under positive dependence $\rho > 0$. Single-step λ-calibration addresses

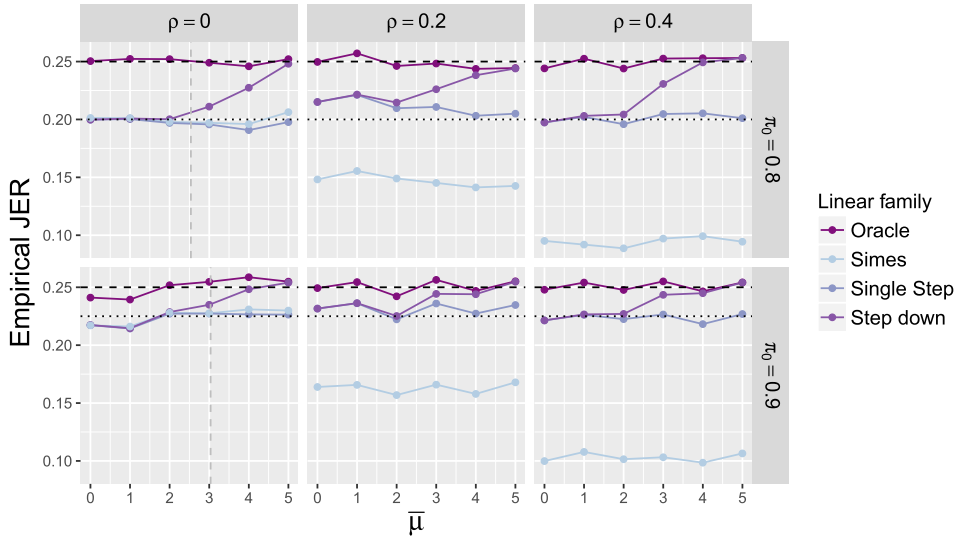

FIG. 4. *JER control based on the linear template for equi-correlated test statistics.*

