Permutation-based true discovery guarantee by sum tests

Permutation-Based True Discovery Guarantee by Sum Tests

Anna Vesely
University of Bremen, Germany
E-mail: vesely@uni-bremen.de
Livio Finos
University of Padua, Italy
Jelle J. Goeman
Leiden University Medical Center, The Netherlands

Summary. Sum-based global tests are highly popular in multiple hypothesis testing. In this paper we propose a general closed testing procedure for sum tests, which provides lower confidence bounds for the proportion of true discoveries (TDP), simultaneously over all subsets of hypotheses. These simultaneous inferences come for free, i.e., without any adjustment of the α-level, whenever a global test is used. Our method allows for an exploratory approach, as simultaneity ensures control of the TDP even when the subset of interest is selected post hoc. It adapts to the unknown joint distribution of the data through permutation testing. Any sum test may be employed, depending on the desired power properties. We present an iterative shortcut for the closed testing procedure, based on the branch and bound algorithm, which converges to the full closed testing results, often after few iterations; even if it is stopped early, it controls the TDP. We compare the properties of different choices for the sum test through simulations, then we illustrate the feasibility of the method for high dimensional data on brain imaging and genomics data.

Keywords: closed testing, multiple testing, permutation test, selective inference, sum test, true discovery proportion

1. Introduction

In high-dimensional data analysis, researchers are often interested in detecting subsets of features that are associated with a given outcome. For instance, in functional magnetic resonance imaging (fMRI) data the objective may be to identify a brain region that is activated by a stimulus; in genomics data one may want to find a biological pathway that is differentially expressed. In this context, global tests allow to aggregate signal from multiple features and make meaningful statements at the set level. A diverse range of global tests has been proposed in literature: well-known examples are p-value combinations, described and compared in Pesarin (2001), Loughin (2004), Won et al. (2009) and Pesarin and Salmaso (2010); other popular methods are Simes test (Simes, 1986), the global test of Goeman et al. (2006), the sequence kernel association test (SKAT) (Wu et al., 2011) and higher criticism (Donoho and Jin, 2015). A substantial proportion, including many of the above-mentioned methods, is sum-based, meaning that the global test statistic may be written as a sum of contributions per feature. In this paper we restrict to such sum-based tests.

The probability distribution of a global statistic depends not only on the marginal distributions of the data, but also on the joint distribution; for this reason, many sum tests only have a known null distribution under independence. Approaches that deal with the a-priori unknown joint distribution are worst-case distributions, defined either generally or under restrictive assumptions (Vovk and Wang, 2020), and nonparametric permutation testing (Fisher, 1936; Ernst, 2004). As worst-case distributions tend to be very conservative, the latter approach is preferable; it relies on minimal assumptions (Hemerik and Goeman, 2018a), and generally offers an improvement in power over the parametric approach, especially when multiple hypotheses are considered (Westfall and Young, 1993; Pesarin, 2001; Hemerik and Goeman, 2018b; Hemerik et al., 2019).

Rejecting a null hypothesis, however, gives little information on the corresponding set. A significant p-value only indicates that there is at least one true discovery, i.e., one feature associated with the outcome, but does not give any information on the proportion of true discoveries (TDP), nor their localization. This becomes problematic especially for large sets (Woo et al., 2014). Moreover, since interest is usually not just in the set of all features, but in several subsets, a multiple testing procedure is necessary (Nichols, 2012; Meijer and Goeman, 2016). Finally, when researchers do not know a priori which subsets they are interested in, they may want to test many and then make the selection post hoc. The case for the use of TDPs in large-scale testing problems was argued by Rosenblatt et al. (2018) in neuroimaging and by Ebrahimpoor et al. (2020) in genomics.

This paper presents a general approach for inference on the TDP. The method allows any sum-based test, requiring only that critical values are determined by permutations. It provides TDPs not only for the full testing problem, but also simultaneously for all subsets, allowing subsets of interest to be chosen post hoc.

We will rely on the closed testing framework (Marcus et al., 1976), which allows to construct confidence sets for the TDP simultaneously over all possible subsets (Genovese and Wasserman, 2006; Goeman and Solari, 2011; Goeman et al., 2019). These additional simultaneous inferences on all subsets come for free, i.e., without any adjustment of the α-level, whenever a global test is applied. Simultaneity ensures that the procedure is not compromised by post-hoc selection, therefore researchers can postpone the choice of the subset until after seeing the data, while still obtaining valid confidence sets; used in this way, closed testing allows a form of post-hoc inference. Furthermore, closed testing has been proven to be the optimal way to construct multiple testing procedures, as all family-wise error rate (FWER), TDP and related methods are either equivalent to or can be improved by it (Goeman et al., 2021). The main challenge is the computational complexity, which is extremely high when considering many hypotheses, and when using many permutations. Permutation-based closed testing for the TDP so far mostly focused on Simes-based test procedures, while sum tests were approached under independence or with worst-case distributions (Vovk and Wang, 2020; Wilson, 2019; Tian et al., 2022), that are simpler as critical values depend only on the size of the subset.

We propose a general closed testing procedure for sum-based permutation tests, which provides simultaneous confidence sets for the TDP of all subsets of the testing problem. We develop two shortcuts to make this procedure feasible for large-scale problems. First, we develop a quick shortcut that approximates closed testing and has worst-case complexity of order $m \log^2 m$ in the number m of individual hypotheses, and linearithmic in

the number of permutations. Next, we embed this shortcut within a branch and bound algorithm, obtaining an iterative procedure that converges to full closed testing, often after few iterations; even if it is stopped early, it still controls the TDP. This procedure is exact and extremely flexible, as it applies to any sum test and adapts to the correlation structure of the data. It can be scaled up to high-dimensional problems, such as fMRI data, whose typical dimension is of order 10^5. Finally, we show that particular choices of the sum test statistic, namely statistics based on truncation, result in faster procedures.

The structure of the paper is as follows. First, we briefly discuss related works in Section 2. Then we introduce sum tests in Section 3, and we review the properties of permutation testing and closed testing in Sections 4 and 5. We derive the single-step shortcut in Section 6, and characterize when it is equivalent to closed testing in Section 7. In Section 8 we define the iterative shortcut, and finally in Section 9 we introduce refinements that improve the computational complexity. In the remaining section we compare the properties of different sum tests through simulations, and explore an application to fMRI data. Proofs and some additional results are postponed to the supplementary material; the corresponding sections are referred to with an additional S- in the numbering.

2. Related work

In this section we discuss related work, highlighting the contribution of the proposed method and its relevance in applications. As argued in Section 1, in this paper we focus on permutation-based tests. Here we justify the choice of closed testing procedures that give lower $(1-\alpha)$-confidence bounds for the TDP simultaneously over all subsets of hypotheses, which we will refer to as procedures with true discovery guarantee as in Goeman et al. (2021). Then we argue that it is worthwhile to construct such procedures for global tests that are frequently used, many of which are sum-based.

Genovese and Wasserman (2006) and Goeman and Solari (2011) showed that all global tests automatically come with an inbuilt selective inference method; they can be embedded in the closed testing framework to obtain procedures with true discovery guarantee without any adjustment of the α-level. Furthermore, a great number of multiple testing methods, including all those controlling FWER, generalized FWER (k-FWER), false discovery proportion (FDP), false discovery exceedance (FDX) and joint error rate (JER), can be written as procedures with true discovery guarantee. Among these, however, only closed testing procedures are admissible, i.e., cannot be uniformly improved (Goeman et al., 2021). This motivates the study of closed testing procedures for popular global tests.

So far, most procedures that explicitly give true discovery guarantee (Meinshausen, 2006; Rosenblatt et al., 2018; Hemerik et al., 2019; Ebrahimpoor et al., 2020; Blanchard et al., 2020; Andreella et al., 2020; Blain et al., 2022) were constructed using critical vectors for ordered p-values, e.g., based on variants of Simes (1986) or higher criticism (Donoho and Jin, 2015). With the exception of higher criticism, the global tests implicit in these procedures have seldom been considered as global tests in application contexts, and their popularity in multiple testing procedures is partly motivated by mathematical convenience. In contrast, tests based on sums are natural and popular as global tests.

This broad class includes many popular p-value combination tests, such as the classical Fisher combination (Fisher, 1925), as well as recent proposals such as Wilson (2019), Liu and Xie (2020), the global test of Goeman et al. (2006), SKAT (Wu et al., 2011), and e-value combinations (Vovk and Wang, 2021). Though closed testing procedures for sum-based tests were proposed in general in the parametric approach (Tian et al., 2022) and for some particular cases (Goeman and Solari, 2011; Blanchard et al., 2020), general scalable procedures in the permutation framework were lacking. In this paper we fill this gap, providing a procedure that can be applied to any sum-based test, as long as permutations are used to calculate the critical values.

Among permutation-based procedures, we mention especially the methods of Blanchard et al. (2020) and Andreella et al. (2020), using tests based on critical vectors of ordered p-values. First, we remark that our proposed method is not a competitor but complementary, as it deals with a different choice of the underlying test with different power properties. Subsequently, we observe that these methods do not perform full closed testing, and thus may be conservative. Blanchard et al. (2020) and the single-step version in Andreella et al. (2020) have computation times primarily related to computing and sorting permutation test statistics; we will show that the computation time of our single-step shortcut is comparable. The iterative method of Andreella et al. (2020) uniformly improves the corresponding single-step version and Blanchard et al. (2020), but requires a high computational time and is still not guaranteed to converge to closed testing. On the contrary, the proposed iterative shortcut converges to closed testing and so cannot be uniformly improved.

