A Framework for Monte Carlo based Multiple Testing

AXEL GANDY and GEORG HAHN

Department of Mathematics, Imperial College London

Abstract

We are concerned with a situation in which we would like to test multiple hypotheses with tests whose p-values cannot be computed explicitly but can be approximated using Monte Carlo simulation. This scenario occurs widely in practice. We are interested in obtaining the same rejections and non-rejections as the ones obtained if the p-values for all hypotheses had been available. The present article introduces a framework for this scenario by providing a generic algorithm for a general multiple testing procedure. We establish conditions which guarantee that the rejections and non-rejections obtained through Monte Carlo simulations are identical to the ones obtained with the p-values. Our framework is applicable to a general class of step-up and step-down procedures which includes many established multiple testing corrections such as the ones of Bonferroni, Holm, Sidak, Hochberg or Benjamini-Hochberg. Moreover, we show how to use our framework to improve algorithms available in the literature in such a way as to yield theoretical guarantees on their results. These modifications can easily be implemented in practice and lead to a particular way of reporting multiple testing results as three sets together with an error bound on their correctness, demonstrated exemplarily using a real biological dataset.

Keywords: algorithm, framework, hypothesis testing, monte carlo, multiple testing procedure, p-value

1 Introduction

We would like to test m hypotheses $H_{01}, \ldots, H_{0m}$ for statistical significance using a multiple testing procedure given by a mapping

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

which takes a vector of m p-values $p \in [0,1]^m$ and a threshold $\alpha \in [0,1]$ and returns the set of indices of hypotheses to be rejected, where $\mathcal{P}$ denotes the power set. As we

consider control of both the familywise error or the false discovery rate, this procedure could, for instance, be the Bonferroni (1936) correction, the Sidak (1967) correction or the procedures of Holm (1979), Hochberg (1988) or Benjamini & Hochberg (1995).

We assume that the p-values $p^* = (p_1^*, \ldots, p_m^*)$ of the underlying tests cannot be computed explicitly. Moreover, the threshold α^* at which we would like to test may depend on the unknown p^* and thus may be unknown itself. Our aim is to compute $h(p^*, \alpha^*)$.

We assume that we can approximate p^* and α^* through Monte Carlo simulations from a well defined null distribution. Such Monte Carlo simulations are assumed to be carried out separately for each hypothesis. We define *Monte Carlo based multiple testing* to be the evaluation of multiple hypotheses using a multiple testing correction applied to p^* and α^* approximated via Monte Carlo simulations. Assuming a test statistic is available to test all hypotheses, the Monte Carlo simulations considered in the present article correspond to the simulation of independent datasets under the null and to the evaluation of the test statistics on the simulated data. A Monte Carlo p-value can then be computed as a proportion of simulated test statistics exceeding the test statistic evaluated on the observed data (Sandve et al., 2011).

Monte Carlo simulations are commonly carried out by resampling the data in case of bootstrap tests or by generating permutations when using permutation tests. This scenario occurs widely in practical situations (Chen et al., 2013; Nusinow et al., 2012; Rahmatallah et al., 2012).

More precisely, we assume that we can compute sequences of nested confidence intervals $I_1^i \supseteq I_2^i \supseteq \cdots$, $i = 1, \ldots, m$, for $p_1^*, \ldots, p_m^*$ and $I_1^{m+1} \supseteq I_2^{m+1} \supseteq \cdots$ for α^*. For some of our results, we require that these intervals $(I_n^i)_{i,n}$ have a positive joint coverage probability and that their lengths uniformly go to zero.

Being able to compute $h(p^*, \alpha^*)$ through Monte Carlo simulation ensures the repeatability and objectivity of the test result (Gandy & Hahn, 2014). Moreover, all guarantees of h, in particular the type I and type II error control, are valid (up to an error probability) even when using approximations of p^* and α^*.

Existing methods (Besag & Clifford, 1991; Lin, 2005; van Wieringen et al., 2008; Guo & Peddada, 2008; Sandve et al., 2011) return a set of rejected hypotheses which, at least to a certain extent, is random, where the randomness is coming from the Monte Carlo simulations and not from the underlying data. Consider the following example which will be revisited in Section 5. Sandve et al. (2011) use their method `MCFDR` to classify a genome dataset of Pekowska et al. (2010) with the aim to test if a gene modification appears more often in certain gene regions. Each gene region corresponds to one hypothesis. Sandve et al. (2011) report 2747 significant hypotheses out of 3466 hypotheses without providing guarantees on the stability of their finding. Recomputing

the decisions on all hypotheses shows considerable variability: around 353 of the 3466 hypotheses are randomly classified in the sense that they switch from being rejected to non-rejected in more than 1% of all cases when repeatedly applying `MCFDR`. Conclusions based on the significances of these genes should therefore be questioned.

In the present article we will show how algorithms such as `MCFDR` can be modified to give a guarantee on the stability of their findings and thus how to reduce the Monte Carlo randomness in their results. These guarantees are conditional on the data on which the testing is carried out. Although in principle, increased stability can be achieved by augmenting the number of Monte Carlo samples, merely increasing the number of samples does not provide a guarantee on the decision of each hypothesis. The guarantees provided in the present article are effective for any number of samples.

The contribution of the article is threefold. First, Section 2 provides a framework for multiple hypothesis testing under the assumption that p-values are not available and thus have to be approximated using Monte Carlo methods. The framework is phrased as a generic algorithm which, under conditions, computes sub- and supersets of $h(p^*, \alpha^*)$ that converge to $h(p^*, \alpha^*)$ (Lemma 1 and Theorem 1 in Section 2.3). Theoretical bounds on the control of the false discovery rate (fdr) are provided for the proposed algorithm (Lemma 2 in Section 2.3) given it is used in connection with a suitable procedure h controlling the fdr. The framework also incorporates multiple testing at a (possibly unknown) corrected testing threshold, for instance using an estimate of the proportion of true null hypotheses.

Second, we show how to use the framework to modify established algorithms in such a way as to provide certain proven guarantees on their test results (Section 3).

Third, we simplify the condition on the multiple testing procedure in Section 4, yielding an easy-to-check criterion for an arbitrary step-up or step-down procedure (Section 4.1). We then use the simplified criterion to show that many widely used procedures can be employed in our framework (Section 4.2).

One specific implementation of our generic algorithm is the `MMCTest` algorithm of Gandy & Hahn (2014). `MMCTest` uses similar results to the ones in this article to prove the correctness of its test result up to a pre-specified error probability. In contrast to the present article which presents results for a generic algorithm and a generic multiple testing procedure, `MMCTest` focuses on one specific implementation only as well as on the two specific multiple testing procedures of Bonferroni (1936) and Benjamini & Hochberg (1995). Gandy & Hahn (2014) do not prove that a correct test result can also be obtained through appropiate modifications of existing methods. Moreover, hypothesis testing at a variable testing threshold is not possible with `MMCTest`, theoretical bounds on the control of the false discovery rate are not given, and Gandy & Hahn (2014) do not provide a simple criterion to prove whether an arbitrary step-up

or step-down procedure allows one to classify hypotheses without knowledge of the p-values.

In Section 5 we pick up our discussion of the biological dataset of Pekowska et al. (2010). We show that our proposed modifications can easily be implemented in practice, come at virtually no additional computational cost and lead to a certain way of reporting multiple testing results as three sets together with an error bound on their correctness.

The article concludes with a discussion in Section 6. All proofs are included in the Appendix.

Throughout the article, let $|\cdot|$ denote the length of an interval or the size of a set. Let $\|\cdot\|$ denote the Euclidean norm. For an interval $I \subset \mathbb{R}$, let $\min I$ and $\max I$ denote its lower and upper limit, respectively. For any set $S \subseteq \{1, \ldots, m\}$, where $m \in \mathbb{N}$, let S^c denote the complement of S with respect to $\{1, \ldots, m\}$. We abbreviate $(x_1, \ldots, x_n)$ by $x_{1:n}$, where $x_{1:0} = \emptyset$.

2 The framework

Our framework includes two components, the multiple testing procedure h and a generic algorithm presented in Section 2.1. Section 2.2 discusses a concrete implementation of our generic algorithm with the aim to exemplarily improve an established method. Combining both the testing procedure and the algorithm yields a framework which, under conditions, guarantees the correctness of its test result (Section 2.3).

2.1 The generic algorithm

We propose to use the following generic sequential algorithm to draw samples for each hypothesis. As the p-values are unknown, in each iteration n, the generic algorithm computes intervals I_n^i for each p_i^*, $i \in \{1, \ldots, m\}$, as well as an interval I_n^{m+1} for α^*. Usually, these intervals will be confidence intervals, in which case our algorithm will compute sub- and supersets of $h(p^*, \alpha^*)$ (Section 2.3).

Although the threshold will be, in most cases, a function $\alpha^* = g(p^*)$ of the p-values p^* (see Section 3), it is sensible to not restrict the multiple testing procedure to $h(p^*) = h(p^*, g(p^*))$ and to keep a separate interval I_n^{m+1} for α^* instead: Naturally, one could use confidence bounds on p^* to obtain a plug-in interval for α^* (provided that g is monotonic). However, Example 2 demonstrates that for the testing threshold of Pounds & Cheng (2006), Hoeffding's inequality (Hoeffding, 1963) allows one to construct a tighter confidence interval for α^* than the plug-in interval, thus yielding a faster convergence to $h(p^*, \alpha^*)$ as well as considerably more decisions on individual

hypotheses in a real-data study (Section 5).