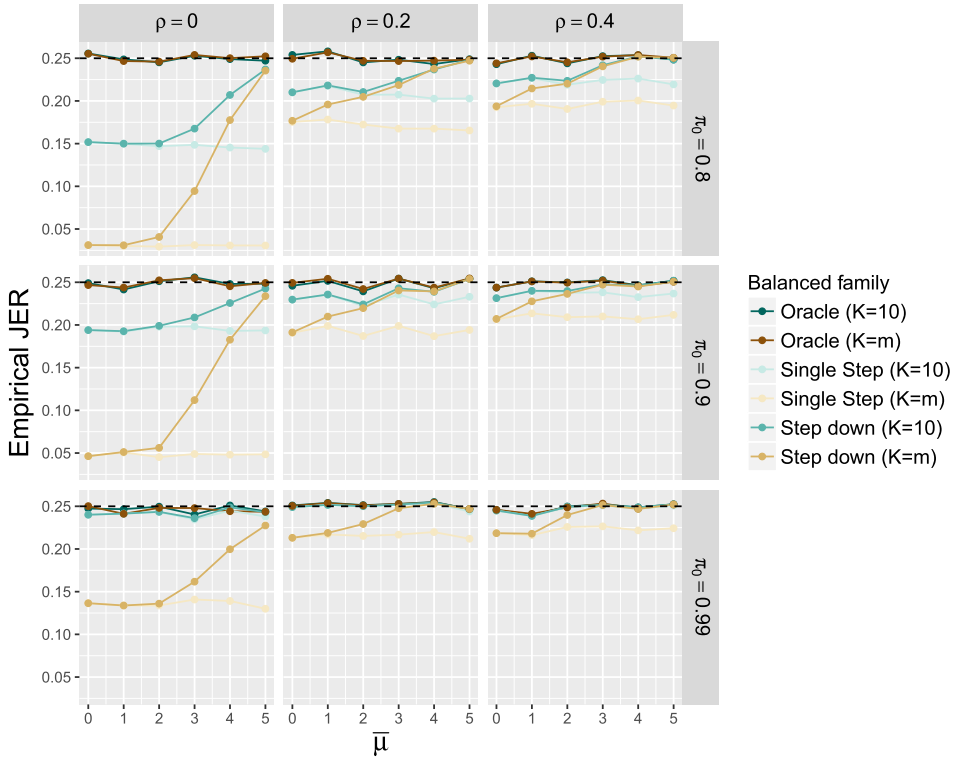

FIG. 5. *JER control based on the balanced template for equi-correlated test statistics, with* $K = m$ *and* $K = 10$.

this conservativeness by adapting to the (unknown) dependence: it yields JER control at $\pi_0\alpha$ in all settings considered. Finally, as the signal-to-noise ratio $\overline{\mu}$ gets larger, the step-down λ-calibration yields a JER closer to the nominal level α in nonsparse situations ($\pi_0 \in \{0.8, 0.9\}$). In a sparse situation ($\pi_0 = 0.99$), corresponding to $m_1 = 10$, the single-step procedure is already quite sharp and essentially indistinguishable from its Oracle counterpart, so we decided to omit this setting from Figure 4.

The results for the balanced template are summarized by Figure 5. First, the JER is empirically controlled at the target level α in all situations. This is worth noting because as discussed in the preceding section, our results do not cover the case of unknown dependence for the balanced template. Looking at the (brown) curves corresponding to $K = m$, single-step λ-calibration leads to a much more conservative JER control than for the linear template, especially under independence or for small values of ρ, even when π_0 is close to one. For example, when $\pi_0 = 0.99$ ($m_1 = 10$ out of $m = 1000$), the JER achieved by the single-step λ-calibration of the balanced family is of the order of $\alpha/2 (\ll \pi_0\alpha)$. When the signal-to-noise ratio is large, our proposed step-down adjustment catches up with the target JER level. This effect is further discussed and formalized in Section S-3.2.

Interestingly, the JER control offered by the balanced family with $K = 10$ (green curves in Figure 5) is much less conservative than with $K = m$, even for the single-step λ-calibration. The magnitude of the λ-adjustment is further discussed in Section S-3.1, and the question of how to choose K is discussed in Section 7.1.

Additional numerical experiments. The experiments reported here are carried out only in the equi-correlated setting and assuming that the mean signal under the alternative is constant: $\mu_i = \overline{\mu}$ for all $i \in \mathcal{H}_1$. We have performed other experiments, where μ_i is uniformly distributed between 0 and $\overline{\mu}$, and/or where the test statistics have a Toeplitz covariance, for which $\Sigma_{i,j} = |i - j|^\theta$, where $\theta \in \{-2, -1, -0.5, -0.2\}$ controls the range of dependency.

The results obtained for both types of signals and for both types of dependency are qualitatively similar, so we have only reported the results for the parameter combination: constant signal/equi-correlated dependency.

6.2. *Power.* In the preceding section, the quality of a JER controlling procedure is quantified by the tightness of its JER control. We now compare some JER controlling procedures in terms of power. This comparison is made under independence for simplicity. We focus on the step-down linear reference family (23) with $K = m$, and the step-down balanced reference family (26) with $K \in \{10, 2m_1, m\}$. We consider a notion of power, referred to as "averaged power," that takes into account the amplitude of the lower bound $\overline{S}_{\mathfrak{R}}(\cdot)$. Let us define for some selected set $R \subset \mathbb{N}_m$ (possibly data dependent),

$$\mathrm{Pow}(\mathfrak{R}, P, R) = \mathbb{E}\Big(\frac{\overline{S}_{\mathfrak{R}}(R)}{|R \cap \mathcal{H}_1(P)|} \Big| |R \cap \mathcal{H}_1(P)| > 0 \Big), \tag{27}$$

where we recall that $\overline{S}_{\mathfrak{R}}(R) = |R| - \overline{V}_{\mathfrak{R}}(R)$. The following selected sets $R \subset \mathbb{N}_m$ are considered:

(a) $R = \mathbb{N}_m$. In this case, the averaged power $\mathrm{Pow}(\mathfrak{R}, P, R)$ measures the (relative) performance of $\overline{S}_{\mathfrak{R}}(\mathbb{N}_m)$ as an estimator of $m_1(P) = |\mathcal{H}_1(P)|$;

(b) $R_0 = \{i \in \mathbb{N}_m : p_i \leq 0.05\}$, and R is a random selection of half of the items of R_0. Each hypothesis is given a selection probability proportional to the rank of its p-value;

(c) Same as (b) with R_0 corresponding to the rejections of the BH procedure at level 0.05.

In (b)–(c) above, the sets R are thought to be typical possible choices for the user. We chose to give nonuniform selection probabilities in order to favor sets enriched in lower p-values. The parameter π_0 is taken in the range $\pi_0 \in \{0.8, 0.9, 0.99\}$. We set $\overline{\mu} = \sqrt{-4\log(1-\pi_0)}$ in order to specifically focus on situations where the signal strength lies just above the estimation boundary, which would correspond to $\overline{\mu} = \sqrt{-2\log(1-\pi_0)}$; see Donoho and Jin (2004).

The results are displayed in Figure 6. The average power of the Simes family (light green) and of the reference families obtained by single-step and step-down λ-calibration of the linear template (dark green) are almost identical. This is consistent with the results displayed in the first column of Figure 4, where the three families achieve very similar JER levels for $\overline{\mu} \leq \sqrt{-4\log(1-\pi_0)}$; this value of $\overline{\mu}$ is shown by a dashed gray vertical line. Overall, the averaged power obtained from the balanced template is substantially larger than the averaged power obtained from the linear template. While neither template uniformly dominates the other one, the only situation where the linear template is more powerful is under the most sparse scenario ($\pi_0 = 0.99$), for the two user-defined rejection sets (b) and (c). In particular, the first row of panels in Figure 6 indicates that, except for a very low target JER ($\alpha \leq 0.02$), the bound $\overline{S}_{\mathfrak{R}}(\mathbb{N}_m)$ obtained from the balanced template provides a better estimator of $m_1(P) = |\mathcal{H}_1(P)|$ than the linear template. These experiments also show that, as expected, the choice of K can improve the performance of the balanced procedure. Some suggestions for choosing K are discussed below.

7. Discussion.

7.1. *Choosing the size K.* While the choice $K = m$ seems a priori natural, we have shown throughout this paper that it induces some conservativeness (via the λ-calibration): choosing a smaller value for K can yield a tighter post hoc bound. This effect is particularly marked in the case of the balanced template when p-values are close to independent (see Figure 5). The choice of K is therefore quite important in practice. We underline the following plausible scenarios:

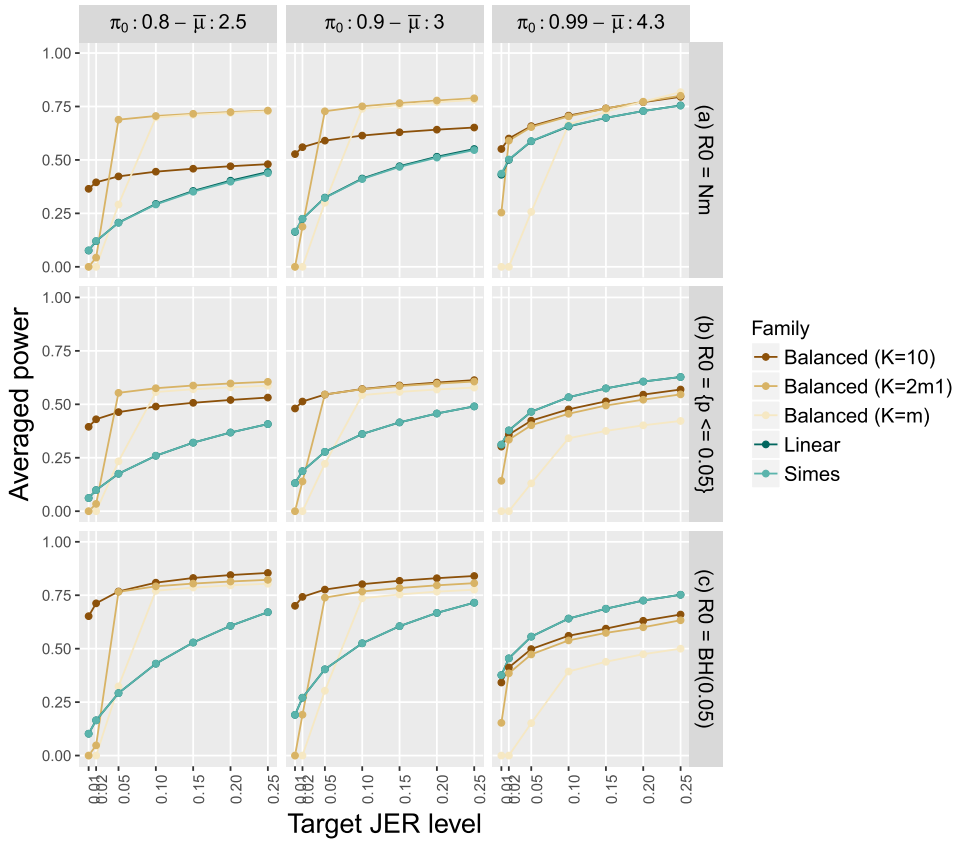

FIG. 6. *Averaged power of JER controlling procedures for independent test statistics.*

- if the user has an *a priori maximum amount of tolerated false discoveries*, then K can be set taken equal to that value. This comes from the following fact: *let* $K_0 \in \mathbb{N}$ *and assume* $\mathfrak{R} = (R_i(X))_{1\le i\le K}$ *is a reference family (using* $\zeta_i = i-1$*) satisfying JER control. Consider any set* $R \subset \mathbb{N}_m$ *such that* $\overline{V}_{\mathfrak{R}}(R) \le K_0 < K$*. Then we have* $\overline{V}_{\mathfrak{R}}(R) = \overline{V}_{\mathfrak{R}^{(K_0)}}(R)$*, where* $\mathfrak{R}^{(K_0)} = (R_i(X))_{1\le i\le K_0+1}$. In words, if the user is only interested in rejected sets R where the bound on the number of false positives is less than K_0, then the family size K can safely be taken equal to K_0+1.
- if the user has some upper bound $\overline{m}_1$ on the number of false hypotheses as prior information, it seems reasonable to take $K_0 = \overline{m}_1$ above (a larger number of false discoveries would mean that more than 50% of the hypotheses in the rejected set are false discoveries). The case $K = 2m_1$ considered in our numerical experiments can be interpreted as such a scenario (assuming a known prior rough upper bound $\overline{m}_1 = 2m_1$).

Designing a theoretically founded data-dependent choice of K is an interesting direction for future efforts. Let us also mention that an alternative to the choice of K is to introduce some smooth decay in the violation probability $\mathbb{P}(|R_k| \ge k)$ as k grows.

7.2. *Step-down algorithm.* The principle of the step-down Algorithm 1 is to approach the oracle value $\lambda(\alpha, \mathcal{H}_0)$ by iterative approximations $\lambda(\alpha, \widehat{A})$. Here, the template $t_k(\cdot)$ is fixed once for all. A seemingly natural extension is to allow the template $t_k(\cdot, A)$ to also depend on subsets $A \subset \mathbb{N}_m$ and to apply the step-down algorithm to the template as well as λ, that is, consider at each step $t_k(\cdot, \widehat{A})$, then apply the λ-calibration step. For instance, for the balanced rejection template, one could define $t_k^B(\lambda, A)$ as the λ-quantile of $q_{k:A}$. From a theoretical

point of view, however, it turns out that the corresponding combined threshold (depending on $\mathcal{H}_0$ both through t_k and λ) loses the monotonicity property with respect to $\mathcal{H}_0$. Hence, our current proof does not extend to that situation and we do not know if the corresponding JER is controlled at level α. This is an interesting (but challenging) issue.

