Post hoc false positive control for spatially structured hypotheses

Guillermo Durand, Gilles Blanchard, Pierre Neuvial, Etienne Roquain

Post hoc false positive control for spatially structured hypotheses

Guillermo Durand

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

Gilles Blanchard

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

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

Etienne Roquain

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

Abstract: In a high dimensional multiple testing framework, we present new confidence bounds on the false positives contained in subsets S of selected null hypotheses. The coverage probability holds simultaneously over all subsets S, which means that the obtained confidence bounds are post hoc. Therefore, S can be chosen arbitrarily, possibly by using the data set several times. We focus in this paper specifically on the case where the null hypotheses are spatially structured. Our method is based on recent advances in post hoc inference and particularly on the general methodology of Blanchard et al. (2017); we build confidence bounds for some pre-specified forest-structured subsets $\{R_k, k \in \mathcal{K}\}$, called the reference family, and then we deduce a bound for any subset S by interpolation. The proposed bounds are shown to improve substantially previous ones when the signal is locally structured. Our findings are supported both by theoretical results and numerical experiments. Moreover, we show that our bound can be obtained by a low-complexity algorithm, which makes our approach completely operational for a practical use. The proposed bounds are implemented in the open-source R package `sansSouci`*.

AMS 2000 subject classifications: Primary 62G10; secondary 62H15.
Keywords and phrases: post hoc inference, selective inference, multiple testing, Simes inequality, Forest structure, DKW inequality.

1. Introduction

1.1. Background

Modern statistical data analysis often involves asking many questions of interest simultaneously, possibly using the data repeatedly, as long as the user feels that this could provide additional information. To avoid selection bias due to various forms of data snooping, specific strategies can be proposed to take into account the procedure as whole, and be investigated as to the statistical guarantees they provide. This problem is often referred to as selective inference, a long standing research field, with a recent renewal of interest. An historical reference is the work of Scheffé

*available from https://github.com/pneuvial/sanssouci.

(1953) (see also Scheffé, 1959, p. 69), which is to our knowledge the earliest work proposing simultaneous selective inference. In the context of linear regression, Berk et al. (2013) proposed an improvement of this Scheffé protection by defining a less conservative correction term (the so-called PoSI constant), see also Bachoc et al. (2018); Bachoc et al. (2018) for recent developments on this issue.

Other strategies perform inference on the observed selection set only, either by a false coverage rate control (Benjamini and Yekutieli, 2005; Benjamini and Bogomolov, 2014) or by a controlling a criterion conditional to a specific initial selection step, see the series of works Fithian et al. (2017); Taylor and Tibshirani (2015); Tibshirani et al. (2016); Choi et al. (2017); Taylor and Tibshirani (2018). In other studies, the selection step is based on sample splitting, see Cox (1975); Bühlmann and Mandozzi (2014); Dezeure et al. (2015), which is another way to tackle selective inference by explicitly avoiding data reuse.

We follow in this paper the aim of establishing confidence bounds on the number of false positives in a multiple testing framework, simultaneously over all possible set of selected hypotheses. If we observe a random variable $X \sim P$, P belonging to some model $\mathcal{P}$, for which m null hypotheses $H_{0,i} \subset \mathcal{P}$, $i \in \mathbb{N}_m = \{1, \dots, m\}$ are under investigation for P, the aim is to build a function $V(X, \cdot) : S \subset \mathbb{N}_m \mapsto V(X, S)$ (denoted by $V(S)$ for short) satisfying

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

where $\mathcal{H}_0(P) = \{i \in \mathbb{N}_m : P \text{ satisfies } H_{0,i}\}$ is the set of true null hypotheses. The bound $V(\cdot)$ will be referred to as a post hoc bound throughout this manuscript.

The problem of constructing post hoc bounds has been first tackled specifically in the case where the selection sets S are of the form of p-value level sets: $\{i : p_i(X) \leq t\}$, $t \in [0, 1]$, where each $p_i(X)$ is a p-value for the null hypothesis $H_{0,i}$, $1 \leq i \leq m$. The resulting bounds are often referred to as confidence envelopes, see Genovese and Wasserman (2004); Meinshausen (2006). Later, Genovese and Wasserman (2006) and Goeman and Solari (2011) proposed to extend this approach to arbitrary subsets S, by using a methodology based on performing $2^m - 1$ local tests (one for each intersection hypothesis), with a possible complexity reduction by using shortcuts. In particular, the approach of Goeman and Solari (2011) extensively relies on the closed testing principle, which was introduced by Marcus et al. (1976). This approach has been further extended in Meijer and Goeman (2015); Meijer et al. (2015) by using the sequential rejection principle of Goeman and Solari (2010). This allows to incorporate structural informations into the post hoc bound. In particular, the method in Meijer et al. (2015), whose goal inspired the present work, deals with geometrically structured null hypotheses, along space or time and shows that incorporating such an external information can substantially improve the detection of signal and thus can increase the accuracy of the resulting post hoc bound.

More recently, Blanchard et al. (2017) (BNR below) have proposed a flexible methodology that adjusts the complexity of the bound by way of a reference family: the post hoc bound is based on a family $\mathfrak{R} = ((R_k(X), \zeta_k(X))_{k\in\mathcal{K}}$ (R_k, ζ_k for short), with $R_k \subset \mathbb{N}_m$ (and $R_k \neq R_{k'}$ if $k \neq k'$), $\zeta_k \in \mathbb{N}$, that satisfies the following joint error rate (JER) control:

$$\forall P \in \mathcal{P}, \qquad \mathbb{P}_{X\sim P}\Big(\forall k \in \mathcal{K},\ |R_k \cap \mathcal{H}_0(P)| \leq \zeta_k\Big) \geq 1 - \alpha, \tag{2}$$

An important difference between (1) and (2) is that S in (1) is let arbitrary and typically chosen by the user, whereas R_k, ζ_k in (2) is part of the methodology and is chosen by the statistician to make (2) hold. Once the reference family is fixed, a post hoc bound is obtained from (2) simply by interpolation, by exploiting the constraints that the event in (2) imposes to the unknown set $\mathcal{H}_0(P)$, namely that it is a subset A with the property "$\forall k \in \mathcal{K},\ |R_k \cap A| \leq \zeta_k$":

$$V^*_{\mathfrak{R}}(S) = \max\big\{\, |S \cap A|\, , A \subset \mathbb{N}_m, \forall k \in \mathcal{K},\ |R_k \cap A| \leq \zeta_k \big\}, \quad S \subset \mathbb{N}_m\, . \tag{3}$$

Hence, if (2) holds, then $V = V^*_{\mathfrak{R}}$ satisfies (1). This post-hoc bound will be referred to as the *optimal bound* (relative to a given reference family).

1.2. Contributions of the paper

In this paper, we propose new post hoc bounds that incorporate the specific spatial structure of the null hypotheses. While this aim is similar in spirit to Meijer et al. (2015), our method is markedly different, as it relies on the general strategy laid down by BNR, with a specifically structured reference family $R_k, k \in \mathcal{K}$ (see Section 6.1 for a comparison between our approach and the one of Meijer et al., 2015). In addition, the way the method is built here is different than the one proposed in Section 3-6 of BNR: the main focus in BNR is the case of (random) reference sets $R_k = R_k(X)$ that are designed in order to satisfy (2) with $\zeta_k = k - 1$ (thus corresponding to a "joint k-family-wise error rate"). By contrast, in the present work the reference sets R_k are fixed in advance, and the (random) bounds on the number false positives $\zeta_k = \zeta_k(X)$ are designed to satisfy the constraint (2). The rationale behind this approach is that the reference sets R_k can be chosen arbitrarily by the statistician, so that it can accommodate any pre-specified structure (reflecting some prior knowledge on the considered problem). Since we are interested in structured signal, we focus on a reference family enjoying a forest structure, meaning that two reference sets are either disjoint or nested.

The second ingredient of our method is the local bounds $\zeta_k(X)$, that should estimate $|R_k \cap \mathcal{H}_0(P)|$ with a suitable deviation term. While any deviation inequality can be used, we have chosen to focus on the DKW inequality (Dvoretzky et al., 1956), that has the advantage to be sub-Gaussian. Hence, the uniformity over the range $k \in \mathcal{K}$ can be obtained by a simple union bound without being too conservative.

Let us mention that using the DKW inequality to obtain a confidence bound for the proportion of null hypotheses is not new, see Genovese and Wasserman (2004) (Equation (16) therein), Meinshausen (2006), and Farcomeni and Pacillo (2011). While our bound is a uniform improvement of the existing version (see Remark 4.3 below for more details), our main innovation is to use the DKW bound in a local manner and to appropriately combine these local bounds to derive an overall post hoc bound. The improvement can be substantial, as illustrated in our numerical experiments.

The paper is organized as follows: precise setup and notation are introduced in Section 2. For any reference family with a forest structure, the optimal post hoc bound is computed in Section 3. The calibration of the local bounds ζ_k and of the overall reference family is done in Section 4. This section also includes a theoretical comparison with previous methods, which quantifies formally the amplitude of the improvement induced by the new method. The latter is supported by numerical experiments in Section 5, where a hybrid approach is also introduced to mimic the best between the new approach and the existing Simes bound (the latter being defined in (7) below). A discussion is given in Section 6 and the proofs are provided in Section 7. Additional technical details are postponed to Appendices A and B.

2. Preliminaries

2.1. Assumptions

We focus on the common situation where a test statistic $T_i(X)$ is available for each null hypothesis $H_{0,i}$. For $i \in \mathbb{N}_m$, each statistic $T_i(X)$ is transformed into a p-value $p_i(X)$, satisfying the following assumptions:

$$\forall i \in \mathcal{H}_0,\ \ \forall t \in [0,1],\ \ \mathbb{P}\left(p_i(X) \le t\right) \le t; \tag{Superunif}$$

$$\{p_i(X)\}_{i \in \mathcal{H}_0} \text{ is a family of independent } p\text{-values and is independent of } \{p_i(X)\}_{i \in \mathcal{H}_1}. \tag{Indep}$$

Extending our results to the case where (Indep) fails is possible, see the discussion in Section 6.

2.2. Classical post hoc bounds

As argued in BNR, computing the optimal post hoc bound (3) relative to a given reference family $(R_k, \zeta_k)_{k \in \mathcal{K}}$ can be NP-hard, and simpler, more conservative versions can be provided, that is, bounds V such that for all $S \subset \mathbb{N}_m$, $V^*_{\mathfrak{R}}(R) \leq V(R)$. A simple upper-bound for $V^*_{\mathfrak{R}}$ is given by

$$\overline{V}_{\mathfrak{R}}(S) = |S| \wedge \min_{k \in \mathcal{K}} \{\zeta_k + |S \setminus R_k|\}\,, \quad S \subset \mathbb{N}_m\,. \tag{4}$$

It is straightforward to check that

$$V^*_{\mathfrak{R}}(S) \leq \overline{V}_{\mathfrak{R}}(S), \quad S \subset \mathbb{N}_m. \tag{5}$$

While this inequality is strict in general, BNR established that it is an equality if the reference family is nested, that is,

$$\mathcal{K} = \{1, \dots, K\} \text{ and } R_k \subset R_{k+1} \text{ for } 1 \leq k \leq K-1. \tag{Nested}$$

Condition (Nested) is mild when the sequence ζ_k is nondecreasing, e.g., $\zeta_k = k - 1$.

A consequence of (5) is that $\overline{V}_{\mathfrak{R}}$ is a post hoc bound in the sense of (1) as soon as the reference family $\mathfrak{R}$ is such that (2) holds. A simple union bound under (Superunif) yields that (2) holds with $\mathfrak{R} = \{(R_1, \zeta_1)\}$, $R_1 = \{i \in \mathbb{N}_m \,:\, p_i \leq \alpha/m\}$, $\zeta_1 = 0$. This leads to the Bonferroni post hoc bound

$$V_{\text{Bonf}}(S) = \sum_{i \in S} \mathbb{1}\{p_i(X) > \alpha/m\}, \quad S \subset \mathbb{N}_m. \tag{6}$$

The more subtle Simes inequality (Simes, 1986), valid under (Superunif)–(Indep), ensures that (2) holds with $\mathfrak{R} = \{(R_k, \zeta_k), 1 \leq k \leq m\}$, $R_k = \{i \in \mathbb{N}_m \,:\, p_i \leq \alpha k/m\}$, $\zeta_k = k - 1$. This leads to the Simes post hoc bound

$$V_{\text{Simes}}(S) = \min_{1 \leq k \leq m} \left\{ \sum_{i \in S} \mathbb{1}\{p_i(X) > \alpha k/m\} + k - 1 \right\}, \quad S \subset \mathbb{N}_m. \tag{7}$$

As noted in BNR, this bound is identical to post hoc bound of Goeman and Solari (2011), which will be used as a benchmark in this paper.