The generic algorithm draws Monte Carlo samples in each iteration n, denoted by the observations O_n. These are typically sets of samples drawn for all hypotheses or for a subset of the hypotheses. The decision for which hypotheses to sample new observations may depend on the history of observations drawn up to iteration $n-1$. For instance, in Section 2.2 which considers the evaluation of multiple tests based on a test statistic, each observation O_n is a vector of indicators signaling if the test statistic evaluated on the n^{th} sample drawn for each of the m hypothesis exceeds the observed test statistic.

Each (confidence) interval I_n^i is computed by a function F_i, $i \in \{1,\dots,m+1\}$, using the current history of observations $O_{1:n}$, $n \in \mathbb{N}$. For generality, we do not impose that F_i computes any specific type of confidence interval. Intersecting the intervals in Algorithm 1 produces a nested sequence of $(I_n^i)_{n\in\mathbb{N}}$ for each $i \in \{1,\dots,m+1\}$.

Algorithm 1 (Generic algorithm)**.**

$$
\begin{aligned}
&\underline{A}_0 = \emptyset, \overline{A}_0 = \{1,\dots,m\}, I_0^i = [0,1], i \in \{1,\dots,m\}, I_0^{m+1} = \mathbb{R}.\\
&\textit{For } n = 1,2,\dots: \quad \textit{Choose which } O_n \textit{ to sample based on } O_{1:n-1},\\
&\qquad\qquad\qquad\quad \textit{Sample } O_n,\\
&\qquad\qquad\qquad\quad I_n^i = F_i(O_{1:n}) \cap I_{n-1}^i, i \in \{1,\dots,m+1\},\\
&\qquad\qquad\qquad\quad \overline{A}_n = h((\min I_n^i)_{i\in\{1,\dots,m\}}, \max I_n^{m+1}),\\
&\qquad\qquad\qquad\quad \underline{A}_n = h((\max I_n^i)_{i\in\{1,\dots,m\}}, \min I_n^{m+1}).
\end{aligned}
$$

In each iteration n, Algorithm 1 uses the history of samples observed up to iteration $n-1$ to determine a new set of observations O_n to be sampled. The key idea of Algorithm 1 is to apply the multiple testing procedure h to lower ($\min I_n^i$) and upper ($\max I_n^i$) confidence limits of the $(I_n^i)_{n\in\mathbb{N}}$, $i \in \{1,\dots,m\}$. This yields two sets $\overline{A}_n$ and $\underline{A}_n$.

2.2 The improved naïve method

A widely used method in practice to estimate $h(p^*, \alpha^*)$ is to draw a constant number of samples s for each hypothesis H_{0i}, where $i \in \{1,\dots,m\}$, then compute a point estimate of each p-value and classify all hypotheses at a constant threshold α^* based on these point estimates (Nusinow et al., 2012; Gusenleitner et al., 2012; Rahmatallah et al., 2012; Zhou et al., 2013; Li et al., 2012; Cohen et al., 2012). We will call this the naïve method. The naïve method can be applied to any multiple testing procedure h.

In the following we present an improvement of the naïve method by stating a concrete implementation of Algorithm 1. As shown in Section 3, under conditions on h, the sets $\underline{A}_n$ ($\overline{A}_n$) defined in Algorithm 1 will be subsets (supersets) of $h(p^*, \alpha^*)$ in each iteration n of our improved naïve method up to a pre-specified error probability.

One key ingredient of the improved naïve method are the confidence sequences given in Lai (1976): for independent $Y_1, Y_2, \ldots \sim \text{Bernoulli}(p)$,

$$\mathbb{P}(g_n^\beta(S_n) < p < f_n^\beta(S_n)\ \forall n \geq 1) \geq 1 - \beta,$$

where $S_n = \sum_{i=1}^n Y_i$ and $g_n^\beta(x) < f_n^\beta(x)$ are the two distinct (Lai, 1976) roots of $(n+1)\binom{n}{x}p^x(1-p)^{n-x} = \beta$ for a given $\beta \in (0,1)$.

In many applications of the naïve method, multiple tests are based on a test statistic and it is possible to sample under the null hypothesis. Let $X_n^i = 1$ if the test statistic evaluated on the n^{th} sample drawn for hypothesis H_{0i} exceeds the observed test statistic, otherwise $X_n^i = 0$. For our improved naïve method, we draw one new sample per hypothesis in each iteration n.

The improved naïve method is obtained by defining

$$\begin{aligned} O_n &= (X_n^1, \ldots, X_n^m), \\ F_i(O) &= \left[g_{|O|}^\beta \left(\sum_{j=1}^{|O|} O_j^i \right), f_{|O|}^\beta \left(\sum_{j=1}^{|O|} O_j^i \right) \right],\ i = 1, \ldots, m, \\ F_{m+1}(O) &= \{\alpha^*\}, \end{aligned}$$

where $O_j^i = X_j^i$ and $|O_{1:n}| = n$.

Although the above method is open-ended, we usually stop the improved naïve method after a pre-specified total number of iterations s. In this case, solely the two test results in $\underline{A}_s$ and $\overline{A}_s$ based on the intervals of the last iteration will be returned as result of the algorithm.

In the improved naïve method, the testing threshold α^* is assumed to be constant. However, the interval $F_{m+1}(O_{1:n})$ for α^* is needed if α^* depends on p^*. For instance, this is the case for thresholds depending on an estimate of the proportion of true null hypotheses which is usually a functional of p^*. Using such an estimated threshold potentially results in more significant hypotheses which is desired in practice.

Starting with the work of Schweder & Spjøtvoll (1982), many authors have investigated estimators of the proportion of true null hypotheses, such as Storey (2002), Langaas et al. (2005), Pounds & Cheng (2006), Finner & Gontscharuk (2009) and Friguet & Causeur (2011).

2.3 Convergence results

This section states our main results for which we need the following monotonicity property:

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

A multiple testing procedure is thus monotonic if smaller p-values (as introduced in Tamhane & Liu, 2008) or a higher testing threshold (see Roth, 1999) lead to more rejections. The estimators of the proportion of true null hypotheses listed in the last paragraph of Section 2.1 all depend on the p-values p^* only (and some tuning parameters) and are monotonically increasing in p^*. When combined with a multiple testing procedure they thus preserve the monotonicity in the threshold argument.

Suppose in each iteration $n \in \mathbb{N}$, each p-value p_i^* is contained in its interval $F_i(O_{1:n})$, $i \in \{1, \ldots, m\}$, and the testing threshold α^* is contained in the interval $F_{m+1}(O_{1:n})$, expressed as the event

$$R_1 = \{\alpha^* \in F_{m+1}(O_{1:n}), p_i^* \in F_i(O_{1:n}) \; \forall i \in \{1, \ldots, m\}, n \in \mathbb{N}\}\,.$$

The following lemma shows that on the event R_1, classifying hypotheses based on upper and lower interval bounds allows Algorithm 1 to compute sub- and supersets of $h(p^*, \alpha^*)$ for monotonic multiple testing procedures h.

Lemma 1. *Let h be a monotonic multiple testing procedure. Then,*

1. *$\underline{A}_n \nearrow$ and $\overline{A}_n \searrow$ as $n \to \infty$,*
2. *$\underline{A}_n \subseteq h(p^*, \alpha^*) \subseteq \overline{A}_n \; \forall n \in \mathbb{N}$ on the event R_1.*

The first part of Lemma 1 is not dependent on the event R_1. It follows purely from the construction of Algorithm 1 which computes nested intervals for each p_i^*, $i \in \{1, \ldots, m\}$. The second part of Lemma 1 shows that on R_1, in any iteration n, all the hypotheses in the set $\underline{A}_n$ ($\overline{A}_n^c$) can already be classified as being rejected (non-rejected).

Additional properties of Algorithm 1 can be derived for any monotonic multiple testing procedure h and choice of p^*, α^* which satisfy 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. *There exists $\delta > 0$ such that $p \in [0,1]^m$, $\alpha \in [0,1]$ and $\|p - p^*\| \vee |\alpha - \alpha^*| < \delta$ imply $h(p, \alpha) = h(p^*, \alpha^*)$.*

Condition 1 ensures that lowering (increasing) the p-value of any rejected (non-rejected) hypothesis does not affect the result of h. Moreover, we require that there exists a neighborhood of p^* and α^* on which h is constant. In Section 4 we will simplify Condition 1 for so-called step-up and step-down procedures.

We will call a monotonic multiple testing procedure h *well-behaved* for p^* and α^* if it satisfies Condition 1. The multiple testing procedures we consider in this article, such as the common procedures of Bonferroni (1936), Sidak (1967), Holm (1979), Hochberg (1988) or the one of Benjamini & Hochberg (1995) are well-behaved for all but a null set of p^* and α^* (with respect to the Lebesgue measure). A (non-exhaustive) list of well-behaved procedures can be found in Section 4.2.

