Post hoc inference via joint family-wise error rate control

Gilles Blanchard, Pierre Neuvial, Etienne Roquain

Post hoc inference via joint family-wise error rate control

Gilles Blanchard

Universität Potsdam, Institut für Mathematik
Karl-Liebknecht-Straße 24-25 14476 Potsdam, Germany
e-mail: gilles.blanchard@math.uni-potsdam.de

and

Pierre Neuvial

Institut de Mathématiques de Toulouse;
UMR 5219, Université de Toulouse, CNRS
UPS IMT, F-31062 Toulouse Cedex 9, France
e-mail: pierre.neuvial@math.univ-toulouse.fr

and

Etienne Roquain

Sorbonne Université (Université Pierre et Marie Curie), LPSM,
4, Place Jussieu, 75252 Paris cedex 05, France
e-mail: etienne.roquain@upmc.fr

Abstract: We introduce a general methodology for post hoc inference in a large-scale multiple testing framework. The approach is called "user-agnostic" in the sense that the statistical guarantee on the number of correct rejections holds for any set of candidate items selected by the user (after having seen the data). This task is investigated by defining a suitable criterion, named the joint-family-wise-error rate (JER for short). We propose several procedures for controlling the JER, with a special focus on incorporating dependencies while adapting to the unknown quantity of signal (via a step-down approach). We show that our proposed setting incorporates as particular cases a version of the higher criticism as well as the closed testing based approach of Goeman and Solari (2011). Our theoretical statements are supported by numerical experiments.

AMS 2000 subject classifications: Primary 62G10; secondary 62H15.
Keywords and phrases: post hoc inference, multiple testing, Simes inequality, family-wise error rate, step-down algorithm, dependence, higher criticism.

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 look like significant measurements. 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.

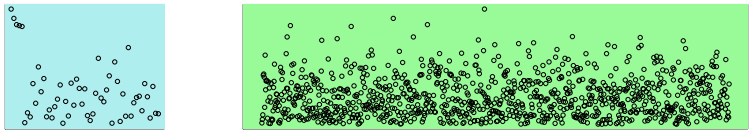

FIG 1. *Illustration of the post hoc selection effect. Right: virtual data set with* 1000 *measurements. Left: data set of* 55 *measurements selected from the right dataset. Measures have been generated as i.i.d. absolute values of* $\mathcal{N}(0,1)$.

A particular case of post hoc inference is faced when the selection step R is a pre-specified selection method, see Benjamini and Yekutieli (2005); Lockhart et al. (2014); Fithian et al. (2014); Bühlmann and Mandozzi (2014); Belloni et al. (2014); Taylor and Tibshirani (2015), among others. However, 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 et al., 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, e.g., Goeman and Solari (2011); Berk et al. (2013); Bachoc et al. (2016) 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 paper Goeman and Solari (2011). More formally, our goal is to build functional $V(\cdot)$ defined on all subset of hypotheses, such that the following uniform guarantee holds:

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

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.

Our construction of post hoc bounds relies on the control of a multiple testing criterion that we call "joint (family-wise) error rate" (JER for short), which was implicitly defined by Meinshausen (2006) for building false discovery proportion confidence envelopes (see also Genovese and Wasserman, 2004, 2006 for more details on this topic). The JER has a particularly simple expression in the case of p-value thresholding: given a family $\{p_i(X), 1 \leq i \leq m\}$ of m p-values and a family of thresholds $\mathcal{T} = (t_k)_{1\leq k\leq K}$, the JER of $\mathcal{T}$ is related to the distribution of $p_{(k:\mathcal{H}_0)}$, the k-th smallest value in the set $\{p_i, i \in \mathcal{H}_0\}$:

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

where $m_0 = |\mathcal{H}_0|$ is the number of true null hypotheses. It turns out that finding $\mathcal{T}$ such that $\mathrm{JER}(\mathcal{T}) \leq \alpha$ provides that the functional

$$V(R) = \min_{k\in\{1,\dots,K\}} \left\{ \sum_{i\in R} \mathbb{1}\left\{p_i(X) \geq t_k\right\} + k - 1 \right\}, \quad R \subset \{1,\dots,m\} \tag{2}$$

is a valid post hoc bound (see Section 2 for a proof in a general context). Hence, 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.

The contributions of the present work are the following:

- a general framework to build post hoc bounds, that generalizes the method of Goeman and Solari (2011) and does not rely on closed testing but on JER control;
- JER controlling procedures, with adaptivity to dependence and to the proportion of true null hypotheses. These procedures are implemented in an open-source R (R Core Team, 2017) package (Blanchard et al., 2017a);
- reproducible numerical experiments to illustrate our theoretical statements.

In addition, this study connects former (*a priori* unrelated) concepts: the closed testing-based method of Goeman and Solari (2010, 2011), the confidence envelopes of Meinshausen (2006) and the higher criticism of Donoho and Jin (2004).

The paper is organized as follows. In Section 2, we expose the general approach to post hoc multiple test inference based on JER control. In the following sections, we develop this point of view in some specific exemplary models under known or unknown dependence structure; the models are presented in Section 3 and the basic JER control obtained using the classical Simes inequality is analyzed in Section 4. In Section 5, we present improvements to this basic case by considering more general threshold families, incorporating adaptation to noise dependence structure, and a step-down principle. Two specific examples of this improved methodology are developed in Section 6. In Section 7, we present the results of numerical simulations illustrating and comparing the developed methods. We conclude with a discussion of various points in Section 8. Due to space constraints, proofs as well as some additional results (including a detailed comparison to the work of Goeman and Solari, 2010, 2011, to the higher criticism of Donoho and Jin, 2004, related optimality properties for detection purposes, and algorithmic details concerning Monte-Carlo and permutation-based calibration) are postponed to the supplementary material Blanchard et al. (2017b). 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 will 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 (1) and (2) (Section 2.4). Formal proofs for theoretical claims in this section are found in Section S-6.1.

