HAL
open science

Post hoc inference via joint family-wise error rate control

Gilles Blanchard, Pierre Neuvial, Etienne Roquain

To cite this version:

Gilles Blanchard, Pierre Neuvial, Etienne Roquain. Post hoc inference via joint family-wise error rate control. 2017. hal-01483585v2

HAL Id: hal-01483585

https://hal.science/hal-01483585v2

Preprint submitted on 9 Mar 2017 (v2), last revised 10 Apr 2019 (v5)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

UPMC Université Paris 6, LPMA,
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

A major challenge in modern high-dimensional statistics is that of large-scale multiple inference with a rigorous statistical guarantee. 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, then they may tend to "snoop" in the data, which will invalidate the 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 inference".

A particular case of post hoc inference is faced when the selection step R is a pre-specified selection method, see Lockhart et al. (2014), Fithian et al. (2014), Bühlmann and Mandozzi (2014),

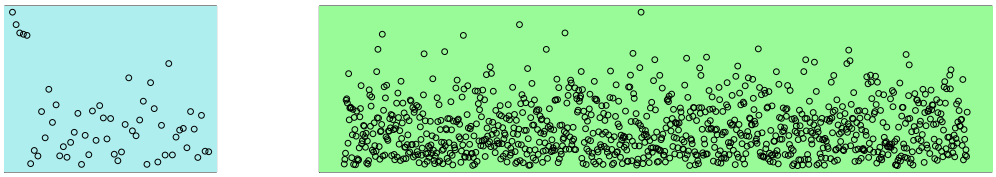

FIG 1. *Illustration of the post hoc selection effect. Right: virtual data set with* 1000 *measurements. Left: data set of* 55 *measurements selected from the right dataset. Measures have been generated as i.i.d. absolute values of* $\mathcal{N}(0,1)$.

Belloni et al. (2014) and Taylor and Tibshirani (2015), among others. In this case, the significance is established conditionally on the fact that the items have been previously selected in the first round. However, this does not allow fully for "data snooping", since the selection step is fixed.

By contrast, at a pure exploratory research step, many kinds of selection procedure R can be employed by the user, possibly many times, and it is desirable to provide a statistical guarantee *simultaneously* on any of them. This point of view has been proposed in several papers for various statistical guarantees, e.g., Benjamini and Yekutieli (2005); Goeman and Solari (2011); Berk et al. (2013). In this paper, we 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). We like to think to this general principle as "user-agnostic" or *in omnia paratus*, because the provided inference is "ready for any selected set".

Maybe at a more technical level, another inspiration for our work is the criterion that we call "joint family wise error rate" (joint error rate, JER, in short), which was implicitly defined in Meinshausen (2006) (in a less general form) for building false discovery proportion confidence envelopes (see also Genovese and Wasserman, 2004, 2006 for more details on this topic).

In a nutshell, the contributions of our work are listed as follows:

- we provide a general framework for obtaining post hoc bounds, that generalizes the method of Goeman and Solari (2011) and does not rely on closed testing but on JER control;
- new procedures controlling the JER are introduced, by incorporating the (known) dependence of the data, by introducing an additional parameter K and by defining a devoted step-down algorithm. Combining some of these three factors makes the obtained post hoc bounds potentially much sharper;
- we provide power optimality theoretical statements for detection purpose;
- our theoretical statements and the advantages of the new proposed procedures are illustrated with extensive numerical experiments.

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

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 (Section 2.4). Formal proofs for theoretical claims in this section are found in Appendix B.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 for a collection of null hypotheses $H_{0,i} \subset \mathcal{P}$ indexed by $i \in \mathbb{N}_m := \{1, \ldots, m\}$. 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 $\mathcal{H}_1 = \mathbb{N}_m \backslash \mathcal{H}_0$ be the set of (indices of) false nulls and $m_1 = m - m_0$ its cardinality.

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

$$\text{For all } P \in \mathcal{P} \qquad \mathbb{P}_{X\sim P}\Big(\forall R \subset \mathbb{N}_m,\ |R \cap \mathcal{H}_0(P)| \leq V(R)\Big) \geq 1 - \alpha. \tag{PH$_\alpha$}$$

or, equivalently, a function $S(R)(= |R| - V(R))$ satisfying

$$\text{For all } P \in \mathcal{P} \qquad \mathbb{P}_{X\sim P}\Big(\forall R \subset \mathbb{N}_m,\ |R \cap \mathcal{H}_1(P)| \geq S(R)\Big) \geq 1 - \alpha, \tag{PH$'_\alpha$}$$

where $\mathcal{H}_1(P) = \mathbb{N}_m \backslash \mathcal{H}_0(P)$.

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. In particular, this bound will apply (with probability at least $1 - \alpha$) to any arbitrary data-dependent choice of R made by the user (including of course choosing after looking at the value of the bound itself for different candidates for R).

2.2. General principle

The question of how to obtain a control of the general form (PH$_\alpha$) 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 $\overline{V}(R)$ 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 controlled joint Family-Wise Error Rate. 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, $|A \cap \mathcal{H}_0(P)| \leq 5$. Then this also provides information on other subsets $R \subset \mathbb{N}_m$ with $R \neq A$. Namely, for any $R \subset \mathbb{N}_m$, $|R \cap \mathcal{H}_1(P)| \geq |R \cap A| - 5$. Of course, while this information is useful for R if $|R \cap A| \geq 6$, it is not if $|R \cap A| \leq 5$ (nonpositive bound), as in the scheme below. Next, if we want to improve the bound, we can consider another set B (here including A) with the property $|B \cap \mathcal{H}_0(P)| \leq 7$ (say). In the situation pictured in the scheme below, 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,

$$\text{For all } P \in \mathcal{P}, \qquad \mathrm{JER}(\mathfrak{R}, P) \leq \alpha, \tag{1}$$

where we have denoted

$$\mathrm{JER}(\mathfrak{R}, P) := 1 - \mathbb{P}_{X\sim P}(\mathcal{E}(\mathfrak{R}, \mathcal{H}_0(P)))\,, \tag{2}$$

with the event

$$\mathcal{E}(\mathfrak{R}, \mathcal{H}_0) := \{\forall k = 1, \ldots, K,\ |R_k(X) \cap \mathcal{H}_0| \leq \zeta_k(X)\}\,. \tag{3}$$

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

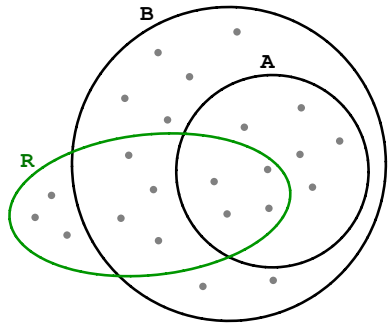

FIG 2. *Toy example: use of a reference family with two subsets A and B for the construction of post hoc bounds on the number of true positives in an arbitrary candidate rejection set R.*

(or size) K of the reference family is also allowed to be data-dependent in the most general form. 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:

- $\zeta_k = |R_k| - 1$ for all k: in this case, we if we associate to each $R \subset \mathbb{N}_m$ the *intersection hypothesis* $H_{0,R} := \bigcap_{i \in R} H_{0,i}$, the statement (1)-(3) can be equivalently interpreted as follows: if we reject all intersection hypotheses H_{0,R_k}, $k = 1, \dots, K$, then the corresponding family-wise error is controlled at level α. Rejecting certain intersection hypotheses with controlled family-wise error is the principle underlying closed testing, see Section 8 for a more detailed discussion.
- $\zeta_k = k - 1$ for all k: in this case, taken individually, each rejection set R_k has controlled k-FWER, and the above corresponds to a joint control of the error related to k-FWER, uniformly for all $k \in \{1, \dots, K\}$.

We will mainly focus on the second situation $\zeta_k = k - 1$ in the rest of the paper, and therefore assume this setting by default unless otherwise specified. However, in the present section and the next one, we analyze principles and properties that hold in relation to (1) in its general form.

How can we "interpolate" from the control on a reference family (1) to a control on all possible rejection sets (PH_α)? If our only available information is that event (3) is satisfied, then the best we can do to bound $|R \cap \mathcal{H}_0|$ for a proposed rejection set R is a worst-case bound given the known constraints:

$$V^*_{\mathfrak{R}}(R) := \max_{A \in \mathcal{A}(\mathfrak{R})} |R \cap A| \,, \quad R \subset \mathbb{N}_m \,, \tag{4}$$

$$S^*_{\mathfrak{R}}(R) = |R| - V^*_{\mathfrak{R}}(R) := \min_{A \in \mathcal{A}(\mathfrak{R})} |R \setminus A| \,, \quad R \subset \mathbb{N}_m \,, \tag{5}$$

where

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

is the collection of all possible configurations for $\mathcal{H}_0$ that are compatible with the event $\mathcal{E}(\mathfrak{R}, \mathcal{H}_0)$. Obviously, $V^*_{\mathfrak{R}}(R)$ is then the largest number of false rejections (resp. $S^*_{\mathfrak{R}}(R)$ the smallest number of correct rejections) in the set R that are compatible with this event.

A significant problem is that $V^*(R), S^*(R)$ (we will sometimes drop the index $\mathfrak{R}$ for simplicity) may not be easy to compute in general (see Proposition 2.2 below). We therefore introduce the following coarser but simpler bounds:

$$\overline{V}_{\mathfrak{R}}(R) := \min_{k \in \{1, \dots, K\}} (|R \setminus R_k| + \zeta_k) \wedge |R| \,, \ \ R \subset \mathbb{N}_m \,; \tag{6}$$

$$\overline{S}_{\mathfrak{R}}(R) = |R| - \overline{V}_{\mathfrak{R}}(R) := \max_{k \in \{1, \dots, K\}} (|R \cap R_k| - \zeta_k) \vee 0 \,, \ \ R \subset \mathbb{N}_m \,. \tag{7}$$

Observe that $\overline{V}(R)$ and $\overline{S}(R)$ lie in $\{0,\dots,|R|\}$ and their sum is $|R|$. Also, both of these functions are non-decreasing in the sense that $R \subset R'$ implies $\overline{V}(R) \le \overline{V}(R')$ as well as $\overline{S}(R) \le \overline{S}(R')$. The next result, proved in Section B.1, formalizes the link between JER control and the associated post hoc bounds.

Proposition 2.1. *Let $\mathfrak{R} = (R_k(X), \zeta_k(X))_{1\le k\le K}$ 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 following events are identical:*

$$\mathcal{E}(\mathfrak{R}, \mathcal{H}_0) = \left\{\forall R \subset \mathbb{N}_m,\ |R \cap \mathcal{H}_1| \ge \overline{S}(R)\right\} = \left\{\overline{S}(\mathcal{H}_0) = 0\right\} \tag{8}$$

$$= \left\{\forall R \subset \mathbb{N}_m,\ |R \cap \mathcal{H}_1| \ge S^*(R)\right\} = \left\{S^*(\mathcal{H}_0) = 0\right\}. \tag{9}$$

In particular, $\mathfrak{R}$ satisfies the JER control (1) if and only if $\overline{S}(R)$ or $S^*(R)$ satisfies (PH'_α) (or, equivalently, if and only if $\overline{V}(R)$ or $V^*(R)$ satisfies (PH_α)).

Proposition 2.1 shows that the probability of violating the post hoc statement (using either the bounds V^*, S^* or $\overline{V}, \overline{S}$) is exactly equal to the JER criterion (3). To this extent, the main aim of the rest of the paper will be to find a suitable reference family $\mathfrak{R}$ (which may be seen as the "procedure") that controls the joint family-wise error rate at some pre-specified level 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 $\overline{V}$ (resp. lower-bound $\overline{S}$), which we can see as approximations of the optimal bounds V^*, S^*. A first legitimate question is whether approximations of the optimal bounds are necessary in the first place, and then whether these approximations possess favorable properties. In this section, we provide arguments in this direction.

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

Proposition 2.2. *The problem of computing $V^*(R)$ 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 2.2 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 — in fact, we will discuss precisely such a situation below. On the other hand, it is worth noting that the proof of the above result establishes NP-hardness for the 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 8 that in this case, the bounds V^*, S^* coincide with those 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^*, S^*. We turn to general properties of the proposed bounds $\overline{V}, \overline{S}$.

Self-consistency of $\overline{V}, \overline{S}$. A first desirable property is *self-consistency*. Given some reference family $\mathfrak{R} = (R_i, \zeta_i)_{1\le i\le K}$, 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 2.3. *For any collection $\mathfrak{R} = (R_i, \zeta_i)_{1\le i\le K}$, define $(\tilde{\zeta}_i)_{1\le i\le K}$ by (10). Denoting $\widetilde{\mathfrak{R}} = (R_i, \tilde{\zeta}_i)_{1\le i\le K}$, 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, this implies that the $\tilde{\zeta}_k$s satisfy the following "self-consistency" equation:

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

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

Proposition 2.4. *For any collection $\mathfrak{R} = (R_k, \zeta_k)_{1\le k\le K}$ such that $R_k \subset R_{k'}$ whenever $k \le k'$, we have $\overline{V}_{\mathfrak{R}}(R) = V^*_{\mathfrak{R}}(R)$.*

2.4. Thresholding-based reference family

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 ("joint k-FWER control" setting) and therefore omit the ζs and use the simplified notation $\mathfrak{R} = (R_1(X), \dots, R_K(X))$ for the reference family (also called "procedure"). 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. The corresponding post hoc bounds are given by:

$$\overline{V}_{\mathfrak{R}}(R) = \min_{k\in\{1,\dots,K\}} \left\{ \sum_{i\in R} \mathbf{1}\{p_i(X) \ge t_k\} + k - 1 \right\}, \;\; R \subset \mathbb{N}_m;$$

$$\overline{S}_{\mathfrak{R}}(R) = \max_{k\in\{1,\dots,K\}} \left\{ \sum_{i\in R} \mathbf{1}\{p_i(X) < t_k\} - (k-1) \right\}, \;\; R \subset \mathbb{N}_m.$$

The JER control is related to the distribution of $p_{(k:\mathcal{H}_0)}$, the k-th smallest value in the set $\{p_i(X), i \in \mathcal{H}_0(P)\}$:

$$\mathrm{JER}(\mathfrak{R}, P) = \mathbb{P}_{X\sim P}\Big(\exists k \in \{1,\dots,K \wedge m_0\} \,:\, p_{(k:\mathcal{H}_0)} < t_k\Big). \tag{14}$$

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 uniformity in k.

2.5. Location model

Throughout the paper, we consider as a concrete testbed the location model

$$X_i = \mu_i + \varepsilon_i, \quad i \in \mathbb{N}_m, \tag{15}$$

where the ε_i are identically distributed, centered, and have a *known* joint distribution. We consider the one-sided (resp. two-sided) testing problem with null hypotheses $H_{0,i}$: "$\mu_i \le 0$" (resp. $H_{0,i}$:"$\mu_i = 0$") versus the alternative hypotheses $H_{1,i}$:"$\mu_i > 0$" (resp. $H_{1,i}$:"$\mu_i \ne 0$") for all $i \in \mathbb{N}_m$. The p-values are given by $p_i(X) = \overline{F}(X_i)$, where $\overline{F}(x) = \mathbb{P}(\varepsilon_1 \ge x)$ (resp. $\overline{F}(x) = \mathbb{P}(|\varepsilon_1| \ge |x|)$). We denote also by $q_i(X) = \overline{F}(\varepsilon_i)$ the "null-shifted p-values" in which the signal has been removed. While the quantities $q_i(X)$ are not observed and will be used purely as a technical device, note that their joint distribution is known (i.e. it is the p-value distribution under

the full null). This joint distribution will be denoted by ν_m. The q_i's have $U(0,1)$ marginals whenever $\overline{F}$ is continuous, which we will assume in the sequel for simplicity. Hence, under independence of the ε_i's, the q_i's are i.i.d. $U(0,1)$.