A second condition is necessary to obtain convergence of the two bounds $\underline{A}_n$ and $\overline{A}_n$ established in Lemma 1 to $h(p^*, \alpha^*)$ as $n \to \infty$. Whereas on the event R_1, all hypotheses in $\underline{A}_n$ ($\overline{A}_n^c$) can already be rejected (non-rejected), we additionally require that the length of each interval belonging to a yet undecided hypothesis in the set $\overline{A}_n \setminus \underline{A}_n$ or to the threshold goes to zero:

$$R_2 = \left\{\max\{|F_i(O_{1:n})| : i \in \overline{A}_n \setminus \underline{A}_n \cup \{m+1\}\} \to 0 \text{ as } n \to \infty\right\}.$$

The following theorem improves upon Lemma 1 on the more restrictive event $R = R_1 \cap R_2$:

Theorem 1. *Let h be a well-behaved multiple testing procedure for p^* and α^*. On the event R, both sequences $(\underline{A}_n)_{n\in\mathbb{N}}$ and $(\overline{A}_n)_{n\in\mathbb{N}}$ converge to $h(p^*, \alpha^*)$, i.e. there exists $n_0 \in \mathbb{N}$ such that $\underline{A}_n = h(p^*, \alpha^*) = \overline{A}_n \ \forall n \geq n_0$.*

In the next section, we will use Lemma 1 and Theorem 1 to establish guarantees on the test result of existing algorithms.

Suppose Algorithm 1 is used in connection with a well-behaved multiple testing procedure controlling the familywise error rate (fwer). Then at any stage, the fwer is also controlled for all the rejections in $\underline{A}_n \subseteq h(p^*, \alpha^*)$. This is easily proven using Boole's inequality.

A similar statement, however, is not true for well-behaved multiple testing procedures controlling the false discovery rate (fdr). Although the fdr is not generally controlled for subsets $\underline{A}_n \subseteq h(p^*, \alpha^*)$ or supersets $\overline{A}_n \supseteq h(p^*, \alpha^*)$, the following guarantees hold if Algorithm 1 is run with suitable stopping times.

Lemma 2. *Let h control the fdr at level α, let $\underline{V}_n$ ($\overline{V}_n$) be the set of rejected true null hypotheses in $\underline{A}_n$ ($\overline{A}_n$) for $n \in \mathbb{N}$ and let $\eta \geq 1$, $\xi \geq 0$.*

1. $\mathbb{E}\left(|\underline{V}_s|/|\underline{A}_s|\right) \leq \eta\alpha$ *for the stopping time* $s = \min\{n \in \mathbb{N} : |\overline{A}_n|/|\underline{A}_n| \leq \eta\}$.
2. $\mathbb{E}\left(|\overline{V}_t|/|\overline{A}_t|\right) \leq \alpha + \xi$ *for* $t = \min\{n \in \mathbb{N} : (|\overline{A}_n| - |\underline{A}_n|)/|\overline{A}_n| \leq \xi\}$.

In Lemma 2, we define the fraction in the definition of the stopping time s (time t) to be zero if $|\underline{A}_n|$ ($|\overline{A}_n|$) is zero as in this case, false rejection errors are impossible.

Lemma 2 thus provides two different guarantees on the fdr, a multiplicative one on the set of rejected hypotheses $\underline{A}_s$ and an additive guarantee on the rejections in $\overline{A}_t$ with respect to the two stopping times s and t.

3 Improving existing algorithms

In this section, we introduce a class of established methods which estimate $h(p^*, \alpha^*)$ and show how the framework given by Algorithm 1 can be used to modify these methods in such a way as to provide a guarantee on the correctness of their test results. We will demonstrate our proposed modifications by extending the improved naïve method presented in Section 2.2 to the situation of an estimated testing threshold.

Consider an existing method to compute $h(p^*, \alpha^*)$. The threshold α^* can either be constant or given by a monotonic (increasing or decreasing) function $g : [0,1]^m \to \mathbb{R}$, thus $\alpha^* = g(p^*)$. In the latter case, α^* is a function of p^* and thus unknown itself.

Methods working with bootstrap point estimates of p^* (Besag & Clifford, 1991; van Wieringen et al., 2008; Sandve et al., 2011; Jiang & Salzman, 2012), fitted distributions (Knijnenburg et al., 2009) or permutation based methods (Westfall & Young, 1993; Westfall & Troendle, 2008; Meinshausen, 2006) can be phrased in the following way: Draw independent samples $X_{ij} \sim \text{Bernoulli}(p_i^*)$, $j \in \mathbb{N}$, for each $i \in \{1, \ldots, m\}$. Use a finite number S_i of these samples $X_{i1}, \ldots, X_{i,S_i}$ to compute a p-value estimate $\hat{p}_i$ of p_i^*, where S_i is a (random) index and $i \in \{1, \ldots, m\}$. Estimate the testing threshold α^* using the plug-in estimate $\hat{\alpha} = g(\hat{p})$, where $\hat{p} = (\hat{p}_1, \ldots, \hat{p}_m)$. Return $h(\hat{p}, \hat{\alpha})$ as the test result.

Based on Algorithm 1 we propose to modify any method of the above type by

1. Maintaining a confidence sequence (Lai, 1976) with a coverage probability of $1 - \epsilon/m$ for each p-value p_i^*, $i \in \{1, \ldots, m\}$, and by using each sequence as $F_i(O_{1:n})$ in Algorithm 1. The overall error probability ϵ is chosen by the user.

2. Computing plug-in bounds $F_{m+1}(O_{1:n})$ for α^* using the monotonicity of g and the above confidence sequences.

3. Reporting hypotheses in $\underline{A}_n$ as rejected and in $\overline{A}_n^c$ as non-rejected. The remaining hypotheses are still undecided.

As the confidence sequence of Lai (1976) satisfies $\mathbb{P}(\exists n : p_i^* \notin F_i(O_{1:n})) < \beta$ for each p_i^* (see Section 2.2), the choice $\beta = \epsilon/m$ yields

$$\mathbb{P}(\exists i, n : p_i^* \notin F_i(O_{1:n})) \leq \sum_{i=1}^{m} \mathbb{P}(\exists n : p_i^* \notin F_i(O_{1:n})) \leq \sum_{i=1}^{m} \epsilon/m = \epsilon,$$

and hence $\mathbb{P}(p_i^* \in F_i(O_{1:n})\ \forall i \in \{1,\dots,m\}, n \in \mathbb{N}) \geq 1-\epsilon$. The event R_1 thus occurs with probability at least $1-\epsilon$.

Consequently, any modified method of the above type has the following advantage over its unimproved counterpart:

Remark 1. *By Lemma 1, a modified method of the above type has the property that all the hypotheses in the set $\underline{A}_n$ ($\overline{A}_n^c$) which are rejected (non-rejected) in any iteration n are indeed correctly rejected (non-rejected) with probability at least $1-\epsilon$.*

Remark 1 applies to the improved naïve method (Section 2.2) upon stopping in iteration s as well as to the methods presented in the following two examples. First, we generalize Section 2.2 to the situation where the testing threshold is unknown.

Example 1. Additionally to the setting of Section 2.2, we assume that multiple testing is carried out at the corrected testing threshold $\alpha^* = t^*/\hat{\pi}_0(p^*)$, where t^* is an uncorrected threshold (typically $t^* = 0.05$ or $t^* = 0.1$) and $\hat{\pi}_0(p) = \min\left(1, \frac{2}{m}\sum_{i=1}^m p_i\right)$ is an estimator of the proportion of true null hypotheses (Pounds & Cheng, 2006). Recent applications of this threshold include Han & Dalal (2012), Lu et al. (2011), Jupiter et al. (2010), Cheng (2009). As $\hat{\pi}_0(p^*)$ depends on the p-values, the corrected threshold α^* is unknown in practice. We thus need to compute a confidence interval for it. The interval can be constructed using the monotonicity of $\hat{\pi}_0(p)$: in iteration n, $\underline{\pi}_n = \hat{\pi}_0(\min I_n^1, \dots, \min I_n^m)$ is a lower bound on $\hat{\pi}_0(p^*)$, likewise $\overline{\pi}_n = \hat{\pi}_0(\max I_n^1, \dots, \max I_n^m)$ is an upper bound. This immediately translates to the interval $F_{m+1}(O_{1:n}) = [t^*/\overline{\pi}_n, t^*/\underline{\pi}_n]$ for α^*.

We try to improve Example 1 by using a (hopefully) tighter confidence interval $F_{m+1}(O_{1:n})$ for α^* based on Hoeffding's inequality (Hoeffding, 1963).

Example 2. Suppose we have observed s samples $X_1^i, \dots, X_s^i$ per hypothesis H_{0i}, where X_j^i is the indicator of an exceedance for the jth sample drawn for H_{0i} (see Section 2.2). Then, $\mathbb{P}\left(\left|\frac{1}{ms}\sum_{i=1}^m\sum_{j=1}^s X_j^i - \frac{1}{m}\sum_{i=1}^m p_i^*\right| \geq u\right) \leq 2\exp\left(-2msu^2\right)$ for all $u > 0$ by Hoeffding's inequality. Thus for a given $\eta \in [0,1]$, $\frac{1}{ms}\sum_{i=1}^m\sum_{j=1}^s X_j^i \pm \sqrt{-\log(\eta/2)/(2ms)}$ are boundaries of a $1-\eta$ confidence interval for $\frac{1}{m}\sum_{i=1}^m p_i^*$. Using the monotonicity of the mapping $x \mapsto t^*/\min(1,2x)$, this immediately translates to a $1-\eta$ confidence interval for α^*. When using Hoeffding's interval in the improved naïve method, we allocate an error of $\eta = \epsilon/(m+1)$ to the computation of Hoeffding's interval as well as to the computation of each of the m confidence sequences for the p-values. As the improved naïve method is open-ended, we use a non-negative real sequence $(\eta_n)_{n\in\mathbb{N}}$ satisfying $\sum_{n=1}^\infty \eta_n = \eta$ to distribute η for Hoeffding's interval over all iterations of the algorithm, thus computing it at level η_n in each iteration n.