2.3. Improved interpolation bound

When the sequence ζ_k is not nondecreasing, inequality (5) can be far too conservative. We introduce the following extension: for a reference family $\mathfrak{R} = (R_k(X), \zeta_k(X))_{k \in \mathcal{K}}$ of cardinal $K = |\mathcal{K}|$,

$$\widetilde{V}^q_{\mathfrak{R}}(S) = \min_{Q \subset \mathcal{K}, |Q| \leq q} \left(\sum_{k \in Q} \zeta_k \wedge |S \cap R_k| + \left| S \setminus \bigcup_{k \in Q} R_k \right| \right), \quad 1 \leq q \leq K, \quad S \subset \mathbb{N}_m\,; \tag{8}$$

$$\widetilde{V}_{\mathfrak{R}}(S) = \widetilde{V}^K_{\mathfrak{R}}(S), \quad S \subset \mathbb{N}_m\,. \tag{9}$$

Obviously, we have $\widetilde{V}^1_{\mathfrak{R}} = \overline{V}_{\mathfrak{R}}$ and $\widetilde{V}^q_{\mathfrak{R}}$ is non-increasing in q. The following result shows that these bounds are all conservative versions of $V^*_{\mathfrak{R}}$.

Lemma 2.1. *For any reference family $\mathfrak{R}$, we have*

$$V^*_{\mathfrak{R}}(S) \leq \widetilde{V}_{\mathfrak{R}}(S) \leq \widetilde{V}^q_{\mathfrak{R}}(S) \leq \overline{V}_{\mathfrak{R}}(S), \quad 1 \leq q \leq K, \quad S \subset \mathbb{N}_m. \tag{10}$$

In particular, if $\mathfrak{R}$ is such that (2) holds, then $\widetilde{V}_{\mathfrak{R}}$ is a post hoc bound in the sense of (1).

Lemma 2.1 is proved in Section 7.1. The inequality $V^*_{\mathfrak{R}}(S) \leq \widetilde{V}_{\mathfrak{R}}(S)$ in (10) is strict in general, see Example 2.2. As we will show in the next section, this relation is nevertheless an equality when $\mathfrak{R}$ has a specific forest structure, which makes $\widetilde{V}_{\mathfrak{R}}$ a particularly interesting bound.

Example 2.2. Let $m = 4$, $K = 3$, $R_1 = \{1, 2, 4\}$, $R_2 = \{2, 3, 4\}$, $R_3 = \{1, 3, 4\}$. Consider the event where $\zeta_1(X) = \zeta_2(X) = \zeta_3(X) = 1$. For $S = \mathbb{N}_4$, we easily check that $V^*_{\mathfrak{R}}(S) = 1$ and $\widetilde{V}_{\mathfrak{R}}(S) = 2$.

3. Post hoc bound for forest structured reference family

3.1. Forest structure

Definition 3.1. A reference family $\mathfrak{R} = (R_k, \zeta_k)_{k\in\mathcal{K}}$ is said to have a forest structure if following property is satisfied:

$$\forall k, k' \in \mathcal{K},\ \ R_k \cap R_{k'} \in \{R_k, R_{k'}, \varnothing\}, \tag{Forest}$$

that is, two elements of $\{R_k\}_{k\in\mathcal{K}}$ are either disjoint or nested.

The forest structure is general enough to cover a wide range of different situations, as for instance the disjoint case

$$\forall k, k' \in \mathcal{K},\ \ k \neq k' \Rightarrow R_k \cap R_{k'} = \varnothing. \tag{Disjoint}$$

and the nested case (Nested). In general, if each R_k is considered as a node and if an oriented edge $R_k \leftarrow R_{k'}$ is depicted between two different sets R_k and $R_{k'}$ if and only if $R_k \subset R_{k'}$ and there is no $R_{k''}$ such that $R_k \subsetneq R_{k''} \subsetneq R_{k'}$; the obtained graph correspond to a (directed) forest in the classical graph theory sense, see e.g. Kolaczyk (2009). An illustration is given in Figure 1. The positions of the nodes in this picture rely on the depth of $\mathfrak{R}$, which can be defined as the function

$$\phi\ :\ \left\{\begin{array}{ccl} \mathcal{K} & \to & \mathbb{N}^* \\ k & \mapsto & 1 + \left|\{k' \in \mathcal{K} : R_{k'} \supsetneq R_k\}\right|. \end{array}\right. \tag{11}$$

For instance, under (Disjoint), $\phi(k) = 1$ for all $k \in \mathcal{K}$, while under (Nested), $\phi(k) = K + 1 - k$ for all $1 \leq k \leq K$.

Example 3.2. Let $m = 25$, $R_1 = \{1, \dots, 20\}$, $R_2 = \{1, 2\}$, $R_3 = \{3, \dots, 10\}$, $R_4 = \{11, \dots, 20\}$, $R_5 = \{5, \dots, 10\}$, $R_6 = \{11, \dots, 16\}$, $R_7 = \{17, \dots, 20\}$, $R_8 = \{21, 22\}$, $R_9 = \{22\}$. Then the corresponding reference family $\mathfrak{R} = (R_k, \zeta_k)_{1\leq k\leq 9}$ satisfies (Forest). The sets R_1, R_8 are of depth 1; the sets R_2, R_3, R_4, R_9 are of depth 2; the sets R_5, R_6, R_7 are of depth 3.

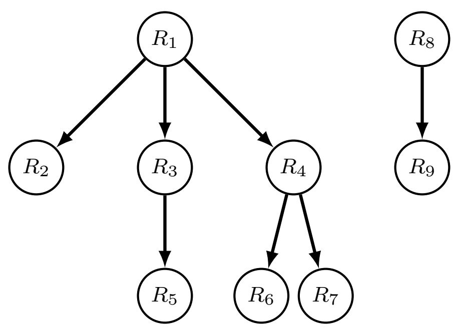

FIG 1. *Graph corresponding to the reference family given in Example 3.2.*

A useful characterization of a forest-structure reference family is given in the next lemma.

Lemma 3.3. *For any reference family $\mathfrak{R} = (R_k, \zeta_k)_{k\in\mathcal{K}}$ having the structure (Forest), there exists a partition $(P_n)_{1\leq n\leq N}$ of $\mathbb{N}_m$ such that for each $k \in \mathcal{K}$, there exists some (i, j) with $1 \leq i \leq j \leq N$ and $R_k = P_{i:j}$, where we denote*

$$P_{i:j} = \bigcup_{i\leq n\leq j} P_n, \qquad 1 \leq i \leq j \leq N. \tag{12}$$

Conversely, for some partition $(P_n)_{1\leq n\leq N}$ of $\mathbb{N}_m$, consider any reference family of the form $\mathfrak{R} = (P_{i:j}, \zeta_{i,j})_{(i,j)\in\mathcal{C}}$ with $\mathcal{C} \subset \{(i, j) \in \mathbb{N}_N^2\ :\ i \leq j\}$ such that for $(i, j), (i', j') \in \mathcal{C}$, we have

$$[\![i, j]\!] \cap [\![i', j']\!] = \varnothing;\ or\ [\![i, j]\!] \subset [\![i', j']\!];\ or\ [\![i', j']\!] \subset [\![i, j]\!]\ ,$$

where $[\![i, j]\!]$ denotes the set of all integers between i and j. Then $\mathfrak{R}$ has the structure (Forest).

For the ease of notation, the set $\mathcal{C}$ will be identified to $\mathcal{K}$ throughout the paper, which leads to the following slight abuse: denoting indifferently $k \in \mathcal{K}$ or $(i,j) \in \mathcal{K}$, and

$$\mathfrak{R} = (R_k, \zeta_k)_{k\in\mathcal{K}} \quad \text{or} \quad \mathfrak{R} = (P_{i:j}, \zeta_{i,j})_{(i,j)\in\mathcal{K}}. \tag{13}$$

We call "atoms" the elements of the underlying partition $(P_n)_{1\leq n\leq N}$ because they have the thinnest granularity in the structure and because any subset R_k of the family can be expressed as a combination of these atoms. Note however that this partition is not unique. A simple algorithm to compute $(P_n)_n$ and the proof of Lemma 3.3 are provided in Appendix B. An example of such a partition is given in Example 3.4 and Figure 2.

Example 3.4. For the reference family given in Example 3.2, a partition as in Lemma 3.3 is given by $P_1 = R_2$, $P_2 = R_3 \setminus R_5$, $P_3 = R_5$, $P_4 = R_6$, $P_5 = R_7$, $P_6 = R_8 \setminus R_9$, $P_7 = R_9$, $P_8 = \mathbb{N}_m \setminus \{R_1 \cup R_8\}$.

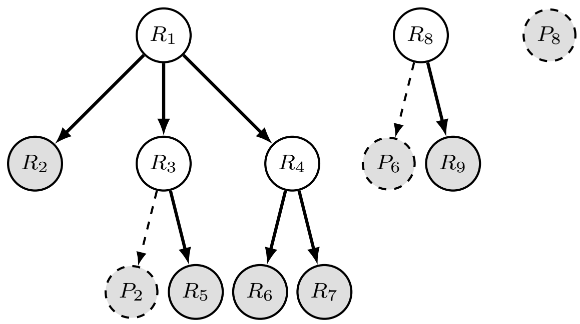

FIG 2. *Graph corresponding to the reference family given by Example 3.2, with the associated partition (atoms) $\{P_n, 1 \leq n \leq N\}$, displayed by light gray nodes and given in Example 3.4. The nodes that correspond to atoms that are not in the reference family are depicted with a dashed circle.*

An important particular case in our analysis is the case where the forest structure includes all atoms, that is

$$\forall n \in \{1, \dots, N\},\ \ P_n \in \{R_k, k \in \mathcal{K}\}. \tag{All-atoms}$$

When (All-atoms) does not hold (as in Example 3.4), we can impose this condition by adding them to the structure, building in this way the completed reference family:

Definition 3.5. Consider any reference family $\mathfrak{R} = (P_{i:j}, \zeta_{i,j})_{(i,j)\in\mathcal{K}}$ satisfying (Forest) and associated to atoms $(P_n)_{1\leq n\leq N}$ by (13). Let $\mathcal{K}^+ = \{(i,i), 1 \leq i \leq N : (i,i) \notin \mathcal{K}\}$, $\zeta_{i,i} = |P_{i:i}| = |P_i|$ for all $(i,i) \in \mathcal{K}^+$, and $\mathcal{K}^\oplus = \mathcal{K} \cup \mathcal{K}^+$. Then the completed version of $\mathfrak{R}$ is given by $\mathfrak{R}^\oplus = (P_{i:j}, \zeta_{i,j})_{(i,j)\in\mathcal{K}^\oplus}$.

For the reference family $\mathfrak{R}$ given by Example 3.2, the completed version $\mathfrak{R}^\oplus$ is depicted in Figure 3.

3.2. Deriving the optimal post hoc bound

The next result shows that the expression of the optimal post hoc bound $V^*_{\mathfrak{R}}$ can be simplified when $\mathfrak{R}$ satisfies (Forest).

Theorem 3.6. *Let $\mathfrak{R}$ be a reference family having the structure* (Forest)*. Then the optimal bound $V^*_{\mathfrak{R}}$* (3) *can be derived from the bounds $\widetilde{V}^q_{\mathfrak{R}}$* (8) *and $\widetilde{V}_{\mathfrak{R}}$* (9) *in the following way:*

$$V^*_{\mathfrak{R}}(S) = \widetilde{V}_{\mathfrak{R}}(S), \quad S \subset \mathbb{N}_m; \tag{14}$$

$$V^*_{\mathfrak{R}}(S) = \widetilde{V}^d_{\mathfrak{R}}(S), \quad S \subset \mathbb{N}_m, \tag{15}$$

where d is the maximum number of disjoint sets that can be found in the reference family, that is,

$$d = \max\{|Q|, Q \subset \mathcal{K} : \forall k, k' \in Q,\ \ k \neq k' \Rightarrow R_k \cap R_{k'} = \varnothing\}.$$

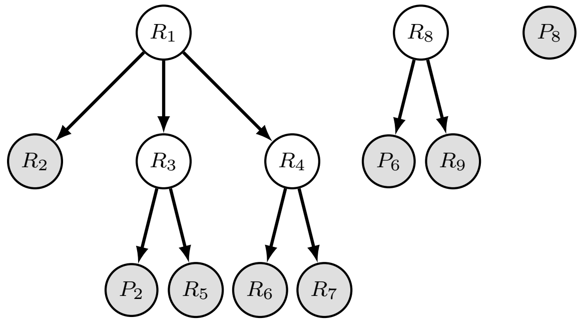

FIG 3. *Graph corresponding to the completed version $\mathfrak{R}^{\oplus}$ of the reference family $\mathfrak{R}$ given by Example 3.2 with the atoms given in Example 3.4.*

A byproduct of Theorem 3.6 is that, if (Nested) holds, $V^*_{\mathfrak{R}} = \widetilde{V}^1_{\mathfrak{R}}(S) = \overline{V}_{\mathfrak{R}}$ and we recover Proposition 2.5 of BNR. Another interesting case is the structure (Disjoint), where $\widetilde{V}_{\mathfrak{R}}$ has a simpler form. This is summarized in the following result.

