Flexible Multiple Testing with the FACT Algorithm

Edgar Dobriban*

July 21, 2018

Abstract

Modern high-throughput science often leads to multiple testing problems: researchers test many hypotheses, wishing to find the significant discoveries. The development of flexible multiple testing methods is thus a central problem in statistics. In this paper, we introduce the new *Fast Closed Testing (FACT)* method for multiple testing, controlling the family-wise error rate. Our method relies on symmetry and monotonicity to enable the classical closed testing principle in the important setting of large datasets. As the closed testing principle is more than 40 years old, we find it surprising that this simple and fundamental algorithm has not been described before. Our FACT method is general and flexible, and can be used to design powerful new architectures for multiple testing. We showcase it by proposing the *Simes-Higher Criticism* fusion test, which is powerful for detecting both a few strong signals, and also many moderate signals. We illustrate the method in simulations and in a genome-wide association study of coronary artery disease, and obtain more power than with existing methods.

1 Introduction

Understanding how to make multiple decisions, with guarantees on the number of errors, is an important problem in statistics and data science. Here we study this problem, proposing a general framework for constructing multiple tests that is flexible, powerful, and computationally efficient.

1.1 Large-scale inference and multiple testing

In many areas of science, researchers test candidate hypotheses, wishing to find the significant discoveries. In such a multiple testing problem, we have several hypotheses,

---

*Department of Statistics, The Wharton School, University of Pennsylvania. E-mail: `dobriban@wharton.upenn.edu`. The author thanks Emmanuel Candes, from whose course at Stanford University he learned about the closure principle, and Jelle Goeman, Jiangtao Gou, and Ajit Tamhane for valuable comments on an earlier draft of the manuscript. He also thanks Nancy Zhang, William Leeb, Jingshu Wang, and Min Xu for helpful discussions.

for instance about the association between explanatory variables and an observed outcome. Based on the observed data, we want to discover the variables truly associated with the outcome, while controlling the erroneous discoveries.

The classical error criterion to control in multiple testing is the family-wise error rate (FWER), which is the probability of any erroneous discoveries. Family-wise error rate control requires that the probability of making *any false positives* is small, which is very appealing in critical application areas where mistakes are expensive, such as human deaths due to inadequate medical treatments, or large financial losses due to mistaken investment decisions.

While false discovery rate (FDR) control is preferred in some application areas, FWER control remains the standard in high-stakes areas where more rigorous error control is desired, e.g. end-stage genome-wide association studies (Sham and Purcell, 2014), neuroimaging (Eklund et al., 2016), as well as medical and pharmaceutical applications (Dmitrienko et al., 2009). FWER control leads to much stronger reproducibility for the individual discoveries.

1.2 Closed testing

The *closure principle* is an important general framework for constructing multiple tests controlling the FWER (Marcus et al., 1976). Suppose that we have individual null hypotheses $H_1, H_2, \ldots, H_n$, and that we want to control the FWER at level α. For instance, H_i may correspond to the null hypothesis that the i-th explanatory variable is not associated with the outcome.

In the closure principle, we start with valid level α *local testing rules* for all nulls that are intersections of subsets of individual null hypotheses. An intersection null H_J states that all of the hypotheses $i \in J$ are null. In closed testing, we reject the individual null H_i if, for all sets J such that $i \in J$, the intersection null H_J is rejected. To argue that we have an individual discovery, we must be able to "see" this discovery in all groups of individual hypotheses. This method controls FWER strongly, under any configuration of true and false individual null hypotheses (Marcus et al., 1976).

Clearly, this method is computationally intractable in general, because it requires testing all $2^n - 1$ intersection nulls, corresponding to all nonempty subsets of hypotheses. In practice, popular computationally efficient *shortcuts* are used in special cases. For instance, the important Holm's procedure (Holm, 1979)—used in tens of thousands of scientific studies—is the closure of Bonferroni's method, and Hommel's procedure (Hommel, 1988) is the closure of Simes' method (Simes, 1986). Hochberg's procedure Hochberg (1988) is a simple method that is strictly more conservative than Hommel's procedure. Liu's method Liu (1996) gives conditions under which the closure principle can be reduced to step-up or step-down methods.

However, these uses of the closed testing principle are very limited, and they do not allow the user to design multiple testing methods for specific applications. For instance, it would be desirable to choose local testing rules that can use information in an adaptive way. If there are only a small number of large effects, we should use a local testing rule that only looks at the smallest p-values, for instance Bonferroni's or Simes' method. If there are many moderately small effects, then we should use a local testing rule like that uses many of the small p-values, like the chi-squared or Higher

Criticism tests. There are currently no known computationally efficient methods that achieve this.

To broaden the methods available for multiple testing, in this paper we introduce *Fast Closed Testing (FACT)*, a new computationally efficient framework for multiple testing based on the closure principle. Our algorithm allows the user to design powerful multiple testing architectures for specific applications. The key requirement on the problems where it can be used is the *symmetry* of the hypotheses – that is, we treat the hypotheses as if there was little or no distinction between them. This is a reasonable first order approximation in many applications. The use of prior information can be difficult to quantify and can bias results; see below for more discussion. Therefore, symmetry is often a good assumption even in the presence of some prior information.

The closed testing principle is more than 40 years old. While there is a lot of related work (see Section 2.3), it appears that the FACT algorithm has not been reported before. We found this surprising, and closing this gap, along with deepening our understanding of multiple testing, were our main motivations for writing the paper.

More specifically, our contributions and the structure of the paper are as follows:

1. We introduce the new Fast Closed Testing (FACT) method for multiple testing (Algorithm 1 and Theorem 2.1). This requires using monotone symmetric rules for testing subsets of hypotheses. We also give an algorithm for computing adjusted p-values for FACT.

2. As an important example, we show the flexibility of the FACT algorithm by designing tests that are powerful against both sparse and dense alternatives (Section 2.2). Specifically, we propose the *Simes-Higher Criticism* fusion rule, which uses Simes' method for testing intersection nulls of a small size, and the Higher Criticism for intersection nulls of a large size (Algorithm 3). We also show how to extend the method to truly massive datasets by subgroup splitting.

3. We show how to construct monotone symmetric rules using appropriate test statistics (Section 3). We first show that the class of such statistics has several invariance properties (Lemma 3.1), and then give an extensive list of examples (Section 4). The examples include Bonferroni-type rules, Simes-type rules, monotone sums, and monotone functions of order statistics (such as the higher criticism).

4. We discuss the notion of consonance, showing that the FACT method is not always consonant, and hence cannot be deduced from the known fast algorithms for consonant methods (Section 5).

5. We perform numerical simulations supporting our theoretical results (Section 6). We also illustrate our method in a Genome-Wide Association Study (GWAS) of coronary artery disease, where it leads to stronger association evidence for secondary loci (Section 7).

Software implementing our method, and reproducing our results, is available from the author's GitHub page, `http://github.com/dobriban/FACT`.

2 Fast closed testing

More formally, let p_i, $i = 1, \ldots, n$ be the p-values for the individual null hypotheses H_i. The p-values are assumed to be uniform random variables whenever the null hypotheses H_i are true. Depending on the context, we will need specific assumptions on the joint distribution of the p-values (these will range from no additional assumptions to joint independence).

The model where only p-values are available is important and very general, because it encompasses many practical problems. In particular, if we have test statistics for testing individual nulls, with known null distributions, then we can reduce the problem to p-values with the appropriate transformation. The "p-values only" model is suitable for challenging statistical problems where we may not have enough data to estimate the individual effects accurately.

Denote by p_J the vector of p-values with indices in the set $J \subset \{1, 2, \ldots, n\}$, and consider a local testing rule for the intersection null

$$H_J = \cap_{i \in J} H_i,$$

with decision function

$$\Phi(p_J) : [0,1]^J \to \{0,1\}.$$

The definition of the decision function is that the intersection null H_J is rejected based on the p-values p_J if $\Phi(p_J) = 1$.

The closed testing (CT) method (Marcus et al., 1976) defines a multiple testing procedure for the individual null hypotheses H_i based on the local testing rules. In words, we reject H_i if, for all sets J such that $i \in J$, the null H_J is rejected. Formally, the CT decision rule Φ_c for the i-th null is

$$\Phi_c(p_i) = \prod_{J : i \in J} \Phi(p_J).$$

Some of the null hypotheses H_i are true, while some are false, but we do not know which ones. An intersection null H_J is true if and only if all of the individual hypotheses H_i with $i \in J$ are true. The local testing rules are said to have level α if the probability of rejecting H_J when it is true is at most α. The family-wise error rate (FWER) of a multiple testing rule is defined as the probability of rejecting any individual null hypothesis H_i when it is true. The CT method controls the FWER, if each local testing rule has level α (see Section 9.1 for the well-known argument).

To implement CT, we need to find the indices i for which all parent subsets are rejected. In general, we would need to check each of the 2^{n-1} subsets containing an index. However, this can simplify under some conditions, as there can be many sets that we do not need to check.

We assume that the problem has a global symmetry structure, meaning that we treat the hypotheses identically. In that case, it makes sense to use the same local testing rule T_j for each subset of a fixed size j. Suppose also that the local testing rules satisfy the following two conditions:

1. **Monotonicity**: If we reject the intersection null and decrease any p-value, we still reject it. If for all coordinates i, $p_i \leq p'_i$, then

$$\Phi(p_J) \geq \Phi(p'_J).$$

2. **Symmetry**: Rejection only depends on the set of p-values, and not on the labels of their null indices. For any permutation π of the indices J,

$$\Phi(p_J) = \Phi\left(p_{\pi(J)}\right).$$

Under these conditions only the sizes of the p-values—and not their indices—matter when determining which hypotheses get rejected. Moreover, smaller p-values are always better. This implies that *some group of the smallest p-values* will be rejected.

Algorithm 1 Fast Closed Testing (FACT)

```
1: input: $p$-values $p_i$, $i = 1, \ldots, n$. local testing rules $T_k$ for testing subsets of size $k$.
   Desired FWER $\alpha$
2: Sort the $p$-values: $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(n)}$
3: for $k = 1, \ldots, n+1$ do
4:     for $j = k, \ldots, n$ do
5:         if the null with $p$-values $p_{(k)}, p_{(j+1)}, p_{(j+2)}, \ldots, p_{(n)}$ is not rejected
6:             using local testing rule $T_{n-j+1}$ then
7:              go to line (8)
8: return: Reject the hypotheses corresponding to the $p$-values $p_{(1)}, \ldots, p_{(k-1)}$
```

It remains to find this group. Intuitively, we should start with the smallest p-value, and check if the "hardest" subset of every given size containing it will be rejected. We show that this intuition is in fact correct, and we give an equivalent sequential algorithm relying on two *for loops*. We call this the Fast Closed Testing (FACT) algorithm (see Alg. 1).

