Exact testing with random permutations

Jesse Hemerik and Jelle Goeman
Radboudumc, The Netherlands

December 1, 2014

Abstract

The way in which random permutations have been used in various permutation-based methods leads to anti-conservativeness, especially in multiple testing contexts. Problems arise in particular for Westfall and Young's max T method, a more recent method by Meinshausen and a global test that we introduce. We illustrate this using simulations. We solve the problem of anti-conservativeness by proving that an exact test is obtained, if the identity map is added to the randomly drawn permutations.

1 Introduction

Permutation tests are nonparametric tests that can be used when under the null hypothesis the distribution of the data is invariant under certain transformations. Permutation tests are frequently used in for example omics. Not only permutations can be used, but other groups of transformations as well. Examples are rotations (Langsrud, 2005) and, when under the null the data distribution is symmetric, multiplication of part of the data by -1. We will often use the term 'permutation test' when we consider tests based on other transformations as well. Traditionally, Fisher's *Lady Tasting Tea* experiment from 1935 is mentioned as the first permutation test (Hoeffding, 1952; Ernst et al., 2004; Phipson and Smyth, 2010). As we will argue, however, the first permutation test was described in 1936.

Apart from the permutation test for a single hypothesis, there are permutation-based methods for familywise error rate (FWER) and false discovery

MSC 2010 subject classifications: Primary 62G09; secondary 62G10
Keywords and phrases: Permutation tests, Random permutations, Multiple testing, Nonparametric methods, False discovery proportion

proportion (FDP) control in multiple testing contexts. Generally, permutations of the data are used to simulate realizations of the null distribution. This has been done by Pawitan et al. (2006) to estimate the FDP in certain contexts. Well-known FWE-controlling methods that use permutations are the min$P$ and max$T$ methods by Westfall and Young (1993). Another, more recent method, which gives uniform upper bounds for the FDP, has been published by Meinshausen (2006). This method is closely related to the basic permutation test for one hypothesis, as is explained in this paper.

As we will illustrate, it is important that the set of transformations used is a group. To our knowledge, the first author who explicitly assumed a group structure is Hoeffding (1952). The importance of the group structure has only recently been emphasized, by e.g. Southworth et al. (2009) and Goeman and Solari (2010). The latter note that in particular the set of *balanced permutations* cannot be used, since it does not form a group.

Often it is computationally infeasible to use the whole group of permutations, hence random permutations are used. The set of randomly drawn permutations is usually not a group. Using random permutations was first proposed by Dwass (1957) and is often done in the various permutation-based methods. The question how and why random permutations can be used, has until now not fully been answered. Dwass (1957) gives a proof, but bases it on an incorrect assumption. Consequently his test is anti-conservative.

Corresponding to the test that Dwass defined, permutation p-values are often calculated incorrectly and are consequently too small. In particular, p-values can become zero, as e.g. Knijnenburg et al. (2009) and Phipson and Smyth (2010) note. The latter illustrate that in combination with Bonferroni's method, this can lead to excessive rejection probabilities.

The permutation-based multiple testing methods that we will consider, are also anti-conservative when random permutations are employed in the usual way (although these methods are not based on permutation p-values). For the methods of Westfall and Young and Meinshausen, the anti-conservativeness is limited if α and the number of permutations are not too small. However, some methods can become highly anti-conservative, as we will illustrate with a new global test.

Phipson and Smyth (2010) have noted and solved part of the problem of anti-conservativeness. For the usual single-hypothesis permutation test, they give an elegant, exact formula for the permutation p-values, when the random permutations are drawn without replacement. They do this under the assumption that the drawn permutations give distinct test statistics. From this formula we can see that – under this assumption – random permutations, drawn without replacement, do give an exact permutation test, as long as the identity permutation is added to the collection of randomly drawn permutations.

Until now is has not been clear whether a similar result holds when permutations are drawn *with* replacement. This is, however, what is usually done in practice. Phipson and Smyth prove a formula for the exact p-value when permutations are drawn with replacement (under the assumption that under the permutation distribution, all possible test statistics are equally likely). This formula however is quite involved and may result in increased computation times, compared to the simpler solution that we will provide. Also, the approach op Phipson and Smyth does not easily generalize to permutation-based multiple testing methods.

The main purpose of this paper is to show in a general way how random transformations, either drawn with or without replacement, can be used in a permutation test and generalizations thereof, such as Meinshausen's (2006) method and Westfall and Young's max$T$ method to obtain exact tests. The main message is that in all these methods the identity transformation should be added to the set of randomly drawn transformations, both when drawing with and without replacement. When drawn without replacement, the random transformations are not allowed to be the identity. We do not need the assumptions on the test statistics that Phipson and Smyth (2010) use. Some authors, e.g. Ge et al. (2003) and Knijnenburg et al. (2009), already added the identity, for intuitive reasons.

This paper is built up as follows. We will first discuss the importance of the group structure of the set of transformations used. We will then formulate how random permutations can be used. As examples of methods that use random permutations, we will consider a new global test, Westfall and Young's maxT-method and Meinshausen's method for FDP control. We will use simulations to show that these methods are anti-conservative, when the identity permutation is not added.

2 The role of the group structure in the permutation test

In this section we discuss the importance of the group structure of the set of transformations used for a permutation test.

2.1 In the Lady Tasting Tea experiment no group structure is needed

Many authors (Hoeffding, 1952; Ernst et al., 2004; Phipson and Smyth, 2010) mention Fisher's famous Lady Tasting Tea experiment, from his 1935 book *The Design of Experiments*, as the first permutation test in the literature. The Lady Tasting Tea experiment is designed to determine whether a lady can distinguish two types of cups of tea with milk: cups in which the milk was poured first, and cups in which the tea was poured first. While this experiment can indeed be formulated as a permutation test, there is an important difference with the usual permutation test as it has been described by Fisher (1936), Pitman (1937) and many authors since.

The setup of the experiment is as follows. The lady is given eight cups – there are four of both kinds, which the lady is told beforehand – and she has to label the cups correctly as being of the first or the second kind. There are $\binom{8}{4} = 70$ ways to label the cups. It is obvious that under the hypothesis that the lady cannot distinguish between the two types of cups, the probability that she guesses correct is $\frac{1}{70}$. Note that the lady could have a preference for picking certain patterns, but that this probability would still be $\frac{1}{70}$, as long as the researcher picks each possible pattern of cups with probability $\frac{1}{70}$.

There are $8! = 40320$ permutations of eight cups, and each pattern corresponds to $\frac{8!}{70}$ of these permutations. It is not essential to use all 70 patterns, i.e. the whole permutation group; the researcher could also use a smaller set of $n < 70$ patterns. As long as he would tell the lady from which set of patterns he randomly picked one, the lady would guess correctly with probability $\frac{1}{n}$. Thus, in the Lady Tasting Tea experiment, it is not important whether the transformations form a group. This is due to the randomization inherent in the experimental design.

