Leiden University Medical Center

Universiteit Leiden

Master Thesis

Permutation Tests and Multiple Testing

Jesse Hemerik
Leiden University
Mathematical Institute
Track: Applied Mathematics

December 2013

Thesis advisor: Prof. dr. J.J. Goeman
Leiden University Medical Center

Thesis supervisor: Prof. dr. A.W. van der Vaart
Leiden University, Mathematical Institute

Acknowledgements

First of all I would like to thank my supervisor, professor Aad van der Vaart, for helping me find this research project at the LUMC and for his advice and interest. Also, I am deeply grateful to my advisor at the LUMC, professor Jelle Goeman, for all his enthousiastic guidance.

I would also like to express my appreciation to Vincent van der Noort for his time and the inspiring conversations on permutation tests. To professor Aldo Solari I am also grateful, for his advice and for sharing a very interesting unpublished manuscript. Finally, the support of family and friends has been invaluable to me.

Contents

Introduction

Permutation tests are statistical procedures used to investigate correlations within random data. For example, they are often used to compare gene expressions between two groups of people (for instance a group of patients with a certain illness and a group of healthy patients.) In the most basic kind of permutation test, the whole group of permutations (or other 'null-invariant' group of transformations) is used. In many cases, like when examining gene expressions, one wants to test hundreds or thousands of null hypotheses at once, instead of one. This is called multiple testing and calls for new ways to control the amount of type I errors.

Here we will investigate how we can define valid permutation tests that do *not* use all permutations. We will look for ways to only use a subset of the whole permutation group and for methods that use randomly picked permutations. We will construct methods using random permutations not only for single hypothesis testing contexts, but also in the context of multiple testing. The main advantage of not using all permutations (or, more generally, transformations), is that a lot of computation time can be saved. When the permutation group is big or when a lot of hypotheses are tested, it is often simply infeasible to use all permutations. We will also compare existing multiple testing methods and improve them.

Basic single-hypothesis permutation tests using the whole group of permutations have been discussed in e.g. Lehmann and Romano (2005), Southworth et al. (2009) and Phipson and Smyth (2010). We will define various methods that allow the user to only use part of the permutations, thus saving a lot of computation time. It has been noted in the literature that random permutations can be used for permutation tests. Phipson and Smyth (2010), for instance, write that "it is common practice to examine a random subset of the possible permutations". We will show however, that it is necessary to add the identity permutation to the set of random permutations used. This has never been stated in the literature to our knowledge. Phipson and Smyth (2010) have calculated correct p-values for random permutation tests. However, it hasn't been stated that one can use the basic permutation test (where one uses the whole permutation group) also for random permutations, as long as one adds the identity permutation. Phipson and Smith calculate an exact p-value corresponding to the amount of test statistics exceeding the original test statistic, when permutations are picked with replacement. Computing this p-value can be time-consuming, so in practise one wouldn't want to use it. In Section 3 we give a method that avoids using this computation and guarantees a type I error probability of exactly α under the null hypothesis.

For a multiple testing context, Meinshausen (2005) has constructed a method for finding a lower bound for the amount of true discoveries, i.e. the amount of rightfully rejected hypotheses. This method uses randomly drawn permutations, and we will show that it is necessary to add the identity permutation. We will also discuss a related, as of yet unpublished, method by Goeman and Solari, expand it and compare it to Meinshausen's method.

1 Basic permutation tests and relabelling

1.1 Introduction

We start with an example of how a basic permutation test may be used. Suppose we are interested in making a certain species of plants grow faster. We have two types of soil, type I and type II, and we want to know whether the type of soil a plant grows in, influences its length. makes the plants grow taller within one month than the other type. A way to investigate this, is to put twenty small plants of equal size in pots and let them grow in equal conditions, except that we put the first ten flowers in soil of type I and the last ten in soil of type II. After one month we compare the heigths of all the plants. If most of the first ten plants are taller than the other ten, then this suggests that the type of soil influences the length. How can we use statistics to say something about the significance of the results?

A way to say something about this is to perform a permutation test. We define $X = (X_1, ..., X_{20})$ to be the hights of the plants, where $X_1, ..., X_{10}$ are the hights of the plants in type I soil. We define a test statistic T by

$$T(X) = |\sum_{i=1}^{10} X_i - \sum_{i=11}^{20} X_i|.$$

Note that high values of T are evidence that one type of soil is better than the other. We can reorder the plants in 20! ways; correspondingly, we can permute X in 20! ways. We do this, and for each permuted version of X we calculate $T(X)$. We end up with 20! test values and we let

$$T^{(1)}(X) \leq ... \leq T^{(20!)}(X)$$

be the ordered test values. We are interested in whether the null hypothesis H_0 that the type of soil doesn't influence the length after one month, is true. Note that if H_0 is true, then all the X_i are i.i.d. distributed. To test H_0 with a false rejection probability of at most 0.1, we can do the following: we reject H_0 if and only if the original test statistic $T(X)$ is larger than $T^{(0.9 \cdot 20!)}(X)$. We will show in Theorem 1.2 that, if H_0 is true, $P(T(X) > T^{(0.9 \cdot 20!)}(X)) \leq 0.1$. That is, if H_0 is true, the false rejection probability is at most 0.1, as we wanted.

The example we have given, gives an idea of the use of permutation tests. The null hypothesis was that the X_i were i.i.d.. We didn't make any other assumptions. The fact that very little assumptions are needed, is one of the benefits of permutation testing. Indeed, suppose that we needed to make very specific assumptions in the null hypothesis, for example normality. Then discovering that the null hypothesis doesn't hold, wouldn't give us much information, since we wouldn't know what aspects of the null hypothesis were false. Indeed, maybe it was just the assumption of normality that was false. In the example above however, knowing that the null hypothesis is false gives us the useful information that the type of soil influences the length of a plant.

To perform the test described in the example, we would need to calculate 20! test statistics: one for each permutation. As this simple example already illustrates, the total number of permutations is often much too big (for computation), and we need to limit the amount of permutatins used somehow. This section discusses how this can be done.

1.2 The basic permutation test

We will now give the general definition of a basic permutation test and show that the rejection probability under H_0 is α. Note that throughout this thesis, we will often say 'permutation test', when actually our statement holds for tests using certain other groups of transformations too. The reason we do this, is that the term 'permutation test' is common in the literature, while the term '(null-invariant) transformation test' is not. We need the following lemma.

Lemma 1.1. *Let $G = \{g_1, ..., g_n\}$ be a group and let $g \in G$. Write $Gg = \{g_1 g, g_2 g, ..., g_n g\}$. It holds that $G = Gg$.*

Proof. G is closed, so $Gg^{-1} \subseteq G$. Hence $G = (Gg^{-1})g \subseteq Gg$. That $Gg \subseteq G$ directly follows from the fact that G is closed. □

Definition of the basic permutation test

Theorem 1.2. *Let X be data with any distribution and let G be a* group *of transformations on the range of X (with composition of maps as the group operation). Let H_0 be a null hypothesis and let T be a test statistic on the range of X, high values of which are evidence against H_0. Let $M = \#G$ and let $T^{(1)}(X) \leq ... \leq T^{(M)}(X)$ be the ordered values $T(gX)$, $g \in G$. Suppose that H_0 is such that if it is true, $T(X) \stackrel{d}{=} T(gX)$ for all $g \in G$. Note that this holds in particular when $X \stackrel{d}{=} gX$ for all $g \in G$. Let α be the desired type I error rate and let $k = M - \lfloor M\alpha \rfloor$. Define*

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

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

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

Let

$$\phi(X) = \mathbb{1}_{\{T(X) > T^{(k)}(X)\}} + a(X)\mathbb{1}_{\{T(X) = T^{(k)}(X)\}}.$$

Then $0 \leq \phi \leq 1$. Reject H_0 when $\phi = 1$. Reject H_0 with probability $a(X)$ when $\phi(X) = a(X)$. (This is the boundary case $T(X) = T^{(k)}(X)$.) That is, reject with probability ϕ. Then, under H_0, $P(\text{reject } H_0) = E\phi = \alpha$.

Proof. By Lemma 1.1, $G = Gg$ for every $g \in G$, and consequently

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

Hence $T^{(k)}(X) = T^{(k)}(gX)$, $M^0(X) = M^0(gX)$, $M^+(X) = M^+(gX)$ and $a(X) = a(gX)$. So

$$\sum_{g \in G} \phi(gX) =$$

$$\sum_{g \in G} \mathbb{1}_{\{T(gX) > T^{(k)}(gX)\}} + a(gX)\mathbb{1}_{\{T(gX) = T^{(k)}(gX)\}} =$$

$$\sum_{g \in G} \mathbb{1}_{\{T(gX) > T^{(k)}(X)\}} + a(X)\mathbb{1}_{\{T(gX) = T^{(k)}(X)\}}.$$

By construction, this equals

$$M^+(X) + a(X)M^0(X) = M \cdot \alpha.$$

For every $g \in G$, it holds under H_0 that

$$(T(X), T^{(k)}(X), a(X)) \stackrel{d}{=} (T(gX), T^{(k)}(gX), a(gX))$$

and consequently $\phi(X) \stackrel{d}{=} \phi(gX)$, so $E\phi(X) = E\phi(gX)$. Hence, under H_0,

$$E\phi(X) = \frac{1}{M} E \sum_{g \in G} \phi(gX) = \alpha,$$

as we wanted to show. □

Remark 1. Note that if we had simply defined $\phi = \mathbb{1}_{\{T(X) > T^{(k)}(X)\}}$, we would have had a simpler, valid test with $P(\text{reject } H_0) \leq \alpha$ under H_0. (The example in subsection 1.1 is an example of such a test.) The advantage of the method above is that there this probability is exactly α instead of at most α. When one is only interested in keeping this probability smaller than α, it suffices to take $\phi = \mathbb{1}_{\{T(X) > T^{(k)}(X)\}}$. Note that as long as $\#\{g \in G : \ T(gX) = T^{(k)}(X)\}$ is in expectance relatively small (compared to $\#G$), the type I error probability under H_0 will be close to α anyway.

Note that when $M^+(X) = M\alpha$, it holds that $a(X) = 0$, so then $\phi(X) = \mathbb{1}_{\{T(X) > T^{(k)}(X)\}}$ in Theorem 1.2. However, $M^+(X) = M\alpha$ only holds when $M\alpha \in \mathbb{N}$ and $T^{k+1}(X) > T^k(X)$. So when with probability one all transformations give distinct test statistics and M is chosen such that $\alpha \in \mathbb{N}/M$, then it holds that $E\mathbb{1}_{\{T(X) > T^{(k)}(X)\}} = \alpha$ under H_0.

Remark 2. The function ϕ in Theorem 1.2 is based on the ordered test statistics. We can also adapt de definition of ϕ and base it on the ordered p-values resulting from the test statistics.

Example
The test that we now define is a specific case of the basic test defined in Theorem 1.2. It is an example of a test that uses a different transformation group than the permutation group.

For the following, we define multiplication of vectors pointwise, such that for all $x, y \in \mathbb{R}^n$,

$$xy = (x_1 y_1, ..., x_n y_n).$$

Let the null hypothesis H_0 be that $X = (X_1, ..., X_{2m}) \in \mathbb{R}^{2m}$ are i.i.d. and symmetric around 0. Let the test statistic be

$$T(X) = \sum_{i=1}^{m} X_i - \sum_{i=m+1}^{2m} X_i.$$

Let $R = \{(x_1, ..., x_{2m}) \in \mathbb{R}^{2m} : \ x_i \in \{-1, 1\} \text{ for all } 1 \leq i \leq 2m\}$. R is a group under the multiplication we just defined; each element has itself as the inverse. Each $r = (r_1, ..., r_{2m}) \in R$ can be seen as a 'relabelling' of the X_i in light of the test statistic. Write $M = \#R$. Let

$$T^{(1)}(X) \leq ... \leq T^{(M)}(X)$$

be the ordered test values $\in \{T(rX) : r \in R\}$. Let $k = M - \lfloor M\alpha \rfloor$. Define

$$M^+(X) = \#\{r \in R : \ T(X) > T(rX)\},$$

$$M^0(X) = \#\{r \in R : \ T(X) = T(rX)\},$$

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

$$\phi(X, r) = \mathbb{1}_{\{T(X) > T^{(k)}(X)\}} + a(X)\mathbb{1}_{\{T(X) = T^{(k)}(X)\}}.$$

Reject H_0 with probability ϕ. (So we always reject when $\phi = 1$.) Then, under H_0, $E\phi(X) = \alpha$.

Proof. Let $G = \{g^r : r \in R\}$, where $g^r : \mathbb{R}^{2m} \to \mathbb{R}^{2m}$ is given by $x \mapsto rx$. Under the null hypothesis, $X \stackrel{d}{=} rX = g^r(X)$ for all $r \in R$. Apply Theorem 1.2. □

Another example of a group of transformations that can sometimes be used in Theorem 1.2, are rotations (of a matrix) [11]. They are useful for testing intersection hypotheses. In section 5 we introduce the concept of closed testing, which is a multiple testing procedure. The closed testing method is based on tests of intersection hypotheses, which are single hypotheses tests. The use of Theorem 1.2 is certainly not limited to single hypothesis testing contexts.

1.3 The importance of the group structure

In the proof of the basic permutation test (Theorem 1.2), it was essential that $(T^{(1)}(X), ..., T^{(M)}(X))$ was invariant under all transformations in G of X, i.e.

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

for all $g \in G$. This property was guaranteed because it holds for a group G that $Gg = G$ for all $g \in G$. We now show that any set G of transformations (of which at least one is onto) which satisfies $Gg = G$ for all $g \in G$, is a group.

Proposition 1.3. *Let A be a nonempty set and let G be a set of maps $g : A \to A$. Assume that at least one element of G is onto. If $G = G \circ g$ for all $g \in G$, then G is a group (under composition of maps).*

Proof. Pick an element $g \in G$ that is onto. Since $G = Gg$, it holds that $g \in Gg$. Choose $g' \in G$ such that $g = g'g$. Let $y \in A$. Using that g is onto, choose $x \in A$ with $g(x) = y$. Thus $g'(y) = g'g(x) = g(x) = y$. So g' is the identity map on A.

For every $g \in G$ it holds that $Gg = G$, so there exists a $g' \in G$ with $g'g = id$. So every element of G has a left inverse and consequently is injective. Each $g \in G$ is surjective, because otherwise its left inverse would not be injective. So each element of G is a bijection. It follows that the left inverse of g is also the right inverse. We conclude that every element in G has an inverse in G.

