CrossMark

ORIGINAL PAPER

Closure properties of classes of multiple testing procedures

Georg Hahn[1]

Abstract Statistical discoveries are often obtained through multiple hypothesis testing. A variety of procedures exists to evaluate multiple hypotheses, for instance the ones of Benjamini–Hochberg, Bonferroni, Holm or Sidak. We are particularly interested in multiple testing procedures with two desired properties: (solely) monotonic and well-behaved procedures. This article investigates to which extent the classes of (monotonic or well-behaved) multiple testing procedures, in particular the subclasses of so-called step-up and step-down procedures, are closed under basic set operations, specifically the union, intersection, difference and the complement of sets of rejected or non-rejected hypotheses. The present article proves two main results: First, taking the union or intersection of arbitrary (monotonic or well-behaved) multiple testing procedures results in new procedures which are monotonic but not well-behaved, whereas the complement or difference generally preserves neither property. Second, the two classes of (solely monotonic or well-behaved) step-up and step-down procedures are closed under taking the union or intersection, but not the complement or difference.

Keywords Multiple hypothesis testing · Statistical significance · Step-up procedure · Set operations · Monotonicity

Mathematics Subject Classification 62G10

1 Introduction

Multiple testing is a widespread tool to evaluate scientific studies (Westfall and Young 1993; Hsu 1996; Hochberg and Tamhane 2008). We are interested in testing $m \in$

✉ Georg Hahn
    g.hahn11@ic.ac.uk

1   Department of Mathematics, Imperial College London, London SW7 2AZ, UK

$\mathbb{N}$ hypotheses $H_{01}, \ldots, H_{0m}$ with corresponding p-values $p_1, \ldots, p_m$ for statistical significance while controlling an error criterion such as the familywise error (FWER) or the false discovery rate (FDR). Following Gandy and Hahn (2016), we define a multiple testing procedure as a mapping

$$h : [0, 1]^m \times [0, 1] \rightarrow \mathcal{P}(\{1, \ldots, m\})$$

whose input is a vector of m p-values $p \in [0, 1]^m$ and a significance level $\alpha \in [0, 1]$ and whose output is the set of indices of rejected hypotheses, where $\mathcal{P}$ denotes the power set.

Many procedures of the above form are available in the literature in order to correct for multiple tests, for instance the procedures of Bonferroni (1936), Sidak (1967), Holm (1979), Hochberg (1988) or Benjamini and Hochberg (1995). Many common procedures, including the ones aforementioned, belong to a certain class of procedures, called step-up and step-down procedures (Romano and Shaikh 2006). It is assumed throughout the article that only the m p-values which serve as input to h are used as a basis for making decisions, dependencies between elementary hypotheses are not considered explicitly. Apart from defining properties on p imposed by those multiple testing procedures to which the results of this article are applied, no additional conditions on p are required.

This article focuses on two types of multiple testing procedures: *monotonic* procedures defined in Roth (1999) and Tamhane and Liu (2008) as well as *well-behaved* procedures (Gandy and Hahn 2016). We investigate to which extent the class of solely monotonic and the class of well-behaved multiple testing procedures is closed under the computation of the union, intersection, difference or the complement of sets of rejected or non-rejected hypotheses.

A multiple testing procedure is said to be monotonic if smaller p-values (Tamhane and Liu 2008) or a higher significance level (Roth 1999) lead to more rejections. Gandy and Hahn (2016) call a monotonic multiple testing procedure well-behaved if p-values corresponding to rejected hypotheses can be lowered and p-values corresponding to non-rejected hypotheses can be increased while leaving all rejections and non-rejections invariant.

For a set of given hypotheses, the closed testing procedure (CTP) of Marcus et al. (1976) (also referred to as the *closure principle*) and the partitioning principle (PP) of Finner and Strassburger (2002) provide means to efficiently construct a simultaneous hypothesis test controlling the FWER. The CTP is based on enforcing *coherence* (Gabriel 1969): An intersection hypothesis H_I, that is a hypothesis of the form $H_I = \cap_{i \in I} H_i$ for $I \subseteq \{1, \ldots, m\}$, is rejected if and only if all intersection hypotheses implying H_I are rejected by their local tests (Hommel et al. 2007). Many common procedures such as the one of Holm (1979) can be constructed using the CTP. The PP divides the parameter space underlying the hypotheses of interest into disjoint subsets which are then tested independently at level α. Since the partitioned hypotheses are disjoint, no multiplicity correction is necessary and at most one of the mutually exclusive hypotheses is true. Whereas CTP and PP can only be used to construct procedures with FWER control, the present article offers a means to combine procedures controlling several criteria such as the FDR into one procedure (see the example in

Sect. 4.5). In case of the CTP, the exponential number of tests to be carried out might also pose a problem: The present article considers the direct construction of step-up and step-down procedures which allow for efficient testing of multiple hypotheses.

The motivation for the present article is as follows:

1. Investigating closure properties (in a set theoretical sense) of a class, in the case of the present article certain classes of multiple testing procedures, is of interest in its own right: The closure of step-up and step-down procedures allows us to construct new multiple testing procedures of the same (step-up/step-down) form from existing ones; moreover, the resulting procedure will be given explicitly.