2.2 For the usual permutation test a group structure is needed

For a permutation test as it is usually defined, it is less immediately obvious than for the Lady Tasting Tea experiment, that the test has the stated level: we need to use the group structure of the set of transformations. As an example of the usual permutation test, we consider a hypothetical experiment that Fisher described in 1936, which we consider to be the first permutation test. This was only a thought-experiment, since it would have been computationally infeasible at that time. The null hypothesis to be tested is that the statures of Frenchmen and Englishmen are distributed equally. To test the hypothesis, a hundred Frenchmen and a hundred Englishmen are sampled. Write the data as $X = (X_1, ..., X_{200})$, where $X_1, ..., X_{100}$ are the statures of the Frenchmen. The test statistic is

$$T(X) = |\sum_{i=1}^{100} X_i - \sum_{i=101}^{200} X_i|.$$

For all permutations of X, we compute the corresponding test statistics. Then to have an α-level test, we should reject when $T(X)$ is among the highest $\alpha \cdot 100\%$ of the computed test statistics. (This simple test can be conservative when α is not chosen suitably, or due to tied values.) Pitman descibes a mathematically analogous experiment in 1937 (p. 122). The data which he uses in his example is a vector of only eight numbers, so that he is able to actually execute the test. What these authors do not note, is that it is less obvious than in the Lady Tasting Tea Experiment, that the desired level is indeed obtained.

To prove that the usual permutation test has the desired level, the explicit use of the *group structure* of the set of permutation maps G is needed. Indeed, the group structure guarantees that $Gg = G$ for all $g \in G$ and consequently

$$\big(T^{(1)}(X), ..., T^{(M)}(X)\big) = \big(T^{(1)}(gX), ..., T^{(M)}(gX)\big),$$

where $T^{(1)} \leq ... \leq T^{(M)}$ are the ordered test statistics for the $M := \#G$ permutated versions of the data. Due to the above property, it holds that

$$T^{(k)}(X) = T^{(k)}(gX)$$

for all $g \in G$, where $k = \lceil (1-\alpha)M \rceil$. Under the null hypothesis that $X \overset{d}{=} gX$ for all $g \in G$ it follows that

$$P\big(T(X) > T^{(k)}(X)\big) = \frac{1}{M}\sum_{g \in G} P\big(T(X) > T^{(k)}(X)\big) =$$

$$\frac{1}{M}\sum_{g \in G} P\big(T(gX) > T^{(k)}(gX)\big) = \frac{1}{M}\sum_{g \in G} P\big(T(gX) > T^{(k)}(X)\big) =$$

$$\frac{1}{M}\sum_{g \in G} E\mathbb{1}_{\{T(gX) > T^{(k)}(X)\}} = \frac{1}{M} E \sum_{g \in G} \mathbb{1}_{\{T(gX) > T^{(k)}(X)\}} = \frac{1}{M}(M-k) \leq \alpha.$$

The proof we have given is essentially the same as the one that Hoeffding (1952) and Lehmann and Romano (2005) give, but spells out the role of the group structure more explicitly.

A sidenote is, that to guarantee a rejection probability of *exactly* α, it suffices to reject with a suitable probability a in the boundary case that $T(X) = T^{(k)}(X)$, as Hoeffding (1952) shows. a is given by

$$a = \frac{M\alpha - M^{+}(X)}{M^{0}(X)}, \tag{1}$$

where

$$M^{+}(X) := \#\{g \in G : T(gX) > T^{(k)}(X)\},$$
$$M^{0}(X) := \#\{g \in G : T(gX) = T^{(k)}(X)\}.$$

When the set of transformations G is *not* a group, the permutation test usually doesn't attain the desired level. Indeed, the ratio of the desired level α and the real level, can be arbitrarily large or small. This we prove with the following examples.

Take $n \geq 3$ and let $X = (X_1, ..., X_{2n})$ be a any random vector in $\mathbb{R}^{2n}$, where $X_1, ..., X_{2n}$ are continuous. Let G be the group of all permutation maps $\mathbb{R}^{2n} \to \mathbb{R}^{2n}$ of the form $(x_1, ..., x_{2n}) \mapsto (x_{i_1}, ..., x_{i_{2n}})$, where $(i_1, ..., i_{2n})$ is a permuted version of $(1, ..., 2n)$. Let H_0 be that the X_i are i.i.d.. Define the test statistic by $T(X) = \sum_{i=1}^{n} X_i - \sum_{i=n+1}^{2n} X_i$. Let $\hat{\pi} \in G$ be the permutation that interchanges the first and the $(n+1)$-th element of its argument. Take $U := \{id\} \cup A\hat{\pi}$, where

$$A := \{g \in G : \; T(gx) = T(x) \text{ for all } x \in \mathbb{R}^{2n}\}.$$

Note that $\#U = n!n! + 1$. Moreover, $x_{n+1} < x_1$ implies that $T(ux) < T(x)$ for all $u \in U \setminus \{id\}$. If we take $\alpha = \alpha_n = \frac{1}{n!n!+1}$, then for the usual permutation test, using only the permutations in U instead of all of G, under H_0, $P(\text{reject } H_0) =$

$$P\big(T(X) > T(uX) \text{ for all } u \in U \setminus \{id\}\big) =$$

$$P(x_{n+1} < x_1) = \frac{1}{2}.$$

Thus, under H_0, as $n \to \infty$, $\frac{P(\text{reject } H_0)}{\alpha_n} \to \infty$.

We now give a counterexample where $\frac{P(\text{reject } H_0)}{\alpha_n} \to 0$. Take X and H_0 as above, take as the set of permutations

$$U = \{id, \pi_1, ..., \pi_n\},$$

where π_i is the permutation map that interchanges the i-th and the $(i+n)$-th element of its argument, and let $\alpha = \alpha_n = \frac{1}{n+1}$. Then, under H_0,

$P(\text{reject } H_0) = \frac{1}{2}^n$, so $P(\text{reject } H_0) \cdot {\alpha_n}^{-1} \to 0$. In these examples we only added the identity to U to show that we get a completely wrong rejection probability, even if we add the identity. We conclude that the relative difference between the desired level and the actual level of the test, can be arbitrarily small or large if we do not require the set of transformations to be a group.

Another example of a set of permutations that is not a group is the set of *balanced permutations*, which Southworth et al. (2009) discuss. These permutations have been used in various papers since they can have an intuitive appeal. However, they tend to give rejection probabilities larger than α under H_0. Had more emphasis been put on the importance of the group structure, then the use of balanced permutations might have been prevented.

The role of the group structure has not been given the attention it deserves, since the first permutation test was formulated. This is possibly due to the confusion caused by viewing the Lady Tasting Tea experiment as a permutation test. The first writing we know of that explicitly assumes that the transformations are a group is by Hoeffding (1952). He does not make the importance of the group structure very explicit. Southworth et al. (2009) show this role explicitly. Goeman and Solari (2011) explicitly show the role of the group structure in the permutation-based maxT method by Westfall and Young.