That G is closed follows immediately from the fact that $Gg = G$ for all $g \in G$. It follows that G is a group. □

Remark. In the proof of Theorem 1.2 we use that the distribution of $T(X)$ is invariant under transformations (in G) of the data. This essentially comes down to assuming that the distribution of the data themselves is invariant under transformations in G. This assumption implies that the transformations, restricted to the range of the data, are all onto. So the assumption in Proposition 1.3, that at least one transformation should be onto, is not restrictive at all in this context.

Southworth, Kim and Owen (2009) show that *balanced permutations* can't be used for a permutation test, since the set of balanced permutations is not a group. We will give an example of another situation, where using a set of permutations that is not a group, leads to a false rejection probability which is much too large. It illustrates that one should be very careful when using a subset of the permutation group that isn't a subgroup: usually such a subset will not give a false rejection probability of α. (There are exceptions though. One important exception is the subject of sections 1.5 and 1.6.)

Example
Let $X = (X_1, ..., X_6)$ be a random vector in $\mathbb{R}^6$, such that $X_1, ..., X_6$ are continuously distributed. Let the null hypothesis H_0 be that $X_1, ..., X_6$ are i.i.d.. Let $T(X) = X_1 + X_2 + X_3 - X_4 - X_5 - X_6$ be the test statistic. Let G be the set of all permutation maps on $\mathbb{R}^6$.

Let

$$A := \{g \in G : \ T(gx) = T(x) \text{ for all } x \in \mathbb{R}^6\}$$

and

$$B := \{(14), (25), (36), (14)(25), (25)(36), (14)(36), (14)(25)(36)\}.$$

Let $U := \{id\} \cup \{ab : \ a \in A, b \in B\}$, where $id \in G$ is the identity permutation. Observe that $\#U = 3! \cdot 3! \cdot 7 + 1 = 253$. Note that for all $b \in B$, $x_{i+3} < x_i$ for all $i \in \{1, 2, 3\}$ implies that $T(bx) < T(x)$. Hence for all $u \in U \setminus \{id\}$, $x_{i+3} < x_i$ for all $i \in \{1, 2, 3\}$ implies that $T(ux) < T(x)$. But $P(x_{i+3} < x_i \text{ for all } i \in \{1, 2, 3\}) = \frac{1}{8}$. So with probability at least $\frac{1}{8}$, $T(ux) < T(x)$ for all $u \in U$. Take $\alpha = \frac{1}{253}$ and consider the basic test defined in section 1.2. Instead of using all permutations though, we only use the permutations in U. Then under H_0,

$$P(\text{reject } H_0) = P(T(X) > T(ux) \text{ for all } u \in U \setminus \{id\}) = \frac{1}{8},$$

which is much larger than α. (If we had excluded id from U, then even for arbitrarily small $\alpha > 0$, it would have held under H_0 that $P(\text{reject } H_0) = \frac{1}{8}$.) We conclude that the basic permutation method can go very wrong for some sets of permutations that aren't groups.

We will now generalize this example to show that the relative difference between α and $P(\text{reject } H_0)$ under H_0 can become arbitrarily large, even if we include the identity in the set of permutations used. For each $n \geq 3$, take $X = (X_1, ..., X_{2n})$ to be a random vector in $\mathbb{R}^{2n}$ such that $X_1, ..., X_{2n}$ are continuous. Let G be the group of all permutation maps on $\mathbb{R}^{2n}$. Let H_0 be that the X_i are i.i.d.. Take $T(X) = \sum_{i=1}^{n} X_i - \sum_{i=n+1}^{2n} X_i$ and define a set $U \in G$ with $\#U \geq n!n!$, $id \in U$ and such that $x_{i+n} < x_i$ for all $i \in \{1, ..., n\}$ implies that $T(ux) < T(x)$ for all $u \in U \setminus \{id\}$. [1] If we then take $\alpha_n = \frac{1}{n!n!}$, then for the basic permutation test, however using only the permutations in U (and with $\alpha = \alpha_n$), under H_0, $P(\text{reject } H_0) \geq$

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

$$P(x_{i+n} < x_i \text{ for all } i \in \{1, ..., n\}) = \frac{1}{2^n}.$$

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

We see that using certain sets of permutations, that aren't groups, can give a completely wrong false rejection probability. So using a random set of permutations seems to be generally a bad idea. However, some sets of permutations will give a rejection probability under H_0 that is too large, but other sets will give a rejection probability smaller than α. Thus, one might ask whether under

[1] To see that such a U exists, e.g. take $\pi \in G$ such that $x_{i+n} < x_i$ for all $i \in \{1, ..., n\}$ implies that $T(\pi x) < T(x)$. Then define $U = id \cup A\pi$, where $A := \#\{g \in G : \ T(gx) = T(x) \text{ for all } x \in \mathbb{R}^{2n}\}$.

H_0, $P(\text{reject } H_0)$ is equal to α *on average* over all random sets of permutations. This is indeed the case (when we add the identity permutation) and we will exploit this fact to construct a test (see Section 3) that gives the correct false rejection probability for a randomized set of transformations.

1.4 How to choose a subgroup

Suppose we have randomly distributed data X and a group G of transformations on the range of X that we would like to use for a basic permutation test as defined in Theorem 1.2. However, suppose this group is too large, such that a permutation test using all transformations in G would take too much time. A solution to this problem is to use a subgroup $S \subset G$, since this is still a group and thus gives a valid test. Using a subgroup of G reduces the computation time. Indeed, usually the computation time is roughly proportional to the amount of transformations used.

There are also other solutions, that decrease the computation time. First of all, one could use a completely different set of transformations. For instance, in the example at the end of Section 1.2, we could have used all permutations as the transformation group, but instead we used a different kind of maps. This group is much smaller than the group of all permutations. Another way to reduce the amount of transformations, is to use the fact that there (sometimes) are cosets of equivalent transformations. We explain this in Section 1.6. Finally, a way to decrease the computation time is to pick random transformations from a group (and add the identity transformation). A test using random permutations is defined in Theorem 3.1.

Power considerations
Here we will give some advice on how to choose a subgroup of a given group G of transformations, to be used for the test defined in Theorem 1.2. It is important to choose such a subgroup carefully, since the type II error probability varies depending on which subgroup is chosen. (The type I error probability is always α under H_0, so we only need to worry about the type II error probability.) It is certainly not the case that the biggest of two subgroups, is always the best.

To optimize the power, one wants to maximize the probability that the original test statistic $T(X)$ is among the $\alpha \cdot 100\%$ highest test statistics, if H_0 is false (where we assume that high values of $T(X)$ are evidence against H_0). We think that optimizing this probability largely comes down to avoiding that too many transformations are 'similar' to the identity transformtation, since these have a relatively high probability of giving test statistics higher than $T(X)$ if H_0 is false. We think the best way to do that, considering that we require $S \subset G$ to be a group, is to make sure that the elements in S are well 'spread out' across G, i.e. to make sure that every two elements of S are as 'dissimilar' as possible (in light of the test statistic). We will now give an example where we do this.

Example
Suppose our data are $X = (X_1, ..., X_{80}) \in \mathbb{R}^{80}$, G is the set of permutation maps

on $\mathbb{R}^{80}$ and we must choose a subgroup $S \subseteq G$, to be used for a permutation test, with $\#S < C$ for a given number C. A way to define a subgroup is to choose k with $\frac{80}{k} \in \mathbb{N}$ and define

$$Z_1 = (X_1, ..., X_k), Z_2 = (X_{k+1}, ..., X_{2k}), ..., Z_{\frac{80}{k}} = (X_{80-k+1}, ..., X_{80}).$$

We now define $S \subseteq G$ to be the set of all permutations g on the range of X that permute the Z_i, i.e. of the form

$$g(Z_1, ..., Z_{\frac{80}{k}}) = (Z_{i_1}, ..., Z_{i_{\frac{80}{k}}}),$$

where $(i_1, ..., i_{\frac{80}{k}})$ is a permuted version of $(1, 2, ..., \frac{80}{k})$. Thus $S \subseteq G$ is clearly a group and has $\frac{80}{k}!$ elements, which is much less than $\#G = 80!$ if $k > 1$. To guarantee that $\#S < C$, we simply choose a suitable k.

Note that we can often save even more computation time by letting each Z_i be the *average* of $X_{(i-1)k+1}, ..., X_{ik}$, and using permutations of $(Z_1, ..., Z_{\frac{80}{k}}) \in \mathbb{R}^{\frac{80}{k}}$. (We will have to redefine the test statistic as a function on $\mathbb{R}^{\frac{80}{k}}$ instead of $\mathbb{R}^{80}$.)

The set S of permutations on $\mathbb{R}^{80}$, that we just defined, seems to be a fairly good choice of a subgroup (depending on the test statistic), since it is well 'spread out' across G. However, if the test statistic is given by e.g.

$$T(X) = |\sum_{i=1}^{40} X_i - \sum_{i=41}^{80} X_i|,$$

then there are many permutations which are equivalent in light of the test statistic. For example, the permutation in S that simply swaps Z_1 and Z_2, is equivalent to the identity permutation in the sense that they always give the same test statistic. A way to use such observations to greatly reduce the number of permutations used, is given in Section 1.6. For $k = 10$, this method uses instead of S a set $S' \subset S$, which contains $\binom{8}{4} = 70$ instead of $\#S = 8!$ elements. S' contains one element from each coset of equivalent permutations in S. S' is (usually) not a subgroup of S.

A different subset of G which gives a valid permutation test (since it is a group) is the group L generated by the left shift $f : \mathbb{R}^{80} \to \mathbb{R}^{80}$ given by

$$f(x_1, ..., x_{80}) = (x_2, x_3, ..., x_{80}, x_1).$$

L contains 80 elements, which is slightly more than $\#S' = 70$.

Consider the case that $T(X)$ is a as defined above, $\alpha = 0.05$ and H_0 is the hypothesis that all X_i are i.i.d.. Suppose that it is given that $X_1, ..., X_{40}$ are i.i.d. and that $X_{41}, ..., X_{80}$ are i.i.d., and that all X_i are normally distributed with standard deviation 1 (and unknown expectance). Then, despite of the fact that L is bigger than S', L seems to give a less powerful test than S', since the set L contains significantly more transformations that are very similar to the identity map (in light of the test statistic). (Note that we are only speculating

here. More work is required to prove this.) For example, $T(f(X))$, $T(f \circ f(X))$, $T(f^{-1}(X))$ and $T(f^{-1} \circ f^{-1}(X))$ will often be close to $T(X)$, and the risk that some of these values are higher than $T(X)$ can be quite large. For every permutation g in $S' \setminus \{id\}$, however, it is more unlikely that $T(gX) \geq T(X)$. So S' seems to be a better choice than L (that is, S' gives more power), even though $\#S' < \#L$. (Again, this is speculation.)

1.5 Cosets of equivalent transformations

Introductory example
Consider again the example from section 1.1. As data $X = (X_1, ..., X_{20})$ we had the length of 20 plants and the test statistic was $T(X) = \sum_{i=1}^{10} X_i - \sum_{i=11}^{20} X_i$. As the group G of transformations, we used all permutation maps on $\mathbb{R}^{20}$. Now, to perform a basic permutation test as defined in section section 1.2, we need to calculate $T(gX)$ for all $g \in G$, where $\#G = 20! \approx 2.4 \cdot 10^{18}$, which is a lot. However, a lot of permutations are equivalent in light of the test statistic. Indeed, if π is the permutation that swaps X_{11} with X_1, then π and $\pi \circ (23)$ are equivalent, in the sense that for every realization x of X, $T(\pi x)$ is equal to $T(\pi(23)X)$ under H_0. That π and $\pi(23)$ are equivalent is because of the fact that they regroup the X_i in the same way, if we see $X_1, ..., X_{10}$ and $X_{11}, .., X_{20}$ as two groups. The shuffling that occurs *within* the two groups, doesn't affect the test statistic; only which values from group one are placed in group two and the other way around, affects the test statistic. In other words, only the relabelling of the X_i (as part of group one or two) is what the test statistic recognizes. There are $\binom{20}{10}$ ways to relabel the X_i. We will show in Theorem 1.5 (which assumes a more general setting), that instead of using thw whole given group of transformations, it suffices to use all 'relabellings'. This doesn't affect the type-I or type-II error probabilities at all and it saves a lot of computation time, since $\binom{20}{10} < 2^{20}$, which is much smaller than 20!, the total amount of permutations.

The following lemma makes the idea of equivalent transformations in light of the test statistic, more precise. We will use it in the proof of Theorem 1.5.

Lemma 1.4. *Let $G = \{g_1, ..., g_M\}$ be a group of maps (with composition as the group operation) from a measurable space A to itself. Let $T : A \to \mathbb{R}$ be a measurable map. Let $H := \{h \in G : T \circ h = T\}$. Then H is a subgroup of G. For all $g_1, g_2 \in G$, either $Hg_1 = Hg_2$ or $Hg_1 \cap Hg_2 = \emptyset$.*

Let $R \subseteq G$ be such that it contains exactly one element of each set of the form Hg, $g \in G$. Then the sets Hr, $r \in R$ are a partition of G. They all have $\#H$ elements, so $\#R = \frac{\#G}{\#H}$.

Proof. Note that $id \in H$ and H is closed. Let $h \in H$. Then $Th^{-1} = Thh^{-1} = T$, so $h^{-1} \in H$. Thus H is a group.

Let $g_1, g_2 \in G$ and suppose that $Hg_1 \cap Hg_2 \neq \emptyset$. Choose $h_1, h_2 \in H$ with $h_1 g_1 = h_2 g_2$. So $h_2^{-1} h_1 g_1 = g_2$, hence $g_2 \subseteq Hg_1$. Analogously it follows that $g_1 \subseteq Hg_2$, so $Hg_1 = Hg_2$, which proves the second claim of the lemma.

To see that the sets Hr, $r \in R$, are disjoint, note that for $r_1, r_2 \in R$, $Hr_1 \cap Hr_2 \neq \emptyset \implies \exists h_1, h_2 \in H : h_1 r_1 = h_2 r_2 \implies r_1 \in Hr_2$ and $r_2 \in Hr_1 \implies Hr_1 = Hr_2$. Let $g \in G$ and choose $r \in Hg$. Choose $h \in H$ with $r = hg$. So $g = h^{-1}r \in Hr$. So $G \subseteq \cup_{r \in R} Hr$. Hence the sets Hr, $r \in R$, are a partition of G.

To see that Hg has $\#H$ elements, note that for $h_1, h_2 \in H$, $h_1 g = h_2 g \implies h_1 = h_2$, so $h_1 \neq h_2 \implies h_1 g \neq h_2 g$.