Corollary 3.7. *Let $\mathfrak{R}$ be a reference family.*

(i) if $\mathfrak{R}$ satisfies (Nested), then $V^*_{\mathfrak{R}} = \overline{V}_{\mathfrak{R}}$.
(ii) if $\mathfrak{R}$ satisfies (Disjoint), then $V^*_{\mathfrak{R}}(S) = \sum_{k=1}^K \zeta_k \wedge |S \cap R_k| + \left|S \setminus \bigcup_{k=1}^K R_k\right|, \quad S \subset \mathbb{N}_m$.

Theorem 3.6 and Corollary 3.7 are respectively proved in Section 7.2 and Section 7.3.

The proof of Theorem 3.6 being constructive, it provides an algorithm to compute easily $V^*_{\mathfrak{R}}(S)$, that we now describe. Let us first introduce an additional piece of notation. For some reference family $\mathfrak{R} = (P_{i:j}, \zeta_{i,j})_{(i,j)\in\mathcal{K}}$ of depth function ϕ (see (11)), we denote

$$\mathcal{K}^h = \{(i,j) \in \mathcal{K} : \phi(i,j) = h \text{ or } (i = j \text{ and } \phi(i,i) \le h)\}, \quad h \ge 1.$$

Hence, each $\mathcal{K}^h$ contains the indexes of the sets of depth h and also the atoms with an inferior depth. Figure 4 displays some $\mathcal{K}^h$ for the reference family of Example 3.2.

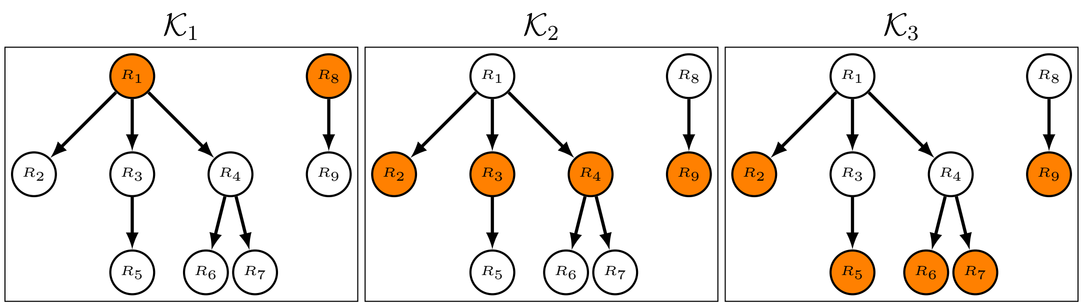

FIG 4. *Display of the nodes corresponding to $\mathcal{K}^1$, $\mathcal{K}^2$, $\mathcal{K}^3$ (in orange) for the reference family given in Example 3.2.*

Algorithm 1 below gives the steps to compute $V^*_{\mathfrak{R}}(S)$: first, complete the family $\mathfrak{R}$ by adding all the members of the partition, as explained in Definition 3.5, in order to get $\mathfrak{R}^{\oplus}$. By Lemma A.4, we have $V^*_{\mathfrak{R}^{\oplus}}(S) = V^*_{\mathfrak{R}}(S)$, so that this operation does not change the targeted quantity. In particular, (All-atoms) holds after this step. Second, the algorithm uses a reverse loop, which successively updates a vector V whose components correspond to active nodes; the current value of the bound is equal to the sum of the components of V. Each step of the loop will update the value of V to make the bound possibly smaller, to obtain at the end $V^*_{\mathfrak{R}}(S)$. The time complexity of the Algorithm 1 for a given S is $O(Hm)$, where $H = \max_{k\in\mathcal{K}} \phi(k)$ is the maximal depth of the reference family. where ϕ is the depth function defined by (11).

Let us describe the loop in more detail by using the particular situation of Figure 5. Initialization: $H = 3$ and $\mathcal{K}^H = \mathcal{K}^3$, which corresponds to the active nodes in the rightmost graph. Hence, V is equal to the vector of values $\zeta_k \wedge |S \cap R_k|$ among these nodes. First step: $h = 2$ hence $\mathcal{K}^h = \mathcal{K}^2$, for which the active nodes are displayed in the middle graph. Each of these nodes $k \in \mathcal{K}^2$, gives a bound $\zeta_k \wedge |S \cap R_k|$ that should be compared with the one of the previous step, that is, $\sum_{k' \in Succ_k} V_{k'}$, where $Succ_k$ denotes the offspring of R_k. The vector V is defined by the best choice among these two. Second (and final) step: $h = 1$ hence $\mathcal{K}^h = \mathcal{K}^1$ (leftmost graph) which only contains the roots of the forest and where V is updated following the same process. The algorithm then returns $V^*_{\mathfrak{R}}(S) = \sum_{k \in \mathcal{K}^1} V_k$.

Algorithm 1: Computation of $V^*_{\mathfrak{R}}(S)$

Data: $\mathfrak{R} = (P_{i:j}, \zeta_{i,j})_{(i,j)\in\mathcal{K}}$ and $S \subset \mathbb{N}_m$.
Result: $V^*_{\mathfrak{R}}(S)$.

$\mathfrak{R} \longleftarrow \mathfrak{R}^{\oplus}$; $\mathcal{K} \longleftarrow \mathcal{K}^{\oplus}$ (completion, see Definition 3.5);
$H \longleftarrow \max_{k\in\mathcal{K}} \phi(k)$, see (11);
$V \longleftarrow (\zeta_k \wedge |S \cap R_k|)_{k\in\mathcal{K}^H}$;
for $h \in \{H-1, \dots, 1\}$ **do**
  $newV \longleftarrow (0)_{k\in\mathcal{K}^h}$;
  for $k \in \mathcal{K}^h$ **do**
    $Succ_k \longleftarrow \{k' \in \mathcal{K}^{h+1} : R_{k'} \subset R_k\}$;
    $newV_k \longleftarrow \min\left(\zeta_k \wedge |S \cap R_k|, \sum_{k' \in Succ_k} V_{k'}\right)$;
  end
  $V \longleftarrow newV$;
end
return $\sum_{k\in\mathcal{K}^1} V_k$.

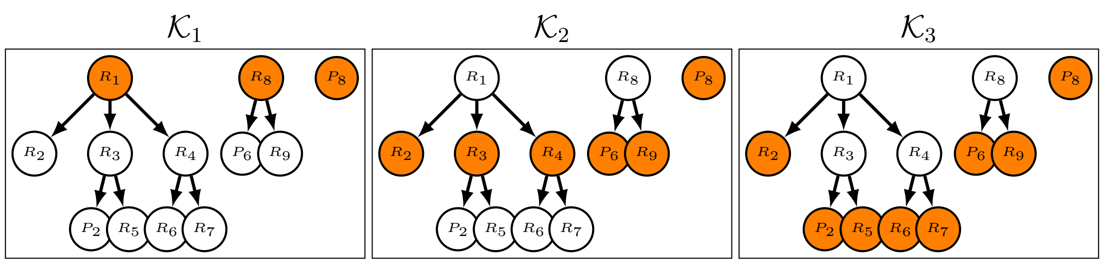

FIG 5. *Same as Figure 4 but for the completed version.*

4. Local calibration of the reference family

In this section, we explain how to build a reference family $\mathfrak{R}$ such that (2) holds. The results presented in this section hold for any deterministic $(R_k)_k$ and the calibration concerns only $(\zeta_k)_k$ here.

4.1. Calibration of ζ_k by DKW inequality

In this section, we estimate $|S \cap \mathcal{H}_0|$ by using an approach close in spirit to the so-called Storey estimator (Storey, 2002). The latter depends on a parameter, denoted by t here, that has to be chosen appropriately (see Blanchard and Roquain, 2009 for a discussion on this issue). To avoid this caveat while improving accuracy, we can derive an estimator uniform on t by using the DKW inequality (Dvoretzky et al., 1956), with the optimal constant of Massart (1990).

For any deterministic subsets $R_k \subset \mathbb{N}_m$, $k \in \mathcal{K}$, $K = |\mathcal{K}|$, let

$$\zeta_k(X) = |R_k| \wedge \min_{t\in[0,1)} \left\lfloor \frac{C}{2(1-t)} + \left(\frac{C^2}{4(1-t)^2} + \frac{\sum_{i\in R_k} \mathbf{1}\{p_i(X) > t\}}{1-t} \right)^{1/2} \right\rfloor^2, \quad k \in \mathcal{K}, \tag{16}$$

where $C = \sqrt{\frac{1}{2}\log\left(\frac{K}{\alpha}\right)}$ and $\lfloor x \rfloor$ denotes the largest integer smaller than or equal to x.

Proposition 4.1. *Consider any deterministic (different) subsets $R_k \subset \mathbb{N}_m$, $k \in \mathcal{K}$ ($K = |\mathcal{K}|$) and assume $\alpha/K < 1/2$. Assume that for all $k \in \mathcal{K}$, the p-value family $\{p_i(X),\ i \in R_k\}$ satisfies (Superunif) and (Indep). Then the JER control (2) holds for the reference family $\mathfrak{R} = (R_k, \zeta_k(X))_{k\in\mathcal{K}}$, for which the local bounds ζ_k are given by (16).*

Combining Proposition 4.1 with Lemma 2.1, we obtain that, under the assumptions of Proposition 4.1, the bound

$$V_{\text{DKW}} = \widetilde{V}_{\mathfrak{R}} \text{ given by (9) with } \mathfrak{R} = (R_k, \zeta_k(X))_{k\in\mathcal{K}} \text{ and } \zeta_k(X) \text{ given by (16)}, \tag{17}$$

satisfies (1) and thus is a valid post hoc bound.

Proposition 4.1 is proved in Section 7.4. Note that $\zeta_k(X) \geq \lfloor \log(K/\alpha)/2 \rfloor \geq 1$ as soon as $\alpha \leq e^{-2}K$. Hence, this contrasts with previous approaches (Blanchard et al., 2017; Goeman and Solari, 2011), for which $\zeta_k = 0$ was included in the reference family. This means that using this reference family induces a minimum cost. In the next section, we will see that this cost is generally compensated by the accuracy of the joint estimation of $|R_k \cap \mathcal{H}_0|$, $k \in \mathcal{K}$.

Remark 4.2. In practice, $\zeta_k(X)$ in (16) can be computed as

$$\zeta_k(X) = s \wedge \min_{0\leq \ell \leq s} \left\lfloor \frac{C}{2(1-p_{(\ell)})} + \left(\frac{C^2}{4(1-p_{(\ell)})^2} + \frac{s-\ell}{1-p_{(\ell)}} \right)^{1/2} \right\rfloor^2,$$

where $s = |R_k|$ and $0 = p_{(0)} \leq p_{(1)} \leq \cdots \leq p_{(s)}$ are the ordered p-values of $\{p_i(X), i \in R_k\}$.

Remark 4.3. With our notation, the previous $(1-\alpha)$-confidence bound of Genovese and Wasserman (2004) (Equation (16) therein) corresponds to take

$$\zeta_k^{GW}(X) = |R_k| \wedge \min_{t\in[0,1)} \left\lfloor \frac{\sum_{i\in R_k} \mathbf{1}\{p_i(X) > t\} + |R_k|^{1/2} C}{1-t} \right\rfloor.$$

By using (33) in Lemma A.1 with $a = 1-t$, $b = C$, $c = \sum_{i\in R_k} \mathbf{1}\{p_i(X) > t\}$, and $d = |R_k|$, we can see that the quantity $\zeta_k^{GW}(X)$ is always larger than the $\zeta_k(X)$ given by (16). Hence our result is a uniform improvement of Genovese and Wasserman (2004).

Remark 4.4. The local bounds ζ_k in (16) depend on the target level α only through C, where $2C^2 = \log(K/\alpha)$. Therefore, the post hoc bounds derived from Proposition 4.1 are expected to depend only weakly on α. This important point is illustrated in our numerical experiments (Section 5), where this property is used to propose a hybrid post hoc bound taking the best of both the Simes and the DKW-based bounds.

4.2. Comparison to existing post hoc bounds

To explore the benefit of the new reference family when the signal is localized, let us consider a stylized model where the signal is localized according to a regular partition

$$R_k = \{1 + (k-1)s, \ldots, ks\},\ \ 1 \leq k \leq K, \tag{18}$$

composed of K regions of equal size s. In particular, this reference family satisfies (Disjoint). Among the regions R_k, only R_1 contains false nulls, and $r \in (0,1)$ denotes the proportion of signal in R_1, that is

$$r = |R_1 \cap \mathcal{H}_1|/|R_1|. \tag{19}$$

The remaining regions contain no signal, that is $|R_k \cap \mathcal{H}_1| = 0$, for $k \geq 2$.