2. Being able to perform set operations with multiple testing procedures is useful in practice: Many multiple testing procedures exist to test hypotheses according to various criteria, each of which might prove beneficial in certain applications. Whereas hypotheses can also be tested sequentially using several procedures, it is non-trivial a priori that procedures can be combined to test multiple hypotheses in a single run while drawing benefits of several criteria simultaneously. This feature is similar to using (stepwise) "shortcut procedures" (Romano and Wolf 2005; Hommel et al. 2007) which aim to reduce the (potentially) exponential number of tests required by the CTP for FWER control to a polynomial number of tests.

3. Monotonic and well-behaved procedures have already been of interest in the literature. For instance, Gordon (2007) uses the idea of monotonicity to show that there is no monotonic step-up procedure which improves upon the Bonferroni (1936) procedure in the sense that it always returns the same rejections or possibly more. Gordon and Salzman (2008) show that the classical Holm (1979) procedure dominates all monotonic step-down multiple testing procedures in the above sense. Proving that certain classes of procedures (for instance, monotonic procedures) are closed renders the applicability of known results more apparent.

4. The results discussed in this paper extend the methodology developed in Gandy and Hahn (2014) and Gandy and Hahn (2016) which relies on well-behaved procedures. Briefly, the authors consider a scenario in which the p-value underlying each hypothesis is unknown, but can be estimated through Monte Carlo samples drawn under the null, for instance using bootstrap or permutation tests. Instead of using estimated p-values to obtain ad hoc decisions on all hypotheses, the authors prove that it is possible to improve existing algorithms designed for Monte Carlo-based multiple testing (Besag and Clifford 1991; Lin 2005; Wieringen et al. 2008; Guo and Peddada 2008; Sandve et al. 2011): the proposed modifications guarantee that the test results of published algorithms are identical (up to an error probability pre-specified by the user) to the ones obtained with the unknown p-values. This ensures the repeatability and objectivity of multiple testing results even in the absence of p-values.

The article is structured as follows. Section 2 provides formal definitions of the two properties of a multiple testing procedure under investigation. Section 3 considers arbitrary (solely monotonic or well-behaved) multiple testing procedures and demonstrates that solely the monotonicity is preserved when taking unions and intersections. The difference and complement are neither monotonic nor well-behaved. Section 4 focuses on step-up and step-down procedures and shows that both classes of (solely

monotonic or well-behaved) step-up and step-down procedures are closed under the union or intersection operation, but not the complement or difference. The article concludes with a short discussion in Sect. 5. All proofs are given in Appendix 6. In the entire article, $|\cdot|$ and $\|\cdot\|$ denote the absolute value and the Euclidean norm, respectively, and $M := \{1, \ldots, m\}$.

2 Basic definitions

Consider a step-up (h^u) and step-down (h^d) procedure

$$h^u(p, \alpha) = \left\{ i \in \{1, \ldots, m\} : p_i \leq \max\{p_{(j)} : p_{(j)} \leq \tau_\alpha(j)\} \right\},$$
$$h^d(p, \alpha) = \left\{ i \in \{1, \ldots, m\} : p_i < \min\{p_{(j)} : p_{(j)} > \tau_\alpha(j)\} \right\},$$

returning the set of indices of rejected hypotheses (Gandy and Hahn 2016), where $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$ refers to the ordered p-values. Any procedure of the above form is fully characterised by a threshold function $\tau_\alpha : \{1, \ldots, m\} \to [0, 1]$ returning the critical value $\tau_\alpha(i)$ each $p_{(i)}$ is compared to. A step-up procedure first determines the largest $j \in M$ such that the p-value $p_{(j)}$ lies below $\tau_\alpha(j)$ and then rejects all hypotheses having p-values up to $p_{(j)}$. Likewise, a step-down procedure non-rejects all those hypotheses with p-values larger or equal to the smallest p-value above the threshold function.

We now consider two useful properties of arbitrary multiple testing procedures. The first one, monotonicity, states that smaller p-values (Tamhane and Liu 2008) or a higher significance level (Roth 1999) lead to more rejections:

Definition 1 A multiple testing procedure h is *monotonic* if $h(p, \alpha) \subseteq h(q, \alpha')$ for $p \geq q$ and $\alpha \leq \alpha'$.

The monotonicity in α introduced by Roth (1999), also called α-consistency (Hommel and Bretz 2008), is a natural property desired for any testing procedure since testing at a more stringent significance level should never result in more rejections (Dmitrienko and Tamhane 2013).

Gandy and Hahn (2016) introduce another useful property, the class of well-behaved multiple testing procedures. Such procedures, in connection with a generic algorithm presented in Gandy and Hahn (2016), allow to use p-value estimates obtained with independent samples under the null to compute test results which are proven to be identical (up to a pre-specified error probability) to the ones obtained with the unknown p-values. A monotonic multiple testing procedure h is well-behaved if it additionally satisfies the following condition.