In this algorithm, we make the following conventions: The indices in the loops run until $n+1$, but the p-value $p_{(n+1)}$ is taken to be the empty set. That is, this element is not included in the list of p-values to be tested. Moreover, the p-values $p_{(j+1)}, \ldots, p_{(n)}$ are also taken to be the empty set when $j = n$, i.e., they are not included in the list. Finally, when $k = n+1$ and $j = n$, the above conventions specify that the p-values to be tested are the empty set, i.e., the list is empty. In this case, we make the convention that this set is always rejected.

Our first main result shows that the FACT algorithm controls the FWER, and characterizes the running time of the algorithm. See Section 9.2 for the proof.

Theorem 2.1 (Correctness of the FACT algorithm)**.** *Consider the FACT algorithm from Alg. 1. Suppose the intersection null hypotheses are tested with* monotone symmetric *testing rules T_j with level α, using the same rule for all sets with the same size j. Then:*

1. *The FACT algorithm controls the family-wise error rate (FWER) at level α.*

2. *Moreover, suppose that applying the local testing rule T_j to j hypotheses takes linear time $O(j)$. Then, the FACT algorithm takes $O(sn^2)$ time, where s is the number of significant discoveries (rejections).*

The proof also shows that FACT leads *exactly* to the same rejections as closed testing.

The significance of the FACT method is that it enables using the closed testing framework in a much more flexible way than what was known before. It is possible to design closed testing architectures based on the FACT algorithm for many specific applications. Before the FACT method, the applications of closed testing were limited to either very special cases like Holm's method, or were computationally intractable. We show later that the FACT algorithm recovers several well-known shortcuts for closed testing (Section 4). We emphasize again that we find it surprising that such a simple algorithm for fast closed testing has not been reported before.

By choosing the local testing rules appropriately for specific applications, practitioners can design many powerful new closed testing architectures. In the next section we will illustrate the steps in a specific application. In later sections, we will give a broader set of examples. We also point out that the error control for FACT holds under the same conditions as for the local testing rules used.

Algorithm 2 Adjusted p-values for FACT

```
input: p-values $p_i$, $i = 1, \dots, n$. local testing rules $T_k$ for testing subsets of size $k$.
    Methods for evaluating p-values for testing intersection nulls with $T_k$.
Sort the p-values: $p_{(1)} \le p_{(2)} \le \dots \le p_{(n)}$
for $k = 1, \dots, n+1$ do
    for $j = k, \dots, n$ do
        Let $q_{kj}$ be the p-value for the null with p-values $p_{(k)}, p_{(j+1)}, p_{(j+2)}, \dots, p_{(n)}$
            using local testing rule $T_{n-j+1}$
    Let $\tilde{p}_{(k)} = \max_{j:j \ge k} q_{kj}$
return: Adjusted p-values $\tilde{p}_{(1)}, \dots, \tilde{p}_{(n)}$
```

It is often of interest to compute adjusted p-values for a multiple testing procedure. An adjusted p-value of a hypothesis H_i is the smallest critical value at which the hypothesis is rejected. To complement FACT, we also show how to compute adjusted p-values for the method in Algorithm 2. This algorithm requires as input methods for evaluating p-values for the local testing rules T_k. Its computational complexity is $O(n^3)$ for testing rules for which finding critical values takes linear time. While computing the adjusted p-values is based on the same idea as the FACT algorithm itself, we present this method separately, because in principle it can be run independently from FACT, and because it takes time $O(n^3)$ instead of $O(sn^2)$.

2.1 Simes-Higher Criticism (Simes-HC) fusion rule

The greatest flexibility of the FACT algorithm is achieved by using different local testing rules for intersection nulls of different subset sizes. We will show here that this enables the design of tests that can be powerful against both sparse and dense alternatives.

Suppose that we are in a setting where we have n total p-values and the model is s-sparse, in the sense that there are s false nulls; or equivalently nonzero effect sizes. If we are testing an intersection null of size j containing a specific false null, then in the worst case there are

$$A = \max(j - [n-s], 1)$$

false nulls in this set. Now, the closure principle must reject each subset of size j containing the hypothesis in order to reject the individual null. Therefore, to maximize the chances that we reject this false null, we should use only the smallest A p-values out of the total of j p-values in the test statistic T_j.

The above reasoning gives a heuristic for the maximal number of non-nulls that each local testing rule can use. When testing subsets of a small size, we should use only a small number, and possibly only one p-value. This suggests that we should use the Bonferroni or the Simes rules for small subsets. Specifically, given a guess s for the sparsity, in this section we propose to use the Simes rule for subsets of size at most $n-s+1$. The reason why we use Simes is that it is strictly more powerful than the Bonferroni method, and it is known to work under general correlation structures (see Section 4.2).

Given an intersection null of size j, and p-values p_i for this null, Simes rule sorts them, and rejects the intersection null if, for any p-value, $p_{(i)}/i$ is at most α/j (Simes, 1986). Formally, it rejects if

$$\min_i \frac{p_{(i)}}{i} \le \frac{\alpha}{j}.$$

The Simes test is more powerful than the Bonferroni test, and has correct level under a broad class of positive dependence structures (see e.g., Goeman and Solari, 2014; Tamhane and Gou, 2018, for details). We will see later that both the Bonferroni and Simes rules are symmetric and monotone, so they can be used with FACT (Section 4.1).

When the subset size is large, we should use local testing rules that are powerful against relatively denser alternatives. Given that the Higher Criticism (HC) rule for local testing is known to be effective against many types of alternatives (Donoho and Jin, 2004), we propose to use it for large subsets. The Higher Criticism for testing a global null based on j p-values works as follows. For a fixed critical value β, under the global null, the number of p-values less than β follows a binomial distribution with j trials and success probability β. Therefore, we can compute the fraction f_β of p-values below β, standardize it by its standard deviation, and obtain the test statistic

$$C(\beta) = \sqrt{j}\frac{f_\beta - \beta}{\sqrt{\beta(1-\beta)}}.$$

If this test statistic is large, then the fraction of p-values below β is large, suggesting evidence against the global null. In order to be adaptive to the number of nonzero effects, the HC test takes the largest of these statistics over a range of β-s. This ensures that we can detect both a few large effect sizes, as well as a larger number of moderate effect sizes (see Donoho and Jin (2004) and Section 4.1). We will see that the HC rule is symmetric and monotone, and so it can be used with FACT (Section 4.1).

This leads to the **Simes-Higher Criticism** (Simes-HC) fusion rule, which is summarized in Algorithm 3. We will use this algorithm in our simulations and data analysis example.

Algorithm 3 Simes-Higher Criticism (Simes-HC) fusion for FACT

```
1: input: number of nulls $n$, preliminary estimate for sparsity $s$.
2: For $j \leq n - s + 1$, choose $T_j$ to be Simes' rule
3: For $j > n - s + 1$, choose $T_j$ to be the Higher Criticism rule
4: Run FACT with this choice of local testing rules
```

2.2 Scaling to massive data by subgroup splitting

In this section, we show how to scale the FACT method to massive data by subgroup splitting. If we have an extremely large number of hypotheses, it may be too conservative to control the FWER, and the computational cost of FACT may be prohibitive. However, we can overcome these difficulties if we can naturally group the hypotheses into subgroups (or sub-families), each of which has the desired symmetry properties outlined above. Then we can run FACT on each subfamily, and report the merged list of significant discoveries. This is described in detail below.

Suppose we have a collection of hypotheses H_i^j, which can be grouped in sub-families $F_j = \{H_i^j\}_i$, $j = 1, 2, \ldots, m$. Each sub-family contains n_j null hypotheses, $i = 1, \ldots, n_j$. To perform multiple testing, we propose to apply the FACT method to each sub-family separately, with appropriate local testing rules T_k^j, which may differ across families. For each family, we choose the local testing rules to control the FWER for the individual family, each at level α_j. We then report the overall list of discoveries.

Importantly, this method controls the overall FWER at $\alpha = \sum_j \alpha_j$. This follows by the union bound, because the probability of a false positive overall is at most the sum of the probabilities of false positives in each set.[1]

Moreover, the computational complexity of the subgroup splitting algorithm is $\sum_j n_j^2 s_j$, which can be much smaller than the complexity for running the FACT algorithm on all hypotheses, $n^2 s = (\sum_j n_j)^2 s$. This shows that the splitting algorithm can scale to massive data.

2.3 Some related work

Here we review some of the most closely related work in the literature. For broader reviews of multiple testing, see e.g., Hochberg and Tamhane (1987); Goeman and Solari (2014); Bretz et al. (2016). See also Tamhane and Gou (2018) for a review of p-value based methods focusing on more recent methods. The relationship between closure and

---

[1] We thank Ajit Tamhane and Jiangtao Gou for pointing out this result for FWER control.

computationally efficient algorithms has been studied from various perspectives. For instance, Grechanovsky and Hochberg (1999) establish conditions when a closure has a step-down sequentially rejective shortcut. Our work has a broader scope, because step-down procedures are a very special case of efficient algorithms. The FACT method is often not a step-down algorithm. Gou et al. (2014) propose improved hybrid Hochberg–Hommel type step-up multiple test procedures, which are also different from the FACT Simes-HC hybrid.

A different line of work aims to develop *interpretable* closed testing methods using graphical approaches (Bretz et al., 2009). This is important because these methods can be easily explained to practitioners. However, from a methodological and computational point of view, it is also important to develop new powerful methods such as those in our work.

Monotonicity ideas have appeared in the literature on multiple testing. A related monotonicity condition appeared in Birnbaum (1954), however, it was used for a completely different purpose than in our work. Indeed, there it was used as a condition under which meta-analysis methods are optimal. A monotonicity of the resulting closed test has been discussed in (Dmitrienko et al., 2009, Sec. 2.3.4). However, this is a *global monotonicity* condition, different from ours, because it applies to the overall multiple testing procedure, as opposed to the local tests. In particular, our methods are always monotone in the global sense. Hommel and Bretz (2008) discuss several monotonicity requirements for tests, and mention the present notion too (See Sec. 3.2 "Monotonicity within the same hypothesis"). However, we also develop explicit algorithms based on this condition.

Monotonicity has also appeared as a condition for error control in the sequential testing principle of Goeman and Solari (2010). However, the algorithms presented there are even more general than the closed testing principle, and thus not always computable in polynomial time. A similar observation about computationally efficient closed tests was made by Henning and Westfall (2015) (see their Section 3), who also noticed that for deciding whether or not to reject H_i, one must identify the "hurdle" subset for each subset size. However, it appears that they did not explicitly describe an algorithm to do so in full detail.