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

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

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:

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

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 on $R \subset \mathbb{N}_m$.

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 family-wise Error Rate (JER). The basic argument is illustrated by Figure 2. 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 2. 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

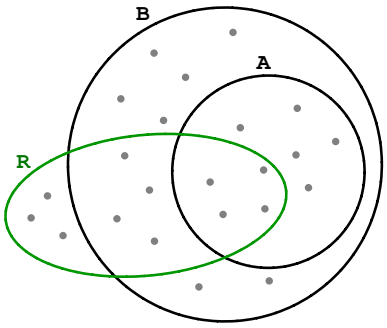

FIG 2. *Toy example: use of a reference family with two subsets A and B to build a post hoc bound on the number of true positives in an arbitrary candidate rejection set R.*

pictured in Figure 2, this ensures that R contains at least one element which is in $\mathcal{H}_1(P)$.

More generally, let us assume that we have at hand $\mathfrak{R} = ((R_1(X), \zeta_1(X)), \ldots, (R_K(X), \zeta_K(X)))$ 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,

$$\text{For all } P \in \mathcal{P}, \qquad \mathrm{JER}(\mathfrak{R}, P) \leq \alpha, \tag{3}$$

where we have denoted

$$\mathrm{JER}(\mathfrak{R}, P) := 1 - \mathbb{P}_{X\sim P}(\mathcal{E}(\mathfrak{R}, \mathcal{H}_0(P)))\,, \tag{4}$$

with the event

$$\mathcal{E}(\mathfrak{R}, \mathcal{H}_0) := \{\forall k = 1, \ldots, K,\ |R_k(X) \cap \mathcal{H}_0| \leq \zeta_k(X)\}\,. \tag{5}$$

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. Different choices are possible for $\mathfrak{R}$, allowing to recover as particular cases settings considered in previous literature. Let us mention two important cases concerning the bounds ζ_k:

- $\zeta_k = k-1$ for all k: in this case, each individual rejection region R_k has controlled k-FWER, and the control is uniform over the regions.
- $\zeta_k = |R_k| - 1$ for all k: adopting a different point of view, 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 (3)-(5) is interpreted as saying that with high probability,

each individual rejection region R_k has 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 family-wise error rate control over this family of hypotheses.

Our first goal in this paper is to analyze how to go from the JER control (3)-(5) to a post hoc statement (PH_α). This will be done in the present section in a general setting. In the remaining sections, we will concentrate on how to obtain the JER control itself. For this, we will focus on the first situation above ($\zeta_k = k - 1$) and therefore assume this setting by default unless otherwise specified. In the second situation ($\zeta_k = |R_k| - 1$), JER control can in particular be obtained via closed testing, thus recovering the setting of Goeman and Solari (2011), see Section S-1 for a more detailed discussion.

How can we "interpolate" from the control on a reference family (3) to a control on all possible rejection sets (PH_α)? On the event (5), the only available information on the unknown subset $\mathcal{H}_0$ is that it is an element of the collection of subsets

$$\begin{aligned}\mathcal{A}(\mathfrak{R}) &:= \{A \subset \mathbb{N}_m : \mathcal{E}(\mathfrak{R}, A) \text{ holds } \} \\ &= \{A \subset \mathbb{N}_m : \ \forall k = 1, \dots, K, \ |R_k \cap A| \leq \zeta_k\} .\end{aligned}$$

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:

$$V^*_{\mathfrak{R}}(R) := \max_{A \in \mathcal{A}(\mathfrak{R})} |R \cap A| , \quad R \subset \mathbb{N}_m . \tag{6}$$

A significant problem is that $V^*(R)$ (we will sometimes drop the index $\mathfrak{R}$ for simplicity) may not be easy to compute in general (see Proposition 2.3 below). We therefore introduce the following coarser but simpler bound:

$$\overline{V}_{\mathfrak{R}}(R) := \min_{k \in \{1,\dots,K\}} (|R \setminus R_k| + \zeta_k) \wedge |R| , \quad R \subset \mathbb{N}_m . \tag{7}$$

Observe that $\overline{V}(R)$ is non-decreasing in the sense that $R \subset R'$ implies $\overline{V}(R) \leq \overline{V}(R')$. The next result formalizes the link between JER control and the associated post hoc bounds. This result is proved in Section S-6.1, along with all of the other results of the present section.

Proposition 2.1. *Let $\mathfrak{R} = (R_k(X), \zeta_k(X))_{1 \leq k \leq K}$ be a data-dependent collection of subsets R_k of $\mathbb{N}_m$ and of integers ζ_k. Then for any $\mathcal{H}_0 \subset \mathbb{N}_m$, $\mathcal{H}_1 = \mathbb{N}_m \setminus \mathcal{H}_0$, the event $\mathcal{E}(\mathfrak{R}, \mathcal{H}_0)$ defined in (5) is such that*

$$\mathcal{E}(\mathfrak{R}, \mathcal{H}_0) = \left\{\forall R \subset \mathbb{N}_m, \ |R \cap \mathcal{H}_0| \leq \overline{V}_{\mathfrak{R}}(R)\right\} \tag{8}$$

$$= \{\forall R \subset \mathbb{N}_m, \ |R \cap \mathcal{H}_0| \leq V^*_{\mathfrak{R}}(R)\} . \tag{9}$$

In particular, Proposition 2.1 shows that $\mathfrak{R}$ satisfies the JER control (3) if and only if $\overline{V}_{\mathfrak{R}}(\cdot)$ or $V^*_{\mathfrak{R}}(\cdot)$ satisfies (PH_α).

2.3. General properties

In this section, we further discuss general properties of the obtained post hoc bounds. The JER control gives rise to the post hoc upper-bound $\overline{V}$, which we can see as an approximation of the optimal bound V^*. A first legitimate question is whether an approximation of the optimal bound is necessary in the first place, and then whether these approximations possess favorable properties. In this section, we provide arguments in this direction.