In addition, we consider an independent Gaussian one-sided setting where the false nulls have mean $\mu > 0$, that is, we assume that $X_i \sim \mathcal{N}(0,1)$ if $i \in \mathcal{H}_0$ and $X_i \sim \mathcal{N}(\mu,1)$ if $i \in \mathcal{H}_1$, and the p-values are derived as $p_i(X) = \bar{\Phi}(X_i)$, $i \in \mathbb{N}_m$, where $\bar{\Phi}$ denotes the upper-tail of the standard normal distribution.

Proposition 4.5. *Let us consider the post hoc bounds V_{Bonf} (6); V_{Simes} (7) and the new post hoc bound V_{DKW} given by (17) and associated to the reference regions R_k defined above. In the setting defined above, we have*

$$\frac{\mathbb{E}(V_{\text{DKW}}(R_1))}{|R_1|} \leq 1 \wedge \left(1 - r + 2r\,\overline{\Phi}(\mu) + \frac{4C}{\sqrt{s}}\left(1 + \frac{C}{\sqrt{s}}\right)\right) \tag{20}$$

$$\frac{\mathbb{E}(V_{\text{Simes}}(R_1))}{|R_1|} \geq (1-r)(1-\alpha s/m) + r\,\overline{\Phi}(\mu - \overline{\Phi}^{-1}(\alpha s/m)); \tag{21}$$

$$\frac{\mathbb{E}(V_{\text{Bonf}}(R_1))}{|R_1|} = (1-r)(1-\alpha/m) + r\,\overline{\Phi}(\mu - \overline{\Phi}^{-1}(\alpha/m)). \tag{22}$$

This proposition is proved in Section 7.5. In particular, combining (20) and (21) yields

$$\frac{\mathbb{E}(V_{\text{DKW}}(R_1))}{\mathbb{E}(V_{\text{Simes}}(R_1))} \leq \frac{1 \wedge \left(1 - r + 2r\,\overline{\Phi}(\mu) + \frac{4C}{\sqrt{s}}\left(1 + \frac{C}{\sqrt{s}}\right)\right)}{(1-r)(1-\alpha s/m) + r\,\overline{\Phi}(\mu - \overline{\Phi}^{-1}(\alpha s/m))}. \tag{23}$$

This ratio is displayed in Figure 6 for a choice of model parameters. The new bound can substantially improve the Simes bound over a wide range of effect sizes.

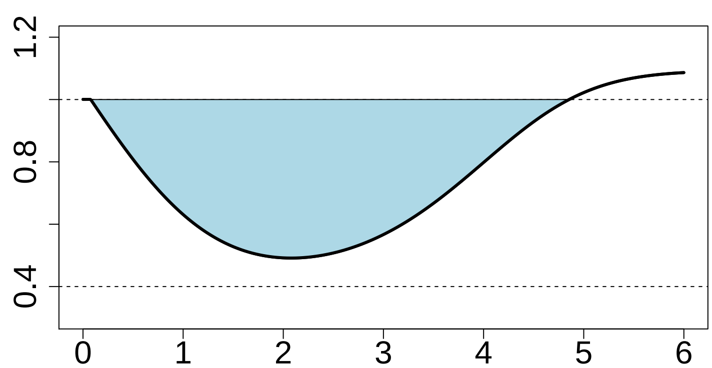

Fig 6. *Y-axis: upper bound of the ratio between the new bound and the Simes bound, see (23). X-axis: effect size μ. $m = 10^7$, $s = m^{2/3}$, $K = m/s$, $r = 3/5$, $\alpha = 0.1$.*

This improvement can also be put forward by an asymptotic approach.

Corollary 4.6. *Let us consider the framework of Proposition 4.5. In the asymptotic setting in m where s tends to infinity with $s \gg \log K$ and μ tends to infinity with $\mu - \overline{\Phi}^{-1}(\alpha/m) \to -\infty$, we have*

$$\limsup_m \left\{\frac{\mathbb{E}(V_{\text{DKW}}(R_1))}{|R_1|}\right\} \leq 1 - r, \quad \textit{and} \quad \limsup_m \left\{\frac{\mathbb{E}(V_{\text{Bonf}}(R_1))}{|R_1|}\right\} = 1.$$

If moreover $s \ll m$ (i.e., $K \to \infty$) and $\mu - \overline{\Phi}^{-1}(\alpha s/m) \to -\infty$, we have

$$\limsup_m \left\{\frac{\mathbb{E}(V_{\text{DKW}}(R_1))}{|R_1|}\right\} \leq 1 - r, \quad \textit{and} \quad \limsup_m \left\{\frac{\mathbb{E}(V_{\text{Simes}}(R_1))}{|R_1|}\right\} = 1.$$

In particular, this corollary establishes that the order of the new bound can improve the Simes bound by a factor $1 - r$.

5. Numerical experiments

5.1. Setting

In this section we perform numerical experiments to compare our new post hoc bound V_{DKW} (17) with Simes post hoc bound (7). Let q be some fixed integer, say larger than 1. We consider two versions of our new bound:

- The first version of our post hoc bound, denoted V_{part}, is defined by (17) in which the reference family $\mathfrak{R}^{\text{part}}$ is the regular partition of $\mathbb{N}_m$ given by (18) for $K^{\text{part}} = 2^q$ ($s = m/2^q$ being assumed to be an integer).
- The second version of our post hoc bound, denoted V_{tree}, is defined similarly by (17), but the reference family $\mathfrak{R}^{\text{tree}}$ is given this time by the perfect binary tree whose leaves are the elements of $\mathfrak{R}^{\text{part}}$. Hence, by using the notation of Lemma 3.3, this means $P_k = \{1 + (k - 1)s, \dots, ks\}$, $1 \le k \le 2^q$. The cardinal of the reference family is thus $K^{\text{tree}} = 2^{q+1} - 1$.

The true/false null hypothesis configuration is as follows: the false null hypotheses are contained in P_k for $1 \le k \le K_1$, for some fixed value of K_1. The quantity r is defined similarly as in (19), as the fraction of false null hypotheses in those P_k, and is set to $r \in \{0.5, 0.75, 0.9, 1\}$. All of the other partition pieces only contain true null hypotheses. Finally, the true null p-values are distributed as i.i.d. $\mathcal{N}(0,1)$, and false null p-values are distributed as i.i.d. $\mathcal{N}(\bar{\mu}, 1)$, where $\bar{\mu}$ is a fixed value in $\{2, 3, 4\}$. This construction is illustrated in Figure 7 for $q = 3$ (leading to $K^{\text{part}} = 8$ and $K^{\text{tree}} = 15$) and $K_1 = 2$. In our experiments, we have chosen $q = 7$ and $s = 100$ (corresponding to $K^{\text{part}} = 128$ and $K^{\text{tree}} = 255$ and $m = 12800$), and $K_1 = 8$.

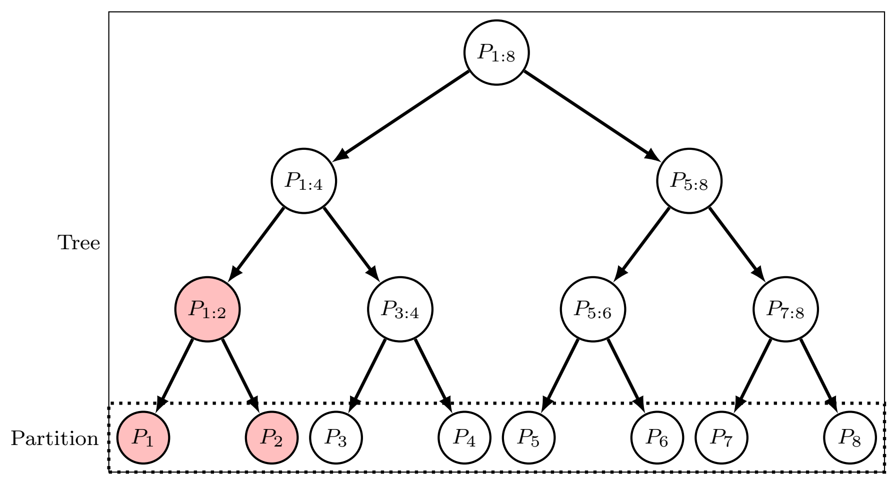

FIG 7. *Partition and perfect binary tree structures used in simulations, here with $q = 3$ and $K_1 = 2$ ($K^{\text{part}} = 8$ and $K^{\text{tree}} = 15$). The pink nodes are those containing some signal.*

We also performed numerical experiments with $s \in \{10, 20, 50\}$ and $K_1 \in \{1, 4, 16\}$, and with Poisson- and Gaussian-distributed $\bar{\mu}$. Because the results are qualitatively similar, we only report the above-described setting.

5.2. Comparing confidence envelopes

One possible way to evaluate the performance of post hoc bounds is to consider the associated confidence envelopes on the number of true discoveries among the most significant hypotheses. Formally, for $k = 1, \dots, m$, we let $S_k = \{i_1, \dots, i_k\}$, where i_j is the index of the j^{th} smallest p-value. Note that focusing on such sets is *a priori* favorable to the Simes bound, for which the reference family are among the S_k. In Figure 8, each panel corresponds to a particular choice of the model parameters d (in rows) and $\bar{\mu}$ (in columns). Each panel compares the actual number of true positives $(k - |\mathcal{H}_0 \cap S_k|)$, $k = 1, \ldots, m$ (labelled "Oracle") to post hoc bounds of the form $(k - V(S_k))$, $k = 1, \ldots, m$, where V is V_{Simes}, V_{part}, or V_{tree}. In this figure, the confidence level is set to $1 - \alpha = 95\%$.

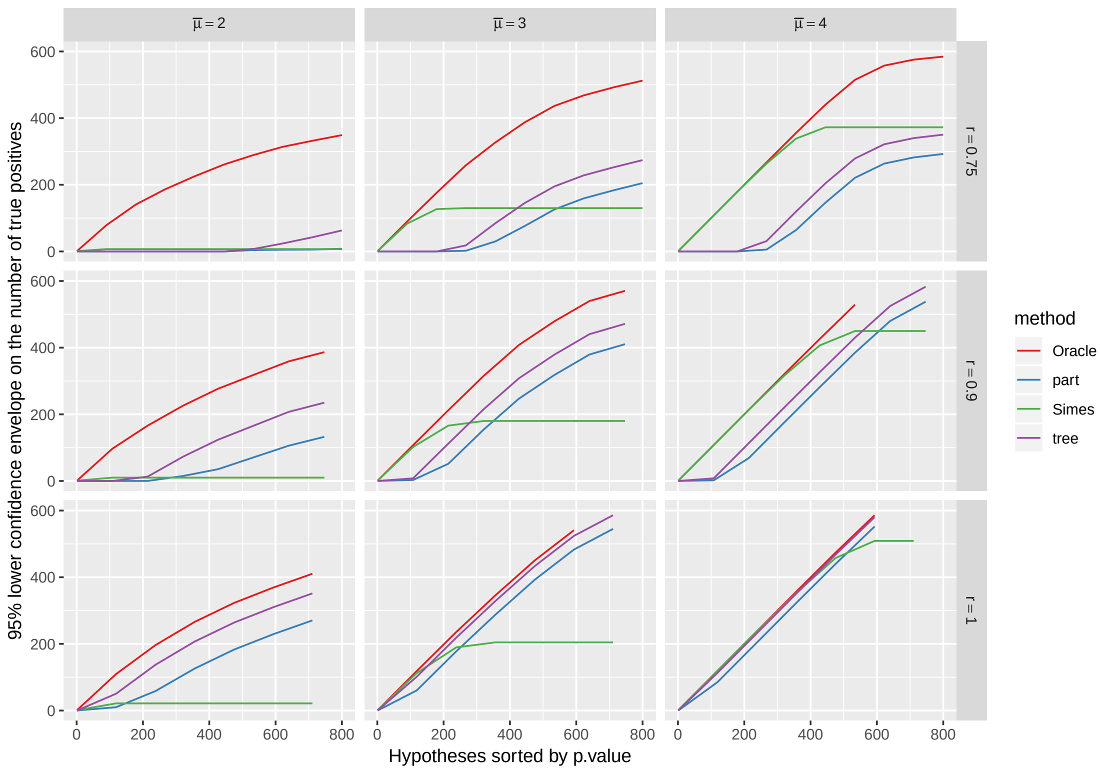

FIG 8. *95% lower confidence envelopes on the number of true positives obtained from Simes inequality and from the proposed methods are compared to the actual (Oracle) number of true positives.*

The chosen model parameters span a wide range of situations between very low and very high signal. For very low signal ($\bar{\mu} = 2, r = 0.75$, top-left panel), all the bounds are trivial, i.e. provide $V(S_k)$ close to $|S_k|\,(= k)$. As expected, all the bounds get sharper as the signal to noise ratio increases, that is, as $\bar{\mu}$ or r increase, and for very high signal ($\bar{\mu} = 4, r = 1$, bottom-right panel), all the bounds are very close to the actual number of true positives. The tree-based bound dominates the partition-based bound, which is expected because in this particular experiment, the regions P_k containing signal are adjacent (see Figure 7), and the multiscale nature of the tree-based bound allows it to take advantage of large-scale clusters. When the signal regions are not adjacent, these two bounds are very close (additional numerical experiments not shown). Our proposed bounds are more sensitive to the proportion of signal in each active region, while the Simes bound is more sensitive to the strength of the signal in those regions. As a result, none of the Simes and the "tree" bound is uniformly better than the other one. The Simes bound is typically sharper than the "tree" bound for small values of k, but becomes more conservative for larger values of k. This is expected, because the "tree" bound is based on *estimating the proportion of* non-null items, while the Simes bound is based on *pinpointing* non-null items.