Both the plug-in interval (Example 1) and Hoeffding's confidence interval (Example 2) will be evaluated in Section 5.

4 Well-behaved step-up and step-down procedures

Although the multiple testing procedure h does not have to be of a special form, many procedures used in practice such as the ones of Bonferroni (1936), Sidak (1967), Holm (1979), Hochberg (1988) or the one of Benjamini & Hochberg (1995) belong to a certain class of procedures, called step-up and step-down procedures. We will simplify Condition 1 for step-up and step-down procedures in Section 4.1 and use the simplified condition in Section 4.2 to verify that many widely used procedures are well-behaved. As shown in Section B in the Appendix, the Hommel (1988) procedure is an example of a procedure which is not well-behaved.

4.1 Condition 1 can be simplified for step-up and step-down procedures

Suppose we are given an arbitrary step-up procedure h_u or step-down procedure h_d (Romano & Shaikh, 2006) returning the set of rejected indices. For our purposes, we phrase these two procedures in terms of a threshold function $\tau_\alpha : \{1, \dots, m\} \to [0, 1]$ which depends on a threshold $\alpha \in [0, 1]$ and returns the critical value $\tau_\alpha(i)$ each $p_{(i)}$ is compared to:

$$h_u(p, \alpha) = \left\{ i \in \{1, \dots, m\} : p_i \leq \max\{p_{(j)} : p_{(j)} \leq \tau_\alpha(j)\} \right\}, \tag{2}$$

$$h_d(p, \alpha) = \left\{ i \in \{1, \dots, m\} : p_i < \min\{p_{(j)} : p_{(j)} > \tau_\alpha(j)\} \right\}, \tag{3}$$

where $\max \emptyset := 0$, $\min \emptyset := 1$, and where the order statistic of $p_1, \dots, p_m$ is denoted by $p_{(1)} \leq \dots \leq p_{(m)}$.

We assume that the threshold function τ_α satisfies the following condition.

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.

By the following lemma, a step-up or step-down procedure is well-behaved if the threshold function τ_α defining it satisfies Condition 2.

Lemma 3. *If τ_α satisfies Condition 2 then the corresponding h_u and h_d are monotonic and satisfy the first part of Condition 1. If moreover $\tau_{\alpha^*}(i) \neq p^*_{(i)}$ for all $i \in \{1, \dots, m\}$, h_u and h_d also satisfy the second part of Condition 1 for p^* and α^*.*

We investigate in which cases the condition $\tau_{\alpha^*}(i) \neq p^*_{(i)}$ for all $i \in \{1, \dots, m\}$ in Lemma 3 is satisfied if p^* are random.

If p^* come from a discrete distribution, the p-values satisfying $\tau_{\alpha^*}(i) \neq p^*_{(i)}$ do not necessarily form a null set. For a fixed α^*, however, the p-values not satisfying the

conditions of Lemma 3 form a null set if p^* are random with a distribution that is absolutely continuous with respect to the Lebesgue measure.

We now consider the case of a threshold α^* given by a deterministic function of the p-values p^*. We show that the p-values p^* not satisfying the condition $\tau_{\alpha^*}(i) \neq p^*_{(i)}$ for all $i \in \{1, \dots, m\}$ in Lemma 3 form a null set if p^* come from an absolutely continuous distribution with respect to the Lebesgue measure, and if the Benjamini & Hochberg (1995) or Bonferroni (1936) procedure applied to the p^* dependent threshold of Pounds & Cheng (2006) is used to test the hypotheses.

The Benjamini & Hochberg (1995) procedure is characterized by the threshold function $\tau_\alpha(i) = i\alpha/m$ (see Section 4.2). The threshold of Pounds & Cheng (2006) is given by $\alpha^*(p^*) = t^* / \min\left(1, \frac{2}{m}\sum_{r=1}^m p^*_r\right)$ (see Examples 1 and 2).

Fix $i \in \{1, \dots, m\}$. We use the fact that $\sum_{r=1}^m p^*_r = \sum_{r=1}^m p^*_{(r)}$ and that either $\tau_{\alpha^*(p^*)}(i) = it^*/m$ or $\tau_{\alpha^*(p^*)}(i) = \frac{it^*}{2}(\sum_{r=1}^m p^*_r)^{-1}$. Conditional on $\{p^*_{(r)} : r \neq i\}$,

$$\tau_{\alpha^*(p^*)}(i) = p^*_{(i)} \Rightarrow \frac{it^*}{m} = p^*_{(i)} \vee \frac{it^*}{2} = p^*_{(i)}(p^*_{(i)} + s_i) \Leftrightarrow p^*_{(i)} \in \left\{\frac{it^*}{m}, \rho_1, \rho_2\right\},$$

where $s_i = \sum_{r \neq i} p^*_{(r)}$ and ρ_1, ρ_2 are the two solutions of $it^*/2 = \rho(\rho + s_i)$. Two distinct solutions always exist given $t^* > 0$.

Using that $p^*_{(1)} \leq \cdots \leq p^*_{(m)}$ are also random with a distribution that is absolutely continuous with respect to the Lebesgue measure, given p^* come from an absolutely continuous distribution, implies $\mathbb{P}(p^*_{(i)} \in \{it^*/m, \rho_1, \rho_2\} \mid p^*_{(r)} : r \neq i) = 0$.

The previous result immediately extends to $\mathbb{P}(\exists i : \tau_{\alpha^*(p^*)}(i) = p^*_{(i)}) = 0$, hence the p-values $p^*_{(i)}$ which coincide with their critical value $\tau_{\alpha^*(p^*)}(i)$ form a null set.

As the threshold function of the Bonferroni (1936) correction can be recovered from the one of the Benjamini & Hochberg (1995) procedure by removing the dependence of $\tau_\alpha(i) = i\alpha/m$ on i (see Section 4.2), the above result also holds true for the Bonferroni (1936) correction.

A similar argumentation can be used to extend the above result to other common estimators of α^* and threshold functions τ_α.

4.2 Examples of well-behaved step-up and step-down procedures

This section shows that a variety of commonly used step-up and step-down procedures are monotonic and satisfy Condition 2.

The following multiple testing procedures are determined by $\tau_\alpha(i)$, where $i \in \{1, \dots, m\}$, and control the fwer or the fdr at a threshold α. We denote the hypothesis corresponding to the ordered p-value $p_{(i)}$ by $H_{0(i)}$, $i \in \{1, \dots, m\}$.

In most cases, Condition 2 can be checked by considering the derivatives of $\tau_\alpha(i)$ with respect to α and i, thus regarding i as a continuous parameter. Unless stated otherwise, all the threshold functions listed below are clearly non-decreasing in both i and α as well as continuous in α and thus satisfy Condition 2.

The Bonferroni (1936) correction can be derived from either a step-up or a step-down procedure using the constant threshold function $\tau_\alpha(i) = \alpha/m$.

The following step-up procedures are well-behaved:

1. The Simes (1986) procedure rejects $\cap_{i\in\{1,\ldots,m\}} H_{0i}$ if there exists $k \in \{1,\ldots,m\}$ such that $p_{(k)} \leq k\alpha/m$. It can be used in our framework with the help of the following modification: Once $h_u(p,\alpha)$ for a step-up procedure with threshold function $\tau_\alpha(i) = i\alpha/m$ is correctly determined, the Simes (1986) procedure rejects $\cap_{i\in\{1,\ldots,m\}} H_{0i}$ if and only if $|h_u(p,\alpha)| > 0$.

2. The Hochberg (1988) procedure uses $\tau_\alpha(i) = \alpha/(m+1-i)$.

3. The Rom (1990) procedure increases the power of the Hochberg (1988) procedure by replacing its critical values $\tau_\alpha(i) = \alpha/(m+1-i)$ by "sharper" values $\tau_\alpha(i) = c_i$. The c_i are computed recursively as given in Rom (1990) and satisfy $c_i \nearrow$ for a fixed α. Moreover, the c_i are non-decreasing in α.

4. The choice $\tau_\alpha(i) = i\alpha/m$ yields the Benjamini & Hochberg (1995) procedure.

5. The Benjamini & Yekutieli (2001) procedure controls the fdr under arbitrary dependence by applying the Benjamini & Hochberg (1995) procedure at the corrected constant threshold $\alpha/\left(\sum_{i=1}^m i^{-1}\right)$.