Remark 2.2. 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.

Computing the optimal bounds is NP-hard The claim that computing the optimal bound V^* is computationally difficult in general is supported by the following NP-hardness result:

Proposition 2.3. *The problem of computing $V^*_{\mathfrak{R}}(R)$ given an 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 2.3 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 — in fact, we will discuss precisely such a situation below (nested regions). Still, it is worth noting that the proof of the above result establishes NP-hardness for the more specific case $\zeta_k = |R_k| - 1$, where the reference family is interpreted as tests of certain intersection hypotheses. We show in Section S-1 that in this case, the bound V^* coincides with the bound that can be derived from the closed testing approach of Goeman and Solari (2011).

In general, it is therefore sensible in practice to look for computable approximations of V^*. We turn to general properties of the proposed bound $\overline{V}$.

Self-consistency Given some reference family $\mathfrak{R} = (R_k, \zeta_k)_{1\le k\le K}$, on the large probability event (8) 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

$$\tilde{\zeta}_k := \overline{V}_{\mathfrak{R}}(R_k) = \min_{j\in\{1,\dots,K\}} (|R_k \setminus R_j| + \zeta_j) \wedge |R_k|, \quad 1 \le k \le K. \tag{10}$$

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 $\tilde{\zeta}$.

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 (10). Denoting $\widetilde{\mathfrak{R}} = (R_k, \tilde{\zeta}_k)_{1\le k\le K}$, we have*

$$\overline{V}_{\mathfrak{R}}(R) = \min_{k\in\{1,\dots,K\}} \left(|R \setminus R_k| + \tilde{\zeta}_k\right) \wedge |R| = \overline{V}_{\widetilde{\mathfrak{R}}}(R), \quad R \subset \mathbb{N}_m. \tag{11}$$

In particular, the $\tilde{\zeta}_k$s satisfy the following "self-consistency" equation:

$$\tilde{\zeta}_k = \min_{j\in\{1,\dots,K\}} \left(|R_k \setminus R_j| + \tilde{\zeta}_j\right) \wedge |R_k|, \quad 1 \le k \le K. \tag{12}$$

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 $\overline{V}_{\mathfrak{R}}(R) = V^*_{\mathfrak{R}}(R)$.*

The more specific reference families studied in the remainder of the paper will satisfy the nestedness assumption, but it is in general not the case for closed testing-based families.

2.4. Focus of the paper

The aim of the rest of the paper is to find suitable reference families $\mathfrak{R}$ (which may be seen as "procedures") that control the joint family-wise error rate at some pre-specified level α.

A variety of choices are possible for the reference family. In 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 $i \in \mathbb{N}_m$. As announced earlier, we will also always choose $\zeta_k = k-1$, $1 \le k \le K$ from now on and therefore omit the ζ and use the simplified notation $\mathfrak{R} = (R_1(X), \dots, R_K(X))$ for the reference family. We will also assume that K is non-random and has been fixed in advance. In this situation, a simple way to build a reference family is to use p-value thresholding:

$$R_k(X) = \{i \in \mathbb{N}_m \,:\, p_i(X) < t_k\}, \;\; k \in \{1, \dots, K\}, \tag{13}$$

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

$$X_i = \mu_i + \varepsilon_i, \quad i \in \mathbb{N}_m\,, \tag{14}$$

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 point-wise 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 (14). The p-values are assumed to be given by $p_i(X) = \overline{G}\left(|T(X_{i,j}, 1 \leq j \leq n)|\right)$, 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 $\varepsilon_j, 1 \leq j \leq n$ of ε_1. Then, by a standard argument (see, e.g., Arlot et al., 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:

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

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 et al. (2012). On the other hand, the unknown dependence case corresponds to the general situation where we have no information on Σ. A suitable statistics 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-equi-correlated 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 $(q_i(X))_{1\leq i\leq m}$ such that for all $P \in \mathcal{P}$,

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

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

Unknown dependence We assume that there is 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,

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

where $g.X$ denotes X that has been transformed by g. This assumption has been introduced in Hemerik and Goeman (2017) 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-4.

4. JER control based on classical inequalities

In this section, we present an elementary approach where JER control (3) is derived from probabilistic inequalities that are well-known in multiple testing literature.

4.1. Simes reference family

Theorem 4.1 (Simes and Hommel inequalities)**.** *Let $(p_i(X))_{i\in\mathbb{N}_m}$ be a p-value family for the null hypotheses $(H_{0,i})_{i\in\mathbb{N}_m}$, satisfying the characteristic property*

$$\forall P \in \mathcal{P}, \forall i \in \mathcal{H}_0(P),\quad \forall t \in [0,1],\ \mathbb{P}_{X\sim P}(p_i(X) \leq t) \leq t. \tag{15}$$

Then it holds that $\forall P \in \mathcal{P}$,

$$\mathbb{P}_{X\sim P}\left(\exists k \in \{1,\dots,m_0\} \,:\, p_{(k:\mathcal{H}_0)} \leq \frac{\alpha k}{m_0 c_m}\right) \leq \alpha\,, \tag{16}$$

where:

(i) $c_m = C_m := \sum_{i=1}^{m} 1/i$ *under arbitrary dependency of the p-value family;*
(ii) $c_m = 1$ *if for all* $P \in \mathcal{P}$*, the p-value family is positive regression dependent on each element of the subset* $\mathcal{H}_0(P)$ *(in short, PRDS on* $\mathcal{H}_0(P)$*).*

Moreover, (16) *is an equality (with* $c_m = 1$*) when the* p_i*,* $i \in \mathcal{H}_0(P)$*, are i.i.d.* $U(0,1)$*.*

The inequalities corresponding to items (i) and (ii) are often referred to as the Hommel inequality (Hommel, 1983) and the Simes inequality (Simes, 1986), respectively. We refer to Benjamini and Yekutieli (2001) for a formal definition of the PRDS property. We recall that in the Gaussian model defined in Section 3.1 (one-sided), the PRDS assumption is valid if $\Sigma_{i,j} \geq 0$ for all $i, j \in \mathbb{N}_m$.