Condition 1 *1. Let $p, q \in [0, 1]^m$ and $\alpha \in \mathbb{R}$. If $q_i \leq p_i \ \forall i \in h(p, \alpha)$ and $q_i \geq p_i$ $\forall i \notin h(p, \alpha)$, then $h(p, \alpha) = h(q, \alpha)$.*
2. Fix $p^* \in [0, 1]^m$ and $\alpha^* \in [0, 1]$. Then, there exists $\delta > 0$ such that $p \in [0, 1]^m$, $\alpha \in [0, 1]$ and $\max(\|p - p^*\|, |\alpha - \alpha^*|) < \delta$ imply $h(p, \alpha) = h(p^*, \alpha^*)$.

Well-behaved procedures stay invariant if rejected (non-rejected) p-values are replaced by smaller (larger) values. Moreover, well-behaved procedures are constant on a δ-neighbourhood around fixed inputs p^* and α^*.

The level α is a parameter in Condition 1 to account for settings in which α is unknown a-priori: This can occur, for instance, when the significance level depends on an estimate of the proportion of true null hypotheses which is often a functional of p (Gandy and Hahn 2016, Sect. 2.2). Condition 1 is a generalisation of (Gandy and Hahn 2014, Condition 1) which states the same invariance property for the case that α is a given constant: In this case, h is solely a function of p and the condition $|\alpha - \alpha^*| < \delta$ in the second part of Condition 1 can be ignored.

3 Arbitrary multiple testing procedures

We define the union, intersection, difference and the complement of two procedures to be the equivalent operations on the sets of rejected hypotheses returned by the two procedures. Formally, for two multiple testing procedures h_1 and h_2 we define

$$h_1 \cup h_2 : [0, 1]^m \times [0, 1] \rightarrow \mathcal{P}(\{1, \ldots, m\}),$$
$$h_1 \cup h_2(p, \alpha) := h_1(p, \alpha) \cup h_2(p, \alpha),$$

and similarly $h_1 \cap h_2$, $h_1 \backslash h_2$ and the complement $h_i(p, \alpha)^c := \{1, \ldots, m\} \backslash h_i(p, \alpha)$, where $i \in \{1, 2\}$.

In what follows, we sometimes drop the dependence of $h(p, \alpha)$ on p, on α, or on both parameters. The following lemma summarises the results.

Lemma 1 *Let h_1 and h_2 be two well-behaved multiple testing procedures.*

1. *$h_1 \cup h_2$ and $h_1 \cap h_2$ are monotonic and satisfy part 2. of Condition 1.*
2. *$h_i(p, \alpha)^c$ and $h_1 \backslash h_2$ are not monotonic, $i \in \{1, 2\}$.*

As well-behaved procedures are also monotonic, the complement or difference of two procedures is also not well-behaved.

Although by Lemma 1, both the union and the intersection are monotonic, they do not necessarily allow to lower the p-values of rejected hypotheses or to increase the p-values of non-rejected hypotheses (first part of Condition 1) as demonstrated in the following two counterexamples.

Example 1 Let $p^* = (0.034, 0.06, 1)$ and $\alpha^* = 0.1$. Let h_1 be the Benjamini and Hochberg (1995) step-up procedure, h_2 be the Sidak (1967) step-down procedure and $h(p, \alpha) = h_1(p, \alpha) \cap h_2(p, \alpha)$. Then, $h_1(p^*, \alpha^*) = \{1, 2\}$, $h_2(p^*, \alpha^*) = \{1\}$ and thus $2, 3 \notin h(p^*, \alpha^*)$. However, increasing p^* to $q = (0.034, 1, 1)$ results in $h_1(q, \alpha^*) = \emptyset$ and thus $h(q, \alpha^*) = \emptyset \neq h(p^*, \alpha^*)$.

Example 2 Let p^* and α^* be as in Example 1. Let h_1 be a step-up procedure which uses the same threshold function as the (step-down) Sidak (1967) correction, and likewise h_2 be a step-down procedure using the same threshold function as the (step-up) Benjamini and Hochberg (1995) procedure—using (Gandy and Hahn 2016, Lemma 3), it is straightforward to show that both procedures are well-behaved. Let $h(p, \alpha) = h_1(p, \alpha) \cup h_2(p, \alpha)$. Then, $h_1(p^*, \alpha^*) = \{1\}$, $h_2(p^*, \alpha^*) = \emptyset$ and thus $h(p^*, \alpha^*) = \{1\}$. However, decreasing p^* to $q = (0, 0.06, 1)$ results in $h_2(q, \alpha^*) = \{1, 2\}$ and thus $h(q, \alpha^*) = \{1, 2\} \neq h(p^*, \alpha^*)$.

Examples 1 and 2 also demonstrate that both the union and the intersection of a well-behaved step-up and a well-behaved step-down procedure are not necessarily well-behaved any more.

Although neither the class of well-behaved multiple testing procedures of general form nor the combination of a well-behaved step-up and a well-behaved step-down procedure is closed under the four set operations aforementioned, the next section proves that this holds true for the special classes of well-behaved step-up and step-down procedures individually (when taking unions and intersections).

4 Step-up and step-down procedures

Gandy and Hahn (2016) show that any step-up or step-down procedure (characterised by its threshold function τ_α) which satisfies the following condition is well-behaved:

Condition 2 *1. $\tau_\alpha(i)$ is non-decreasing in i for each fixed α.*
2. $\tau_\alpha(i)$ is continuous in α and non-decreasing in α for each fixed i.