A simple particular case is the Gaussian location model for which $\varepsilon \sim \mathcal{N}(0,\Sigma)$ for some known covariance matrix Σ with $\Sigma_{i,i} = 1$ for $i \in \mathbb{N}_m$. This instance arises in a standard Gaussian linear model or in marginal regression, see Fan et al. (2012). 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 $\rho \in [0,1]$. Finally, we note that our approach is not restricted to models of the form (15), as we discuss in Section 9.

3. JER control based on classical inequalities

In this section, we present an elementary approach where JER control (1) is derived from probabilistic inequalities that are well-known in multiple testing literature.

3.1. Simes reference family

Proposition 3.1 (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*

$$\forall P \in \mathcal{P}, \forall i \in \mathcal{H}_0(P), \quad \forall t \in [0,1],\ \mathbb{P}_{X\sim P}(p_i(X) \leq t) \leq t. \tag{16}$$

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

$$\mathbb{P}_{X\sim P}\left(\exists k \in \{1,\ldots,m_0\} \,:\, p_{(k:\mathcal{H}_0)} \leq \frac{\alpha k}{m_0 c_m}\right) \leq \alpha\,, \tag{17}$$

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 positively regressively dependent on each one of $\mathcal{H}_0(P)$ (in short, PRDS on $\mathcal{H}_0(P)$).

Moreover, (17) 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 2.5 (one-sided), the PRDS assumption is valid if $\Sigma_{i,j} \geq 0$ for all $i,j \in \mathbb{N}_m$.

In view of (14), Inequality (17) implies that the JER control (1) is satisfied for $K = m$ (under the appropriate conditions) by the reference family $\mathfrak{R}^0 = (R_1^0(X),\ldots,R_m^0(X))$ given by

$$R_k^0(X) = \left\{ i \in \mathbb{N}_m \,:\, p_i < \frac{\alpha k}{m c_m} \right\}, 1 \leq k \leq m. \tag{18}$$

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

$$\overline{V}_{\mathfrak{R}^0}(R) = \min_{k\in\{1,\ldots,m\}} \left\{ \sum_{i\in R} \mathbf{1}\left\{p_i(X) \geq \alpha k/m\right\} + k - 1 \right\}, \quad R \subset \mathbb{N}_m\,. \tag{19}$$

This bound is considered as a baseline for our work. As shown in Section 8, this bound is in fact equivalent to the one proposed in Goeman and Solari (2011) from Simes local tests.

3.2. 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 Proposition 3.1 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, 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.

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.

3.3. Unbalancedness

Let us consider a "favorable" case P for the Simes procedure, for which the p-values are all independent and associated to true null hypotheses ("full null" configuration). In this case, the Simes inequality is an equality

$$\mathbb{P}_{X\sim P}\left(\exists k \in \{1, \ldots, m\} \,:\, p_{(k:m)} < \frac{\alpha k}{m}\right) = \alpha\,. \tag{20}$$

In particular, the conservativeness described in Section 3.2 is not true here, and we might conclude that the family reference $\mathfrak{R}^0$ given by (18) can be suitably used for our aim. However, we argue that the errors in the event described in (20) 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 (20) 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 Appendix A, see (44).

k	1	2	3	4	5	10	100	1000
$\mathbb{P}(p_{(k:m)} \leq \alpha k/m)$	$4.9e^{-2}$	$4.7e^{-3}$	$5.0e^{-4}$	$5.7e^{-5}$	$6.6e^{-6}$	$1.6e^{-10}$	$5.8e^{-93}$	0

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 $\alpha = 0.05$.

4. 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 P". To this end, we focus on the location model given in Section 2.5 for which the distribution of the noise is *known*. We combine two approaches:

- build a reference family $\mathfrak{R}$ that incorporates the known dependence structure of the noise;
- use a step-down algorithm to adapt to $\mathcal{H}_0$.

4.1. Kernel and pivotal function

The starting point is to consider a reference family $\mathfrak{R}$ of the form (13) (with a deterministic size K) , based on thresholds $t_k(\lambda)$, $1 \leq k \leq K$, for some functions $t_k : \lambda \in [0,1] \mapsto t_k(\lambda)$ and then to choose $\lambda = \lambda(\alpha)$ so that the JER control (1) is satisfied.

Definition 4.1. A *(rejection) kernel* is a family of functions $t_k(\lambda)$, $\lambda \in [0,1]$, $1 \leq k \leq K$, such that $K \in \{1, \ldots, m\}$ and for all $k \in \{1, \ldots, K\}$, $t_k(0) = 0$ and $t_k(\cdot)$ is non-decreasing and left-continuous on $[0,1]$. The parameter K is called the *size* of the kernel.

Note that, for a given kernel, when λ is fixed, we refer to $t_k(\lambda)$, $1 \leq k \leq K$, as thresholds. Several choices of kernel are possible as we will see in Section 5. Here, we work with a generic, fixed kernel $t_k(\lambda)$, $\lambda \in [0,1]$, $1 \leq k \leq K$. We denote the generalized inverse of $t_k(\cdot)$ by $t_k^{-1}(y) = \max\{x \in [0,1] \,:\, t_k(x) \leq y\}$, for any $y \in \mathbb{R} \cup \{-\infty, +\infty\}$.

Lemma 4.2. *Consider the location model (15), any $\lambda \in [0,1]$ and any reference family $\mathfrak{R}_\lambda$ based on thresholds $t_k(\lambda)$, $k \in \{1, \ldots, K\}$. Then the error rate (2) can be written as follows: for any $P \in \mathcal{P}$,*

$$\mathrm{JER}(\mathfrak{R}_\lambda, P) = \mathbb{P}_{q \sim \nu_m}\left(\min_{1 \leq k \leq K \wedge m_0} \left\{ t_k^{-1}\left(q_{(k:\mathcal{H}_0)}\right) \right\} < \lambda \right), \tag{21}$$

where $q_{(k:\mathcal{H}_0)}$ denotes the k-th minimum of the set $\{q_i, i \in \mathcal{H}_0\}$.

Proof. By definition, $t_k(\lambda) > q_{(k:\mathcal{H}_0)}$ if and only if $\lambda > t_k^{-1}(q_{(k:\mathcal{H}_0)})$. Also, for all $i \in \mathcal{H}_0$, we have $p_i(X) \geq q_i(X)$ (as defined in Section 2.5) both in the one-sided and two-sided cases. Hence, in view of (14), we obtain

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

which proves the result. □

4.2. Single-step and step-down λ-adjustments

An important consequence of Lemma 4.2 is that JER control (1) can be achieved by choosing λ equal to the α-quantile of the distribution of $\min_{1 \leq k \leq K \wedge m_0} \left\{ t_k^{-1}\left(q_{(k:\mathcal{H}_0)}\right) \right\}$, that is, equal to $\lambda(\alpha, \mathcal{H}_0)$, where for all $\mathcal{C} \subset \{1, \ldots, m\}$, we denoted

$$\lambda(\alpha, \mathcal{C}) = \max\left\{ \lambda \geq 0 \,:\, \mathbb{P}_{q \sim \nu_m}\left(\min_{1 \leq k \leq K \wedge |\mathcal{C}|} \left\{ t_k^{-1}\left(q_{(k:\mathcal{C})}\right) \right\} < \lambda \right) \leq \alpha \right\}. \tag{22}$$

Note that $\lambda(\alpha, \mathcal{C})$ depends on ν_m and on the kernel, although it is not explicitly underlined in the notation for simplicity. Unfortunately, since $\mathcal{H}_0$ is unknown, so is $\lambda(\alpha, \mathcal{H}_0)$. Therefore, a useful property at this stage is that the functional $\mathcal{C} \mapsto \lambda(\alpha, \mathcal{C})$ is nonincreasing:

$$\forall \mathcal{C}, \mathcal{C}' \subset \{1, \ldots, m\}, \text{ with } \mathcal{C} \subset \mathcal{C}', \quad \lambda(\alpha, \mathcal{C}') \leq \lambda(\alpha, \mathcal{C}). \tag{NI}$$

This is a direct consequence of the fact that, pointwise, if $\mathcal{C} \subset \mathcal{C}'$, then for all fixed $k \leq |\mathcal{C}|$, $q_{(k:\mathcal{C}')} \leq q_{(k:\mathcal{C})}$ which in turn implies $t_k^{-1}\left(q_{(k:\mathcal{C}')}\right) \leq t_k^{-1}\left(q_{(k:\mathcal{C})}\right)$ and thus

$$\min_{1\leq k\leq K\wedge|\mathcal{C}'|}\left\{t_k^{-1}\left(q_{(k:\mathcal{C}')}\right)\right\} \leq \min_{1\leq k\leq K\wedge|\mathcal{C}|}\left\{t_k^{-1}\left(q_{(k:\mathcal{C}')}\right)\right\} \leq \min_{1\leq k\leq K\wedge|\mathcal{C}|}\left\{t_k^{-1}\left(q_{(k:\mathcal{C})}\right)\right\}.$$

A consequence of (NI) is that $\lambda(\alpha, \mathbb{N}_m) \leq \lambda(\alpha, \mathcal{H}_0)$ and $\lambda(\alpha, \mathbb{N}_m)$ can be used as a conservative substitute for $\lambda(\alpha, \mathcal{H}_0)$. Now, $\lambda(\alpha, \mathbb{N}_m)$ (often denoted by $\lambda(\alpha)$ for short) can be computed from the distribution of the distribution ν_m. This provides the following result.

Proposition 4.3. *In the framework of Lemma 4.2, consider $\lambda(\alpha) = \lambda(\alpha, \mathbb{N}_m)$ defined by (22). Then the procedure $\mathfrak{R}_{\lambda(\alpha)}$ controls the JER criterion at level α in the sense of (1).*

Above, we have used $\lambda(\alpha, \mathbb{N}_m)$ as a conservative substitute for $\lambda(\alpha, \mathcal{H}_0)$. This induces a loss in the JER control, that is sometimes substantial (see Sections 5.3 and 6 for more details). This loss can be reduced by using $\lambda(\alpha, \widehat{\mathcal{C}})$, where $\widehat{\mathcal{C}}$ comes from the following step-down algorithm.

Algorithm 4.4. (General step-down algorithm)

- initialization: let $\mathcal{C}^{(0)} = \mathbb{N}_m$;
- step $j \geq 1$: compute $\lambda_j = \lambda(\alpha, \mathcal{C}^{(j-1)})$ by using (22) and consider

$$\mathcal{C}^{(j)} = \{i \in \mathbb{N}_m \,:\, p_i(X) \geq t_1(\lambda_j)\}, \tag{23}$$

  If $\mathcal{C}^{(j)} = \mathcal{C}^{(j-1)}$, stop and let $\widehat{\mathcal{C}} = \mathcal{C}^{(j)}$. Otherwise, continue and go to step $j+1$;

Note that, while the t_k's are used in (23) only through $k = 1$, $\widehat{\mathcal{C}}$ depends on all the t_k's through the functional $\lambda(\alpha, \cdot)$.

Proposition 4.5. *In the framework of Lemma 4.2, consider the functional $\lambda(\alpha, \cdot)$ defined by (22) and compute $\widehat{\mathcal{C}}$ by Algorithm 4.4. Then the procedure $\mathfrak{R}_{\lambda(\alpha,\widehat{\mathcal{C}})}$ controls the JER at level α in the sense of (1).*

Proof. Consider the event Ω for which

$$\forall k \in \{1, \ldots, K\}, \quad p_{(k:\mathcal{H}_0)}(X) \geq t_k(\lambda(\alpha, \mathcal{H}_0)), \tag{24}$$

which occurs with probability at least $1-\alpha$ by (22). Now, since $t_1(\lambda(\alpha, \cdot))$ is a non-decreasing function on the subsets of $\mathbb{N}_m$, we have on Ω, for all $j \geq 0$,

$$\mathcal{H}_0 \subset \mathcal{C}^{(j-1)} \;\Rightarrow\; p_{(1:\mathcal{H}_0)}(X) \geq t_1(\lambda(\alpha, \mathcal{C}^{(j-1)})) \;\Rightarrow\; \mathcal{H}_0 \subset \mathcal{C}^{(j)},$$

and thus $\mathcal{H}_0 \subset \widehat{\mathcal{C}}$, which itself entails

$$\forall k \in \{1, \ldots, K\}, \quad p_{(k:\mathcal{H}_0)}(X) \geq t_k(\lambda(\alpha, \widehat{\mathcal{C}})).$$

Since Ω is of probability at least $1-\alpha$, the result is proved. □

Remark 4.6. When we choose $K = 1$, Algorithm 4.4 reduces to the usual FWER controlling step-down algorithm (see, e.g., Romano and Wolf, 2005).

5. Application : two examples of kernel-based reference families

In this section, we apply the methodology presented in the previous section to two particular instances of rejection kernels.

5.1. Linear kernel

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

$$t_k^L(\lambda) = \lambda k/m, \quad \lambda \in [0,1], \quad 1 \le k \le K. \tag{25}$$

Since $(t_k^L)^{-1}(u) = 1 \wedge (\frac{m}{k}u)$, our methodology is to perform a λ-adjustment according to the following relation (see (22)): for all $\mathcal{C} \subset \{1,\dots,m\}$,

$$\lambda^L(\alpha,\mathcal{C}) = \max\left\{\lambda \ge 0 \,:\, \mathbb{P}_{q\sim\mathcal{N}(0,\Sigma)}\left(\min_{1\le k\le K\wedge|\mathcal{C}|}\left\{m\, q_{(k:\mathcal{C})}/k\right\} < \lambda\right) \le \alpha\right\}. \tag{26}$$

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 $\mathfrak{R}^L = (R_1^L(X), \dots, R_K^L(X))$, where

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

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

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

  where $\hat{\mathcal{C}}$ is derived from Algorithm 4.4, used with $\lambda(\cdot) = \lambda^L(\cdot)$ and $t_1(\cdot) = t_1^L(\cdot)$.

Propositions 4.3 and 4.5 ensure that the reference families $\mathfrak{R}^L$ and $\mathfrak{R}^{L,sd}$ both control the JER at level α in the location model. Note that $\mathfrak{R}^L = \mathfrak{R}^0$ (Simes family, see (18)) when $\lambda^L(\alpha,\mathbb{N}_m) = \alpha$ and $K = m$, which arises in the independent case by Proposition 3.1.