□

Example

Note that in the example above Lemma 1.4, the set $H := \{h \in G : T \circ h = T\}$ would be all maps of the form $h(X) = (\pi_1(X_1, ..., X_{10}), \pi_2(X_{11}, ..., X_{20}))$, where $\pi_1, \pi_2 : \mathbb{R}^{10} \to \mathbb{R}^{10}$ are permutation maps. So $\#H = 10!10! \approx 1.3 \cdot 10^{13}$. H contains all the permutations that would keep the order of the labels unchanged if $X_1, ..., X_{10}$ had been labeled '1' and $X_{11}, ..., X_{20}$ had been labeled '2'.

Correspondingly, for the general case where the null hypothesis is that $X = (X_1, ..., X_{2n})$ are i.i.d. and the test statistic is

$$T(X) = \sum_{i=1}^{n} X_i - \sum_{i=n+1}^{2n} X_i,$$

we could define the set R as follows. For $i \in \{1, 2\}$ let $v_i = (i, ..., i)$ have length n, let G be the set of all permutation maps on $\mathbb{R}^{2n}$ and let $S = \{g(v_1, v_2) : g \in G\}$ be the set of all vectors of length $2n$ containing n ones and n twos. For each $s = \{s_1, ..., s_{2n}\} \in S$, let $f^s : \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ be a permutation map in G such that for each $z = \{z_1, ..., z_{2n}\} \in \mathbb{R}^{20}$, the first ten elements of $f^s(z)$ are the z_j with (indices j with) $s_j = 1$. Again let $H := \{h \in G : T \circ h = T\}$. Then for every $g \in G$, there are *unique* $s \in S$, $h \in H$ for which $g = h \circ f^s$, as we now prove.

It is clear that for each $g \in G$ there are such r and h. We now show that they are always unique. Choose s_1, $s_2 \in S$ and h_1, $h_2 \in H$ such that $h_1 \circ f^{s_1} = h_2 \circ f^{s_2}$. Suppose $s_1 \neq s_2$. Let $z = (z_1, ..., z_{2n}) \in \mathbb{R}^{2n}$ be a vector with $z_i \neq z_j$ for all $1 \leq i \leq j \leq 2n$. Choose an $1 \leq i \leq 2n$ such that $s_{1i} \neq s_{2i}$. But then it is clear that in exactly one of the vectors $h_1 \circ f^s(z)$ and $h_2 \circ f^s(z)$, the value z_i is among the first n arguments. Contradiction with $h_1 \circ f^{s_1} = h_2 \circ f^{s_2}$. So $s_1 = s_2$. But then it is clear that h_1 and h_2 must also be equal. This finishes the proof that g can be uniquely written as hf^s, with $h \in H$ and $s \in S$.

Thus $\{Hf^s : s \in S\}$ is a partition of G. by Lemma 1.4, for all g_1, $g_2 \in G$, $Hg_1 = Hg_2$ or $Hg_1 \cap Hg_2 = \emptyset$. Hence, as the set R (see Lemma 1.4), we could have chosen $\{f^s : s \in S\}$ in this example.

Each $s \in S$ can be seen as corresponding to a relabelling of the X_i. The test statistic $T(f^s(X))$ is the sum of the X_i labelled '1' minus the sum of the X_i labelled '2'.

1.6 A test method using cosets of equivalent transformations

We are now ready to prove the following theorem, which gives a permutation method that exploits the fact that there are subgroups of transformations that are equivalent in light of the test statistic. This method allows the user to only use one transformation from each coset of equivalent permutations (without sacrificing any power). That is, one only needs the transformations in the set R defined in Lemma 1.4.

Theorem 1.5. *Let X be data with any distribution and let T be a measurable test statistic from the range A of X to $\mathbb{R}$. Let $G = \{g_1, ..., g_M\}$ be a group of transformations (with composition as the group operation) from A to A. Let $T^{(1)}(X) \leq ... \leq T^{(M)}(X)$ be the ordered test values $T(gX)$, $g \in G$. Suppose that H_0 is such that if it is true, $T(X) \stackrel{d}{=} T(gX)$ for all $g \in G$. Note that this holds in particular when $X \stackrel{d}{=} gX$ for all $g \in G$.*

Let $H := \{h \in G : T \circ h = T\}$. By Lemma 1.4, H is a subgroup of G and for all $g_1, g_2 \in G$, either $Hg_1 = Hg_2$ or $Hg_1 \cap Hg_2 = \emptyset$. Let $R \subseteq G$ be such that it contains exactly one element of each set of the form Hg, $g \in G$. Then the sets Hr, $r \in R$ are a partition of G and each set Hr has $\#H$ elements, by Lemma 1.4.

Define a basic transformation test as in section 1.2, yet using only the transformations in R instead of all transformation in G (so M becomes $\#R$). That is, let $T'^{(1)}(X) \leq ... \leq T'^{(\#R)}(X)$ be the ordered test statistics $T(rX)$, $r \in R$. Let $k' = \#R - \lfloor (\#R)\alpha \rfloor$. Let

$$M'^{+}(X) = \#\{r \in R : \ T(rX) > T'^{(k')}(X)\},$$

$$M'^{0}(X) = \#\{r \in R : \ T(rX) = T'^{(k')}(X)\},$$

$$a'(X) = \frac{\#R\alpha - M'^{+}}{M'^{0}}$$

and define

$$\phi'(X) = \mathbb{1}_{\{T(X) > T'^{(k')}(X)\}} + a'(X)\mathbb{1}_{\{T(X) = T'^{(k')}(X)\}}.$$

Reject H_0 with probability $\phi'(X)$. Then $P(\text{reject } H_0) = \alpha$ under H_0.

Proof. Let $T^{(i)}$, M^0, M^+, a and ϕ be as in section 1.2. By Lemma 1.1, $G = Gg$ for every $g \in G$, and consequently

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

Note that

$$(T'^{(1)}(X), ..., T'^{(\#R)}(X)) = (T^{(1 \cdot \#H)}(X), T^{(2 \cdot \#H)}(X), ..., T^{(\#R \cdot \#H)}(X)),$$

since for each $r \in R$, $T(h_1 rX) = T(h_2 rX)$ for all $h_1, h_2 \in H$. Hence

$$(T'^{(1)}(X), ..., T'^{(\#R)}(X)) = (T'^{(1)}(rX), ..., T'^{(\#R)}(rX))$$

for all $r \in R$. Consequently $T'^{(k')}(X) = T'^{(k')}(rX)$, $M'^0(X) = M'^0(rX)$, $M'^+(X) = M'^+(rX)$ and thus $a'(X) = a'(rX)$ for all $r \in R$. So

$$\sum_{r\in R} \phi'(rX) =$$

$$\sum_{r\in R} \mathbb{1}_{\{T(rX)>T'^{(k')}(rX)\}} + a'(rX)\mathbb{1}_{\{T(rX)=T'^{(k')}(rX)\}} =$$

$$\sum_{r\in R} \mathbb{1}_{\{T(rX)>T'^{(k')}(X)\}} + a'(X)\mathbb{1}_{\{T(rX)=T'^{(k')}(X)\}}.$$

By construction, this equals

$$M'^+(X) + a'(X)M'^0(X) = \#R \cdot \alpha.$$

For every $g \in G$, $T(X) \stackrel{d}{=} T(gX)$, so

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

Hence for every $r \in R$

$$(T(X), T'^{(1)}(X), T'^{(2)}(X), ..., T'^{(\#R)}(X)) \stackrel{d}{=} (T(rX), T'^{(1)}(rX), T'^{(2)}(rX), ..., T'^{(\#R)}(rX))$$

and consequently $E\phi'(X) = E\phi'(rX)$.

Hence $E\phi'(X) = \frac{1}{\#R} E \sum_{r\in R} \phi'(rX) = \alpha$, as we wanted to show. □

Note that no power is sacrificed by using this method instead of the method given in Theorem 1.2. Indeed, this method gives exactly the same rejection function ϕ.

Remark. Instead of taking R to contain exactly one element from each coset Hg, $g \in G$, we could have taken R to contain exactly n elements from each coset Hg (for $n \leq \#H$). This doesn't have any advantages though in practise.

2 Preparations

Parts of the proofs of Theorems 3.1, 6.1 (the second proof) and the result in section 7.2 are essentially the same. So to avoid repeating ourselves, we prove this part in Theorem 2.2 and Corollary 2.3. We will use the following lemma.

Lemma 2.1. *Let G be a group and let Π be the vector $(id, g_2, ..., g_w)$, where $g_2, ..., g_w$ are random elements from G and id is the identity in G. Write $g_1 = id$. Either draw the permutations with or without replacement. If the g_i are drawn* with *replacement, for each $2 \leq i \leq w$, g_i is uniformly distributed on G and the g_i are i.i.d. If the g_i are drawn* without *replacement, then Π is uniformly distributed on*

$$W := \{(id, f_2, ..., f_w) : \ f_i, f_j \in G \setminus \{id\} \text{ and } f_i \neq f_j \text{ for all } 2 \leq i < j \leq w\}.$$

Then for every $1 \leq i \leq w$, there is a permuted version $\hat{\Pi}$ of Π such that $\hat{\Pi} \stackrel{d}{=} \Pi g_i^{-1} = \{g_1 g_i^{-1}, ..., g_w g_i^{-1}\}$. More precisely, $\Pi \stackrel{d}{=} \pi_i(\Pi g_i^{-1})$ where[2] *$\pi_i : G^w \to G^w$ is the map given by $\pi_i(h_1, ..., h_w) = (h_i, h_2, ..., h_{i-1}, h_1, h_{i+1}, ..., h_w)$,*[3] *i.e. π_i is the map that swaps the first and the i-th element of its argument.*

Proof. We give one proof for both the case of drawing without replacement and the case of drawing with replacement. Let W be the range of Π. For every $2 \leq i \leq w$, define

$$F_i : W \to W \text{ by } F_i(f) = \pi_i(f f_i^{-1}),$$

where $f = (id, f_2, ..., f_w)$. [4] Note (for $i > 4$ and $w > 7$) that

$$F_i(f) = (id, f_2 f_i^{-1}, ..., f_{i-1} f_i^{-1}, f_i^{-1}, f_{i+1} f_i^{-1}, ..., f_w f_i^{-1}).$$

So $F_i(f)$ is contained in W. It is easy to show that F_i is onto. Hence F_i is a bijection.

To show that $\Pi \stackrel{d}{=} \pi_i(\Pi g_i^{-1})$, we must show that $\pi_i(\Pi g_i^{-1})$ is uniformly distributed on W. Note that for all $f \in W$,

$$P(\pi(\Pi g_i^{-1}) = f) = P(F_i(\Pi) = f) = P(\Pi = F_i^{-1}(f)) = \frac{1}{\#W},$$

where the last equality follows from the fact that Π is uniformly distributed on W. So $\pi_i(\Pi g_i^{-1})$ is uniformly distributed on W, as we wanted.

□

Theorem 2.2. *Let X be data with any distribution. Suppose G is a group (under composition of maps) of measurable transformations on the range of X. Let m, $w \in \mathbb{Z}_{>0}$. Let Π be the vector $(id, g_2, ..., g_w)$, where $g_2, ..., g_w$ are random elements from G, independent of X, and id is the identity in G. Write $g_1 = id$. Either draw the permutations with or without replacement. If the g_i are drawn* with *replacement, then we assume that for each $2 \leq i \leq w$, g_i is uniformly distributed on G and the g_i are i.i.d. If the g_i are drawn* without *replacement, then we take Π to be uniformly distributed on*

$$\{(id, f_2, ..., f_w) : \ f_i, f_j \in G \setminus \{id\} \text{ and } f_i \neq f_j \text{ for all } 2 \leq i < j \leq w\}.$$

Let S be the range of X. Let $f^1 : S \to \mathbb{R}^m$ and $f^2 : S \times G^w \to \mathbb{R}^m$ be measurable maps. f^2 is also allowed to depend on additional randomness. [5] *Let $f^1_{(1)}(X) \leq ... \leq f^1_{(m)}(X)$ be the ordered values in $\{f^1_1(X), ..., f^1_m(X)\}$. Let $\alpha \in (0, 1)$. Define*

$$M^+(X, \Pi) = \#\{1 \leq j \leq w : \ f^1_{(i)}(g_j X) > f^2_i(X, \Pi) \text{ for all } 1 \leq i \leq m\}$$

[2] G^w is the Cartesian product $G \times G \times ... \times G$

[3] We slightly abuse notation here. The notation is only correct for $i > 4$ and $w > 7$.

[4] We write $f f_i^{-1} = (f_1 f_i^{-1}, f_2 f_i^{-1}, ..., f_w f_i^{-1})$. Note that the i-th element of $f f_i^{-1}$ is id. Hence the first element of $\pi_i(f f_i^{-1})$ is id.

[5] That is, it is allowed that f^2 depends on a third random variable Z. We will write $f^2(\cdot, \cdot)$ instead of $f^2(\cdot, \cdot, Z)$ for short. Everywhere $f^2(\cdot, \cdot, Z)$ should be read instead of $f^2(\cdot, \cdot)$.

and suppose that M^+ *is bounded from above by* $w\alpha$. *Define* $M^0(X,\Pi) =$

$\#\{1 \le j \le w : \ f^1_{(i)}(g_j X) \ge f^2_i(X,\Pi)$ *for all* $1 \le i \le m$, *with equality for at least one* $i\}$

and suppose $M_0 > 0$.

Define $a(X,\Pi) = \frac{\alpha w - M^+}{M^0}$ *and*

$$\phi(X,\Pi) := \mathbb{1}_{E^+}(X,\Pi) + a(X,\Pi)\cdot \mathbb{1}_{E^0}(X,\Pi),$$

where $E^+(X,\Pi)$ *is the event that*

$$f^1_{(i)}(X) > f^2_i(X,\Pi) \text{ for all } 1 \le i \le m,$$

(– denote this by $f^1(X) > f^2(X,\Pi)$ *for short –) and* $E^0(X,\Pi)$ *is the event that*

$$f^1_{(i)}(X) \ge f^2_i(X,\Pi) \text{ for all } 1 \le i \le m, \text{ with equality for at least one } i.$$

(Denote this by $f^1(X) \ge f^2(X,\Pi)$ *for short.) Write* $\{h_1, ..., h_{\#G}\} := G$. *Let* H_0 *be a null hypothesis such that if* H_0 *is true, the following hold.*

- *Property 1: Given any* $\Theta \in G^w$, *for all* $g \in G$,

  $$(f^1(h_1 X), ..., f^1(h_{\#G} X), f^2(X,\Theta)) \overset{d}{=} (f^1(h_1 gX), ..., f^1(h_{\#G} gX), f^2(gX,\Theta)).$$

  Note that this holds in particular when $X \overset{d}{=} gX$ *for all* $g \in G$.