3. Sum tests

We start with a general definition of a sum test statistic. Throughout the paper, we will refer to null hypotheses simply as hypotheses, and we will denote both variables and sets with capital letters, leaving the distinction to context. Let $\mathbf{X} = (X_1, \ldots, X_m)$ be a collection of observable variables from m testing units, having indices in $M = \{1, \ldots, m\}$ and taking values in a sample space $\mathcal{X}$. We are interested in studying m corresponding univariate hypotheses $H_1, \ldots, H_m$ with confidence $1 - \alpha$, where $\alpha \in [0, 1)$. Let $N \subseteq M$ be the unknown subset of true hypotheses. A generic subset $S \subseteq M$, with size $|S| = s$, defines an intersection hypothesis $H_S = \bigcap_{i \in S} H_i$, which is true if and only if $S \subseteq N$. In the particular case of $S = \emptyset$, we take $H_\emptyset$ as usual to be a hypothesis that is always true.

For each univariate hypothesis H_i, let $T_i : \mathcal{X} \to \mathbb{R}$ be a test statistic. The general form of a sum test statistic for H_S is

$$T_S = g\left(\sum_{i \in S} f_i(T_i)\right),$$

where $f_i : \mathbb{R} \to \mathbb{R}$ are generic functions, and $g : \mathbb{R} \to \mathbb{R}$ is strictly monotone. Usually the functions f_i are also taken as monotone, so that high values of T_S give evidence against H_S. Moreover, as f_i may depend on i, the contributions $f_i(T_i)$ may have different distributions, as in the case of weighted sums. Examples include p-value combinations such as Fisher (1925), Pearson (1933), Liptak/Stouffer (Liptak, 1958), Lancaster (1961), Edgington (1972), and Cauchy (Liu and Xie, 2020). We mention especially the generalized

mean family (Vovk and Wang, 2020) with $f_i(y) = y^r$ and $g(z) = z^{1/r}$, where $r \in \mathbb{R}$, for which Wilson (2019) studied the harmonic mean ($r = -1$).

Since we can always re-write $\tilde{T}_i = f_i(T_i)$ and $\tilde{T}_S = g^{-1}(T_S)$, without loss of generality we can assume that f_i and g are the identity, so that $T_S = \sum_{i \in S} T_i$. In particular, for the empty set we obtain $T_\emptyset = 0$. Furthermore, we assume that the signs of the statistics T_i are chosen in such a way that high values of T_i, and therefore T_S, correspond to evidence against H_i and H_S, respectively.

4. Permutation testing

To test H_S with significance level α we will use permutations. Let Π be a collection of transformations $\pi : \mathcal{X} \to \mathcal{X}$ of the sample space; these may be permutations, but also other transformations such as rotations (Langsrud, 2005; Solari et al., 2014) and sign flipping (Hemerik et al., 2020). We assume that Π is an algebraic group with respect to the operation of composition of functions. The group structure is important as, without it, the resulting test may be highly conservative or anti-conservative (Hoeffding, 1952; Southworth et al., 2009).

Denote with $T_i = T_i(\mathbf{X})$ and $T_i^\pi = T_i(\pi\mathbf{X})$, with $\pi \in \Pi$, the statistics for the original and transformed variables, respectively, and with t_i and t_i^π the values computed on the observed and transformed data. The main assumption of permutation testing is the following.

Assumption 1. *The joint distribution of the statistics T_i^π, with $i \in N$ and $\pi \in \Pi$, is invariant under all transformations in Π of $\mathbf{X}$: $(T_i)_{i \in N} \overset{d}{=} (T_i^\pi)_{i \in N}$ for each $\pi \in \Pi$, where $\overset{d}{=}$ denotes equality in distribution.*

This assumption is common to most permutation-based multiple-testing methods, such as maxT-method (Westfall and Young, 1993; Meinshausen, 2006; Goeman and Solari, 2010; Hemerik et al., 2019). For some choices of the group Π, the assumption holds only asymptotically (Winkler et al., 2014; Solari et al., 2014; Hemerik et al., 2020). Detailed illustration and examples can be found in Pesarin (2001), Huang et al. (2006) and Hemerik and Goeman (2018a). Even if the invariance assumption is common and reasonable in many contexts, in applications an argument must be given for it; in some cases, it is violated even asymptotically (e.g., for Behrens-Fisher problem (Schildknecht et al., 2015)).

A slightly stronger assumption, that is easier to check, is that the statistic $T_S = T_S(\mathbf{X}_S)$ is a function of $\mathbf{X}_S = (X_i : i \in S)$ only, and $\mathbf{X}_N \overset{\mathrm{d}}{=} \pi\mathbf{X}_N$ for each π. Note that the assumption holds also when the distributions of the individual statistics T_i are different, as in the case of weighted sums. Moreover, it holds in the particular case when H_S true implies that $\mathbf{X}_S \overset{\mathrm{d}}{=} \pi\mathbf{X}_S$ for each π.

If the cardinality of Π is large, a valid α-level test may use B randomly chosen elements (Hemerik and Goeman, 2018b). The value of B does not need to grow with m or s; to have non-zero power we must only have $B \geq 1/\alpha$, though larger values of B give more power. For $\alpha = 0.05$, $B \geq 200$ is generally sufficient (see Section 10.2). Consider a vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_B)$, where $\pi_1 = \mathrm{id}$ is the identity in Π, and $\pi_2, \ldots, \pi_B$ are random

Table 1. Toy example: original and centered test statistics.

	original t_i^π					centered c_i^π				
	H_1	H_2	H_3	H_4	H_5	H_1	H_2	H_3	H_4	H_5
id	6	5	4	1	1	0	0	0	0	0
π_2	1	2	1	0	4	5	3	3	1	-3
π_3	8	3	0	2	1	-2	2	4	-1	0
π_4	8	1	0	1	0	-2	4	4	0	1
π_5	0	6	1	1	2	6	-1	3	0	-1
π_6	7	0	1	2	1	-1	5	3	-1	0

elements drawn with replacement from a uniform distribution on Π. A test for H_S may be defined taking as critical value the $\lceil (1-\alpha)B \rceil$-th quantile, where $\lceil \cdot \rceil$ represents the ceiling function, and $t_S^{(1)} \leq \ldots \leq t_S^{(B)}$ are the sorted values t_S^π, with $\pi \in \boldsymbol{\pi}$.

LEMMA 1. *Under Assumption 1, the test that rejects H_S when $t_S > t_S^{(\lceil (1-\alpha)B \rceil)}$ is an α-level test.*

The test is defined conditionally on $\mathbf{X}$, but it becomes unconditional if we take the expected value on both sides of the inequality. Note that both the test statistic and the critical value are random variables. For our method it will be convenient to use an equivalent characterization of the test with a non-random critical value. Therefore, for each π we define the centered statistic $C_S^\pi = T_S - T_S^\pi$, so that the observed value $c_S = c_S^{\text{id}}$ is always zero, and so no longer random. We give a permutation test based on these new statistics, using $\omega = \lfloor \alpha B \rfloor + 1$ to obtain the quantile, where $\lfloor \cdot \rfloor$ is the floor function.

THEOREM 1. *Under Assumption 1, the test that rejects H_S when $c_S^{(\lfloor \alpha B \rfloor + 1)} > 0$ is an α-level test.*

For illustration, we introduce a recurring toy example with $m = 5$ univariate hypotheses and $B = 6$ transformations (Table 1). Given the subset $S = \{1, 2\}$, we are interested in testing H_S with significance level $\alpha = 0.4$. The statistics t_S^π and c_S^π are obtained summing columns 1 and 2 by row. Since $\omega = 3$ and $c_S^{(\omega)} = 2$, the test of Theorem 1 rejects H_S.

5. True discovery guarantee

Based on the notation introduced above, consider the number of true discoveries $\delta(S) = |S \setminus N|$ made when rejecting H_S. We are interested in deriving simultaneous $(1-\alpha)$-confidence sets for this number, so that the simultaneity makes their coverage robust against post-hoc selection. This way, the rejected hypothesis can be selected after reviewing all confidence sets, while still keeping correct $(1-\alpha)$-coverage of the corresponding confidence set (Goeman and Solari, 2011).

Let $d : 2^M \to \mathbb{R}$ be a random function, where 2^M is the power set of M. We say that d has true discovery guarantee if $d(S)$ are simultaneous lower $(1-\alpha)$-confidence bounds for $\delta(S)$, i.e.,

$$P\left(\delta(S) \geq d(S) \text{ for each } S \subseteq M\right) \geq 1 - \alpha.$$

An equivalent condition is that $\{d(S), \dots, s\}$ is a $(1-\alpha)$-confidence set for $\delta(S)$, simultaneously for all $S \subseteq M$. Notice that the resulting confidence sets are one-sided, since hypothesis testing is focused on rejecting, not accepting. From $d(S)$ simultaneous $(1-\alpha)$-confidence sets can be immediately derived for other quantities of interest such as the TDP and the number or proportion of false discoveries (Goeman and Solari, 2011).

A general way to construct procedures with true discovery guarantee is provided by closed testing, based on the principle of testing different subsets by means of a valid α-level local test, which in this case is the permutation test. Throughout this paper, we will loosely say that a set S is rejected when the corresponding hypothesis H_S is. Hence denote the collection of sets rejected by the permutation test of Theorem 1 by

$$\mathcal{R} = \left\{ S \subseteq M \,:\, c_S^{(\omega)} > 0 \right\}.$$

Genovese and Wasserman (2006) and Goeman and Solari (2011) equivalently define a procedure d with true discovery guarantee as $d(S) = s - q(S)$, where

$$q(S) = \max\left\{ |V \cap S| \,:\, V \subseteq M,\, V \notin \mathcal{R} \right\} \tag{1}$$

is the maximum intersection between S and a set not rejected by the permutation test. The equivalence of the two methods is shown in Goeman et al. (2021).

The main challenge of this method is its exponential complexity in the number of hypotheses. Indeed, the number of tests that must be evaluated to determine $d(S)$ may be up to order 2^m. In the toy example, where $m = 5$, this number is 32; it is immediate that it quickly grows to an infeasible size as m increases.