7.3. *Choice of the reference family.* In the general setting presented in Section 2, although the aim is to obtain a uniform guarantee for any possible rejected set, a tradeoff is implicitly present in the choice of the reference family. The post hoc bounds (7), (8) can be understood as interpolation bounds relating an arbitrary R to sets of the reference family $\mathfrak{R}$, so that generally speaking they will be more accurate for rejection sets that are "well approximated" by sets of the reference family. From the definition of the JER control (4), it is clear that there is a tradeoff between the cardinality of the reference family and the conservativeness of the bound, which requires a uniform control over the family. Depending on the specific application, reference families corresponding to different expected tradeoffs can be considered. In the running example considered in this paper, the choice of K (discussed above) represents precisely such a tradeoff; so does the choice of the template, as we have already argued. Adequate choice of reference families for specific applications and goals, and an appropriate notion of which sets well approximated by the reference family, remains an important avenue to explore. The specific case of a spatially-structured signal is studied in Durand et al. (2018).

7.4. *Principled use of user-agnostic bounds and admissible sets.* This point stems from an insightful remark by an anonymous reviewer. If there are no constraints on the rejected set R selected by the user, and a post hoc bound $V(\cdot)$ is available, it seems sensible to require that one should not be able to add hypotheses to the rejected set without increase of the bound on false discoveries, nor exclude hypotheses from it without decrease of the bound on true discoveries; otherwise, the choice of R would obviously be suboptimal given the information given by the bound. Formally, call R admissible with respect to bound $V(\cdot)$ if

(i) $\forall R' \supsetneq R$, $V(R') > V(R)$;
(ii) $\forall R' \subsetneq R$, $S(R') < S(R)$.

We leave to the reader to check the following result: *the only sets admissible with respect to* $\overline{V}_{\mathfrak{R}}$ *(of* (8)*) belong to the reference family.* (In particular, for nested reference families, only the reference sets are admissible with respect to the optimal post hoc bound $V^*_{\mathfrak{R}}$.) This property emphasizes the role played by the choice of reference family—while also putting into question to allow rejection sets not belonging to it in the first place. Concerning this last point, we argue that additional constraints (sometimes only implicitly defined by the selection procedure used) often restrict the rejection sets under consideration of the user (this is the case in the two exemplary applications mentioned in the Introduction). In such a situation, the reference sets might not satisfy the constraints, which justifies the interest of a bound for more general Rs. One may in this case adapt the above definition of admissible sets by restricting comparisons to sets satisfying the constraints; which sets are then admissible would have to be investigated in specific situations.

In any case, introducing flexibility in the bound to allow for arbitrary rejection sets should not be interpreted as absolving the user of any responsibility: they should still expose the selection protocol they used—even if only heuristically motivated—in a convincing manner.

7.5. *Optimality in detection power.* Numerical experiments of Section 6.2 show that, while the balanced post hoc bound seems to improve over the Simes bound in many cases, neither bound uniformly outperforms the other in terms of averaged power (27). By contrast,

consider the *detection* power, defined as the probability that the bound, applied to the entire set of hypotheses $\mathbb{N}_m$, is nontrivial and indicates at least a nonnull hypothesis (see (S-15)). We show in Section S-4 that the Simes post hoc bound is always more conservative than the balanced one in a certain asymptotic regime. In a nutshell, the reason is that the balanced post hoc bound is related to the higher criticism method described in Donoho and Jin (2004) (optimal for detection), while the Simes post hoc bound is related to the Benjamini–Hochberg procedure of Benjamini and Hochberg (1995) (suboptimal for detection). While the detection power is certainly a somewhat coarse way to measure the quality of a post hoc bound, this once more underlines the potential advantage of the balanced bound over the Simes one. It is also of interest to note that while providing much more detailed information than mere detection, the post hoc bound retains optimal detection power in the considered setting.