- *Property 2: Given* $x \in S$ *and* $\Theta \in G^w$, *for each permuted version* $\hat{\Theta}$ *of* Θ,[6]

  $$(f^1(h_1 x), ..., f^1(h_{\#G} x), f^2(x,\Theta)) \overset{d}{=} (f^1(h_1 x), ..., f^1(h_{\#G} x), f^2(x,\hat{\Theta})).$$

- *Property 3:* $f^2(gX,\Theta) = f^2(X,\Theta g)$ *for all* $g \in G$ *and* $\Theta \in G^w$.

Then, if H_0 *is true,* $E\phi = \alpha$ *and* $0 \le \phi \le 1$. *(Hence rejecting* H_0 *with probability* $\phi(X,\Pi)$ *gives a rejection probability of* α *under* H_0.*)*

Proof. First consider the term

$$a(X,\Pi)\cdot \mathbb{1}_{\{f^1(X) \ge f^2(X,\Pi)\}}$$
$$= \frac{\alpha w - \#\{g \in \Pi : f^1(gX) > f^2(X,\Pi)\}}{\#\{g \in \Pi : f^1(gX) \ge f^2(X,\Pi)\}} \cdot \mathbb{1}_{\{f^1(X) \ge f^2(X,\Pi)\}}.$$

By Lemma 2.1, for each $2 \le i \le w$, $\Pi \overset{d}{=} \pi_i(\Pi g_i^{-1})$, so the above is in distribution equal to

$$\frac{\alpha w - \#\{g \in \pi_i(\Pi g_i^{-1}) : f^1(gX) > f^2(X, \pi_i(\Pi g_i^{-1}))\}}{\#\{g \in \pi_i(\Pi g_i^{-1}) : f^1(gX) \ge f^2(X, \pi_i(\Pi g_i^{-1}))\}} \cdot \mathbb{1}_{\{f^1(X) \ge f^2(X, \pi_i(\Pi g_i^{-1}))\}}.$$

[6] i.e. for $\hat{\Theta} = \rho(\Theta)$, where ρ is any permutation map on G^w

By Property 2, this is equal in distribution to

$$\frac{\alpha w - \#\{g \in \pi_i(\Pi g_i^{-1}) : f^1(gX) > f^2(X, \Pi g_i^{-1})\}}{\#\{g \in \pi_i(\Pi g_i^{-1}) : f^1(gX) \geq f^2(X, \Pi g_i^{-1})\}} \cdot \mathbb{1}_{\{f^1(X) \geq f^2(X, \Pi g_i^{-1})\}}.$$

Since $\pi_i(\Pi g_i^{-1})$ and Πg_i^{-1} contain the same elements, this equals

$$\frac{\alpha w - \#\{g \in \Pi g_i^{-1} : f^1(gX) > f^2(X, \Pi g_i^{-1})\}}{\#\{g \in \Pi g_i^{-1} : f^1(gX) \geq f^2(X, \Pi g_i^{-1})\}} \cdot \mathbb{1}_{\{f^1(X) \geq f^2(X, \Pi g_i^{-1})\}}.$$

It follows from Property 1 that this is equal in distribution to

$$\frac{\alpha w - \#\{g \in \Pi g_i^{-1} : f^1(gg_iX) > f^2(g_iX, \Pi g_i^{-1})\}}{\#\{g \in \Pi g_i^{-1} : f^1(gg_iX) \geq f^2(g_iX, \Pi g_i^{-1})\}} \cdot \mathbb{1}_{\{f^1(g_iX) \geq f^2(g_iX, \Pi g_i^{-1})\}}.$$

By Property 3 this equals

$$\begin{aligned}
&\frac{\alpha w - \#\{g \in \Pi g_i^{-1} : f^1(gg_iX) > f^2(X, \Pi)\}}{\#\{g \in \Pi g_i^{-1} : f^1(gg_iX) \geq f^2(X, \Pi)\}} \cdot \mathbb{1}_{\{f^1(g_iX) \geq f^2(X, \Pi)\}} \\
&= \frac{\alpha w - \#\{g \in \Pi : f^1(gX) > f^2(X, \Pi)\}}{\#\{g \in \Pi : f^1(gX) \geq f^2(X, \Pi)\}} \cdot \mathbb{1}_{\{f^1(g_iX) \geq f^2(X, \Pi)\}} \\
&= a(X, \Pi) \mathbb{1}_{\{f^1(g_iX) \geq f^2(X, \Pi)\}}.
\end{aligned}$$

In a similar way it follows that

$$\mathbb{1}_{\{f^1(X) > f^2(X, \Pi)\}} \stackrel{d}{=} \mathbb{1}_{\{f^1(g_iX) > f^2(X, \Pi)\}}.$$

Thus, for all $2 \leq i \leq w$,

$$E\phi(X, \Pi) = E\mathbb{1}_{\{f^1(g_iX) > f^2(X, \Pi)\}} + Ea(X, \Pi)\mathbb{1}_{\{f^1(g_iX) \geq f^2(X, \Pi)\}}.$$

Hence $E\phi(X, \Pi)$

$$\begin{aligned}
&= \frac{1}{w}\Big(\sum_{i=1}^{w} E\mathbb{1}_{\{f^1(g_iX) > f^2(X, \Pi)\}} + \sum_{i=1}^{w} \big(Ea(X, \Pi)\mathbb{1}_{\{f^1(g_iX) \geq f^2(X, \Pi)\}}\big)\Big) \\
&= \frac{1}{w}\Big(E\sum_{i=1}^{w} \mathbb{1}_{\{f^1(g_iX) > f^2(X, \Pi)\}} + Ea(X, \Pi)\big(\sum_{i=1}^{w} \mathbb{1}_{\{f^1(g_iX) \geq f^2(X, \Pi)\}}\big)\Big) \\
&= \frac{1}{w}\Big(EM^+(X, \Pi) + E\big(\frac{\alpha w - M^+(X, \Pi)}{M^0(X, \Pi)} \cdot M^0(X, \Pi)\big)\Big) \\
&= \frac{1}{w}\big(EM^+(X, \Pi) + \alpha w - EM^+(X, \Pi)\big) = \alpha.
\end{aligned}$$

□

It is important to add the identity transformation
We defined Π to be a vector of random transformations, with the identity permutation added to it (i.e. we let $g_1 = id$). For the proof, it was in particular

important that

$$E\mathbb{1}_{\{T(X)>T^{(k)}(X,V)\}} = \frac{1}{w}\sum_{j=0}^{w} E\mathbb{1}_{\{T(g_jX)>T^{(k)}(X,V)\}}.$$

This followed from the fact that for each $1 \leq j \leq w$,

$$E\mathbb{1}_{\{T(X)>T^{(k)}(X,V)\}} = E\mathbb{1}_{\{T(g_jX)>T^{(k)}(X,V)\}}.$$

In deriving this equality, we used the essential fact that, as is stated in Lemma 2.1, Π and Πg_j^{-1} are 'equal' in distribution if we don't pay attention to the order of the elements (but only to the amount of times each transformation $g \in G$ occurs in Π).

As we have seen in Theorem 1.2, a permutation test can be defined when we – instead of using random permutations – just use all permutations in the permutation group exactly once. This is a consequence of the group structure. When using random permutations, one loses this group structure. Though when we add the identity to the vector of random permutations, we get at least some of the nice structure back: we get the property that Π and Πg_j^{-1} have the same 'distribution', when we don't pay attention to order. This would also hold if Π was simply the group of all permutations and g_j any element of this group. So by adding the identity, we have made sure Π has a nice property that groups have and which is essential in this context.

We will need the following alternative, simpler version of Theorem 2.2.

Corollary 2.3. *Make the same assumptions and use the same definitions as in Theorem 2.2, except for the definitions of M^+ and ϕ (and E^+). Define*

$$M^+(X,\Pi) = \#\{1 \leq j \leq w : \ f^1_{(i)}(g_jX) \geq f^2_i(X,\Pi) \text{ for all } 1 \leq j \leq m\}.$$

Let $\hat{\alpha} \in (0,1)$ and suppose $M^+ \geq \hat{\alpha}w$. Define

$$\phi(X,\Pi) := \mathbb{1}_{E^+},$$

where E^+ is the event that

$$f^1_{(i)}(X) \geq f^2_i(X,\Pi) \text{ for all } 1 \leq i \leq m.$$

Then $E\phi \geq \hat{\alpha}$.

Proof. As in the proof of Theorem 2.2 it follows here that

$$E\mathbb{1}_{\{f^1_{(i)}(X)\geq f^2_i(X,\Pi) \text{ for all } 1\leq i\leq m\}} = E\frac{M^+(X,\Pi)}{w}.$$

Now use that $M^+ \geq \hat{\alpha}w$.

□

3 A permutation test using random permutations

We now state our permutation method using random permutations (or other random transformations from a group). It is basically the same as the basic permutation test defined in Theorem 1.2, apart from the fact that random transformations (with the identity added) are used.

Theorem 3.1. *Let X be data with any distribution. Let G be a group (with composition as the group operation) of transformations from the range of X to itself. Write $G =: \{h_1, ..., h_{\#G}\}$. Let T be a test statistic and let the null hypothesis H_0 be such that if it is true, then*

$$(T(h_1X), ..., T(h_{\#G}X)) \stackrel{d}{=} (T(h_1gX), ..., T(h_{\#G}gX))$$

for all $g \in G$. Note that this holds in particular when $X \stackrel{d}{=} gX$ for all $g \in G$.

Let Π be the vector $(id, g_2, ..., g_w)$, where $g_2, ..., g_w$ are random elements from G, independent of X, and id is the identity in G. Write $g_1 = id$. Either draw the permutations with or without replacement. If the g_i are drawn with *replacement, then for each $2 \leq i \leq w$, g_i is uniformly distributed on G and the g_i are i.i.d. If the g_i are drawn* without *replacement, then Π is uniformly distributed on*

$$\{(id, f_2, ..., f_w) : \ f_i, f_j \in G \setminus \{id\} \text{ and } f_i \neq f_j \text{ for all } 2 \leq i < j \leq w\}.$$

Let $T^{(1)}(X, \Pi) \leq ... \leq T^{(w)}(X, \Pi)$ be the w ordered test statistics $\in \{T(g_1X), ..., T(g_wX)\}$. Let $k = w - \lfloor w\alpha \rfloor$.
Let

$$M^+(X, V) = \#\{1 \leq i \leq w : \ T^{(i)}(X, \Pi) > T^{(k)}(X, \Pi)\}$$

and

$$M^0(X, V) = \#\{1 \leq i \leq w : \ T^{(i)}(X, \Pi) = T^{(k)}(X, \Pi)\}.$$

Let

$$a(X, V) = \frac{w\alpha - M^+}{M^0}.$$

Define

$$\phi(X, \Pi) = \mathbb{1}_{\{T(X) > T^{(k)}(X, \Pi)\}} + a(X, \Pi)\mathbb{1}_{\{T(X) = T^{(k)}(X, \Pi)\}}.$$

Reject H_0 when $\phi(X, \Pi) = 1$. Reject H_0 with probability $a(X, \Pi)$ when $\phi(X, \Pi) = a(X, \Pi)$. That is, we reject with probability ϕ.
Then $0 \leq \phi \leq 1$ and under H_0, $E\phi(X, V) = \alpha$.

Proof. Take $f^1(\cdot) = T(\cdot)$ and $f^2(\cdot, \cdot) = T^{(k)}(\cdot, \cdot)$. Note that the assumptions in Theorem 2.2 hold. [7]

[7] In Theorem 2.2, Property 1 follows from the fact that

$$(T(h_1X), ..., T(h_{\#G}X)) \stackrel{d}{=} T(h_1gX), ..., T(h_{\#G}gX))$$

for all $g \in G$, together with the fact that $T^{(k)}(X)$ is a function of $(T(h_1X), ..., T(h_{\#G}X))$.
Property 2 holds since the order of the random permutations doesn't influence $T^{(k)}$.
Property 3 holds since $T^{(k)}(gX, \Pi) = T^{(k)}(X, \Pi g)$.

The desired properties follow immediately from this theorem.

□

4 Exploratory research in multiple testing

4.1 Multiple testing and exploratory research

Suppose we are testing multiple hypotheses and want to keep the probability of any type I errors below α. This means that we are interested in controllong the *familywise error rate* (FWER), the probability that there is at least one true hypothesis among the rejected hypotheses. Especially when the amount of hypotheses is large, such tests will often result in a high amount of type II errors. Indeed, when there are many hypotheses, it is to be expected that there are some p-values that are quite low, but don't correspond to hypotheses that are false. Thus, often only hypotheses with extremely low p-values can be rejected, if we want to keep the FWER small.

We now give an example of a simple test controlling the FWER. Say we are testing hypotheses $H_1, ..., H_m$ and find corresponding p-values $p_1, ..., p_m$. For each $1 \leq i \leq m$, if we reject H_i if and only if $p_i \leq \alpha$, then the type I error probability for this single hypothesis is bounded by α. However, if we have this rejection rule for every hypothesis, then the FWER will usually be too high. (Of course, when all null hypotheses are false, the FWER is zero.) A way to control the FWER is to reject only the hypotheses with indices in $\{1 \leq i \leq m : p_i \leq \frac{\alpha}{m}\}$. Indeed, if $q_1, ..., q_{m_0}$ are the p-values corresponding the true null hypotheses, then the FWER equals

$$P(\bigcup_{1 \leq i \leq m_0} \{q_i \leq \frac{\alpha}{m}\}) \leq \sum_{i=1}^{m_0} P(q_i \leq \frac{\alpha}{m}) \leq m_0 \frac{\alpha}{m} \leq \alpha.$$

As the number of hypotheses m increases, $\frac{\alpha}{m}$ decreases, so the type II error probability for each hypothesis increases. This is also the case for more sophisticated FWER-controlling multiple hypothesis tests, like Holm's method, although these can give a lower type II error rate.

Often (for example in genetic research) statisticians are interested in testing thousands of null hypotheses. Then a test controlling the FWER would lead to very little rejections, if any at all. Therefore it is often better to first select a smaller set of hypotheses that look particularly promising, and continue only testing those. To do this, researchers have come up with methods that control the *amount* of type I errors. Benjamini and Hochberg (1995) have introduced the notion of the *false discovery rate* (FDR), defined as E(FDP), where the FDP is the *false discovery proportion*, the ratio of the number of true hypotheses among all rejections. (The FDP is a property of the specific rejected set, while the FDR is a property of the testing method.)