Studying a different problem, that of constructing a confidence statement on the number of false rejections incurred, Goeman and Solari (2011) also construct shortcuts for exchangeable local tests, such as the Fisher's test (see their Section 4 and Appendix A). While they are based on the same principle, looking at the worst case set at each level, the two algorithms are different.

3 Constructing monotone symmetric rules

How should we construct monotone symmetric rules? In this section we discuss some general principles. Suppose we use a test-statistic based rule, where we compute some test statistic $T = T(p)$ based on the p-values, and reject if this test statistic is less than some critical value c_α. Formally, the rejection rule has the form

$$\Phi(p) = I(T(p) \leq c_\alpha).$$

When does such a test become monotone and symmetric? It is easy to see that this will hold if the test statistic T itself is also monotone increasing and symmetric, in the following sense:

1. **Monotonicity**: If we decrease any p-value, the value of the test statistic T decreases. Formally, $T(p) \le T(p')$ if on all coordinates i, $p_i \le p_i'$.
2. **Symmetry**: The test statistic T only depends on the set of p-values, and not on their indices. Formally, $T(p) = T(p_\pi)$ for any permutation π.

It turns out that moving from rejection rules to test statistics is valuable, because one can naturally combine such test statistics in various ways. This is useful, because one can start with simple test statistics, and combine them to design more powerful testing rules. The following result summarizes the operations that one can take. See Section 9.3 for the proof.

Lemma 3.1 (Constructing monotone symmetric test statistics)**.** *Let $T^i, i \in I$ be any collection of monotone symmetric test statistics. Then, one can construct new monotone symmetric test statistics by taking:*

1. *Minima: $\min_{i\in I} T^i$ is monotone symmetric.*
2. *Maxima: $\max_{i\in I} T^i$ is monotone symmetric.*
3. *Non-negative linear combinations: $\sum_{i=1}^k \lambda_i T^i$ is monotone symmetric for any set of $\lambda_i \ge 0$.*
4. *Monotone functions: $g(T^1, \ldots, T^k)$ is monotone symmetric if the function g is coordinate-wise monotone, in the sense that $g(x_1, \ldots, x_k) \le g(x_1', \ldots, x_k')$ if on all coordinates i, $x_i \le x_i'$.*

An important example that we will use several times in the future are the order statistics. See Section 9.4 for the proof.

Lemma 3.2 (Order statistics)**.** *Any order statistic $T(p) = p_{(i)}$ is monotone symmetric.*

In the next section, we will discuss several examples of monotone symmetric rules.

4 Examples

In this section, we will give several examples of monotone symmetric rules. We will discuss both known examples (such as Bonferroni), as well as examples whose use in closed testing is new (such as the Higher Criticism). We consider test-statistic based local rules. Suppose that the local testing rule for subset J has the form

$$\Phi(p_J) = I\left(T(p_J) \le c_{|J|,\alpha}\right).$$

Here $I(E)$ is the indicator function of an event E, so that $I(E) = 1$ if E occurs, and $I(E) = 0$ otherwise. Also, T is a collection of testing rules T_j for each subset size $j = 1, 2, \ldots, n$, but for simplicity we omit the subscript. If T is monotone symmetric for all j, then such a local testing rule becomes monotone and symmetric.

4.1 Bonferroni-type rules

The simplest class of monotone symmetric rules is the **Bonferroni**-type rules:

$$T(p_J) = \min_{j \in J} p_j.$$

The critical values are $c_{|J|,\alpha} = \alpha/|J|$. The Bonferroni rule controls the type I error under any dependence structure. This rule is clearly symmetric and monotone. This can also be seen by Lemma 3.2, because it is based on the smallest order statistic of the p-values. It is well known that the closure principle applied to the Bonferroni test becomes Holm's procedure (Holm, 1979).

Moreover, we show that the FACT algorithm reduces to Holm's procedure. See Section 9.5 for the proof. This result is important because it shows that our FACT algorithm is sensible, as it recovers the most well-known example of an efficient closed testing algorithm.

Proposition 4.1. *The FACT algorithm for closing the Bonferroni method leads to the same rejections as Holm's procedure.*

The Bonferroni-type rules can also be written equivalently in terms of maxima of test statistics, which are known as **max-t procedures**. Suppose we have test statistics t_i for testing the hypotheses H_i, such that large t_i are indicative of a non-null H_i. Any p-value based multiple testing procedure can be represented equivalently in terms of the test statistics t_i. This representation is especially popular in contexts where the null distribution of the test statistics is not known. In such a setting, resampling methods can be used under some conditions to obtain valid critical values (e.g., Westfall and Young, 1993).

A popular method, sometimes called the max-t method, is to use $T_J = \max_{j \in J} t_j$ as the test statistic for testing the null H_J. The subset pivotality condition (Westfall and Young, 1993) ensures the symmetry of the local tests. Thus, the FACT algorithm is applicable, and recovers the well-known step-down max-t method (Westfall and Young, 1993). From an algorithmic perspective, the resulting closed testing procedure is equivalent to the Holm step-down method with a new set of critical values.

4.2 Simes-type rules

A second class of monotone symmetric rules are **Simes**-type rules, which are based on comparing the order statistics of the p-values to increasing thresholds:

$$\Phi(p_J) = I\left(\min_i \frac{p_{(i)}}{i} \le \frac{\alpha}{|J|}\right).$$

Therefore, $T(p_J) = \min_{i \in J} p_{(i)}/i$, and $c_{|J|,\alpha} = \alpha/|J|$ (Simes, 1986). The Simes test is more powerful than the Bonferroni test. Moreover, Simes' test has exact level α for the intersection null under independence (Simes, 1986). Under many types of positive dependence, Simes is conservative (Samuel-Cahn, 1996; Sarkar and Chang, 1997), and the known situations in which it is anti-conservative occur under quite pathological negative dependence structures (Rødland, 2006). See also Goeman and Solari (2014); Tamhane and Gou (2018).

As a first remark, we observe that Simes' method is symmetric and monotone by inspection. More formally, this holds because we are taking a minimum of non-negative scalings of order statistics. The order statistics are monotone symmetric by Lemma 3.2, so we can apply the rules for constructing monotone symmetric test statistics (Lemma 3.1).

Proposition 4.2. *Simes' method is symmetric and monotone.*

The next problem is to apply closed testing to the Simes test. Hommel (1988) gives the following algorithm for the closure of Simes: Let j be the largest index such that

$$p_{(n-j+k)} > k\alpha/j$$

for all $k = 1, \dots, j$. If j does not exist, reject all H_i. Otherwise, reject all H_i with $p_i \le \alpha/j$. This algorithm takes $O(n^2)$ in the worst case.

The FACT method, as given in Algorithm 1, takes $O(n^3)$ in the worst case, and thus clearly its steps do not agree with Hommel's method. However, we now show that there is a simplification, and the FACT method reduces to Hommel's procedure.

Proposition 4.3. *The FACT algorithm for closing Simes' method leads to the same rejections as Hommel's procedure.*

See Section 9.6 for the proof. Again, this result is important because it shows that our FACT method recovers another well-known example of an efficient closed testing algorithm.

More recently, Meijer et al. (2017) gave a more efficient, $O(n \log n)$ algorithm for closing Simes' method. This is faster than FACT, but is limited to Simes' method. In contrast, FACT is applicable to any monotone symmetric local testing rule, including Bonferroni and monotone combinations (see below).

Another important class of monotone local testing rules comes from the **Generalized Simes Test (GST)**. Grechanovsky and Hochberg (1999) introduce the GST, which tests a intersection null H_J in the following way. Let

$$d_{|J|,1}, \dots, d_{|J|,1}$$

be a sequence of critical values. Reject H_J if, there is $j \in J$ with

$$p_{(j)} \le d_{|J|,j}.$$

Here $p_{(j)}$ refers to the ordering of the p-values within the set J. Note that for Simes test, $d_{|J|,j} = \alpha j/|J|$. Liu (1996) gave conditions for reducibility of a closure based on GST to general sequentially rejective step-down or step-up procedures. Here we consider a more general set of GST procedures.

We next show that the Generalized Simes Test is symmetric and monotone. Therefore, we can use GST as a component in the FACT algorithm. This paves the way to a wide variety of new methods for closed testing. See Section 9.7 for the proof.

Proposition 4.4. *The Generalized Simes Test is symmetric and monotone.*

4.3 Monotone sums

Another broad class of monotone symmetric rules is the set of **monotone sums**, in which the test statistics are sums of monotone functions of the p-values. These are especially appealing if the effect sizes are "dense", in the sense that we expect to have many nonzero effects. Monotone sums include the following test statistics:

1. **Fisher's combination**:

   $$T(p_J) = 2 \sum_{i \in J} \ln(p_i).$$

   Fisher's test has a $-\chi^2_{2k}$ distribution under the intersection null when all p-values are uniform and independent (Fisher, 1970). Thus. $c_{k,\alpha} = -\chi^2_{2k}(1-\alpha)$, the $100(1-\alpha)$-th percentile of the χ^2_{2k} distribution.

   Fisher's combination is monotone, because the function $2\ln(x)$ is monotone increasing for $x \in [0,1]$.

2. **Stouffer's combination**:

   $$T(p_J) = \sum_{i \in J} \Phi^{-1}(p_i),$$

   where Φ is the standard normal cdf (Stouffer et al., 1949). Stouffer's combination has a $\mathcal{N}(0,k)$ distribution under the intersection null when all p-values are uniform and independent. Thus, $c_{k,\alpha} = k^{1/2}\Phi^{-1}(\alpha)$.

   Stouffer's combination test is symmetric and monotone, for the same reasons as Fisher's test.

3. **Wilkinson's combination**:

   $$T(p_J) = -\sum_{i \in J} I(p_i \le d),$$

   where $d > 0$ is some constant (Wilkinson, 1951). Wilkinson's combination has a sign-flipped Binomial distribution with n trials and success probability d under the intersection null when all p-values are uniform and independent. Thus, its critical values can be found from the distribution of the Binomial.

4. **Truncated Product Method**:

   $$T(p_J) = \sum_{i \in J} \ln(p_i) I(p_i \le \tau),$$

   where $\tau > 0$ is some constant (Zaykin et al., 2002). The value $\tau = \alpha$ is suggested as a default (Zaykin et al., 2002). The critical values of this test can be found numerically.