6. Shortcut

Fix the set of interest S, so that any dependence on it may be omitted in the notation. We propose a shortcut that quickly evaluates whether $q < z$ for any value z. This will allow to approximate q, and eventually define a procedure with true discovery guarantee. First, we will re-write q as the unique change-point of an increasing function:

$$\phi : \{0, \dots, s+1\} \longrightarrow \{0, 1\}, \qquad \phi(z) = 1 \quad \text{if and only if} \quad q < z \tag{2}$$

$$q = \max\left\{ z \in \{0, \dots, s+1\} \,:\, \phi(z) = 0 \right\}. \tag{3}$$

Then we will approximate q from above with the change point $q^{(0)}$ of a second increasing function:

$$\underline{\phi} : \{0, \dots, s+1\} \longrightarrow \{0, 1\}, \qquad \underline{\phi}(z) \leq \phi(z) \tag{4}$$

$$q^{(0)} = \max\left\{ z \in \{0, \dots, s+1\} \,:\, \underline{\phi}(z) = 0 \right\}. \tag{5}$$

We start by giving an equivalent characterization of the quantity of interest q. For any $z \in \{0, \dots, s+1\}$, we define the collection $\mathcal{V}_z = \{V \subseteq M \,:\, |V \cap S| \geq z\}$ of sets that have at least size z overlap with S, and investigate whether all its elements are rejected. We define ϕ so that it represents such rejection, taking

$$\phi(z) = \mathbf{1}\{\mathcal{V}_z \subseteq \mathcal{R}\} \qquad (z \in \{0, \dots, s+1\}), \tag{6}$$

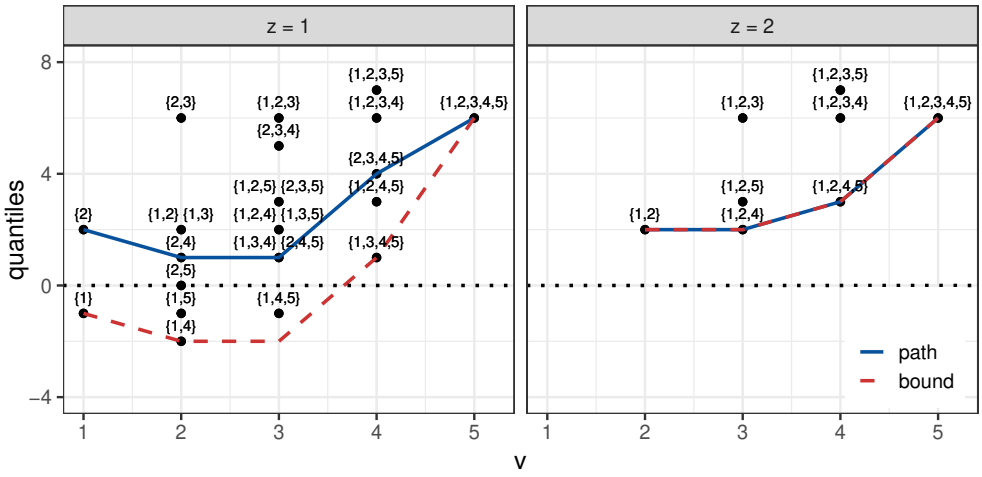

Fig. 1. Toy example with $S = \{1, 2\}$: shortcut to evaluate $\phi(z)$ in $z = 1$ and $z = 2$. Points denote the quantiles for the sets in $\mathcal{V}_z$. The dashed and solid lines represent the bound ℓ_z (8) and the path u_z (15), respectively.

where $\mathbf{1}\{\cdot\}$ denotes the indicator function. The following lemma shows that q can be written as in (3).

LEMMA 2. *$\phi(0) = 0$ and $\phi(s + 1) = 1$. Moreover, $\phi(z) = 0$ if and only if $z \in \{0, \dots, q\}$.*

Now we fix a value $z \in \{1, \dots, s\}$ and derive the shortcut to make statements on $\phi(z)$ without testing all the sets contained in $\mathcal{V}_z$. We do this by partitioning $\mathcal{V}_z$ by the size of its elements, obtaining

$$\mathcal{V}_z = \bigcap_{v=z}^{m} \mathcal{V}_z(v), \qquad \mathcal{V}_z(v) = \{V \in \mathcal{V}_z \,:\, |V| = v\}. \tag{7}$$

Each $\mathcal{V}_z(v)$ is the sub-collection of all sets of size v that have at least size z overlap with S. We can analyse these sub-collections separately and combine the results, noting that $\phi(z) = 1$ if and only if $\mathcal{V}_z(v) \subseteq \mathcal{R}$ for all $v \in \{z, \dots, m\}$.

By definition, $\mathcal{V}_z(v) \subseteq \mathcal{R}$ when all sets in the sub-collection have positive quantiles, i.e., $c_V^{(\omega)} > 0$ for each $V \in \mathcal{V}_z(v)$. The main idea of the shortcut is to obtain information on each sub-collection $\mathcal{V}_z(v)$ by bounding the corresponding quantiles from below. In particular, we will construct a bound

$$\ell_z \,:\, \{z, \dots, m\} \longrightarrow \mathbb{R}, \qquad \ell_z(v) \leq c_V^{(\omega)} \quad \text{for each } V \in \mathcal{V}_z(v). \tag{8}$$

This way, if $\ell_z(v) > 0$, we know that all sets in $\mathcal{V}_z(v)$ have positive quantiles. If ℓ_z is positive in its entire domain, then $\mathcal{V}_z(v) \subseteq \mathcal{R}$ for each v, and so $\phi(z) = 1$. Figure 1 displays the bound, which we will define in the following paragraphs, in the toy example

Table 2. Toy example with $S = \{1, 2\}$: matrix of the sorted centered statistics to compute the bound ℓ_1. The value $\ell_1(v)$ is obtained summing the first v columns by row, and then taking the quantile.

	selected in S	remaining			
	$i_1(\pi)$	$j_1(\pi)$	$j_2(\pi)$	$j_3(\pi)$	$j_4(\pi)$
id	0 (H_1)	0 (H_2)	0 (H_3)	0 (H_4)	0 (H_5)
π_2	3 (H_2)	-3 (H_5)	1 (H_4)	3 (H_3)	5 (H_1)
π_3	-2 (H_1)	-1 (H_4)	0 (H_5)	2 (H_2)	4 (H_3)
π_4	-2 (H_1)	0 (H_4)	1 (H_5)	4 (H_2)	4 (H_3)
π_5	-1 (H_2)	-1 (H_5)	0 (H_4)	3 (H_3)	6 (H_1)
π_6	-1 (H_1)	-1 (H_4)	0 (H_5)	3 (H_3)	5 (H_2)

for $z = 1$ and $z = 2$. Note that indeed all quantiles lie on it or above; the bound can be loose, as seen with $\ell_1(3)$. Since ℓ_2 lies entirely in the positive half-space, we know that $\phi(2) = 1$. In contrast, we cannot make a statement on $\phi(1)$ based on ℓ_1.

Fix a size $v \in \{z, \dots, m\}$. To define an $\ell_z(v)$ that does not exceed the minimum quantile over all sets in $\mathcal{V}_z(v)$, as required in (8), we approximate the minimum quantile from below with the quantile of the minimum. We do this by taking the smallest centered statistics for each transformation π, with some constraints from the structure of $\mathcal{V}_z(v)$.

In the toy example, choose $z = 1$, and let V be any set in the sub-collection $\mathcal{V}_1(v)$ of interest. Note that V must contain v indices, at least $z = 1$ of which is in S. Consider the centered statistics $c_i^{\pi_2}$ for transformation π_2 (second row in Table 1, right). First, we select the lowest value in S, then we sort the remaining values in ascending order, as in the second row of Table 2. If $b_v^{\pi_2}$ is the sum of the first v elements of the row, we know that $b_v^{\pi_2} \leq c_V^{\pi_2}$. After constructing the other rows of Table 2 according to the same principle, we define $\ell_1(v) = b_v^{(\omega)}$. Since $b_v^\pi \leq c_V^\pi$ for each π, we obtain $\ell_1(v) \leq c_V^{(\omega)}$.

In general, for each $\pi \in \boldsymbol{\pi}$, we select the z smallest centered statistics in S, and then the $v - z$ remaining smallest statistics. We define two permutations of the indices:

$$S = \{i_1(\pi), \dots, i_s(\pi)\} \quad : \quad c^\pi_{i_1(\pi)} \leq \dots \leq c^\pi_{i_s(\pi)} \tag{9}$$

$$M \setminus \{i_1(\pi), \dots, i_z(\pi)\} = \{j_1(\pi), \dots, j_{m-z}(\pi)\} \quad : \quad c^\pi_{j_1(\pi)} \leq \dots \leq c^\pi_{j_{m-z}(\pi)}. \tag{10}$$

The set $\{i_1(\pi), \dots, i_z(\pi)\}$ is a subset of S, containing the indices of the z smallest values in S (for transformation π). For instance, in the toy example we have $S = \{2, 1\}$, and $M \setminus \{2\} = \{5, 4, 3, 1\}$. Then the value of the bound is defined as

$$\ell_z(v) = b_v^{(\omega)} \qquad \text{where} \qquad b_v^\pi = \sum_{h=1}^{z} c^\pi_{i_h} + \sum_{h=1}^{v-z} c^\pi_{j_h} \quad (\pi \in \boldsymbol{\pi}). \tag{11}$$

LEMMA 3. *$\ell_z(v) \leq c_V^{(\omega)}$ for all $V \in \mathcal{V}_z(v)$. Hence $\min_v \ell_z(v) > 0$ implies $\phi(z) = 1$.*

Now we use the bound to define a function $\underline{\phi}$ as in (4). In the extremes, where the value of ϕ is known, we set $\underline{\phi}(0) = \phi(0) = 0$ and $\underline{\phi}(s+1) = \phi(s+1) = 1$ (see Lemma 2).