5.3. Hybrid approach

An interesting question raised in Section 4.1 (Remark 4.4) is how these bounds are influenced by the target confidence level, which is fixed to $1 - \alpha = 95\%$ in Figure 8. In Figure 9 we compare the bounds obtained across values of α (corresponding to different line types) for $\bar{\mu} \in \{3, 4\}$ and $r \in \{0.75, 0.9\}$. The influence of α on the Simes bound is quite substantial. This is consistent with the shape of the bound (7), the p-values are directly compared to α. The influence of α on the bounds derived from (16) is much weaker, as expected from Remark 4.4. In particular, the envelopes derived from the "tree" method are very close to each other when α varies from 0.001 to 0.05. These striking differences suggest to introduce hybrid confidence envelopes that could take

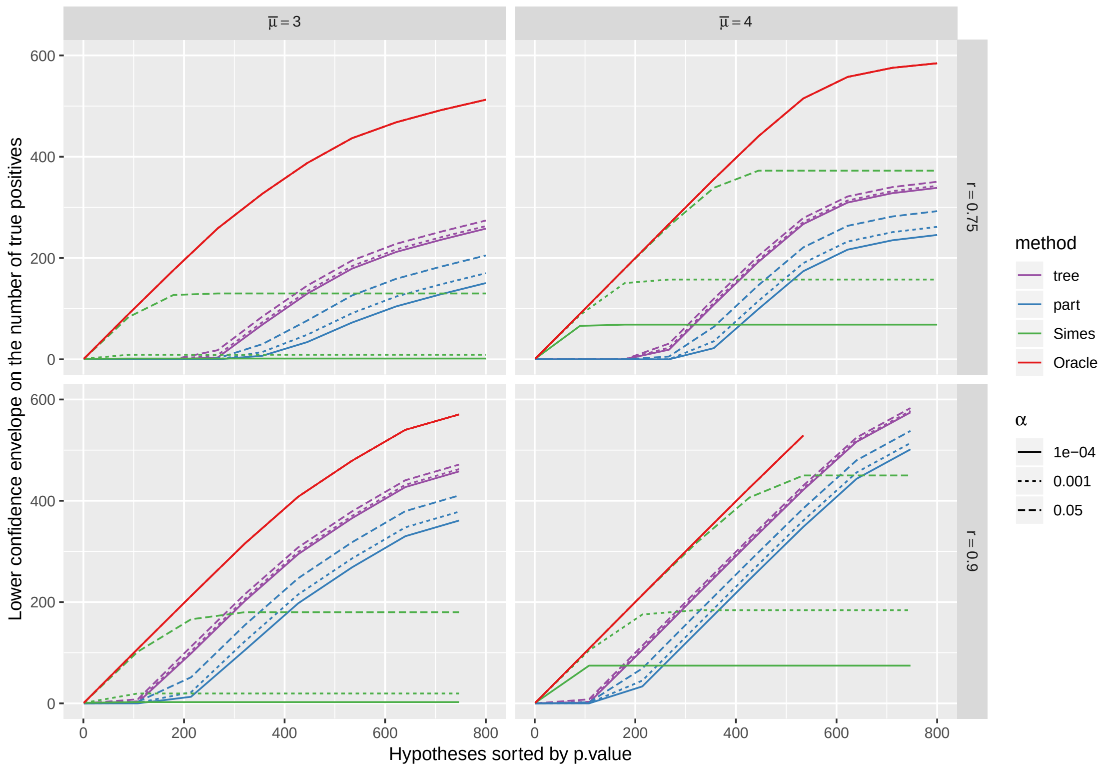

FIG 9. *Influence of the target level parameter α on upper confidence envelopes on the number of true positives.*

advantage of the superiority of the Simes bound on sets S_k for small k with that of the DKW-tree-based bound on sets S_k for larger k. For a fixed $\gamma \in [0, 1]$, let us define the bound $V^{\gamma}_{\text{hybrid}}$ as follows. For $S \subset \mathbb{N}_m$,

$$V^{\gamma}_{\text{hybrid}}(\alpha, S) = \min\left(V_{\text{Simes}}((1-\gamma)\alpha, S), V_{\text{tree}}(\gamma\alpha, S)\right),$$

where the notation in the bounds explicitly acknowledges the dependence of the bounds in the target level α. By an union bound, $V^{\gamma}_{\text{hybrid}}(\alpha, \cdot)$ is a $(1-\alpha)$-level post hoc bound. Figure 10 gives an illustration with $\alpha = 0.05$ and $\gamma = 0.02$. In this case, the hybrid envelope is the minimum of the Simes envelope at level $(1-\gamma)\alpha = 0.049$ and the DKW-tree-based envelope at level 0.001. Because $(1-\gamma)\alpha$ is very close to α, the confidence envelope $V^{0.02}_{\text{hybrid}}$ is essentially equivalent to the Simes-based confidence envelope for small k; for larger values of k, $V^{0.02}_{\text{hybrid}}$ is only slightly worse than the DKW-tree-based confidence envelope at level $\gamma\alpha = 0.001$.

6. Discussion

6.1. Comparison to Meijer et al. (2015)

Since our aim is similar to the one of Meijer et al. (2015) (denoted MKG below for short), let us make a short qualitative comparison between MKG and our study. First, while both approaches

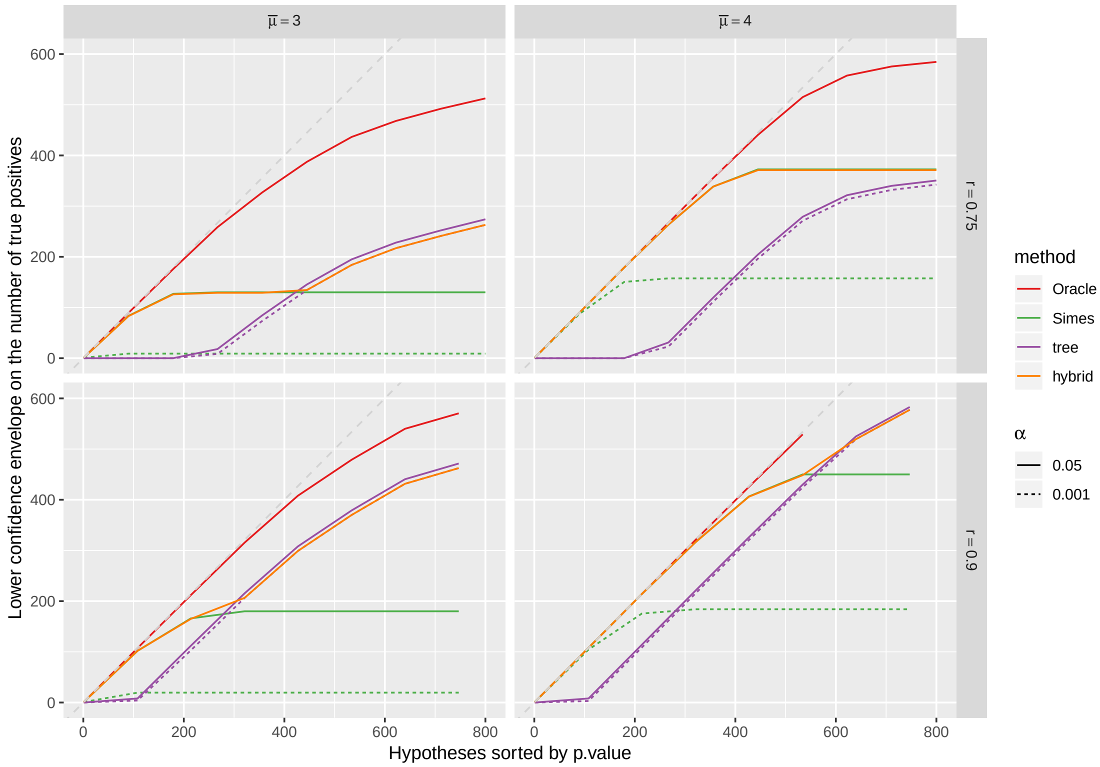

FIG 10. *Combining Simes and tree-based confidence envelopes on the number of true positives into a hybrid confidence envelope.*

are based on graph-structured subsets $\{R_k, k \in \mathcal{K}\}$, the geometrical shapes of the nodes R_k are different: the nodes in MKG correspond to all possible consecutive intervals, possibly overlapping, while our regions are based on partitioned regions at different resolutions. Our approach avoids redundancies of the tests but is suitable when the signal is structured according to the pre-specified partition structure, and may lead to a less accurate bound otherwise. This in turn impacts the way the local pieces of information are combined. The MKG approach uses a sequential, top-down algorithm, with an α-recycling method (that allows, for instance, to spend the same nominal level α both for a parent and its child). By contrast, our approach uses a bottom-up algorithm, with an overall nominal level adjusted by a simple overall union bound, which is generally conservative but seems fair here as the nodes are disjoint (at each resolution).

Second, the criteria used are different: MKG focus on simultaneous FWER control of local tests of intersections of null hypotheses $\cap_{i \in R_k} H_{0,i}$, $k \in \mathcal{K}$, while our statistical criterion ensures with high probability $|\mathcal{H}_0 \cap R_k| \leq \zeta_k$, for all $k \in \mathcal{K}$, for some bounds ζ_k. As already noted in BNR (see the supplementary file therein), the two approaches coincide when $\zeta_k = |R_k| - 1$, because $|\mathcal{H}_0 \cap R_k| > |R_k| - 1$ is equivalent to the fact that $\cap_{i \in R_k} H_{0,i}$ is true. Hence, a family $\{R_k, k \in \mathcal{K}\}$ violating $|\mathcal{H}_0 \cap R_k| > |R_k| - 1$ for some k will also wrongly reject $\cap_{i \in R_k} H_{0,i}$ for some k. However, when using another type of ζ_k, such as the DKWM device used here, such a connection is not valid anymore and the two criteria does not incorporate the local structure of the nodes in the same way. Here, using ζ_k's based on classical estimators will in principle lead to better post hoc bounds.

Third, within each node, the local statistics used are not of the same nature: in MKG, the local tests are based on a multivariate χ^2-type test, see Goeman et al. (2004). Here, we use an estimator relying on individual p-values that exploits the independence structure. This means that the assumptions made in MKG are much weaker, since it is valid under arbitrary dependence. Our approach can in principle also accommodate such a distributional setting, but this needs additional investigations, see the discussion in Section 6.2.

Finally, let us mention a setting for which the two methods can be fairly compared. First take the MKG method with Bonferroni local tests. As proved in MKG, the resulting FWER controlling procedure (reject the $H_{0,i}$ for which $V(\{i\}) = 0$) then reduces to the Holm procedure Holm (1979). By contrast, if we consider ζ_k equals to the number of accepted null hypotheses by the Holm procedure restricted to R_k (satisfying (Disjoint)), our methodology induces another overall FWER controlling procedure: simply the one rejecting all the null hypotheses rejected by the local Holm procedures. Both FWER controlling procedures are valid under arbitrary independence. Interestingly, if the signal is sparse but localized in one of the pre-specified R_k, the new procedure will dominate the Holm procedure (this is supported by a numerical experiment and a theoretical study, not reported here for short). This illustrates, once again, that our methodology can improve the state of the art, even in a very elementary framework.

6.2. Extension to general local confidence bounds

In this work, the local bounds ζ_k have been designed by using the DKW inequality. This can be straightforwardly extended to the case where the bound (16) is replaced by $\zeta_k(X) = L_k(\alpha/K)$, for which the function $L_k(\cdot)$ is a local bound satisfying the condition

$$\forall \lambda \in (0,1), \qquad \forall k \in \mathcal{K}, \qquad \forall P \in \mathcal{P}, \qquad \mathbb{P}_{X\sim P}\Big(|R_k \cap \mathcal{H}_0(P)| \leq L_k(\lambda)\Big) \leq \lambda\,. \tag{24}$$

The properties of the final post hoc bound will obviously depend on the choice of L_k. For instance, the validity of our post hoc bounds relies on (Indep), which is a strong assumption. The latter is only used to make the DKW inequality valid. If this assumption is violated, we should use another local bound L_k, that satisfies condition (24) under the specific dependence setting of the data. For instance, when the dependence is known or satisfies a randomization hypothesis (see Hemerik and Goeman, 2018), such a local bound can be easily constructed by applying the λ-calibration methodology of BNR (e.g., the one corresponding to the balanced template therein). However, the computational complexity of the final post hoc bound will substantially increase, which will make such an approach difficult to use in practice. Solving this problem seems challenging and is left for future work.