3 Exact test with random permutations

It is often computionally infeasible to use the whole group of permutations in a permutation test. A vector of 20 numbers can already be permuted in $20! \approx 2.4 \cdot 10^{18}$ ways. When the test statistic is given by $T(x_1, ..., x_{20}) = x_1 + ... + x_{10}$ and we use only one permutation from each class of $10!10!$ equivalent permutations, then we still use $\binom{20}{10} = 184756$ different permutations.

A possible solution is to use a *subgroup* of the whole set of transformations. A simple example is to partition the data vector into m subvectors of equal size and to consider the $m!$ permutations of these chunks of data. In practice, researchers choose to use random permutations to limit the computation time. As Phipson and Smyth (2010) explain, the way in which random permutations have usually been used in permutation tests, leads to anti-conservativeness.

Applying Bonferroni's method to a set of p-values as they are often calculated, can lead to much too large rejection probabilities. In the following sections we will discuss three other permutation-based multiple testing methods that become anti-conservative when random permutations are employed in the usual way.

In this section, we discuss the use of random permutations in the basic permutation test. We first discuss Phipson and Smyth's formulas for exact permutation p-values. Then in Theorem 3.1 we formulate a basic test with random permutations which generalizes some of Phipson and Smyth's results. In particular we do not assume that all permutations drawn give distinct test statistics and we allow the permutations to be drawn with replacement. The multiple testing methods considered in the following sections are based on this basic test.

3.1 Exact p-values

Phipson and Smyth (2010) provide a simple formula for exact permutation p-values, for when the permutations are drawn *without replacement*. The (slightly generalized) setting is as follows. Let X be data with any distribution, T a test statistic and G a finite group of transformations from and to the range of X. Assume that all these transformations in G give distinct test statistics. This is in a sense a large assumption, even when the data are continuous, since there are often equivalent transformations that *always* give the same test statistic. However, e.g. in case that the test statistic is the difference of the averages of two groups of size m and n, it can be shown that we are allowed to take G such that it contains exactly one element from each of the classes of equivalent permutations. We then essentially take G to be the $\binom{m+n}{m}$ different relabellings.

The formula for the permutation p-value that researchers (e.g. Dwass) typically use is

$$p = \frac{b}{m},$$

where b is the number of random permutations g in G for which $T(gX) \geq T(X)$ and m the total number of random permutations. However, as Phipson and Smyth explain, the correct formula is (under the above assumption)

$$p = \frac{b+1}{m+1},$$

where the definition of b remains the same except that the random permutations are drawn from $G \setminus \{id\}$. Note that if we would add the identity transformation to the random permutations, then this would be equivalent to adding +1 in the numerator and the denominator of $\frac{b}{m}$. But this implies that (under the assumption that all transformations in G give distinct test statistics) we can use random transformations, drawn without replacement from $G \setminus \{id\}$, in the basic permutation test, if we add the identity transformation.

As Phipson and Smyth note, the formula for the permutation p-value above is analogous to the formula for the Monte Carlo p-value. This is because in both settings, b has the uniform distribution on $\{0, ..., m\}$. We should add however that the reason why b is uniform in the permutation context is quite different from why it is uniform in the Monte Carlo setting: that b is uniform in the permutation context is a consequence of the group structure of the set from which the random permutations have been drawn.

Phipson and Smyth also provide a formula for the p-value in the case of drawing *with* replacement. In this case the random permutations are drawn from G instead of $G \setminus \{id\}$. The assumption that is necessary in this case is that under the permutation distribution there are m_t possible distinct values of the test statistic (not counting the original value) and all (including the original value) are equally likely. This is often satisfied, due to equally sized classes of equivalent permutations. The formula is computationally involved though. It is not necessary to use this p-value, due to Theorem 3.1. For this theorem we do not need the assumptions on the test statistics, that Phipson and Smyth use.

3.2 The exact test

We show in Theorem 3.1 how random transformations can be used in the basic permutation test for a single hypothesis. We first define the vector of random transformations.

Definition 3.1. Given a finite group of transformations G, let G' be the vector $(id, g_2, ..., g_w)$, where id is the identity in G and $g_2, ..., g_w$ are random elements from G. Write $id =: g_1$. The transformations can be drawn either with or without replacement: the statements in this paper hold for both cases. If we draw $g_2, ...g_w$ *without* replacement, then we take them to be uniformly distributed on $G \setminus \{id\}$, otherwise uniform on G.

Theorem 3.1. *Let X be data with any distribution. Suppose G is a finite group (under composition of maps) of measurable transformations from and to the range of X. Let G' be as in Definition 3.1.*

Let T be a test statistic on the range of X. Let $T^{(1)}(X, G') \leq ... \leq T^{(w)}(X, G')$ be the ordered test statistics $T(g_i' X)$, $1 \leq i \leq w$. Let $\alpha \in [0, 1]$ and $k = \lceil (1 - \alpha) w \rceil$.

Let H_0 be a null hypothesis such that if H_0 is true, then the joint distribution of the test statistics $T(gX)$, $g \in G$, is invariant under all transformations in G of X. This holds in particular if $X \overset{d}{=} gX$ for all $g \in G$. Reject H_0 when $T(X, G') > T^{(k)}(X, G')$. Then the rejection probability under H_0 is at most α.

Proof. From the group structure of G, it follows that for all $1 \leq j \leq w$, $G' g_j^{-1}$ and G' have the same distribution, if we disregard the order of the elements. Let j have the uniform distribution on $\{1, ..., w\}$ and write $h = g_j$. Under H_0,

$$\begin{aligned}
&P\big(T(X) > T^{(k)}(X, G')\big) = \\
&P\big(T(X) > T^{(k)}(X, G'h^{-1})\big) = \\
&P\big(T(hX) > T^{(k)}(hX, G'h^{-1})\big).
\end{aligned}$$

Since $(G'h^{-1})(hX) = G'(h^{-1}hX)$, the above equals

$$\begin{aligned}
&P\big(T(hX) > T^{(k)}(h^{-1}hX, G')\big) = \\
&P\big(T(hX) > T^{(k)}(X, G')\big) \leq \alpha.
\end{aligned}$$

□

We end this section with some remarks. The test above can be slightly conservative if w and α are not chosen suitably or due to tied values of the test statistics. However, the level will be exactly α if we randomly reject with a suitable probability $a = a(X, G')$ in the boundary case that $T(X) = T^{(k)}(X, G')$. a is given in (1), with the adjustment that w now takes the role of M and G' takes the role of G.

It is possible to define a permutation test that needs no group structure whatsoever, by slightly randomizing the rejection rule as follows. Let G be *any* finite, nonempty set of invertible transformations from and to the range of X. Let h have the uniform distribution on G. Let T, $T^{(k)}$ and H_0 be as usual. Reject H_0 only if

$$T(X) > T^{(k)}(X, Gh^{-1}).$$

This is a randomized rejection rule, since it depends on h, which is randomly drawn each time the test is executed. The rejection probability is at most α, which follows from an argument analogous to the last four steps of the proof of Theorem 3.1. Note that if G *is* a group, then $Gh^{-1} = G$ and this test coincides with the basic permutation test. Thus it is a generalization thereof.