Elsewhere, we set

$$\underline{\phi}(z) = \mathbf{1}\left\{\min_v \ell_z(v) > 0\right\} \qquad (z \in \{1, \dots, s\}). \tag{12}$$

This function may not be monotone, but we are only interested in its smallest change point; indeed, if $\underline{\phi}(z) = 1$ for a value z, we know that $q < z$. We make it increasing and obtain a single change point in $q^{(0)}$, as defined in (5), by imposing

$$\underline{\phi}(z) = 1 \quad \text{if} \quad \underline{\phi}(z^*) = 1 \text{ for some } z^* \leq z \qquad (z \in \{1, \dots, s\}). \tag{13}$$

PROPOSITION 1. *As $\underline{\phi}(z) \leq \phi(z)$ for each $z \in \{0, \dots, s+1\}$, $q^{(0)} \geq q$.*

For instance, in the toy example of Figure 1, $\underline{\phi}(1) = 0$ and $\underline{\phi}(2) = 1$, and so $q^{(0)} = 1$. Finally, from this result we can approximate d from below with $d^{(0)} = s - q^{(0)}$.

THEOREM 2. *$d^{(0)} \leq d$.*

To summarise, Proposition 1 represents the basis of the shortcut. For any value z, it allows to make statements on the value of $\phi(z)$ by constructing $\underline{\phi}(z) \leq \phi(z)$; it requires to evaluate a number of tests which is linear in the total number m of hypotheses, in contrast to the exponential number required by closed testing. Theorem 2 employs the shortcut to provide a lower $(1-\alpha)$-confidence bound $d^{(0)}$ for the number of true discoveries δ. The theorem holds for all $S \subseteq M$, hence the procedure $d^{(0)}$ has true discovery guarantee. In Section S-1 we propose an algorithm for the shortcut, then we embed it into a binary search to approximate q with reduced complexity. We prove that in the worst case the computational complexity is of order $mB(\log^2 m + \log B)$. Moreover, we show how the method can be combined with an algorithm of Tian et al. (2022) to find the largest set with given TDP among a collection of incremental sets.

7. Equivalence to closed testing

The shortcut of Proposition 1 defines $\underline{\phi}(z) \leq \phi(z)$ for any z. For those values of z for which $\underline{\phi}(z) = 1$, we know that also $\phi(z) = 1$. Where $\underline{\phi}(z) = 0$, however, there are two distinct cases. If $\phi(z) = 0$, the shortcut is equivalent to closed testing; otherwise, if $\phi(z) = 1$, it is conservative, as it does not reject all sets in $\mathcal{V}_z$ while closed testing does. In the toy example with $z = 1$ we are in the first case (Figure 1, left), but we cannot see that from the bound only. Now we propose a sufficient condition to state that $\underline{\phi}(z) = \phi(z)$. This will play an important role in the iterative shortcut of Section 8. We will define an increasing function

$$\overline{\phi} : \{0, \dots, s+1\} \longrightarrow \{0, 1\}, \qquad \underline{\phi}(z) \leq \phi(z) \leq \overline{\phi}(z). \tag{14}$$

This way, if $\underline{\phi}(z) = \overline{\phi}(z)$ for a value z, we know that $\underline{\phi}(z) = \phi(z)$. Note that this holds in particular when either $\underline{\phi}(z) = 1$ or $\overline{\phi}(z) = 0$.

Fix $z \in \{1, \dots, s\}$. Based on partition (7) of $\mathcal{V}_z$, the main idea is to construct a greedy path of sets $V_z \subset \dots \subset V_m$, with $V_v \in \mathcal{V}_z(v)$ for each v, and check whether

Table 3. Toy example with $S = \{1, 2\}$: matrix of the sorted centered statistics to compute the path u_1. The value $u_1(v)$ is obtained summing the first v columns by row, and then taking the quantile.

	selected in S	remaining			
	i_1 (H_2)	j_1 (H_4)	j_2 (H_5)	j_3 (H_3)	j_4 (H_1)
id	0	0	0	0	0
π_2	3	1	-3	3	5
π_3	2	-1	0	4	-2
π_4	4	0	1	4	-2
π_5	-1	0	-1	3	6
π_6	5	-1	0	3	-1

their quantiles are all strictly positive. If we find a non-positive quantile, then we have established that $\mathcal{V}_z \not\subseteq \mathcal{R}$, and so $\underline{\phi}(z) = \phi(z) = 0$; the shortcut is equivalent to closed testing for this value of z. We will define the path

$$u_z : \{z, \dots, m\} \longrightarrow \mathbb{R}, \qquad u_z(v) = c_{V_v}^{(\omega)} \quad \text{with} \quad V_v \in \mathcal{V}_z(v) \tag{15}$$

that connects these quantiles. This way, if $u_z(v) \leq 0$, we know that $\mathcal{V}_z(v)$ contains a non-rejected set, and so $\phi(z) = 0$. Figure 1 displays the bound ℓ_z and the path u_z, which we will define in the next paragraphs, for the toy example. The path connects some of the quantiles, one for each size v, and so is never smaller than the bound. From ℓ_2 we already had $\phi(2) = 1$; as u_1 is entirely positive, results on $\phi(1)$ are still unsure.

Fix a size $v \in \{z, \dots, m\}$. We define $u_z(v)$ as the quantile of a set $V_v \in \mathcal{V}_z(v)$, as required in (15), choosing V_v such that it is unlikely to be rejected. We take V_v as the set containing the smallest observed non-centered statistics, with the constraint that V_v is an element of $\mathcal{V}_z(v)$. This is a heuristic choice: t_i by itself does not give full information on the rejection of H_i; still, if t_i is small, generally H_i is less likely to be rejected.

In the toy example, choose $z = 1$. The set $V_v \in \mathcal{V}_1(v)$ must contain v indices, at least $z = 1$ of which is in S. Consider the observed statistics t_i (first row in Table 1, left). First, we select the column of the smallest value in S, then sort the remaining columns so that their values are in ascending order. Table 3 presents the centered statistics c_i^π according to this new order. We define V_v as the set of the indices of the first v columns, obtaining $V_1 = \{2\}$, $V_2 = \{2, 4\}$, $V_3 = \{2, 4, 5\}$, $V_4 = \{2, 4, 5, 3\}$ and $V_5 = M$.

In general, we select the z smallest observed non-centered statistics in S, and then the $v - z$ remaining smallest statistics. We define two permutations of the indices:

$$S = \{i_1, \dots, i_s\} \quad : \quad t_{i_1} \leq \dots \leq t_{i_s} \tag{16}$$

$$M \setminus \{i_1, \dots, i_z\} = \{j_1, \dots, j_{m-z}\} \quad : \quad t_{j_1} \leq \dots \leq t_{j_{m-z}}. \tag{17}$$

The set $\{i_1, \dots, i_z\}$ is a subset of S, containing the indices of the z smallest values in S. For instance, in the toy example we have $S = \{2, 1\}$, and $M \setminus \{2\} = \{4, 5, 3, 1\}$. The value of the path is then defined as

$$u_z(v) = c_{V_v}^{(\omega)} \qquad \text{where} \qquad V_v = \{i_1, \dots, i_z\} \cup \{j_1, \dots, j_{v-z}\}. \tag{18}$$

It is immediate that $V_v \in \mathcal{V}_z(v)$ and $u_z(v) \geq \ell_z(v)$.

LEMMA 4. $\min_v u_z(v) \leq 0$ *implies* $\phi(z) = 0$.

The path is used to define a function $\overline{\phi}$ as in (14). Similarly to the definition of $\underline{\phi}$ in the previous section, first we set $\overline{\phi}(0) = \phi(0) = 0$, $\overline{\phi}(s+1) = \phi(s+1) = 1$, and

$$\overline{\phi}(z) = \mathbf{1}\left\{\min_v u_z(v) > 0\right\} \qquad (z \in \{1, \ldots, s\}). \tag{19}$$

Then we make the function increasing by taking only its largest change point, imposing

$$\overline{\phi}(z) = 0 \quad \text{if} \quad \overline{\phi}(z^*) = 0 \text{ for some } z^* \geq z \qquad (z \in \{1, \ldots, s\}). \tag{20}$$

PROPOSITION 2. $\underline{\phi}(z) \leq \phi(z) \leq \overline{\phi}(z)$ *for each* $z \in \{0, \ldots, s+1\}$. *Hence* $\underline{\phi}(z) = \overline{\phi}(z)$ *implies* $\underline{\phi}(z) = \phi(z)$, *i.e., equivalence between the shortcut and closed testing.*

For instance, in the toy example of Figure 1 we obtain $\underline{\phi}(1) = 0 < \overline{\phi}(1) = 1$ and $\underline{\phi}(2) = \overline{\phi}(2) = 1$. Hence the shortcut is equivalent to closed testing for $z = 2$, as we already observed, but we cannot establish equivalence for $z = 1$.

8. Iterative shortcut

The shortcut we have described in Section 6 approximates closed testing and efficiently computes $q^{(0)} \geq q$; however, as seen in Section 7, it may be conservative. In this section we improve this single-step shortcut by embedding it into a branch and bound algorithm. We obtain an iterative shortcut which defines closer approximations of q, and thus smaller confidence sets for δ, as the number of steps increases. Eventually, after a finite number of steps, it reaches the same results as full closed testing.

At each step $n \in \mathbb{N}$, we will define two increasing functions

$$\underline{\phi}^{(n)}, \overline{\phi}^{(n)} : \{0, \ldots, s+1\} \longrightarrow \{0, 1\}, \qquad \underline{\phi}^{(n)}(z) \leq \phi(z) \leq \overline{\phi}^{(n)}(z). \tag{21}$$

We will approximate q from above with the change point of the first function,

$$q^{(n)} = \max\left\{z \in \{0, \ldots, s+1\} \,:\, \underline{\phi}^{(n)}(z) = 0\right\}. \tag{22}$$

Then we will use the second to assess possible equivalence to closed testing. If $\underline{\phi}^{(n)}(z) = \overline{\phi}^{(n)}(z)$ for a value z, then $\underline{\phi}^{(n)}(z) = \phi(z)$ and so results cannot be further improved. Moreover, these functions will be defined so that $q^{(n)}$ becomes a better approximation of q as n increases, and finally converges to it after at most m steps:

$$q^{(n)} \geq q^{(n+1)} \geq q^{(m)} = q \qquad (n \in \mathbb{N}). \tag{23}$$

In the next sections we introduce the structure of the branch and bound algorithm, then use it to construct the functions $\underline{\phi}^{(n)}$ and $\overline{\phi}^{(n)}$ with the desired properties.

