2

Extending Hommel's procedure: Efficient algorithms for closed testing based on Simes' test

Abstract

We present an algorithm to calculate adjusted p-values as given by Hommel's procedure more efficiently. Furthermore, we explain how one can easily verify whether an intersection hypothesis can be rejected by a closed testing procedure in combination with a Simes' test, without having to carry out this full procedure. Using this observation, we subsequently develop an algorithm that, based on this same closed testing procedure, calculates confidence sets for the number of true (or false) hypotheses within any set of elementary hypotheses.

This chapter has been submitted as: Rosa J. Meijer, Thijmen J.P. Krebs, Aldo Solari and Jelle J. Goeman. Extending Hommel's procedure: Efficient algorithms for closed testing based on Simes' test.

2.1 Introduction

When multiple null-hypotheses are tested, a multiple testing correction is needed in order to prevent false positive results. For this reason, methods that control the familywise error rate (FWER) have been developed. The familywise error rate is defined as the probability of at least one type I error, where a type I error is the rejection of a null-hypothesis that was actually true. Controlling the FWER on significance level α thus means that the probability of having only correct rejections is at least $1 - \alpha$.

In this paper, we will focus on the multiple testing method of Hommel (1988); a method that controls the FWER under certain assumptions. Hommel's method is a shortcut for the closed testing procedure, as developed by Marcus et al. (1976), in which in addition to all elementary null-hypotheses each intersection hypothesis is tested with a Simes' test, as we will explain in more detail in the next section. Whereas carrying out a closed testing procedure takes exponential time with respect to the number of elementary null-hypotheses, say n, the current implementation of Hommel's procedure will only take $\Theta(n^2)$ time. This can still be slow however, when the number of elementary hypotheses is very large. This can happen in a genomic context for example. One could think of research in which the aim is to find certain SNPs (Single Nucleotide Polymorphisms) or genes that are associated with an interesting outcome variable, such as a specific disease. Often there are many candidate SNPs or genes, which results in many elementary hypotheses. To be able to still use Hommel's procedure in such situations, we developed an algorithm that carries out Hommel's procedure in $O(n \log(n))$ time. With this new procedure, it will be feasible to use Hommel's multiple testing method even if there are millions of elementary hypotheses.

Hommel's method is specifically developed to test the elementary null-hypotheses, but for genomic data not only the elementary hypotheses are of interest but looking for an association between the outcome and for example sets of interacting genes can be of interest as well. These gene sets could be presented as intersection hypotheses, which are the hypotheses tested in a closed testing procedure. To find out whether a certain intersection hypothesis can be rejected within a closed testing procedure will usually take a lot of time, but in this article we will show that, if the Simes' test is used throughout the full closed testing procedure, there is an easy rule that can be used to decide whether a certain hypothesis can be rejected by the closed testing procedure, without having to carry out this full procedure.

If the intersection hypothesis can be rejected, or in our example, if the gene set is shown to be associated with the outcome, the next question could be how many of the individual genes are at least associated with the outcome. Goeman and Solari (2011) showed that this question can be answered by using the outcome of a full closed testing procedure. Using this procedure, we can make confidence sets for the number of true (or false) hypotheses within every arbitrarily chosen set of elementary null-hypotheses. Because the full closed testing procedure will again be too time-consuming to carry out fully, once more we will use the fact that we are using a Simes' test within this procedure.

From this it will follow that not all intersection hypotheses have to be tested and we will show that the confidence statements can be obtained by means of a fast algorithm.

In the remainder of this article we will describe two algorithms: one that can be used to carry out Hommel's method efficiently, and one that can be used to derive confidence statements. In Section 2.2, we will first give a short introduction to Hommel's method and the corresponding closed testing procedure. In section 2.3 we will subsequently present a faster way to carry out Hommel's procedure. In section 2.4 we will say more about confidence statements and we will describe the algorithm to derive them. In the Application section, we will further illustrate our algorithms on the basis of some real and simulated data examples and we will conclude with a short discussion.

2.2 Hommel's procedure

In this article, we will focus on one of the four most well-known FWER controlling multiple testing methods, namely the procedure of Hommel (1988). The other three are the methods of Bonferroni (Bonferroni, 1935) , Holm (Holm, 1979) and Hochberg (Hochberg, 1988). All these methods start from the n raw (i.e. unadjusted) p-values $p_1, \ldots, p_n$ corresponding to the n null-hypotheses of interest $H_1, \ldots, H_n$. In the remainder of this article, we will for convenience assume that $p_1 \leq p_2 \leq \ldots \leq p_n$, where potential ties are broken arbitrarily. Bonferroni's and Holm's method both control the FWER without making assumptions on the dependence of the raw p-values. Although this is a strong property of these methods, the methods can be rather conservative for this same reason. The methods of Hochberg and Hommel on the other hand are only valid under some assumptions on the underlying dependency structure of the p-values, but these assumptions are usually not considered too restrictive, and the methods are less conservative compared to Bonferroni's and Holm's method. Hochberg's procedure is faster than Hommel's procedure, but is less powerful (Hommel, 1989). We will show however that Hommel's procedure can be implemented in such a way that it has the same time complexity as Hochberg's procedure which will make it the preferred method of the two in every situation.

Hommel's procedure is a shortcut of the closed testing procedure of Marcus et al. (1976). In the closed testing procedure, all possible intersection hypotheses $H_I = \bigcap_{i \in I} H_i$, with $I \subseteq \{1, \ldots, n\}$ nonempty, are tested on level α with a certain α-level test which is called the local test. In Figure 2.1 the collection of all intersection hypotheses in the situation in which we have 3 elementary null-hypotheses is displayed. An intersection hypothesis H_I (note that an elementary hypothesis H_i is also an intersection hypothesis) is subsequently rejected if and only if all intersection hypotheses H_J with $I \subseteq J$ can be rejected by this local test. If we denote the intersection hypothesis of all n_0 *true* null-hypotheses by H_T, where T thus denotes the index set consisting of all indices corresponding to an H_i that is in fact true, we see that the closed testing procedure controls the FWER if H_T is tested with a valid α-level test. Each true (intersection) null-hypothesis $H_{T'}$ can only be rejected if the same holds for H_T, since $T' \subseteq T$, and if H_T is tested