7.6. *Further perspective.* In recent work of Katsevich and Ramdas (2018), false positive bounds are established uniformly over paths of rejection sets induced by several standard multiple testing procedures. Interestingly, they proved that the price to pay for this uniformity is generally quite low. This can be fruitfully combined to the bounds in the current paper to obtain user-agnostic bounds (note that the rejection paths are usually naturally nested).

Acknowledgments. We would like to thank an Associate Editor and the referees for their insightful comments. We are also most grateful to Aaditya Ramdas for pointing to us important connections between the present work and that of Genovese and Wasserman (2006). We also thank Yoav Benjamini and Jelle Goeman for interesting discussions, and Guillermo Durand for a careful reading of the manuscript. This work was finished while the first author was a guest at the Institut des Hautes Études Scientifiques, Université Paris-Saclay, whose support is also gratefully acknowledged.

Acknowledgments. This work was supported by CNRS (PEPS FaSciDo), ANR-16-CE40-0019 (SansSouci) and the French ministry of foreign and European affairs (EGIDE—PROCOPE project number 21887 NJ).

The first author was supported by the German DFG, under the Research Unit FOR-1735 "Structural Inference in Statistics and Adaptation and Efficiency," and under the Collaborative Research Center SFB-1294 "Data Assimilation."

The third author was supported by ANR-17-CE40-0001 (BASICS).

SUPPLEMENTARY MATERIAL

Supplement to "Post hoc confidence bounds on false positives using reference families" (DOI: 10.1214/19-AOS1847SUPP; .pdf). The supplement provides additional materials: relation to previous work, additional properties of templates and reference families, algorithms, proofs and numerical experiments.

REFERENCES

ARLOT, S., BLANCHARD, G. and ROQUAIN, E. (2010). Some nonasymptotic results on resampling in high dimension. II. Multiple tests. *Ann. Statist.* **38** 83–99. MR2589317 https://doi.org/10.1214/08-AOS668

BACHOC, F., PREINERSTORFER, D. and STEINBERGER, L. (2019). Uniformly valid confidence intervals post-model-selection. *Ann. Statist.* To appear.

BELLONI, A., CHERNOZHUKOV, V. and HANSEN, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Rev. Econ. Stud.* **81** 608–650. MR3207983 https://doi.org/10.1093/restud/rdt044

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

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

BENJAMINI, Y. and YEKUTIELI, D. (2005). False discovery rate-adjusted multiple confidence intervals for selected parameters. *J. Amer. Statist. Assoc.* **100** 71–93. With comments and a rejoinder by the authors. MR2156820 https://doi.org/10.1198/016214504000001907

BERK, R., BROWN, L., BUJA, A., ZHANG, K. and ZHAO, L. (2013). Valid post-selection inference. *Ann. Statist.* **41** 802–837. MR3099122 https://doi.org/10.1214/12-AOS1077

BLANCHARD, G., NEUVIAL, P. and ROQUAIN, E. (2019). R package sansSouci version 0.8.0. https://github.com/pneuvial/sanssouci.

BLANCHARD, G., NEUVIAL, P. and ROQUAIN, E. (2019). Supplement to "Post hoc confidence bounds on false positives using reference families." https://doi.org/10.1214/19-AOS1847SUPP.