Under dependence, what is the magnitude of $\lambda^L(\alpha,\mathbb{N}_m)$? Figure 3 displays $\lambda^L(\alpha,\mathbb{N}_m)$ in the (one-sided) Gaussian ρ-equi-correlated setting, for different values of ρ. The influence of the size K is also illustrated. In a nutshell, we see that the influence of K and ρ is moderate for, say, $\rho \le 0.2$ (a somewhat realistic range for the dependency strength). The lack of sensitivity with respect to K is not surprising because for the linear kernel, only the very first k will be important inside the probability of relation (26). However, the influence of K is more pronounced as ρ gets larger, say $\rho \ge 0.4$ (which is arguably a less realistic range of values for ρ).

5.2. Balanced kernel

Considering a linear kernel is not always appropriate : as already mentioned, under independence and $K = m$, $\mathfrak{R}^L$ corresponds to the Simes reference family $\mathfrak{R}^0$ (18), and thus suffers from a kind of unbalancedness, as underlined in Section 3.3. To address this issue, we introduce another choice for the kernel. For each $k \in \{1,\dots,m\}$, let us define $F_k(x) = \mathbb{P}_{q\sim\nu_m}(q_{(k:m)} \le x)$, $x \in [0,1]$, which is the c.d.f. of $q_{(k:m)}$. The *balanced kernel* (of size K) is then given by

$$t_k^B(\lambda) = F_k^{-1}(\lambda) = \min\{t \in \mathbb{R} \,:\, F_k(x) \ge \lambda\}, \quad \text{with} \quad k \in \{1,\dots,K\}. \tag{29}$$

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, because it ensures $\mathbb{P}(q_{(k:m)} < t_k^B(\lambda)) \le \lambda$ for $1 \le k \le K$. It is straightforward to check that $t_k^B(\cdot)$ fulfills the requirements of Definition 4.1. Since $(t_k^B)^{-1}(x) = F_k(x)$ for all $x \in [0,1]$, the definition of $\lambda(\alpha,\mathcal{C})$ in (22) can be rewritten as follows: for all $\mathcal{C} \subset \mathbb{N}_m$,

$$\lambda^B(\alpha,\mathcal{C}) = \max\left\{\lambda \in [0,1] \,:\, \mathbb{P}_{q\sim\nu_m}\left(\min_{1\le k\le K\wedge|\mathcal{C}|}\left\{F_k\left(q_{(k:\mathcal{C})}\right)\right\} < \lambda\right) \le \alpha\right\}. \tag{30}$$

For each K, this gives rise to two new reference families:

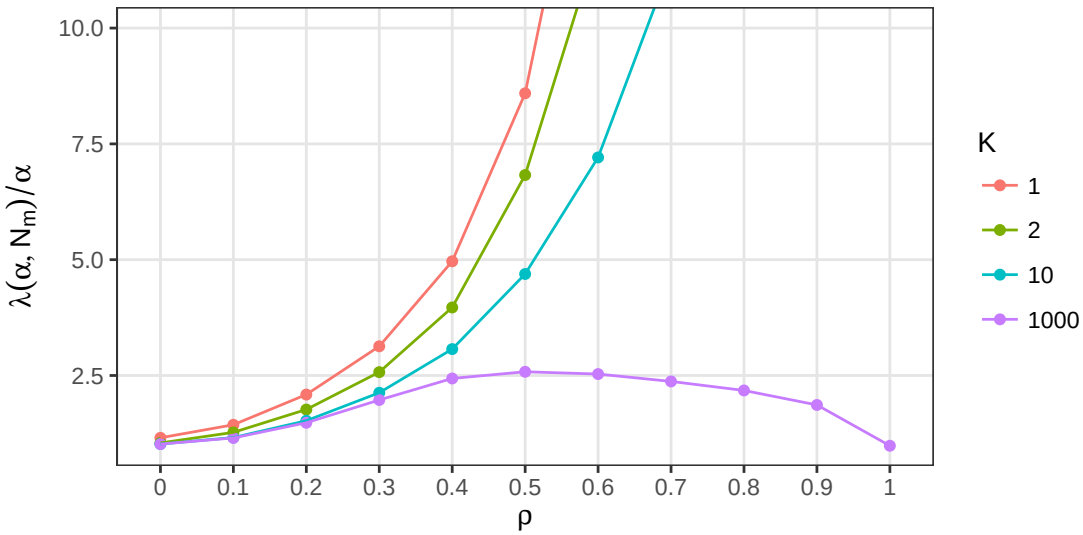

FIG 3. *Influence of the equi-correlation level ρ on the adjustment factor $\lambda^L(\alpha, \mathbb{N}_m)$ for linear kernel. Different values of K are used. $m = 1,000$; $\alpha = 0.2$; $\pi_0 = 1$. $\lambda^L(\alpha, \mathbb{N}_m)$ was estimated based on $B = 10^4$ Monte-Carlo samples of the joint null distribution.*

- 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

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

- 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

$$R_k^{B,sd}(X) = \left\{ i \in \mathbb{N}_m \ : \ p_i < t_k^B(\lambda^B(\alpha, \hat{\mathcal{C}})) \right\} \ , 1 \leq k \leq K, \tag{32}$$

  where $\hat{\mathcal{C}}$ is derived from Algorithm 4.4, used with $\lambda(\cdot) = \lambda^B(\cdot)$ and $t_1(\cdot) = t_1^B(\cdot)$.

Propositions 4.3 and 4.5 ensure that the reference families $\mathfrak{R}^B$ and $\mathfrak{R}^{B,sd}$ both control the JER at level α in the location model. Let us also mention that since the reference family $\mathfrak{R}^{B,sd}$ can appear as quite complex to compute, we provide a full Monte-Carlo approximation scheme in Appendix A.3.

What is the magnitude of $\lambda^B(\alpha, \mathbb{N}_m)$? Since each of the $F_k\left(q_{(k:m)}\right)$ is uniformly distributed on $(0, 1)$, a simple union bound argument provides the following bounds:

$$\alpha/K \leq \lambda^B(\alpha, \mathbb{N}_m) \leq \alpha. \tag{33}$$

Under independence and for $K = m$, Lemma A.2 shows the more accurate upper-bound $\lambda^B(\alpha, \mathbb{N}_m) \leq 1/(\log m)^{1/4}$ (for m large enough). This in particular shows that $\lambda^B(\alpha, \mathbb{N}_m)$ tends to zero when m tends to infinity. However, when the size K is kept fixed, say $K = 10$, (33) ensures that $\lambda^B(\alpha, \mathbb{N}_m)$ is bounded away from zero.

More specifically, Figure 4 shows the influence of ρ and K on the value of $\lambda^B(\alpha, \mathbb{N}_m)$ under (one-sided) Gaussian ρ-equi-correlated dependence. Compared to the linear kernel, we see the sensitivity of $\lambda^B(\alpha, \mathbb{N}_m)$ w.r.t. K and ρ is more substantial. When $\rho = 0$, the value of $\lambda^B(\alpha, \mathbb{N}_m)$ is small for $K = m$ and increases as K becomes smaller, which supports the above theoretical statements. Also, even moderate values of ρ (say, $\rho \leq 0.2$) have a large impact on the value of $\lambda^B(\alpha, \mathbb{N}_m)$.

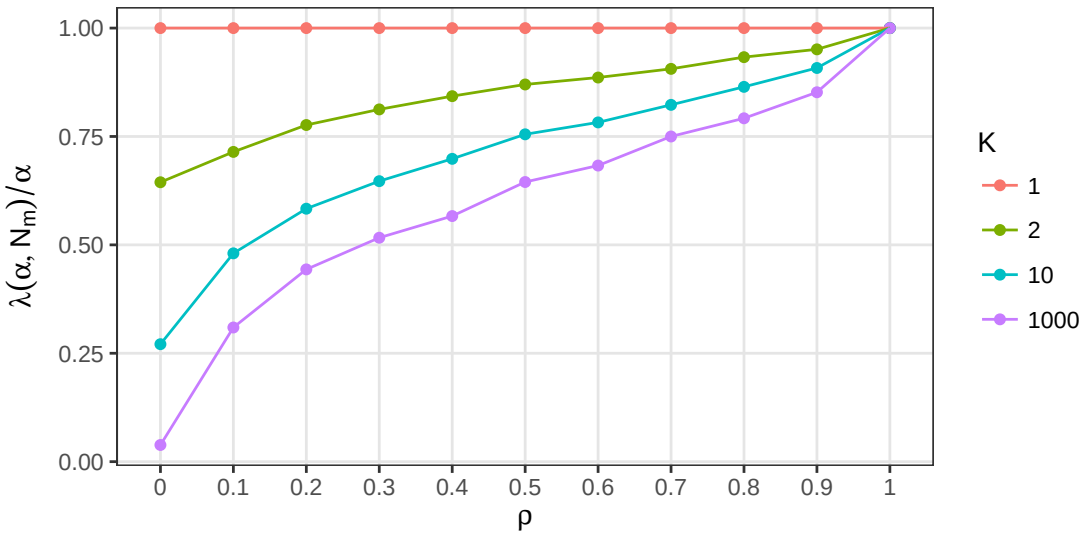

FIG 4. *Influence of the equi-correlation level ρ on the adjustment factor $\lambda^B(\alpha, \mathbb{N}_m)$ for the balanced kernel. Different values of K are used. $m = 1,000$; $\alpha = 0.2$; $\pi_0 = 1$. $\lambda^B(\alpha, \mathbb{N}_m)$ was estimated based on $B = 10^4$ Monte-Carlo samples of the joint null distribution.*

5.3. Effects of step-down algorithm

Remember that the rationale behind the proposed step-down algorithm is that, when $\pi_0 = m_0/m$ is smaller than 1, some of the hypotheses will be rejected at each step, which will improve the value of the λ-adjustement by replacing $\lambda(\alpha, \mathbb{N}_m)$ by $\lambda(\alpha, \widehat{\mathcal{C}})$ for $\widehat{\mathcal{C}}$ smaller than $\mathbb{N}_m$. How large is the magnitude of the improvement, for instance under independence (and $K = m$) ? It turns out that the step-down refinement has a stronger influence for the balanced kernel than for the linear kernel :

- For the linear reference family, we have $\lambda(\alpha, \mathbb{N}_m) = \alpha$ and the family reduces to the Simes family in this case. From (17), the JER achieved is

$$\mathbb{P}\left(\exists k \in \{1, \ldots, m_0\} \ : \ q_{(k:m_0)} < \alpha k/m\right) = \pi_0 \alpha.$$

  As a consequence, the criterion as a linear dependence w.r.t. π_0. The best improvement that the step-down algorithm can provide is thus $\lambda = \alpha/\pi_0$.
- By contrast, for the balanced reference family, the influence of π_0 is more substantial. To see this, remember that $\lambda^B(\alpha, \mathbb{N}_m)$ is calibrated so that

$$\mathbb{P}\Big(\exists k \in \{1, \ldots, m\} \ : \ q_{(k:m)} < t_k^B\left(\lambda^B(\alpha, \mathbb{N}_m)\right)\Big) = \alpha.$$

  while the achieved JER is

$$\mathbb{P}\Big(\exists k \in \{1, \ldots, m_0\} \ : \ q_{(k:m_0)} < t_k^B\left(\lambda^B(\alpha, \mathbb{N}_m)\right)\Big),$$

  where q_i, $i \in \mathbb{N}_m$, are i.i.d. $U(0,1)$. We show in Appendix A that, for all $\lambda \leq 0.5$, the probability $\mathbb{P}(p_{(k:m_0)} \leq t_k^B(\lambda))$ decreases exponentially with $k(1-\pi_0)^2$ and thus becomes small when k is "not small", see (43). A consequence is that, when π_0 is bounded away from (but however close to) 1, the achieved JER tends to zero at a $(\log m)^{1/8}$ rate, see (46). This shows that the influence of $\pi_0 < 1$ on the achieved JER is substantial; this makes the potential improvement of the step-down algorithm all the more important. Of course, the

amplitude of this phenomenon decreases as π_0 gets closer to 1, but our numerical experiments suggest that it still exists for cases where $\pi_0 \approx 1$ (sparsity). This is a new feature of step-down type algorithms to the best of our knowledge.

6. Numerical experiments

We report numerical experiments to illustrate the performance of the procedures defined in the previous section. In our setting, the test statistics are distributed as multivariate equi-correlated Gaussian $\mathcal{N}(\mu, \Sigma)$, where the mean vector μ is unknown and the covariance Σ satisfies $\Sigma_{i,i} = 1$ and $\Sigma_{i,j} = \rho$ for $1 \leq i \neq j \leq m$, where ρ is *known*. The number of tested hypotheses is set to $m = 1,000$. We set $\mu_i = 0$ for all $i \in \mathcal{H}_0$, and $\mu_i = \overline{\mu}$ for all $i \in \mathcal{H}_1$, where $\overline{\mu} > 0$ quantifies the signal-to-noise ratio (SNR) of the setting.

6.1. JER control

The target JER level is set to $\alpha = 0.25$. In order to study a wide range of scenarios from sparse to non-sparse signals, we have considered the following range of parameter values:

- $\rho \in \{0, 0.2, 0.4\}$;
- $\pi_0 \in \{0.8, 0.9, 0.99\}$, corresponding to $m_1 \in \{200, 100, 10\}$;
- $\overline{\mu} \in \{0, 1, 2, 3, 4, 5\}$

For each setting, we report the proportion $\hat{p}$ of simulation runs (out of a total of $n = 1,000$ runs) for which the bound $|R_k(X) \cap \mathcal{H}_0(P)| \geq k$ does not hold for at least one $k \in \{1, \ldots, K\}$. This proportion is an estimate of the achieved JER. The results are summarized by Figure 5 for the linear kernel, and by Figure 6 for the balanced kernel. Each figure is a 3×3 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 of these panels, 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 kernel, the level $\pi_0\alpha$ is represented by a horizontal dotted line. In both figures, each color corresponds to a different reference family:

- *Simes* : the reference family $\mathfrak{R}^0$ given by (18) and obtained by the Simes inequality;
- *Linear Single Step* : the single-step linear reference family $\mathfrak{R}^L$ (of size $K = m$), given by (27) and obtained from a single-step λ-adjustment $\lambda(\alpha, \mathbb{N}_m)$;
- *Linear Step-down* : the step-down linear reference family $\mathfrak{R}^{L,sd}$ (of size $K = m$), given by (28) and obtained from a step-down λ-adjustment $\lambda(\alpha, \widehat{\mathcal{C}})$;
- *Linear Oracle* : same as linear step-down with an oracle λ-adjustment $\lambda(\alpha, \mathcal{H}_0)$;
- *Balanced Single Step* : the single-step balanced reference family $\mathfrak{R}^B$ (of size $K = m$), given by (31) and obtained from a single-step λ-adjustment $\lambda(\alpha, \mathbb{N}_m)$;
- *Balanced Step-down* : the step-down balanced reference family $\mathfrak{R}^{B,sd}$ (of size $K = m$), given by (32) and obtained from a step-down λ-adjustment $\lambda(\alpha, \widehat{\mathcal{C}})$;
- *Balanced Oracle* : same as balanced step-down with an oracle λ-adjustment $\lambda(\alpha, \mathcal{H}_0)$.