5. **Romano-Shaikh-Wolf combination**: Suppose we observe independent random variables $X_i \sim \mathcal{N}(\mu_i, 1)$ and we wish to test $H_i : \mu_i = 0$ against $\mu_i \neq 0$. Romano et al. (2011) study test statistics of the form

   $$T(X_J) = \sum_{i \in J} \cosh(\varepsilon |X_i|)$$

for $\varepsilon > 0$. They show that the closure of these test statistics has a maximin optimality property against subsets of the alternative of the form $\gamma(\varepsilon) = \{\mu : |\mu_i| \geq \varepsilon, \text{all } i\}$, for large enough ε. This follows from their more general result that maximin optimality is inherited under closure if the resulting multiple test is consonant. Clearly, their test statistics are monotone in $|X_i|$, so they fit in our framework.

6. **Monotone combination**: More generally, we can use sums a of monotone increasing function f of the p-values:

$$T(p_J) = \sum_{i \in J} f(p_i).$$

   The above tests are special cases. The critical values of this test can be found numerically.

   Any monotone combination tests takes $T_n = O(n)$ to apply to n p-values. Thus, the overall running time of the closed testing method is $O(n^2 k)$ if k nulls are rejected, and $O(n^3)$ in the worst case.

4.4 Monotone functions of order statistics

The power and flexibility of our method is showcased by the ability to use local testing rules that go beyond the classical ones (such as Bonferroni and Simes). Indeed, we can use any **monotone functions of the order statistics**. We give several examples below:

1. **Rank-sum type statistics**. We can use test statistics of the form

$$T(p) = \sum_i f_i(p_{(i)}),$$

   where f_i are for monotone increasing functions, possibly changing with i. The Bonferroni method is a special case, where $f_1(x) = x$, and the other functions are zero. Broader examples of rank-sum type statistics include **linear weighted rank-sum statistics**. For any weights $w_i \geq 0$, we can use the linear weighted rank-sum statistics

$$T(p) = \sum_i w_i \cdot p_{(i)}.$$

   For instance, if we want to emphasize not just the smallest, but also the second smallest p-value, we may use $p_{(1)} + \varepsilon p_{(2)}$.

2. **Min/max type statistics**. We can also use test statistics of the form

$$T(p) = \max_i f_i(p_{(i)})$$

   and $T(p) = \min_i f_i(p_{(i)})$ where f_i are monotone increasing functions, possibly changing with i. The following test statistics are examples:

   (a) **Generalized Simes Test (GST)**. Recall that the GST rejects H_J if there is $j \in J$ with $p_{(j)} \leq d_{|J|,j}$ (Grechanovsky and Hochberg, 1999). This falls in the max-category, where $f_i(x) = I(x > d_{|J|,i})$.

(b) **Higher Criticism (HC)**: The higher criticism (HC) test statistic was introduced by Tukey in the 1960s, and experienced a resurgence of interest after its study by Donoho and Jin (2004). The local test can be described using the functions

$$g_i(x) = \sqrt{n}\frac{x - i/n}{\sqrt{x(1-x)}},$$

and the test statistic equals, for some $0 < \alpha_0 < 1$,

$$T(p) = \min_{i \le \alpha_0 n} g_i(p_{(i)}).$$

The critical value for the test can be chosen as $-\sqrt{2\log\log n}(1 + o(1))$ (Donoho and Jin, 2004), but this may need some adjustments in finite samples. This test falls in the min-category, where $f_i = g_i$ for $i \le \alpha_0 n$, and $f_i = 0$ otherwise. The higher criticism has originally been studied under independence, but there are extensions allowing some degree of dependence.

A challenge with these general monotone functions is that the critical values are typically not available in closed form. However they can usually be evaluated numerically.

3. **Hybrid Hochberg-Hommel**. Gou et al. (2014) proposed improved hybrid Hochberg–Hommel type step-up multiple test procedures, and showed that they are the closures of the following local tests. We reject the intersection null, i.e., $T(p) = 1$ if one of the following mutually exclusive events occurs:

$$E_i = \begin{cases} p_{(n)} \le \alpha & i = 1, \\ p_{(n)} > \alpha, p_{(n-1)} > c_2\alpha, \ldots, p_{(n-i+2)} > c_{i-1}\alpha & \\ p_{(n-i+1)} \le c_i\alpha, p_{(1)} \le d_i\alpha & i \ge 2. \end{cases}$$

Here c_i, d_i are two monotone decreasing sequences of critical constants with $1 \ge c_i \ge d_i$. It is easy to see that this local test is monotone and symmetric. Indeed, we only need to observe that if we are in E_i, and decrease any p-value, we will either stay in E_i, or move to some E_k with a smaller index $k < i$.

5 Consonance

The notion of consonance is fundamental in multiple testing. A multiple testing rule is said to be consonant when, for any set A, if the global null H_A is rejected, there is at least one singleton $i \in A$ such that the individual null H_i is also rejected (Gabriel, 1969). Formally, if the multiple test has decision rule Φ, then it is consonant if for any A, there is an index $i \in A$ such that $\Phi(p_A) \le \Phi(p_i)$.

Consonance also leads to computationally efficient closed testing rules under some conditions. Hommel et al. (2007) show that consonance leads to a shortcut of order n for closed testing. However, their shortcut is only feasible if one can identify the elementary hypothesis in an efficient way. Unfortunately, this is only known for weighted Bonferroni tests. Hommel et al. (2007) acknowledge this limitation, writing that "*after rejecting H_A it may sometimes remain difficult to identify an elementary hypothesis H_i, $i \in A$, to be rejected. In such cases the short-cut can still be computer intensive*". See also Brannath and Bretz (2010) for methods based on local consonance for restricted hypotheses.

Going back to the main topic of the paper, suppose now that the local test statistics used in closed testing are also consonant. Assuming in addition that they are symmetric and monotone, as in the previous sections, it is easy to see that one obtains an efficient dynamic programming-type shortcut for computing the closed testing method. Indeed, the closed testing decision rule for the smallest p-value reduces to

$$\Phi_c(p_{(1)}) = \Phi(p_{(1)}, \ldots, p_{(n)}).$$

To see this, first we notice that the decision rule clearly must include the above factor, i.e., the intersection null with all n p-values must be rejected in order for the smallest p-value to be rejected. Next, by consonance, if the null with these p-values is rejected, then there must be a singleton j that is rejected. Now, since $p_{(1)} \le p_j$, we obtain that $p_{(1)}$ is also rejected. This shows that the decision rule for the smallest p-value has the above form.

With a similar reasoning, we obtain that the decision rule for the second p-value has the form

$$\Phi_c(p_{(2)}) = \Phi(p_{(1)}, \ldots, p_{(n)}) \cdot \Phi(p_{(2)}, \ldots, p_{(n)}).$$

Therefore, we must only test the null with p-values $p_{(2)}, \ldots, p_{(n)}$ to compute the decision rule for the second smallest p-value. Continuing, this shows that an algorithm of complexity $O(ns)$ exists for computing consonant closed tests based on monotone symmetric testing rules. For instance, the Bonferroni rule with $\Phi(p_1, \ldots, p_k) = I(\min p_i \le \alpha/k)$ satisfies these properties. It is also well known that the Hommel procedure is a consonant closed testing procedure (Sonnemann, 1982, 2008). Therefore, consonance and closed testing lead to extremely fast algorithms.

However, we emphasize that the scope of this paper goes *much beyond* consonance. In this section, we will show here that there are important examples of closed tests based on monotone and symmetric rules that are *not* consonant. We will establish conditions needed for consonance, and then show that specific tests do not satisfy them.

For this we take a systematic approach. Suppose we are testing n null hypotheses using the closure of monotone combination tests $T(p_J) = \sum_{i \in J} f(p_i)$. Recall that the global rules are $\Phi(p_J) = I(T(p_J) \le c_{|J|})$, where $c_{|J|} = c_{|J|,\alpha}$. We have the following key result, which clarifies the conditions on closures of monotone sums under which we have consonance. We call this result the Consonance-Closure-Monotonicity (CCM) lemma. See Section 9.8 for the proof.

Lemma 5.1 (Consonance-Closure-Monotonicity (CCM) lemma)**.** *Let F_k be the cdf of an average of k random variables $f(P_i)$, where P_i are independent p-values uniformly distributed on* $[0, 1]$*:*

$$F_k(c) = \Pr\left(k^{-1}\sum_{i=1}^{k} f(P_i) \le c\right).$$

Then the closure of monotone combination tests $\Phi(p_J) = I(T(p_J) \le c_{|J|})$ is consonant if and only if the following two conditions hold:

(a) **Level**: *Each local test has level α:*

$$F_k(c_k) \leq \alpha$$

for all $k \leq n$.

(b) **Sub-linear critical value growth**: *The critical values c_k grow at most linearly:*

$$c_k \leq kc_1.$$

From the CCM lemma, we can derive several important and interesting results. First, we study the consonance of two specific monotone combination rules, Stouffer's and the Truncated Product Method, and then we give a more general result. See Section 9.9 for the proof.

Proposition 5.2 (Examples of Consonance)**.** *1. The closure of Stouffer's combination, where $T(p_J) = \sum_{i \in J} \Phi^{-1}(p_i)$, is not consonant if*

$$\alpha < 1/2.$$

2. The closure of the Truncated Product Method, where $T(p_J) = \sum_{i \in J} \ln(p_i) I(p_i \leq \tau)$, is consonant if

$$1 - \sqrt{1 - \alpha} \leq \tau \leq \alpha.$$

3. More generally, suppose that f is a strictly increasing continuous function such that $\mathbb{E}|f(P)| < \infty$ for a uniform p-value P. Suppose we are testing n null hypotheses. The closure of monotone combination tests $T(p_J) = \sum_{i \in J} f(p_i)$ based on independent p-values can only be consonant *for all n if*

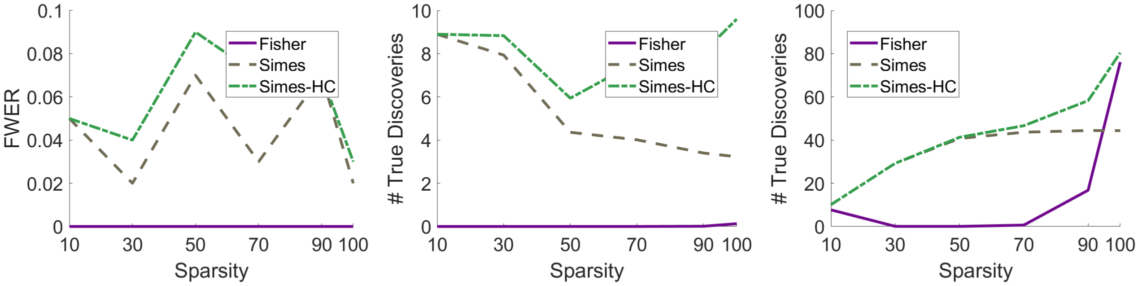
$$\mathbb{E}f(P) \leq f(\alpha).$$

This result gives clear conditions under which closures of monotone sums are consonant, and gives specific examples. In particular, the third statement implies that the closure of Fisher's combination test is not consonant for all n if $\alpha > 1/e$.