7. Proofs

7.1. Proof of Lemma 2.1

The second and third inequalities in (10) are straightforward from the fact that $\widetilde{V}^q_{\mathfrak{R}}$ is non-increasing in q and $\widetilde{V}^1_{\mathfrak{R}} = \overline{V}_{\mathfrak{R}}$. For the first inequality, let $S \subset \mathbb{N}_m$ and consider $A \subset \mathbb{N}_m$ such that $\forall k \in \mathcal{K}$, $|R_k \cap A| \leq \zeta_k$. For any $Q \subset \mathcal{K}$, we get

$$\begin{aligned}
|S \cap A| &\leq \sum_{k \in Q} |S \cap A \cap R_k| + \left| S \cap A \cap \Big(\bigcup_{k\in Q} R_k \Big)^{\mathsf{c}} \right| \\
&\leq \sum_{k\in Q} \zeta_k \wedge |S \cap R_k| + \left| S \setminus \bigcup_{k \in Q} R_k \right|,
\end{aligned}$$

which implies the result.

7.2. Proof of Theorem 3.6

In this proof, we fix $S \subset \mathbb{N}_m$. Also, we let

$$\mathcal{A}(\mathfrak{R}) = \{ A \subset \mathbb{N}_m : \forall k \in \mathcal{K},\ |R_k \cap A| \leq \zeta_k \}\,, \tag{25}$$

so that $V^*_{\mathfrak{R}}(S) = \max_{A\in\mathcal{A}(\mathfrak{R})} |S\cap A|$. Also note that (8)–(9) can be rewritten as

$$\widetilde{V}_{\mathfrak{R}}(S) = \min_{\mathcal{K}'\subset\mathcal{K}} \left(\sum_{k\in\mathcal{K}'} \zeta_k \wedge |S\cap R_k| + \left| S \setminus \bigcup_{k\in\mathcal{K}'} R_k \right| \right). \tag{26}$$

7.2.1. Proof of (14)

First, by Lemma A.4, it is sufficient to prove (14) for $\mathfrak{R}^{\oplus}$. Hence, we can focus without generality on the case where (All-atoms) holds. Recall that this means that $(i,i)\in\mathcal{K}$ for all $1\le i\le N$. Now, to prove that $\widetilde{V}_{\mathfrak{R}}(S) = V^*_{\mathfrak{R}}(S)$, it suffices to build $A\subset S$ such that $A\in\mathcal{A}(\mathfrak{R})$ and $|A| = \widetilde{V}_{\mathfrak{R}}(S)$. The key point is that for any h, A is the disjoint union of the $A\cap R_k$, $k\in\mathcal{K}^h$, because the R_k, $k\in\mathcal{K}^h$, form a partition of $\mathbb{N}_m$ (by Lemma A.2). Let $H = \max_{k\in\mathcal{K}}\phi(k)$ be the greater depth of the Forest structure, we will construct A with a decreasing recursion over $h\in\{1,\dots,H\}$. To this end, we need some additional notation: first, for any $k\in\mathcal{K}$, let $\mathcal{K}_k = \{k'\in\mathcal{K} : R_{k'}\subset R_k\}$ be the set of indexes of elements that are subsets of R_k. Then, for any h, let $\mathcal{K}^{\ge h} = \bigcup_{h\le h'\le H}\mathcal{K}^{h'}$. Note that $\mathcal{K}^{\ge 1} = \mathcal{K}$. Finally let

$$\mathfrak{P}^h = \{\mathcal{P}\subset\mathcal{K}^{\ge h} :\ \text{the } R_k,\ k\in\mathcal{P},\ \text{form a partition of } \mathbb{N}_m\},$$

and note that the result of Lemma A.3 (that is, equation (34)) can be rewritten in

$$\widetilde{V}_{\mathfrak{R}}(S) = \min_{\mathcal{P}\in\mathfrak{P}^1} \sum_{k\in\mathcal{P}} \zeta_k\wedge|S\cap R_k|. \tag{27}$$

The decreasing recursion starts like this: noting that $\mathcal{K}^H$ is the set of all the (i,i)'s, $1\le i\le N$, we define A^H by choosing (arbitrarily) $\zeta_{i,i}\wedge|S\cap P_{i:i}|$ distinct elements of $S\cap P_{i:i}$ for each $1\le i\le N$. Note that we have both

$$\forall k\in\mathcal{K}^{\ge H},\ \ |A^H\cap R_k|\le\zeta_k,$$

and

$$|A^H| = \sum_{k\in\mathcal{K}^H}\zeta_k\wedge|S\cap R_k| = \min_{\mathcal{P}\in\mathfrak{P}^H}\sum_{k\in\mathcal{P}}\zeta_k\wedge|S\cap R_k|,$$

since $\mathfrak{P}^H = \{\mathcal{K}^H\}$.

Now let h be given and assume we have constructed an $A^{h+1}\subset S$ such that both

$$\forall k\in\mathcal{K}^{\ge h+1},\ \ |A^{h+1}\cap R_k|\le\zeta_k,$$

and

$$\begin{aligned}|A^{h+1}| &= \min_{\mathcal{P}\in\mathfrak{P}^{h+1}}\sum_{k\in\mathcal{P}}\zeta_k\wedge|S\cap R_k| \\ &= \sum_{k\in\mathcal{P}^{h+1}}\zeta_k\wedge|S\cap R_k|,\end{aligned} \tag{28}$$

for a given $\mathcal{P}^{h+1}\in\mathfrak{P}^{h+1}$. Using that $|A^{h+1}| = \sum_{k\in\mathcal{P}^{h+1}}|A^{h+1}\cap R_k|$ and that $|A^{h+1}\cap R_k|\le\zeta_k\wedge|S\cap R_k|$ for all $k\in\mathcal{P}^{h+1}$, we deduce that $|A^{h+1}\cap R_k| = \zeta_k\wedge|S\cap R_k|$ for all $k\in\mathcal{P}^{h+1}$.

Now we want to construct A^h by defining all the $A^h\cap R_k$, $k\in\mathcal{K}^h$. By writing that $R_k = \bigcup_{k'\in\mathcal{P}^{h+1}\cap\mathcal{K}_k} R_{k'}$, the union being disjoint, we have first that, for all $k\in\mathcal{K}^h$,

$$\begin{aligned}|A^{h+1}\cap R_k| &= \sum_{k'\in\mathcal{P}^{h+1}\cap\mathcal{K}_k}|A^{h+1}\cap R_{k'}| \\ &= \sum_{k'\in\mathcal{P}^{h+1}\cap\mathcal{K}_k}\zeta_{k'}\wedge|S\cap R_{k'}|.\end{aligned}$$

Second, we have that:

$$\min_{\mathcal{P}\in\mathfrak{P}^h}\sum_{k\in\mathcal{P}}\zeta_k\wedge|S\cap R_k| = \sum_{k\in\mathcal{K}^h}\min_{\mathcal{P}\in\mathfrak{P}^h}\left(\sum_{k'\in\mathcal{P}\cap\mathcal{K}_k}\zeta_{k'}\wedge|S\cap R_{k'}|\right) \tag{29}$$

$$= \sum_{k\in\mathcal{K}^h}\min\left(\zeta_k\wedge|S\cap R_k|, \min_{\mathcal{P}\in\mathfrak{P}^{h+1}}\left(\sum_{k'\in\mathcal{P}\cap\mathcal{K}_k}\zeta_{k'}\wedge|S\cap R_{k'}|\right)\right) \tag{30}$$

$$= \sum_{k\in\mathcal{K}^h}\min\left(\zeta_k\wedge|S\cap R_k|, \sum_{k'\in\mathcal{P}^{h+1}\cap\mathcal{K}_k}\zeta_{k'}\wedge|S\cap R_{k'}|\right) \tag{31}$$

$$= \sum_{k\in\mathcal{K}^h}\min\left(\zeta_k\wedge|S\cap R_k|, |A^{h+1}\cap R_k|\right).$$

In the above, (29) holds by additivity and because for every $\mathcal{P}\in\mathfrak{P}^h$, any element of $\mathcal{P}$ is also an element of one of the $\mathcal{P}\cap\mathcal{K}_k$, $k\in\mathcal{K}^h$. Moreover, for every $\mathcal{P}\in\mathfrak{P}^h$ and $k\in\mathcal{K}^h$, $\mathcal{P}\cap\mathcal{K}_k$ is either $\{k\}$, either a set of elements of depth $\geq h+1$, hence (30). Finally, (31) holds because all the minima in (30) are realized in $\mathcal{P}^{h+1}$, otherwise the minimality of $\mathcal{P}^{h+1}$ in (28) would be contradicted.

We finally construct all the $A^h\cap R_k$, $k\in\mathcal{K}^h$, in the following way: if $|A^{h+1}\cap R_k|\leq\zeta_k\wedge|S\cap R_k|$, we let $A^h\cap R_k = A^{h+1}\cap R_k$, else we let $A^h\cap R_k$ be a subset of $\zeta_k\wedge|S\cap R_k|$ distinct elements of $A^{h+1}\cap R_k$. This both ensures that

$$|A^h| = \min_{\mathcal{P}\in\mathfrak{P}^h}\sum_{k\in\mathcal{P}}\zeta_k\wedge|S\cap R_k|,$$

and that

$$\forall k\in\mathcal{K}^{\geq h},\ \ |A^h\cap R_k|\leq\zeta_k,$$

because $\mathcal{K}^{\geq h} = \mathcal{K}^h\cup\mathcal{K}^{\geq h+1}$ and $A^h\subset A^{h+1}$, which ends the recursion.

Now letting $A = A^1$, we have found an $A\subset S$ such that $A\in\mathcal{A}(\mathfrak{R})$ and $|A| = \widetilde{V}_{\mathfrak{R}}(S)$ (by (27)).

7.2.2. Proof of (15)

By (14) and Lemmas A.3 and A.4, we have

$$V^*_{\mathfrak{R}}(S) = V^*_{\mathfrak{R}^\oplus}(S) = \widetilde{V}_{\mathfrak{R}^\oplus}(S) = \sum_{k\in\overline{\mathcal{K}}}\zeta_k\wedge|S\cap R_k|,$$

for some $\overline{\mathcal{K}}\subset\mathcal{K}^\oplus$ such that the R_k, $k\in\overline{\mathcal{K}}$, form a partition of $\mathbb{N}_m$. Hence,

$$\begin{aligned}
V^*_{\mathfrak{R}}(S) &= \sum_{k\in\mathcal{K}\cap\overline{\mathcal{K}}}\zeta_k\wedge|S\cap R_k| + \sum_{k\in\overline{\mathcal{K}}\setminus\mathcal{K}}\zeta_k\wedge|S\cap R_k| \\
&= \sum_{k\in\mathcal{K}\cap\overline{\mathcal{K}}}\zeta_k\wedge|S\cap R_k| + \sum_{k\in\overline{\mathcal{K}}\setminus\mathcal{K}}|S\cap R_k| \\
&= \sum_{k\in\mathcal{K}\cap\overline{\mathcal{K}}}\zeta_k\wedge|S\cap R_k| + \left|S\setminus\bigcup_{k\in\mathcal{K}\cap\overline{\mathcal{K}}}R_k\right|,
\end{aligned}$$

because the R_k, $k\in\overline{\mathcal{K}}\setminus\mathcal{K}$ are all disjoint. Now, $|\mathcal{K}\cap\overline{\mathcal{K}}|\leq d$ by definition of d, which means that the latter display is larger than or equal to $\widetilde{V}^d_{\mathfrak{R}}(S)$, which proves the result.

7.3. Proof of Corollary 3.7

Proof of *(i)* This is a direct byproduct of Theorem 3.6, because if (Nested) holds, then $d=1$ and thus $V^*_{\mathfrak{R}} = \widetilde{V}^d_{\mathfrak{R}} = \widetilde{V}^1_{\mathfrak{R}} = \overline{V}_{\mathfrak{R}}$.