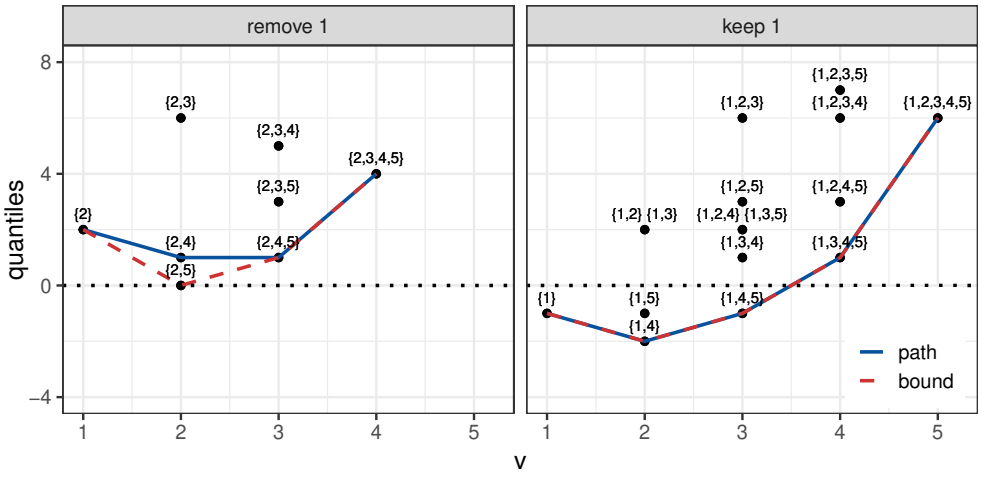

Fig. 2. Toy example with $S = \{1, 2\}$: iterative shortcut at step $n = 1$ to evaluate $\phi(z)$ in $z = 1$. Points denote the quantiles for the sets in $\mathcal{V}_1^-$ and $\mathcal{V}_1^+$. The dashed and solid lines represent the bound and the path, respectively.

8.1. Branch and bound

The branch and bound algorithm (Land and Doig, 1960; Mitten, 1970) is used when exploring a space of elements in search of a solution, and is based on the following principle. The space is partitioned into two subspaces, and each subspace is systematically evaluated; the procedure can be iterated until the best solution is found. Hence the algorithm consists of a branching rule, which defines how to generate subspaces, and a bounding rule, which gives bounds on the solution. This way, one can discard entire subspaces that, according to the bounding rule, cannot contain the solution.

Here, we want to evaluate $\phi(z)$ for any value z, i.e., determine whether the space $\mathcal{V}_z$ contains a non-rejected set (see definition (6)). The bounding rule that allows to make statements on the existence of such a set is the single-step shortcut of Propositions 1 and 2. If the shortcut is equivalent to closed testing, meaning that we are able to determine $\phi(z)$, the procedure stops; otherwise, we partition $\mathcal{V}_z$ and apply the shortcut within each resulting subspace. This procedure may be iterated as needed.

For instance, in the toy example, the single-step shortcut gives $\phi(2) = 1$ but cannot determine $\phi(1)$ (Figure 1). At step $n = 1$, we partition $\mathcal{V}_1$ into two subspaces $\mathcal{V}_1^-$ and $\mathcal{V}_1^+$, according to the inclusion of index $j^* = 1$: $\mathcal{V}_1^-$ contains all sets that do not include j^*, and $\mathcal{V}_1^+$ contains the others. We choose $j^* \in M$ as the index of the hypothesis that we believe we have most evidence against, i.e., having the greatest value t_i (first row in Table 1, left). Subsequently, we use the shortcut to examine each subspace. Figure 2 shows the bound ℓ_1 and the path u_1 in the two subspaces; the path indicates that $\mathcal{V}_1^+$ contains a non-rejected set, therefore we conclude that $\phi(1) = 0$.

In general, the branching rule is chosen to find an eventual non-rejected set with the smallest number of steps. Fix $z \in \{1, \dots, s\}$, as by Lemma 2 there is no need to partition

$\mathcal{V}_0$ or $\mathcal{V}_{s+1}$. The space $\mathcal{V}_z$ of interest is partitioned into

$$\mathcal{V}_z^- = \{V \in \mathcal{V}_z : j^* \notin V\}, \qquad \mathcal{V}_z^+ = \{V \in \mathcal{V}_z : j^* \in V\}$$

where j^* is the index of the greatest observed non-centered statistic, with the constraint that the procedure cannot generate empty subspaces. Recall that any set $V \in \mathcal{V}_z$ has at least size z overlap with S. Hence, with the notation of (16) and (17), we fix the indices $\{i_1, \dots, i_z\}$ of the z smallest observed statistics in S, then we take $j^* = j_{m-z}$ as the index of the greatest remaining observed statistic. The same principle may be applied to partition any subspace.

At any step $n \in \mathbb{N}$, the procedure partitions $\mathcal{V}_z$ into $K_{n,z}$ subspaces $\mathcal{V}_z^1, \dots, \mathcal{V}_z^{K_{n,z}}$ without any successors, where $K_{n,z} \in \{1, \dots, 2^n\}$. Suppose to apply the single-step shortcut within a subspace $\mathcal{V}_z^k$. If the result is $\phi(z) = 0$, then $\mathcal{V}_z^k$ contains a non-rejected set, and we stop with $\phi(z) = 0$. In contrast, if the shortcut determines that $\phi(z) = 1$, all sets in $\mathcal{V}_z^k$ are rejected, and we may explore other subspaces. Finally, if the shortcut produces an unsure outcome, i.e., $\phi(z)$ is still unknown, $\mathcal{V}_z^k$ can be partitioned again.

8.2. Structure of the iterative shortcut

Fix a step $n \in \mathbb{N}$. For every z, the branching rule partitions $\mathcal{V}_z$ into $K_{n,z}$ subspaces $\mathcal{V}_z^1, \dots, \mathcal{V}_z^{K_{n,z}}$, and the bounding rule applies the shortcut within them. We use this structure to define the functions $\underline{\phi}^{(n)}$ and $\overline{\phi}^{(n)}$ introduced in (21). We consider the point-wise minimums of $\underline{\phi}$ and $\overline{\phi}$ within the different subspaces, and so we take

$$\underline{\phi}^{(n)}(z) = \min_k \left\{ \underline{\phi}(z) \text{ in } \mathcal{V}_z^k \right\}, \qquad \overline{\phi}^{(n)}(z) = \min_k \left\{ \overline{\phi}(z) \text{ in } \mathcal{V}_z^k \right\}.$$

Since $\underline{\phi}$ and $\overline{\phi}$ are increasing functions, also $\underline{\phi}^{(n)}$ and $\overline{\phi}^{(n)}$ are increasing. The following proposition shows that property (21) holds, so that we can approximate q from above with $q^{(n)}$, and we can assess possible equivalence to closed testing for any z. Moreover, the proposition gives property (23) by showing that $\underline{\phi}^{(n)}$ and $\overline{\phi}^{(n)}$ become closer to ϕ as n increases, and finally converge to it after at most m steps.

PROPOSITION 3. *For any $n \in \mathbb{N}$ and any $z \in \{0, \dots, s+1\}$,*

$$\underline{\phi}^{(n)}(z) \le \underline{\phi}^{(n+1)}(z) \le \underline{\phi}^{(m)}(z) = \phi(z) = \overline{\phi}^{(m)}(z) \le \overline{\phi}^{(n+1)}(z) \le \overline{\phi}^{(n)}(z).$$

Hence $\underline{\phi}^{(n)}(z) = \overline{\phi}^{(n)}(z)$ implies $\underline{\phi}^{(n)}(z) = \phi(z)$, i.e., equivalence between the iterative shortcut and closed testing. Moreover, $q^{(n)} \ge q^{(n+1)} \ge q^{(m)} = q$.

In the toy example, consider step $n = 1$ of the iterative shortcut. For $z = 2$, from results of the single-step shortcut we have $\underline{\phi}^{(1)}(2) = \overline{\phi}^{(1)}(2) = \phi(2) = 1$ without partitioning $\mathcal{V}_2$. For $z = 1$, from Figure 2 we have $\underline{\phi}^{(1)}(1) = \overline{\phi}^{(1)}(1) = \phi(1) = 0$. After one step we obtain the same results as full closed testing, with $q^{(1)} = q = 1$. Then, similarly to Theorem 2, at each step n we may approximate d from below with $d^{(n)} = s - q^{(n)}$.

THEOREM 3. *$d^{(n)} \leq d^{(n+1)} \leq d^{(m)} = d$ for each $n \in \mathbb{N}$.*

Proposition 3 is the basis of the iterative shortcut. At any step n and for any z, it allows to make statements on the value of $\phi(z)$ by applying the single-step shortcut within at most 2^n subspaces. Then Theorem 3 gives lower $(1-\alpha)$-confidence bounds for the number of true discoveries δ. Even if the iterative shortcut is stopped early, before reaching convergence, $d^{(n)}$ is always a valid lower confidence bound; we have increasingly better approximations of d as n increases, and obtain full closed testing results after at most m steps. As the theorem may be applied to any $S \subseteq M$, the procedure $d^{(n)}$ has true discovery guarantee. In Section S-1 we provide an algorithm for the iterative shortcut. In the worst case, the complexity of each iteration, i.e., each application of the shortcut in a subspace, is of order $mB\log(mB)$. The algorithm converges to full closed testing results after a number of iterations of order 2^m.

9. Refinements

In this section we show two strategies that reduce the computational time of the shortcut. First we modify the ordering of the statistics used to define the path in Section 7 and the branching in Section 8.1; then we introduce truncated test statistics.