The condition above gives a lower bound on the critical value α for which the tests are consonant. The values of interest for us are typically $\alpha = 0.05$ or $\alpha = 0.01$. For these values, however, the classical combination tests (Fisher and Stouffer), are *not* consonant. The closure of the Truncated Product Method is consonant for the choices of τ specified in Proposition 5.2. The FACT algorithm applies to the closure of Fisher's and Stouffer's test, neither of which are consonant in general. This shows that our FACT algorithm has a broader scope than consonance.

A key idea about consonance is *consonantization* (Romano et al., 2011), which shows that under some conditions, any closed testing method can be replaced with a consonant one. Specifically, following the above reference, a family $\{H_i\}_{i=1,\dots,n}$ of hypotheses is called elementary if there is no $i \neq j$ such that $H_i \subset H_j$. Any closed testing method of elementary hypotheses can be made consonant, in the following way. Suppose H_K is rejected when $\Phi_K = 1$. Define

$$\Phi'_K = \max_{i: i \in K} \prod_{J: i \in J} \Phi_J.$$



(a) $M = 0$ (b) $M = 1$ (c) $M = 2$

Figure 1: FWER and average number of hypotheses rejected by the FACT method for the closure of Fisher's method, Simes' method, and the Simes-HC fusion rule. The displays are as a function of the signal strength M and sparsity s.

Then, the closure of Φ_K and Φ'_K reaches the same decisions about H_i, and the closure of Φ'_K is consonant.

While this is an intriguing idea, it is not clear to us if it can be computed efficiently in a general context. This is an interesting problem that falls beyond our current scope.

6 Numerical experiments

We perform numerical simulations to understand and compare the behavior of our proposed methods.

6.1 Fusion rules are more powerful

We compare the closure of Fisher's method, Simes' method, and FACT with the Simes-Higher Criticism (Simes-HC) fusion rule. The reason for performing this simulation is that we would like to understand under what condition the new fusion rule can perform better than Simes method or Fisher's method.

We do not compare Holm's method, because the closure of Simes' method is more powerful. Moreover, we do not include the closure of the HC method, because the critical values of that test are only asymptotic, and the level control is not guaranteed for small sizes of the intersection null. When we use the HC method in the fusion rule, we always use it for intersection nulls of a large size, where the asymptotics are more accurate.

In the simulation, we use the normal means model, where the data X_i is independent and normally distributed with $X_i \sim \mathcal{N}(\mu_i, 1)$ for $i = 1, \ldots, 100$. The null hypotheses considered are that $\mu_i = 0$. The alternative hypotheses are that $\mu_i > 0$.

We consider both sparse and dense models. We change the sparsity—the number of nonzero μ_i effect sizes—of the model on a grid from zero to 100. We tune the effect sizes so that the power is in a non-trivial regime, and comparable across the different sparsities. Specifically, it is well known that the chi-squared test for a global null based on k normal means behaves approximately as

$$\mathcal{N}(k + |\mu|_2^2, 2k)$$

for large k, where $|\mu|_2$ is the Euclidean norm of the vector of means. Therefore we choose our effect sizes in the following way. For a global effect size M, and sparsity s, we set each nonzero effect size to be equal to $\mu_i = (2p/s)^{1/2}M$. This ensures that the Euclidean norm of the effects is the same for different sparsities. We take the signal strength M to be 0, 1, and 2, respectively.

We then run the Fast Closed Testing method using various local testing methods. We always use the global significance level $\alpha = 0.05$. We average the results over 100 independent Monte Carlo trials.

We compare the closure of Fisher's method, Simes' method, and the FACT algorithm using the Simes-Higher Criticism (Simes-HC) fusion rule. For the Simes-HC rule, we assume that the preliminary estimate of the sparsity is correct. The results are displayed in Figure 1. We observe that all methods empirically control the family-wise error rate, up to random sampling error. This can be seen on the left plot showing the FWER, where $M = 0$, so that we are under the global null. On the remaining plot, we show the expected number of true discoveries.

We observe that the closure of the Simes-HC fusion rule is more powerful than the other methods. First, the closure of Fisher's method only has power against fully dense alternatives, and only when the effect sizes are large ($M = 2$). Second, as seen in the middle plot, the closure of Simes' method can lose power when the sparsity increases.

Finally, FACT with the Simes-HC rule has more power than the other methods. This can be seen best on the rightmost plot, where Simes HC tracks Simes for sparse alternatives, and tracks Fisher for dense alternatives. This shows that the FACT method, when used with the appropriate local testing rules, can be more powerful that the closure of a minimum or sum-based local testing rule, and demonstrates the power of our approach.

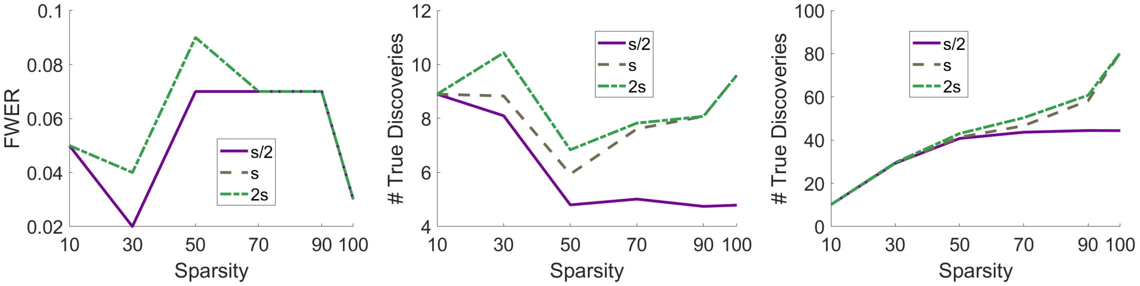
6.2 Robustness to mis-specifying the sparsity

We examine the robustness of the Simes-HC fusion rule to the sparsity tuning parameter. The reason for performing this simulation is that we would like to understand how the performance of the method depends on our prior guess for the sparsity s. We use the same simulation setup as in the previous section. We now run the Simes-HC rule with sparsity tuning parameters equal to $s/2, s$, and $\min(s, n)$. The results are displayed in Figure 2.

We observe that the method is not too sensitive to the sparsity tuning parameter. Regardless of the value of that parameter, the Simes-HC rule always has good power. Underestimating the sparsity seems to lead to a smaller number of rejections. This suggests that in practice one should use over-estimates of the sparsity for robust performance.

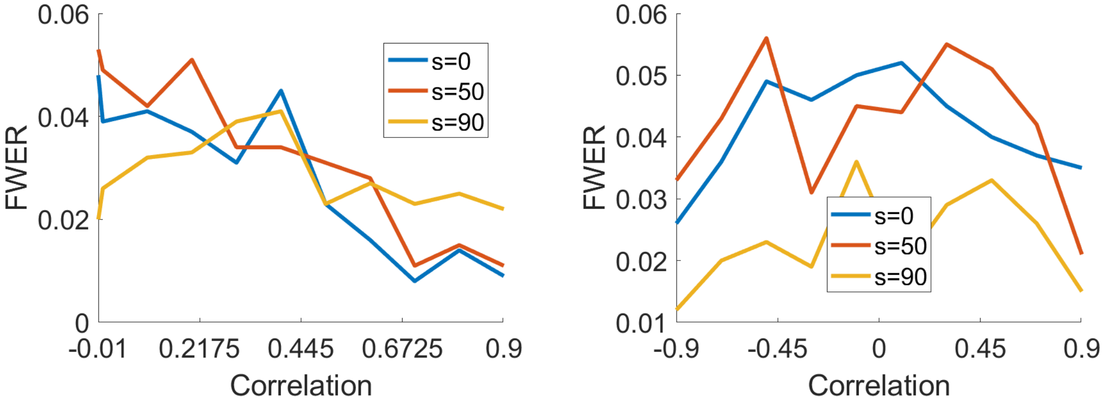
6.3 Robustness to correlated test statistics

We examine the robustness of the FACT method to with the Simes-HC fusion rule to correlated test statistics. The reason is to understand how the FWER depends on the



(a) $M = 0$ (b) $M = 1$ (c) $M = 2$

Figure 2: FWER and average number of hypotheses rejected by the FACT method with the Simes-HC fusion rule, for potentially mis-specified sparsity levels. The displays are as a function of the signal strength M and sparsity s.



(a) Spiked model (b) AR-1

Figure 3: FWER of the FACT method with the Simes-HC fusion rule, for correlated test statistics. The displays are as a function of the correlation coefficient.

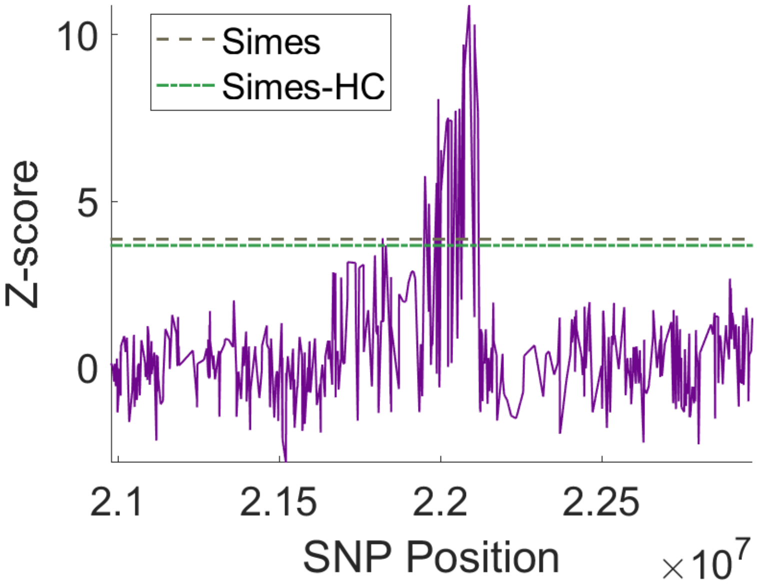


Figure 4: Plot of Z-scores for association with coronary artery disease in the C4D GWAS dataset, focusing on the 9p21.3 locus. The horizontal lines are the critical values of the closure of Simes' method (i.e., Hommel's method), and the FACT algorithm with the Simes-HC fusion rule. The Simes-HC method discovers a second associated SNP in the secondary cluster.

correlation structure of the tests.