Proof of *(ii)* By Theorem 3.6, $V^*_{\mathfrak{R}} = \widetilde{V}_{\mathfrak{R}} = \widetilde{V}^K_{\mathfrak{R}}$ defined by (8)–(9). Now, for any $S \subset \mathbb{N}_m$, for any $Q \subset \mathcal{K}$ with $|Q| \leq K-1$, by denoting k_0 any element not in Q, we have

$$R_{k_0} \cap \left(\bigcup_{k\in Q} R_k \right) = \varnothing,$$

by (Disjoint), and

$$\begin{aligned}
\sum_{k\in Q} \zeta_k \wedge |S\cap R_k| + \left| S \setminus \bigcup_{k\in Q} R_k \right| &= |S\cap R_{k_0}| + \sum_{k\in Q} \zeta_k \wedge |S\cap R_k| + \left| S \setminus \left(\bigcup_{k\in Q} R_k \cup R_{k_0} \right) \right| \\
&\geq \zeta_{k_0} \wedge |S\cap R_{k_0}| + \sum_{k\in Q} \zeta_k \wedge |S\cap R_k| + \left| S \setminus \left(\bigcup_{k\in Q} R_k \cup R_{k_0} \right) \right| \\
&= \sum_{k\in Q\cup\{k_0\}} \zeta_k \wedge |S\cap R_k| + \left| S \setminus \bigcup_{k\in Q\cup\{k_0\}} R_k \right|.
\end{aligned}$$

Hence, the minimum in (8) within the $\widetilde{V}^K_{\mathfrak{R}}$ expression is attained for $Q = \mathcal{K}$ and the result is proved.

7.4. Proof of Proposition 4.1

Let us show that for all $\lambda \in (0, 1/2)$, for any $S \subset \mathbb{N}_m$ with cardinal $s = |S|$, we have with probability at least $1-\lambda$ that

$$|S\cap\mathcal{H}_0| \leq \min_{t\in[0,1)} \left(\frac{\sqrt{\log(1/\lambda)/2}}{2(1-t)} + \left\{ \frac{\log(1/\lambda)/2}{4(1-t)^2} + \frac{N_t(S)}{1-t} \right\}^{1/2} \right)^2, \tag{32}$$

for $N_t(S) = \sum_{i\in S} \mathbf{1}\{p_i(X) > t\}$. Let $v = |S\cap\mathcal{H}_0|$ (assumed to be positive without loss of generality) and $U_1, \dots, U_v$ being v i.i.d. uniform random variables. The DKW inequality (with the optimal constant of Massart, 1990) ensures that, with probability at least $1-\lambda$, for all $t\in[0,1]$, we have

$$v^{-1} \sum_{i=1}^{v} \mathbf{1}\{U_i > t\} - (1-t) \geq -\sqrt{\log(1/\lambda)/(2v)}.$$

Now using Lemma A.1 with $x = v^{1/2}$, $a = 1-t$, $b = \sqrt{\log(1/\lambda)/2}$ and $c = \sum_{i=1}^{v} \mathbf{1}\{U_i > t\}$ provides (32) but with $N_t(S)$ replaced by c. Since $p_i(X)$ stochastically dominates U_i, by independence $N_t(S)$ also dominates c, which yields

$$\forall k\in\mathcal{K},\ \mathbb{P}\left(|R_k\cap\mathcal{H}_0| > \zeta_k(X)\right) \leq \frac{\alpha}{K},$$

by choosing $\lambda = \frac{\alpha}{K}$. Then (2) follows by a classical union bound argument.

7.5. Proof of Proposition 4.5

We have for any $t\in[0,1)$,

$$\begin{aligned}
\frac{\mathbb{E}(V_{\text{Bonf}}(R_1))}{|R_1|} &= s^{-1} \sum_{i\in R_1\cap\mathcal{H}_0} \mathbb{P}(p_i(X) > \alpha/m) + s^{-1} \sum_{i\in R_1\cap\mathcal{H}_1} \mathbb{P}(p_i(X) > \alpha/m) \\
&= (1-r)(1-\alpha/m) + r\left(1 - \overline{\Phi}(\overline{\Phi}^{-1}(\alpha/m) - \mu)\right),
\end{aligned}$$

which gives (22). Next,

$$\begin{aligned} V_{\text{Simes}}(R_1) &= \min_{1\leq k\leq s} \left\{ \sum_{i\in R_1} \mathbb{1}\left\{p_i(X) > \alpha k/m\right\} + k - 1 \right\} \\ &\geq \sum_{i\in R_1} \mathbb{1}\left\{p_i(X) > \alpha s/m\right\}, \end{aligned}$$

which gives (21). Finally, for all $t \in [0, 1)$, by denoting $N = \sum_{i\in R_1} \mathbf{1}\{p_i(X) > t\}$, we have

$$\begin{aligned} \mathbb{E}(V_{\text{DKW}}(R_1)) &\leq \mathbb{E}\left[\left(\frac{C}{2(1-t)} + \left\{ \frac{C^2}{4(1-t)^2} + \frac{N}{1-t} \right\}^{1/2} \right)^2 \right] \\ &\leq \mathbb{E}\left[\left(\frac{C}{1-t} + \left\{ \frac{N}{1-t} \right\}^{1/2} \right)^2 \right] \\ &\leq \frac{C^2}{(1-t)^2} + \frac{\mathbb{E}N}{1-t} + \frac{2C}{(1-t)^{3/2}}\, \mathbb{E}\left[N^{1/2}\right] \\ &\leq \frac{C^2}{(1-t)^2} + \frac{\mathbb{E}N}{1-t} + \frac{2C}{1-t} \left(\frac{\mathbb{E}N}{1-t} \right)^{1/2}, \end{aligned}$$

where we used $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ for all $x, y \geq 0$ and that $x \mapsto x^{1/2}$ is concave. Since

$$\mathbb{E}\left[N/|R_1|\right] = (1-r)(1-t) + r\left(1 - \overline{\Phi}(\overline{\Phi}^{-1}(t) - \mu)\right),$$

and $\mathbb{E}[N] \leq s(1-t)$, this provides

$$\frac{\mathbb{E}(V_{\text{DKW}}(R_1))}{|R_1|} \leq \min_t \left\{ s^{-1}\frac{C^2}{(1-t)^2} + 1 - r + r\frac{\overline{\Phi}(\mu - \overline{\Phi}^{-1}(t))}{1-t} + s^{-1/2}\frac{2C}{1-t} \right\}.$$

Taking $t = 1/2$ gives (20).

Acknowledgements

This work has been supported by ANR-16-CE40-0019 (SansSouci) and ANR-17-CE40-0001 (BASICS).

References

Bachoc, F., Blanchard, G., and Neuvial, P. (2018). On the Post Selection Inference constant under Restricted Isometry Properties. *arXiv preprint arXiv:1804.07566*.

Bachoc, F., Leeb, H., and Pötscher, B. M. (2018+). Valid confidence intervals for post-model-selection predictors. *Annals of Statistics (to appear)*.

Benjamini, Y. and Bogomolov, M. (2014). Selective inference on multiple families of hypotheses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):297–318.

Benjamini, Y. and Yekutieli, D. (2005). False discovery rate–adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, 100(469):71–81.

Berk, R., Brown, L., Buja, A., Zhang, K., Zhao, L., et al. (2013). Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837.

Blanchard, G., Neuvial, P., and Roquain, E. (2017). Post hoc inference via joint family-wise error rate control. *arXiv preprint arXiv:1703.02307*.

Blanchard, G. and Roquain, É. (2009). Adaptive false discovery rate control under independence and dependence. *Journal of Machine Learning Research*, 10(Dec):2837–2871.

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.

Choi, Y., Taylor, J., Tibshirani, R., et al. (2017). Selecting the number of principal components: estimation of the true rank of a noisy matrix. *The Annals of Statistics*, 45(6):2590–2617.

Cox, D. R. (1975). A note on data-splitting for the evaluation of significance levels. *Biometrika*, 62(2):441–444.

Dezeure, R., Bühlmann, P., Meier, L., and Meinshausen, N. (2015). High-dimensional inference: confidence intervals, p-values and `r`-software `hdi`. *Statist. Sci.*, 30(4):533–558.

Dvoretzky, A., Kiefer, J., and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pages 642–669.

Farcomeni, A. and Pacillo, S. (2011). A conservative estimator for the proportion of false nulls based on Dvoretzky, Kiefer and Wolfowitz inequality. *Statistics & Probability Letters*, 81(12):1867–1870.

Fithian, W., Sun, D., and Taylor, J. (2017). Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*.

Genovese, C. and Wasserman, L. (2004). A stochastic process approach to false discovery control. *Annals of Statistics*, pages 1035–1061.

Genovese, C. R. and Wasserman, L. (2006). Exceedance control of the false discovery proportion. *Journal of the American Statistical Association*, 101(476):1408–1417.

Goeman, J. J. and Solari, A. (2010). The sequential rejection principle of familywise error control. *The Annals of Statistics*, pages 3782–3810.

Goeman, J. J. and Solari, A. (2011). Multiple testing for exploratory research. *Statistical Science*, pages 584–597.

Goeman, J. J., Van De Geer, S. A., De Kort, F., and Van Houwelingen, H. C. (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1):93–99.

Hemerik, J. and Goeman, J. J. (2018). False discovery proportion estimation by permutations: confidence for significance analysis of microarrays. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):137–155.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.*, 6(2):65–70.

Kolaczyk, E. D. (2009). *Statistical analysis of network data: methods and models*. Springer Science & Business Media.

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

Massart, P. (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability*, pages 1269–1283.

Meijer, R. J. and Goeman, J. J. (2015). A multiple testing method for hypotheses structured in a directed acyclic graph. *Biometrical Journal*, 57(1):123–143.

Meijer, R. J., Krebs, T. J., and Goeman, J. J. (2015). A region-based multiple testing method for hypotheses ordered in space or time. *Statistical Applications in Genetics and Molecular Biology*, 14(1):1–19.

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

Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance. *Biometrika*, 40(1-2):87–110.

Scheffé, H. (1959). *The analysis of variance*. John Wiley & Sons, Inc., New York; Chapman & Hall, Ltd., London.

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

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498.

Taylor, J. and Tibshirani, R. (2018). Post-selection inference for-penalized likelihood models. *Canadian Journal of Statistics*, 46(1):41–61.

Taylor, J. and Tibshirani, R. J. (2015). Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25):7629–7634.

Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514):600–620.

Appendix A: Auxiliary lemmas

The following lemma holds.

Lemma A.1. *For all $a > 0$ and $b, c, x \geq 0$, the two following assertions are equivalent*

(i) $c - ax^2 \geq -bx$;

(ii) $x \leq \frac{b}{2a} + \sqrt{\frac{b^2}{4a^2} + \frac{c}{a}}$.

In particular, we have for all $d \geq 0$,

$$d \wedge \left(\frac{b}{2a} + \sqrt{\frac{b^2}{4a^2} + \frac{c}{a}}\right)^2 \leq d \wedge \left(\frac{c + d^{1/2}b}{a}\right). \tag{33}$$

Proof. The equivalence between (i) and (ii) is obvious. For $d \geq 0$, if we have the inequality $\left(b/(2a) + \sqrt{b^2/(4a^2) + c/a}\right)^2 \geq d$, then (ii) is satisfied with $x = d^{1/2}$, which entails $c - ad \geq -bd^{1/2}$ and gives $d \leq (c + d^{1/2}b)/a$. If, on the contrary, $\left(b/(2a) + \sqrt{b^2/(4a^2) + c/a}\right)^2 \leq d$, then

$$\begin{aligned}\left(\frac{b}{2a} + \sqrt{\frac{b^2}{4a^2} + \frac{c}{a}}\right)^2 &= \frac{b^2}{2a^2} + \frac{c}{a} + \frac{b}{a}\sqrt{b^2/(4a^2) + c/a} \\ &= \frac{c}{a} + \frac{b}{a}\left(b/(2a) + \sqrt{b^2/(4a^2) + c/a}\right) \leq \frac{c}{a} + \frac{b}{a}d^{1/2}.\end{aligned}$$

This entails the result. □

The two following lemmas are used in the proof of Theorem 3.6, in the case where condition (All-atoms) holds.

Lemma A.2. *For a reference family that has a Forest structure, if (All-atoms) holds, then for any $h \geq 1$, the $P_{i:j}$, $(i, j) \in \mathcal{K}^h$, form a partition of $\mathbb{N}_m$.*

Proof. Let $h \geq 1$. Let $(i, j), (i', j') \in \mathcal{K}^h$ such that $(i, j) \neq (i', j')$. By (Forest), either $P_{i:j}$ and $P_{i':j'}$ are disjoint, or, without loss of generality, $P_{i:j} \subset P_{i':j'}$. If $\phi(i', j') = h$ then the latter is not possible because that would mean that $\phi(i, j) \geq h + 1$. If $i' = j'$, then $P_{i:j} \subset P_{i':j'}$ would imply that $P_i \cup \cdots \cup P_j \subset P_{i'}$ which in turn implies $i = j = i' = j'$ which is also impossible. So $P_{i:j}$ and $P_{i':j'}$ are disjoint.