The following comments can be made from Figures 5 and 6. As expected, JER is controlled at the target level α in all situations, and the Oracle family yields exact JER control, up to sampling fluctuations. As discussed in Section 3.2, the Simes reference family with parameter α yields a JER equal to $\pi_0\alpha$ under independence ($\rho = 0$), while it is more conservative under positive dependence $\rho > 0$ (Figure 5). The (single-step) λ-adjustment procedure described in Section 4.2 yields JER control at $\pi_0\alpha$ in all dependency settings considered (we recall that the dependency, which is parametrized by ρ, is assumed to be known here). Finally, as the signal-to-noise ratio $\overline{\mu}$ gets larger, the step-down refinement proposed in Section 4.4 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

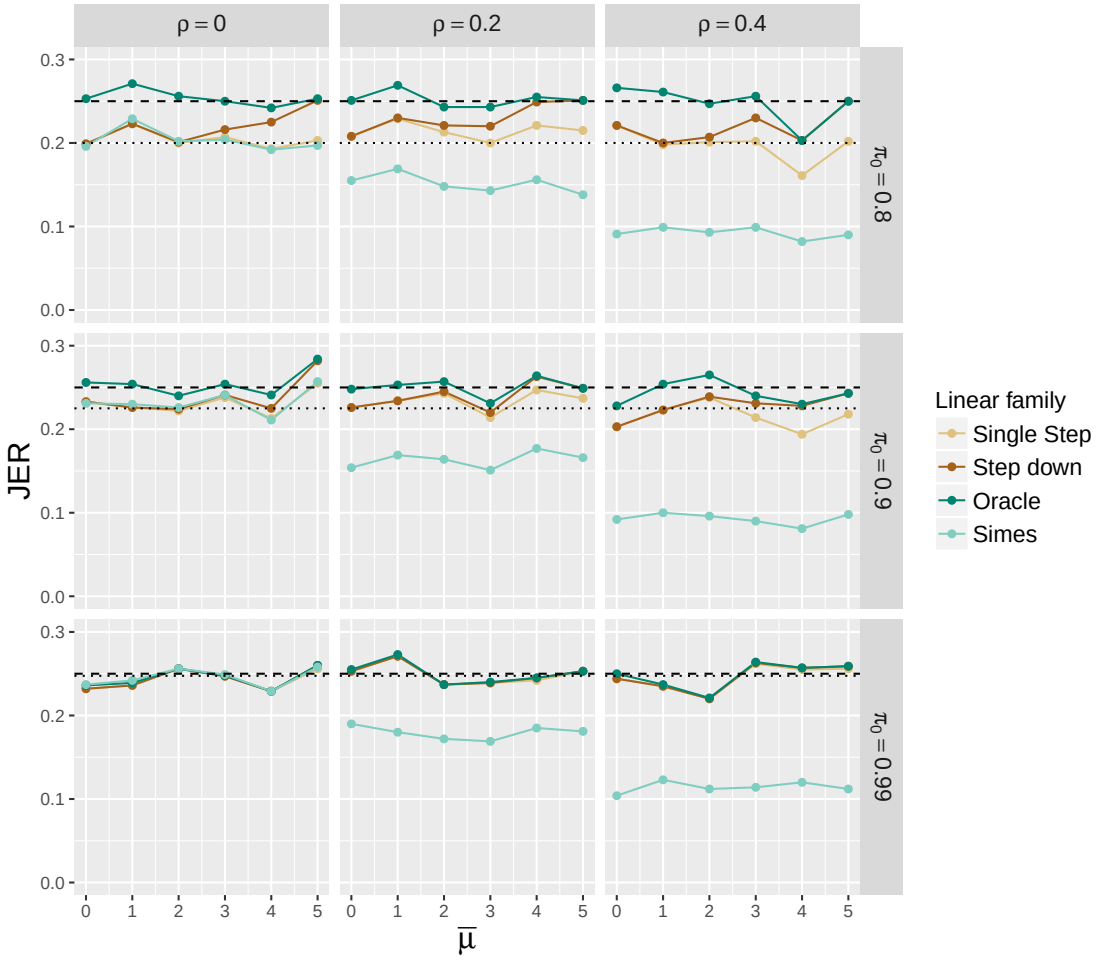

FIG 5. *JER control based on the linear kernel for equi-correlated test statistics.*

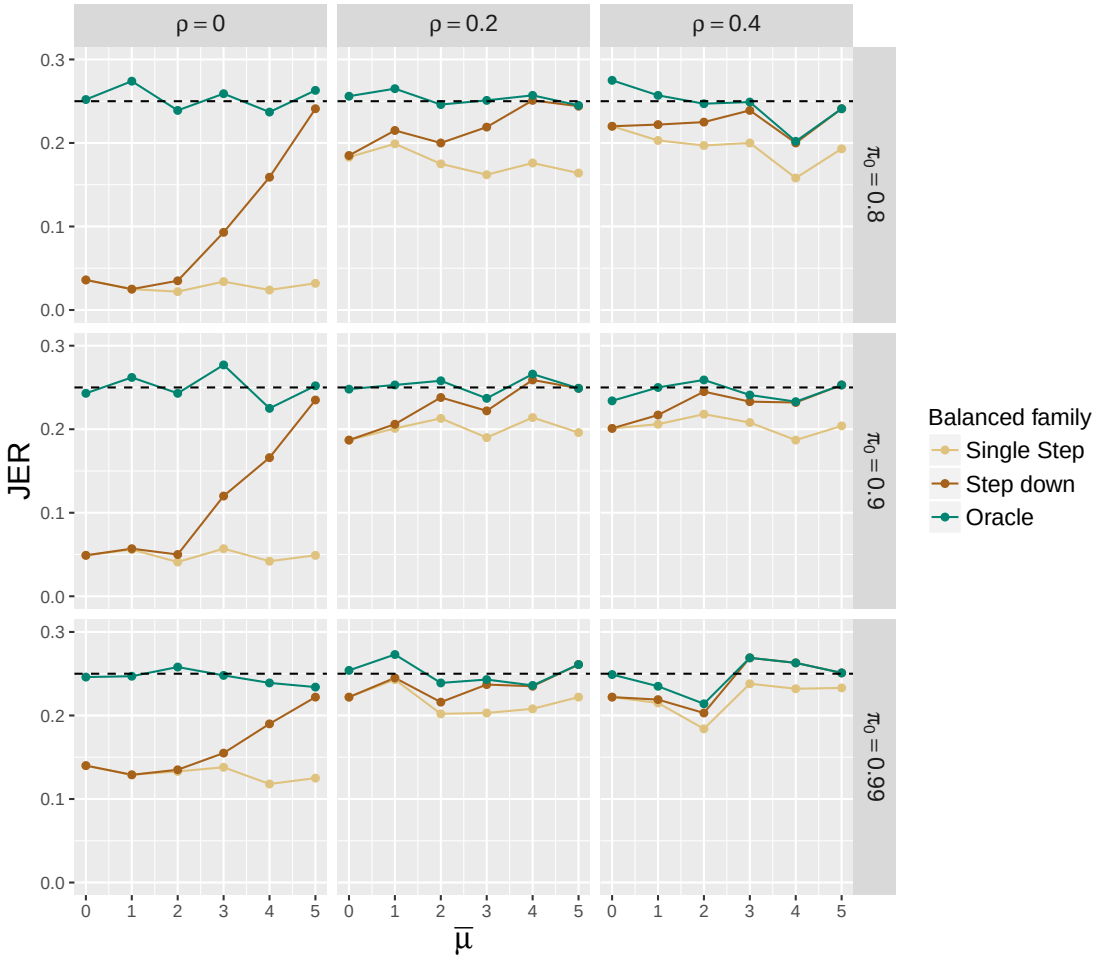

FIG 6. *JER control based on the balanced kernel for equi-correlated test statistics.*

to $m_1 = 10$, the single-step procedure is already quite sharp and essentially indistinguishable from its Oracle counterpart.

For the balanced kernel, as summarized by Figure 6, the single step λ-adjustment leads to much more conservative JER control than for the linear kernel, especially under independence or weak dependence. This is a numerical illustration of the effect discussed in Section 5.3: while the JER achieved by the single step λ-adjustment of the Simes family is close to $\pi_0\alpha$, the JER achieved by the single step λ-adjustment of the balanced family is quite conservative, 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-adjustment of the balanced family is of the order of $\alpha/2$. Therefore, a step-down adjustment is required in order to catch up with the target JER level. When the signal-to-noise ratio is very large, the JER of the step-down procedure is indeed much closer to α.

The JER control in the above-described experiments has been obtained with $K = m$. We have performed further numerical experiments to illustrate the influence of K. The results are consistent with the observations made in Section 5 (Figures 3 and 4): for the linear family, the influence of K on the achieved JER is negligible for the considered range of values of ρ. For the balanced family, the influence of K is more substantial. We report in Figure 7 the results obtained for the balanced family with $K = 10$. The JER control offered by the balanced family with $K = 10$ is much less

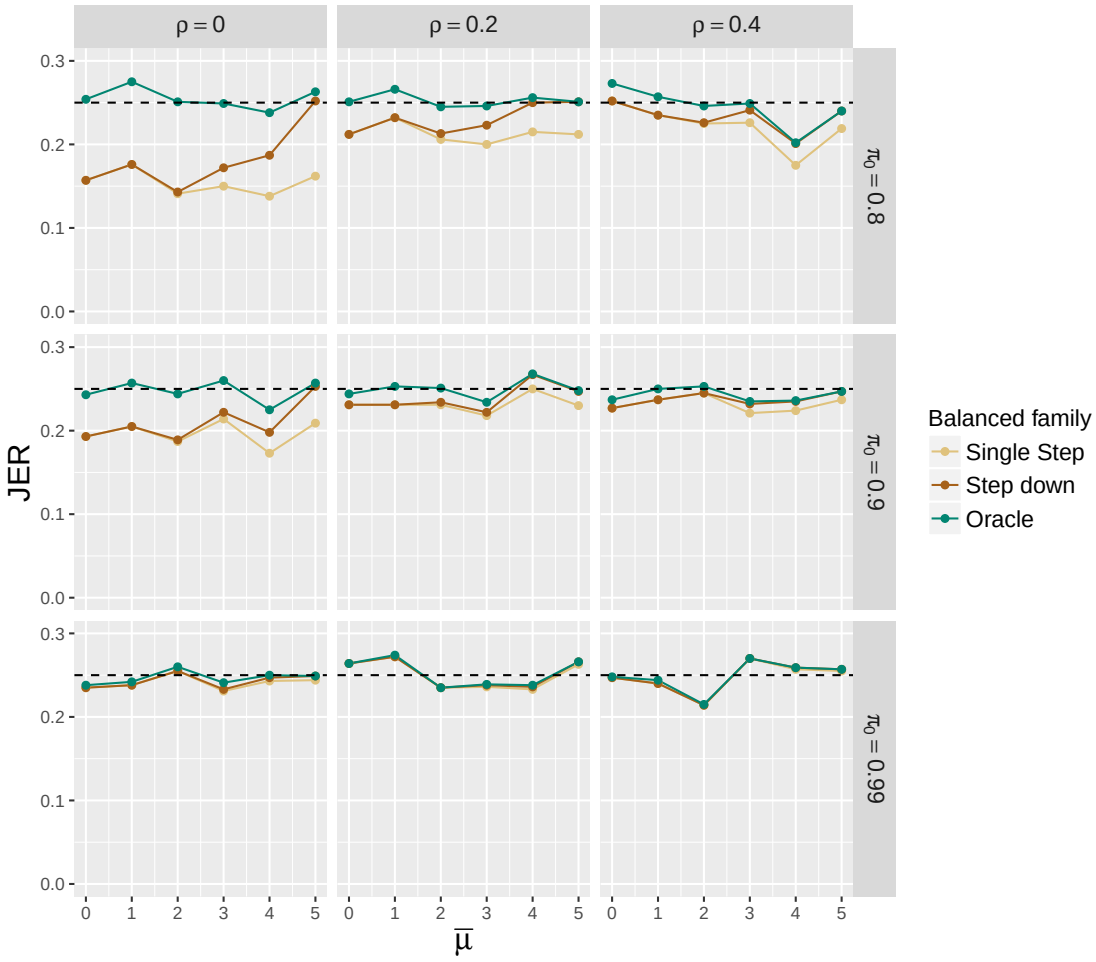

FIG 7. *JER control based on the balanced kernel for equi-correlated test statistics, with $K = 10$.*

conservative than with $K = m$ (Figure 6), even for the single step λ-adjustment. The question of how to choose K is briefly mentioned in the Discussion (Section 9).

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 $\mu_i = \overline{\mu}$ spans a range of values between 0 and $\overline{\mu}$, and/or where the test statistics have a (known) Toeplitz covariance, for which $\Sigma_{i,j} = |i-j|^\theta$, where $\theta \in \{-2, -1, -0.5, -0.2\}$ controls the range of dependency. Smaller values of θ correspond to short-range dependency while larger values of θ correspond to longer-range dependency. Because the results obtained for both types of signals and for both types of dependency are qualitatively similar, we have only reported the results for the parameter combination: constant signal/equi-correlated dependency in this manuscript.

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 consider the following reference families:

- Linear (K) : the reference family defined with a linear kernel in its step-down form (28) ($K = m$ or $K = 10$);
- Balanced (K) : the reference family defined with a balanced kernel in its step-down form (32) ($K = m$ or $K = 10$);

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 $R \subset \{1, \ldots, m\}$ (possibly data dependent),

$$\mathrm{Pow}(\mathfrak{R}, P) = \mathbb{E}\left(\frac{\overline{S}_{\mathfrak{R}}(R))}{|R \cap \mathcal{H}_1(P)|} \,\middle|\, |R \cap \mathcal{H}_1(P)| > 0 \right). \tag{34}$$

The following selected sets $R \subset \{1, \ldots, m\}$ are considered:

(a) $R = \mathbb{N}_m$. In this case, the averaged power $\mathrm{Pow}(\mathfrak{R}, P)$ measures the (relative) performance of $\overline{S}_{\mathfrak{R}}(\mathbb{N}_m)$ as an estimator of $m_1(P) = |\mathcal{H}_1(P)|$;
(b) $R_0 = \{i \in \{1, \ldots, m\} : p_i \leq 0.05\}$, 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 8. The average power of the Simes family (light green) and of the reference families obtained by single step and step-down λ-adjustment of the linear kernel (dark green) are almost identical. This is consistent with the results displayed in the first column of Figure 5, where the three families achieve very similar JER levels for $\overline{\mu} \leq 4$ when $\rho = 0$. Overall, the averaged power obtained from the balanced kernel are substantially larger than the averaged power obtained from the linear kernel. The only situation where the linear kernel 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 8 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 kernel provides a better estimator of $m_1(P) = |\mathcal{H}_1(P)|$ than the linear kernel. 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 in Section 9.