We use the same simulation setup as in the previous section, with effect size $M = 1$, and varying sparsity. Moreover, we sample the test statistics as

$$X \sim \mathcal{N}(\mu, \Sigma),$$

where Σ is a covariance matrix. We choose Σ to be either a spiked covariance matrix $\Sigma = (1-\rho)I_n + \rho 1 1^\top$, where 1 is the vector or all ones, or an autoregressive covariance matrix of order one (AR-1), $\Sigma_{ij} = \rho^{|i-j|}$. We let the correlation coefficient ρ vary over the entire range where Σ is non-negative definite. This includes both positive and negative correlation structures. The results are displayed in Figure 3. We average over 1000 Monte Carlo trials. We observe that the method essentially controls the FWER for all correlation structures in this case.

7 Data Analysis

We illustrate the FACT method on a Genome-Wide Association Study (GWAS) of coronary artery disease. In the last decade, GWAS have become the backbone of modern medical genomics. Thousands of such studies have been performed, and have led to hundreds of novel associations between common traits and genetic variants (see e.g., Visscher et al., 2012, for a review). Thus there is a great deal of interest in improving statistical inference via multiple testing in this area (e.g. Efron, 2012; Dobriban et al., 2015; Fortney et al., 2015; Dobriban, 2017).

GWAS are a general and flexible type of genomic study to understand complex traits and diseases. In a typical GWAS, we collect a large number of cases and controls for a disease of interest, such as coronary artery disease. We also measure the genotypes of the samples for potentially hundreds of thousands of genetic variants known as Single Nucleotide Polymorphisms (SNPs). We then find the most significantly associated SNPs using multiple testing.

In this paper, we will llustrate the FACT method on the GWAS dataset from the C4D consortium for coronary artery disease genetics (Coronary Artery Disease Genetics Consortium, 2011). This dataset contains about 500,000 genotyped SNPs, on 15,420 individuals with CAD (cases) of which 8,424 are Europeans and 6,996 are South Asians, along with 15,062 controls. Because out Simes-HC fusion method can more powerful than the closure of Simes' method (i.e., Hommel's method) for relatively dense alternatives, we focus on a subset of the SNPs that is already known to be enriched for CAD-related loci. Specifically, we focus on the neighborhood of the CDKN2A gene, at the 9p21.3 locus.

This locus is known to be strongly associated with CAD from prior work, however the functional mechanism appears to be not fully known (see e.g., Harismendy et al., 2011; Chen et al., 2014). Therefore, it is of interest to better understand the local genetic architecture of this region. We focus on a 200 kilobase region centered at the position 21967752, which is the location of one of the most significant hits in the current study. There are $J = 452$ SNPs in this window.

We perform multiple testing with the closure of Simes' method (i.e., Hommel's method) and the FACT method using the Simes-HC fusion. For the Simes-HC fusion, we set the sparsity as $s = 0.1J$. Hommel's method finds 24 significant loci, while the FACT with Simes-HC finds 25. The results displayed in Figure 4 show that most of the discoveries are in a contiguous window around the center of the region.

Both Hommel's method and the fusion find significant discoveries in a second cluster, about 20 kilobases away from first one, but the fusion finds two loci there. This gives stronger evidence for the association of that cluster. Regarding interpretation, it is most likely that there are several loosely dependent SNPs associated with CAD in this region, so we think that it is valuable to have stronger evidence near the secondary locus.

Finally, we note that we expect the test statistics in this region to be dependent. However, we do not have access to the full dataset, but only to the list of p-values, and hence performing permutation methods is not feasible. Heuristically, the dependence should be positive, because of linkage disequilibrium between neighboring SNPs. Hence Simes' rule is expected to have the appropriate level. The same statement is less clear about the higher criticism, and this specific question deserves further study.

8 Discussion

In this paper we have proposed the Fast Closed Testing (FACT) method for flexible multiple testing. The FACT method enables the efficient computation of the closure principle for monotone symmetric local testing rules. A key strength of the method is that practitioners can choose which local testing rules to use for any given subset size, for instance Bonferroni-type rules, Simes-type rules, monotone sums, or more general combinations like the Higher Criticism. We have illustrated this by designing the Simes-Higher Criticism fusion rule, which uses Simes' method for testing intersection nulls of a small size, while using the Higher Criticism to test intersection nulls of a larger size. We gave an example using the method in a Genome-Wide Association Study of coronary artery disease.

This work raises a number of interesting questions for future work. In our work, symmetry was crucial for deriving efficient algorithms. However, there are important non-symmetric multiple testing methods, including weighted methods (Holm, 1979; Dobriban et al., 2015; Fortney et al., 2015; Dobriban, 2017), fixed-sequence procedures (Maurer et al., 1995; Westfall and Krishen, 2001), and fallback procedures (Wiens, 2003). These control the FWER, because they are closures of weighted Bonferroni methods (Hommel et al., 2007). However, there is currently no general explanation for why they admit efficient algorithms. Can we derive such principles? Can these principles lead to new multiple testing methods?

Second, closed testing methods have appealing power properties. It is known that any multiple testing method can be replaced by a closed testing method that rejects the same, and possibly more, hypotheses while controlling the FWER (Sonnemann, 1982, 2008; Sonnemann and Finner, 1988). However, it is not known how to compute this closure efficiently. Are there conditions under which it can be computed in polynomial time?

In addition, it would be of major practical interest to study the case of *dependent* p-values.[2] For this we would need a new probability inequality for local tests based on symmetrically distributed–or exchangeable–p-values, similar to Bonferroni's inequality for arbitrary dependence, and Simes' inequality for positive dependence (Goeman and Solari, 2014). Then we could apply the closure principle as in this paper.

Moreover, there are many multiple testing methods that go beyond the closure principle. For instance, the it would be interesting to see when the methods based on the sequential testing principle (Goeman and Solari, 2010) can be computed efficiently. Understanding these questions will help broaden our methods for multiple testing.

---

[2] We thank Jelle Goeman for encouraging us to consider this problem.

9 Proofs

9.1 Review of the well-known argument that CT controls the FWER

This is the well-known argument that CT controls the FWER. let I_0 be the set of all true nulls. Then, for all $i \in I_0$:

$$\Phi_c(p_i) = \prod_{J: i \in J} \Phi(p_J) \le \Phi(p_{I_0}),$$

where in the first step we have used the definition of CT (reject individual null if all sets containing it are rejected), and that all rejection rules take values in $\{0, 1\}$, while in the second step we have used that I_0 is among the sets J containing i. We conclude that if H_i is rejected, then so is H_{I_0}. Since the probability of rejecting H_{I_0} is at most α, this shows that CT controls the FWER.

9.2 Proof of Theorem 2.1

Consider the closed testing method. To decide whether or not we reject the k-th hypothesis, we must decide for every subset containing k whether or not it is rejected based on the local testing rule. Now consider subsets of a fixed size j containing k. By assumption, for each of these subsets, we use the same local testing rule T_j. Also by assumption, these testing rules are symmetric and monotone. It follows that all subsets are rejected if and only if the "worst" one is rejected. The "worst one" has the largest $j - 1$ p-values excluding p_k.

We can formalize this intuition as follows. Let $p^{-k}_{(1)} \le \ldots \le p^{-k}_{(n)}$ be the sorted p-values excluding p_k. Recall that the decision rule Φ_c for the k-th null is

$$\Phi_c(p_k) = \prod_{J: k \in J} \Phi(p_J).$$

We can write this as a product over subsets of each possible size j. By the above discussion, the j-th term equals $\Phi(p_k, p^{-k}_{(n-j+1)}, \ldots, p^{-k}_{(n)})$. Therefore, the entire decision rule for the j-th term has the form

$$\Phi_c(p_k) = \Phi(p_k) \cdot \Phi(p_k, p^{-k}_{(n)}) \cdot \ldots \cdot \Phi(p_k, p^{-k}_{(1)}, \ldots, p^{-k}_{(n)}).$$

Next, we observe that if p_k is less than or equal to p_l, then the term $\Phi(p_k, p^{-k}_{(a)}, \ldots, p^{-k}_{(n)})$ is less than or equal to $\Phi(p_l, p^{-l}_{(a)}, \ldots, p^{-l}_{(n)})$. This simply means that if we reject all subsets of a fixed size containing the l-th hypothesis, then we also reject all subsets of the same size containing the k-th hypothesis. Therefore, if we do not reject the "worst" subset of size j (say) for $p_{(k)}$, we do not reject the "worst" subset for any $p_{(l)}$ with $l \ge k$.

This shows that the following algorithm is equivalent to the closed testing method. Start by sorting the p-values. For each p-value $p_{(k)}$ starting with the smallest one, check if the "worst" subset of each size j is rejected using the local testing rule T_j.

If any such subset is not rejected, stop, and reject the hypotheses with the p-values $p_{(1)}, \ldots, p_{(k-1)}$. This agrees with the FACT algorithm, showing its correctness.

Finally, we study the computational cost of the algorithm. The initial sort takes $O(n \log(n))$ steps. Then at step k, the cost is at most

$$t_1 + t_2 + \ldots + t_{n-k+1},$$

where t_i is the cost of applying T_i to a size i subset. For the total cost, suppose we reject k out of n hypotheses. Then the total cost is at most

$$C_k = O\left(n \log(n) + k\left[t_1 + t_2 + \ldots + t_{n-k}\right] + \sum_{j<k} j t_{n-j+1}\right)$$

For instance, if $t_n = O(n^c)$ with $c \geq 1$, then we get a total cost $O(kn^{c+1})$. When $t_i = O(i)$, i.e., when applying the test takes linear time, the cost is $C_k = O(kn^2)$. This finishes the proof.

9.3 Proof of Lemma 3.1

We first study the properties of the rejection regions of monotone symmetric sets. Then, we apply these results to prove the current result. Consider a monotone symmetric rule $\Phi : [0,1]^n \to \{0,1\}$. Let A be the rejection region, $A = \{p : \Phi(p) = 1\}$. Note that $A \subset [0,1]^n$ must of course be Borel measurable. Then monotonicity and symmetry of Φ are equivalent to the same properties of A:

1. **Set Monotonicity**: $p \in A$ implies $q \in A$, if on all coordinates i, $q_i \leq p_i$.
2. **Set Symmetry**: $p = (p_1, \ldots, p_n) \in A$ implies $p_\pi = (p_{\pi(1)}, \ldots, p_{\pi(n)}) \in A$ for any permutation π of $\{1, \ldots, n\}$.

Let $\mathcal{A}$ be the collection of all such sets. What properties does $\mathcal{A}$ have?