Methods controlling the FDP can be used for *exploratory research*, i.e. for selecting a set of hypotheses (from a larger set) with a large percentage of false

hypotheses. The aim of exploratory research is finding a set of hypotheses that 'look promising'. After selecting such a set, the researcher should continue with *confirmatory research*, i.e. perform tests that control the FWER. (One should not use the same data as were used for the exploratory research. I.e. one should use new realizations.) In exploratory research, the researcher is allowed to use all his knowledge and intuition in choosing the set of hypotheses to follow up on. For example, he might choose to select a hypothesis not because of its p-value, but because he has other knowledge (e.g. biological knowledge, or certain patterns in the other p-values) that suggests that the hypothesis is likely false. Exploratory research can be open-minded, since it is followed by confirmatory research, which rigorously protects against type I errors. Goeman and Solari (2012) give a more detailed introduction to multiple testing (in genomics).

4.2 Meinshausen's method

Typically, exploratory research allows some true hypotheses among the rejected hypotheses. Moreover, it ideally allows the researcher to change the rules of how to choose the rejected set, even after seeing the data and the test results. Preferably, the method nevertheless still gives a mathematical statement about the amount of false positives. An example of such a method is Meinhausen's. For $\{1 \leq k \leq m : p_k \leq t\}$ as (the indices of) the rejected set, Meinshausen's method gives (under additional assumptions) a lower bound $\underline{S}(t)$ for $S(t)$, the amount of false hypotheses among the rejected hypotheses. More precisely, it guarantees that

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

That is, $\underline{S}(t)$ is a lower bound uniformly over all $t \in [0,1]$. Hence t may be chosen *after* seeing the test results, depending on any considerations of the researcher.

Meinshausen's method makes use of random permutations. We will show that it is necessary to always add the identity permutation. We give a proof of our method in Theorem 6.1. A similar method is Goeman and Solari's procedure with random permutations, which we discuss in Sections 7.2 and 9.

5 Closed testing

For the proof in Section 7.2, we will use *closed testing*. Also, we will show the link between Meinshausen's method and closed testing in section 6. We will now give an explanation of closed testing, based on [3]. Let $H_1, ..., H_m$ be a collection of null hypotheses. Let $\mathcal{C}$ be the collection of all nonempty subsets of $\{1, ..., m\}$. For each set $I \in \mathcal{C}$, let H_I be the hypothesis that all hypotheses H_i, $i \in I$, are true. Such a hypothesis is called an intersection hypothesis. Suppose that we have an α-level test for each intersection hypothesis H_I, $I \in \mathcal{C}$. We call them the *local tests*. Let $\mathcal{U} \subseteq \mathcal{C}$ be the collection of subsets $U \in \mathcal{C}$ for which H_U is rejected by the local test. Closed testing is defined as the procedure that

rejects those intersection hypotheses H_I for which $J \in \mathcal{U}$ for every $J \supseteq I$. Write $\mathcal{X} = \{I \in \mathcal{C} : \ I \text{ is rejected by the closed testing procedure}\}$. The occurrence of any incorrect rejections by the closed testing procedure has probability at most α, as has been shown by Marcus, Peritz and Gabriel (1976).

Uniformly over all possible (possibly post hoc selected) rejected sets $R \subseteq \{1, ..., m\}$, an upper bound can be found for the amount of true hypotheses among the rejected hypotheses H_i, $i \in R$. This is stated in the following lemma, which can be found in [3].

Lemma 5.1. *Suppose we have a closed testing procedure as described above. For each nonempty $R \subseteq \{1, ..., m\}$, let*

$$t_\alpha(R) = \max\{\#I : \ I \subseteq R, I \notin \mathcal{X}\}.$$

Let $T \subset \{1, ..., m\}$ be the set of indices corresponding to the true hypotheses. Let $\tau(R) = \#(T \cap R)$ be the number of true hypotheses among the rejected hypotheses. Then

$$P(\tau(R) \leq t_\alpha(R) \text{ for all choices of } R \subset \{1, ..., m\}) \geq 1 - \alpha.$$

Proof. The event $E = \{T \notin \mathcal{U}\}$ happens with probability at least $1-\alpha$. Assume that E is true. Let $I' = T \cap R$. So $I' \subseteq T$ and $I' \notin \mathcal{X}$, since $T \notin \mathcal{U}$. But it also holds that $I' \subseteq R$, so $t_\alpha(R) \geq \#I'$ by definition. That is, $t_\alpha(R) \geq \tau(R)$. This holds for *any* choice of the rejected set $R \subseteq \{1, ..., m\}$, whenever E holds. That is, it holds for all R with probability at least $1 - \alpha$ □

Note that once we have an upper bound $t_\alpha(R)$ to the amount of incorrect rejections, then

$$f_\alpha(R) := \#R - t_\alpha(R)$$

is a lower bound for the amount of correct rejections. Once we have one bound, the other one follows directly. Just like $t_\alpha(R)$, $f_\alpha(R)$ is a bound uniformly over all R:

$$P(R - \tau(R(t)) \geq f_\alpha(R) \text{ for all } t \in [0, 1]) \geq 1 - \alpha,$$

where $R - \tau(R(t))$ is the amount of correct rejections, i.e. the amount of false hypotheses among the rejected hypotheses.

The following theorem gives a shortcut for finding a lower bound for the amount of correct rejections, when the closed testing procedure is of a specific form. It only holds if R is of the form $R(t) = \{1 \leq i \leq m : p_i \leq t\}$. It is an adaptation of a result from section 4.2 of [3].

Theorem 5.2. *Suppose we have a closed testing procedure for hypotheses $H_1, ..., H_m$, that rejects each intersection hypothesis H_I if and only if*

$$p^I_{(i)} < c_i$$

for at least one $1 \leq i \leq \#I$. *Here* $(p^I_{(1)}), ..., p^I_{(\#I)})$ *are the ordered p-values for the hypotheses* H_i, $i \in I$, *and* $c_1, ..., c_m \in \mathbb{R}$ *are allowed to depend on the data and anything else. Let the rejected set*[8] *be* $R(t) = \{1 \leq i \leq m : p_i \leq t\}$ *for* $t \in [0, 1]$. *Then*

$$f_\alpha(R) > \max\{S_r : 1 \leq r \leq \#R(t)\},$$

where $S_r := \max\{s \geq 0 : p_{(r)} < c_{r-s}\}$.

Proof. Write $R = R(t)$, let $r := \#R$ and suppose that $s \geq 0$ is such that $p_{(r)} < c_{r-s}$. Then $f_\alpha(R) > s$, as we now show. Choose any $K \subseteq R$ with $\#K \geq r - s$ and any $J \supseteq K$. Note that

$$p^J_{(r-s)} \leq p^K_{(r-s)} \leq p_{(r)} < c_{r-s},$$

where the second equality holds since $K \subseteq R = \{i : p_i \leq t\}$, which means that $p^K_{(r-s)}$ is among the r smallest p-values. Thus $K \in \mathcal{X}$ for every $K \subseteq R$ with $\#K \geq r-s$, so $t_\alpha(R) < r-s$ by definition. Hence $f_\alpha(R) = \#R - t_\alpha(R) > s$.

So $f_\alpha(R) > \max\{s \geq 0 : p_{(r)} < c_{r-s}\}$. Analogously we find for every $T \subset R$ that $f_\alpha(T) > \max\{s \geq 0 : p_{(\#T)} < c_{\#T-s}\}$. Since $f_\alpha(R) \geq f_\alpha(T)$ for every $T \subset R$, it follows that $f_\alpha(R) > \max\{S_r : \ 1 \leq r \leq \#R(t)\}$. □

6 Meinshausen's method with added identity permutation

6.1 Introduction

Suppose we are interested in null hypotheses $H_1, ..., H_m$. Let $P_1, ..., P_m$ be the corresponding p-values and suppose we reject the hypotheses with indices in

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

(The following may be generalized to hold for some other rejected sets as well.) Then, under certain additional assumptions, Meinshausen's method gives a lower bound $\underline{S}(t)$ for the amount of correct rejections $S(t)$, such that with probability at least $1 - \alpha$, $\underline{S}(t) \leq S(t)$ uniformly over all $t \in (0, 1)$. Since the lower bound is uniform, post hoc selection of t is allowed. Hence the method is well-suited for exploratory research.

We will state Meinshausen's method as set forth in his paper [9]. The only major adjustment is that we add the identity transformation to the set of randomly drawn transformations. Moreover, we give a different proof than [9]. We will also give a (longer) alternative proof, that shows the connection to closed testing.

[8] I.e. the set of indices of the rejected hypotheses

6.2 Definition of the method

Assumptions

Let A be a measurable space and let $X_1, ..., X_m$ be data with any distribution, each taking values in A. Write $X = (X_1, ..., X_m)$. For example, each X_i can be an $n \times 2$-matrix where the first column contains the expressions of a certain gene for n patients and the second column contains class variables (e.g. '1' for healthy patients and '0' for ill patients). Let H be a group of transformations from A to A. For each $h \in H$, write $hX := (h(X_1), ..., h(X_m))$. Suppose we have null hypotheses $H_1, ..., H_m$ and corresponding p-values $P_1(X), ..., P_m(X)$ respectively. Suppose that for each $1 \leq i \leq m$, $P_i(X)$ is a function of X_i only. (We simply mean that given X_i, $P_i(X)$ is known.) Let $\mathcal{N} \subseteq \{1, ..., m\}$ be the indices corresponding to the true null hypotheses.

For each $h \in H$, let $h^{\mathcal{N}}$ be the transformation given by

$$h^{\mathcal{N}}(X) = (\rho_1(X_1), ..., \rho_m(X_m)),$$

where

$$\rho_i = \begin{cases} h, & \text{if } i \in \mathcal{N}. \\ id, & \text{else.} \end{cases}$$

Let $H^{\mathcal{N}} := \{h^{\mathcal{N}} : h \in H\}$.

Suppose that the null hypotheses are such that the *joint* distribution of the p-values $P_i(hX)$ with $1 \leq i \leq m$, $h \in H$, is invariant[9] under all transformations in $H^{\mathcal{N}}$ of the data X.

The algorithm

Under the above assumptions, the following algorithm gives the lower bound $\underline{S}(t)$.

Step 1: Let Π be the vector $(id, h_2, ..., g_w)$, where $h_2, ..., h_w$ are random elements from H, independent of X, and id is the identity in H. Write $h_1 = id$. Either draw the permutations with or without replacement. If the h_i are drawn *with* replacement, then for each $2 \leq i \leq w$, h_i is uniformly distributed on G and the h_i are i.i.d. If the h_i are drawn *without* replacement, then Π is uniformly distributed on

$$\{(id, f_2, ..., f_w) : \ f_i, f_j \in H \setminus \{id\} \text{ and } f_i \neq f_j \text{ for all } 2 \leq i < j \leq w\}.$$

For each randomly drawn transformation, the m p-values resulting from the transformed data must be calculated. The p-values for the random permutations in Π are arranged in a $w \times m$ matrix $\mathbf{P}$ of p-values,

$$\begin{pmatrix} P_1(idX) & P_2(idX) & \cdots & P_m(idX) \\ P_1(h_2X) & P_2(h_2X) & \cdots & P_m(h_2X) \\ \vdots & \vdots & \ddots & \vdots \\ P_1(h_wX) & P_2(h_wX) & \cdots & P_m(h_wX) \end{pmatrix}.$$

[9] I.e. if we would place the elements of $\{P_i(hX) : 1 \leq i \leq m, h \in H\}$ in a vector, then the distribution of this vector would be invariant under transformations in $H^{\mathcal{N}}$ of X.

Step 2: Each column of this matrix is now permuted randomly.[10] Subsequently, for each row these p-values are placed in increasing order. Finally, the values in each column are placed in increasing order. We call the matrix that is thus obtained Q, and let Q_i^l denote the (l,i)-th entry of Q. Write Q^l for the l-th row of Q. From this matrix, a suitable row is chosen in the following step.
Step 3: For each $1 \leq l \leq w$, let

$$\beta(l) = \frac{\#\{1 \leq i \leq w : Q^l(X,\Pi) \leq P(h_i X)\}}{w}.$$

Here $Q^l(X,\Pi) \leq P(h_i X)$ means that $\forall 1 \leq k \leq m : Q_k^l(X,\Pi) \leq P_{(k)}(h_i X)$, where $P_{(1)}(h_i X) \leq ... \leq P_{(m)}(h_i X))$ are the sorted p-values in $\{P_1(h_i X), ..., P_m(h_i X)\}$. Define

$$l(\alpha) = \max\{l \in \{1, ..., w\} : \beta(l) \geq 1 - \alpha\},$$

the largest row index for which $\beta(l)$ is still at least $1-\alpha$. (When $\beta(1) < 1-\alpha$, then don't do as below but simply define $\underline{S}(t)$ to be the trivial lower bound given by $\underline{S}(t) = 0$ for all $t \in [0,1]$.) Define for all $t \in [0,1]$,

$$B(t) := \#\{k \in \{1, ..., m\} : \ Q_k^{l(\alpha)} \leq t\}. \tag{1}$$

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,*

$$P(S(t) \geq \underline{S}(t) \text{ for all } t \in [0,1]) \geq 1 - \alpha,$$
$$P(V(t) \leq \overline{V}(t) \text{ for all } t \in [0,1]) \geq 1 - \alpha,$$

Proof. Define the vector $\tilde{P}(X) = (\tilde{P}_1(X), ..., \tilde{P}_m(X))$ by

$$\tilde{P}_k(X) = \begin{cases} P_k, & \text{if } k \in \mathcal{N}. \\ 1, & \text{otherwise.} \end{cases}$$

Let $(\tilde{P}_{(1)}(X), ..., \tilde{P}_{(m)}(X))$ denote the sorted version of $\tilde{P}(X)$.

For all $h \in H$ and $k \in \mathcal{N}$, $\tilde{P}_k(h^{\mathcal{N}} X) = \tilde{P}_k(hX)$, since we assumed that P_k is a function of its k-th argument only. For $k \in \{1, ..., m\} \setminus \mathcal{N}$, $\tilde{P}_k(h^{\mathcal{N}} X) = 1 = \tilde{P}_k(hX)$. Hence, for all $1 \leq k \leq m$, $\tilde{P}_{(k)}(h^{\mathcal{N}} X) = \tilde{P}_{(k)}(hX) \geq P_{(k)}(hX)$. So it holds that

$$\frac{\#\{1 \leq i \leq w : \tilde{P}_{(k)}(h_i^{\mathcal{N}} X) \geq Q_k^{l(\alpha)} \text{ for all } 1 \leq k \leq m\}}{w} \geq 1 - \alpha.$$

[10] We have found that it is often better not to do this: see Section 8.4.