Furthermore, Gandy and Hahn (2016) verify that a large variety of commonly used procedures satisfies Condition 2 and is hence well-behaved, among them the procedures of Bonferroni (1936), Sidak (1967), Holm (1979), Hochberg (1988) or Benjamini and Hochberg (1995).

Even though (Gandy and Hahn 2016, Lemma 3) only prove that Condition 2 is sufficient for a procedure to be well-behaved, the condition is actually also necessary:

**Lemma 2** *Any well-behaved step-up or step-down procedure satisfies Condition 2.*

Consider two step-up procedures h^u and $\tilde{h}^u$ with threshold functions τ_α^u and $\tilde{\tau}_\alpha^u$ as well as two step-down procedures h^d and $\tilde{h}^d$ with threshold functions τ_α^d and $\tilde{\tau}_\alpha^d$.

In the following subsections, we separately investigate whether the classes of step-up (step-down) procedures are closed under each of the four set operations (union, intersection, difference and complement). Moreover, we investigate whether the subclasses of well-behaved step-up (step-down) procedures are closed. To this end, by Lemma 2, it suffices to show that the classes of step-up (step-down) procedures satisfying Condition 2 are closed.

4.1 Union

The class of step-up procedures is closed under the union operation: To be precise, if h^u and $\tilde{h}^u$ are two step-up procedures, their union is computed by another step-up procedure h with threshold function $\tau_\alpha(i) = \max(\tau_\alpha^u(i), \tilde{\tau}_\alpha^u(i))$ as visualised in Fig. 1 (left).

This is seen as follows: As $\tau_\alpha^u(i), \tilde{\tau}_\alpha^u(i) \leq \tau_\alpha(i)$ for all $i \in M$, all hypotheses rejected by either h^u or $\tilde{h}^u$ are also rejected by h, that is $h^u \cup \tilde{h}^u \subseteq h$. Likewise, as $\tau_\alpha(i)$ takes precisely one of the values $\tau_\alpha^u(i)$ or $\tilde{\tau}_\alpha^u(i)$ for each $i \in M$, any p-value belonging to the non-rejection area of both procedures h^u and $\tilde{h}^u$ also stays non-rejected in h, hence $(h^u)^c \cap (\tilde{h}^u)^c \subseteq h^c$.
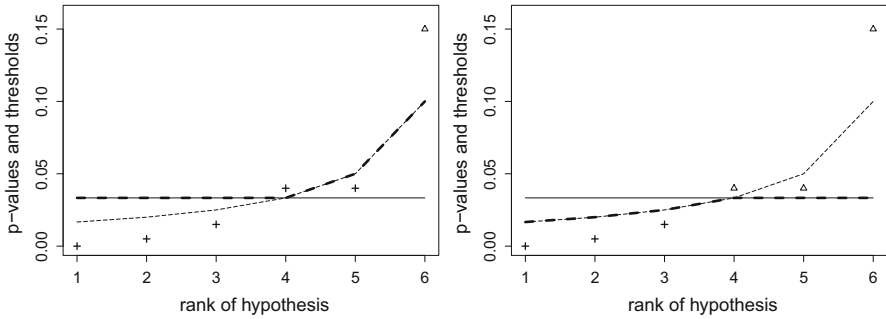
Fig. 1 Combined threshold function (*bold*) for the computation of the union (*left*) and the intersection (*right*) of the Bonferroni (1936) correction (*vertical solid line*) and the Hochberg (1988) (*dashed line*) procedure. The Bonferroni (1936) correction was applied with significance level 0.2, the Hochberg (1988) procedure with level 0.1. *p*-values of rejected (*crosses*) and non-rejected (*triangles*) hypotheses

Moreover, the subclass of well-behaved step-up procedures is also closed under the union operation as proven in the following lemma.

Lemma 3 *If h^u and $\tilde{h}^u$ are two step-up procedures which satisfy Condition* 2 *then so does the union $h^u \cup \tilde{h}^u$.*

Similarly, the union of two step-down procedures h^d and $\tilde{h}^d$ (having threshold functions τ^d_α and $\tilde{\tau}^d_\alpha$) is obtained through another step-down procedure characterised by the threshold function $\tau_\alpha(i) = \max(\tau^d_\alpha(i), \tilde{\tau}^d_\alpha(i))$. Since the proof of Lemma 3 does not use any properties of τ^u_α and $\tilde{\tau}^u_\alpha$ other than that both satisfy Condition 2, the maximum of two step-down threshold functions likewise leads to a threshold function satisfying Condition 2.

4.2 Intersection

Similarly to Sect. 4.1, the intersection of two step-up procedures h^u and $\tilde{h}^u$ is again a step-up procedure h, characterised by the new threshold function $\tau_\alpha(i) = \min(\tau^u_\alpha(i), \tilde{\tau}^u_\alpha(i))$ as visualised in Fig. 1 (right).