Both the path and the branching are constructed sorting the indices as in (16) and (17), with the intuition that a small observed value t_i corresponds to a hypothesis that is less likely to be rejected. This heuristic choice may be improved if we relate the observed value with all the permuted ones, i.e., if we sort $t_i - \text{mean}(t_i^\pi)$ instead of t_i. This modification proved to be slightly more efficient.

Subsequently, recall that the computational complexity of the shortcut increases with m. We argue that this complexity is much reduced if the method is applied to truncated statistics, as it allows to shrink the effective total number of hypotheses from m to $m' \in \{s, \dots, m\}$. In practice, with large B, m' is obtained by taking all statistics in S, and only the non-truncated observed statistics in $M \setminus S$.

Truncation-based statistics were advocated in the truncation product method of Zaykin et al. (2002), in the context of p-value combinations. The main idea was to emphasize smaller p-values by taking into account only p-values smaller than a certain threshold, and setting to 1 the others; a natural, common choice for the threshold is the significance level α. A similar procedure, the rank truncation product (Dudbridge and Koeleman, 2003; Kuo and Zaykin, 2011), takes into account only the k-th smallest p-values, for a given k. Eventually, weights can be incorporated into both analyses. Such procedures provide an increased power in many scenarios, and in particular for signal detection, when there is a predominance of near-null effects. They have been widely applied in literature (Yu et al., 2009; Li and Tseng, 2011; Biernacka et al., 2012; Dai et al., 2014); refer to Zaykin et al. (2007), Finos (2003) and Zhang et al. (2020) for a review of the methods and their applications.

With our notation, we can define a truncation-based statistic for H_S as following. For each hypothesis H_i, we set to a common ground value γ all statistics T_i^π smaller than a threshold τ_i. The threshold τ_i may depend on i, or be a prefixed value, or be the k-th greatest statistic T_i^π ($i \in M$, $\pi \in \boldsymbol{\pi}$) for a given k. The ground value must be $\gamma \leq \min_i \tau_i$; it may be chosen, for instance, as the minimum possible value of the test

Table 4. Toy example with $S = \{1, 2\}$: test statistics after truncation of elements smaller than $\tau = 2$ to the ground value $\gamma = 0$, and after dimensionality reduction.

	truncated $f(t_i^\pi)$					dim. reduction		
	H_1	H_2	H_3	H_4	H_5	H_1	H_2	$H_{4,5}$
id	6	5	4	0	0	6	5	0
π_2	0	2	0	0	4	0	2	4
π_3	8	3	0	2	0	8	3	2
π_4	8	0	0	0	0	8	0	0
π_5	0	6	0	0	2	0	6	2
π_6	7	0	0	2	0	7	0	2

statistics, or set equal to the smallest threshold $\min_i \tau_i$. Then

$$T_S = \sum_{i \in S} f_i(T_i), \qquad f_i(T_i) = \gamma \cdot \mathbf{1}\{T_i < \tau_i\} + T_i \cdot \mathbf{1}\{T_i \geq \tau_i\}.$$

For simplicity of notation, let $\tau_i = \tau$, and so $f_i = f$, be independent of i. Table 4 shows the values $f(t_i)$ in the toy example after truncation with $\tau = 2$ and $\gamma = 0$. Here, τ is set as the k-th greatest statistic, where $k = \lceil Bm\alpha \rceil$ is chosen so that the proportion of non-null contributions $f(t_i^\pi)$ is approximately α. Observe that H_3 is such that the observed truncated statistic is the greatest over all permutations, i.e., $f(t_3) = \max_\pi f(t_3^\pi)$; as a consequence, adding $\{3\}$ to any set V can only increase the number of rejections. On the contrary, H_4 and H_5 are such that the observed statistics are the smallest over all permutations, and so adding $\{4\}$ or $\{5\}$ to any set can only decrease rejections. Truncation makes those two particular cases more common as well as easier to check, through the following conditions:

$$f(t_i^\pi) = \gamma \quad \text{for all } \pi \in \boldsymbol{\pi} \setminus \{\text{id}\} \tag{24}$$

$$f(t_i) = \gamma \tag{25}$$

PROPOSITION 4. *Let $V \subseteq M$ and $i \in M$. If i satisfies condition* (24)*, then $V \in \mathcal{R}$ implies $(V \cup \{i\}) \in \mathcal{R}$. If i satisfies condition* (25)*, then $(V \cup \{i\}) \in \mathcal{R}$ implies $V \in \mathcal{R}$.*

The shortcut examines the collection $\mathcal{V}_z$ of sets that have at least size z overlap with S, searching for a set $V \notin \mathcal{R}$. In this case, the focus is on the number of indices in S, hence we may reduce the dimensionality of the problem by applying Proposition 4 to the remaining indices. If an index $i \in M \setminus S$ satisfies condition (24), then it is not useful for finding a non-rejected set, and so can be removed from M. If two indices $i, j \in M \setminus S$ satisfy condition (25), they may be collapsed into a new index h, so that $H_h = H_{\{i,j\}}$ can only decrease the number of rejections. This allows to reduce the total number of hypotheses from m for computational purposes to a substantially lower $m' \in \{s, \ldots, m\}$. In the toy example column 3 is removed, while columns 4 and 5 are collapsed into a single column, reducing the number of hypotheses from $m = 5$ to $m' = 3$.

10. Applications

In this section, we use the iterative shortcut of Section 8 to analyse simulated and real fMRI data, while in Section S-2.3 we analyse differential gene expression data. We use the `sumSome` package (Vesely, 2021) developed in `R` (R Core Team, 2017), with underlying code in `C++`.

10.1. Simulations

We use the shortcut to compare the performance of different p-value combinations through simulations. When using p-value combinations, the unknown joint distribution of the data is often managed through worst-case distributions, defined either generally or under restrictive assumptions (Vovk and Wang, 2020). However, this approach makes comparisons difficult, since different tests have different worst cases. In contrast, our method adapts to the unknown distribution through permutations, and thus allows to compare the tests on equal footing. Determining which test has the highest power in different settings is a major issue, for which a full treatment is out of the scope of the paper; we present a first exploration.

We simulate n independent observations from a multivariate normal distribution with m variables: $\mathbf{X} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$, with $\mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\varepsilon} \in \mathbb{R}^m$ and $\boldsymbol{\varepsilon} \sim \text{MVN}(0, \Sigma_\rho)$. Here Σ_ρ is an equicorrelation matrix with off-diagonal elements equal to ρ. The mean $\boldsymbol{\mu}$ has a proportion a of non-null entries, with value computed so that the two-sided one-sample t-test with significance level α has a given power β. From the resulting data, we obtain p-values applying a two-sided one-sample t-test for each variable i, with null hypothesis $H_i : \mu_i \neq 0$. P-values are computed for B random permutations. Moreover, we employ truncation, setting to a common ground value γ any p-value greater than a threshold τ.

We analyse the subset S of false hypotheses (active variables), and the complementary subset $M \setminus S$ of true hypotheses (inactive variables), by means of different p-value combinations: Pearson (1933), Liptak (1958), Cauchy (Liu and Xie, 2020), and generalized means with parameter $r \in \{-2, -1, -0.5, 0, 1, 2\}$ (Vovk and Wang, 2020). The latter will be denoted by VW(r). Notice that VW(-1) corresponds to the harmonic mean (Wilson, 2019), VW(0) to Fisher (1925), and VW(1) to Edgington (1972). As a comparison, we also apply the maxT-method of Westfall and Young (1993), corresponding to the limit of VW(r) when r tends to $-\infty$; we apply the usual algorithm for the maxT.

We fix $n = 50$, $m = 1000$, $\alpha = 0.05$, $B = 200$ and $\gamma = 0.5$, then we consider $a \in \{0, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 0.9\}$, $\beta \in \{0.5, 0.8, 0.95\}$, $\rho \in \{0, 0.3, 0.6, 0.9\}$, and $\tau \in \{0.005, 0.01, 0.05, 0.1, 1\}$, where $\tau = 1$ leads to no truncation. For each setting, we simulate data 1000 times, and compute the TDP lower confidence bound for the set S as the mean of $d(S)/s$ over the simulations. Furthermore, we compute the FWER as the proportion of simulations where $d(M \setminus S) > 0$, meaning that the method finds at least one discovery among the true hypotheses. The algorithm is run for a maximum of 1000 iterations.

Figure 3 shows the average TDP lower confidence bounds obtained in different scenarios for $\beta = 0.95$ and $\tau \in \{0.005, 0.05, 1\}$. Certain groups of tests have similar performances: (a) VW(1), VW(2) and Pearson; (b) VW(-1) and Cauchy. For clarity, among these tests, only VW(1) and VW(-1) are displayed in the plots. Results indicate that

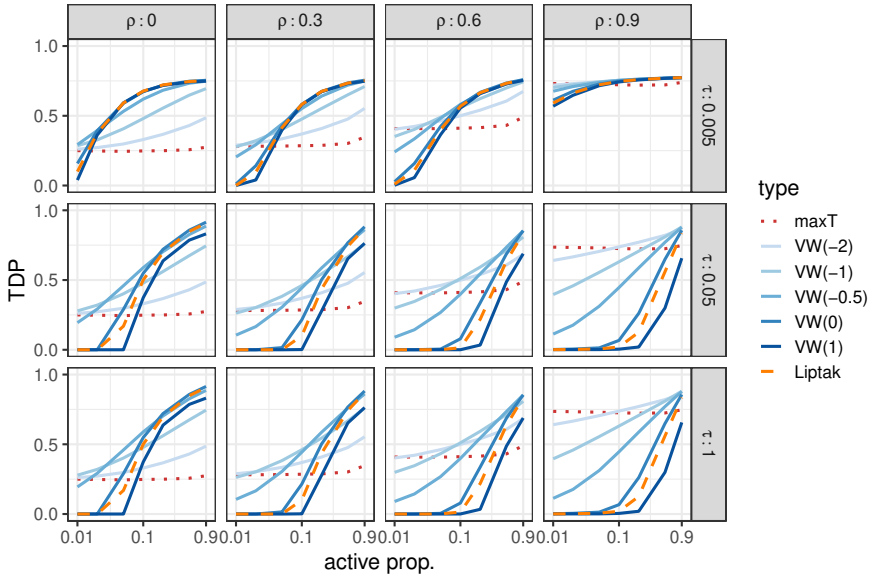

Fig. 3. Simulated data: TDP lower confidence bounds for the set S of active variables, by active proportion a (log scale) and for different p-value combinations. Variables have equicorrelation ρ. P-values greater than τ are truncated.

the intensity of the signal, determined by the parameter β, does not significantly affect the behaviour of the tests; nevertheless, differences between tests are amplified when the signal is high. Furthermore, results suggest that truncation is generally advisable, unless the signal is very dense, i.e., a is high. Indeed, in most cases tests tend to be more powerful when τ is low, and thus more statistics are truncated; the improvement is stronger for sparse signal, and when considering VW(0), VW(1) and Liptak.