Similarly, the following step-down procedures satisfy Condition 2:

1. The Sidak (1967) correction uses $\tau_\alpha(i) = 1-(1-\alpha)^{1/(m+1-i)}$.

2. The choice $\tau_\alpha(i) = \alpha/(m+1-i)$ yields the Holm (1979) procedure.

3. The Shaffer (1986) procedure modifies the Holm (1979) procedure in order to obtain an increase in power. For the tests under consideration, let $0 \leq a_1 < a_2 < \cdots < a_r \leq n$ be all possible numbers of true null hypotheses. Assuming that $H_{0(1)},\ldots,H_{0(i-1)}$ are false, let $t_i = \max\{a_j : a_j \leq n-i+1\}$ be the maximum possible number of true null hypotheses. The Shaffer (1986) procedure determines the minimal index k such that $p_{(k)} > \alpha/t_k$ and then rejects $H_{0(1)},\ldots,H_{0(k-1)}$. It can be obtained from a step-down procedure using $\tau_\alpha(i) = \alpha/t_i$, which is clearly continuous and non-decreasing in α for a fixed i. As $a_i \nearrow$ and thus $t_i \searrow$, $\tau_\alpha(i)$ is also non-decreasing in i for a fixed α.

For a given α^*, by Lemma 3, all the procedures listed above are well-behaved for all but a null set of p-values p^*.

5 Using the framework in practice

The improved naïve method (Section 2.2) is capable of computing test results which consist, up to a pre-specified error probability ϵ, of sets of correctly rejected and correctly non-rejected hypotheses as well as of a set of undecided hypotheses. The following contains an example of such a classification.

Sandve et al. (2011) use their method `MCFDR` to classify a dataset of gene modifications (so-called H3K4me2-modifications) of Pekowska et al. (2010). This dataset consists of gene regions and gene modifications within each region, characterized by their midpoint. The beginning and the end of each region on the genome are normed to 0 and 1, respectively. The authors test if the gene modifications appear more often in a certain part of the gene region.

To be precise, Sandve et al. (2011) observe k random points $Y_1, \ldots, Y_k$ in $[0,1]$ (these are the midpoints of the gene modifications) and test the null hypothesis $H_0 : \mathbb{E}\left(\frac{1}{k}\sum_{i=1}^k Y_i\right) \geq 0.5$ against the alternative $H_1 : \mathbb{E}\left(\frac{1}{k}\sum_{i=1}^k Y_i\right) < 0.5$ using the test statistic $T = \frac{1}{k}\sum_{i=1}^k Y_i$. Each null hypothesis is tested by permuting the midpoints in each region while preserving their inter-point distances.

Sandve et al. (2011) first filter the dataset for genes with at least 10 modifications per gene region. Each such region becomes one hypothesis, leading to $m = 3465$ hypotheses (gene regions) under consideration. They evaluate the data using the procedure of Benjamini & Hochberg (1995) with a corrected testing threshold at level $0.1/\hat{\pi}_0(\hat{p})$, where $\hat{\pi}_0$ is the estimator of Pounds & Cheng (2006) introduced in Example 1 and $\hat{p}$ is an estimate of p^* returned by `MCFDR`. Sandve et al. (2011) report 2747 significant hypotheses.

Nevertheless, the authors do not provide any guarantee on the correctness of their findings. Recomputing the results of Sandve et al. (2011) indeed shows considerable variability. To demonstrate this, we re-classify the H3K4me2 dataset using the `MCFDR` algorithm of Sandve et al. (2011) a total number of $r = 1000$ times. Let p_i^s (p_i^n) be the empirical probability that hypothesis H_{0i} is significant (non-significant) in these r repetitions.

We are interested in measuring the randomness in the output of an algorithm and use $p_i^r = \min(p_i^s, p_i^n)$ as probability of H_{0i} being randomly classified. We call all hypotheses having $p_i^r > 0.01$ "randomly classified" and denote their total number by rc. The choice 0.01 is arbitrary. It depends on how much uncertainty a user is willing to tolerate for a single decision on a hypothesis to be "reasonably firm". For `MCFDR`

Table 1: Repeated application of the improved and the unimproved naïve method to the same data.

	naïve method	improved naïve method							
		with plug-in interval (Ex.1)				with Hoeffding's interval (Ex.2)			
s	*rc*	rejected	non-rej.	undec.	*rc*	rejected	non-rej.	undec.	*rc*
10^2	349	0	161.8	3303.2	0	0	372.0	3093.0	0
10^3	107	2386.0	487.5	591.5	0	2568.5	576.0	320.5	0
10^4	33	2649.0	624.6	191.4	0	2697.3	661.7	106.0	0

s: number of samples drawn per hypotheses; rc: number of randomly classified hypotheses; rejected, non-rejected and undecided are average numbers based on 1000 repetitions.

we observe that 353 hypotheses remain randomly classified on average.

We first use the (unimproved) naïve method (as defined at the beginning of Section 2.2) with $s \in \{10^2, 10^3, 10^4\}$ samples per hypothesis to classify the same dataset. Table 1 shows the number of randomly classified hypotheses rc observed for the naïve method as a function of s (second column). For $s = 10^2$, the total effort is comparable to the one of `MCFDR` and both methods yield equally high numbers of random decisions ($rc \approx 350$). For high precision ($s = 10^4$), up to 33 hypotheses remain inconsistently classified.

We then apply the improved naïve method (Section 2.2) to the same dataset using an overall error probability of $\epsilon = 0.01$. The improved method is stopped after having drawn s samples per hypothesis. Table 1 shows rejected, non-rejected, undecided (see Remark 1) and randomly classified hypotheses. We evaluate both the plug-in interval for α^* introduced in Example 1 (columns three to six) as well as Hoeffding's confidence interval derived in Example 2 (columns seven to ten). For Hoeffding's interval, we use $\eta_n = \nu_n - \nu_{n-1}$ with $\nu_n = \frac{n}{n+s}\frac{\epsilon}{m+1}$, $n \in \mathbb{N}$.

Using a confidence interval for α^* based on Hoeffding's inequality (as opposed to the plug-in interval) yields considerably more decisions (rejections and non-rejections) and thus less undecided hypotheses for all ranges of precision.

Although for low numbers of samples many hypotheses remain undecided, the test results of the improved naïve method are consistent in the sense that no hypothesis is randomly classified. The improved naïve method therefore provides reliable test results and ensures repeatability. For a high precision ($s = 10^4$), the improved naïve method with Hoeffding's interval for α^* yields around 2700 rejections and 660 non-rejections. The remaining 106 hypotheses are still undecided, meaning that within this limited computational effort, no statement about these hypotheses (gene regions) should be made. The probability of the above results being correct is at least 0.99.

We interpret the set of undecided hypotheses as the set of hypotheses for which

a clear decision exists, even though this decision cannot yet be obtained within the limited computational effort used in a real testing scenario. Consequently, by using more Monte Carlo samples, the decision of any hypothesis will eventually be revealed. Alternatively, one can view the set of undecided hypotheses as a set of hypotheses whose decision is essentially arbitrary.

Finally, the framework is not limited to a strict familywise error control on all its Monte Carlo decisions. It would be possible to relax the guarantee and to control a less conservative criterion instead, for instance the false discovery rate.

6 Discussion

The present article considers p-value based multiple testing under the assumption that the p-value of each hypothesis is unknown and can only be approximated using Monte Carlo simulations. Although widely occurring in experimental studies, common methods for this scenario do not give any guarantee on how their test results relate to the one obtained if all p-values had been known.

The article introduces a framework for Monte Carlo based multiple testing, both in terms of a general multiple testing procedure and a generic algorithm. Conditions on both the multiple testing procedure and the algorithm guarantee that the rejections and non-rejections returned by our generic algorithm are identical to the ones obtained with the p-values. A simplified condition for step-up and step-down multiple testing procedures is derived.

We demonstrate how to use our framework to modify established methods in such a way as to yield theoretical guarantees on their test results. As demonstrated on a class of commonly used methods, these modifications can easily be implemented in practice and come at virtually no additional computational cost.

Improved established methods, such as the improved naïve method evaluated in this article on a real data study, allow one to report multiple testing results as three sets: rejected, non-rejected and undecided hypotheses, together with an error bound on their correctness. We recommend any multiple testing result to be reported in this fashion.

Acknowledgements

We would like to thank the editor, associate editor and the two referees for their constructive comments on the manuscript. The second author was supported by the EPSRC.

References

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

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

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

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubbl. d. R. Ist. Super. di Sci. Econom. e Commerciali di Firenze* **8**, 3–62.

Chen, J., Bushman, F., Lewis, J., Wu, G. & Li, H. (2013). Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. *Biostat.* **14**, 244–258.

Cheng, C. (2009). Internal validation inferences of significant genomic features in genome-wide screening. *Comput. Statist. Data Anal.* **53**, 788–800.

Cohen, O., Ashkenazy, H., Burstein, D. & Pupko, T. (2012). Uncovering the co-evolutionary network among prokaryotic genes. *Bioinform.* **28**, i389–i394.

Finner, H. & Gontscharuk, V. (2009). Controlling the familywise error rate with plug-in estimator for the proportion of true null hypotheses. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71**, 1031–1048.

Friguet, C. & Causeur, D. (2011). Estimation of the proportion of true null hypotheses in high-dimensional data under dependence. *Comput. Statist. Data Anal.* **55**, 2665–2676.

Gandy, A. & Hahn, G. (2014). MMCTest – A Safe Algorithm for Implementing Multiple Monte Carlo Tests. *Scand. J. Stat.* **41**, 1083–1101.