Lemma 9.1 (Union)**.** *The union of any collection of monotone symmetric sets is monotone symmetric.*

Proof. Let A_i, $i \in I$, be a collection of monotone symmetric sets, and suppose $x \in \cup_i A_i$. Let any $y \leq x$, where inequality is meant coordinate-wise, i.e., $y_i \leq x_i$ for all i. Now, we must have $x \in A_i$ for some $i \in I$. Then, since A_i is monotone, we have $y \in A_i$, hence $y \in \cup_i A_i$. Moreover, since A_i is symmetric, we have $x_\pi \in A_i \subset \cup_i A_i$ finishing the proof. ☐

Lemma 9.2 (Intersection)**.** *The intersection of any collection of monotone symmetric sets is monotone symmetric.*

The proof is similar to the previous one, and hence omitted.

For any $a \in [0,1]^n$, let us denote the hyper-rectangle with opposite vertices 0 and a by $H(a)$. Thus

$$H(a) = \{x : 0 \leq x_i \leq a_i, i = 1, \ldots, n\}.$$

Clearly, $H(a)$ is a monotone set. Let moreover $H_\pi(a)$ be the symmetrization of $H(a)$, that is

$$H_\pi(a) = \cup_{\pi \in S_n} H(a_\pi).$$

Thus $H_\pi(a)$ is monotone and symmetric. Moreover, $H_\pi(a)$ is a minimal monotone symmetric set, in the sense that if a belongs to a monotone symmetric set, then $H_\pi(a)$ must also belong to it.

Lemma 9.3 (Representation)**.** *A set A is monotone symmetric if and only if it can be written as*

$$A = \cup_{a \in I} H_\pi(a),$$

for some measurable set $I \subset [0,1]^n$.

Proof. Clearly, if this representation holds, then by the union lemma, Lemma 9.1, A is monotone symmetric. On the other hand, if A is monotone symmetric, then clearly A has the given representation, with $I = A$. This finishes the argument. ☐

9.3.1 Final proof of Lemma 3.1

Clearly, a test statistic T is monotone and symmetric if and only if all its sub-level sets $S_T(c) = I(T(p) \le c)$ are monotone symmetric in the sense of sets, as defined above.

Now, for the first part, let $T = \min_{i \in I} T^i$ be a minimum of an arbitrary measurable collection of monotone symmetric test statistics. We notice that $\min_{i \in I} T^i \le c$ iff for some $i \in I$ we have $T^i \le c$. Thus

$$S_T(c) = \cup_{i \in I} S_{T^i}(c).$$

Thus, T is monotone symmetric the union lemma, Lemma 9.1.

Similarly, for the second part let $T = \max_{i \in I} T^i$ be a maximum of an arbitrary measurable collection of monotone symmetric test statistics. Clearly, $\max_{i \in I} T^i \le c$ iff for all $i \in I$ we have $T^i \le c$. Thus $S_T(c) = \cap_{i \in I} S_{T^i}(c)$. Thus, this property follows from the intersection lemma, Lemma 9.2.

For the third part, it is clear from the original definition for test statistics that monotonicity and symmetry are preserved under non-negative combinations.

For the fourth part, suppose that on all coordinates i, $p_i \le p'_i$. Then, by monotonicity of the T^j, $T^j(p) \le T^j(p')$ for all j, so that by monotonicity of g,

$$g(T^1(p), \ldots, T^k(p)) \le g(T^1(p'), \ldots, T^k(p')).$$

This shows that $g(T^1, \ldots, T^k)$ is a monotone test statistic. Next, to see the symmetry, we notice that the value of each T^j is unchanged under permutations. Therefore, the value of g is also unchanged, finishing the proof.

9.4 Proof of Lemma 3.2

The proof of this claim follows by examining what happens to $p_{(i)}$ if we decrease any index $p_k = p_{(j)}$. Consider the set $S_i(c) = \{(p_1, \ldots, p_n) : p_{(i)} \le c\}$. If $j < i$, clearly the value of $p_{(i)}$ does not change. If $j = i$, then $p_{(i)}$ decreases, so we stay in the set $S_i(c)$.

Finally, if $j > i$, then the value of $p_{(i)}$ stays fixed initially, and can only decrease after we have decreased the coordinate that used to be $p_{(j)}$ enough to reach $p_{(i)}$. This shows that regardless which coordinate we decrease, we stay in $S_i(c)$, finishing the proof.

9.5 Proof of Proposition 4.1

In general, the decision rule for the hypothesis corresponding to the first p-value equals

$$\Phi(p_{(1)}) \cdot \Phi(p_{(1)}, p_{(n)}) \cdot \ldots \cdot \Phi(p_{(1)}, p_{(2)}, \ldots, p_{(n)}).$$

Since we are working with the Bonferroni method, the i-th factor equals $I(p_{(1)} \leq \alpha/i)$. Since the thresholds are decreasing in i, the last rule is the most stringent one. Thus, we obtain the simplification that the decision rule for the smallest p-value equals

$$I(p_{(1)} \leq \alpha/n).$$

If we have rejected the hypothesis corresponding to the smallest p-value, we continue and examine the second smallest one. This is rejected based on the rule

$$\Phi(p_{(2)}) \cdot \Phi(p_{(2)}, p_{(n)}) \cdot \Phi(p_{(2)}, p_{(n-1)}, p_{(n)}) \ldots \Phi(p_{(2)}, p_{(3)}, \ldots, p_{(n)}).$$

As above, the i-th factor equals $I(p_{(2)} \leq \alpha/i)$. Since the thresholds are decreasing in i, the last rule is the most stringent one. Thus, we obtain that the second decision is based on

$$I(p_{(2)} \leq \alpha/(n-1)).$$

Continuing similarly, we obtain that the decision to reject the hypothesis corresponding to $p_{(i)}$ is only considered if the hypotheses corresponding to the smaller p-values are all rejected. The hypothesis is then rejected if $p_{(2)} \leq \alpha/(n-i+1)$, which agrees with Holm's method. This shows that the steps of the FACT algorithm recover Holm's method.

9.6 Proof of Proposition 4.3

The reasoning proceeds from the last loop of the FACT algorithm sequentially towards the first loop. We show that the decision made at each loop matches the decision at the corresponding step of Hommel's procedure. We view Hommel's procedure as a sequential algorithm proceeding from $j = 0$ to $j = n$. At the j-th step, we check if $p_{(n-j+k)} > k\alpha/j$ for all $k \leq j$. If this is the largest j with this property, we reject all H_i with $p_i \leq \alpha/j$.

Thus, consider the last outer loop of the FACT algorithm, where $k = n$. There, if $p_{(n)} \leq \alpha$, all hypotheses are rejected. This matches the first step in Hommel's procedure, where $j = 0$. Indeed, in that case, there is no j such that $p_{(n-j+k)} > k\alpha/j$ for all $k \leq j$. This means, for $j = 1$, that $p_n \leq \alpha$.

However, we also need to argue that if $p_{(n)} \leq \alpha$, then the FACT algorithm indeed arrives at the n-th loop, and does not stop before that. This is indeed true, because if $p_{(n)} \leq \alpha$, then each of the conditions in each of the previous $n-1$ loops are fulfilled. This shows that the first steps of the two algorithms agree.

It remains to understand the case where $p_{(n)} > \alpha$. In this case, the FACT algorithm stops before the n-th outer loop. As before, there are two conditions for it to stop at the $n-1$-st outer loop. First, both of the conditions

$$\{p_{(n-1)} \le \alpha/2\} \cup \{p_{(n)} \le \alpha\}$$

and $\{p_{(n-1)} \le \alpha\}$ must be satisfied. Since $p_{(n)} > \alpha$, this means that the required condition is $p_{(n-1)} \le \alpha/2$. The second condition is that the algorithm must not stop before the $n-1$-st outer loop. However, we claim that the condition $p_{(n-1)} \le \alpha/2$ guarantees that.

Indeed, consider a previous outer loop, say the k-th one with $k \le n-2$. Then, we claim that for any inner loop j, there is a p-value satisfying the Simes constraint. Indeed, consider first the loops j for which $p_{(n-1)}$ is one of the p-values considered in $p_{(k)}, p_{(j+1)}, \ldots, p_{(n)}$. Thus, $j+1 \le n-1$, or $j \le n-2$. In this case, the threshold to which $p_{(n-1)}$ is compared in the Simes test is $\alpha \cdot (n-j)/(n-j+1)$. Since this is greater than $\alpha/2$, we have that

$$p_{(n-1)} \le \alpha \cdot (n-j)/(n-j+1),$$

and so the Simes constraint is satisfied in this case.

Consider next the loops j for which $j \ge n-1$. In this case, the threshold to which $p_{(1)}$ is compared in the Simes test is $\alpha/(n-j+1)$. Since this is greater than or equal to $\alpha/2$, and $p_{(1)} \le p_{(n-1)}$ we have that the Simes constraint is satisfied in this case.

In conclusion, we have shown that the FACT algorithm stops at the $n-1$-st loop precisely when $p_{(n)} > \alpha$ and $p_{(n-1)} \le \alpha/2$. The algorithm stops before this loop when $p_{(n-1)} > \alpha/2$. This agrees with the second step in Hommel's algorithm.

By a similar inductive argument, we obtain that the FACT algorithm stops at the $n-j+1$-st loop precisely when

$$p_{(n)} > \alpha, p_{(n-1)} > \alpha \cdot (j-1)/j, \ldots, \text{and } p_{(n-j+1)} \le \alpha/j.$$

This agrees with the j-th step in Hommel's algorithm, and finishes the proof.

9.7 Proof of Proposition 4.4

Let us write $d_i = d_{|J|,i}$ for simplicity. Define the regions

$$S_i(d_i) = \{p : p_{(i)} \le d_i\}.$$

The acceptance region of the Generalized Simes Test equals $S = \cup_i S_i(d_i)$. The order statistics are monotone symmetric by Lemma 3.2, hence the sets $S_i(d_i)$ are monotone symmetric. So by the closure properties (Lemma 3.1) of monotone symmetric sets, S is monotone symmetric, finishing the proof.

9.8 Proof of Lemma 5.1

For a test of the form $\Phi(p_J) = I(T(p_J) \le c_{|J|,\alpha})$, consonance requires that if $T(p_J) \le c_{|J|,\alpha}$, then there is an index $i \in J$ such that $T(p_i) \le c_{1,\alpha}$. For monotone combination tests $T(p_J) = \sum_{i \in J} f(p_i)$ with $k = |J|$, consonance requires that if

$$\sum_{i \in J} f(p_i) \leq c_{k,\alpha},$$

then $f(p_{(1)}) \leq c_{1,\alpha}$.

Now, in the worst case, we can take the p-values to be equal. Thus, consonance can only hold for all p-values if $\frac{c_{k,\alpha}}{k} \leq c_{1,\alpha}$. This finishes the proof.