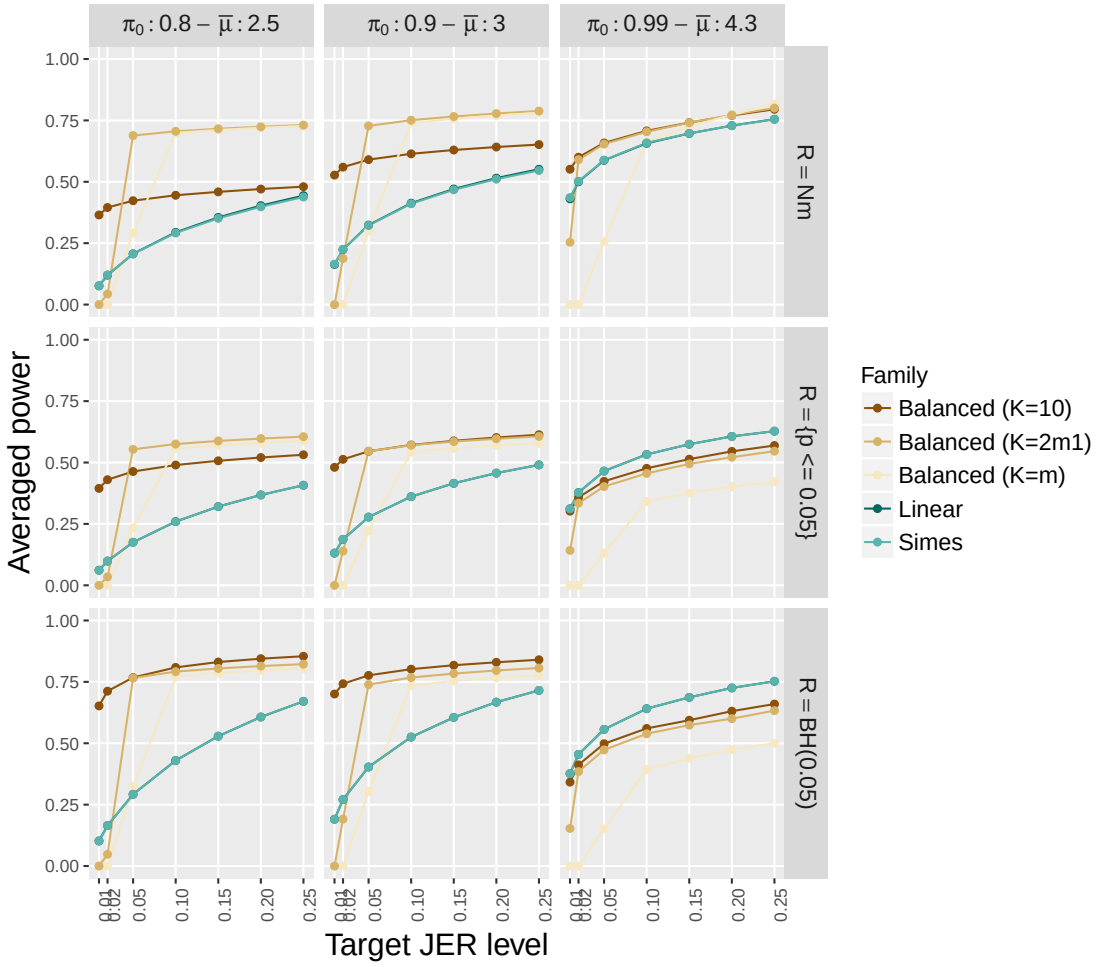

FIG 8. *Averaged power of JER controlling procedures for independent test statistics.*

7. Relation to higher criticism and detection power optimality

In a nutshell, we show in this section that, as a detection procedure, $\mathfrak{R}^B$ shares some similarity with the calibration of the higher criticism (HC for short) method of Donoho and Jin (2004), DJ04 for short. By contrast, $\mathfrak{R}^L$ (with $K = m$), which is equal to $\mathfrak{R}^0$ in the setting of this section, is connected to the procedure of Benjamini and Hochberg (1995), BH for short. This induces specific power properties.

In this section, we evaluate the power of a reference family through its detection capability of any false null:

$$\mathrm{Pow}^\star(\mathfrak{R}, P) = \mathbb{P}(\overline{S}_{\mathfrak{R}}(\mathbb{N}_m) \geq 1) = \mathbb{P}\left(\exists k \in \{1, \ldots, K\} \,:\, |R_k| \geq k\right). \tag{35}$$

Note that this can be seen as the power of the single test rejecting the null H_0: "$\forall i \in \mathbb{N}_m, H_{0,i}$ is true" if there exists $k \in \{1, \ldots, m\}$ such that $|R_k| \geq k$. With respect to this criterion, and in a special regime, we show in this section that $\mathfrak{R}^B$ is optimal, while $\mathfrak{R}^L$ is suboptimal.

Note that the step-down algorithm does not give any improvement in terms of detection power: the step-down can potentially make the sets R_k in the reference family larger in comparison to the single-step procedure, but by construction such an improvement can only take place if $|R_1| \geq 1$ in the first place for the single-step procedure (which is the first iteration of the step-down). Hence, we focus on the single-step versions in this section.

7.1. Framework

We consider the location model (15) in the Gaussian independent one-sided framework, with the special setting considered in DJ04 where the trueness/falseness of the null hypotheses are randomized with a distribution belonging to some sparse regime. Specifically, we consider the hierarchical model where H_i are i.i.d. $\mathcal{B}(\pi_{1,m})$ and the p-values are independent conditionally on the H_i's, with

- $p_i(X) \mid H_i = 0 \sim U(0,1)$;
- $p_i(X) \mid H_i = 1$ has for c.d.f. $F_{1,m}(t) = \overline{\Phi}(\overline{\Phi}^{-1}(t) - \mu_m)$.

Hence, overall, the p-values $(p_i, i \in \mathbb{N}_m)$ are i.i.d. and of common c.d.f. $G_{1,m}(t) = \pi_{0,m} t + \pi_{1,m} F_{1,m}(t)$, where $\pi_{0,m} = 1 - \pi_{1,m}$. The parameters $\pi_{1,m}, \mu_m$ are taken in the asymptotic range where $\pi_{1,m} = m^{-\beta}$ and $\mu_m = \sqrt{2r \log m}$ for two parameters $\beta \in (1/2, 1)$ and $r \in (0, 1)$.

Let us also recall the optimal asymptotic detection boundary:

$$\rho^\star(\beta) = \begin{cases} \beta - 1/2 & \text{if } \beta \in (1/2, 3/4]; \\ (1 - \sqrt{1-\beta})^2 & \text{if } \beta \in (3/4, 1). \end{cases} \tag{36}$$

Note that for all $\beta \in (1/2, 3/4]$, the range where $\beta - 1/2 \leq r \leq (1 - \sqrt{1-\beta})^2$ is usually referred to as sparse/weak, that is, with sparsity and low signal strength. This regime is of interest because DJ04 showed that, in this regime, BH has asymptotically no detection power while HC has optimal asymptotic power.

7.2. Test statistic of the balanced detection procedure

By definition, $\mathfrak{R}^B$ makes a detection if there exists k such that $p_{(k:m)} < t_k^B(\lambda^B(\alpha))$. Furthermore, from (33) and (42), we have the lower bound

$$t_k^B(\lambda^B(\alpha)) \geq \frac{k}{m+1} - \left\{ \frac{k}{m+1} \left(1 - \frac{k}{m+1}\right) \right\}^{1/2} m^{-1/2} \left(4 \log(K/\alpha)\right)^{1/2}. \tag{37}$$

Hence, $\mathfrak{R}^B$ makes a detection whenever the test statistic

$$\max_{1\leq k\leq K}\left\{m^{1/2}\frac{\frac{k}{m+1}-p_{(k:m)}}{\left\{\frac{k}{m+1}\left(1-\frac{k}{m+1}\right)\right\}^{1/2}}\right\}$$

exceeds $(4\log(K/\alpha))^{1/2}$. This is close to the higher criticism procedure of DJ04.

Remark 7.1. Note that in the definition of Higher Criticism considered in DJ04, the authors have similarly restricted the range of the indices considered to $\{1,\ldots,\alpha_0 m\}$, that is, $\alpha_0 m$ plays the a similar role as K does here. This is useful to tune the power detection ability, see Appendix D.

7.3. Optimality results

By adapting the proof of DJ04, we can show the following result (see Appendix B.2 for a proof):

Theorem 7.2. *Consider the asymptotic setting of Section 7.1. Then, the two following facts hold.*

(i) Consider any family $\mathfrak{R}$ with thresholds t_k, $1\leq k\leq m$, that controls the JER at level α in the sense

$$\mathbb{P}(\exists k\in\{1,\ldots,m\}\,:\,U_{(k:m)}\leq t_k)\leq\alpha,$$

for $U_i, i\in\mathbb{N}_m$ i.i.d. uniformly distributed on $(0,1)$. Then we have

$$\limsup_m Pow^\star(\mathfrak{R},P)\leq\alpha$$

whenever P is such that $r<\rho^\star(\beta)$.

(ii) For the balanced family $\mathfrak{R}^B$ with $K=m$, we have $Pow^\star(\mathfrak{R}^B,P)\to 1$ whenever P is such that $r>\rho^\star(\beta)$.

By contrast, the family $\mathfrak{R}^L$ is sub-optimal, as we now show. It will make a detection if there exists k such that $p_{(k:m)}<\alpha k/m$, that is, if the Benjamini-Hochberg procedure rejects at least one null hypothesis. The following result is in fact a reformulation of Theorem 1.4 in Donoho and Jin (2004); its proof is given in Appendix B.3 for completeness.

Theorem 7.3. *Consider the asymptotic setting of Section 7.1. Then the linear reference family $\mathfrak{R}^L$ satisfies the following:*

(i) for $r>(1-\sqrt{1-\beta})^2$, $\lim Pow^\star(\mathfrak{R}^L,P)=1$;
(ii) for $r<(1-\sqrt{1-\beta})^2$, $\limsup Pow^\star(\mathfrak{R}^L,P)\leq\alpha$.

From an intuitive point of view, the threshold is $\alpha k/m=k/m-(1-\alpha)k/m$, so the deviation term is not of the correct order. This implies a lack of detection power which makes this procedure miss the optimal boundary.

Let us finally emphasize that the domination of the balanced family/HC w.r.t. the linear family/BH in terms of detection power is less obvious for a moderate value of m, as illustrated in the numerical experiments of Appendix D with $m=1,000$. This suggests that the asymptotical regime described in Theorems 7.2 and 7.3 is not fully reached for such a value of m (while it seems reached for $m=10^6$ in DJ04).

8. Relation to Goeman and Solari (2011)

To the best of our knowledge, the only existing user-agnostic post hoc approach to multiple testing is the method of Goeman and Solari (2011) (GS11 below for short), which served as an inspiration for the present work. This method is based on closed testing (Marcus et al., 1976), which relies, in principle, on testing all 2^m-1 possible intersections between m hypotheses. In this section, we discuss connections between that work and ours.

8.1. Closed testing principle and GS approach

We come back to the general setting of Section 2.1. For any subset $I \subset \mathbb{N}_m$, define the associated intersection hypothesis as $H_{0,I} = \bigcap_{i \in I} H_{0,i}$. Therefore, $H_{0,I}$ is true iff $P \in H_{0,i}$ for all $i \in I$, or equivalently iff $I \subset \mathcal{H}_0(P)$. We will often informally identify the index subset I and the corresponding intersection hypothesis $H_{0,I}$ in the text to simplify statements. Assume that for any index subset I, the intersection null $H_{0,I}$ can be tested by a so-called *local test* $\phi_I(X) \in \{0,1\}$ of (individual) level α. From the collection of intersection hypotheses that are rejected by their respective local tests, the classical closed testing principle (Marcus et al., 1976) extracts a subcollection[1] $\mathfrak{R}$ of subsets of $\mathbb{N}_m$, such that rejecting all intersection hypotheses of the collection $\mathfrak{R}$ has now controlled family-wise error rate at level α.

Equivalently, the complementary $\mathcal{X} = \mathfrak{R}^c$, i.e. the collection of intersection hypotheses not rejected by the closed testing procedure, satisfies for all $P \in \mathcal{P}$:

$$\mathbb{P}_{X \sim P}\Big(\forall I \subset \mathcal{H}_0(P), I \in \mathcal{X}\Big) \geq 1 - \alpha. \tag{38}$$

From this, GS11 construct a user-agnostic post hoc bound in the following way: on the event (38), for any arbitrary $R \subset \{1, \ldots, m\}$, $I_0 = R \cap \mathcal{H}_0(P)$ is a subset of $\mathcal{H}_0(P)$ and thus must be included in the collection $\mathcal{X}$. This entails that (PH_α) is satisfied with

$$\overline{V}_\alpha^{GS}(R) = \max\{|I| \,:\, I \in \mathcal{X}, I \subset R\}. \tag{39}$$

8.2. Closed testing as a particular JER control

We now justify that the inequality (38) resulting from closed testing can be seen as a JER control. As a result, the GS11 post hoc bound (39) can be seen as a particular case of the JER post hoc approach developed here, and more precisely a particular case of the optimal bound V^* defined in (4).

For this, consider $\mathfrak{R} = (R_1, \ldots, R_K)$ returned by the closed testing procedure as in Section 8.1. The closed testing principle inplies that $\mathfrak{R}$ is closed by the superset operation, i.e., $I \in \mathfrak{R}$ implies $\forall J \supset I,\ J \in \mathfrak{R}$. Now let $\zeta(k) := |R_k| - 1$, $1 \leq k \leq K$. Finally, we have

$$\begin{aligned}
\Big\{\forall I \subset \mathcal{H}_0(P), I \in \mathcal{X}\Big\}^c &= \Big\{\exists I \subset \mathcal{H}_0(P), I \in \mathfrak{R}\Big\} \\
&= \Big\{\mathcal{H}_0(P) \in \mathfrak{R}\Big\} \\
&= \Big\{\exists k \in \{1, \ldots, K\} \,:\, |\mathcal{H}_0(P) \cap R_k| = |R_k|\Big\} \\
&= \Big\{\exists k \in \{1, \ldots, K\} \,:\, |\mathcal{H}_0(P) \cap R_k| > \zeta(k)\Big\} = \mathcal{E}(\mathfrak{R}, \mathcal{H}_0)^c.
\end{aligned}$$

Concerning the post hoc bounds, note that

$$\begin{aligned}
\mathcal{A}(\mathfrak{R}) = \{A \subset \mathbb{N}_m :\ \forall k \in \mathbb{N}_K\ |R_k \cap A| \leq \zeta_k\} &= \{A \subset \mathbb{N}_m :\ \forall k \in \mathbb{N}_K\ |R_k \cap A| \leq |R_k| - 1\} \\
&= \{A \subset \mathbb{N}_m :\ \forall k \in \mathbb{N}_K\ R_k \not\subset A\} \\
&= \mathfrak{R}^c = \mathcal{X},
\end{aligned}$$

where we have used the fact that $\mathfrak{R}$ is closed by superset operation. Hence

$$V^*_{\mathfrak{R}}(R) = \max_{A \in \mathcal{A}(\mathfrak{R})} |R \cap A| = \max_{A \in \mathcal{X}} |R \cap A| = \max_{A \in \mathcal{X}, A \subset R} |A| = \overline{V}_\alpha^{GS}(R)\,,$$

since $\mathcal{X}$ is closed by subset operation.

[1] The rejected collection $\mathfrak{R}$ is the maximal subcollection of hypotheses $H_{0,I}$ rejected by their local test that is closed by the superset operation on their indices. In other words, the index set I corresponding to intersection hypothesis $H_{0,I}$ is selected to be included in the rejected collection $\mathfrak{R}$ by the closed testing principle iff $\phi_I(X) = 1$, as well as all $\phi_{I'}(X) = 1$ for all $I \subset I'$.