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

Gusenleitner, D., Howe, E., Bentink, S., Quackenbush, J. & Culhane, A. (2012). iBBiG: iterative binary bi-clustering of gene sets. *Bioinform.* **28**, 2484–2492.

Han, B. & Dalal, S. (2012). A Bernstein-type estimator for decreasing density with application to p-value adjustments. *Comput. Statist. Data Anal.* **56**, 427–437.

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

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58**, 13–30.

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

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* **75**, 383–386.

Jiang, H. & Salzman, J. (2012). Statistical properties of an early stopping rule for resampling-based multiple testing. *Biometrika* **99**, 973–980.

Jupiter, D., Sahutoglu, J. & VanBuren, V. (2010). TreeHugger: A new test for enrichment of gene ontology terms. *INFORMS Journal on Computing* **22**, 210–221.

Knijnenburg, T., Wessels, L., Reinders, M. & Shmulevich, I. (2009). Fewer permutations, more accurate P-values. *Bioinform.* **25**, i161–i168.

Lai, T. (1976). On Confidence Sequences. *Ann. Statist.* **4**, 265–280.

Langaas, M., Lindqvist, B. & Ferkingstad, E. (2005). Estimating the proportion of true null hypotheses, with application to dna microarray data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67**, 555–572.

Li, G., Best, N., Hansell, A., Ahmed, I. & Richardson, S. (2012). BaySTDetect: detecting unusual temporal patterns in small area data via bayesian model choice. *Biostat.* **13**, 695–710.

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

Lu, H.-F., Dong, H.-T., Sun, C.-B., Qing, D.-J., Li, N., Wu, Z.-K., Wang, Z.-Q. & Li, Y.-Z. (2011). The panorama of physiological responses and gene expression of whole plant of maize inbred line yq7-96 at the three-leaf stage under water deficit and re-watering. *Theor. Appl. Genet.* **123**, 943–958.

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

Nusinow, D., Kiezun, A., O'Connell, D., Chick, J., Yue, Y., Maas, R., Gygi, S. & Sunyaev, S. (2012). Network-based inference from complex proteomic mixtures using SNIPE. *Bioinform.* **28**, 3115–3122.

Pekowska, A., Benoukraf, T., Ferrier, P. & Spicuglia, S. (2010). A unique h3k4me2 profile marks tissue-specific gene regulation. *Genome Research* **20**, 1493–1502.

Pounds, S. & Cheng, C. (2006). Robust estimation of the false discovery rate. *Bioinform.* **22**, 1979–1987.

Rahmatallah, Y., Emmert-Streib, F. & Glazko, G. (2012). Gene set analysis for self-contained tests: complex null and specific alternative hypotheses. *Bioinform.* **28**, 3073–3080.

Rom, D. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika* **77**, 663–665.

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

Roth, A. (1999). Multiple comparison procedures for discrete test statistics. *J. Statist. Plann. Inference* **82**, 101–117.

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

Schweder, T. & Spjøtvoll, E. (1982). Plots of p-values to evaluate many tests simultaneously. *Biometrika* **69**, 493–502.

Shaffer, J. (1986). Modified sequentially rejective multiple test procedures. *J. Amer. Statist. Assoc.* **81**, 826–831.

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

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

Storey, J. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64**, 479–498.

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

van Wieringen, W., van de Wiel, M. & van der Vaart, A. (2008). A test for partial differential expression. *J. Amer. Statist. Assoc.* **103**, 1039–1049.

Westfall, P. & Troendle, J. (2008). Multiple testing with minimal assumptions. *Biom J.* **50**, 745–755.

Westfall, P. & Young, S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*. Wiley, New York.

Zhou, Y.-H., Barry, W. & Wright, F. (2013). Empirical pathway analysis, without permutation. *Biostat.* **14**, 573–585.

Address of corresponding author

Georg Hahn, Department of Mathematics, Imperial College London, London SW7 2AZ, U.K., e-mail: g.hahn11@imperial.ac.uk

A Proofs

For simplicity of notation we sometimes drop the dependence of the multiple testing procedure $h(p, \alpha)$ on the threshold α.

A.1 Proofs of Section 2.3

Proof of Lemma 1. 1. By construction, Algorithm 1 computes nested intervals, thus $\overline{p}_n = (\max I_n^i)_{i\in\{1,\dots,m\}} \searrow$ and $\overline{\alpha}_n = \max I_n^{m+1} \searrow$ as well as $\underline{p}_n = (\min I_n^i)_{i\in\{1,\dots,m\}} \nearrow$ and $\underline{\alpha}_n = \min I_n^{m+1} \nearrow$. Hence,

$$\underline{A}_n = h(\overline{p}_n, \underline{\alpha}_n) \subseteq h(\overline{p}_{n+1}, \underline{\alpha}_n) \subseteq h(\overline{p}_{n+1}, \underline{\alpha}_{n+1}) = \underline{A}_{n+1},$$
$$\overline{A}_n = h(\underline{p}_n, \overline{\alpha}_n) \supseteq h(\underline{p}_{n+1}, \overline{\alpha}_n) \supseteq h(\underline{p}_{n+1}, \overline{\alpha}_{n+1}) = \overline{A}_{n+1},$$

where the first (second) subset relation follows from the monotonicity of h (Condition 1) in the first (second) argument.

2. On the event R_1, $p_i^* \in I_n^i$ and $\alpha^* \in I_n^{m+1}$ for all i and n, thus $\overline{p}_n \geq p_n^* \geq \underline{p}_n$ and $\underline{\alpha}_n \leq \alpha^* \leq \overline{\alpha}_n$. By monotonicity of h (Condition 1), $\underline{A}_n = h(\overline{p}_n, \underline{\alpha}_n) \subseteq h(p^*, \alpha^*) \subseteq h(\underline{p}_n, \overline{\alpha}_n) = \overline{A}_n \ \forall n \in \mathbb{N}$. □

Proof of Theorem 1. Let $\overline{\alpha}_n = \max I_n^{m+1}$, $\underline{\alpha}_n = \min I_n^{m+1}$ as well as $B_n = \overline{A}_n \backslash \underline{A}_n$. Suppose $\exists i \in \limsup_{n\to\infty} B_n$. On the event R_2, $|I_n^i| \to 0$ as $n \to \infty$ for $i \in \limsup_{n\in\mathbb{N}} B_n$ as well as $|I_n^{m+1}| \to 0$ as $n \to \infty$. Let δ be as given in Condition 1. As $B_n \subseteq \{1, \dots, m\}$ is finite $\forall n \in \mathbb{N}$, there exists $n_0 \in \mathbb{N}$ such that $|I_n^i|^2 < \delta^2/m$ and $|\underline{\alpha}_n - \overline{\alpha}_n| < \delta$ for $n \geq n_0$ and all $i \in \limsup_{n\to\infty} B_n$.

We show that for all $n \geq n_0$,

$$\overline{A}_n = h((\min I_n^i)_{i\in\{1,\dots,m\}}, \overline{\alpha}_n) = h(p^*, \alpha^*) = h((\max I_n^i)_{i\in\{1,\dots,m\}}, \underline{\alpha}_n) = \underline{A}_n.$$

To do this, we show

$$\begin{aligned} h(p^{(1)}, \overline{\alpha}_n) &= h(p^{(2)}, \overline{\alpha}_n) = h(p^{(3)}, \overline{\alpha}_n), \\ h(p^{(3)}, \overline{\alpha}_n) &= h(p^{(4)}, \alpha^*) = h(p^{(5)}, \underline{\alpha}_n), \\ h(p^{(5)}, \underline{\alpha}_n) &= h(p^{(6)}, \underline{\alpha}_n) = h(p^{(7)}, \underline{\alpha}_n), \end{aligned}$$

where

$$\begin{aligned} p^{(1)} &:= (\min I_n^i)_{i\in\{1,\dots,m\}}, & p^{(4)} &:= p^*, \\ p^{(2)} &:= \begin{cases} \min I_n^i & i \in \overline{A}_n, \\ p_i^* & i \notin \overline{A}_n, \end{cases} & p^{(5)} &:= \begin{cases} \max I_n^i & i \in B_n, \\ p_i^* & i \notin B_n, \end{cases} \\ p^{(3)} &:= \begin{cases} \min I_n^i & i \in B_n, \\ p_i^* & i \notin B_n, \end{cases} & p^{(6)} &:= \begin{cases} \max I_n^i & i \in \overline{A}_n, \\ p_i^* & i \notin \overline{A}_n, \end{cases} \end{aligned}$$

and $p^{(7)} := (\max I_n^i)_{i\in\{1,\dots,m\}}$. The following holds true on the event R_1.

(1) By definition, $\overline{A}_n = h(p^{(1)}, \overline{\alpha}_n)$. As $p_j^{(2)} = p_j^* \geq \min I_n^j = p_j^{(1)}\ \forall j \notin \overline{A}_n$ and $p_j^{(2)} = p_j^{(1)}\ \forall j \in \overline{A}_n$, the first part of Condition 1 yields $\overline{A}_n = h(p^{(1)}, \overline{\alpha}_n) = h(p^{(2)}, \overline{\alpha}_n)$ for a fixed $\overline{\alpha}_n$.