In case the transformations are drawn with replacement in Theorem 3.1, a simple upper bound for the permutation p-value is

$$\frac{\#\{1 \leq i \leq w : T(g_i X) \geq T(X)\}}{w}.$$

As discussed, Phipson and Smyth (2010) give formulas for the exact p-value for both the cases of drawing with and without replacement, under certain additional assumptions. The formula for the p-value for the case of drawing with replacement is rather involved. We can also use an alternative definition of the p-value, which is simpler. Consider the exact variant of the test in Theorem 3.1 that rejects with a suitable probability a when $T(X) = T^{(k)}(X, G')$. Suppose w.l.o.g. that when $T(X) = T^{(k)}$, the test rejects if and only if $a > u$, where u is uniform on $[0, 1]$ and independent. Denote the rejection function, which takes the values 0 and 1, by $\phi_\alpha = \phi_\alpha(X, G')$. Define the p-value by

$$\tilde{p} = \frac{\#\{1 \leq i \leq w : T(g'_i X) > T(X)\}}{w} + u \cdot \frac{\#\{1 \leq i \leq w : T(g'_i X) = T(X)\}}{w}.$$

This formula is simple indeed. This p-value depends on u, so it is randomized. This is in itself not objectionable when random transformations are used, since the p-value is randomized anyway due to the random transformations. It can be seen as follows that under H_0, $\tilde{p}$ has the uniform distribution on $[0, 1]$. Observe that given u, for all $c \in [0, 1]$, $P(\tilde{p} \leq c) = P(\phi_c = 1)$. Thus, under H_0,

$$P(\tilde{p} \leq c) = \int_0^1 P\big(\phi_c = 1 | u = y\big) dy = P(\phi_c = 1) = c.$$

A result analogous to Theorem 3.1 holds for Monte Carlo testing: the original observation should be added to the random draws from the null distribution. Again we do not need to assume that the test statistics are distinct. We can again guarantee exactness of the test by introducing a randomized decision.

4 Application: a permutation-based global test

In certain permutation-based multiple testing contexts, adding the identity permutation can be of great importance. The anti-conservativeness can become huge when we do not add the identity, even if we use many random permutations. To illustrate this, we consider the following procedure. It is perhaps the most obvious permutation-based global test, next to Westfall and Young's single-step maxT method. In fact, we used this method ourselves, before we found that we needed to add the identity. Therefore this section can serve as a caveat to the research community. It illustrates that a seemingly reasonable method may become wildly anti-conservative if random permutations are used incorrectly.

Consider data X with any distribution, hypotheses $H_1, ..., H_m$ and corresponding p-values $P_1(X)$,...,$P_m(X)$. Let G be a group of transformations from and to the range of X. Suppose that $H_0 := \cap_{i=1}^{m} H_i$ implies that $X \overset{d}{=} gX$ for all $g \in G$. The method that we will define permutes the data w times, each time obtaining a p-value curve. In this manner a permutation distribution is obtained for the p-value curve. Next a critical curve $\mathbf{k}$ is constructed with the property that $\lceil (1-\alpha)w \rceil$ of the w p-value curves lie everywhere above $\mathbf{k}$. Thus, under the permutation distribution, the probability that the p-value curve lies below $\mathbf{k}$ somewhere is at most α. Assuming that under H_0 the permutation distribution is a good approximation of the true distribution, the probability under H_0 that the original p-value curve lies below $\mathbf{k}$ somewhere is at most α. Thus rejecting H_0 when the original p-value curve lies below $\mathbf{k}$ somewhere gives a valid α-level global test. See figure 1 for a visualization of the method.

The precise definition of the method is the following. Let G' be as defined in Definition 3.1. Define a critical curve $\mathbf{k}(X, G') \in \mathbb{R}^m$ as follows. For each $1 \leq j \leq w$, consider the corresponding curve of sorted p-values,

$$\big(P_{(1)}(g_j X), ..., P_{(m)}(g_j X)\big).$$

Pick $r \in \{1, ..., w\}$ such that $\#J \geq (1-\alpha)w$, where

$$J := \big\{1 \leq j \leq w : P_{(\lceil \frac{m}{2} \rceil)}(g_j X) \geq P_{(\lceil \frac{m}{2} \rceil)}(g_r X)\big\}.$$

Define $\mathbf{k}(X, G') = (\mathbf{k}_1, ..., \mathbf{k}_m)$ by

$$\mathbf{k}_i := \min\{P_{(i)}(g_j X) : j \in J\}.$$

Note that at least $(1-\alpha)100\%$ of the p-value curves lie everywhere above $\mathbf{k}$. Reject H_0 when

$$P_{(i)}(X) < \mathbf{k}_i \text{ for at least one } 1 \leq i \leq m.$$

Suppose we do not add the identity, i.e. we let g_1 be random instead of taking $g_1 = id$. As we will see in section 7.1, the method is then very

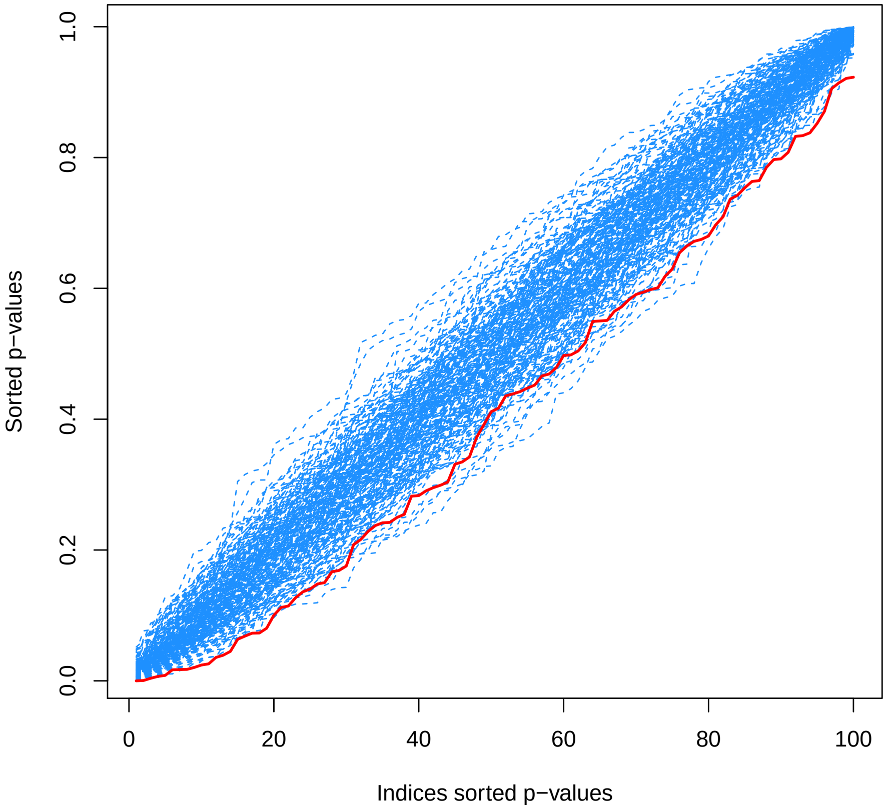

Figure 1: *The critical curve* $\mathbf{k}$ *(red and solid) of the global test is the pointwise minimum of the p-value curves with indices in* J*. The global null hypothesis is rejected only if the original p-value curve lies below* $\mathbf{k}$ *somewhere.*

anti-conservative. The method is the most anti-conservative when there are many hypotheses and the number of permutations is limited, but even if the number of hypotheses is small (e.g. $m = 50$) and the number of permutations large (e.g. $w = 1000$), the anti-conservativess is substantial. We conclude that the permutation distribution is apparently not a good approximation of the true distribution in this context.