In view of (1), inequality (16) implies that the JER control (3) is satisfied for $K = m$ (under the appropriate conditions) by the reference family $\mathfrak{R}^0 = (R_1^0(X), \dots, R_m^0(X))$ given by

$$R_k^0(X) = \left\{ i \in \mathbb{N}_m \,:\, p_i < \frac{\alpha k}{m c_m} \right\}, 1 \leq k \leq m. \tag{17}$$

Above, we have upper-bounded m_0 by m because m_0 is generally unknown. The Hommel inequality is known to be exaggeratedly conservative, because the correction term C_m is of the order of $\log(m)$. Therefore, we will only use in the sequel the reference family $\mathfrak{R}^0$ when $c_m = 1$ and refer to it as the *Simes reference family*. The corresponding bound is given by

$$\overline{V}_{\mathfrak{R}^0}(R) = \min_{k\in\{1,\dots,m\}} \left\{ \sum_{i\in R} \mathbb{1}\left\{p_i(X) \geq \alpha k/m\right\} + k - 1 \right\}\,,\quad R \subset \mathbb{N}_m\,. \tag{18}$$

This bound is considered as a baseline for our work. As shown in Section S-1, this bound is in fact equivalent to the one proposed in Goeman and Solari (2011) for Simes local tests.

4.2. Sharpness and conservativeness

An important limitation of the reference family $\mathfrak{R}^0$ is its conservativeness and lack of adaptiveness, that is, even if $\max_{P\in\mathcal{P}} \mathrm{JER}(\mathfrak{R}^0, P)$ is close to α, $\mathrm{JER}(\mathfrak{R}^0, P)$ can be far from α for the P that truly generated the data. Indeed, both inequalities stated in Theorem 4.1 are adjusted to a *worst case dependency*, thus do not adapt or take into account the dependence between the tested hypotheses. For example, when the test statistics are strongly positively dependent, the Simes

inequality may be too conservative, and the associated post hoc bounds will inherit this conservativeness.

To illustrate this point, we carried out a simulation study in the Gaussian equi-correlated model where the one-sided test statistics follow the distribution $\mathcal{N}(0, \Sigma)$ with $\Sigma_{ii} = 1$ and $\Sigma_{ij} = \rho$ for $i \neq j$, for some $\rho \geq 0$. As noted above, this p-value family is PRDS. We consider a white setting (that is, all null hypotheses are true, $m_0 = m = 1,000$). In Table 1, we quantify the conservativeness of JER control in this model as the ratio of the JER actually achieved (estimated from $1,000$ simulations) to the target JER level α (for $\alpha = 0.2$). For example, we observe that for $\rho = 0.2$, the JER actually achieved by the canonical reference family $\mathfrak{R}^0$ is only 73% of the target JER level.

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

TABLE 1

Conservativeness of JER control based on Simes inequality in the Gaussian equi-correlated model. Here, $m_0 = m = 1,000$ and $\alpha = 0.2$. The standard error estimate is below 0.001 in all cases.

4.3. Unbalancedness

Let us consider a "favorable" case P for the Simes procedure, for which the p-values are i.i.d. uniform on $(0, 1)$. In this case, the Simes inequality is an equality

$$\mathbb{P}_{X\sim P}\left(\exists k \in \{1, \dots, m\} \,:\, p_{(k:m)} < \frac{\alpha k}{m}\right) = \alpha\,, \tag{19}$$

where $p_{(k:m)}$ is the k-th smallest p-value. In particular, the conservativeness described in Section 4.2 is not true here, and we might conclude that the family reference $\mathfrak{R}^0$ given by (17) can be suitably used for our aim. However, we argue that the errors in the event described in (19) are *not balanced* w.r.t. the parameter k. As an illustration, $\mathbb{P}(p_{(1:m)} < \alpha/m) = 1 - \left(1 - \frac{\alpha}{m}\right)^m = \alpha + o(\alpha)$, hence the probability of the event in (19) 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 $m = 1,000$, where $p_{(k:m)} \sim \mathrm{Beta}(k, m+1-k)$. 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-2.3, see (S-10).

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

TABLE 2

Values of $\mathbb{P}(p_{(k:m)} < \alpha k/m)$ for several k when $p_{(k:m)} \sim \mathrm{Beta}(k, m+1-k)$, $m = 1,000$ and $\alpha = 0.05$.

5. Methodology for adaptive JER control

In this section, we aim at building a thresholding-based reference family $\mathfrak{R}$ for which the quantity $\mathrm{JER}(\mathfrak{R}, P)$ is as close as possible to α, for "many interesting Ps". 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 $\mathcal{H}_0$.

5.1. Threshold templates

We start with considering a reference family $\mathfrak{R}_\lambda$ of the form (13), 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 (3) is satisfied, which we call λ-*calibration.*

Definition 5.1. 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 non-decreasing 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 (13). Several choices of template are possible as we will see in Section 6. 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 $y \in \mathbb{R} \cup \{-\infty, +\infty\}$.

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 (1), we obtain

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

This proves the following result.

Lemma 5.2. *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 (4) of the reference family $\mathfrak{R}_\lambda$ given by (13) can be written as follows: for any $P \in \mathcal{P}$,*

$$\mathrm{JER}(\mathfrak{R}_\lambda, P) = \mathbb{P}_{X\sim P}\Big(\min_{1\le k\le K\wedge m_0}\left\{t_k^{-1}\left(p_{(k:\mathcal{H}_0)}(X)\right)\right\} < \lambda\Big). \tag{20}$$

5.2. Single-step and step-down procedures by λ-calibration

The JER control (3) can now be achieved by choosing λ in an appropriate way.