This is seen as follows: As $\tau^u_\alpha(i), \tilde{\tau}^u_\alpha(i) \geq \tau_\alpha(i)$ for all $i \in M$, any hypothesis non-rejected by either procedure h^u or $\tilde{h}^u$ is also non-rejected by h, that is $(h^u)^c \cup (\tilde{h}^u)^c \subseteq h^c$. Likewise, as $\tau_\alpha(i)$ takes precisely one of the values $\tau^u_\alpha(i)$ or $\tilde{\tau}^u_\alpha(i)$ for each $i \in M$, any *p*-value in the rejection area of both procedures remains rejected when tested with h, thus $h^u \cap \tilde{h}^u \subseteq h$.

Similarly to Lemma 3, the subclass of well-behaved step-up procedures is again closed under the intersection operation.

Lemma 4 *If h^u and $\tilde{h}^u$ are two step-up procedures which satisfy Condition* 2 *then so does the intersection $h^u \cap \tilde{h}^u$.*

The intersection of two step-down procedures h^d and $\tilde{h}^d$ is again obtained with another step-down procedure using the threshold function $\tau_\alpha(i) = \min(\tau^d_\alpha(i), \tilde{\tau}^d_\alpha(i))$. Analogously to Sect. 4.1, the proof of Lemma 4 does not use any properties of τ^u_α and $\tilde{\tau}^u_\alpha$ other than that both satisfy Condition 2, thus the minimum of two step-down threshold functions again leads to a threshold function satisfying Condition 2.

4.3 Complement

Whereas the complement is generally neither well-behaved nor monotonic, it can be computed for step-up and step-down procedures using the following construction.

Let α be a known constant. We re-consider the step-up procedure h^u with threshold function τ_α^u. Then, the step-down procedure $h^d(1-p)$ with threshold function $\tau_\alpha^d(i) = 1 - \tau_\alpha^u(m+1-i)$ applied to $1 - p$ (instead of p) computes the complement of $h^u(p)$, where $1 - p$ for $p \in [0, 1]^m$ is understood coordinate-wise.

The reasoning behind this is as follows: For any hypothesis with p-value $p_{(i)}$ below $\tau_\alpha^u(i)$, $1 - p_{(i)}$ (having rank $m+1-i$ in the sorted sequence of values $1 - p$) is above $\tau_\alpha^d(m+1-i)$ by construction of τ_α^d. Therefore, all former rejections of h^u turn into non-rejections of h^d and vice versa.

Likewise, the complement of a step-down procedure h^d with threshold function τ_α^d and constant α is computed by a step-up procedure h^u with threshold function $\tau_\alpha^u(i) = 1 - \tau_\alpha^d(m+1-i)$. Condition 2 is again satisfied:

Lemma 5 *Let α be a known constant. If the step-up procedure h^u with threshold function τ_α^u satisfies Condition 2, then so does its step-down complement h^d (defined with threshold function $\tau_\alpha^d(i) = 1 - \tau_\alpha^u(m+1-i)$).*

The requirement that α be a known constant is crucial since τ_α^d is not non-decreasing in α for a fixed i as required in the second part of Condition 2. However, Lemma 5 is made possible by the fact that for a given constant α (that is, if h and the threshold function seize to be a function of α), all the parts in Condition 1 (and likewise, Condition 2) which involve α can be ignored (see remark at the end of Sect. 2).

4.4 Difference

Following the notation of Sect. 3, the difference $h_1 \backslash h_2$ of two procedures h_1 and h_2 can equivalently be written as $h_1 \cap h_2^c$ using the complement of h_2. If h_2 is a step-up procedure, h_2^c turns into a step-down procedure (see Sect. 4.3).

Therefore, in case both h_1 and h_2 are step-up (step-down) procedures satisfying Condition 2, Lemma 1 yields that $h_1 \backslash h_2$ is still monotonic but not well-behaved any more. However, if h_1 is a step-down and h_2 is a step-up procedure (or vice versa), the results from Sect. 4.2 apply and yield that $h_1 \backslash h_2$ a well-behaved step-up/step-down procedure with explicit threshold function.

4.5 Example

Suppose we are interested in testing $H_{01}, \ldots, H_{0m}$ for statistical significance while ensuring FDR control at a pre-specified level 0.05, for instance using the Benjamini and Hochberg (1995) procedure. Additionally, we are interested in only selecting those $k \in \mathbb{N}$ hypotheses having the lowest p-values (assuming there are no ties), for instance due to the fact that budget constraints only allow follow-up studies for k hypotheses. We thus look to construct an intersection procedure which returns the indices of hypotheses satisfying both requirements simultaneously.

To this end, let h^1 be the Benjamini and Hochberg (1995) step-up procedure controlling the FDR at level 0.05, defined through the threshold function $\tau^1(i) = 0.05 \cdot i/m$ for $i \in \{1, \ldots, m\}$. Moreover, let h^2 be the (step-up) Bonferroni (1936) correction with constant but p-dependant threshold function $\tau_p^2(i) = p_{(k)}$ for $i \in \{1, \ldots, m\}$, where $p_{(k)}$ denotes the k'th smallest entry of vector $p = (p_1, \ldots, p_m)$. By construction, all rejected hypotheses by h^2 are precisely the ones with the k lowest p-values. Threshold functions τ_α for which $\alpha = \alpha(p)$ is a function of p are widely used in practice, for instance when using an estimate of the proportion of true null hypotheses to correct the level α (see, for instance, Example 1 in Gandy and Hahn (2016)). Both the Benjamini and Hochberg (1995) procedure h^1 and the Bonferroni (1936) correction h^2 satisfy Condition 2 and are thus well-behaved.