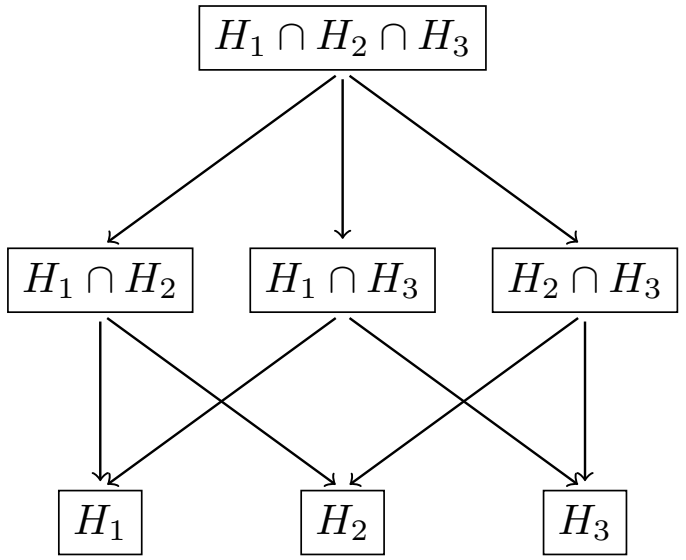


Figure 2.1: All intersection hypotheses for a set of 3 elementary hypotheses in the form of a graph. The nodes represent the hypotheses, the edges denote underlying subset relationships.

with a valid test, this can only happen with probability at most α.

Hommel's procedure is based on the closed testing procedure in combination with a Simes' test as developed by Simes (1986). This test is based on the following inequality:

$$P\left(\bigcup_{i\in I}\left\{p_i \leq \frac{\text{rk}_I(i)\alpha}{|I|}\right\}\right) \leq \alpha, \tag{2.1}$$

where $|I|$ denotes the size of set I and $\text{rk}_I(i)$ is the rank of p-value p_i within the set of p-values $\{p_i : i \in I\}$, that is, $\text{rk}_I(i) = |\{i' \in I : i' <= i\}|$. Based on this Simes' inequality, an intersection hypothesis H_I will be rejected by a Simes' test if and only if there is at least one p-value p_i with $i \in I$ for which $p_i \leq \text{rk}_I(i)\alpha/|I|$. To have a valid α-level test for the intersection hypothesis H_T of all true null-hypotheses, inequality (2.1) must hold for the n_0 p-values corresponding to the n_0 true null-hypotheses. The assumptions on the dependence of the n_0 p-values needed to make this inequality hold are technical but fairly general as well (see e.g. Sarkar, 1998; Goeman and Solari, 2014) and the Simes' test will for that reason be valid in many situations. For example in the situation where we have p-values from identically distributed, non-negatively correlated test statistics (Goeman and

Solari, 2011). Because Hommel's multiple testing method is based on the Simes' test, it will control the FWER given that the assumptions underlying the validity of the Simes' test are met.

Although Hommel's method will reject the same elementary hypotheses as the closed testing procedure in combination with a Simes' test, the fact that the method is a shortcut of the closed testing procedure means that not all $2^n - 1$ intersection hypotheses have to be tested in order to find out which of the n elementary hypotheses can be rejected. For n null-hypotheses $H_1, \ldots, H_n$ with ordered p-values $p_1 \leq \ldots \leq p_n$, using Hommel's method to control the FWER on level α comes down to computing

$$j(\alpha) = \max \{s \in \{1, \ldots, n\} : p_{n-s+k} > k\alpha/s, \text{ for } k = 1, \ldots, s\}, \tag{2.2}$$

and rejecting all H_i with $p_i \leq \alpha/j(\alpha)$, or all H_i if this maximum does not exist. The maximum $j(\alpha)$ can be seen as the size of the largest intersection hypothesis H_I, with $I \subseteq \{1, \ldots, n\}$ that cannot be rejected with a Simes' test on level α. If such an hypothesis does not exist, this means that all intersection hypotheses and for that reason all elementary hypotheses can be rejected. From the formula it is clear that in order to find whether there is any intersection hypothesis H_I of a certain size $|I| = s$ that cannot be rejected by a Simes' test, Hommel only looks at the hypothesis H_I that is most difficult to reject. This will always be the intersection $H_I = \bigcap_{i=n-s+1}^{n} H_i$ of the s elementary hypotheses with the s largest unadjusted p-values (Wright, 1992). If this hypothesis can be rejected, the same will hold for all intersection hypotheses of the same size.

Using this procedure will give the elementary hypotheses that can be rejected while controlling the FWER on one specific α-level. To have more information than just the statement whether to reject a specific hypothesis or not, for each H_i we can calculate an *adjusted* p-value instead, which is the smallest overall significance level at which the corresponding hypothesis would be rejected *within the multiple testing setting* (see e.g. Wright, 1992). In case of a closed testing procedure, the adjusted p-value for an intersection hypothesis H_I would thus be the minimum value of α needed to reject all intersection hypotheses H_J with $J \supseteq I$.

In Hommel's procedure, to find adjusted p-values for all elementary hypotheses H_i, we need to compute the maximum $j(\alpha)$ as given in equation (2.2) not only for one value of α, but for all values of $\alpha \in [0, 1]$. This will result in a step-function j, where $j(\alpha)$ is the size of the largest intersection hypothesis that cannot be rejected (with a Simes' test) on this α-level. As j is a step function, it suffices to find the values of α on which the function jumps. We will see that those jump values can be easily determined once we computed the minimum α-level needed to reject all intersection hypotheses H_I of size $s = |I|$, which we will denote by α_s, for each value $s \in \{1, \ldots, n\}$.

As explained before, the minimum α needed to reject all intersection hypotheses of size s will equal the minimum α needed to reject the intersection hypothesis $H_I = \bigcap_{i=n-s+1}^{n} H_i$ which is the intersection of the s elementary hypotheses with the largest p-values. To find the minimum α needed to reject this hypothesis with a Simes' test comes down to finding the smallest value of α for which $p_{n-s+k} = k\alpha/s$ for at least one

$k \in \{1, \ldots, s\}$. So, for each $s \in \{1, \ldots, n\}$, α_s is given by:

$$\alpha_s = \min_{k \in \{1,\ldots,s\}} \frac{s}{k} \cdot p_{n-s+k} = s \cdot \min_{k \in \{1,\ldots,s\}} \frac{p_{n-s+k}}{k}. \tag{2.3}$$

Computing n of those minima will normally take $\Theta(n^2)$ time and this is also how Hommel's procedure is currently implemented in the `R`-function `p.adjust`. In the next section, we will explain how the complexity can be reduced from $\Theta(n^2)$ to $O(n \log(n))$ by using information from the location of previously calculated minima to make computing new minima easier.