When the signal is sparse, VW(r) with $r < 0$ performs best; the most powerful test is VW(-1) for low correlation, and VW(-2) for high correlation. The remaining tests perform well when the signal is dense; among those, in the considered scenarios VW(0) is the most powerful, but the powers of these tests become more similar as the signal becomes denser. These results confirm that the test is more directed towards sparse alternatives when the individual contributions, i.e., the transformed p-values, have heavy-tailed distributions, and towards dense alternatives otherwise (Vovk and Wang, 2020). Computation time is between 0.04 and 20 seconds. Moreover, simulations confirm that the method controls the FWER. Plots for the computation time, rates of convergence and the FWER are provided in Section S-2.1.

Finally, Section S-2.1 contains a comparison with closed testing based on worst-case distributions (Tian et al., 2022) for generalized means VW(r) (Vovk and Wang, 2020). As expected, worst-case distributions tend to be very conservative, and are never more powerful than the shortcut. The difference in power varies according to the choice of r and the setting. The largest differences are observed for $r = 1$ in settings with dense signal and medium-low correlation, for which only the shortcut has non-zero power.

10.2. fMRI data

In this section we apply the shortcut to fMRI brain imaging data, demonstrating feasibility of the method on large datasets, adaptation to the correlation structure and post-hoc flexibility. In fMRI imaging, Blood Oxygen Level Dependent (BOLD) response is measured, i.e., changes in blood flow in the brain induced by a sequence of stimuli, at the level of small volume units called voxels. Brain activation is then inferred as correlation between the stimuli and the BOLD response. Researchers are interested in studying this activation within different clusters, brain regions of connected voxels.

Typically, voxels are highly correlated. This is usually taken into account by means of cluster extent thresholding (Nichols, 2012; Woo et al., 2014; Rosenblatt et al., 2018). However, when the method finds activation in a given cluster, it only indicates that the cluster contains at least one active voxel, but does not provide any information on the proportion of active voxels (TDP) nor their spatial location. This leads to the spatial specificity paradox, the counter-intuitive property that activation in a large cluster is a weaker finding than in a small cluster (Woo et al., 2014). Moreover, follow-up inference inside a cluster leads to inflated Type I error rates (Kriegeskorte et al., 2009). In contrast, our approach not only adapts to the high correlation, but also provides confidence sets for the TDP, and allows for post-hoc selection and follow-up inference inside clusters.

We analyse data collected by Pernet et al. (2019), which compares subjects examined while listening to vocal and non-vocal sounds. Data consists of brain images for 140 subjects, each composed of 168,211 voxels. As for any standard fMRI analysis (Lindquist, 2008), as first-level analysis for each subject we estimate the contrast map that describes the difference in activation during vocal and non-vocal stimuli, with the same procedure of Andreella et al. (2020). Then these contrast maps are used to run the second-level analysis; for each voxel we compute a test statistic by means of a two-sided one-sample t-test, with the null hypothesis that the voxel's mean contrast between subjects is zero. Finally, we define the global test statistic for a cluster as the sum of its voxels' t-statistics.

We examine supra-threshold clusters with threshold 3.2, chosen by convention, and then we make follow-up inference inside those by studying clusters with threshold 4. The significance level is taken as $\alpha = 0.05$. We construct statistics for the permutation test by using B elements from the group of sign-flipping transformations, which satisfies Assumption 1 (Winkler et al., 2014). Moreover, we employ truncation as in Section 9 by setting to $\gamma = 0$ any statistic smaller than $\tau = 3.2$; this way, we take into account only statistics at least as extreme as the cluster-defining threshold. We use two settings. First, we apply a 'quick' analysis, fast and feasible on a standard machine, by using $B = 200$ transformations and stopping after 50 iterations of the single-step shortcut. Subsequently, we consider a 'long' analysis, run on the platform CAPRI (University of Padua, 2017), that employs $B = 1000$ transformations and stops after 1000 iterations. Computation time for the 'quick' setting is less than 8 minutes on a standard PC, while the 'long' setting requires around 9 hours for clusters with threshold 3.2, and 36 hours for follow-up inference on clusters with threshold 4.

Results, shown in Section S-2.2, indicate that the setting of the 'long' analysis does not provide larger TDP values than the 'quick'. Notice that the method provides valid $(1 - \alpha)$-confidence bounds for the TDP in all settings. In Section S-2.2 we further investigate the role of the numbers of iterations and permutations, confirming that, even

though larger values give greater mean power and less variability, the 'quick' setting provides suitable power. Moreover, our method finds activation in concordance with previous studies. An extensive comparison with other methods is beyond the scope of this paper, however our results can be immediately compared to those in Andreella et al. (2020), since the same data was used. For the particular settings used in the analyses, the proposed method is more powerful in detecting signal in bigger clusters, while loses power in smaller ones. In general, however, results strongly depend on the choice of the tests: the sum test in the proposed method, and the critical vector in Andreella et al. (2020). A preliminary study is shown in Vesely et al. (2021).

11. Discussion

We have proposed a new perspective on the age-old subject of global testing, arguing that all global tests automatically come with an inbuilt selective inference method, allowing many additional inferences to be made without paying a price in terms of the global test's α-level. Our proposed approach provides not just p-values but gives a confidence bound for the TDP, which is considerably more informative; indeed, reporting a p-value only infers the presence of some discoveries, while the TDP allows to quantify the proportion of these discoveries. Such TDP confidence bounds come not just for the full testing problem, but also simultaneously for all subsets of hypotheses; this way, subsets of interest may be chosen post hoc, without compromising the validity of the method.

To construct simultaneous confidence bounds for the TDP of all subsets, we have provided a general closed testing procedure for sum tests, a broad class of global tests that includes many p-value combinations and other popular multiple testing methods. The procedure uses permutation testing to adapt to the unknown joint distribution of the data, avoiding strong assumptions or potential loss of power due to worst-case distributions. We have presented an iterative shortcut for this procedure, where the complexity of each iteration is linearithmic both in the numbers m of hypotheses and B of permutations. Moreover, we have argued that $B = 200$ permutations are generally sufficient for the usual significance level $\alpha = 0.05$. The shortcut converges to full closed testing results after a finite, but possibly exponential in m, number of iterations; furthermore, it may be stopped at any time while still providing control of the TDP. As shown in simulations, when studying 1000 hypotheses, in many cases the procedure converges to closed testing in seconds. Moreover, the method is feasible in high-dimensional settings, as shown in applications on fMRI data and differential gene expression data. An implementation is available in the `sumSome` package (Vesely, 2021) in `R`, with underlying code in `C++`.

Our method is extremely flexible, allowing any sum test of choice; different choices of the sum test have very different power properties, as we have illustrated. More research is needed on the performance of different sum tests in different scenarios. Notice that the test statistic, including the eventual truncation, needs to be chosen a priori, before performing the analysis. Moreover, permutations are known to have a better performance than worst-case distributions under general dependence structure, but we have performed only a preliminary investigation to quantify the improvement given by permutations in the case of sum tests. Finally, a comparison with other permutation-based procedures that rely on bounding functions (Blanchard et al., 2020; Andreella

et al., 2020; Blain et al., 2022) would be of great interest, but would be extensive for two main reasons. First, all these procedures do not represent single methods but families of methods, allowing different choices for the test (i.e., sum test statistic in our case, and critical vector in the others); where and how the signal is distributed strongly influences the power of each method. Hence a fair study would require to first choose a proper test within each family, depending on many different characteristics of the problem, and only then compare results. Furthermore, the methods give statements for each of the 2^m possible subsets of hypotheses. Depending on the loss function chosen to summarize these statements, different methods could result to be preferable. In consequence, such an analysis is left for future work.

Data availability and funding

The data underlying this article are available in the OpenNeuro dataset ds000158 at `https://doi.org/10.18112/openneuro.ds000158.v1.0.0` (raw data), and at `https://github.com/angeella/fMRIdata/tree/master/data-raw/AuditoryData` (pre-processed data). The code for data simulation and analysis is available in the `R` package `sumSome`, at `https://CRAN.R-project.org/package=sumSome`.

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. Some of the analyses were carried out using the University of Padua Strategic Research Infrastructure Grant 2017: 'CAPRI: Calcolo ad Alte Prestazioni per la Ricerca e l'Innovazione', `https://capri.dei.unipd.it`.

References

Andreella, A., Hemerik, J., Weeda, W., Finos, L. and Goeman, J. J. (2020) Permutation-based true discovery proportions for fMRI cluster analysis. ArXiv: 2012.00368.

Biernacka, J. M., Jenkins, G. D., Wang, L., Moyer, A. M. and Fridley, B. L. (2012) Use of the gamma method for self-contained gene-set analysis of SNP data. *Eur. J. Hum. Genet.*, **20**, 565–571.

Blain, A., Thirion, B. and Neuvial, P. (2022) Notip: Non-parametric true discovery proportion control for brain imaging. *NeuroImage*, **260**, 119492.