(2) As $(\max I_n^i)_{i\in\{1,\dots,m\}} \geq p^{(3)}$ and as h is monotonic by Condition 1, $\underline{A}_n \subseteq h(p^{(3)}, \underline{\alpha}_n) \subseteq h(p^{(3)}, \overline{\alpha}_n)$. As $p_j^{(2)} = \min I_n^j \leq p_j^* = p_j^{(3)}\ \forall j \in \underline{A}_n$ and $p_j^{(2)} = p_j^{(3)}\ \forall j \notin \underline{A}_n$, the first part of Condition 1 yields $h(p^{(2)}, \overline{\alpha}_n) = h(p^{(3)}, \overline{\alpha}_n)$.

(3) On the event R_1, $|\underline{\alpha}_n - \overline{\alpha}_n| < \delta$ implies $|\alpha^* - \overline{\alpha}_n| < \delta$ and $|I_n^i|^2 < \delta^2/m$ implies $\|p^{(3)} - p^*\| < \delta$. The second part of Condition 1 thus yields $h(p^{(3)}, \overline{\alpha}_n) = h(p^{(4)}, \alpha^*) = h(p^*, \alpha^*)\ \forall n \geq n_0$.

Arguing similarly to (1), (2), (3) we can show $h(p^{(4)}, \alpha^*) = h(p^{(5)}, \underline{\alpha}_n)$ as well as $h(p^{(5)}, \underline{\alpha}_n) = h(p^{(6)}, \underline{\alpha}_n)$ and $h(p^{(6)}, \underline{\alpha}_n) = h(p^{(7)}, \underline{\alpha}_n) = \underline{A}_n$. □

Proof of Lemma 2. Let R be the set of rejected hypotheses and V be the set of rejected true null hypotheses. As $\underline{A}_n \subseteq R \subseteq \overline{A}_n$ for all $n \in \mathbb{N}$, $|\underline{A}_n| \leq |R| \leq |\overline{A}_n|$. Moreover, $\underline{V}_n \subseteq \underline{A}_n \subseteq R$ implies $|\underline{V}_n| \leq |V|$, and as the difference in numbers of rejected true null hypotheses in V and $\overline{V}_n$ cannot differ by more than the number of undecided hypotheses $|\overline{A}_n| - |\underline{A}_n|$ for any $n \in \mathbb{N}$, $|V| \leq |\overline{V}_n| \leq |V| + (|\overline{A}_n| - |\underline{A}_n|)$.

1. Using the above,

$$\frac{|\underline{V}_n|}{|\underline{A}_n|} = \frac{|V|}{|R|} + \frac{|\underline{V}_n||R| - |\underline{A}_n||V|}{|\underline{A}_n||R|} \leq \frac{|V|}{|R|} + \frac{|V|}{|R|}\frac{|R| - |\underline{A}_n|}{|\underline{A}_n|} \leq \frac{|V|}{|R|}\frac{|R|}{|\underline{A}_n|} \leq \frac{|V|}{|R|}\frac{|\overline{A}_n|}{|\underline{A}_n|}$$

for all $n \in \mathbb{N}$, thus $\mathbb{E}(|\underline{V}_s|/|\underline{A}_s|) \leq \eta\mathbb{E}(|V|/|R|) = \eta\alpha$.

2. Similarly,
$$\frac{|\overline{V}_n|}{|\overline{A}_n|} \leq \frac{|V| + (|\overline{A}_n| - |\underline{A}_n|)}{|\overline{A}_n|} \leq \frac{|V|}{|R|} + \frac{|\overline{A}_n| - |\underline{A}_n|}{|\overline{A}_n|}$$
for all $n \in \mathbb{N}$, thus $\mathbb{E}(|\overline{V}_t|/|\overline{A}_t|) \leq \mathbb{E}(|V|/|R|) + \xi = \alpha + \xi$.

□

A.2 Proofs of Section 4.1

The following two Lemmas will be needed for the proof of Lemma 3. First, Lemma 4 proves three properties of step-up and step-down procedures which are slightly stronger than the requirements stated in Condition 1. For a vector $p = (p_1, \ldots, p_m)$, we denote the rank of p_i in the sorted sequence $p_{(1)} \leq \ldots \leq p_{(m)}$ by $r_p(i)$.

Lemma 4. *Let $p, q \in [0,1]^m$. Let h_u (h_d) be a step-up (step-down) procedure defined through a threshold function τ_α satisfying Condition 2.*

1. *h_u is monotonic.*
2. *If $q_i \leq \tau_\alpha(|h_u(p)|)\ \forall i \in h_u(p)$ and $q_i = p_i\ \forall i \notin h_u(p)$, then $h_u(p) = h_u(q)$.*
3. *If $q_i = p_i\ \forall i \in h_u(p)$ and $q_i > \tau_\alpha(r_p(i))\ \forall i \notin h_u(p)$, then $h_u(p) = h_u(q)$.*
4. *h_d is monotonic.*
5. *If $q_i \leq \tau_\alpha(r_p(i))\ \forall i \in h_d(p)$ and $q_i = p_i\ \forall i \notin h_d(p)$, then $h_d(p) = h_d(q)$.*
6. *If $q_i = p_i\ \forall i \in h_d(p)$ and $q_i > \tau_\alpha(|h_d(p)| + 1)\ \forall i \notin h_d(p)$, then $h_d(p) = h_d(q)$.*

Proof. As h_u and h_d are invariant to permutations, we may assume $p_1 \leq \cdots \leq p_m$.

1. Let $p \in [0,1]^m$ and $i \in \{1, \ldots, m\}$. It suffices to show that $h_u(p) \supseteq h_u(q)$ for any $q \in [0,1]^m$ given by $q_j = p_j\ \forall j \neq i$ and $q_i > p_i$.

Let $k := |h_u(p)|$ be the largest rejected index. We need to show that $j \notin h_u(q)$ $\forall j \geq k+1$. Let α be fixed.

Case 1: $r_q(i) \leq k$. This implies $r_q(j) = j\ \forall j \geq k+1$ and hence $q_j = p_j > \tau_\alpha(j) = \tau_\alpha(r_q(j))$. Therefore, $j \notin h_u(q)\ \forall j \geq k+1$.

Case 2: $r_q(i) \geq k+1$. Let $j \geq k+1$, $j \neq i$. Then the rank of the jth p-value can only drop by one when p_i is replaced by q_i, i.e. $r_q(j) \in \{j-1, j\}$. Thus $q_j = p_j > \tau_\alpha(j) \geq \tau_\alpha(r_q(j))$ by Condition 2 (using that $\tau_\alpha(i)$ is non-decreasing in i). Furthermore, as $r_q(i) \geq k+1$, q_i takes the position of the former $p_{r_q(i)}$ in the ordered sequence of values from q, i.e. $q_i \geq p_{r_q(i)}$. Hence, $r_q(i) \notin h_u(p)$ because of $r_q(i) \geq k+1$ and thus $q_i \geq p_{r_q(i)} > \tau_\alpha(r_q(i))$. Therefore, $\{k+1, \ldots, m\} \cup \{i\} \notin h_u(q)$. This proves the monotonicity in the first argument of h_u.

The monotonicity in the second argument of h is immediate as $p_i \leq \max\{p_{(j)} : p_{(j)} \leq \tau_\alpha(j)\}$ for all $i \in h_u(p, \alpha)$. On Condition 2, and using that τ_α is non-decreasing in α, $\alpha \leq \alpha'$ implies $\tau_\alpha(j) \leq \tau_{\alpha'}(j)\ \forall j$, hence $i \in h_u(p, \alpha')$. This proves 1.

2. All $i \notin h_u(p)$ satisfy $p_i > \tau_\alpha(r_p(i)) > \tau_\alpha(|h_u(p)|)$ whereas by assumption, $q_i \leq \tau_\alpha(|h_u(p)|)\ \forall i \in h_u(p)$. Hence, using $q_i = p_i\ \forall i \notin h_u(p)$, it follows that $r_q(i) = r_p(i)\ \forall i \notin h_u(p)$. Thus, $q_i = p_i > \tau_\alpha(r_p(i)) = \tau_\alpha(r_q(i))$ for all $i \notin h_u(p)$. Hence $h_u(p)^c \subseteq h_u(q)^c$.

Conversely, define $\tilde{q} := \max\{q_i : i \in h_u(p)\}$. As $\tilde{q} \leq \tau_\alpha(|h_u(p)|) < q_i$ for all $i \notin h_u(p)$ and as there are precisely $|h_u(p)|$ values $q_i \leq \tilde{q}$, the rank of $\tilde{q}$ in q is precisely $|h_u(p)|$. As $q_i \leq \tilde{q} \leq \tau_\alpha(|h_u(p)|)\ \forall i \in h_u(p)$, all $\{q_i\}_{i \in h_u(p)}$ are rejected, so $h_u(p) \subseteq h_u(q)$. This proves 2.

3. As $q_i = p_i$ for all $i \in h_u(p)$, have $h_u(p) \subseteq h_u(q)$.

Let $i \notin h_u(p)$. If $r_q(i) \leq r_p(i)$, then $q_i > \tau_\alpha(r_p(i)) \geq \tau_\alpha(r_q(i))$ by Condition 2. If $r_q(i) > r_p(i)$, q_i replaces a $q_j > \tau_\alpha(r_p(j))$ at rank $r_p(j)$ in the sorted sequence of q, hence $r_q(i) = r_p(j)$ and $q_i \geq q_j > \tau_\alpha(r_p(j)) = \tau_\alpha(r_q(i))$. Thus $q_i > \tau_\alpha(r_q(i))\ \forall i \notin h_u(p)$, which implies $h_u(p)^c \subseteq h_u(q)^c$. This proves 3.