2.3 More efficient implementation of Hommel's procedure

Calculating all minima given in equation (2.3) is equivalent to finding a minimum in each column of the following matrix M:

$$M = \begin{pmatrix} p_1 & & & & & \\ \frac{p_2}{2} & p_2 & & & & \\ \frac{p_3}{3} & \frac{p_3}{2} & p_3 & & & \\ \frac{p_4}{4} & \frac{p_4}{3} & \frac{p_4}{2} & p_4 & & \\ \vdots & \vdots & \vdots & \vdots & \ddots & \\ \frac{p_n}{n} & \frac{p_n}{n-1} & \frac{p_n}{n-2} & \frac{p_n}{n-3} & \cdots & p_n \end{pmatrix}.$$

To see how the the number of steps needed to find a minimum in each column can be reduced from $\Theta(n^2)$ to $O(n \log(n))$, we formulate the following lemma and corresponding proof.

Lemma 2.3.1. If, for matrix M, a minimum in column c is found on the i^{th} row, a minimum in column c' with $c' < c$ will be found on a row j with $j \leq i$ and a minimum in column c' with $c' > c$ will be found on a row j with $j \geq i$.

Proof. Suppose a minimum in column c is found on row i, which means the minimum is given by $\frac{p_i}{i-c+1}$. Being a minimum means that:

$$\forall j \neq i, j \in \{c, \ldots, n\} : \frac{p_i}{i-c+1} \leq \frac{p_j}{j-c+1}. \tag{2.4}$$

For a column $c' < c$, it now suffices to show that

$$\forall j > i \colon \frac{p_i}{i-c'+1} \leq \frac{p_j}{j-c'+1},$$

which leads to the first statement of lemma 2.3.1.
To show that this holds, we take $j > i$ and start by multiplying both sides of equation (2.4) by $(i - c + 1)(j - c + 1)$ which gives:

$$p_i(j - c + 1) \leq p_j(i - c + 1).$$

Since $(c - c')p_i \leq (c - c')p_j$ because $p_i \leq p_j$ for $i < j$ and $c' < c$, adding this inequality gives:

$$p_i(j - c' + 1) \leq p_j(i - c' + 1).$$

Dividing both sides by $(j - c' + 1)(i - c' + 1)$ now gives the desired result:

$$\frac{p_i}{i - c' + 1} \leq \frac{p_j}{j - c' + 1}.$$

The second statement in lemma 2.3.1 now immediately follows from a proof by contradiction. Given that a minimum in column c is found on row i, assume that a minimum in column c' with $c' > c$ is found in row j with $c' \leq j < i$. From the first statement in lemma 2.3.1 it would now follow that a minimum in column c would lie in a row k with $c \leq k \leq j < i$ which is a contradiction. A minimum in column c' must thus be found on a row j with $j \geq i$. □

By using Lemma 2.3.1, finding the minimum in every column of the previously defined matrix M can be done in $O(n \log(n))$ steps by dividing the matrix in sub matrices in every step and always computing the minimum in the middle column of a sub matrix. We start by taking the minimum of the middle column m, which we will assume to be at row i. This will take $O(n)$ steps. The matrix is now naturally divided in two sub matrices; the matrix consisting of column $1, \ldots, m-1$ and row $1, \ldots, i$ because all minima in those columns can, by Lemma 2.3.1, only be in these rows, and the matrix consisting of column $m + 1, \ldots, n$ and row $i, \ldots, n$. For both sub matrices we take the minimum of their middle column. This time, calculating both minima simultaneously will take $O(n)$ steps. Both matrices are now divided in two new sub matrices, which gives four sub matrices for which we can find the four minima of their middle columns in $O(n)$ time and so we continue. In each step, twice the number of minima compared to the previous step are calculated, and finding the minima for all n columns will thus take $\lceil \log(n) \rceil$ steps. In each step, all minima can be simultaneously calculated in $O(n)$ time which brings the complexity of the whole procedure to $O(n \log(n))$.

Finding all values of α_s as given in equation (2.3) can thus be done in $O(n \log(n))$ time. We can order those n values from α_n to α_1 and subsequently take a cumulative maximum to get the minimum α-values on which all intersection hypotheses H_I of size $|I|$ and larger can be rejected. We will denote the sequence of cumulative maxima by $\alpha_n^*, \alpha_{n-1}^*, \ldots, \alpha_1^*$. Whereas, for the Simes critical values, it is easy to show that $\alpha_s^* = \alpha_s$ for all $s \in \{1, \ldots, n\}$, we keep the distinction between α_s and α_s^* for conceptual reasons and for easier generalizability of the algorithm to other critical values. From the values of α_s^* with $s \in \{1, \ldots, n\}$ we can directly calculate the function j. For $\alpha < \alpha_n^*$, $j(\alpha) = n$

because the overall intersection hypothesis cannot be rejected on these levels. At $\alpha = \alpha_n^*$, j jumps to $s-1$, where s is the smallest value for which $\alpha_s^* = \alpha_n^*$. Indeed, all intersection hypotheses H_I for which $|I| \geq s$ can be rejected on this α-level but there is at least one intersection hypothesis H_I with $|I| = s - 1$ which cannot be rejected on this level. The next jump of j will subsequently be at α_{s-1}^*, et cetera. The function j will thus jump on all distinct values in the sequence $\alpha_n^*, \ldots, \alpha_1^*$ and the size of the jump will depend on the number of successive equal values in this sequence.

From j we can subsequently calculate adjusted p-values for each elementary hypothesis H_i with corresponding p-value p_i by finding the minimum value of α for which $p_i \leq \alpha/j(\alpha)$. For all n elementary hypotheses this can be done simultaneously in $O(n)$ time. Given n elementary hypotheses $H_1, \ldots, H_n$ with corresponding (sorted) p-values $p_1, \ldots, p_n$, we thus showed that finding the adjusted p-values based on Hommel's multiple testing procedure can be done in $O(n \log(n))$ time, which is the same order that is already needed for sorting the p-values, which is a requirement for Hochberg's and Hommel's procedure as well. In the Application section, we will provide some data examples to illustrate the gain in computation time when using this new algorithm compared to the current implementation as found in the `R`-function `p.adjust`.

In addition to rejecting elementary hypotheses, we also want to look at intersection hypotheses H_I and formulate confidence statements for the number of elementary hypotheses H_i with $i \in I$ that have to be false with probability $1 - \alpha$ for a given α. For this purpose, we will again use the function j as we will explain in the next section.

2.4 Confidence Sets

Given the n elementary hypotheses $H_1, \ldots, H_n$, we know there is a subset of the actual *true* hypotheses. Let $T \subseteq \{1, \ldots, n\}$ denote the unknown index set corresponding to these true hypotheses. In this section, we will discuss how confidence sets can be constructed for the number of true hypotheses $\tau(S) = |T \cap S|$ within an arbitrary set $\{H_i \colon i \in S\}$ of elementary hypotheses with $S \subseteq \{1, \ldots, n\}$.