Following Sect. 4.2, the step-up procedure h defined through the threshold function $\tau_p(i) = \min(\tau^1(i), \tau_p^2(i)) = \min(0.05 \cdot i/m, p_{(k)})$ computes $h^1 \cap h^2$. Moreover, h is well-behaved by Lemma 4.

Consider the numerical example of 15 ordered p-values (here denoted as $\tilde{p}$) given in Sect 3.2 of Benjamini and Hochberg (1995). In agreement with Benjamini and Hochberg (1995), who test $\tilde{p}$ while controlling the FDR at level 0.05 and observe four rejections (of the first four hypotheses), h^1 applied to $\tilde{p}$ yields $h^1(\tilde{p}) = \{1, 2, 3, 4\}$. Applying the intersection procedure h constructed above with $k = 3$ to $\tilde{p}$ yields $h(\tilde{p}) = \{1, 2, 3\}$, that is h indeed yields those $k = 3$ hypotheses having the lowest p-values which are also significant under FDR control at level 0.05.

5 Discussion

This article investigates closure properties of general multiple testing procedures, step-up and step-down procedures as well as subclasses of (solely) monotonic and well-behaved procedures under four set operations (union, intersection, complement and difference).

The article shows that for general multiple testing procedures, solely the class of monotonic procedures is closed under taking the union and intersection. However, the subclass of well-behaved step-up (step-down) procedures is closed under taking the union and intersection.

The implications of the closure properties proven in this article are threefold: They provide a tool to construct new procedures of known form and with known properties, they render theoretical results (Gordon 2007; Gordon and Salzman 2008) instantly applicable to a large class of multiple testing procedures and they allow to combine the benefits of various multiple testing procedures in practice.

6 Appendix: Proofs

The appendix contains all proofs sorted by section.

6.1 Proofs of Section 3

Proof of Lemma 1 We prove both assertions.

1. Monotonicity. If $p \leq q$ and $\alpha \leq \alpha'$ then $h_1(q, \alpha) \subseteq h_1(p, \alpha')$, $h_2(q, \alpha) \subseteq h_2(p, \alpha')$ and thus $h_1(q, \alpha) \cup h_2(q, \alpha) \subseteq h_1(p, \alpha') \cup h_2(p, \alpha')$ as well as $h_1(q, \alpha) \cap h_2(q, \alpha) \subseteq h_1(p, \alpha') \cap h_2(p, \alpha')$.
   The second statement of Condition 1. As h_1 satisfies Condition 1, there exists δ_1 such that $\max(\|p - p^*\|, |\alpha - \alpha^*|) < \delta_1$ implies $h_1(p, \alpha) = h_1(p^*, \alpha^*)$. Likewise for h_2 with a suitable δ_2. For $\delta = \min(\delta_1, \delta_2)$ and $\max(\|p - p^*\|, |\alpha - \alpha^*|) < \delta$, we have $h_1(p, \alpha) = h_1(p^*, \alpha^*)$ and $h_2(p, \alpha) = h_2(p^*, \alpha^*)$ and thus $h_1 \cup h_2(p, \alpha) = h_1 \cup h_2(p^*, \alpha^*)$. Likewise for the intersection.
2. Fix α. If $q \leq p$ then $h_i(p, \alpha) \subseteq h_i(q, \alpha)$, but $h_i(p, \alpha)^c \supseteq h_i(q, \alpha)^c$ for $i \in \{1, 2\}$. The complement is thus not monotonic. The operation $h_1(p, \alpha) \backslash h_2(p, \alpha)$ is equivalent to $h_1(p, \alpha) \cap (h_2(p, \alpha))^c$ and thus also not monotonic.

6.2 Proofs of Section 4

Proof of Lemma 2 Let h be a step-up (step-down) procedure characterised through its threshold function τ_α. We now verify Condition 2.