Blanchard, G., Neuvial, P. and Roquain, E. (2020) Post hoc confidence bounds on false positives using reference families. *Ann. Statist.*, **48**, 1281–1303.

Dai, H., Leeder, J. S. and Cui, Y. (2014) A modified generalized fisher method for combining probabilities from dependent tests. *Front. Genet.*, **5**.

Donoho, D. and Jin, J. (2015) Higher criticism for large-scale inference, especially for rare and weak effects. *Statist. Sci.*, **30**, 1–25.

Dudbridge, F. and Koeleman, B. P. C. (2003) Rank truncated product of p-values, with application to genomewide association scans. *Genet. Epidemiol.*, **25**, 360–366.

Ebrahimpoor, M., Spitali, P., Hettne, K., Tsonaka, R. and Goeman, J. J. (2020) Simultaneous enrichment analysis of all possible gene-sets: unifying self-contained and competitive methods. *Brief. Bioinform.*, **21**, 1302–1312.

Edgington, E. S. (1972) An additive method for combining probability values from independent experiments. *J. Psychol.*, **80**, 351–363.

Ernst, M. D. (2004) Permutation methods: a basis for exact inference. *Statist. Sci.*, **19**, 676–685.

Finos, L. (2003) *Metodi Non Parametrici per l'Analisi Multi-Focus e per il Controllo della Molteplicità con Applicazioni in Ambito Biomedico*. Ph.D. thesis, Dep. of Statistical Sciences, University of Padua.

Fisher, R. A. (1925) *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.

— (1936) "The coefficient of racial likeness" and the future of craniometry. *J. R. Anthropol. Inst. of Great Britain and Ireland*, **66**, 57–63.

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

Goeman, J. J., van de Geer, S. A. and van Houwelingen, H. C. (2006) Testing against a high dimensional alternative. *J. R. Statist. Soc. B*, **68**, 477–493.

Goeman, J. J., Hemerik, J. and Solari, A. (2021) Only closed testing procedures are admissible for controlling false discovery proportions. *Ann. Statist.*, **49**, 1218–1238.

Goeman, J. J., Meijer, R. J., Krebs, T. J. P. and Solari, A. (2019) Simultaneous control of all false discovery proportions in large-scale multiple hypothesis testing. *Biometrika*, **106**, 841–856.

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

— (2011) Multiple testing for exploratory research. *Statist. Sci.*, **26**, 584–597.

Hemerik, J. and Goeman, J. J. (2018a) Exact testing with random permutations. *TEST*, **27**, 811–825.

— (2018b) False discovery proportion estimation by permutations: confidence for significance analysis of microarrays. *J. R. Statist. Soc. B*, **80**, 137–155.

Hemerik, J., Goeman, J. J. and Finos, L. (2020) Robust testing in generalized linear models by sign flipping score contributions. *J. R. Statist. Soc. B*, **82**, 841–864.

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

Hoeffding, W. (1952) The large-sample power of tests based on permutations of observations. *Ann. Math. Statist.*, **23**, 169–192.

Huang, Y., Xu, H., Calian, V. and Hsu, J. C. (2006) To permute or not to permute. *Bioinformatics*, **22**, 2244–2248.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F. and Baker, C. I. (2009) Circular analysis in systems neuroscience: the dangers of double dipping. *Nat. Neurosci.*, **12**, 535–540.

Kuo, C.-L. and Zaykin, D. V. (2011) Novel rank-based approaches for discovery and replication in genomewide association studies. *Genetics*, **189**, 329–340.

Lancaster, H. O. (1961) The combination of probabilities: an application of orthonormal functions. *Aust. J. Statist.*, **3**, 20–33.

Land, A. H. and Doig, A. G. (1960) An automatic method of solving discrete programming problems. *Econometrica*, **28**, 497–520.

Langsrud, Ø. (2005) Rotation tests. *Statist. Comput.*, **15**, 53–60.

Li, J. and Tseng, G. C. (2011) An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. *Ann. Appl. Statist.*, **5**, 994–1019.

Lindquist, M. A. (2008) The statistical analysis of fMRI data. *Statist. Sci.*, **23**, 439–464.

Liptak, T. (1958) On the combination of independent tests. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, **3**, 1971–1977.

Liu, Y. and Xie, J. (2020) Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *J. Am. Statist. Ass.*, **115**, 393–402.

Loughin, T. M. (2004) A systematic comparison of methods for combining p-values from independent tests. *Comput. Statist. Data Anal.*, **47**, 467 485.

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

Meijer, R. J. and Goeman, J. J. (2016) Multiple testing of gene sets from gene ontology: possibilities and pitfalls. *Brief. Bioinformatics*, **17**, 808–818.

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

Mitten, L. G. (1970) Branch-and-bound methods: general formulation and properties. *Ops. Res.*, **18**, 24–34.

Nichols, T. E. (2012) Multiple testing corrections, nonparametric methods, and random field theory. *NeuroImage*, **62**, 811–815.

Pearson, K. (1933) On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. *Biometrika*, **25**, 379–410.

Pernet, C. R., Belin, P., McAleer, P., Gorgolewski, K. J., Valdes-Sosa, M., Latinus, M., Charest, I., Bestelmeyer, P. E. G., Watson, R. H. and Fleming, D. (2019) The human voice areas: spatial organisation and inter-individual variability in temporal and extra-temporal cortices. OpenNeuro, dataset ds000158.v1.0.0.

Pesarin, F. (2001) *Multivariate Permutation Tests: with Applications in Biostatistics*. New York: Wiley.

Pesarin, F. and Salmaso, L. (2010) *Permutation Tests for Complex Data: Theory, Applications and Software*. New York: Wiley.

R Core Team (2017) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. URL: `https://www.R-project.org/`.

Rosenblatt, J. D., Finos, L., Weeda, W. D., Solari, A. and Goeman, J. J. (2018) All-Resolutions Inference for brain imaging. *NeuroImage*, **181**, 786–796.

Schildknecht, K., Olek, S. and Dickhaus, T. (2015) Simultaneous statistical inference for epigenetic data. *PLOS ONE*.

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

Solari, A., Finos, L. and Goeman, J. J. (2014) Rotation-based multiple testing in the multivariate linear model. *Biometrics*, **70**, 954–961.

Southworth, L. K., Kim, S. K. and Owen, A. B. (2009) Properties of balanced permutations. *J. Comput. Biol.*, **16**, 625–638.

Tian, J., Chen, X., Katsevich, E., Goeman, J. J. and Ramdas, A. (2022) Large-scale simultaneous inference under dependence. *Scand. J. Statist.*, 1–47.

University of Padua (2017) *CAPRI: Calcolo ad Alte Prestazioni per la Ricerca e l'Innovazione*. Strategic Research Infrastructure Grant. URL: `https://capri.dei.unipd.it`.

Vesely, A. (2021) *sumSome: permutation true discovery guarantee by sum-based tests*. URL: `https://CRAN.R-project.org/package=sumSome`. R package.

Vesely, A., Finos, L., Goeman, J. J. and Andreella, A. (2021) Valid double-dipping via permutation-based closed testing. In *Book of Short Papers SIS 2021* (eds. C. Perna, N. Salvati and F. S. Spagnolo), 776–781.

Vovk, V. and Wang, R. (2020) Combining p-values via averaging. *Biometrika*, **asaa027**.

— (2021) E-values: Calibration, combination and applications. *Ann. Statist.*, **49**, 1736–1754.

Westfall, P. H. and Young, S. S. (1993) *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. New York: Wiley.

Wilson, D. J. (2019) The harmonic mean p-value for combining dependent tests. *Proc. Natl. Acad. Sci.*, **116**, 1195–1200.

Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M. and Nichols, T. E. (2014) Permutation inference for the general linear model. *NeuroImage*, **92**, 381–397.

Won, S., Morris, N., Lu, Q. and Elston, R. C. (2009) Choosing an optimal method to combine p-values. *Statist. Med.*, **28**, 1537–1553.

Woo, C.-W., Krishnan, A. and Wager, T. D. (2014) Cluster-extent based thresholding in fMRI analyses: pitfalls and recommendations. *Neuroimage*, **33**, 412–419.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M. and Lin, X. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.*, **89**, 82–93.

Yu, K., Li, Q., Bergen, A. W., Pfeiffer, R. M., Rosenberg, P. S., Caporaso, N., Kraft, P. and Chatterjee, N. (2009) Pathway analysis by adaptive combination of p-values. *Genet. Epidemiol.*, **33**, 700–709.

Zaykin, D. V., Zhivotovsky, L. A., Czika, W., Shao, S. and Wolfinger, R. D. (2007) Combining p-values in large-scale genomics experiments. *Pharm. Statist.*, **6**, 217–226.

Zaykin, D. V., Zhivotovsky, L. A., Westfall, P. H. and Weir, B. S. (2002) Truncated product method for combining p-values. *Genet. Epidemiol.*, **22**, 170–185.

Zhang, H., Tong, T., Landers, J. and Wu, Z. (2020) TFisher: a powerful truncation and weighting procedure for combining p-values. *Ann. Appl. Statist.*, **14**, 178–201.

List of figures

Figure 1: Toy example with $S = \{1, 2\}$: shortcut to evaluate $\phi(z)$ in $z = 1$ and $z = 2$. Points denote the quantiles for the sets in $\mathcal{V}_z$. The dashed and solid lines represent the bound ℓ_z (8) and the path u_z (15), respectively. (p. 8)

Figure 2: Toy example with $S = \{1, 2\}$: iterative shortcut at step $n = 1$ to evaluate $\phi(z)$ in $z = 1$. Points denote the quantiles for the sets in $\mathcal{V}_1^-$ and $\mathcal{V}_1^+$. The dashed and solid lines represent the bound and the path, respectively. (p. 13)

Figure 3: Simulated data: TDP lower confidence bounds for the set S of active variables, by active proportion a (log scale) and for different p-value combinations. Variables have equicorrelation ρ. P-values greater than τ are truncated. (p. 18)