8.3. Shortcut as using a particular reference family

However, as recognized by GS11, computation of the closed testing output $\mathfrak{R}$ is (in general) not feasible when m is larger than a few dozens. And even if $\mathfrak{R}$ can be computed, we have shown that the calculation of the post hoc bound (39) is itself NP-hard in a generic setting. To circumvent this complexity issue, a less time-consuming conservative version of the bound (39) has been proposed by GS11 for Simes-type local tests:

$$\phi_I(X) = \mathbf{1}\left\{\exists i \in \{1, \ldots, |I|\} : p_{(i:I)} \leq c_i^{|I|}\right\}, \quad I \subset \mathbb{N}_m, \tag{40}$$

with the assumptions $c_i^\ell \leq c_i^k$, for $\ell \geq k$ and $c_i^\ell \leq c_j^\ell$, for $i \leq j$. The corresponding bound takes the form:

$$\overline{V}_\alpha^{GS}(R) = |R| - (1 + \max\{S_r, 1 \leq r \leq |R|\}) \vee 0, \tag{41}$$

where $S_r = \max\{0 \leq s \leq r-1 : p_{(r:R)} \leq c_{r-s}^m\}$ (with $\max \emptyset = -\infty$). We argue below in Lemma 8.2 that this bound is in fact equivalent to the post hoc bound $\overline{V}_{\mathfrak{R}}(R)$ defined in (6), for the family $\mathfrak{R} = (R_k, \zeta_k := k-1)_{1 \leq k \leq m}$ defined by

$$R_k = \{i \in \mathbb{N}_m : p_i \leq c_k^m\}, \quad 1 \leq k \leq m.$$

The next lemma establishes that JER control holds for this family:

Lemma 8.1. *Assume that the tests $(\phi_I)_{I \subset I_m}$ form a family of local tests at level α for the considered model, i.e., for any $P \in H_{0,I}$, it holds $\mathbb{P}_{X \sim P}(\phi_I(X) = 1) \leq \alpha$. Then joint control of the k-FWER of R_k at level α, uniformly over $k \in \mathbb{N}_m$, holds; in other words, equation (1) holds for the reference family $\mathfrak{R} = (R_k, \zeta_k := k-1)_{1 \leq k \leq m}$.*

Proof. For any given distribution P in the model, we have for $I = \mathcal{H}_0 = \mathcal{H}_0(P)$ the local test control

$$\mathbb{P}_{X \sim P}\left[\exists k \leq m_0 : p_{(k:\mathcal{H}_0)} \leq c_k^{m_0}\right] \leq \alpha\,,$$

implying by the monotonicity assumption $c_i^\ell \leq c_i^j$ for $\ell \geq j$:

$$\mathbb{P}_{X \sim P}\left[\exists k \leq m_0 : p_{(k:\mathcal{H}_0)} \leq c_k^{m}\right] \leq \alpha\,.$$

As we argued in Section 2.4, this is equivalent to $\mathrm{JER}(\mathfrak{R}, P)$ for the threshold-based reference family $\mathfrak{R} = (R_k, k-1)_{1 \leq k \leq m}$ using thresholds $t_k := c_k^m, k \in \mathbb{N}_m$, see (13)-(14). □

Now, we establish the equivalence of the two bounds:

Lemma 8.2. *For any $R \subset \mathbb{N}_m$, $\overline{V}_{\mathfrak{R}}(R) = \overline{V}_\alpha^{GS}(R)$.*

Proof. The result comes from

$$\begin{aligned}
\max\{S_r, 1 \leq r \leq |R|\} &= \max\{s \geq 0 : \exists r \text{ s.t. } 1 \leq r \leq |R| \text{ and } 0 \leq s \leq r-1 \text{ and } p_{(r:R)} \leq c_{r-s}^m\} \\
&= \max\{s \geq 0 : \exists r \text{ s.t. } s+1 \leq r \leq |R| \text{ and } |R_{r-s} \cap R| \geq r\} \\
&= \max\{s \geq 0 : \exists r \leq m \text{ s.t. } s+1 \leq r \leq |R_{r-s} \cap R|\} \\
&= \max\{s \geq 0 : \exists k \leq m \text{ s.t. } 1 \leq k \leq |R_k \cap R| - s\} \\
&= \max\{|R_k \cap R| - k, 1 \leq k \leq m\},
\end{aligned}$$

by letting $k = r - s$. □

A consequence is that using this GS11 shortcut, which is again the only computable way to use close-testing for m large (to our knowledge), reduces to the post hoc bound studied in this paper. (A remark pointing in that direction is also mentioned at the end of Section 4.2 of GS11.) In particular, for $c_k^m = \alpha k/m$, the reference family reduces to the Simes reference family $\mathfrak{R}_0$ (18), and the bound $\overline{V}_\alpha^{GS}(R)$ has the simple equivalent form given by (19).

9. Discussion

Beyond the location model We have investigated the performance of the proposed approach in a location model with a known distribution of the noise, see (15). Our theoretical findings can be easily extended to any statistical model $(P_\theta, \theta \in \Theta)$, with $\Theta \subset \mathbb{R}^m$, where

- we aim at testing $H_{0,i}$: "$\theta_i = \theta_i^0$" against $H_{1,i}$: "$\theta_i \neq \theta_i^0$", for some fixed $\theta^0 \in \mathbb{R}^m$;
- there exist statistics $S_i(X)$, $i \in \mathbb{N}_m$, and a $\mathbb{R}$-valued pivotal function Ψ such that

$$(\Psi(S_i(X), \theta_i))_{i \in \mathbb{N}_m}$$

  has a known distribution Q on $\mathbb{R}^m$, not depending on θ.

Then, we can build a test by rejecting each $H_{0,i}$ if $\Psi\left(S_i(X), \theta_i^0\right)$ is large (say). More specifically, denoting the i-th marginal of Q as Q_i with c.d.f. F_i, the p-values can be set as $p_i(X) = 1 - F_i\left(\Psi\left(S_i(X), \theta_i^0\right)\right)$. Lemma 4.2 and thus the whole λ-adjustment process easily extends to that situation because $(p_i(X))_{i \in \mathcal{H}_0}$ follows the distribution of $(Q_i)_{i \in \mathcal{H}_0}$, where Q is known.

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-adjustment): choosing a smaller value for K can yield a tighter post hoc bound. This effect is particularly marked in the case of the balanced kernel 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 $\mathfrak{R} = (R_i(X))_{1 \leq i \leq K}$ is a reference family (using $\zeta_i = i - 1$) satisfying JER control. Consider any set $R \subset \mathbb{N}_m$ such that $\overline{V}_{\mathfrak{R}}(R) \leq K_0 \leq K$. Then we have $\overline{V}_{\mathfrak{R}}(R) = \overline{V}_{\mathfrak{R}^{(K_0)}}(R)$, where* $\mathfrak{R}^{(K_0)} = (R_i(X))_{1 \leq i \leq K_0}$. 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.
- 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.

Step-down algorithm The principle of the step-down Algorithm 4.4 is to approach the oracle value $\lambda(\alpha, \mathcal{H}_0)$ by iterative approximations $\lambda(\alpha, \widehat{\mathcal{C}})$. Here the kernel $t_k(\cdot)$ is fixed once for all. A seemingly natural extension is to allow the kernel $t_k(\cdot, \mathcal{C})$ to also depend on subsets $\mathcal{C} \subset \mathbb{N}_m$ and to apply the step-down algorithm to the kernel as well as λ, that is, consider at each step $t_k(\cdot, \widehat{\mathcal{C}})$, then apply the λ-adjustment step. For instance, for the balanced rejection kernel, one could define $t_k^B(\lambda, \mathcal{C})$ as the λ-quantile of $q_{k:\mathcal{C}}$. 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.

Permutation-based approaches The λ-adjustment methodology presented in this paper extensively uses the fact that the distribution of the noise is known. When this is not the case, permutation-based approaches (also known as exact tests) provide a good alternative in many cases. However, the non-asymptotical theoretical study of permutation-based procedures is challenging for JER control. A general framework to define and analyze such procedures is to assume

that the joint distribution of the null p-values is invariant under transformation of the data through a transformation group $\mathcal{G}$. The principle is to replace the distribution of the null p-values by their empirical distribution under transformation of the data by a random element of $\mathcal{G}$; see, e.g., Romano and Wolf (2005) in the context of FWER control. It seems natural to aim at extending this approach to JER control.

For a deterministic kernel (e.g., linear), we can prove that the λ-adjustment can be done by the permutation method while maintaining exact JER control. Nevertheless, for a data-driven kernel, such as the balanced kernel that would itself be calibrated using random permutations, things are getting more complicated. A possible algorithm could be similar to the one proposed in Appendix A.3, except that the Monte-Carlo samples are be replaced by randomly transformed samples. This would also follow the lines 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), which seems to have been overlooked so far[2]. As a consequence, the usual argument to theoretically control error rates with permutation-based procedures apparently falls short in our setting. Hence, proving a theoretical JER control for such a procedure remains an open issue – one of primary importance.

Acknowledgements

This work has been supported by CNRS (PEPS FaSciDo) and ANR-16-CE40-0019. GB acknowledges the support from the german DFG, under the Research Unit FOR-1735 "Structural Inference in Statistics – Adaptation and Efficiency". We also thank Prof. Yoav Benjamini for interesting discussions.

Appendix A: Some properties of linear and balanced kernels and reference families

A.1. Properties of linear and balanced kernels

The following result gather some of the properties of the balanced kernel under independence :

Proposition A.1. *Consider the location model (15), assume independence between the components of the noise ε and consider $t_k(\lambda)$ the threshold given by (29). Then :*

(i) for all $k \in \{1, \dots, m\}$ and $\lambda \in [0, 1]$, $t_k(\lambda)$ is the λ-quantile of the distribution $\mathrm{Beta}(k, m + 1 - k)$.

(ii) the following relations holds : for any $\lambda \leq 0.5$,

$$t_k(\lambda) \leq \frac{k}{m+1};$$
$$t_k(\lambda) \geq \frac{k}{m+1} - \left\{ \frac{k}{m+1} \left(1 - \frac{k}{m+1} \right) \right\}^{1/2} m^{-1/2} \left(4 \log 1/\lambda \right)^{1/2}; \tag{42}$$

(iii) for all $m_0 \in \{1, \dots, m\}$ and $k \in \{1, \dots, m_0\}$, for all $\lambda \leq 0.5$,

$$\mathbb{P}\left(p_{(k:m_0)} \leq t_k(\lambda) \right) \leq \exp\left(-\frac{k}{32} \left(1 - \frac{m_0}{m} \right)^2 \right). \tag{43}$$

(iv) for all $k \in \{1, \dots, m\}$ and $\alpha \leq 0.5$,

$$\mathbb{P}\left(p_{(k:m)} \leq \alpha \frac{k}{m} \right) \leq \exp\left(-\frac{k}{4} \left(1 - \alpha - \frac{1}{m+1} \right)^2 \right). \tag{44}$$

[2] More precisely, Equation (12) of Meinshausen (2006) is not fully justified: the fact that for all $g \in \mathcal{G}$,

- $(t_k(g.X))_{1 \leq k \leq K} = (t_k(X))_{1 \leq k \leq K}$;
- $(p_i(X))_{i \in \mathcal{H}_0} \sim (p_i(g.X))_{i \in \mathcal{H}_0}$,

does not imply equality of joint distributions $((t_k(X))_{1 \leq k \leq K}, (p_i(X))_{i \in \mathcal{H}_0}) \sim ((t_k(g.X))_{1 \leq k \leq K}, (p_i(g.X))_{i \in \mathcal{H}_0})$.

Proof. Item (i) just follows from the definition. The proof of item (ii) is straightforward from a classical bound for beta distribution, see relation (49) and Lemma C.3 in Appendix C. For item (iii), we use item (ii) and $m/(m+1) \geq 1/2$ and $m_0/(m_0+1) \geq 1/2$, to write

$$\begin{aligned}
\mathbb{P}\left[p_{(k:m_0)} \leq t_k(\lambda)\right] &\leq \mathbb{P}\left[p_{(k:m_0)} \leq \frac{k}{m+1}\right] \\
&= \mathbb{P}\left[m_0^{1/2}\left(p_{(k:m_0)} - \frac{k}{m_0+1}\right) \leq -(k)^{1/2}(km_0)^{1/2}\left(\frac{1}{m_0+1} - \frac{1}{m+1}\right)\right] \\
&\leq \mathbb{P}\left[m_0^{1/2}\left(p_{(k:m_0)} - \frac{k}{m_0+1}\right) \leq -\left\{\frac{k}{m_0+1}\right\}^{1/2}(k/2)^{1/2}\left(\frac{m-m_0}{m+1}\right)\right] \\
&\leq \mathbb{P}\left[m_0^{1/2}\left(p_{(k:m_0)} - \frac{k}{m_0+1}\right) \leq -\left\{\frac{k}{m_0+1}\right\}^{1/2}(k/2)^{1/2}\left(1-\frac{m_0}{m}\right)/2\right],
\end{aligned}$$

and we conclude by using (49). For (iv), the reasoning is similar, using

$$\mathbb{P}\left[p_{(k:m)} \leq \alpha\frac{k}{m}\right] = \mathbb{P}\left[m^{1/2}\left(p_{(k:m)} - \frac{k}{m+1}\right) \leq -\left\{\frac{k}{m}\right\}^{1/2}\sqrt{k}\left(\left(1-\frac{1}{m+1}\right)-\alpha\right)\right].$$

□

A.2. Properties of the single-step balanced reference family of size $K = m$

Under independence, we can have a bound measuring how severe the λ adjustment can be when m is large:

Lemma A.2. *In the framework of Proposition A.1, consider $\lambda^B(\alpha) = \lambda^B(\alpha, \mathbb{N}_m)$ defined by (30) for $K = m$. Then for m large enough, we have*

$$\lambda^B(\alpha) \leq 1/(\log m)^{1/4}. \tag{45}$$

In particular, this result shows that $\lambda(\alpha)$ tends to zero as m grows to infinity.

Proof. Let $\lambda_0 = 1/(\log m)^{1/4}$ and consider $U_1, \ldots, U_m$ i.i.d. $\sim U(0,1)$. By definition of $\lambda(\alpha)$, it is sufficient to prove that for m large enough, the probability $\mathbb{P}(\exists k \in \{1,\ldots,m\} : U_{(k:m)} < t_k(\lambda_0))$ is larger than α. For this, use the lower bound (42) to write for a large m,

$$\begin{aligned}
&\mathbb{P}(\exists k \in \{1,\ldots,m\} : U_{(k:m)} < t_k(\lambda_0)) \\
&= \mathbb{P}(\exists k \in \{1,\ldots,m\} : U_{(k:m)} \leq t_k(\lambda_0)) \\
&\geq \mathbb{P}\left(\exists k \in \{1,\ldots,m\} : U_{(k:m)} \leq \frac{k}{m+1} - \left\{\frac{k}{m+1}\left(1-\frac{k}{m+1}\right)\right\}^{1/2} m^{-1/2}\left(4\log(1/\lambda_0)\right)^{1/2}\right) \\
&= \mathbb{P}\left(Z_m \geq \left(4\log(1/\lambda_0)\right)^{1/2}\right)
\end{aligned}$$

where we let

$$Z_m = \max_{1\leq k\leq m}\left\{\frac{m^{1/2}}{\left\{\frac{k}{m+1}\left(1-\frac{k}{m+1}\right)\right\}^{1/2}}\left(\frac{k}{m+1} - U_{(k:m)}\right)\right\}.$$

Since $\left(4\log(1/\lambda_0)\right)^{1/2} = (\log\log m)^{1/2}$, we conclude by applying Lemma C.2. □

The next lemma shows that using the substitute $\lambda^B(\alpha, \mathbb{N}_m)$ instead of $\lambda^B(\alpha, \mathcal{H}_0)$ for the balanced kernel (in the case $K = m$) results in a JER tending asymptotically to 0 with m if π_0 is bounded away from 1. Thus, this justifies the importance of trying to use some kind of adaptive procedure, such as the step-down.