Define

$$f^1(X) = (\tilde{P}_{(1)}(X), ..., \tilde{P}_{(m)}(X)),$$

$$f^2(X, \Pi) = (Q_1^{l(\alpha)}(X, \Pi), ..., Q_m^{l(\alpha)}(X, \Pi)).$$

We assumed that the null hypotheses are such that the joint distribution of the p-values $P_k(hX)$ with $1 \leq k \leq m$, $h \in H$, is invariant under all transformations in $H^{\mathcal{N}}$ of the data X. Given $\Pi \in G^w$ (and the permutation maps applied to the columns of $\mathbf{P}$ in step 2 of the algorithm) and the p-values $P_k(hX)$ with $1 \leq k \leq m$, $h \in H$, $f^1(X)$ and $f^2(X, \Pi)$ are known.

It follows that given $\Pi \in G^w$ (and the permutation maps applied to the columns of $\mathbf{P}$ in step 2 of the algorithm), the distribution of

$$(f^1(h_1 X), ..., f^1(h_{\#H}(X)), f^2(X, \Pi))$$

is invariant under transformations in $H^{\mathcal{N}}$ of the data X. Let $G = H^{\mathcal{N}}$ and for each $1 \leq i \leq m$ define $g_i = h_i^{\mathcal{N}}$. Then Property 1 in Theorem 6.1 is satisfied. Property 2 is also satisfied: in particular $f^2(X, \Pi) \stackrel{d}{=} f^2(X, \hat{\Pi})$ for every permuted version $\hat{\Pi}$ of Π, since each column of $\mathbf{P}$ is shuffled in step 2 of the algorithm anyway.[11] Property 3 clearly holds as well, since for each $g \in G$, (gX, Π) and $(X, \Pi g)$ give the same initial matrix $\mathbf{P}$.

It follows that the assumptions needed for Corollary 2.3 hold if we take $\hat{\alpha} := 1 - \alpha$. Thus, by this corollary,

$$P(f_k^1(X) \geq f_k^2(X, \Pi) \text{ for all } 1 \leq k \leq m) \geq 1 - \alpha.$$

That is,

$$P(\tilde{P}_{(k)}(X) \geq Q_k^{l(\alpha)}(X, \Pi) \text{ for all } 1 \leq k \leq m) \geq 1 - \alpha.$$

We can now proceed as in [9]. Note that

$$V(t) = \#\{1 \leq k \leq m : \tilde{P}_{(k)} \leq t\}. \tag{2}$$

Comparing (1) and (2), we see that

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

So $B(t)$ is an upper bound for the number of incorrect rejections. It follows that

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

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

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

[11] If we would leave out the step that randomly permutes the columns (i.e. the first part of step 2), then it even holds that $f^2(X, \Pi) = f^2(X, \hat{\Pi})$, because each column is sorted (after each row has been sorted).

Hence

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

as we wanted to show.

□

6.3 The relation to closed testing

We will show that in the setting of Theorem 6.1 – with a small assumption added that often is not restricting in practise –, we can define an α-level test for each intersection hypothesis $H_I = \bigcap_{i \in I} H_i$ and thus a closed testing procedure is defined for these hypotheses. As we will show, we can use Theorem 5.2 to get the lower bound $\underline{S}(t)$ to the amount of correct rejections, defined in Theorem 6.1. So the following is (almost) an alternative proof of Theorem 6.1, using the connection to closed testing. The proof we have already given is shorter, but since closed testing is such an important tool in multiple testing, we think it is worthwile to show the link between Theorem 6.1 and closed testing.

The following proof is partly similar to that of Theorem 7.2. For Theorem 7.2 we could have used a proof similar to the proof of Theorem 6.1, if we had made the same assumptions about the data and null hypotheses as the assumptions used in Theorem 6.1. However, we make less assumptions in Theorem 7.2 (e.g. we don't assume that X can be written as $X = (X_1, ..., X_m)$ where the i-th p-value only depends on X_i), because this makes the result more general. We haven't been able to find a proof of Theorem 6.1 in the more general setting of Theorem 7.2.

We now give the alternative proof of Theorem 6.1 using closed testing. We first state the additional property that we assume:

For each $I \in \mathcal{N}$ and $h \in H$, let h^I be the transformation given by

$$h^I(X) = (\rho_1^I(X_1), ..., \rho_m^I(X_m)),$$

where

$$\rho_i^I = \begin{cases} h, & \text{if } i \in I. \\ id, & \text{else.} \end{cases}$$

Let $H^I := \{h^I : h \in H\}$. Suppose that the null hypotheses are such that for each $I \in \mathcal{N}$ the *joint* distribution of the p-values $P_i(hX)$ with $1 \leq i \leq m$, $h \in H$, is invariant under all transformations in H^I of the data X.

Proof. Step 1.
We will use the closed testing theory we developed in Section 5. For each nonemtpy $I \in \{1, ..., m\}$, we will define a local a test for $H_I = \cap_{i \in I} H_i$. Let $P^I = (P^I_{(1)}, ..., P^I_{(\#I)})$ be the sorted version of $\{P_k : k \in I\}$. Reject H_I if and only if $P^I_{(k)}(X) < Q_k^{l(\alpha)}(X, \Pi)$ for at least one $1 \leq k \leq \#I$. We now show that the rejection probability is at most α under H_I.

Note that for all $1 \leq k \leq \#I$, $P^I_{(k)}(X) \geq P_{(k)}(X)$. By definition of $l(\alpha)$,

$$\frac{\#\{1 \leq i \leq w : P_{(k)}(h_i X) \geq Q_k^{l(\alpha)}(X, \Pi) \text{ for all } 1 \leq k \leq m\}}{w} \geq 1 - \alpha.$$

Hence it also holds that

$$\frac{\#\{1 \leq i \leq w : P^I_{(k)}(h_i X) \geq Q_k^{l(\alpha)} \text{ for all } 1 \leq k \leq \#I\}}{w} \geq 1 - \alpha.$$

Note that for each $h \in H$ and $1 \leq i \leq \#I$, $P^I(hX) = P^I(h^I X)$, since for all $k \in I$, $P_k(h^I X) = P_k(hX)$. Hence

$$\frac{\#\{1 \leq i \leq w : P^I_{(k)}(h^I_i X) \geq Q_k^{l(\alpha)} \text{ for all } 1 \leq k \leq \#I\}}{w} \geq 1 - \alpha.$$

Define

$$f^1(X) = (P^I_{(1)}(X), ..., P^I_{(\#I)}(X)),$$

$$f^2(X, \Pi) = (Q_1^{l(\alpha)}(X, \Pi), ..., Q_{\#I}^{l(\alpha)}(X, \Pi)).$$

We assumed that for all $I \subseteq \mathcal{N}$, the joint distribution of all the p-values $P_k(hX)$ with $1 \leq k \leq m$, $h \in H$, is invariant under transformations in H^I of the data X. If H_I is true, then $I \subseteq \mathcal{N}$. Given $\Pi \in G^w$ (and the permutation maps applied to the columns of $\mathbf{P}$ in step 2 of the algorithm) and the p-values $P_k(hX)$ with $1 \leq k \leq m$, $h \in H$, $f^1(X)$ and $f^2(X, \Pi)$ are known. Thus it follows that (under H_I) given $\Pi \in G^w$ (and the permutation maps applied to the columns of $\mathbf{P}$ in step 2 of the algorithm), the distribution of

$$(f^1(h_1 X), ..., f^1(h_{\#H} X), f^2(X, \Pi))$$

is also invariant under transformations in H^I of the data X. Let $G = H^I$ and for each $1 \leq i \leq m$ define $g_i = h^I_i$. Then Property 1 in Theorem 6.1 is satisfied. Property 2 is also satisfied: in particular $f^2(X, \Pi) \stackrel{d}{=} f^2(X, \hat{\Pi})$ for every permuted version $\hat{\Pi}$ of Π, since each column of $\mathbf{P}$ is shuffled in step 2 of the algorithm anyway.[12] Property 3 clearly holds as well, since for each $g \in G$, (gX, Π) and $(X, \Pi g)$ give the same initial matrix $\mathbf{P}$.

Take $\hat{\alpha} := 1 - \alpha$. It thus follows from Corollary 2.3 that

$$P(f^1_k(X) \geq f^2_k(X, \Pi) \text{ for all } 1 \leq k \leq m) \geq 1 - \alpha.$$

That is,

$$P(P^I_{(k)}(X) \geq Q_k^{l(\alpha)}(X, \Pi) \text{ for all } 1 \leq k \leq m) \geq 1 - \alpha.$$

That is, under H_I, $P(\text{reject } H_I) \leq \alpha$. Thus we have an α-level local test for each nonempty $I \subseteq \{1, ..., m\}$.

[12] It also holds if we don't randomly shuffle the columns, since the columns are sorted (after each row has been sorted).

Step 2.
A closed testing procedure is defined, based on these local tests. Theorem 5 gives a lower bound for the amount of correct rejections among the rejected hypotheses (i.e. the hypotheses with indices in $R(t) = \{1 \leq k \leq m : P_k \leq t\}$). This theorem implies that

$$f_\alpha(R(t)) > \max\{S_r : 1 \leq r \leq \#R(t)\}$$

where $S_r = \max\{s \geq 0 : \ P_{(r)} < Q_{r-s}^{l(\alpha)}\}$.
This means that $P(S(t) > \max\{S_r : 1 \leq r \leq \#R(t)\} \text{ for all } t) \geq 1 - \alpha$. To see that $\underline{S}(t) \leq f_\alpha(R(t))$, note that $\underline{S}(t)$

$$= \max_{0 \leq q \leq t} (\#R(q) - B(q))$$

$$= \max_{0 \leq q \leq t} (\#\{1 \leq k \leq m : \ P_{(k)} \leq q\} - \#\{1 \leq k \leq m : Q_k^{l(\alpha)} \leq q\}).$$

Since the maximum is taken at $P_{(r)}$ for some $1 \leq r \leq \#R$, this equals

$$= \max_{1 \leq r \leq \#R} (\#\{1 \leq k \leq m : \ P_{(k)} \leq P_{(r)}\} - \#\{1 \leq k \leq m : Q_k^{l(\alpha)} \leq P_{(r)}\}))$$

$$= \max_{1 \leq r \leq \#R} (\#\{1 \leq k \leq r : Q_k^{l(\alpha)} > P_{(r)}\})$$

$$\leq \max_{1 \leq r \leq \#R} (\#\{s \geq 0 : Q_{r-s}^{l(\alpha)} > P_{(r)}\} + 1)$$

$$= \max\{S_r : 1 \leq r \leq \#R(t)\} + 1 \leq f_\alpha(R(t)).$$

Since $f_\alpha(R(t))$ is a uniform lower bound, $\underline{S}(t)$ is too, i.e.

$$P(S(t) \geq \underline{S}(t) \text{ for all } t) \geq 1 - \alpha,$$

as we wanted.

□

7 Goeman and Solari's method

7.1 Goeman and Solari's original method

We wil state an as of yet unpublished multiple testing method, based on a manuscript by Goeman and Solari. Their method has the same purpose as Meinshausen's method: to give a lower bound for the amount of correct rejections, if the rejected set is

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

The methods are similar. However, in this method the columns of the intitial matrix of p-values are not randomly permuted. In the next section we will prove an adaptation of this method that uses random transformations instead of the whole group of transformations.

Definition of the method
Suppose we have data X, null hypotheses $H_1, ..., H_m$ and the m corresponding p-values $P_1(X), ..., P_m(X)$. We reject the hypotheses with indices in

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

Let $P_{(1)}(X) \leq ... \leq P_{(m)}(X)$ denote the ordered p-values. Let $G = \{\tau_1, ..., \tau_r\}$ be a *group* of transformations (with composition as the group operation) from A to A, where A is the range of X. Define the $r \times m$ matrix $\mathbf{P}$ by $p_{i,j} = P_j(\tau_i X)$ and the $r \times m$ matrix $\mathbf{Q}$ by $q_{i,j} = P_{(j)}(\tau_i X)$. Let $\alpha \in (0, 1)$ and let J be a subset of $\{1, ..., r\}$ with $\#J \geq (1 - \alpha)r$. Define the curve $\mathbf{k} = (k_1, ..., k_m)$ by

$$k_j = \min\{q_{i,j}, i \in J\}.$$

That is, k_j is the local minimum of $\#J$ curves. Suppose the following hold.

- Assumption 1: $\mathbf{k}$ is a function of $\mathbf{Q}$ only: given $\mathbf{Q}$, $\mathbf{k}$ is known. Since $\mathbf{Q}$ is a function of X, $\mathbf{k} = \mathbf{k}(X)$ is as well.
- Assumption 2: $\mathbf{k}$ is invariant under shuffling[13] of the rows of $\mathbf{Q}$.

These are implicit assumptions on J. The two assumptions are equivalent with the following: $\mathbf{k}$ only depends on the *set* of p-value curves. (Note that if only this set is known, then it isn't generally known *which* p-value curve is the original 'unpermuted' p-value curve.)

Note that for each $i \in J$, $q_{i,j} \geq k_j$ for all j. Hence

$$M^+(X) := \#\{i : \bigcap_{j=1}^{m} \{q_{i,j} \geq k_j\}\} \geq \#J \geq r(1 - \alpha).$$

Suppose that the null hypotheses are such that the joint distribution of the $P_i(g(X))$ with $i \in \mathcal{N}, g \in G$ is invariant under all transformations in G of X.

We can then find a lower bound for the amount of correct rejections like with Meinshausen's method, as we will show.

Theorem 7.1. *Define for all* $t \in [0, 1]$,

$$B(t) := \#\{j \in \{1, ..., m\} : \ k_j \leq t\}.$$

Define the lower bound as

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

Define $V(t) = \#R(t) - S(t)$ *to be the number of incorrect rejections. Define* $\overline{V}(t) = \#R(t) - \underline{S}(t)$. *Under the assumptions above,*

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

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

[13] i.e. interchanging of the rows, not of the elements within a row

Proof. We will use the closed testing theory we developed in Section 5. For each nonemtpy $I \in \{1, ..., m\}$, we will define a local a test for $H_I = \cap_{i \in I} H_i$. Let $(P^I_{(1)}, ..., P^I_{(\#I)})$ be the ordered version of $\{P_k : \ k \in I\}$. Reject H_I if and only if $P^I_{(j)} < k_j$ for at least one $1 \leq j \leq \#I$. Let $\mathbf{P}^I$ be the $r \times \#I$ matrix given by $p^I_{i,j} = P^I_j(\tau_i X)$, i.e. $\mathbf{P}^I =$

$$\begin{pmatrix} P^I_1(\tau_1 X) & P^I_2(\tau_1 X) & \cdots & P^I_{\#I}(\tau_1 X) \\ P^I_1(\tau_2 X) & P^I_2(\tau_2 X) & \cdots & P^I_{\#I}(\tau_2 X) \\ \vdots & \vdots & \ddots & \vdots \\ P^I_1(\tau_r X) & P^I_2(\tau_r X) & \cdots & P^I_{\#I}(\tau_r X) \end{pmatrix}.$$

Write $I = \{i_1, ..., i_{\#I}\}$. Note that the j-th column of $\mathbf{P}^I$ is the i_j-th column of the matrix $\mathbf{P}$ defined in the algorithm. Define the $r \times \#I$ matrix $\mathbf{Q}^I$ by $q^I_{i,j} = P^I_{(j)}(\tau_i X)$, where $P^I = (P^I_{(1)}, ..., P^I_{(\#I)})$ are the sorted p-values in $\{P^I_1, ..., P^I_{\#I}\}$.