Definition 5.3. 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 non-increasing in A, that is,

$$\forall \alpha \in (0,1), \forall A, A' \subset \{1,\dots,m\}, \text{ with } A \subset A', \quad \lambda(\alpha, A') \le \lambda(\alpha, A), \tag{21}$$

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

$$\mathbb{P}_{X\sim P}\left(\min_{1\le k\le K\wedge m_0} \left\{ t_k^{-1}\left(p_{(k:\mathcal{H}_0(P))}(X)\right)\right\} < \lambda(\alpha, \mathcal{H}_0(P))\right) \le \alpha. \tag{22}$$

Two examples of possible λ-calibrations will be provided in Sections 5.3 and 5.4. In the remaining of this section, 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 5.2 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 (21) 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 5.4. *In the framework of Lemma 5.2, consider $\lambda(\alpha) = \lambda(\alpha, \mathbb{N}_m)$ for some λ-calibration as in Definition 5.3. Then the procedure $\mathfrak{R}_{\lambda(\alpha)}$ controls the JER criterion at level α in the sense of (3).*

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 by substantial, as illustrated with numerical experiments in Section 7; this effect is further studied theoretically in Section S-2.2. This loss can be reduced by using $\lambda(\alpha, \widehat{A})$, where $\widehat{A}$ is the output of the the following step-down algorithm.

Algorithm 1: General step-down algorithm

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

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-6.2.

Proposition 5.5. *In the framework of Lemma 5.2, consider any λ-calibration as in Definition 5.3 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 (3).*

Remark 5.6. When we choose $K = 1$, Algorithm 1 reduces to the usual FWER controlling step-down algorithm (see, e.g., Romano and Wolf, 2005).

5.3. 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) and Lemma 5.2 thus give

$$\mathrm{JER}(\mathfrak{R}_\lambda, P) \leq \mathbb{P}_{q\sim\nu_m}\left(\min_{1\leq k\leq K\wedge m_0}\left\{t_k^{-1}\left(q_{(k:\mathcal{H}_0)}\right)\right\} < \lambda\right), \tag{23}$$

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

$$\lambda(\alpha, A) = \max\left\{\lambda \geq 0 \,:\, \mathbb{P}_{q\sim\nu_m}\left(\min_{1\leq k\leq K\wedge|A|}\left\{t_k^{-1}\left(q_{(k:A)}\right)\right\} < \lambda\right) \leq \alpha\right\}. \tag{24}$$

Property (21) 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 5.7 (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 (24) is a λ-calibration in the sense of Definition 5.3 and thus $\mathfrak{R}_{\lambda(\alpha,\mathbb{N}_m)}$ and $\mathfrak{R}_{\lambda(\alpha,\widehat{A})}$ both control the JER at level α.*

5.4. 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 $A \subset \mathbb{N}_m$,

$$\Psi(X, A) = \min_{1\leq k\leq K\wedge|A|}\left\{t_k^{-1}\left(p_{(k:A)}(X)\right)\right\}.$$

Now introduce the (data-dependent) λ-calibration

$$\lambda(\alpha, A) = \max\left\{\lambda \geq 0 \,:\, B^{-1}\sum_{j=1}^{B} \mathbb{1}\left\{\Psi(g_j.X, A) < \lambda\right\} \leq \alpha\right\}. \tag{25}$$

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 A.

Theorem 5.8 (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* (25) *is a λ-calibration in the sense of Definition 5.3 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 et al. (2017) to build confidence envelopes for the False Discovery Proportion.

6. 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 (24) (known dependence) or by (25) (unknown dependence).

6.1. Linear template

We define the *linear template* (of size K) by

$$t_k^L(\lambda) = \lambda k/m, \quad \lambda \in [0,1], \quad 1 \leq k \leq K. \tag{26}$$

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

$$R_k^L(X) = \left\{i \in \mathbb{N}_m \,:\, p_i < \lambda^L(\alpha, \mathbb{N}_m)\frac{k}{m}\right\}, 1 \leq k \leq K. \tag{27}$$

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

$$R_k^{L,sd}(X) = \left\{i \in \mathbb{N}_m \,:\, p_i < \lambda^L(\alpha, \hat{A})\frac{k}{m}\right\}, 1 \leq k \leq K, \tag{28}$$

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

Theorems 5.7 and 5.8 ensure that the reference families $\mathfrak{R}^L$ and $\mathfrak{R}^{L,sd}$ both control the JER at level α both in the known and unknown dependent case. The magnitude of $\lambda^L(\alpha, \mathbb{N}_m)$ is studied in Section S-2.1 in a simple case. It shows that our λ-calibration adapts to the dependence structure and addresses the conservativeness issue raised in Section 4.2.

6.2. Balanced template

Considering a linear template is not always appropriate: as mentioned above, under independence and $K = m$, $\mathfrak{R}^L$ corresponds to the Simes reference family $\mathfrak{R}^0$ (17), and thus suffers from a kind of unbalancedness, as underlined in Section 4.3. Ideally, a *balanced* reference family R_k would have the property that $\mathbb{P}(|R_k| \geq k)$ 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

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

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

$$t_k^B(\lambda) = F_k^{-1}(\lambda) = \min\{x \in [0,1] \,:\, F_k(x) \geq \lambda\}, \quad \text{with} \quad k \in \{1, \ldots, K\}. \tag{29}$$

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 5.1 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 $\mathfrak{R}^B = (R_1^B(X), \ldots, R_K^B(X))$, where

$$R_k^B(X) = \left\{i \in \mathbb{N}_m \,:\, p_i < t_k^B(\lambda^B(\alpha, \mathbb{N}_m))\right\}, 1 \leq k \leq K. \tag{30}$$

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

$$R_k^{B,sd}(X) = \left\{i \in \mathbb{N}_m \,:\, p_i < t_k^B(\lambda^B(\alpha, \hat{A}))\right\}, 1 \leq k \leq K, \tag{31}$$

  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-5 a detailed construction of the reference families $\mathfrak{R}^B$ and $\mathfrak{R}^{B,sd}$. Theorem 5.7 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 5.8 cannot be directly applied to the balanced template. Indeed, although this is not acknowledged by the notation for simplicity, F_k and thus $t_k^B(\lambda)$ depend on the observation X. Our proof does not generalize easily to such a data-dependent rejection template, although the numerical experiments of Section 7 suggest that the JER control is also valid in that situation.