If Hommel's procedure is used on the elementary hypotheses, there will usually be some hypotheses that cannot be rejected on a certain α-level. This does not imply however that these hypotheses are true hypotheses. Carrying out the full closed testing procedure in which all possible intersection hypotheses $H_I = \bigcap_{i \in I} H_i$ with $I \subseteq \{1, \ldots, n\}$ nonempty are tested with a Simes' test on this same α-level can result in the rejection of intersection hypotheses H_I for which none of the corresponding elementary hypotheses H_i with $i \in I$ can be rejected. Because an intersection hypothesis H_I is true if and only if all H_i with $i \in I$ are true, rejecting H_I within the closed testing procedure means that with probability at least $1 - \alpha$ one or more of the corresponding elementary hypotheses H_i were false, even though the exact hypotheses cannot be identified by the closed testing procedure, or equivalently by Hommel's procedure. Using the full closed testing procedure instead of Hommel's procedure can thus result in more information about the number of true (and

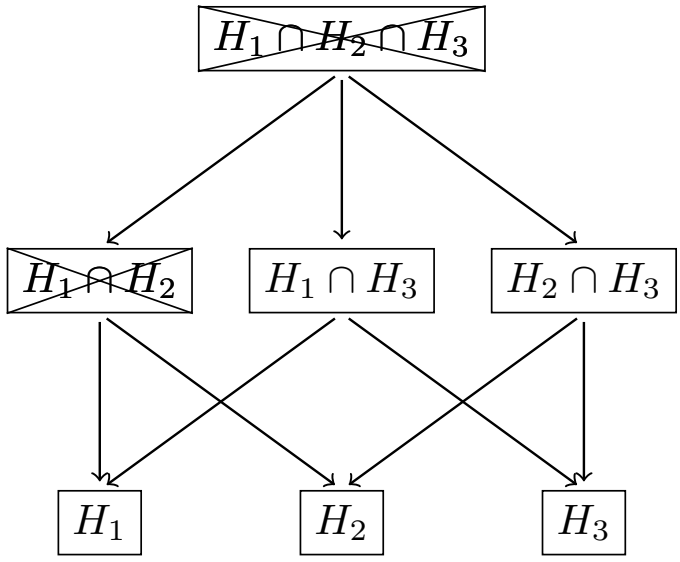


Figure 2.2: All intersection hypotheses for a set of 3 elementary hypotheses. When $p_1 = p_2 = \frac{2}{3}\alpha$ and $p_3 > \alpha$, using the closed testing procedure in combination with a Simes' test results in two rejections (indicated by the crosses) but none of the elementary hypotheses are rejected. Hommel's method will for that reason result in no rejections, while the results of the closed testing procedure show that at least one of the first two elementary hypotheses has to be false with probability at least $1 - \alpha$.

false) elementary hypotheses.

An example of such a situation is given in Figure 2.2. In this example, the intersection hypothesis $H_1 \cap H_2$ was rejected by the closed testing procedure in combination with a Simes' test, but neither H_1 nor H_2 could be rejected by the same procedure. Still, we know that at least one of these two hypotheses has to be false with probability at least $1 - \alpha$, because if $H_1 \cap H_2$ is justly rejected this can only be because at least one of its components is a non-true hypothesis. This statement can be seen as a confidence statement about the number of false null-hypotheses within the set $\{H_1, H_2\}$. Similar statements, all based on the results of the full closed testing procedure, could be made for any arbitrary set of elementary hypotheses, as proposed by Goeman and Solari (2011).

Goeman and Solari (2011) showed that, given the results of one closed testing procedure, for each arbitrary set of elementary hypotheses a confidence set can be constructed for the number of true hypotheses within this set. All confidence sets are simultaneously

valid, because they all follow from the same closed testing procedure of which the results are valid with probability at least $1 - \alpha$. If we denote the set of all possible intersection hypotheses, also called the closure, by $\mathcal{C}$ and if we denote the set of all hypotheses that are rejected by the closed testing procedure on level α by $\mathcal{R} \subseteq \mathcal{C}$, Goeman and Solari (2011) showed that each $100(1-\alpha)\%$ confidence set for the number of true hypotheses $\tau(S)$ within a set $\{H_i : i \in S\}$ can be given by

$$\{0, \ldots, t_\alpha(S)\}, \tag{2.5}$$

with

$$t_\alpha(S) = \max\{|I| : I \subseteq S, H_I \notin \mathcal{R}\}, \tag{2.6}$$

where $t_\alpha(S) = 0$ if such an H_I does not exist. The quantity $t_\alpha(S)$ is the size of the largest subset of S for which the corresponding intersection hypothesis is not rejected by the closed testing procedure.

Often, not the number of true hypotheses within a given set, but the number of false hypotheses within this set will be of interest because false hypotheses usually correspond with actual findings. The $100(1-\alpha)\%$ confidence set for the number of false hypotheses $\phi(S) = |S| - \tau(S)$ follows immediately from (2.5):

$$\{f_\alpha(S), \ldots, |S|\}, \tag{2.7}$$

where $f_\alpha(S) = |S| - t_\alpha(S)$.

To construct both confidence sets, as given in (2.5) and (2.7), all that is needed is thus the quantity $t_\alpha(S)$ which is the largest subset of S for which the corresponding intersection hypotheses cannot be rejected by the closed testing procedure. In the remainder of this section we will discuss a procedure to compute $t_\alpha(S)$ in an efficient way. For this, we first need a method that allows us to easily verify whether an hypothesis H_I can be rejected by the closed testing procedure. Based on this method, we can subsequently develop an algorithm that calculates $t_\alpha(S)$ and $f_\alpha(S)$.

2.4.1 Determining whether an hypothesis can be rejected by the closed testing procedure in an efficient way

To find the size of the largest intersection hypothesis H_I, $I \subseteq S$ that cannot be rejected by the closed testing procedure, we could proceed as before, and check for each s from $|S|$ to 1 whether the intersection hypothesis H_I, $I \subseteq S$ and $|I| = s$, that is most difficult to reject, i.e. the intersection of s elementary hypotheses H_i, $i \in S$ corresponding to the largest possible p-values p_i, can be rejected *within the full closed testing procedure*. The difficult part here is that we have to decide whether H_I can be rejected within the closed testing procedure, meaning that we need to know whether all hypotheses H_J with $J \supset I$ can be rejected with a Simes' test on a certain α-level.

Verifying naively whether all hypotheses H_J with $J \supset I$ can be rejected with a Simes' test will generally not be feasible. Motivated by the flow chart (Fig. 3) of Hommel (1986),

we propose a simple procedure, as given in Procedure 2.4.1, that can be used to directly determine whether an hypothesis H_I can be rejected within the closed testing procedure. This procedure is preceded by Theorem 2.4.1. Throughout, we will assume that $j(\alpha)$ exists, since otherwise all elementary hypotheses and thereby all intersection hypotheses can be rejected.