1. We show that $\tau_\alpha(i)$ must be non-decreasing in i for a fixed α. Indeed, suppose τ_α is decreasing for some i. Then, h cannot be monotonic for all inputs: Assume that $m = 2$, $p = (0.5, 0.5)$ and h is of step-up type with $\tau_\alpha(1) = 1$ and $\tau_\alpha(2) = 0$. Then, $h(p) = \{1\}$ but increasing p to $q = (1, 0.5)$ results in $h(q) = \{2\} \nsubseteq h(p)$, thus contradicting monotonicity.
2. We show that $\tau_\alpha(i)$ must also be non-decreasing in α for any fixed i. Indeed, for a fixed i, suppose $\tau_\alpha(i) > \tau_{\alpha'}(i)$ for $\alpha < \alpha'$. Then, h can again not be monotonic for all inputs: Assume we test $m = 1$ hypothesis H_{01} with p-value $p = \tau_\alpha(1) > \tau_{\alpha'}(1)$. Then, H_{01} is rejected at $\tau_\alpha(1)$ but non-rejected at $\tau_{\alpha'}(1)$ even though $\alpha < \alpha'$, thus contradicting monotonicity.
3. We show that $\tau_\alpha(i)$ is continuous in α for a fixed i. Let $\epsilon > 0$ be given. Fix i and α^*. We show continuity of the threshold function at α^* as $\alpha \to \alpha^*$.
   Case 1: $\alpha^* > \alpha$. Then, $\tau_{\alpha^*}(i) \geq \tau_\alpha(i)$ by monotonicity. Define $p^* = (0, \ldots, 0, p_i^*, 1, \ldots, 1)$ for any $p_i^* \in [0, \tau_{\alpha^*}(i))$ (i.e., p^* contains p_i^* as ith entry, zeros before and ones after). Since h is well-behaved it satisfies the second part of Condition 1, hence for the fixed p^* and α^* there exists $\delta > 0$ such that for all α and p satisfying $|\alpha - \alpha^*| < \delta$, $\|p - p^*\| < \delta$ we have $h(p, \alpha) = h(p^*, \alpha^*)$. Assume $|\alpha - \alpha^*| < \delta$. Define $p = (0, \ldots, 0, p_i^* - \gamma, 1, \ldots, 1)$ for any $0 < \gamma < \min(\delta, \epsilon)$. Since $|\alpha - \alpha^*| < \delta$ and $\|p - p^*\| = \gamma < \delta$, $h(p, \alpha) = h(p^*, \alpha^*)$ by Condition 1: As the ith hypothesis is rejected in $h(p^*, \alpha^*)$ and hence also in $h(p, \alpha)$, it follows that $\tau_{\alpha^*}(i) \geq \tau_\alpha(i) \geq p_i = p_i^* - \gamma$. This holds true for all $p_i^* \in [0, \tau_{\alpha^*}(i))$, thus

$\tau_{\alpha^*}(i) \geq \tau_\alpha(i) \geq \tau_{\alpha^*}(i) - \gamma$ and hence $|\tau_{\alpha^*}(i) - \tau_\alpha(i)| \leq \gamma < \epsilon$.

Case 2: $\alpha^* \leq \alpha$. Then, $\tau_{\alpha^*}(i) \leq \tau_\alpha(i)$. Using $p^* = (0, \ldots, 0, p_i^*, 1, \ldots, 1)$ with $p_i^* \in (\tau_{\alpha^*}(i), 1]$ and $p = (0, \ldots, 0, p_i^* + \gamma, 1, \ldots, 1)$ with $0 < \gamma < \min(\delta, \epsilon)$, the same argument as in Case 1 yields $\tau_{\alpha^*}(i) \leq \tau_\alpha(i) < \tau_{\alpha^*}(i) + \gamma$.

Proof of Lemma 3 Let $h = h^u \cup \tilde{h}^u$ be defined through the threshold function $\tau_\alpha(i) = \max(\tau_\alpha^u(i), \tilde{\tau}_\alpha^u(i))$. First, h is monotonic by Lemma 1. We now verify Condition 2.

1. The function $\tau_\alpha(i)$ is non-decreasing in i: Suppose w.l.o.g. $\tau_\alpha(i) = \tau_\alpha^u(i)$. If $\tau_\alpha^u(i+1) \geq \tilde{\tau}_\alpha^u(i+1)$ then $\tau_\alpha(i) = \tau_\alpha^u(i) \leq \tau_\alpha^u(i+1) = \tau_\alpha(i+1)$ by definition of τ_α as the maximum of τ_α^u and $\tilde{\tau}_\alpha^u$. If $\tau_\alpha^u(i+1) < \tilde{\tau}_\alpha^u(i+1)$ then $\tau_\alpha(i) = \tau_\alpha^u(i) \leq \tau_\alpha^u(i+1) < \tilde{\tau}_\alpha^u(i+1) = \tau_\alpha(i+1)$.
2. τ_α is continuous in α as the maximum of two continuous functions (in this case in α) is continuous. The function τ_α is also non-decreasing in α: Indeed, fix i, let $\alpha \leq \alpha'$ and suppose w.l.o.g. $\tau_\alpha(i) = \tau_\alpha^u(i)$. If $\tau_{\alpha'}^u(i) \leq \tilde{\tau}_{\alpha'}^u(i)$ then $\tau_\alpha(i) = \tau_\alpha^u(i) \leq \tau_{\alpha'}^u(i) \leq \tilde{\tau}_{\alpha'}^u(i) = \tau_{\alpha'}(i)$ by definition of τ_α as the maximum of τ_α^u and $\tilde{\tau}_\alpha^u$. Otherwise, $\tau_\alpha(i) = \tau_\alpha^u(i) \leq \tau_{\alpha'}^u(i) = \tau_{\alpha'}(i)$.

Proof of Lemma 4 Let $h = h^u \cap \tilde{h}^u$ be defined through the threshold function $\tau_\alpha(i) = \min(\tau_\alpha^u(i), \tilde{\tau}_\alpha^u(i))$. Again, h is monotonic by Lemma 1. We now verify Condition 2.