Remark 6.1. The step-down refinement can be substantial for a balanced template, as further discussed in Section S-2.2.

Remark 6.2. By considering the two-sample setting with unknown dependency structure (see Section S-4) 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) there), 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 $g \in \mathcal{G}$, $(t_k(g.X, \lambda))_{1\le k\le K} = (t_k(X, \lambda))_{1\le k\le K}$ 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 $((t_k(X, \lambda))_{1\le k\le K}, (p_i(X))_{i\in\mathcal{H}_0})$ and $((t_k(g.X, \lambda))_{1\le k\le K}, (p_i(g.X))_{i\in\mathcal{H}_0})$.

7. Numerical experiments

We report numerical experiments performed in the two-sided location model (14) 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 5.7. The results of these experiments are quite similar to those reported here for unknown dependence.

7.1. JER control

The target JER level is set to $\alpha = 0.25$, and the simulation parameters are: $m = n = 1,000$, $\rho \in \{0, 0.2, 0.4\}$, $\pi_0 \in \{0.8, 0.9, 0.99\}$ (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 3 for the linear template, and by Figure 4 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 3 illustrates that the JER is controlled at the target level α

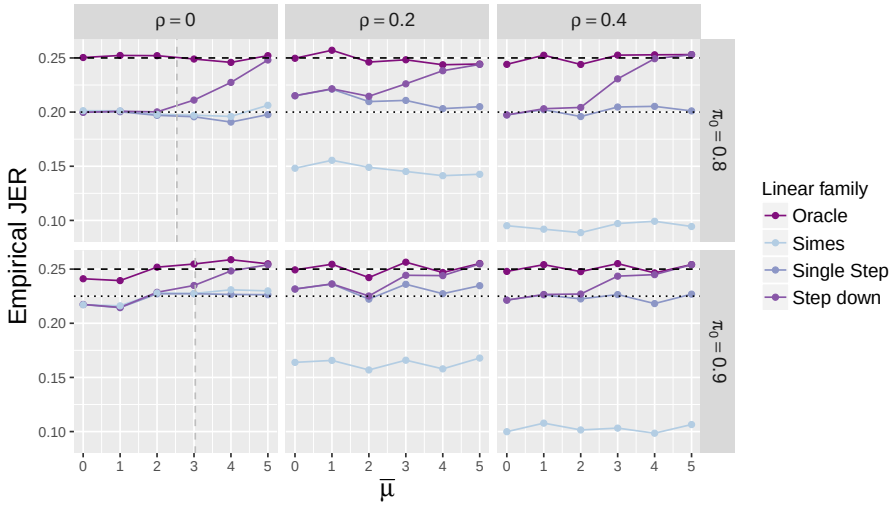

FIG 3. *JER control based on the linear template for equi-correlated test statistics.*

in all situations for the linear template, which is expected according to Proposition 5.8. Oracle calibration yields exact JER control, up to sampling fluctuations. As discussed in Section 4.2, 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 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 non-sparse 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 3.

The results for the balanced template are summarized by Figure 4. 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

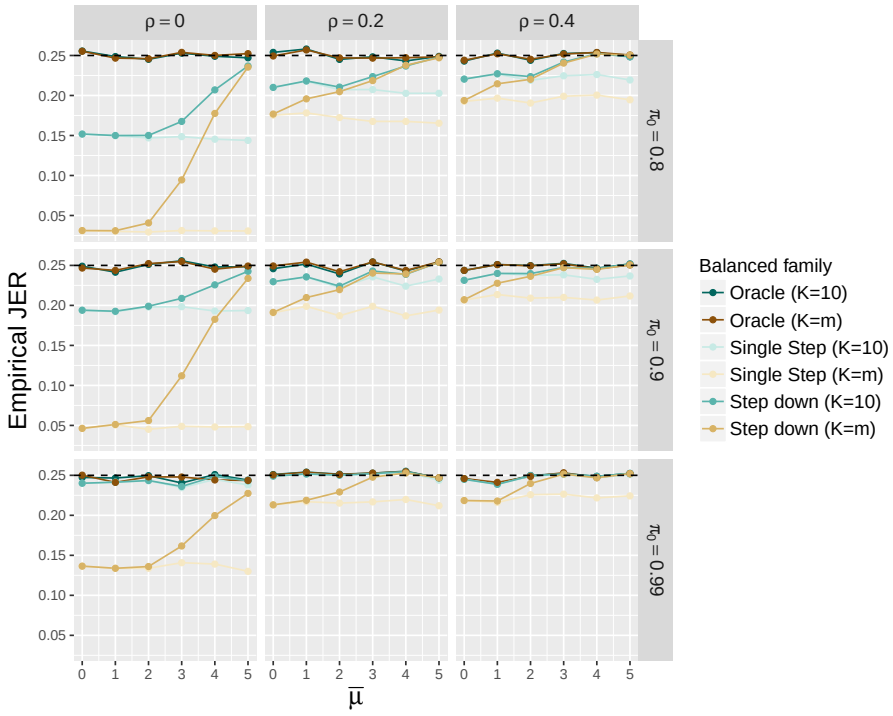

FIG 4. *JER control based on the balanced template for equi-correlated test statistics, with* $K = m$ *and* $K = 10$.

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 $m = 1,000$), 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-2.2.