If we take g_1 te be the identity, then the method becomes correct. In this case the method can actually be rewritten in a simpler form: it becomes a basic permutation test as defined in Theorem 3.1. Indeed, define the test statistic T' on the range of X by $T'(X) = -P_{(\lceil \frac{m}{2} \rceil)}(X)$. Let

$$T'^{(1)}(X, G') \leq ... \leq T'^{(w)}(X, G')$$

be the sorted test statistics $T'(g_i X)$, $1 \leq i \leq w$. Note that the method rejects H_0 if and only if $T'(X) > T'^{(\#J)}(X, G')$, i.e. it is a basic permutation test as defined in Theorem 3.1. As noted below that theorem, the test can be easily made exact by introducing a randomized decision.

The reason that the test becomes correct when we add the identity, is not that under H_0 the permutation distribution is then a better approximation of the true distribution. Instead, the method becomes correct due to the fact that the original p-value curve is among the w curves and under H_0 all p-value curves have the same probability of lying below $\mathbf{k}$ somewhere. The latter is in turn a consequence of the group structure.

5 Application: Westfall and Young's max T-method

Westfall and Young's max T-method is a well-known procedure for strong FWER control, which uses no assumptions on the correlations between the p-values. Beside the single-step variant of this method, there is also a less conservative, sequential variant, which has been shown to be asymptotically optimal in certain cases (Meinshausen et al., 2011). The sequential method coincides with a closed testing procedure where for each intersection hypotheses the single-step test is used.

In the single-step method, all hypotheses among $H_1, ..., H_m$ are rejected for which the test statistic is is higher than the $(1-\alpha)$-quantile of the global null distribution of $\max_{i=1}^{m} T_i$, where the T_i are the test statistics corresponding to the H_i. Often though the global null hypothesis doesn't imply a specific distribution of $\max_{i=1}^{m} T_i$, such that this method can't be used. However if the data corresponding to the true hypotheses are permutation invariant, we can instead use the permutation distribution of $\max_{i=1}^{m} T_i$.

If the identity map is not added to the random permutations though, the method can become anti-conservative, especially when the number of permutations is limited. In section 7 we use simulations to show this. It follows that the – less conservative – sequential procedure must also become anti-conservative.

Using Theorem 3.1, we now show that the single-step method becomes correct if we add the identity to the vector of random transformations.

Theorem 5.1. *Let X be data with any distribution and $\alpha \in [0,1]$. Let G be a finite group of transformations from and to the range of X. Consider hypotheses $H_1, ..., H_m$ and let $\mathcal{N} \subseteq \{1, ..., m\}$ be the indices of the true hypotheses. Suppose that for each $1 \le i \le m$ a test statistic $T_i(X)$ for H_i is defined.*

Assume that the null hypotheses are such that the joint *distribution of the test statistics $T_i(gX)$ with $i \in \mathcal{N}$, $g \in G$, is invariant under all transformations in G of the data X.*

Let G' be as in Definition 3.1. For each $1 \le j \le w$, consider $\mu(g_jX) := \max\{T_i(g_jX) : 1 \le i \le m\}$. Let

$$\mu^{(1)}(X, G') \le ... \le \mu^{(w)}(X, G')$$

be the w sorted maxima $\mu(g_jX)$, $1 \le j \le w$.

Let $k = \lceil (1-\alpha)w \rceil$ and reject the hypotheses H_i with $T_i(X) > \mu^{(k)}(X, G')$. Then the FWER is at most α.

Proof. When $\mathcal{N} = \emptyset$ there are no false positives, so assume $\mathcal{N} \neq \emptyset$. For each $1 \le j \le w$ consider

$$\mu_{\mathcal{N}}(g_jX) := \max\{T_i(g_jX) : i \in \mathcal{N}\}.$$

Let

$$\mu_{\mathcal{N}}^{(1)}(X, G') \le ... \le \mu_{\mathcal{N}}^{(w)}(X, G')$$

be the sorted values $\mu_{\mathcal{N}}(g_jX)$, $1 \le j \le w$. Theorem 3.1 implies that

$$P\big(\mu_{\mathcal{N}}(X) > \mu_{\mathcal{N}}^{(k)}(X, G')\big) \le \alpha. \tag{2}$$

Note that for each $1 \le j \le w$, $\mu(g_jX) \ge \mu_{\mathcal{N}}(g_jX)$. Consequently,

$$\mu_{\mathcal{N}}^{(k)}(X, G') \le \mu^{(k)}(X, G'). \tag{3}$$

(2) and (3) imply that

$$P\big(\mu_{\mathcal{N}}(X) > \mu^{(k)}(X, G')\big) \le \alpha,$$

which means that the FWER is at most α.

□

In the sequential variant of the $\max T$ method, random permutations can be used in the same way. The false rejection probability of the above method is exactly α if all hypotheses are true, with probability one there are no ties among the test statistics and $\alpha \in \{\frac{0}{w}, \frac{1}{w}, ..., \frac{w}{w}\}$. If the latter two conditions are not both satisfied, then the rejection probability becomes α if we reject the hypotheses H_i with $T_i(X) = \mu^{(k)}(X, G')$ with a suitable probability analogous to (1).

6 Application: Meinshausen's method for uniform FDP control

Meinshausen's (2006) procedure is a permutation-based multiple testing method which provides lower bounds for the number of true discoveries. The bounds are uniform over all rejected sets of the form

$$R(t) = \{k \in \{1, ..., m\} : \ P_k \leq t\},$$

where $P_1, ..., P_m$ are the p-values. More precisely, the method provides a lower bound $\underline{S}(t)$ for the number of correct rejections $S(t)$, such that with probability at least $1-\alpha$, $\underline{S}(t) \leq S(t)$ for all $t \in [0,1]$ simultaneously. Since the lower bound is uniform, post hoc selection of t is allowed. Hence the method is well-suited for exploratory research.

The method makes no assumptions on the correlation structure of the p-values. The procedure is based on the assumption that the distribution of the p-values for the true hypotheses is invariant under a group of transformations. The method transforms the data many times, obtaining a collection of p-value curves. Like the global test of section 4, Meinshausen's method constructs a critical curve, $\mathbf{q}$, such that $(1-\alpha)100\%$ of the p-value curves lie everywhere above $\mathbf{q}$. The lower bound $\underline{S}(t)$ depends on the horizontal distance between $\mathbf{q}$ and the original p-value curve: see figure 2.