In a similar fashion, 4., 5. and 6. can be proven for step-down procedures h_d. □

For step-up procedures h_u, part 2. of Lemma 4 shows that p-values of rejected hypotheses can be increased up to $\tau_\alpha(|h_u(p)|)$, the threshold evaluated at the last rejected hypothesis, without affecting the result of h_u. Part 3. of Lemma 4 shows that h_u is invariant if p-values in the non-rejection area are replaced by arbitrary values above the threshold.

Similarly, step-down procedures h_d are invariant if p-values of rejected hypotheses are replaced by arbitrary values below the threshold (part 5.) or p-values of non-rejected hypotheses are replaced by arbitrary values above $\tau_\alpha(|h_d(p)|+1)$, the threshold evaluated at the first non-rejected hypothesis (part 6.).

The following Lemma 5 will also be needed for the proof of Lemma 3. In the following, $\|\tau_\alpha\|_\infty$ shall denote the maximal value attained by $\tau_\alpha : \{1, \ldots, m\} \to [0, 1]$ on $\{1, \ldots, m\}$.

Lemma 5. *Let h stand for h_u or h_d. If $p^* \in [0, 1]^m$, $\alpha^* > 0$ with $p^*_{(i)} \neq \tau_{\alpha^*}(i)$ $\forall i \in \{1, \ldots, m\}$, then there exists $\delta > 0$ such that $p \in [0, 1]^m$, $\tau_\alpha : \{1, \ldots, m\} \to [0, 1]$ and $\|p^* - p\| \vee \|\tau_{\alpha^*} - \tau_\alpha\|_\infty < \delta\ \forall i \in \{1, \ldots, m\}$ imply $h(p, \alpha) = h(p^*, \alpha^*)$.*

Proof. Let

$$\delta' = \min\left(\left\{\frac{p^*_i - p^*_{i-1}}{2} : p^*_{i-1} < p^*_i\right\}_{i=1,\ldots,m} \cup \{|p^*_i - \tau_{\alpha^*}(i)|\}_{i=1,\ldots,m}\right)$$

and let $\delta = \delta'/2$.

By assumption, $\|p - p^*\| < \delta < \delta'$, hence $p_{i-1} < p^*_{i-1} + \delta' \leq p^*_i - \delta' < p_i$. This means that p^*_i and p_i have the same ranks in p^* and p, respectively.

Moreover, $|p^*_i - \tau_{\alpha^*}(i)| \leq |p^*_i - \tau_\alpha(i)| + |\tau_\alpha(i) - \tau_{\alpha^*}(i)| \leq |p^*_i - \tau_\alpha(i)| + \|\tau_\alpha - \tau_{\alpha^*}\|_\infty \leq |p^*_i - \tau_\alpha(i)| + \delta$. Hence $2\delta = \delta' \leq |p^*_i - \tau_{\alpha^*}(i)| \leq |p^*_i - \tau_\alpha(i)| + \delta$, meaning that $\delta \leq |p^*_i - \tau_\alpha(i)|$ for $i \in \{1, \ldots, m\}$.

So $|p^*_i - p_i| < \delta \leq |p^*_i - \tau_\alpha(i)|$, hence p_i and p^*_i lie on the same side of the testing threshold. □

Proof of Lemma 3. 1. The monotonicity of h_u and h_d follows from Lemma 4 (part 1.) and (part 4.), respectively.

2. To prove that h_u satisfies the first part of Condition 1, it suffices to show that for $p, q \in [0,1]^m$, both $q_i \leq p_i \ \forall i \in h_u(p)$ and $q_i = p_i \ \forall i \notin h_u(p)$ as well as $q_i = p_i \ \forall i \in h_u(p)$ and $q_i \geq p_i \ \forall i \notin h_u(p)$ imply $h_u(p) = h_u(q)$.

Indeed, let $p, q \in [0,1]^m$ be such that $q_i \leq p_i \ \forall i \in h_u(p)$ and $q_i = p_i \ \forall i \notin h_u(p)$. We have $p_i \leq \tau_\alpha(|h_u(p)|) \ \forall i \in h_u(p)$, thus $q_i \leq p_i \leq \tau_\alpha(|h_u(p)|) \ \forall i \in h_u(p)$ and $h_u(p) = h_u(q)$ by Lemma 4 (part 2.).

Similarly, let $p, q \in [0,1]^m$ be such that $q_i = p_i \ \forall i \in h_u(p)$ and $q_i \geq p_i \ \forall i \notin h_u(p)$. Using $p_i > \tau_\alpha(r_p(i)) \ \forall i \notin h_u(p)$, it instantly follows that $q_i \geq p_i > \tau_\alpha(r_p(i)) \ \forall i \notin h_u(p)$ and thus $h_u(p) = h_u(q)$ by Lemma 4 (part 3.).

To prove that h_d satisfies the first part of Condition 1, it equally suffices to show that for $p, q \in [0,1]^m$, both $q_i \leq p_i \ \forall i \in h_d(p)$ and $q_i = p_i \ \forall i \notin h_d(p)$ as well as $q_i = p_i \ \forall i \in h_d(p)$ and $q_i \geq p_i \ \forall i \notin h_d(p)$ imply $h_d(p) = h_d(q)$.

Indeed, let $p, q \in [0,1]^m$ be such that $q_i \leq p_i \ \forall i \in h_d(p)$ and $q_i = p_i \ \forall i \notin h_d(p)$. Using $p_i \leq \tau_\alpha(r_p(i)) \ \forall i \in h_d(p)$, it immediately follows that $q_i \leq p_i \leq \tau_\alpha(r_p(i)) \ \forall i \in h_d(p)$ and thus $h_d(p) = h_d(q)$ by Lemma 4 (part 5.).

Similarly, let $p, q \in [0,1]^m$ be such that $q_i = p_i \ \forall i \in h_d(p)$ and $q_i \geq p_i \ \forall i \notin h_d(p)$. We have $p_i > \tau_\alpha(|h_d(p)| + 1) \ \forall i \notin h_d(p)$, thus $q_i \geq p_i > \tau_\alpha(|h_d(p)| + 1) \ \forall i \notin h_d(p)$ and $h_d(p) = h_d(q)$ by Lemma 4 (part 6.).

3. As $\tau_\alpha(i)$ is continuous in $\alpha \ \forall i \in \{1, \ldots, m\}$ by Condition 2, for each $\epsilon_i > 0$ there exists a $\delta_i > 0$ such that $|\alpha^* - \alpha| < \delta_i$ implies $|\tau_{\alpha^*}(i) - \tau_\alpha(i)| < \epsilon_i$. Applying continuity to $\epsilon_i = \delta$ yields a δ_i for each $i \in \{1, \ldots, m\}$, where $\delta > 0$ is given by Lemma 5. The second part of Condition 1 then follows for all $p \in [0,1]^m$ and $\alpha \in [0,1]$ satisfying $\|p - p^*\| \vee |\alpha - \alpha^*| < \min\{\delta, \delta_1, \ldots, \delta_m\}$. □

B The Hommel procedure is not well-behaved

The Hommel (1988) procedure determines the largest index k satisfying $p_{(m-k+j)} > j\alpha/k$ for all $j = 1, \ldots, k$ and then rejects all the H_{0i} with $p_i \leq \alpha/k$. If no such k

exists, all hypotheses are rejected.

The Hommel (1988) procedure $h(p, \alpha)$ is not a classical step-up or step-down procedure. Given p, determining the index k corresponds to applying m step-up procedures h_j, $j \in \{1, \ldots, m\}$, to $P_j = (p_{(m-j+1)}, \ldots, p_{(m)})$ using the threshold functions $\tau_j(i) = i\alpha/j$, $i \in \{1, \ldots, j\}$. Once $k_p = \max\{j : h_j(P_j) = \emptyset\}$ is determined, rejections are calculated by applying the Bonferroni (1936) correction (defined in Section 4.2) at threshold α/k_p to all p-values p, i.e. $h(p, \alpha) = h_{\text{Bonferroni}}(p, m\alpha/k_p)$.

The Hommel (1988) procedure satisfies the first and the third part of Condition 1. However, for $q_i \geq p_i \ \forall i \notin h(p, \alpha)$, the second part of Condition 1 is not satisfied. Consider $p = [\alpha/3+\epsilon, \alpha/2+\epsilon, 1]$, where $0 < \alpha < 1$ and $0 < \epsilon \leq \alpha/6$. Then $h_1(P_1) = \emptyset$, $h_2(P_2) = \emptyset$, $h_3(P_3) = \{1, 2\}$, so $k_p = 2$. Therefore, $h(p, \alpha) = \{1\}$. Increasing p_2 to $p_2 = 2\alpha/3 + \epsilon$ yields $h_1(P_1) = h_2(P_2) = h_3(P_3) = \emptyset$, hence $k_p = 3$ and $h(p, \alpha) = \emptyset$.