BÜHLMANN, P. and MANDOZZI, J. (2014). High-dimensional variable screening and bias in subsequent inference, with an empirical comparison. *Comput. Statist.* **29** 407–430. MR3261821 https://doi.org/10.1007/s00180-013-0436-3

DEZEURE, R., BÜHLMANN, P., MEIER, L. and MEINSHAUSEN, N. (2015). High-dimensional inference: Confidence intervals, p-values and R-software `hdi`. *Statist. Sci.* **30** 533–558. MR3432840 https://doi.org/10.1214/15-STS527

DONOHO, D. and JIN, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32** 962–994. MR2065195 https://doi.org/10.1214/009053604000000265

DUDOIT, S. and VAN DER LAAN, M. J. (2008). *Multiple Testing Procedures with Applications to Genomics*. *Springer Series in Statistics*. Springer, New York. MR2373771 https://doi.org/10.1007/978-0-387-49317-6

DURAND, G., BLANCHARD, G., NEUVIAL, P. and ROQUAIN, E. (2018). Post hoc false positive control for spatially structured hypotheses. Preprint. Available at arXiv:1807.01470.

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

FITHIAN, W., SUN, D. and TAYLOR, J. (2014). Optimal Inference after model selection. Preprint. Available at arXiv:1410.2597.

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

GENOVESE, C. R. and WASSERMAN, L. (2006). Exceedance control of the false discovery proportion. *J. Amer. Statist. Assoc.* **101** 1408–1417. MR2279468 https://doi.org/10.1198/016214506000000339

GOEMAN, J. J. and SOLARI, A. (2011). Multiple testing for exploratory research. *Statist. Sci.* **26** 584–597. MR2951390 https://doi.org/10.1214/11-STS356

HEMERIK, J. and GOEMAN, J. J. (2018). False discovery proportion estimation by permutations: Confidence for significance analysis of microarrays. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 137–155. MR3744715 https://doi.org/10.1111/rssb.12238

HEMERIK, J., SOLARI, A. and GOEMAN, J. (2019). Permutation-based simultaneous confidence bounds for the false discovery proportion. *Biometrika*. To appear.

KATSEVICH, E. and RAMDAS, A. (2018). Towards "simultaneous selective inference": Post-hoc bounds on the false discovery proportion. Preprint. Available at arXiv:1803.06790.

LEE, J. D., SUN, D. L., SUN, Y. and TAYLOR, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.* **44** 907–927. MR3485948 https://doi.org/10.1214/15-AOS1371

LI, W. (2012). Volcano plots in analyzing differential expressions with mrna microarrays. *Journal of Bioinformatics and Computational Biology* **10** 1231003.

MEINSHAUSEN, N. (2006). False discovery control for multiple tests of association under general dependence. *Scand. J. Stat.* **33** 227–237. MR2279639 https://doi.org/10.1111/j.1467-9469.2005.00488.x

MEINSHAUSEN, N. and BÜHLMANN, P. (2005). Lower bounds for the number of false null hypotheses for multiple testing of associations under general dependence structures. *Biometrika* **92** 893–907. MR2234193 https://doi.org/10.1093/biomet/92.4.893

R CORE TEAM (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

ROMANO, J. P. and WOLF, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *J. Amer. Statist. Assoc.* **100** 94–108. MR2156821 https://doi.org/10.1198/016214504000000539

SIMES, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73** 751–754. MR0897872 https://doi.org/10.1093/biomet/73.3.751

TAYLOR, J. and TIBSHIRANI, R. J. (2015). Statistical learning and selective inference. *Proc. Natl. Acad. Sci. USA* **112** 7629–7634. MR3371123 https://doi.org/10.1073/pnas.1507583112

VAN DER LAAN, M. J., DUDOIT, S. and POLLARD, K. S. (2004). Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Stat. Appl. Genet. Mol. Biol.* **3** Art. 15, 27. MR2101464 https://doi.org/10.2202/1544-6115.1042

WESTFALL, P. H. and YOUNG, S. S. (1993). *Resampling-Based Multiple Testing. Examples and Methods for P-Value Adjustment*. Wiley.

WOO, C.-W., KRISHNAN, A. and WAGER, T. D. (2014). Cluster-extent based thresholding in fMRI analyses: Pitfalls and recommendations. *NeuroImage* **91** 412–419.