9.9 Proof of Proposition 5.2

1. It is easy to see that the condition $c_{k,\alpha} \leq k c_{1,\alpha}$ does not hold for Stouffer's test. We have

   $$c_{k,\alpha} = k^{1/2} \Phi^{-1}(\alpha),$$

   thus

   $$c_{k,\alpha}/k = k^{-1/2} \cdot c_{1,\alpha}.$$

   Since $c_{1,\alpha} = \Phi^{-1}(\alpha) < 0$ if $\alpha < 0$, we thus have $c_{k,\alpha}/k > c_{1,\alpha}$. This finishes the proof of this claim.

2. For the truncated product method, we have $f(x) = \ln(x) I(x \leq \tau)$. For the critical value for one test, $c_1 = c_{1,\alpha}$, we need that

   $$F(c_1) = \Pr\left(\ln(P) I(P \leq \tau) \leq c_1\right) \leq \alpha.$$

   Now, taking $c_1 = \ln \tau$, we see that this probability equals τ. Since the value of the random variable $\ln(P) I(P \leq \tau)$ equals zero for $P > \tau$, it follows that $F(c_1) = \tau$ for any $c_1 \in (\ln \tau, 0)$. Since we are interested in the regime where $\tau \leq \alpha$, it follows that we can take any $c_1 \in (\ln \tau, 0)$ and the level condition is satisfied for subsets J of size one.

   Now consider subsets of size two. Let $X_i = \ln(P_i) I(P_i \leq \tau)$ and $c_2 = c_{2,\alpha}$ be the appropriate critical value. We need that

   $$F_2(c_2) = \Pr\left(X_1 + X_2 \leq c_2\right) \leq \alpha.$$

   Since $X_i \leq 0$, we have that $X_1 + X_2 = 0$ only if $X_1 = 0$ and $X_2 = 0$. The probability of this event is $(1-\tau)^2$. Moreover, the next largest value that $X_1 + X_2$ can take with positive probability equals $\ln \tau$, which happens when one $X_i = 0$ and the other $X_j = \ln \tau$. Therefore, we conclude that

   $$F_2(c_2) = 1 - (1-\tau)^2$$

   for all $c_2 \in (\ln \tau, 0)$. For this to be at most α, we need precisely that $1 - \sqrt{1-\alpha} \leq \tau$, which is the required condition.

   For consonance, it remains to show that one can choose c_2 such that $c_2 \leq 2c_1$. This is clear, because c_1, c_2 are only constrained to be in $(\ln \tau, 0)$.

   Finally, consider subsets of size k. Similarly to above, we derive that

   $$F_k(c_k) = 1 - (1-\tau)^k$$

for all $c_k \in (\ln \tau, 0)$. Thus, the level requirement translates to

$$1 - (1-\alpha)^{1/k} \leq \tau.$$

Since $x \to x^{1/k}$ is increasing for $x \in (0,1)$, this condition is implied by the one for $k = 2$. Moreover, for consonance, we need $c_k \leq kc_1$. This holds similarly to the case $k = 2$, finishing the proof.

3. One can check that $c_{1,\alpha} = f(\alpha)$ is a valid choice of a critical value for strictly increasing continuous f. Indeed, $\Pr(f(P_i) \leq f(\alpha)) = \Pr(P_i \leq \alpha) = \alpha$. Thus, consonance requires equivalently that

$$p_n := \Pr\left(\sum_{i=1}^{n} f(p_i) \leq nf(\alpha)\right) \geq \alpha.$$

By the law of large numbers, $n^{-1}\sum_{i=1}^{n} f(p_i) \to \mathbb{E}f(P)$ almost surely, thus if $\mathbb{E}f(P) > f(\alpha)$, then $\limsup_n p_n = 0$. Therefore, consonance for all n requires that $\mathbb{E}f(P) \leq f(\alpha)$, which finishes the proof.

References

A. Birnbaum. Combining independent tests of significance. *Journal of the American Statistical Association*, 49(267):559–574, 1954.

W. Brannath and F. Bretz. Shortcuts for locally consonant closed test procedures. *Journal of the American Statistical Association*, 105(490):660–669, 2010.

F. Bretz, W. Maurer, W. Brannath, and M. Posch. A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine*, 28(4):586–604, 2009.

F. Bretz, T. Hothorn, and P. Westfall. *Multiple Comparisons using R.* CRC Press, 2016.

H.-H. Chen, N. A. Almontashiri, D. Antoine, and A. F. Stewart. Functional genomics of the 9p21. 3 locus for atherosclerosis: clarity or confusion? *Current Cardiology Reports*, 16(7):502, 2014.

Coronary Artery Disease Genetics Consortium. A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease. *Nature Genetics*, 43(4):339–344, 2011.

A. Dmitrienko, A. C. Tamhane, and F. Bretz. *Multiple Testing Problems in Pharmaceutical Statistics.* CRC Press, 2009.

E. Dobriban. Weighted mining of massive collections of p-values by convex optimization. *Information and Inference: A Journal of the IMA*, 7(2):251–275, 2017.

E. Dobriban, K. Fortney, S. K. Kim, and A. B. Owen. Optimal multiple testing under a gaussian prior on the effect sizes. *Biometrika*, 102(4):753–766, 2015.

D. Donoho and J. Jin. Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32(3):962–994, 2004.

B. Efron. *Large-scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction.* Cambridge University Press, 2012.

A. Eklund, T. E. Nichols, and H. Knutsson. Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, 113(28):7900–7905, 2016.

R. A. Fisher. Statistical methods for research workers. *Edinburgh: Oliver and Boyd, 14th ed.*, 1970.

K. Fortney, E. Dobriban, P. Garagnani, C. Pirazzini, D. Monti, D. Mari, G. Atzmon, N. Barzilai, C. Franceschi, A. B. Owen, et al. Genome-wide scan informed by age-related disease identifies loci for exceptional human longevity. *PLoS Genetics*, 11 (12):e1005728, 2015.

K. R. Gabriel. Simultaneous test procedures–some theory of multiple comparisons. *The Annals of Mathematical Statistics*, 40(1):224–250, 1969.

J. J. Goeman and A. Solari. The sequential rejection principle of familywise error control. *The Annals of Statistics*, 38(6):3782–3810, 2010.

J. J. Goeman and A. Solari. Multiple testing for exploratory research. *Statistical Science*, 26(4):584–597, 2011.

J. J. Goeman and A. Solari. Multiple hypothesis testing in genomics. *Statistics in Medicine*, 33(11):1946–1978, 2014.

J. Gou, A. C. Tamhane, D. Xi, and D. Rom. A class of improved hybrid Hochberg–Hommel type step-up multiple test procedures. *Biometrika*, 101(4):899–911, 2014.

E. Grechanovsky and Y. Hochberg. Closed procedures are better and often admit a shortcut. *Journal of Statistical Planning and Inference*, 76(1):79–91, 1999.

O. Harismendy, D. Notani, X. Song, N. G. Rahim, B. Tanasa, N. Heintzman, B. Ren, X.-D. Fu, E. J. Topol, M. G. Rosenfeld, et al. 9p21 dna variants associated with coronary artery disease impair interferon-γ signalling response. *Nature*, 470(7333): 264, 2011.

K. S. Henning and P. H. Westfall. Closed testing in pharmaceutical research: Historical and recent developments. *Statistics in Biopharmaceutical Research*, 7(2):126–147, 2015.

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

Y. Hochberg and A. C. Tamhane. *Multiple Comparison Procedures*. Wiley, 1987.

S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

G. Hommel. A stagewise rejective multiple test procedure based on a modified bonferroni test. *Biometrika*, 75(2):383–386, 1988.

G. Hommel and F. Bretz. Aesthetics and power considerations in multiple testing–a contradiction? *Biometrical Journal*, 50(5):657–666, 2008.

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

W. Liu. Multiple tests of a non-hierarchical finite family of hypotheses. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(2):455–461, 1996.

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

W. Maurer, L. Hothorn, and W. Lehmacher. Multiple comparisons in drug clinical trials and preclinical assays: a-priori ordered hypotheses. In *Biometrie in der Chemisch-in-Pharmazeutischen Industrie*, volume 6, pages 3–18. Fischer-Verlag, 1995.

R. Meijer, T. Krebs, A. Solari, and J. Goeman. A shortcut for Hommel's procedure in linearithmic time. *arXiv preprint arXiv:1710.08273*, 2017.

E. A. Rødland. Simes' procedure is valid on average. *Biometrika*, 93(3):742–746, 2006.

J. P. Romano, A. Shaikh, and M. Wolf. Consonance and the closure method in multiple testing. *The International Journal of Biostatistics*, 7(1):1–25, 2011.

E. Samuel-Cahn. Is the Simes improved Bonferroni procedure conservative? *Biometrika*, 83(4):928–933, 1996.

S. K. Sarkar and C.-K. Chang. The Simes method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Association*, 92(440):1601–1608, 1997.

P. C. Sham and S. M. Purcell. Statistical power and significance testing in large-scale genetic studies. *Nature Reviews Genetics*, 15(5):335, 2014.

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

E. Sonnemann and H. Finner. Vollständigkeitssätze für multiple testprobleme. In *Multiple Hypothesenprüfung/Multiple Hypotheses Testing*, pages 121–135. Springer, 1988.

E. Sonnemann. *Allgemeine Lösungen multipler Testprobleme.* Universität Bern. Institut für Mathematische Statistik und Versicherungslehre, 1982.

E. Sonnemann. General solutions to multiple testing problems. *Biometrical Journal*, 50(5):641–656, 2008.

S. A. Stouffer, E. A. Suchman, L. C. DeVinney, S. A. Star, and R. M. Williams Jr. *The American soldier: Adjustment during army life. Vol. 1.* Princeton Univ. Press, 1949.

A. C. Tamhane and J. Gou. Advances in p-value based multiple test procedures. *Journal of Biopharmaceutical Statistics*, 28(1):10–27, 2018.

P. M. Visscher, M. A. Brown, M. I. McCarthy, and J. Yang. Five years of GWAS discovery. *The American Journal of Human Genetics*, 90(1):7–24, 2012.

P. H. Westfall and A. Krishen. Optimally weighted, fixed sequence and gatekeeper multiple testing procedures. *Journal of Statistical Planning and Inference*, 99(1): 25–40, 2001.

P. H. Westfall and S. S. Young. *Resampling-based multiple testing: Examples and methods for p-value adjustment.* John Wiley & Sons, 1993.

B. L. Wiens. A fixed sequence bonferroni procedure for testing multiple endpoints. *Pharmaceutical Statistics*, 2(3):211–215, 2003.

B. Wilkinson. A statistical consideration in psychological research. *Psychological bulletin*, 48(2):156, 1951.

D. V. Zaykin, L. A. Zhivotovsky, P. H. Westfall, and B. S. Weir. Truncated product method for combining p-values. *Genetic Epidemiology*, 22(2):170–185, 2002.