We will state Meinshausen's method as set forth in his paper. We make two adjustments. Firstly, we add the identity transformation to the collection of randomly drawn transformations. In section 7 we show that the method can otherwise be anti-conservative. Secondly, we pick the critical curve $Q^{l(\alpha)}$ from a different set of candidate curves, which do not depend on the data. This adjustment is needed for the following reason. In Meinshausen's proof, Q and consequently the critical curve $Q^{\tilde{l}(\alpha)}$ depend not only on the p-values P_i with $i \in \mathcal{N}$, but also on the P_i with $i \notin \mathcal{N}$. Even though the distributions of $\tilde{P}$ and $Q^{\tilde{l}(\alpha)}$ are invariant under permutation of the data, the distribution of $(\tilde{P}, Q^{\tilde{l}(\alpha)})$ is not. Consequently, from the fact that $(1-\alpha)100\%$ of the $\tilde{P}$-curves lie above $Q^{\tilde{l}(\alpha)}$ does not follow that one particular curve lies above $Q^{\tilde{l}(\alpha)}$ with probability $1-\alpha$.

The lower bound $\underline{S}(t)$ that Meinshausen's method constructs, is an example of a bound that Goeman and Solari (2011) derive for a general class of closed testing procedures. We do not explicitly use closed testing in the proof, but we could do this, obtaining the lower bound by using shortcut (7) in section 4.2 of Goeman and Solari (2011).

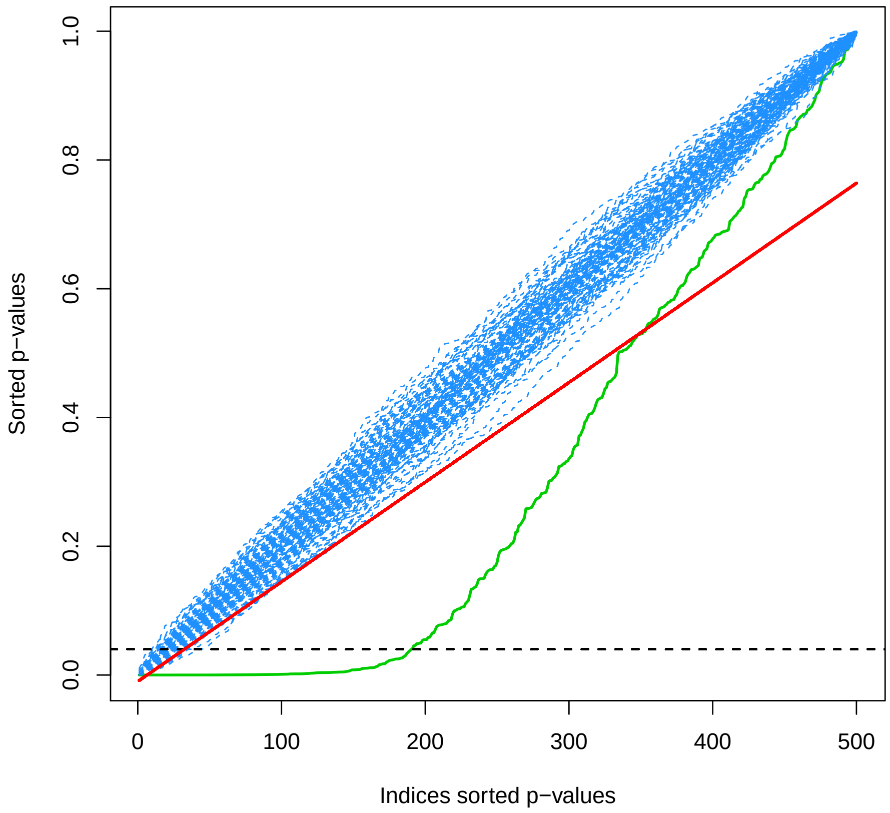

Figure 2: *This figure illustrates Meinshausen's method. The original p-value curve (green and solid) and the $w - 1$ other p-value curves (dashed) are shown. The critical curve* $\mathbf{q}$ *(red and solid) is by definition the highest curve in the set F of candidate curves with the property that $(1 - \alpha)100\%$ of the p-value curves lie above it. In this example we took F to be many straight lines with intercept -0.01. For $t = 0.04$, the lower bound $\underline{S}(t)$ is the length of the longest horizontal line segment that lies between* $\mathbf{q}$ *and the original p-value curve and below the dotted line.*

Assumptions

Let X be data with any distribution and $\alpha \in [0, 1]$. Let G be a finite group of transformations from and to the range of X. Consider hypotheses

$H_1, ..., H_m$ and corresponding p-values $P_1, ..., P_m$. Let $\mathcal{N} \subseteq \{1, ..., m\}$ be the indices of the true hypotheses. Suppose that the null hypotheses are such that the *joint* distribution of the p-values $P_i(gX)$ with $i \in \mathcal{N}$, $g \in G$, is invariant under all transformations in G of the data X.

Reject the hypotheses with indices in

$$R(t) = \{k \in \{1, ..., m\} : \ P_k \leq t\}.$$

Let $S(t)$ be the number of correct rejections.

The algorithm

Under the above assumptions, the following algorithm gives the lower bound $\underline{S}(t)$. Let G' be as in Definition 3.1. For each $1 \leq j \leq w$, let $P_{(1)}(g_jX) \leq ... \leq P_{(m)}(g_jX)$ be the m sorted p-values found after applying transformation g_j to the data. Write $P(g_jX) = (P_{(1)}(g_jX), ..., P_{(m)}(g_jX))$.

Let F be subset of $[0,1]^m$, independent of the data. F is a family of candidate 'curves'. Suppose that it is of the form $F = \{f^\gamma : \gamma \in \Gamma\}$, where $\Gamma \subset \mathbb{R}$ is bounded and closed and f^γ depends continuously on γ. Suppose that $\gamma_1 \leq \gamma_2$ implies that $f^{\gamma_1} \leq f^{\gamma_2}$, i.e. $f_k^{\gamma_1} \leq f_k^{\gamma_2}$ for all $1 \leq k \leq m$.

For each $\gamma \in \Gamma$ define

$$\beta(\gamma) = \frac{\#\{1 \leq i \leq w : P(h_iX) \geq f^\gamma\}}{w}.$$

Let

$$\gamma_{\max} := \max\{\gamma : \beta(\gamma) \geq 1 - \alpha\},$$

the largest γ for which $\beta(\gamma)$ is still at least $1 - \alpha$. Let $\mathbf{q} = f^{\gamma_{\max}}$.