Interestingly, the JER control offered by the balanced family with $K = 10$ (green curves in Figure 4) 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-2.1, and the question of how to choose K is discussed in Section 8.

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.

7.2. 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 (28) with $K = m$, and the step-down balanced reference family (31) 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),

$$\mathrm{Pow}(\mathfrak{R}, P, R) = \mathbb{E}\left(\frac{\overline{S}_{\mathfrak{R}}(R)}{|R \cap \mathcal{H}_1(P)|} \,\middle|\, |R \cap \mathcal{H}_1(P)| > 0 \right), \tag{32}$$

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 non-uniform 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 5. 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 3, 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 5 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.

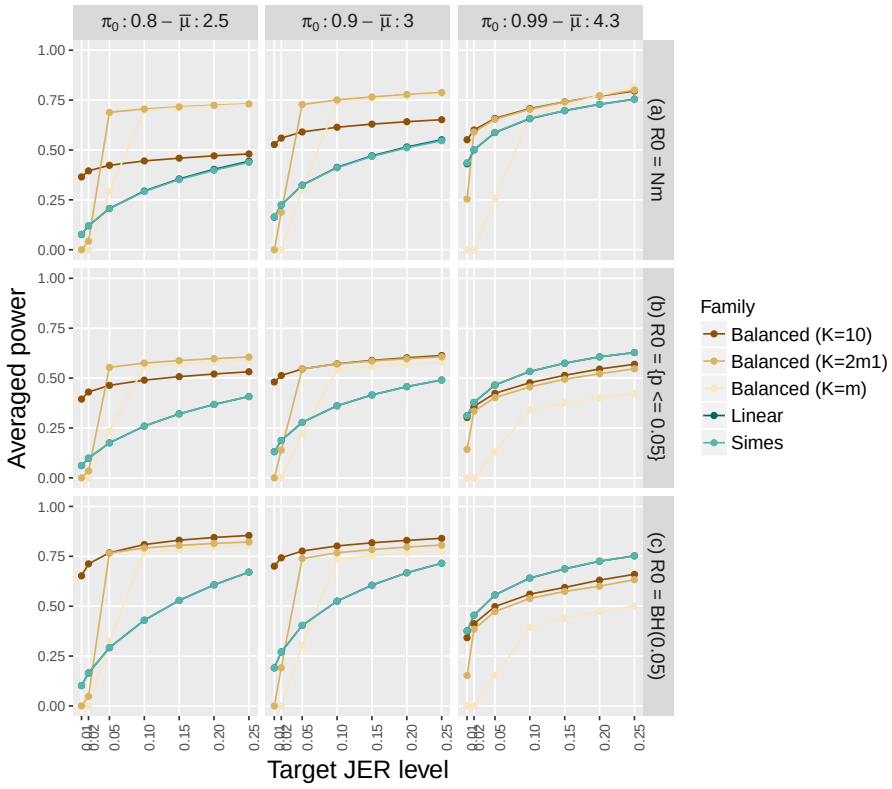

FIG 5. *Averaged power of JER controlling procedures for independent test statistics.*

8. Discussion

8.1. 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 4). The choice of K is therefore quite important in practice. We underline the following plausible scenarios:

- 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 direction to the choice of K is to introduce some smooth decay in the violation probability $\mathbb{P}(|R_k| \ge k)$ as k grows.

8.2. 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.

8.3. 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 (6), (7) 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 (3), 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 calibration function, 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.

8.4. 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 (7)) 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 lay out the protocol they used — even if only heuristically motivated — in a convincing manner.

Appendix A: Proof of Theorem 5.8

We denote in this proof $\lambda(\alpha, X, \mathcal{H}_0)$ instead of $\lambda(\alpha, \mathcal{H}_0)$ to underline the dependence of this functional w.r.t. the data X. By Propositions 5.4 and 5.5, it sufficient to prove that $\lambda(\cdot)$ is a valid λ-calibration, that is, satisfies the requirement of Definition 5.3. Since the monotonic property is clearly satisfied, it remains to establish (22). For this, write

$$\begin{aligned}
\mathbb{P}\Big(\min_{1\le k\le K\wedge m_0} \left\{ t_k^{-1}\left(p_{(k:\mathcal{H}_0)}(X)\right)\right\} < \lambda(\alpha, X, \mathcal{H}_0)\Big) & \\
&= \mathbb{P}\Big(\Psi(X, \mathcal{H}_0) < \lambda(\alpha, X, \mathcal{H}_0)\Big) \\
&\le \mathbb{P}\left(B^{-1}\sum_{j=1}^{B} \mathbb{1}\left\{\Psi(g_j.X, \mathcal{H}_0) \le \Psi(X, \mathcal{H}_0)\right\} \le \alpha\right) \\
&= \mathbb{P}\left(B^{-1}\sum_{j=1}^{B} \mathbb{1}\left\{Y_j \le Y_1\right\} \le \alpha\right),
\end{aligned}$$

where we have used in the inequality the definition of $\lambda(\alpha, X, \mathcal{H}_0)$ (see (25)) and we have let $Y_j = \Psi(g_j.X, \mathcal{H}_0)$, $1 \le j \le m$. Now, by (Rand), we easily check that $(Y_1, \dots, Y_B)$ is an exchangeable random vector: for any g_0 uniformly distributed on $\mathcal{G}$ (and drawn independently of the other variables),

$$\begin{aligned}
(Y_1, \dots, Y_B) &\sim (\Psi(g_1.g_0.X, \mathcal{H}_0), \dots, \Psi(g_B.g_0.X, \mathcal{H}_0)) \\
&\sim (\Psi(g_1'.X, \mathcal{H}_0), \dots, \Psi(g_B'.X, \mathcal{H}_0)),
\end{aligned}$$

where g_j', $1 \le j \le B$, are i.i.d. uniform in $\mathcal{G}$ (independent of X). Above, the first equality in distribution holds because it is true conditionally on $\{g_1, \dots, g_B\}$, and the second one holds because it is true conditionally on X. Since the variables $\Psi(g_j'.X, \mathcal{H}_0)$, $1 \le j \le m$, are i.i.d. conditionally on X, we deduce that $(Y_1, \dots, Y_B)$ is an exchangeable random vector. Hence, for any independent variable U uniformly distributed on $\{1, \dots, B\}$, we obtain

$$\mathbb{P}\left(B^{-1}\sum_{j=1}^{B} \mathbb{1}\left\{Y_j \le Y_1\right\} \le \alpha\right) = \mathbb{P}\left(B^{-1}\sum_{j=1}^{B} \mathbb{1}\left\{Y_j \le Y_U\right\} \le \alpha\right).$$