Lemma A.3. *Consider $\lambda^B(\alpha) = \lambda^B(\alpha, \mathbb{N}_m)$ defined by (30) for $K = m$ and q_i, $i \in \mathbb{N}_m$, i.i.d. $U(0,1)$ variables. Consider $m_0 = \pi_0 m$ for some $\pi_0 \in (0,1)$ fixed with m. Then, for m large enough,*

$$\mathbb{P}\Big(\exists k \in \{1, \dots, m_0\} \,:\, q_{(k:m_0)} < t_k\left(\lambda^B(\alpha)\right)\Big) \leq \frac{C(\pi_0)}{(\log m)^{1/8}}, \tag{46}$$

where $C(\pi_0) = 1 + \frac{64}{(1-\pi_0)^2}\left(1 - e^{-\frac{(1-\pi_0)^2}{32}}\right)$.

Proof. Let $N \in \{1, \dots, m_0 - 1\}$ be some integer to be chosen later. By a union bound argument, we have

$$\mathbb{P}\Big(\exists k \in \{1, \dots, m_0\} \,:\, q_{(k:m_0)} < t_k\left(\lambda^B(\alpha)\right)\Big)$$
$$\leq \sum_{k=1}^{N} \mathbb{P}\Big(q_{(k:m_0)} < t_k\left(\lambda^B(\alpha)\right)\Big) + \sum_{k=N+1}^{m_0} \mathbb{P}\Big(q_{(k:m_0)} < t_k\left(\lambda^B(\alpha)\right)\Big).$$

For the first term, since $q_{(k:m)}$ is stochastically smaller than $q_{(k:m_0)}$, we have

$$\sum_{k=1}^{N} \mathbb{P}\Big(q_{(k:m_0)} < t_k\left(\lambda^B(\alpha)\right)\Big) \leq N\lambda^B(\alpha) \leq N/(\log m)^{1/4},$$

by Lemma A.2. For the second term, by (43) ($\lambda^B(\alpha)$ begin smaller than 0.5 for large enough m) and letting $r = e^{-\frac{(1-\pi_0)^2}{32}}$, we have for large enough m:

$$\sum_{k=N+1}^{m_0} \mathbb{P}\Big(q_{(k:m_0)} < t_k\left(\lambda^B(\alpha)\right)\Big) \leq \sum_{k=N+1}^{m_0} r^k \leq (1-r)^{-1} r^N \leq (1-r)^{-1} \frac{32}{N(1-\pi_0)^2},$$

because $e^{-u} \leq 1/u$ for all $u > 0$. Choosing $N = \lfloor (\log m)^{1/8} \rfloor$ yields the desired result. □

A.3. Monte-Carlo approximation for step-down balanced reference family

Let us consider the reference thresholds $t_k^B(\lambda^B(\alpha, \widehat{\mathcal{C}}))$, $1 \leq k \leq K$, of the balanced reference family $\mathfrak{R}^{B,sd}$ given in Section 5.2. For a practical use, we detail here how to obtain a Monte-Carlo approximation $\widetilde{t}_k^B(\widetilde{\lambda}^B(\alpha, \widetilde{\mathcal{C}}))$ of $t_k^B(\lambda^B(\alpha, \widehat{\mathcal{C}}))$.

1. Start by generating $q^{(1)}, \dots, q^{(G)}$ i.i.d. according to the distribution ν_m (on $[0,1]^m$) and consider the matrix
$$M_0 = \begin{pmatrix} q_1^{(1)} & q_2^{(1)} & \dots & q_m^{(1)} \\ q_1^{(2)} & q_2^{(2)} & \dots & q_m^{(2)} \\ \vdots & \vdots & & \vdots \\ q_1^{(G)} & q_2^{(G)} & \dots & q_m^{(G)} \end{pmatrix}$$
2. Order (some of the elements of) the lines of M_0 and let, for all $\mathcal{C} \subset \{1, \dots, m\}$ of cardinal c,
$$M(\mathcal{C}) = \begin{pmatrix} q_{(1:\mathcal{C})}^{(1)} & q_{(2:\mathcal{C})}^{(1)} & \dots & q_{(c:\mathcal{C})}^{(1)} \\ q_{(1:\mathcal{C})}^{(2)} & q_{(2:\mathcal{C})}^{(2)} & \dots & q_{(c:\mathcal{C})}^{(2)} \\ \vdots & \vdots & & \vdots \\ q_{(1:\mathcal{C})}^{(G)} & q_{(2:\mathcal{C})}^{(G)} & \dots & q_{(c:\mathcal{C})}^{(G)} \end{pmatrix}.$$
3. Consider $M(\mathcal{C})$ for $\mathcal{C} = \mathbb{N}_m$ and approximate $F_k(x)$ by $\widetilde{F}_k(x) = G^{-1}\sum_{g=1}^{G} \mathbf{1}\left\{q_{(k:m)}^{(g)} \leq x\right\}$. For each λ, approximate $t_k(\lambda)$ by $\widetilde{t}_k(\lambda)$ the λ-quantile of the sample
$$\left(q_{(k:m)}^{(1)}, \dots, q_{(k:m)}^{(G)}\right).$$

4. Consider the matrix of 'ranks'

$$S(\mathcal{C}) = \begin{pmatrix} S_{1,1}(\mathcal{C}) & S_{1,2}(\mathcal{C}) & \dots & S_{1,c}(\mathcal{C}) \\ S_{2,1}(\mathcal{C}) & S_{2,2}(\mathcal{C}) & \dots & S_{2,c}(\mathcal{C}) \\ \vdots & \vdots & & \vdots \\ S_{G,1}(\mathcal{C}) & S_{G,2}(\mathcal{C}) & \dots & S_{G,c}(\mathcal{C}) \end{pmatrix}$$

where we let $S_{b,k}(\mathcal{C}) = G^{-1}\sum_{g'=1}^{G} \mathbf{1}\left\{q_{(k:m)}^{(g')} \leq q_{(k:\mathcal{C})}^{(g)}\right\}$, for $1 \leq k \leq K \wedge c$ and $1 \leq b \leq G$.

5. Build the vector

$$V = \left(\min_{1\leq k\leq c}\{S_{1,k}(\mathcal{C})\}, \dots, \min_{1\leq k\leq c}\{S_{G,k}(\mathcal{C})\}\right),$$

by taking the minimum within each line of $S(\mathcal{C})$. Approximate now $\lambda^B(\alpha, \mathcal{C})$ by $\widetilde{\lambda}^B(\alpha, \mathcal{C}) = V_{(\lceil \alpha G \rceil)}$, i.e., the α empirical quantile of the sample $V = (V_1, \dots, V_G)$.

6. Use Algorithm 4.4 with $\widetilde{t}_k^B$ and $\widetilde{\lambda}^B$ instead of t_k^B and λ^B, respectively, to obtain $\widetilde{C}$.

Appendix B: Proofs

B.1. Proofs for Section 2

Proof of Proposition 2.1. We show a circular inclusion of the events in (8), starting with $\mathcal{E}(\mathfrak{R}, \mathcal{H}_0)$. First, for $1 \leq k \leq K$, $|R_k \cap \mathcal{H}_0| \leq \zeta_k$, implies that for any R,

$$|R \cap \mathcal{H}_1| + \zeta_k \geq |R \cap \mathcal{H}_1| + |R_k \cap \mathcal{H}_0| \geq |R \cap R_k|,$$

which entails $|R \cap \mathcal{H}_1| \geq \overline{S}(R)$ by taking a maximum over all possible values of k. Secondly, if $\forall R$, $|R \cap \mathcal{H}_1| \geq \overline{S}(R)$, then for $R = \mathcal{H}_0$, we obtain $\overline{S}(\mathcal{H}_0) = 0$. Finally, $\overline{S}(\mathcal{H}_0) = 0$ implies that for all k, $|\mathcal{H}_0 \cap R_k| - \zeta_k \leq 0$.

Similarly for (9), the event $\mathcal{E}(\mathfrak{R}, \mathcal{H}_0)$ is by definition equivalent to $\mathcal{H}_0 \in \mathcal{A}(\mathfrak{R})$. This implies for any R: $|R \cap \mathcal{H}_1| = |R \setminus \mathcal{H}_0| \geq S^*(R)$. Specializing for $R = \mathcal{H}_0$ as above, this entails $S^*(\mathcal{H}_0) = 0$. Finally, the latter event implies in turn that there must exist $A \in \mathcal{A}(\mathfrak{R})$ with $\mathcal{H}_0 \subset A$, but since any subset of an element of $\mathcal{A}(\mathfrak{R})$ also belongs to $\mathcal{A}(\mathfrak{R})$, we conclude $\mathcal{H}_0 \in \mathcal{A}(\mathfrak{R})$. □

Proof of Proposition 2.2. We prove that the specific subproblem of computing $V^*(\mathfrak{R}, R)$ under the following restrictions is already NP-hard:

- $|R_k| = 2$ for all k;
- $\zeta_k = 1$ for all k;
- $R = \mathbb{N}_m$.

Namely, we can formally embed as an instance of this setting the well-known NP-complete problem of finding a maximal independent set of vertices in an arbitrary graph G, in the following way: let K be the number of edges in the graph; construct the family of sets by associating to each edge e of G the set R_e containing the two vertices it joins, and $\zeta_e = 1$. Then elements of $\mathcal{A}(\mathfrak{R})$ are exactly the subsets of independent vertices of G. Taking $R = \mathbb{N}_m$, computing $V^*(R) = \max_{A\in\mathcal{A}(\mathfrak{R})} |A|$ is then equivalent to finding the maximal size of an independent vertex set in G. □

Proof of Proposition 2.3. Obviously, $\widetilde{\zeta}_k \leq \zeta_k$ and thus $\overline{V}(\widetilde{\mathfrak{R}}, R) \leq \overline{V}(\mathfrak{R}, R)$. Let us prove the reverse inequality:

$$\begin{aligned}\overline{V}(\widetilde{\mathfrak{R}}, R) &= \min_{k\in\{1,\dots,K\}} \left(|R \setminus R_k| + \min_{j\in\{1,\dots,K\}} (|R_k \setminus R_j| + \zeta_j) \wedge |R_k|\right) \wedge |R| \\ &\geq \min_{j,k\in\{1,\dots,K\}} (|R \setminus R_k| + |R_k \setminus R_j| + \zeta_j) \wedge |R| \\ &\geq \min_{j\in\{1,\dots,K\}} (|R \setminus R_j| + \zeta_j) \wedge |R|,\end{aligned}$$

where we used $|E \setminus F| + |F \setminus G| \geq |E \setminus G|$. □

Proof of Proposition 2.4. For convenience, we recall the notation

$$\mathcal{A}(\mathfrak{R}) := \{A \subset \mathbb{N}_m : \ \forall k = 1, \dots, K, \ |R_k \cap A| \le \zeta_k\}$$

in the definition of V^*. Let $R \subset \mathbb{N}_m$; it is straightforward to check that $V^*(\mathfrak{R}, R) \le \overline{V}(\mathfrak{R}, R)$, since V^* is optimal; in fact for all $A \in \mathcal{A}$ and $k \in \{1, \dots, K\}$, we have $|R \cap A| \le |R \cap A \cap R_k| + |R \cap A \cap R_k^c| \le (\zeta_k + |R \cap R_k^c|) \wedge |R|$. We now prove the reverse inequality, by showing that there exists a set $A \in \mathcal{A}(\mathfrak{R})$ such that $A \subset R$ and $|A| \ge \overline{V}(\mathfrak{R}, R)$. For this, note that $\overline{V}(\mathfrak{R}, R) \le |R \setminus (R_K \cap R)| + \tilde{\zeta}_K$, where we let

$$\tilde{\zeta}_k = \min_{1 \le j \le K} \{|(R_k \cap R) \setminus (R_j \cap R)| + \zeta_j\} \wedge |R_k \cap R|, \quad 1 \le k \le K \, .$$

The latter is in accordance with the definition (10), applied to the family $((R_k \cap R), \zeta_k), 1 \le k \le K$. This means that (12) is satisfied and in particular

$$\tilde{\zeta}_k - \tilde{\zeta}_{k-1} \le |(R_k \cap R) \setminus (R_{k-1} \cap R)| \, , \quad 1 \le k \le K,$$

with the conventions $\tilde{\zeta}_0 = 0$ and $R_0 = \emptyset$. Now construct a set A by picking $\tilde{\zeta}_k - \tilde{\zeta}_{k-1}$ elements in each $(R_k \cap R) \setminus (R_{k-1} \cap R)$ for $1 \le k \le K$ (which is possible by the latter display) and add the points of $R \setminus (R_K \cap R)$. We now check that A satisfies the constraints ensuring $A \in \mathcal{A}(\mathfrak{R})$, using the nestedness assumption and the fact that $A \subset R$ by construction:

$$|A \cap R_k| = |R_k \cap R \cap A| = \sum_{j=1}^{k} |(R_j \cap R \cap A) \setminus (R_{j-1} \cap R \cap A)| = \tilde{\zeta}_k \le \zeta_k \, , \quad 1 \le k \le K \, .$$

Moreover, $\overline{V}(\mathfrak{R}, R) \le |R \setminus (R_K \cap R)| + \tilde{\zeta}_K = |R \setminus (R_K \cap R)| + |R_K \cap R \cap A| = |A|$, so the result is proved. □

B.2. Proof of Theorem 7.2

Proof. For proving (i), we note that any family $\mathfrak{R}$ with thresholds t_k, $1 \le k \le m$, controlling the JER at level α induces a test $\varphi = \mathbf{1}\left\{\exists k : p_{(k)} \le t_k\right\}$ of level α of H_0 :"$\mu_m = 0$" (i.e., $p_i, i \in \mathbb{N}_m$ are all i.i.d. uniform) against H_1: "$\mu_m = \sqrt{2r \log m}$". Hence, it will have less power than the likelihood ratio test (LRT) of level α. Now, as claimed in Section 1.1 of Donoho and Jin (2004) (itself referring to Ingster, 1999), the null hypothesis and the alternative hypothesis merge asymptotically whenever $r < \rho^\star(\beta)$. Hence, the asymptotic power of the LRT is less than α.