For all $t \in [0,1]$, let

$$B(t) := \#\{k \in \{1, ..., m\} : \ \mathbf{q}_k \leq t\}. \tag{4}$$

Define the lower bound as

$$\underline{S}(t) := \max_{0 \leq \tau \leq t} \#R(\tau) - B(\tau).$$

Theorem 6.1. *Define $V(t) = \#R(t) - S(t)$ to be the number of false rejections. Let $\underline{S}(t)$ be as above and define $\overline{V}(t) = \#R(t) - \underline{S}(t)$. Under the assumptions above,*

$$P\big(S(t) \geq \underline{S}(t) \textit{ for all } t \in [0,1]\big) \geq 1 - \alpha,$$

$$P\big(V(t) \leq \overline{V}(t) \textit{ for all } t \in [0,1]\big) \geq 1 - \alpha.$$

Proof. Let $P^{\mathcal{N}}(X) = (P^{\mathcal{N}}_{(1)}(X), ..., P^{\mathcal{N}}_{(\#\mathcal{N})}(X))$ be the vector containing the sorted p-values $P_i(X)$ with $i \in \mathcal{N}$. Define the test statistic T by

$$T(X) = \max\{\gamma \in \Gamma : P_{\mathcal{N}}(X) \geq f^{\gamma}\}.$$

Let

$$T^{(1)}(X, G') \leq ... \leq T^{(w)}(X, G')$$

be the sorted test statistics $T(g_j X)$, $1 \leq j \leq w$.

Note that

$$\frac{\#\big\{j : T(g_j X) < T^{(\lfloor \alpha w \rfloor + 1)}(X, G')\big\}}{w} \leq \alpha.$$

It easily follows with Theorem 3.1 that

$$P\big(T(X) < T^{\lfloor \alpha w \rfloor + 1}(X, G')\big) \leq \alpha.$$

Define

$$F^{\mathcal{N}} := \{(f_1, ..., f_{\#\mathcal{N}}) : (f_1, ..., f_m) \in F\}$$

and let $\mathbf{q}^{\mathcal{N}} = (\mathbf{q}^{\mathcal{N}}_1, ..., \mathbf{q}^{\mathcal{N}}_{\#\mathcal{N}})$ be the critical curve $\mathbf{q}$ we would have found if we had used the p-value curves $P^{\mathcal{N}}(g_j X)$ instead of the $P(g_j X)$ and the set of candidate curves $F^{\mathcal{N}}$ instead of F.

Observe that

$$P^{\mathcal{N}}(X) \ngeq \mathbf{q}^{\mathcal{N}}(X, G') \implies T(X) < T^{\lfloor \alpha w \rfloor + 1}(X, G'),$$

where $P^{\mathcal{N}} \ngeq \mathbf{q}^{\mathcal{N}}$ means that $P^{\mathcal{N}}_i < \mathbf{q}^{\mathcal{N}}_i$ for at least one $1 \leq i \leq \#\mathcal{N}$. Hence $P(P^{\mathcal{N}}(X) \ngeq \mathbf{q}^{\mathcal{N}}(X, G')) \leq \alpha$, i.e.

$$P\big(P^{\mathcal{N}}(X) \geq \mathbf{q}^{\mathcal{N}}(X, G')\big) \geq 1 - \alpha. \tag{5}$$

For all $1 \leq k \leq \#\mathcal{N}$, $P^{\mathcal{N}}_{(k)}(g_j X) \geq P_{(k)}(g_j X)$, and consequently

$$\mathbf{q}^{\mathcal{N}}_k(X, G') \geq \mathbf{q}_k(X, G') \text{ for all } 1 \leq k \leq \#\mathcal{N}. \tag{6}$$

From (5) and (6) follows that

$$P\big(P^{\mathcal{N}}_{(k)}(X) \geq \mathbf{q}_k(X, G') \text{ for all } 1 \leq k \leq \#\mathcal{N}\big) \geq 1 - \alpha.$$

This implies that

$$P\big(V(t) \leq B(t) \text{ for all } t \in [0, 1]\big) \geq 1 - \alpha.$$

Hence

$$P\big(\#R(t) - V(t) \geq \#R(t) - B(t) \text{ for all } t\big) \geq 1 - \alpha.$$

$S(t)$ is monotonously increasing in t, so

$$S(t) = \max_{0 \leq \tau \leq t} S(\tau) = \max_{0 \leq \tau \leq t} \#R(\tau) - V(\tau).$$

Thus

$$P\big(S(t) \geq \max_{0 \leq \tau \leq t} \#R(\tau) - B(\tau) \text{ for all } t\big) \geq 1 - \alpha,$$

as we wanted to show.

□

7 Simulations

From the simulation results that follow we will see that the methods of the previous sections can become anti-conservative when the identity isn't added to the random transformations. The simulations were done in R. The goal of this section is not simulate realistic data, but to show in a simple way that the methods considered are incorrect when we do not add the identity map.

7.1 Permutation-based global test

As data we used an $m \times 20$-matrix (X_{ij}) of independent standard normally distributed variables. For each $1 \leq i \leq m$, the hypothesis H_i that the random variables in the i-th row were i.i.d., was tested. Thus all hypotheses were true. For each H_i we calculated the p-value as the probability under H_i of a larger value of $\sum_{j=1}^{10} X_{ij} - \sum_{j=11}^{20} X_{ij}$ than observed. As the group of transformations we used the 20! maps that shuffle the columns of (X_{ij}).

For different values of α, the number of hypotheses m and the number of permutations w, we tested $\cap_{1 \leq i \leq m} H_i$ using the global test of section 4. First we used w random permutations and recorded whether the method rejected the global hypothesis. We then substituted the identity map for the first permutation and did the same. We generated (X_{ij}) many times, each time recording whether the method rejected, to obtain an estimate of the rejection probability. The estimated rejection probabilities (divided by α) are shown in Table 1. From this table we see that the rejection probability was much too large when we did not add the identity. The anti-conservativeness seems to increase as w decreases or m increases. When α decreases, the relative anti-conservativeness seems to increase.

7.2 Westfall and Young's max method

We simulated data using the same distribution as for the permutation-based global test. We used the same group of transformations. The test statistic we used for H_i was $\sum_{j=1}^{10} X_{ij} - \sum_{j=11}^{20} X_{ij}$. In the same way as before we estimated the error rate for different values of w, m and α. The results are shown in Table 2. They suggest that the anti-conservativeness increases as w decreases, and that the relative anti-conservativeness increases as α decreases.

The results suggest that the method is only substantially anti-conservative when few permutations are used. Researchers in various fields use few permutations in permutation tests for computational reasons. Examples of recent papers using few permutations are Byrne et al. (2013) (100 permutations) and Schimanski et al. (2013) (25 to 100 permutations).

α	m	w	without *id*	with *id*
0.1	50	20	$5.32 \pm .10$	$1.00 \pm .06$
0.1	50	200	$1.90 \pm .08$	$.98 \pm .06$
0.1	50	1000	$1.31 \pm .07$	$1.06 \pm .06$
0.1	500	20	$7.22 \pm .09$	$.96 \pm .06$
0.1	500	200	$2.83 \pm .09$	$.98 \pm .06$
0.1	500	1000	$1.53 \pm .07$	$.98 \pm .06$
0.01	50	20	$42.8 \pm .3$	0
0.01	50	200	$10.0 \pm .2$	$1.0 \pm .1$
0.01	500	20	$63.2 \pm .3$	0
0.01	500	200	$18.3 \pm .2$	$1.0 \pm .1$

Table 1: 95%-confidence intervals for (error rate) $\cdot \alpha^{-1}$ for the global test

7.3 Meinshausen's method

We simulated data using the same distribution as for the permutation-based global test and used the same transformations. Again we calculated the p-value for H_i as the probability of a larger value of $\sum_{j=1}^{10} X_{ij} - \sum_{j=11}^{20} X_{ij}$ than observed. As the family F of candidate curves, we used many straight lines through the origin. We took $t = 0.2$. In the same way as before we estimated the error rate (the probability that $\underline{S}(t) > 0$) for different values of w, m and α. The results are shown in Table 3. Again they suggest that the anti-conservativeness increases as w decreases, and the relative anti-conservativeness increases as α decreases.