Let σ any permutation (independent of U) such that $Y_{\sigma(1)} \le \cdots \le Y_{\sigma(B)}$. Since $\sum_{j=1}^{B} \mathbb{1}\left\{Y_j \le Y_U\right\} = \sum_{j=1}^{B} \mathbb{1}\left\{Y_{\sigma(j)} \le Y_U\right\}$ and U and $\sigma(U)$ have the

same distribution conditionally on Y, we have

$$\mathbb{P}\left(B^{-1}\sum_{j=1}^{B}\mathbb{1}\left\{Y_j \le Y_U\right\} \le \alpha \mid Y\right) = \mathbb{P}\left(B^{-1}\sum_{j=1}^{B}\mathbb{1}\left\{Y_{\sigma(j)} \le Y_{\sigma(U)}\right\} \le \alpha \mid Y\right)$$
$$\le \mathbb{P}\left(B^{-1}\sum_{j=1}^{B}\mathbb{1}\left\{j \le U\right\} \le \alpha \mid Y\right) = \mathbb{P}\left(U \le \alpha B \mid Y\right) = \frac{\lfloor \alpha B \rfloor}{B} \le \alpha.$$

We underline that another argument is possible for this proof using a device recently proposed by Hemerik and Goeman (2017), see Section S-6.2 for more details.

Acknowledgements

We would like to acknowledge an associate editor and two referees for their insightful comments. We also thank Prof. Yoav Benjamini for interesting discussions, and Guillermo Durand for a careful reading of the manuscript. This work has been supported by CNRS (PEPS FaSciDo), ANR-16-CE40-0019 (SansSouci) and ANR-17-CE40-0001 (BASICS). The first author acknowledges the support from the german DFG, under the Research Unit FOR-1735 "Structural Inference in Statistics – Adaptation and Efficiency", and under the Collaborative Research Center SFB-1294 "Data Assimilation".

Supplementary Material

Supplement:
(doi: COMPLETED BY THE TYPESETTER; .pdf). The supplement includes relation to previous work (closed testing, higher criticism); general 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(1):83–99.

Bachoc, F., Preinerstorfer, D., and Steinberger, L. (2016). Uniformly valid confidence intervals post-model-selection. *arXiv preprint arXiv:1611.01043*.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Rev. Econ. Stud.*, 81(2):608–650.

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

Benjamini, Y. and Yekutieli, D. (2005). False discovery rate-adjusted multiple confidence intervals for selected parameters. *J. Amer. Statist. Assoc.*, 100(469):71–93. With comments and a rejoinder by the authors.

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013). Valid post-selection inference. *Ann. Statist.*, 41(2):802–837.

Blanchard, G., Neuvial, P., and Roquain, E. (2017a). R package sansSouci version 0.5.0. https://github.com/pneuvial/sanssouci.

Blanchard, G., Neuvial, P., and Roquain, E. (2017b). Supplement to "post hoc inference via joint family-wise error rate control". Submitted to AoS.

Bühlmann, P. and Mandozzi, J. (2014). High-dimensional variable screening and bias in subsequent inference, with an empirical comparison. *Comput. Statist.*, 29(3-4):407–430.

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

Fan, J., Han, X., and Gu, W. (2012). Estimating false discovery proportion under arbitrary covariance dependence. *Journal of the American Statistical Association*, 107(499):1019–1035.

Fithian, W., Sun, D., and Taylor, J. (2014). Optimal Inference After Model Selection. *ArXiv e-prints*.

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

Genovese, C. R. and Wasserman, L. (2006). Exceedance control of the false discovery proportion. *J. Amer. Statist. Assoc.*, 101(476):1408–1417.

Goeman, J. J. and Solari, A. (2010). The sequential rejection principle of familywise error control. *Ann. Statist.*, 38(6):3782–3810.

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

Hemerik, J. and Goeman, J. J. (2017). False discovery proportion estimation by permutations: confidence for significance analysis of microarrays. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Hemerik, J., Solari, A., and Goeman, J. J. (2017). Permutation-based simultaneous confidence bounds for the false discovery proportion. Private communication.

Hommel, G. (1983). Tests of the overall hypothesis for arbitrary dependence structures. *Biometrical J.*, 25(5):423–430.

Li, W. (2012). Volcano plots in analyzing differential expressions with mrna microarrays. *Journal of bioinformatics and computational biology*, 10(06):1231003.

Lockhart, R., Taylor, J., Tibshirani, R. J., and Tibshirani, R. (2014). A significance test for the lasso. *Ann. Statist.*, 42(2):413–468.

Meinshausen, N. (2006). False discovery control for multiple tests of association under general dependence. *Scand. J. Statist.*, 33(2):227–237.

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(469):94–108.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754.

Taylor, J. and Tibshirani, R. J. (2015). Statistical learning and selective infer-

ence. *Proc. Natl. Acad. Sci. USA*, 112(25):7629–7634.

Woo, C.-W., Krishnan, A., and Wager, T. D. (2014). Cluster-extent based thresholding in fMRI analyses: pitfalls and recommendations. *Neuroimage*, 91:412–419.