Theorem 2.4.1. *The closed testing procedure in combination with a Simes' test rejects an intersection hypothesis H_I on level α if and only if there exists some $i \in I$ such that*

$$p_i \leq \frac{rk_I(i)\alpha}{j(\alpha)}.$$

Proof. Suppose there exists some $i \in I$ with $p_i \leq \frac{\text{rk}_I(i)\alpha}{j(\alpha)}$. Given that $|I| > j(\alpha)$, from the definition of $j(\alpha)$ it follows that this hypothesis is also rejected in the closed testing procedure on level α. Given that $|I| \leq j(\alpha)$, let us consider all intersection hypotheses H_J with $J \supseteq I$. If all these intersection hypotheses can be rejected by a Simes' test on level α, we know that the closed testing procedure will reject H_I. Note that, for all intersection hypotheses H_J with $|J| > j(\alpha)$ we already know that they will be rejected by the closed testing procedure, so we can only look at hypotheses H_J with $|J| \leq j(\alpha)$. Given such an H_J, we know that $p_i \leq \frac{\text{rk}_J(i)\alpha}{j(\alpha)}$, since $\text{rk}_I(i) \leq \text{rk}_J(i)$. To reject H_J with a Simes' test on level α there must be a $p_{i'} \leq \frac{\text{rk}_J(i')\alpha}{|J|}$. We have:

$$p_i \leq \frac{\text{rk}_J(i)\alpha}{j(\alpha)} \leq \frac{\text{rk}_J(i)\alpha}{|J|},$$

because $|J| \leq j(\alpha)$, from which it follows that all hypotheses H_J with $J \supseteq I$ (thus including H_I) can be rejected by the closed testing procedure.
Now suppose there is no $i \in I$ with $p_i \leq \frac{\text{rk}_I(i)\alpha}{j(\alpha)}$. Note that this can only happen if $|I| \leq j(\alpha)$, otherwise H_I is rejected by the Simes' test on level α by definition of $j(\alpha)$ and therefore for some $i \in I$ we have that

$$p_i \leq \frac{\text{rk}_I(i)\alpha}{|I|} < \frac{\text{rk}_I(i)\alpha}{j(\alpha)}.$$

Given that $|I| \leq j(\alpha)$, we will show that there is an intersection hypothesis H_J with $J \supseteq I$ such that H_J cannot be rejected by a Simes' test, which implies that H_I cannot be rejected by the closed testing procedure. Consider the intersection hypothesis H_J, with $J \supseteq I$ and $|J| = j(\alpha)$, that is most difficult to reject. This hypothesis will contain, for some l, the l largest p-values $p_{n-l+1}, \ldots, p_n$ and the remaining $j(\alpha) - l$ p-values will all belong to I. The l largest p-values will not be smaller than their corresponding critical values since the hypothesis of size $j(\alpha)$, containing all $j(\alpha)$ largest p-values could not be rejected. H_J can thus only be rejected if one of the first $j(\alpha) - l$ p-values will be smaller than its corresponding critical value, but these are all p-values from I, for which

we already know that there is no $i \in I$ with $p_i \leq \frac{\text{rk}_I(i)\alpha}{j(\alpha)}$. Since $\text{rk}_I(i) = \text{rk}_J(i)$ for these first $j(\alpha) - l$ p-values we thus have that there is no $p_i \leq \frac{\text{rk}_J(i)\alpha}{j(\alpha)}$, which are exactly the critical values for an hypothesis of size $j(\alpha)$. H_J can thus not be rejected by a Simes' test on level α and because $J \supseteq I$, H_I cannot be rejected by the closed testing procedure. □

From Theorem 2.4.1 it follows that, after having calculated $j(\alpha)$ for the desired value of α, deciding whether H_I can be rejected by the closed testing procedure can be done by only looking at the p-values p_i, $i \in I$. The exact values of these p-values are not even important, only whether these values are smaller than their respective boundary values $\alpha/j(\alpha)$, $2\alpha/j(\alpha)$, etcetera. This observation motivates the following definition: the *category* of a p-value p_i is the smallest integer r_i such that $p_i \leq r_i\alpha/j(\alpha)$. Given this definition, Theorem 2.4.1 can be rephrased in terms of these categories which leads to the following procedure:

Procedure 2.4.1. *The closed testing procedure in combination with a Simes' test will reject an intersection hypothesis H_I on level α if and only if the hypothesis will be rejected by the following procedure:*

- *Calculate the size $j(\alpha)$ of the largest intersection hypothesis that cannot be rejected on this α-level by means of the algorithm discussed in the previous section.*
- *For each p_i with $i \in I$, determine the corresponding category value r_i, i.e. the smallest integer r_i such that $p_i \leq r_i\alpha/j(\alpha)$.*
- *Reject H_I if there is a positive integer r for which $\sum_{i \in I} \mathbb{1}\{r_i \leq r\} \geq r$.*

From this Procedure it follows that a p-value p_i with corresponding category $r_i = 1$ is in itself enough to result in the rejection of any hypothesis H_I, with $i \in I$. To see whether an elementary hypothesis H_i can be rejected within the closed testing procedure thus comes down to checking whether the corresponding p_i falls in the first category, i.e. whether $p_i \leq \alpha/j(\alpha)$, which is exactly the same as Hommel's method for the elementary hypotheses. If there are no p-values within H_I that fall in the first category, the question becomes whether there are minimally two p-values in the first two categories, or three in the first three, etcetera, as summarized by the statement that H_I can be rejected if and only if there is an $r \geq 1$ for which $\sum_{i \in I} \mathbb{1}\{r_i \leq r\} \geq r$.

The subsequent question is how we can benefit from Procedure 2.4.1 when we want to construct confidence intervals for the number of true or false hypotheses within a set $\{H_i : i \in S\}$. To construct these intervals, the quantity $t_\alpha(S)$ as given in (2.6) has to be computed, which is the size $|I|$ of the largest hypothesis H_I, $I \subseteq S$, that cannot be rejected by the closed testing procedure. Given Procedure 2.4.1, this is equivalent to finding the size of the largest index set $I \subseteq R$ for which there is no $r \geq 1$ with $\sum_{i \in I} \mathbb{1}\{r_i \leq r\} \geq r$. Based on that observation, we will show that

$$t_\alpha(S) = |S| - d(S), \tag{2.8}$$

with

$$d(S) = \max_{r \geq 1} \left\{ 1 + \sum_{i \in S} \mathbb{1}\{r_i \leq r\} - r \right\}. \tag{2.9}$$