α	m	w	without id	with id
0.1	50	20	$1.41 \pm .02$	$1.00 \pm .02$
0.1	50	200	$1.04 \pm .02$	$1.00 \pm .02$
0.1	500	20	$1.43 \pm .02$	$1.00 \pm .02$
0.1	500	200	$1.04 \pm .02$	$.99 \pm .02$
0.01	50	20	$4.8 \pm .1$	0
0.01	50	200	$1.5 \pm .1$	$1.0 \pm .1$
0.01	500	20	$4.7 \pm .1$	0
0.01	500	200	$1.5 \pm .1$	$1.1 \pm .1$

Table 2: 95%-confidence intervals for $\text{FWER} \cdot \alpha^{-1}$ for the single-step maxT method

α	m	w	without id	with id
0.1	50	20	$1.44 \pm .07$	$1.02 \pm .06$
0.1	50	100	$1.05 \pm .04$	$.98 \pm .04$
0.1	500	20	$1.45 \pm .07$	$1.01 \pm .06$
0.1	500	100	$1.05 \pm .04$	$.98 \pm .04$
0.01	50	20	$4.0 \pm .4$	0
0.01	50	100	$2.0 \pm .3$	$.9 \pm .2$
0.01	500	20	$4.9 \pm .4$	0
0.01	500	100	$1.9 \pm .2$	$1.0 \pm .2$

Table 3: 95%-confidence intervals for (error rate) $\cdot \alpha^{-1}$ for Meinshausen's method

8 Discussion

There are various multiple testing methods that are based on the well-known permutation test. We have discussed three such methods: a new global test, Westfall and Young's max$T$ method for strong FWER control and Meinshausen's method for uniform FDP control. In areas such as omics, the number of hypotheses is often huge. Using the whole group of permutations or other transformations, would then be computationally infeasible. Thus researchers often use random permutations to limit the computation time.

The way in which random permutations have been used however can lead to anti-conservativeness. The anti-conservativeness is minor if a single hypothesis is tested and many permutations are used, but can otherwise be substantial. The anti-conservativeness is limited for Meinshausen's method and Westfall and Young's method when α is not too small and many random permutations are used, but can in other situations be excessive. An example is our permutation-based global test. The cause of the anti-conservativeness is that the permutation distribution is not a good enough approximation of the true distribution.

We have shown that all the permutation-based methods considered become correct when the identity map is added to the random transformations. A fundamental assumption is the group structure of the set from which the random transformations are drawn.

Permutation-based methods can only be used when the null hypotheses imply invariance of the data distribution under transformation. There are many applications where this is the case. Often researchers use random permutations instead of the whole group. This influences the power of the methods, however. Beside using random permutations, a subgroup of the set of all permutations could be used. The question remains open what works best in which situation. As we remarked in section 3.2, we can in fact use any set of invertible transformations – not just groups – if we slightly adapt the test. In particular, when we use permutations we can use any subset of the permutation group. The question how such a subset should be chosen remains an open problem.

In the methods discussed, instead of permutation sampling we could use Monte Carlo sampling. As we remarked in section 3.2, the original observation then needs to be included with the random draws from the null distribution. If the original observation is not included, then we expect problems similar to when permutations are used.

Acknowledgements

We thank Aldo Solari and Vincent van der Noort for their valuable suggestions.

References

Byrne, E., Carrillo-Roa, T., Henders, A., Bowdler, L., McRae, A., Heath, A., Martin, N., Montgomery, G., Krause, L., and Wray, N. Monozygotic twins affected with major depressive disorder have greater variance in methylation than their unaffected co-twin. *Translational psychiatry*, 3(6): e269, 2013.

Dwass, M. Modified randomization tests for nonparametric hypotheses. *The Annals of Mathematical Statistics*, 28:pp. 181–187, 1957.

Ernst, M. D. et al. Permutation methods: a basis for exact inference. *Statistical Science*, 19(4):676–685, 2004.

Fisher, R. A. *The design of experiments.* Oliver and Boyd, 1935.

Fisher, R. A. "the coefficient of racial likeness" and the future of craniometry. *Journal of the Anthropological Institute of Great Britain and Ireland*, 66: 57–63, 1936.

Ge, Y., Dudoit, S., and Speed, T. P. Resampling-based multiple testing for microarray data analysis. *Test*, 12(1):1–77, 2003.

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

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

Hoeffding, W. The large-sample power of tests based on permutations of observations. *The Annals of Mathematical Statistics*, 23:169–192, 1952.

Knijnenburg, T. A., Wessels, L. F., Reinders, M. J., and Shmulevich, I. Fewer permutations, more accurate p-values. *Bioinformatics*, 25(12):i161–i168, 2009.

Langsrud, Ø. Rotation tests. *Statistics and computing*, 15(1):53–60, 2005.

Lehmann, E. L. and Romano, J. P. *Testing statistical hypotheses.* Springer Science & Business Media, 2005.

Meinshausen, N. False discovery control for multiple tests of association under general dependence. *Scandinavian Journal of Statistics*, 33(2):227–237, 2006.

Meinshausen, N., Maathuis, M. H., Bühlmann, P., et al. Asymptotic optimality of the westfall–young permutation procedure for multiple testing under dependence. *The Annals of Statistics*, 39(6):3369–3391, 2011.

Pawitan, Y., Calza, S., and Ploner, A. Estimation of false discovery proportion under general dependence. *Bioinformatics*, 22(24):3025–3031, 2006.

Phipson, B. and Smyth, G. K. Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Statistical applications in genetics and molecular biology*, 9(1):39, 2010.

Pitman, E. J. G. Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society*, 4 (1):119–130, 1937.

Schimanski, L. A., Lipa, P., and Barnes, C. A. Tracking the course of hippocampal representations during learning: when is the map required? *The Journal of Neuroscience*, 33(7):3094–3106, 2013.

Southworth, L. K., Kim, S. K., and Owen, A. B. Properties of balanced permutations. *Journal of Computational Biology*, 16(4):625–638, 2009.

Westfall, P. H. and Young, S. S. *Resampling-based multiple testing: Examples and methods for p-value adjustment*, volume 279. John Wiley & Sons, 1993.

Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Geert Grooteplein 28, 6525 GA Nijmegen, The Netherlands
E-mail: `Jesse.Hemerik@radboudumc.nl, Jelle.Goeman@radboudumc.nl`