Define the curve $\mathbf{k}^I = (k^I_1, ..., k^I_{\#I})$ by

$$k^I_j = \min\{q^I_{i,j}, i \in J\}.$$

Note that for each $i \in I$, $q^I_{i,j} \geq k^I_j$ for all j. Hence

$$M^{+,I}(X) := \#\{i : \bigcap_{j=1}^{\#I} \{q^I_{i,j} \geq k^I_j\}\} \geq \#J \geq r(1-\alpha).$$

Assume that the intersection hypothesis H_I holds. So $I \in \mathcal{N}$. Hence, by Assumption 3, the joint distribution of the $P_i(g(X))$ with $i \in I$, $g \in G$ is invariant under all transformations in G of X. Thus $\mathbf{P}^I(X) \stackrel{d}{=} \mathbf{P}^I(gX)$ for all $g \in G$. So, by Assumption 1, $(P^I(X), \mathbf{k}^I(X)) \stackrel{d}{=} (P^I(gX), \mathbf{k}^I(gX))$ for all $g \in G$. Hence, for every $1 \leq i \leq \#I$,

$$\mathbb{1}_{\{P^I(X) \geq \mathbf{k}(X)\}} \stackrel{d}{=} \mathbb{1}_{\{P^I(\tau_i X) \geq \mathbf{k}(\tau_i X)\}}.$$

Since $Gg = G$ for all $g \in G$, it follows with Assumption 2 that $\mathbf{k}(\tau_i X) = \mathbf{k}(X)$, and consequently $\mathbf{k}^I(\tau_i X) = \mathbf{k}^I(X)$. Hence the above equals

$$\mathbb{1}_{\{P^I(\tau_i X) \geq \mathbf{k}(X)\}}.$$

Since this holds for every $1 \leq i \leq r$, it follows that if H_I is true,

$$E\mathbb{1}_{\{(P^I_{(j)}(X) \geq k^I_j(X) \text{ for all } 1 \leq j \leq \#I\}} =$$

$$\frac{1}{r} E \sum_{j=1}^{r} \mathbb{1}_{\{(P^I_{(j)}(X) \geq k^I_j(X) \text{ for all } 1 \leq j \leq \#I\}} =$$

$$\frac{1}{r} E M^{+,I}(X) \geq \frac{1}{r} r(1-\alpha) = 1 - \alpha.$$

Since for each $1 \leq i \leq \#I$, $k_i^I(X) \geq k_i(X)$, it also holds that

$$P(P_{(j)}^I(X) \geq k_j(X) \text{ for all } 1 \leq j \leq \#I) \geq 1 - \alpha.$$

That is, under H_I, $P(\text{reject } H_I) \leq \alpha$. Thus we have α-level local test for each nonempty $I \subseteq \{1, ..., m\}$.

The rest of the proof is exactly the same as step 2 Section 6.3. □

7.2 Goeman and Solari's method with random permutations

Here we present our adaptation of Goeman and Solari's method. It uses randomly drawn transformations instead of the group of all transformations. Again we must add the identity to the vector of randomly drawn transformations. The method in Section 7.1 is a special case of this method, since if we take $w = \#G$ and we draw without replacement, then Π contains every element of G exactly once.

Definition of the method. Let X be data with any distribution and let $H_1, ..., H_m$ be null hypotheses. Let $P_1(X), ..., P_m(X)$ be the corresponding p-values and $P_{(1)}(X), ..., P_{(m)}(X)$ the sorted p-values. We reject the hypotheses with indices in

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

Let G be a *group* of transformations from A to A, where A is the range of X. Let $\Pi = (id, g_2, ..., g_w)$, where the g_i are random transformations in G, be as always. (Both drawing with replacement and drawing without replacement is allowed.) Write $g_1 := id \in G$. Define the $r \times m$ matrix $\mathbf{P}$ by $p_{i,j} = P_j(g_i X)$ and the $r \times m$ matrix $\mathbf{Q}$ by $q_{i,j} = P_{(j)}(g_i X)$. Let $\alpha \in (0, 1)$ and let J be a subset of $\{1, ..., w\}$ with $\#J \geq (1-\alpha)w$. Define a boundary-p-value curve $\mathbf{k} = (k_1, ..., k_m)$ by

$$k_j = \min\{q_{i,j} : i \in J\}.$$

Note that for each $i \in I$, $q_{i,j} \geq k_j$ for all j. Hence

$$M^+(X, \Pi) := \#\{i : \bigcap_{j=1}^{m} \{q_{i,j} \geq k_j\}\} \geq \#J \geq w(1 - \alpha).$$

Let $\mathcal{N} := \{1 \leq i \leq m : H_i \text{ is true}\}$.

Suppose the following hold.

- Assumption 1: $\mathbf{k}$ is a function of $\mathbf{Q}$ only: given $\mathbf{Q}$, $\mathbf{k}$ is known. Since $\mathbf{Q}$ is a function of (X, Π), $\mathbf{k} = \mathbf{k}(\mathbf{Q}) = \mathbf{k}(X, \Pi)$ is as well.
- Assumption 2: $\mathbf{k}$ is invariant under swapping of the rows of $\mathbf{Q}$.

These are implicit assumptions on J.

Suppose that the null hypotheses are such that the joint distribution of the $P_i(g(X))$ with $i \in \mathcal{N}, g \in G$ is invariant under all transformations in G of X (Assumption 3).

Then, as we will prove, it follows that

$$P(\{P_{(j)} \geq k_j\} \text{ for all } k) \geq 1 - \alpha.$$

It follows that we can find an lower bound for the amount of correct rections as with Meinshausen's method, as we will show.

Theorem 7.2. *Define for all $t \in [0,1]$,*

$$B(t) := \#\{j \in \{1, ..., m\} : \ k_j \leq t\}.$$

Define the lower bound as

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

Define $V(t) = \#R(t) - S(t)$ to be the number of incorrect rejections. Define $\overline{V}(t) = \#R(t) - \underline{S}(t)$. Under the assumptions above,

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

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

Proof. We will use the closed testing theory we developed in Section 5. For each nonemtpy $I \in \{1, ..., m\}$, we will define a local a test for $H_I = \cap_{i \in I} H_i$. Let $i_1 \leq ... \leq i_{\#I}$ be the indices in I. Let $(P^I_{(1)}, ..., P^I_{(\#I)})$ be the sorted version of $(P^I_1, ..., P^I_{\#I})$, where $P^I_j := P_{i_j}$ for all $1 \leq j \leq m$. Reject H_I if and only if $P^I_{(j)} < k_j$ for at least one $1 \leq j \leq \#I$. Define $\mathbf{P}^I =$

$$\begin{pmatrix} P^I_1(X) & P^I_2(X) & \cdots & P^I_{\#I}(X) \\ P^I_1(g_2X) & P^I_2(g_2X) & \cdots & P^I_{\#I}(g_2X) \\ \vdots & \vdots & \ddots & \vdots \\ P^I_1(g_wX) & P^I_2(g_wX) & \cdots & P^I_{\#I}(g_wX) \end{pmatrix}.$$

Note that the j-th column of $\mathbf{P}^I$ is the i_j-th column of the matrix $\mathbf{P}$ defined in the algorithm. Define the $w \times \#I$ matrix $\mathbf{Q}^I$ by $q^I_{i,j} = P^I_{(j)}(g_iX)$. So $\mathbf{Q}^I$ is $\mathbf{P}^I$ with sorted rows.

Define the curve $\mathbf{k}^I(X, \Pi) = (k_1, ..., k_{\#I})$ by

$$k_j = \min\{q^I_{i,j}, i \in J\}.$$

Note that for each $i \in I$, $q^I_{i,j} \geq k^I_j$ for all j. Hence

$$M^{+,I}(X) := \#\{i : \bigcap_{j=1}^{\#I} \{q_{i,j}^I \geq k_j^I\}\} \geq \#J \geq r(1-\alpha).$$

Define $f^1(X) = (P_{(1)}^I(X), ..., P_{(1)}^I(X))$, $f^2(X, \Pi) = \mathbf{k}^I(X, \Pi)$ and $\hat{\alpha} = 1 - \alpha$. Assume H_I. So $I \subseteq \mathcal{N}$. With Assumption 3 it follows that $\mathbf{P}^I(X) \stackrel{d}{=} \mathbf{P}^I(gX)$ for all $g \in G$. It follows that Property 1 in Theorem 2.2 is satisfied. From Assumption 2 it follows that Property 2 holds. Since $\mathbf{P}^I(gX, \Pi) = \mathbf{P}^I(X, \Pi g)$ and hence $\mathbf{Q}^I(gX, \Pi) = \mathbf{Q}^I(X, \Pi g)$ for all $g \in G$, it follows with Assumption 2 that Property 3 in Theorem 2.2 is satisfied.

Thus it follows from Corollary 2.3 (take $\hat{\alpha} = 1 - \alpha$) that

$$P(P_{(j)}^I(X) \geq k_j^I(X, \Pi) \text{ for all } 1 \leq j \leq \#I) \geq 1 - \alpha.$$

Since for each $1 \leq i \leq \#I$, $k_i^I(X, \Pi) \geq k_i(X, \Pi)$, it also holds that

$$P(P_{(j)}^I(X) \geq k_j(X) \text{ for all } 1 \leq j \leq \#I) \geq 1 - \alpha.$$

That is, under H_I, $P(\text{reject } H_I) \leq \alpha$. Thus we have α-level local test for each nonempty $I \subseteq \{1, ..., m\}$.

The rest of the proof is exactly the same as step 2 Section 6.3.

□

This type of proof cannot be used for Theorem 6.1

The algorithm by Meinshausen (see Section 6.2) is similar to Goeman and Solari's method for random permutations (see Section 7.2). The assumptions on the data and p-values in Section 6.2 are a special case of the more general assumptions in Section 7.2. One might ask whether we can't alter the proof of Meinshausen's method in Section 6.3 to be more similar to the proof in Section 7.2 (since then we could relax the assumptions on the data and p-values in Section 6.2). We will now explain why we could not do this.

In the proof of Goeman and Solari's method for random permutations, we constructed a boundary curve $(k_1^I, ..., k_{\#I}^I)$ for each intersection hypothesis H_I, and showed that with probability at least $1 - \alpha$, the curve $(P_{(1)}^I, ..., P_{(\#I)}^I)$ lies above $(k_1^I, ..., k_{\#I}^I)$. We then used that $k_i^I \geq k_i$ for all $1 \leq i \leq \#I$ and concluded that

$$P(P_{(j)}^I(X) \geq k_j(X) \text{ for all } 1 \leq j \leq \#I) \geq 1 - \alpha.$$

In the setting of Section 6.2, we can define a boundary curve for each nonemtpy $I \in \{1, ..., m\}$ in a similar way, as we now illustrate.

For each nonemtpy $I \in \{1, ..., m\}$, define $\mathbf{P}^I =$

$$\begin{pmatrix} P_1^I(X) & P_2^I(X) & \cdots & P_{\#I}^I(X) \\ P_1^I(h_2X) & P_2^I(h_2X) & \cdots & P_{\#I}^I(h_2X) \\ \vdots & \vdots & \ddots & \vdots \\ P_1^I(h_wX) & P_2^I(h_wX) & \cdots & P_{\#I}^I(h_wX) \end{pmatrix},$$

where the P_i^I are the p-values with indices in I. Write $I = \{i_1, ..., i_{\#I}\}$. Note that the j-th column of the above matrix is the i_j-th columns of the matrix $\mathbf{P}$ defined in step 1 of the algorithm. For each $1 \leq j \leq \#I$, apply to the the j-th column of $\mathbf{P}^I$ the permutation that was applied on the i_j-th column of $\mathbf{P}$ (in step 2). Next, order each row and then order each column. We call the resulting matrix Q^I. Write $Q^{I,l}$ for the l-th row of Q^I. For each $1 \leq l \leq w$, let

$$\beta(I,l) = \frac{\#\{1 \leq i \leq w : Q^{I,l}(X,\Pi) \leq P^I(g_i X)\}}{w}.$$

Here $Q^{I,l} \leq P^I$ means that $\forall 1 \leq k \leq m : Q_k^{I,l} \leq P_{(k)}^I$, where $(P_{(1)}^I, ..., P_{(m)}^I)$ is the ordered version of $(P_1^I ..., P_m^I)$.

Define

$$l_I(\alpha) = \max\{l \in \{1, ..., w\} : \beta(I,l) \geq 1 - \alpha\}.$$

Now $Q^{I,l_I(\alpha)} = (Q_1^{I,l_I(\alpha)}, ..., Q_{\#I}^{I,l_I(\alpha)})$ (i.e. the $l_I(\alpha)$-th row of Q^I) is the boundary curve corresponding to the intersection hypothesis H_I

The problem that we now encounter, is that this curve doesn't always lie above the curve $Q^{l(\alpha)}$. I.e. it isn't guaranteed that for all $1 \leq i \leq \#I$, $Q_i^{l(\alpha)} \leq Q_i^{I,l_I(\alpha)}$. Hence, even though we can show that

$$P(P_{(j)}^I(X) \geq Q_j^{I,l_I(\alpha)}(X) \text{ for all } 1 \leq j \leq \#I) \geq 1 - \alpha,$$

it doesn't follow that

$$P(P_{(j)}^I(X) \geq Q_j^{l(\alpha)}(X) \text{ for all } 1 \leq j \leq \#I) \geq 1 - \alpha.$$

This is why we had to use a different proof for Meinshausen's method, using more assumptions than in the proof for Goeman and Solari's method.

To see that it doesn't generally hold that for all $1 \leq i \leq \#I$, $Q_i^{l(\alpha)} \leq Q_i^{I,l_I(\alpha)}$, consider the following counterexample.