Now consider the balanced family $\mathfrak{R}_\alpha^B$ and prove (ii). Write t_k for $t_k^B(\lambda^B(\alpha)$ for simplicity. The basic inequality for our proof is the following: for any $k \in \{1, \dots, m\}$,

$$\mathrm{Pow}^\star(\mathfrak{R}_\alpha^B, P) \ge \mathbb{P}\left(U_{(k:m)} \le \pi_{0,m} t_k + \pi_{1,m} F_{1,m}(t_k)\right) . \tag{47}$$

From (37), now write for any $k \in \{1, \dots, m\}$,

$$\begin{aligned}
&\mathrm{Pow}^\star(\mathfrak{R}_\alpha^B, P) \\
&\ge \mathbb{P}\left(U_{(k:m)} \le \frac{k}{m+1} - \left\{\frac{k}{m+1}\left(1 - \frac{k}{m+1}\right)\right\}^{1/2} m^{-1/2} \left(4 \log(m/\alpha)\right)^{1/2} + \pi_{1,m}\left(F_{1,m}(t_k) - t_k\right)\right) \\
&\ge \mathbb{P}\left(\max_{1 \le k \le m} \left\{ m^{1/2} \frac{U_{(k:m)} - \frac{k}{m+1}}{\left\{\frac{k}{m+1}\left(1 - \frac{k}{m+1}\right)\right\}^{1/2}} \right\} \le -\left(4 \log(m/\alpha)\right)^{1/2} + \pi_{1,m} \frac{F_{1,m}(t_k) - t_k}{k^{1/2}/(m+1)}\right)
\end{aligned} \tag{48}$$

because $k^{1/2}/(m+1) \ge \left\{\frac{k}{m+1}\left(1 - \frac{k}{m+1}\right)\right\}^{1/2} m^{-1/2}$. Let $r_k > 0$ such that $\overline{\Phi}^{-1}(t_k) = \sqrt{2 r_k \log m}$, so that when $k = \lfloor m^{1-q} \log m \rfloor$ for some $q \in (0, 1]$, $r_k \to q$ as m tends to infinity. Then we have

for $q > r$, by denoting ϕ the density of the standard gaussian distribution,

$$\begin{aligned}\pi_{1,m}F_{1,m}(t_k) &= m^{-\beta}\overline{\Phi}(\overline{\Phi}^{-1}(t_k) - \sqrt{2r\log m}) \\ &= m^{-\beta}\overline{\Phi}((q^{1/2} - r^{1/2})\sqrt{2\log m}) \\ &\sim Dm^{-\beta}\phi\left((q^{1/2} - r^{1/2})\sqrt{2\log m}\right)/\sqrt{\log m} \\ &\sim D'm^{-\beta-(q^{1/2}-r^{1/2})^2}/\sqrt{\log m},\end{aligned}$$

for some constants $D, D' > 0$. This entails

$$\pi_{1,m}\frac{F_{1,m}(t_k) - t_k}{k^{1/2}/(m+1)} \sim D'm^{\frac{1+q}{2}-\beta-(q^{1/2}-r^{1/2})^2}/\log m,$$

By choosing $q = (4r) \wedge 1$, we have $\frac{1+q}{2} - \beta - (q^{1/2} - r^{1/2})^2 > 0$ as soon as $r > \rho^\star(\beta)$. Now, (ii) comes from (48) and the fact that the sequence of random variables

$$\max_{1\le k\le m}\left\{m^{1/2}\frac{U_{(k:m)} - \frac{k}{m+1}}{\left\{\frac{k}{m+1}\left(1 - \frac{k}{m+1}\right)\right\}^{1/2}}\right\}/(\log\log m)^{1/2}$$

is tight (see Lemma C.2). □

B.3. Proof of Theorem 7.3

Proof. Let us first prove (i). For any $k \in \{1, \dots, m\}$,

$$\mathrm{Pow}^\star(\mathfrak{R}^0_\alpha, P) \ge \mathbb{P}\left(U_{(k:m)} \le \frac{k}{m}\left(\pi_{0,m}\alpha + m\pi_{1,m}F_{1,m}(\alpha k/m)/k\right)\right)$$

Let $r_k > 0$ such that $\overline{\Phi}^{-1}(\alpha k/m) = \sqrt{2r_k\log m}$, so that when $k = \lfloor \log m \rfloor$, $r_k \to 1$ as m tends to infinity. Then we have for some universal constant $D > 0$,

$$m\pi_{1,m}F_{1,m}(\alpha k/m)/k \sim Dm^{1-\beta-(1-r^{1/2})^2}/(\log m)^{3/2},$$

and thus the latter tends to infinity. Hence, for any $M > 0$, for m large enough, we have

$$\mathrm{Pow}^\star(\mathfrak{R}^0_\alpha, P) \ge \mathbb{P}\left(U_{(k:m)} \le M\frac{k}{m}\right).$$

Then (i) is proved because $mU_{(k:m)}/k$ tends to 1 in probability. Now, let us show (ii). We have

$$\begin{aligned}\mathrm{Pow}^\star(\mathfrak{R}^0_\alpha, P) &= \mathbb{P}\left(\exists k \in \{1, \dots, m\} \,:\, U_{(k:m)} \le \alpha\frac{k}{m}\left(\pi_{0,m} + \pi_{1,m}\frac{F_{1,m}(\alpha k/m)}{\alpha k/m}\right)\right) \\ &\le \mathbb{P}\left(\exists k \in \{1, \dots, m\} \,:\, U_{(k:m)} \le \alpha\frac{k}{m}\left(\pi_{0,m} + \pi_{1,m}\frac{F_{1,m}(\alpha/m)}{\alpha/m}\right)\right)\end{aligned}$$

because $F_{1,m}(x)/x$ is decreasing. Now, we have

$$\pi_{1,m}\frac{F_{1,m}(\alpha/m)}{\alpha/m} \sim Dm^{1-\beta-(1-r^{1/2})^2}/(\log m)^{1/2},$$

for some universal constant $D > 0$, and thus the latter tends to zero as soon as $r < (1 - \sqrt{1-\beta})^2$. Hence, for any $\varepsilon \in (0, 1)$, for m large enough,

$$\mathrm{Pow}^\star(\mathfrak{R}^0_\alpha, P) \le \mathbb{P}\left(\exists k \in \{1, \dots, m\} \,:\, U_{(k:m)} \le \alpha\frac{k}{m}(1+\varepsilon)\right) \le \alpha(1+\varepsilon),$$

by applying the Simes inequality. The result comes by making ε tends to zero. □

Appendix C: Some properties of the beta distribution

We recall the following result (Shorack and Wellner, 1986, p.454-455):

Lemma C.1. *for $U_1, \ldots, U_n$ i.i.d. $\sim U(0,1)$, any $\ell \in \{1, \ldots, n\}$ and $x \geq 1$, we have*

$$\mathbb{P}\left(n^{1/2}\left(U_{(\ell:n)} - \frac{\ell}{n+1}\right) \leq -\left\{\frac{\ell}{n+1}\left(1 - \frac{\ell}{n+1}\right)\right\}^{1/2} x\right) \leq e^{-x^2/4} \tag{49}$$

Here is another lemma, which is a consequence of (24) in (Shorack and Wellner, 1986, p.601):

Lemma C.2. *Let $U_1, U_2, \ldots$ i.i.d. $\sim U(0,1)$ and consider*

$$Z_n = \max_{1 \leq \ell \leq n} \left\{ n^{1/2} \frac{\frac{\ell}{n+1} - U_{(\ell:n)}}{\left\{\frac{\ell}{n+1}\left(1 - \frac{\ell}{n+1}\right)\right\}^{1/2}} \right\},$$

then we have, as n grows to infinity,

$$\mathbb{P}((\log\log n)^{1/2} \leq Z_n \leq 2(\log\log n)^{1/2}) \to 1 \tag{50}$$

Lemma C.3. *Let $U_1, U_2, \ldots$ i.i.d. $\sim U(0,1)$, then, for all $m \geq 2$, for all $k \in \{1, \ldots, m\}$,*

$$\mathbb{P}(U_{(k:m)} \leq (k+1)/m) \geq 0.5. \tag{51}$$

Proof. We can assume $k \leq m-1$. Now, by considering $Z \sim \mathcal{B}(m, (k+1)/m)$, we have

$$\mathbb{P}(U_{(k:m)} \leq (k+1)/m) = \mathbb{P}(Z \geq k) = \mathbb{P}(Z \geq (k+1) - 1) \geq 1/2,$$

where we used that for any binomial distribution, the median and the mean are at a distance at most 1 (see, e.g., Kaas and Buhrman, 1980):

□

Appendix D: Numerical experiments for detection power

Detection power is studied theoretically in Section 7 and formally defined in (35). We consider the independent case, and we calibrate the parameter $\overline{\mu}$ and π_0 according to the regime defined in Section 7, that is, $\pi_0 = 1 - m^{-\beta}$ and $\overline{\mu} = \sqrt{2r\log m}$, for two parameters β (sparsity) and r (signal strength) taken in the range $\beta \in \{0.5, 0.6, 0.8, 1\}$ and $r \in \{0.05, 0.1, 0.2, 0.5, 1\}$. Note that, however, we do not consider an i.i.d. p-value mixture here; we stick to the framework defined in Section 6. For each setting, we estimate this power by its empirical counterpart, the proportion $\hat{q}$ of $1,000$ simulation runs for which at least one of the subsets R_k of the collection $\mathfrak{R}$ contains more than k elements. Our experiments have been made for a range of values of the target JER level $\alpha \in \{0.01, 0.02, 0.05, 0.10, 0.15, 0.20, 0.25\}$. To summarize the results, we plot in Figure 9 (top) the empirical detection power $\hat{q}$ as a function of α for each method.

The parameter configurations (β, r) for which the signal is below the asymptotically optimal detection boundary identified by Donoho and Jin (2004) are represented by blue squares in the bottom panel of Figure 9. As expected from the theory, in such configurations all procedures are powerless, in the sense that the detection power is very close to the JER. Let us focus on the parameter configurations for which detection is asymptotically feasible (green circles and red triangles in the bottom panel of Figure 9). In such configurations, as expected, K has little influence on detection power for the linear kernel. For the balanced kernel, the detection power is substantially higher for $K = 10$ than for $K = m$. This influence of K is consistent with our comments for JER control in the preceding section. Overall, the balanced kernel with $K = m$ has better detection power than the linear kernel for moderate sparsity ($\beta \in \{0.5, 0.6\}$) and signal

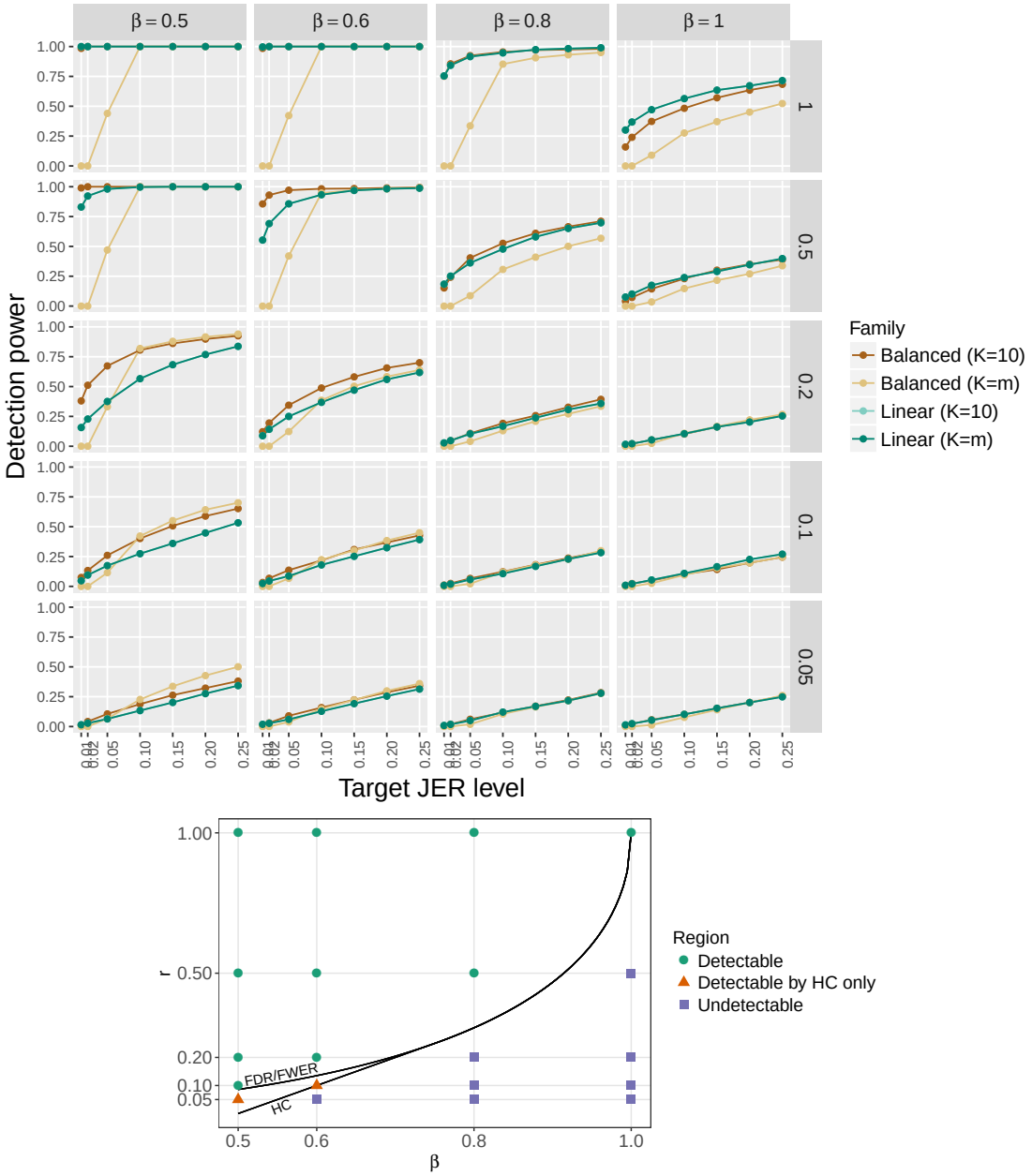

FIG 9. *Top: detection power of JER controlling procedures for independent test statistics in the sparsity range for* 4×5 *parameter configurations for* (β, r) *in the sparsity range* $[1/2, 1] \times [0, 1]$. *Bottom: these* 4×5 *configurations are positioned with respect to the detection boundaries identified in Donoho and Jin (2004).*

($r \leq 1$). However, for sparser settings ($\beta \in \{0.8, 1\}$) the linear kernel performs better than the balanced kernel with $K = m$, and even than the balanced kernel with $K = 10$ in very sparse scenarios. These numerical results provide a useful complement to the asymptotic statements of Donoho and Jin (2004) and of Section 6 of the present paper. In particular, they suggest that for a finite m, the balanced kernel/HC is not always superior to the linear kernel/FWER/FDR. Moreover, in the sparse/weak setting, which is illustrated here by the configurations $\beta = 0.5, r = 0.05$ and $\beta = 0.6, r = 0.1$, the balanced kernel is only marginally superior to the balanced kernel.

References

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

Benjamini, Y. 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.

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. 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.

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

Ingster, Y. I. (1999). Minimax detection of a signal for l^n-balls. *Math. Methods Statist.*, 7(4):401–428.

Kaas, R. and Buhrman, J. (1980). Mean, median and mode in binomial distributions. *Statistica Neerlandica*, 34(1):13–18.

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

Marcus, R., Peritz, E., and Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660.

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

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

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

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