Given the new procedure, H_S can be rejected if and only if $d(S) \geq 1$. In addition, we know that if the smallest $d(S) - 1$ elements are removed from S, for the remaining set S' we have $\sum_{i \in S'} \mathbb{1}\{r_i \leq r\} = r$ for the same r for which $d(S)$ was found, so $H_{S'}$ can still be rejected. If subsequently the smallest element $i \in S'$ would be removed from S', the hypothesis $H_{S''}$ corresponding to the new set S'' can no longer be rejected. The size of the largest set that cannot be rejected, i.e. $t_\alpha(S)$, is thus $|S| - d(S)$. Note that, since we also have that

$$t_\alpha(S) = |S| - f_\alpha(S),$$

where $f_\alpha(S)$ is the lowerbound for the $100(1-\alpha)\%$ confidence interval for the number of false hypotheses within the set S, we have an immediate formula for this lowerbound, namely

$$f_\alpha(S) = d(S).$$

Another remark is that, although we took a maximum over all positive integers r in equation (2.9), we already know that this maximum can never be attained for an $r > |S|$, since for such an r we have that

$$\sum_{i \in S} \mathbb{1}\{r_i \leq r\} - r \leq -1 \leq \sum_{i \in S} \mathbb{1}\{r_i \leq 1\} - 1.$$

Similarly, the maximum will never be attained for an $r > \max_{i \in S} r_i$ or for an $r > j(\alpha)$ because from the definition of $j(\alpha)$ we know that at most $j(\alpha)$ elements will fall in a category $r > j(\alpha)$, which means that for every $r > j(\alpha)$ we will always have

$$\sum_{i \in S} \mathbb{1}\{r_i \leq r\} - r \leq -1.$$

Therefore, instead of taking a maximum over $r \geq 1$ in equation (2.9) we can take this maximum over $r \in \{1, \ldots, k\}$ with

$$k = \min \left\{ |S|, j(\alpha), \max_{i \in S} r_i \right\}. \tag{2.10}$$

In the next subsection, we will describe an algorithm that, given a sequence of categories $(r_i)_{i \in S}$, finds the corresponding value of $d(S) = f_\alpha(S)$ from which the confidence set for the number of true or false hypotheses within the set $\{H_i \colon i \in S\}$ can be computed. In addition to finding the confidence interval for the complete set $\{H_i \colon i \in S\}$, the algorithm will also provide the $|S| - 1$ confidence sets for all subsets $\{H_i \colon i \in S_l\}$ where $S_l \subset S$ is the subset that contains the smallest l elements of S. In this way, many related confidence sets are computed. These confidence sets can be very valuable in the

situation in which the ordering of the hypotheses within S is not randomly chosen, but is chosen based on the relative importance of these hypotheses. If the ordering is for example based on increasing p-values, this approach answers the question how many false hypotheses, i.e. true findings, there are among the l hypotheses with the smallest p-values for all values of $l \leq |S|$. It also answers the question what the minimum size of a set of hypotheses should be if one wants to make sure that this set contains at least m false hypotheses. These are important questions if a validation experiment is planned and the number of hypotheses to follow up on is still undecided.

Because there are certainly orderings imaginable in which the importance of hypotheses does not directly correspond with the size of the corresponding p-values, for example if hypotheses are ordered on corresponding effect size or biological importance, we are interested in an algorithm that can find such confidence sets, even if the sequence $(r_i)_{i \in S}$ is not increasing in the values of the r_i's.

2.4.2 Algorithm

Given a (not necessarily increasing) sequence $\mathbf{r} = (r_1, r_2, \ldots, r_m)$ of categories, we wish to find an algorithm that calculates

$$f_l = \max_{r \in \{1,\ldots,k\}} g_l(r), \tag{2.11}$$

with k as in (2.10) and with

$$g_l(r) = 1 + \sum_{r_i \in \mathbf{r}_l} \mathbb{1}\{r_i \leq r\} - r$$

for all prefixes $\mathbf{r}_l = (r_1, r_2, \ldots, r_l)$ of length l of $\mathbf{r}$ in an efficient way. Of course, we can safely ignore any category bigger than k here, so we will assume categories are at most k. Calculating (2.11) naively involves determining $g_l(r)$ for all $l \in \{1, \ldots, m\}$ and $r \in \{1, \ldots, k\}$, which takes $\Theta(km)$ time by reusing earlier results via

$$g_l(r) = g_{l-1}(r) + \mathbb{1}\{r_l \leq r\}. \tag{2.12}$$

In order to achieve an algorithm that runs in $O(m)$ time, we need an additional observation with respect to (2.11). Suppose that $f_{l-1} = g_{l-1}(r)$ for some maximal r. From (2.12) we can also see that if $r_l \leq r$, then $f_l = g_l(r)$ as well. If on the other hand $r_l > r$, then we may or may not get that $f_l = g_l(r')$ for some maximal $r' > r$, depending on the subsequent maximum of $\big(g_l(r+1), \ldots, g_l(k)\big)$. Let us therefore consider the maxima over each tail of $\big(g_l(1), \ldots, g_l(k)\big)$, and denote these by

$$h_l(s) = \max_{r \in \{s,\ldots,k\}} g_l(r)$$

for $s \in \{1, \ldots, k\}$. In the spirit of (2.12), we will see how to derive h_l from h_{l-1}.

Clearly, we have that h_{l-1} is decreasing and piecewise constant, so the level sets of h_{l-1}, i.e. the parts on which h_{l-1} is constant, form consecutive nonempty subintervals of $\{1, \dots, k\}$. In particular, on each subinterval g_{l-1} takes its maximum at the rightmost endpoint, and the maxima of consecutive intervals decrease by at least 1. Now suppose that r_l is contained in such an interval $\{a, \dots, b\}$. We then conclude from (2.12) that

$$h_l(s) = h_{l-1}(s) + \mathbb{1}\{a \leq s\}. \tag{2.13}$$

Therefore, the level sets of h_l are the same as those of h_{l-1}, except with $\{a, \dots, b\}$ and its possible predecessor merged when the maxima of both intervals differ by 1.

In fact, the last condition is rather redundant as the maxima of consecutive intervals always differ by 1. Namely, initially this invariant holds as we have $h_0(s) = g_0(s) = 1-s$ for all $s \in \{1, \dots, k\}$, and (2.13) preserves it.