Now take any $e \in \mathbb{N}_m$. $(P_n)_{1 \leq n \leq N}$ is a partition so there exists some $n \leq N$ such that $e \in P_n$. If $\phi(n, n) \leq h$ then $(n, n) \in \mathcal{K}^h$. If $\phi(n, n) > h$, then $\{k \in \mathcal{K} : P_n \subsetneq R_k\}$ has at least h elements. Furthermore those elements are nested by (Forest), so there exists $k \in \mathcal{K}$ such that $P_n \subsetneq R_k$ and $\phi(k) = h$, hence $e \in R_k$ with $k \in \mathcal{K}^h$. Finally in both cases $e \in \bigcup_{k \in \mathcal{K}^h} R_k$ so $\mathbb{N}_m = \bigcup_{k \in \mathcal{K}^h} R_k$, which concludes the proof. □

Lemma A.3. *For a reference family that satisfies (Forest) and (All-atoms), we have*

$$\widetilde{V}_{\mathfrak{R}}(S) = \min_{\substack{\overline{\mathcal{K}} \subset \mathcal{K} \\ \text{the } R_k, k \in \overline{\mathcal{K}}, \\ \text{form a partition of } \mathbb{N}_m}} \sum_{k \in \overline{\mathcal{K}}} \zeta_k \wedge |S \cap R_k|. \tag{34}$$

that is, the minimum in (26) is always achieved on a partition of $\mathbb{N}_m$.

Proof. Let any $\mathcal{K}' \subset \mathcal{K}$. Because property (Forest) is true, there exists $\mathcal{K}'_1 \subset \mathcal{K}'$ such that the R_k, $k \in \mathcal{K}'_1$, are pairwise disjoint, and

$$\forall k \in \mathcal{K}', \exists k' \in \mathcal{K}'_1, R_k \subset R_{k'}.$$

Note that this implies that $\bigcup_{k\in\mathcal{K}'_1} R_k = \bigcup_{k\in\mathcal{K}'} R_k$. Likewise, because $\mathcal{K}$ includes all the $(i,i), 1 \leq i \leq N$, there exists $\mathcal{K}'_2 \subset \mathcal{K}$ such that the R_k, $k \in \mathcal{K}'_2$, are pairwise disjoint, and

$$\mathbb{N}_m \setminus \bigcup_{k\in\mathcal{K}'_1} R_k = \bigcup_{k\in\mathcal{K}'_2} R_k.$$

Let $\overline{\mathcal{K}} = \mathcal{K}'_1 \cup \mathcal{K}'_2$ and note that the R_k, $k \in \overline{\mathcal{K}}$, form a partition of $\mathbb{N}_m$. To conclude the proof of (34), we write that

$$\begin{aligned}
&\sum_{k\in\mathcal{K}'} \zeta_k \wedge |S \cap R_k| + \left| S \setminus \bigcup_{k\in\mathcal{K}'} R_k \right| = \\
&\sum_{k\in\mathcal{K}'} \zeta_k \wedge |S \cap R_k| + \left| S \cap \left(\mathbb{N}_m \setminus \bigcup_{k\in\mathcal{K}'_1} R_k \right) \right| \geq \\
&\sum_{k\in\mathcal{K}'_1} \zeta_k \wedge |S \cap R_k| + \sum_{k\in\mathcal{K}'_2} |S \cap R_k| \geq \\
&\sum_{k\in\mathcal{K}'_1} \zeta_k \wedge |S \cap R_k| + \sum_{k\in\mathcal{K}'_2} \zeta_k \wedge |S \cap R_k| = \sum_{k\in\overline{\mathcal{K}}} \zeta_k \wedge |S \cap R_k|.
\end{aligned}$$

□

The last lemma is useful for the general case where (All-atoms) no longer holds, by making use of the completed Forest structure introduced in Definition 3.5.

Lemma A.4. *For a reference family $\mathfrak{R} = (R_k, \zeta_k)_{k\in\mathcal{K}}$ that has a Forest structure, and $\mathcal{K}^+$, $\mathcal{K}^\oplus$, $\mathfrak{R}^\oplus$ as in Definition 3.5, we have for all $S \subset \mathbb{N}_m$:*

$$V^*_{\mathfrak{R}^\oplus}(S) = V^*_{\mathfrak{R}}(S),$$

$$\widetilde{V}_{\mathfrak{R}^\oplus}(S) = \widetilde{V}_{\mathfrak{R}}(S).$$

Proof. It is trivial that $\mathcal{A}(\mathfrak{R}) = \mathcal{A}(\mathfrak{R}^\oplus)$ (see (25)) because $\zeta_k = |R_k|$ for $k \in \mathcal{K}^+$, hence $V^*_{\mathfrak{R}^\oplus}(S) = V^*_{\mathfrak{R}}(S)$. It is also obvious that $\widetilde{V}_{\mathfrak{R}}(S) \geq \widetilde{V}_{\mathfrak{R}^\oplus}(S)$ by (26) and since $\mathcal{K} \subset \mathcal{K}^\oplus$. Now let any $\mathcal{K}' \subset \mathcal{K}^\oplus$. Let $\mathcal{K}'_1 = \mathcal{K}' \cap \mathcal{K}$ and $\mathcal{K}'_2 = \mathcal{K}' \cap \mathcal{K}^+$. Note that $\mathcal{K}'$ is the disjoint union of $\mathcal{K}'_1$ and $\mathcal{K}'_2$. Then,

$$\begin{aligned}
\sum_{k\in\mathcal{K}'} \zeta_k \wedge |S \cap R_k| + \left| S \setminus \bigcup_{k\in\mathcal{K}'} R_k \right| &= \sum_{k\in\mathcal{K}'_1} \zeta_k \wedge |S \cap R_k| + \sum_{k\in\mathcal{K}'_2} |S \cap R_k| + \left| S \setminus \bigcup_{k\in\mathcal{K}'} R_k \right| \\
&\geq \sum_{k\in\mathcal{K}'_1} \zeta_k \wedge |S \cap R_k| + \left| S \setminus \bigcup_{k\in\mathcal{K}'_1} R_k \right| \\
&\geq \widetilde{V}_{\mathfrak{R}}(S),
\end{aligned}$$

because $\zeta_k = |R_k|$ for $k \in \mathcal{K}'_2$. Hence $\widetilde{V}_{\mathfrak{R}^\oplus}(S) \geq \widetilde{V}_{\mathfrak{R}}(S)$, which concludes the proof. □

Appendix B: Material for Lemma 3.3

Algorithm 2 below builds (P_n) and follows directly from the proof. It may be useful for the reader to start by looking the algorithm, in order to get a sense of what the formal proof does.

Proof of Lemma 3.3. Let $H = \max_{k\in\mathcal{K}} \phi(k)$, where ϕ is the depth function defined by (11). We use a recursion to build, for each $1 \le h \le H$, an integer $N^h \ge 1$ and a partition $P^h = (P^h_n)_{1\le n\le N^h}$ which satisfy the following three properties:

$$P^h \text{ is a partition of } \mathbb{N}_m, \tag{$\mathscr{P}^h_1$}$$

$$\forall k \in \mathcal{K} \text{ such that } \phi(k) < h, \exists (i,j) \in \left\{1,\dots,N^h\right\}^2 : R_k = \bigcup_{i\le n\le j} P^h_n, \tag{$\mathscr{P}^h_2$}$$

$$\forall k \in \mathcal{K} \text{ such that } \phi(k) = h, \exists n \in \left\{1,\dots,N^h\right\} : R_k = P^h_n. \tag{$\mathscr{P}^h_3$}$$

We start the recursion with $h = 1$. Let $Succ_1 = \{k \in \mathcal{K} : \phi(k) = 1\}$,

$$New_1 = \{R_k : k \in Succ_1\} \cup \left\{ \mathbb{N}_m \setminus \bigcup_{k\in Succ_1} R_k \right\} \setminus \{\varnothing\},$$

and $N^1 = |New_1|$. We let P^1 be the family of elements of New_1. ($\mathscr{P}^1_1$) is true because, by (Forest), for $k, k' \in Succ_1$, $k \ne k'$, R_k and $R_{k'}$ are disjoint (otherwise they can't have same depth). ($\mathscr{P}^1_2$) and ($\mathscr{P}^1_3$) are trivially true.

Now let $h \in \{2,\dots,H\}$ and assume that there exists N^{h-1} and P^{h-1} satisfying ($\mathscr{P}^{h-1}_1$), ($\mathscr{P}^{h-1}_2$) and ($\mathscr{P}^{h-1}_3$). For all $n \in \{1,\dots,N^{h-1}\}$, let

$$Succ_{h,n} = \left\{k \in \mathcal{K} : \phi(k) = h \text{ and } R_k \subset P^{h-1}_n\right\},$$

$$New_{h,n} = \{R_k : k \in Succ_{h,n}\} \cup \left\{ P^{h-1}_n \setminus \bigcup_{k\in Succ_{h,n}} R_k \right\} \setminus \{\varnothing\},$$

$\mathbb{S}^h_n = \sum_{n'=0}^{n} |New_{h,n'}|$ (with $|New_{h,0}| = 0$ by convention), and $\left(P^h_{\mathbb{S}^h_{n-1}+1}, \dots, P^h_{\mathbb{S}^h_n}\right)$ be the family of the elements of $New_{h,n}$. Then let $N^h = \mathbb{S}^h_{N^{h-1}}$ and $P^h = (P^h_1, \dots, P^h_{N^h})$. Note that for each $1 \le n \le N^{h-1}$, P^{h-1}_n is the disjoint union of $P^h_{\mathbb{S}^h_{n-1}+1}, \dots, P^h_{\mathbb{S}^h_n}$, because by (Forest), for $k, k' \in Succ_{h,n}$, $k \ne k'$, R_k and $R_{k'}$ are disjoint (otherwise they can't have same depth). This and ($\mathscr{P}^{h-1}_1$) imply ($\mathscr{P}^h_1$). Let $k \in \mathcal{K}$ such that $\phi(k) < h$, then ($\mathscr{P}^{h-1}_2$) and ($\mathscr{P}^{h-1}_3$) imply that there exists $(i,j) \in \{1,\dots,N^{h-1}\}^2$ such that $R_k = \bigcup_{i\le n\le j} P^{h-1}_n$. Hence

$$R_k = \bigcup_{\mathbb{S}^{h-1}_{i-1}+1 \le n \le \mathbb{S}^{h-1}_j} P^h_n,$$

and we get ($\mathscr{P}^h_2$). Finally let $k \in \mathcal{K}$ such that $\phi(k) = h$. Let k' be the unique element of $\mathcal{K}$ such that $\phi(k') = h - 1$ and $R_k \subsetneq R_{k'}$. By ($\mathscr{P}^{h-1}_3$), there exists $n \in \{1,\dots,N^{h-1}\}$ such that $R_{k'} = P^{h-1}_n$. Hence $k \in Succ_{h,n}$ and R_k is equal to one of the elements of $New_{h,n}$, which gives us ($\mathscr{P}^h_3$).

Now that the recursion has ended, properties ($\mathscr{P}^H_1$), ($\mathscr{P}^H_2$) and ($\mathscr{P}^H_3$) imply the existence of the desired partition. The proof of the converse statement is straightforward from (12). □

For the purpose of Algorithm 2, we let `len` and `con` be the concatenation and length functions such that, for all $S_1, \dots, S_n, S_{n+1} \subset \mathbb{N}_m$ and $S = (S_1, \dots, S_n)$, $\texttt{len}(S) = n$, $\texttt{con}(S, S_{n+1}) = (S_1, \dots, S_n, S_{n+1})$ if $S_{n+1} \ne \varnothing$ and $\texttt{con}(S, \varnothing) = S$.

Algorithm 2: Computation of $(P_n)_{1\leq n\leq N}$

Data: $\mathfrak{R} = (R_k, \zeta_k)_{k\in\mathcal{K}}$ satisfying (Forest).
Result: $P = (P_n)_{1\leq n\leq N}$ such that for each $k \in \mathcal{K}$, there exists some (i, j) such that $R_k = \bigcup_{i\leq n\leq j} P_n$.

$P \longleftarrow (\mathbb{N}_m)$;
$N \longleftarrow 1$;
$H \longleftarrow \max_{k\in\mathcal{K}} \phi(k)$;
for $h \in (1, \dots, H)$ **do**
  $newP \longleftarrow ()$;
  for $n \in \{1, \dots, N\}$ **do**
    $Succ_{h,n} \longleftarrow \{k \in \mathcal{K} : R_k \subset P_n, \phi(k) = h\}$;
    for $k \in Succ_{h,n}$ **do**
      $newP \longleftarrow \texttt{con}(newP, R_k)$;
    end
    $newP \longleftarrow \texttt{con}\left(P_n \setminus \bigcup_{k\in Succ_{h,n}} R_k, newP\right)$;
  end
  $P \longleftarrow newP$;
  $N \longleftarrow \texttt{len}(P)$;
end
return P