1. The function $\tau_\alpha(i)$ is non-decreasing in i: Suppose w.l.o.g. $\tau_\alpha(i) = \tau_\alpha^u(i)$. If $\tau_\alpha^u(i+1) \geq \tilde{\tau}_\alpha^u(i+1)$ then $\tau_\alpha(i) = \tau_\alpha^u(i) \leq \tilde{\tau}_\alpha^u(i) \leq \tilde{\tau}_\alpha^u(i+1) = \tau_\alpha(i+1)$ by definition of τ_α as the minimum of τ_α^u and $\tilde{\tau}_\alpha^u$. If $\tau_\alpha^u(i+1) < \tilde{\tau}_\alpha^u(i+1)$ then $\tau_\alpha(i) = \tau_\alpha^u(i) \leq \tau_\alpha^u(i+1) = \tau_\alpha(i+1)$.
2. τ_α is continuous in α as the minimum of two continuous functions (in this case in α) is continuous. The function τ_α is also non-decreasing in α: Indeed, fix i, let $\alpha \leq \alpha'$ and suppose w.l.o.g. $\tau_\alpha(i) = \tau_\alpha^u(i)$. If $\tau_{\alpha'}^u(i) \leq \tilde{\tau}_{\alpha'}^u(i)$ then $\tau_\alpha(i) = \tau_\alpha^u(i) \leq \tau_{\alpha'}^u(i) = \tau_{\alpha'}(i)$. Otherwise, $\tau_\alpha(i) = \tau_\alpha^u(i) \leq \tilde{\tau}_\alpha^u(i) \leq \tilde{\tau}_{\alpha'}^u(i) = \tau_{\alpha'}(i)$ (by definition of τ_α as the minimum).

Proof of Lemma 5 Since $\tau_\alpha^u(i)$ is non-decreasing in i, it is immediate to verify that $\tau_\alpha^d(i)$ is also non-decreasing in i. For a given constant α, the second part of Condition 2 can be ignored as shown in (Gandy and Hahn 2014, Condition 1) and is hence automatically satisfied (see Sect. 2).

References

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

Besag, J., Clifford, P.: Sequential Monte Carlo p-values. Biometrika **78**(2), 301–304 (1991)

Bonferroni, C.: Teoria statistica delle classi e calcolo delle probabilità. Pubbl. del R Ist. Super. di Sci. Econ. e Commer. di Firenze **8**, 3–62 (1936)

Dmitrienko, A., Tamhane, A.: General theory of mixture procedures for gatekeeping. Biom. J. **55**(3), 402–419 (2013)

Finner, H., Strassburger, K.: The partitioning principle: a powerful tool in multiple decision theory. Ann. Stat. **30**(4), 1194–1213 (2002)

Gabriel, K.: Simultaneous test procedures—some theory of multiple comparisons. Ann. Math. Stat. **40**(1), 224–250 (1969)

Gandy, A., Hahn, G.: MMCTest—a safe algorithm for implementing multiple Monte Carlo tests. Scand. J. Stat. **41**(4), 1083–1101 (2014)

Gandy, A., Hahn, G.: A framework for Monte Carlo based multiple testing. Scand. J. Stat. **43**(4), 1046–1063 (2016)

Gordon, A.: Unimprovability of the Bonferroni procedure in the class of general step-up multiple testing procedures. Stat. Probab. Lett. **77**(2), 117–122 (2007)

Gordon, A., Salzman, P.: Optimality of the Holm procedure among general step-down multiple testing procedures. Stat. Probab. Lett. **78**(13), 1878–1884 (2008)

Guo, W., Peddada, S.: Adaptive choice of the number of bootstrap samples in large scale multiple testing. Stat. Appl. Genet. Mol. Biol. **7**(1), 1–16 (2008)

Hochberg, Y.: A sharper Bonferroni procedure for multiple tests of significance. Biometrika **75**(4), 800–802 (1988)

Hochberg, Y., Tamhane, A.: Multiple Comparison Procedures. Wiley, New York (2008)

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

Hommel, G., Bretz, F.: Aesthetics and power considerations in multiple testing—A contradiction? Biom. J. **50**(5), 657–666 (2008)

Hommel, G., Bretz, F., Maurer, W.: Powerful short-cuts for multiple testing procedures with special reference to gatekeeping strategies. Stat. Med. **26**(22), 4063–4073 (2007)

Hsu, J.: Multiple Comparisons: Theory and Methods. Chapman and Hall/CRC, Boca Raton (1996)

Lin, D.: An efficient Monte Carlo approach to assessing statistical significance in genomic studies. Bioinformatics **21**(6), 781–787 (2005)

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

Romano, J., Shaikh, A.: Stepup procedures for control of generalizations of the familywise error rate. Ann. Stat. **34**(4), 1850–1873 (2006)

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

Roth, A.: Multiple comparison procedures for discrete test statistics. J. Stat. Plan. Inference **82**(1–2), 101–117 (1999)

Sandve, G., Ferkingstad, E., Nygard, S.: Sequential Monte Carlo multiple testing. Bioinformatics **27**(23), 3235–3241 (2011)

Sidak, Z.: Rectangular confidence regions for the means of multivariate normal distributions. J. Am. Stat. Assoc. **62**(318), 626–633 (1967)

Tamhane, A., Liu, L.: On weighted Hochberg procedures. Biometrika **95**(2), 279–294 (2008)

van Wieringen, W., van de Wiel, M., van der Vaart, A.: A test for partial differential expression. J. Am. Stat. Assoc. **103**(483), 1039–1049 (2008)

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