It is now straightforward to translate the above observation into a simple algorithm to calculate f_l for each $l \in \{1, \dots, m\}$, by keeping track of the level sets of h_l. For $l = 0$ we start with the initial partition

$$\pi_0 = \big(\{1\}, \{2\}, \dots, \{k\}\big).$$

When $l >= 1$, we update our partition π_{l-1} to π_l by picking the interval from π_{l-1} that contains r_l, and merging it with its predecessor (if any). Furthermore, we keep track of f_l as follows. Since $f_l = h_l(1)$, we just start with $f_0 = 0$ and increment f_l by 1 when r_l is contained in the first interval of the partition π_{l-1}, that is,

$$f_l = f_{l-1} + \mathbb{1}\{\pi_{l-1} = \pi_l\}.$$

See table 2.1 for an illustration of the algorithm on a small example.

l	r_l	Partition π_l	f_l	$g_l(1)$	$g_l(2)$	$g_l(3)$	$g_l(4)$	$g_l(5)$
0	-	$\{1\},\{2\},\{3\},\{4\},\{5\}$	0	0	-1	-2	-3	-4
1	4	$\{1\},\{2\},\{3,4\},\{5\}$	0	0	-1	-2	-2	-3
2	2	$\{1,2\},\{3,4\},\{5\}$	0	0	0	-1	-1	-2
3	3	$\{1,2,3,4\},\{5\}$	0	0	0	0	0	-1
4	2	$\{1,2,3,4\},\{5\}$	1	0	1	1	1	0
5	5	$\{1,2,3,4,5\}$	1	0	1	1	1	1
6	3	$\{1,2,3,4,5\}$	2	0	1	2	2	2

Table 2.1: A sample run of the algorithm to calculate f_l from (2.11) on a sequence of categories $\mathbf{r} = (4, 2, 3, 2, 5, 3)$ with $k = 5$, along with the values of g_l for comparison. In each step, either the partition is updated by merging the interval that contains r_l with its predecessor, or f_{l-1} is incremented by 1 if no such predecessor exists. Changes with respect to the previous step are marked in blue.

The algorithm we just described lends itself naturally for implementation with a disjoint-set data structure. Such a data structure allows for efficient queries to which set an element belongs to as well as for taking efficient set unions. Generally, the performance

of a disjoint-set data structure for a linear number of operations is not linear (Fredman and Saks, 1989) but nearly linear in time (Tarjan, 1975). In our specific case it is in fact possible to achieve linear time since we only need to merge adjacent intervals instead of arbitrary sets (Gabow and Tarjan, 1985). However, in practice the gain of the specific linear variant over the general nearly linear disjoint-set data structure is not noticeable so we opted for the latter, simpler implementation.

2.5 Application

In this section, we will use the previously described algorithms for analyzing a gene expression data set from the breast cancer study published by Schmidt et al. (2008). This data is freely available in the `R`-package `breastCancerMAINZ` (Schroeder et al., 2011). The data set contains gene expression data on 22283 features for 200 samples, where the gene expressions are measured by Affymetrix hgu133a technology. In our analysis, for each of the 22283 gene expression variables, we want to test the null-hypothesis that there is no association between this variable and the measured survival time, as given by

$$H_0 : \beta_i = 0$$

where β_i is the regression coefficient in a univariate Cox proportional hazards model for gene i, with $i \in \{1, \ldots, 22283\}$. To find p-values for these 22283 hypotheses, we used the likelihood ratio test from the `survival` package. Given that we have a matrix X of size 200×22283 containing normalized gene expression data and two vectors called $time$ and $event$ of size 200 containing respectively the survival times and the event indicators, we used the following `R`-code to calculate the p-values and the corresponding regression parameters:

```
> n <- ncol(X)
> pvalues <- rep(0,n)
> betas <- rep(0,n)
> for(i in 1:n)
+ {
+   fit <- coxph(Surv(time,event) ~ X[,i])
+   pvalues[i] <- summary(fit)$logtest[3]
+   betas[i] <- abs(summary(fit)$coef[1])
+ }
```

To find adjusted p-values, we can use the `hommelFast` function from the `cherry` package, that will return an object that, among others, contains the original p-values and the adjusted p-values corresponding to Hommel's procedure.

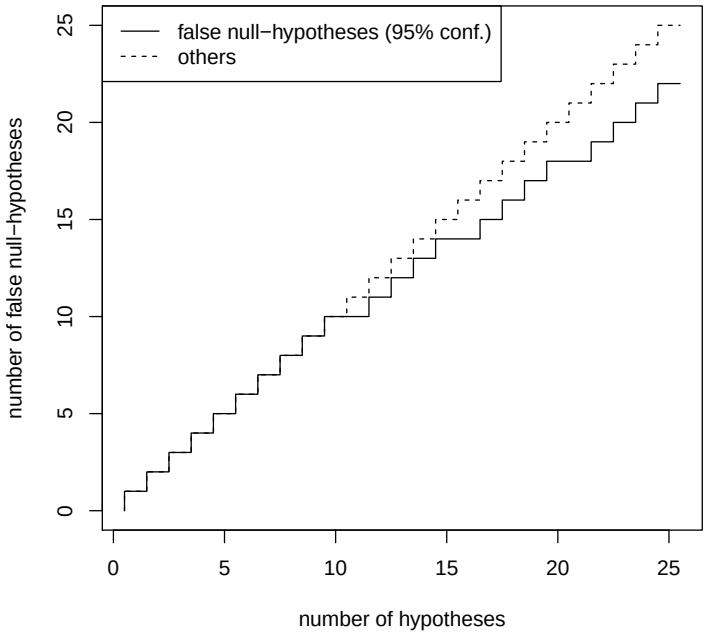
```
> hom <- hommelFast(pvalues)
```

Given this object, we can calculate the number of null-hypotheses that could be rejected on $\alpha = 0.05$ by directly looking at the adjusted p-values. We could also estimate the lower bound of the 95% confidence set of the number of false null-hypotheses, or in other words true findings, by using the `pickSimes` function from the `cherry` package.

```
> sum(pvalue(hom)<=0.05)

[1] 10

> pickSimes(hom,alpha=0.05)

22283 hypotheses selected. At confidence level 0.95:
False null-hypotheses >= 114; True null-hypotheses <= 22169.
```

We see that, while only 10 null-hypotheses could be rejected based on their adjusted p-values, the total number of false null-hypotheses will, with probability at least $1 - \alpha$, be at least 114 . An interesting question could be how many false null-hypotheses there will be among the null-hypotheses with the smallest m p-values. We will answer this question for $m = 114$ and $m = 25$. Our software can either give a lower bound for the number of false hypotheses within a set, where the specific order of hypotheses within this set does not matter, as provided by the function `pickSimes` or it can use the exact order of the hypotheses within this set, and give lower bounds for each subset containing the first l elements, as discussed in the previous section. The function that does this is called `curveSimes`.

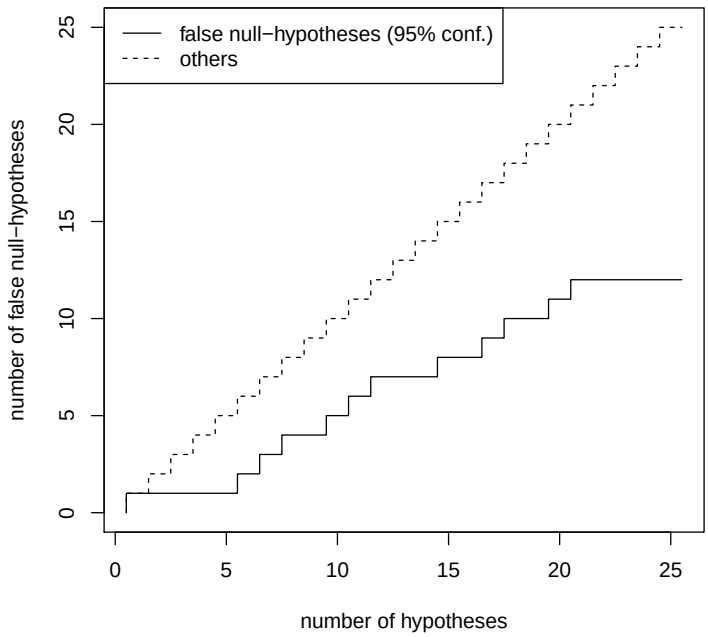
```
> perm <- order(pvalues)
> pickSimes(hom,select = perm[1:114], alpha=0.05)

114 hypotheses selected. At confidence level 0.95:
False null-hypotheses >= 77; True null-hypotheses <= 37.

> curveSimes(hom, order = perm[1:25], alpha=0.05, plot=T)
```



We see that among the 114 null-hypotheses with the smallest p-values, 77 will be false null-hypotheses. From the plot, we see that the first 10 null-hypotheses are, with 95% confidence, false null-hypotheses, whereas for the remaining 15, 12 of them will be false.

We could repeat the same steps for null-hypotheses that are not ordered based on p-value, but on effect size, which can be measured as the absolute values of the regression coefficients.

```
> perm2 <- order(betas,decreasing=T)
> pickSimes(hom,select = perm2[1:114], alpha=0.05)

114 hypotheses selected. At confidence level 0.95:
False null-hypotheses >= 40; True null-hypotheses <= 74.

> curveSimes(hom, order = perm2[1:25], alpha=0.05, plot=T)
```



We see that among the first 114 hypotheses, ordered on decreasing absolute value of the regression coefficient in the Cox model, at least 40 are false null-hypotheses. Among the first 25, this lower bound on the number of false hypotheses equals 12.

Instead of constructing sets of hypotheses based on own criteria, we could also look at sets of hypotheses that correspond to known gene sets, obtained from the Gene Ontology (GO) (Ashburner et al., 2000). As an example we will look at the set of genes that are known for their involvement in apoptosis which is the process of programmed cell death. From the 22283 variables in our data set, 2588 are associated with apoptosis. If we look at the set of corresponding hypotheses, we see that at least 2 hypotheses are false, which means that at least 2 genes within this set will be associated with the survival time.

```
> xx <- as.list(hgu133aGO2ALLPROBES)
> apoptosis <- xx["GO:0006915"]
> pickSimes(hom,select=colnames(X) %in% unlist(apoptosis),
+ alpha=0.05)

2588 hypotheses selected. At confidence level 0.95:
False null-hypotheses >= 2; True null-hypotheses <= 2586.
```

Although we used our own algorithm to calculate the adjusted p-values of Hommel's procedure, as given by the `hommelFast` function, we could also have used the implementation that is available in the `p.adjust` function. To give an idea of the time differences between the two procedures, we simulated 10000 as well as 100000 p-values and used both functions to calculate the corresponding adjusted p-values:

```
> set.seed(1)
> n=10000
> pvalues <- c(runif(0.25*n,0,0.001), runif(0.75*n,0,1))
> system.time(a <- hommelFast(pvalues))

   user  system elapsed
   0.22    0.00    0.22

> system.time(b <- p.adjust(pvalues,method="hommel"))

   user  system elapsed
   6.18    0.00    6.18

> all.equal(pvalue(a),b)

[1] TRUE

> set.seed(1)
> n=100000
> pvalues <- c(runif(0.25*n,0,0.001), runif(0.75*n,0,1))
> system.time(a <- hommelFast(pvalues))

   user  system elapsed
   1.92    0.00    1.91

> system.time(b <- p.adjust(pvalues,method="hommel"))

   user  system elapsed
 588.75    2.24  591.23

> all.equal(pvalue(a),b)

[1] TRUE
```

From this small simulation experiment, it is clear that our new algorithm is the preferred one when the number of null-hypotheses is very large.

2.6 Discussion

In this article, we have shown that Hommel's FWER controlling multiple testing procedure for n elementary hypotheses $H_1, \ldots, H_n$ can be carried out in $O(n \log(n))$ time, which is much faster than the currently available implementation that runs in $\Theta(n^2)$ time. We have also shown that there is an easy procedure to determine whether an arbitrarily chosen intersection hypothesis $H_I = \bigcap_{i \in I} H_i$ can be rejected within the closed testing procedure with a Simes' test as local test. Based on this procedure we developed an algorithm that calculates confidence sets for the number of false (or true) hypotheses within an arbitrarily chosen set $S = \{H_i \colon i \in S\}$.

Throughout the article, we only considered the Simes' test as local test in the closed testing procedure. After applying some minor modifications, our algorithms to find adjusted p-values and confidence sets will however work in a more general setting. An example of such a situation is the situation in which the test proposed by Hommel (1983) is used as a local test in the closed testing procedure. In this test, the critical value for a p-value p_i with rank $\mathrm{rk}_I(i)$ within an intersection hypothesis H_I of size $|I| = s$ is given by

$$\frac{\mathrm{rk}_I(i)\alpha}{C_s \cdot s}, \tag{2.14}$$

where $C_s = \sum_{v=1}^{s} v^{-1}$, as opposed to the critical value $\mathrm{rk}_I(i)\alpha/s$ of the Simes' test. The resulting test is more conservative than the Simes' test, but is valid under any form of dependence of the p-values. Our software, as available in the CRAN package `cherry`, can be used both with a Simes' test as local test and with the just described test of Hommel.

In general, our algorithm to calculate the function j, i.e. the function that gives the size of the largest intersection hypothesis that cannot be rejected on a given α-level within the closed testing procedure, can be used for every local test that has critical values of the form given in (2.14) as long as C_s only depends on the size s of the corresponding intersection hypothesis. The procedure to quickly determine whether H_I can be rejected within the closed testing procedure, as given in Theorem 2.4.1 and the subsequent algorithm to create confidence sets can be used as long as C_s only depends on s and is (weakly) increasing in s.