Counterexample.
Take $\alpha = 1/3$. So $1 - \alpha = 2/3$. Take $w = m = 3$. Suppose the initial matrix of p-values $\mathbf{P}$ is

$$\begin{pmatrix} .1 & .7 & .4 \\ .2 & .8 & .5 \\ .3 & .6 & .9 \end{pmatrix}.$$

The columns must now be randomly permuted. Suppose that we we draw the identity permutation for each column, so the matrix isn't changed. We then sort each row and get

$$\begin{pmatrix} .1 & .4 & .7 \\ .2 & .5 & .8 \\ .3 & .6 & .9 \end{pmatrix}.$$

We then sort each column, keeping the matrix above, which is Q. Note that $l(\alpha) = 2$, since 2/3 of the sorted p-value curves (namely $(.2, .5, .8)$ and $(.3, .6, .9)$) still lie above $Q^2 = (.2, .5, .8)$.

Now take $I = \{1, 2\}$. The corresponding initial p-value matrix $\mathbf{P}^I$ is

$$\begin{pmatrix} .1 & .7 \\ .2 & .8 \\ .3 & .6 \end{pmatrix}.$$

Sorting the rows and then the columns gives $Q^I =$

$$\begin{pmatrix} .1 & .6 \\ .2 & .7 \\ .3 & .8 \end{pmatrix}.$$

Note that $l_I(\alpha) = 1$, so $Q^{I,l_I(\alpha)} = (.1, .6)$, which does not lie above $(Q_1^{l(\alpha)}, Q_2^{l(\alpha)}) = (.2, .5)$.

8 Simulations

We will investigate some properties of Meinshausen's method and the method by Goeman and Solari (with random permutations), using simulations. We will also compare the methods. In Sections 8.1 and 8.2, we illustrate by means of simulations that the identity transformation must be added in both methods. We will see in Sections 8.1, 8.3 and 8.4, that Meinshausen's method gives no non-trivial result for a lot of datasets, since $\beta(1)$ is often already smaller than $1 - \alpha$. See Section 8.4 for more about this.

8.1 Meinshausen's method

Simulations with m=1
In the proof of Meinshausen's method, we used in particular that under the complete null hypothesis,

$$P(P_{(k)} \geq Q_k^{l(\alpha)} \text{ for all } k \in \{1, ..., m\}) \geq 1 - \alpha.$$

We will try to show that this doesn't generally hold when the identity permutation is not added.

To approximate the probability $P(P_{(k)} \geq Q_k^{l(\alpha)}$ for all $k \in \{1, ..., m\})$, we let the software R generate data, perform Meinshausen's method and observed whether the event $\{P_{(k)} \geq Q_k^{l(\alpha)}$ for all $k \in \{1, ..., m\}\}$ occurred. We repeated this many times and counted the number of times that this event occured, thus getting an idea of the probability of this event.

As data we took an $n \times m$ matrix $\hat{X}$ of i.i.d. standard normally distributed random variables and a vector of class variables $Y = \{y_1, ..., y_n\}$, with $y_i = 0$

for i odd and $y_i = 1$ for i even. As the group of transformations we used all permutation maps on $\mathbb{R}^n$, applied on Y. We picked the random permutations $\Pi = \{g_1, ..., g_w\}$ with replacement. For each $1 \leq i \leq m$ we let H_i be the null hypothesis that the distribution of the i-th column of $\hat{X}$ is independent of Y – thus all H_i are true – and let p_i be the corresponding p-value resulting from a t-test using the n values in the i-th column of $\hat{X}$, and Y as the binary class variable.

We took $\alpha = 0.1$, $m = 1$, $n = 20$ and varied w. In this case $\hat{X}$ has only one column, Q has one column and $l(\alpha)$ is simply one number. For both $g_1 = id$ and g_1 random and for various w, we let R generate a realization of $\hat{X}$, run Meinshausen's algorithm and state whether $\{P_{(k)} \geq Q_k^{l(\alpha)} \text{ for all } k \in \{1, ..., m\}\}$ was true. We did this 500 times for each w and stated the results in the table below. Note that for $m = 1$, Q^1 is simply the smallest p-value $P_{(1)}$ and hence $\beta(1)$ equals 1, so $l(\alpha)$ is always defined. (We haven't added the exact R script we used, since Meinshausen's algorithm is already contained in the code in Appendix A.)

	g_1 random	$g_1 = id$
w=1	236/500	500/500
w=2	344/500	500/500
w=3	388/500	500/500
w=5	426/500	500/500
w=8	451/500	500/500
w=12	433/500	451/500
w=16	450/500	477/500
w=20	429/500	443/500
w=30	431/500	454/500
w=45	451/500	455/500

Table 1: *For $\alpha = 0.1$, $m = 1$, $n = 20$ and w specified as well as whether the identity permutation was added, this table shows in how many of the runs the event $\{P_{(j)} \geq k_j$ for all $1 \leq j \leq m\}$ occurred.*

We see in the table above that if we add the identity, in at least about 450 out of 500 runs $P(P_{(i)} \geq Q_i^{l(\alpha)}$ for all $i)$. For $w < 9$ this event always happens, since then $w - \lfloor \alpha w \rfloor = w$, so $l(\alpha)$ is then 1, so for all permutations g_i in Π, $\{P_{(i)}(\hat{X}, g_i(Y)) \geq Q_i^{l(\alpha)}$ for all $i\}$ always occurs.

We see that if we don't add the identity permutation but take g_1 to be a random permutation, especially for small w, the amount of times that P lies above $Q^{l(\alpha)}$ is much smaller than 450. This suggests that then $P(P_{(i)} \geq Q_i^{l(\alpha)}$ for all $i)$ is smaller than 0.9. That this is indeed the case for $m = w = 1$ is easy to see: $Q^{l(\alpha)}$ is then simply equal to $P(\hat{X}, g_1(Y))$, and $P(P(\hat{X}, Y) \geq P(\hat{X}, g_1(Y)))$ is smaller than 0.9. For w large this probability will usually not be far below 0.9, since then taking g_1 random instead of taking $g_1 = id$ usually doesn't have a big influence on $Q^{l(\alpha)}$.

Simulations with m larger than 1. When we perform simulations of Meinshausen's method just like above, with $g_1 = id$, but for larger m , we see that often $\beta(1)$ is already smaller than 0.9, so Meinshausen's method gives a trivial lower bound $\underline{S}(t) = 0$ and $Q^{l(\alpha)}$ isn't defined. For example, for $m = 3$ and $w = 8$, in only 196 out of 500 cases was $\beta(1)$ at least 0.9. $\beta(1)$ is practically always larger than 0.9 for w large enough, though. However, for w large, $P(P_{(k)} \geq Q_k^{l(\alpha)}$ for all $k \in \{1, ..., m\})$ will be very close to $1 - \alpha$ (even when we don't add the identity), since not adding the identity will not affect $Q^{l(\alpha)}$ much. For example, for $n = 20$, $m = 3$, $w = 100$ and without the added identity permutation, in 912 out of 1000 runs $\{P_{(k)} \geq Q_k^{l(\alpha)}$ for all $k \in \{1, ..., m\}\}$ occurred.

8.2 Goeman and Solari's method with random permutations

We have seen that Meinshausen's method often gives no non-trivial result for the data that we used (i.e. $\beta(1)$ was already smaller than $1 - \alpha$). Goeman and Solari's method (with random permutations, as we defined it in section 7.2) however does always give a real result, by construction. For the simulations we will now discuss, we took the same data $\hat{X}, Y$ as above and the same test statistic (a t-test). Here we define J in Goeman and Solari's method to be

$$\{1 \leq i \leq w : q_{i,1} \geq v\},$$

where

$$v(X, \Pi) = \max\{v' \in \{q_{1,1}, ...q_{w,1}\} : \#\{1 \leq i \leq w : q_{i,1} \geq v'\} \geq (1 - \alpha)w\},$$

where $q_{i,j}$ is the (i, j)-th element of the matrix $\mathbf{Q}$. Note that with this definition of J, $\#J \geq (1 - \alpha)w$ and Assumptions 1 and 2 in Section 7.2 are satisfied.

We first performed four types of simulations, where we always took $n = 20$ and $m = 40$. (We haven't added the exact R script we used, since Goeman and Solari's algorithm is already contained in the code in Appendix A.) For $w = 10$, we first performed 1000 runs with the identity added to the random permutations. In 896 out of 1000 cases it held that $P_{(j)} \geq k_j$ for all $1 \leq j \leq m\}$. This is close to $0.9 \cdot 1000$, as expected. However, when we ran the same simulation, but without an added identity permutation (i.e. taking g_1 random), it held in only 380 out of 1000 cases that $P_{(j)} \geq k_j$ for all $1 \leq j \leq m$.

Next, we changed w to 100 and performed 200 runs with the identity added to the random permutations. (More runs would take too much time.) In 179 out of 200 cases it held that $P_{(j)} \geq k_j$ for all $1 \leq j \leq m$. This amount is close to $0.9 \cdot 200$, as expected. However, when we ran the same simulation, but without an added identity permutation (i.e. taking g_1 random), it held in only 164 out of 200 cases that $P_{(j)} \geq k_j$ for all $1 \leq j \leq m$. See the table below. These simulations suggest that when we do not add the identity to the vector of random

permutations, then $P(P_{(j)} \geq k_j \text{ for all } 1 \leq j \leq m)$ becomes too small. The reason is of course, that when the identity isn't added, then the $\#J$ p-value curves of which the minimum is computed to construct the boundary curve, very often don't contain the original p-value curve.

	w=10	w=100
with added id (i.e. with $g_1 = id$)	896/1000	179/200
without added id	380/1000	164/200

Table 2: *For $\alpha = 0.1$, $m = 40$, $n = 20$ and w specified as well as whether the identity permutation was added, this table shows in how many of the runs the event $\{P_{(j)} \geq k_j$ for all $1 \leq j \leq m\}$ occurred.*

For larger w and without the added identity permutation, $P(P_{(j)} \geq k_j \text{ for all } j)$ seems to be higher, judging from the table above. This seems to be caused by the fact that for larger w the boundary curve $(k_1, ..., k_m)$ depends less on each single permutation g_i, because there are more g_i (so it makes less of a difference whether the identity is added or not).

As m gets higher, not adding the identity permutation seems to get more problematic: $P(P_{(j)} \geq k_j \text{ for all } 1 \leq j \leq m)$ gets further away from $1 - \alpha$. See Table 3. This can be explained by noting that when m is larger, the p-value curves have more points, and (if the identity is not added) the probability that the original p-value curve $(P_{(1)}, ..., P_{(m)})$ is below the boundary curve at least somewhere, is larger.

	m=8	m=20	m=40
with added id (i.e. with $g_1 = id$)	895/1000	914/1000	890/1000
without added id	552/1000	456/1000	366/1000

Table 3: *For $\alpha = 0.1$, $w = 10$, $n = 20$ and m specified as well as whether the identity permutation was added, this table shows in how many of the runs the event $\{P_{(j)} \geq k_j$ for all $1 \leq j \leq m\}$ occurred.*

8.3 Comparison of the two methods

We are interested in which method – Meinshausen's or Goeman and Solari's – is best, i.e., which method gives the highest lower bound $\underline{S}(t)$ for the number of correct rejections. (We have already proven that $P(\underline{S}(t) \leq S(t) \text{ for all } t) \geq 1-\alpha$ for both methods.)

Let $m = 8, n = 20, w = 600$, and $\alpha = 0.1$. (We could not take m larger since then Meinshausen's method often gives a trivial result. Making w larger would give a too large computation time.) Let $Y = \{y_1, ..., y_{20}\}$ with $y_i = 0$ for i odd and $y_i = 1$ for i even. We defined $\hat{X}'$ to be an $n \times m$ matrix of i.i.d. standard

normally distributed variables. To give some correlation structure to each row, we define the $n \times m$ matrix $\hat{X}''$ by

$$\hat{X}''_{i,j} = \begin{cases} \hat{X}'_{i,j} + Z_i & \text{for } 3 \leq j \leq 6 \\ \hat{X}'_{i,j} & \text{for } j \in \{1, 2, 7, 8\} \end{cases}$$

for all $1 \leq i \leq 20$, where $Z_1, ..., Z_{20}$ are i.i.d. and normally distributed with $\mu = 0, \sigma = 0.5$.

Finally, to make the first four columns of the matrix dependent of Y, we defined the $n \times m$ matrix $\hat{X}$ by

$$\hat{X}_{i,j} = \begin{cases} \hat{X}''_{i,j} + 2y_i & \text{for } 1 \leq j \leq 4 \\ \hat{X}''_{i,j} & \text{for } 5 \leq j \leq 8. \end{cases}$$

As before, for each $1 \leq i \leq m = 8$ we let H_i be the null hypothesis that the i-th column of $\hat{X}$ is independent of Y. (So $H_1, ..., H_4$ are false and $H_5, ..., H_8$ are true.) For each $1 \leq i \leq 8$ let $P_i = P_i(\hat{X}, Y)$ be the corresponding p-value, resulting from a t-test using the i-th column of $\hat{X}$ as values, and Y as the binary class variables. As the group of transformations, we again used all permutation maps on $\mathbb{R}^m = \mathbb{R}^8$, applied to Y.

For this setting, we performed both Meinshausen's method (as defined in Section 6.2) and Goeman and Solari's method (as defined in Section 7.2) and compared the lower bounds they gave. For the set J in Goeman and Solari's mehod, we again used the definition stated in Section 8.2. We used the R-script in Appendix A. It repeatedly generates a realization of $\hat{X}$ and calculates $\#R(t)$ and the lowerbounds that Meinshausen's method and Goeman and Solari's method give.

We found that (for this specific setting) Goeman and Solari's method gives better (or at least as good) lower bounds $\underline{S}(t)$ for $t = 0.05$, $t = 0.01$ and $t = 0.0005$, which were all the values of t we evaluated the bound at. See Tables 4, 5 and 6. (We used separate realizations for each table.) Naturally, the performance of Goeman and Solari's method is dependent on the definition of J, among other factors.

8.4 Comparison of Meinshausen's method without column-shuffling and Goeman and Solari's method

During simulating, we have seen that Meinshausen's method often gives no non-trivial result when the amount of hypotheses is not very small. Above, we could not take m larger than 8. If we would make it higher, then $\beta(1)$ would often be already smaller than $1 - \alpha$, so the method would give no non-trivial result.

In his 2006 paper, where he defines his method, Meinshausen has presented the results of a simulation he did himself, where $m = 1000$. Naturally we performed this simulation as well, to check if we also get these results. However, we again got no non-trivial result, contrary to Meinshausen.

realization	$\#R(t)$	$\underline{S}(t)$ for G&S	$\underline{S}(t)$ for M.
1	4	4	4
2	4	4	4
3	5	4	3
4	4	4	3
5	4	4	3
6	4	3	2
7	4	4	3
8	4	4	3
9	4	4	3
10	4	4	3

Table 4: *For $t = 0.05$ and ten realizations $\hat{X}$, this table shows the amount of rejections $\#R(t)$ and the lower bounds $\underline{S}(t)$ found with Goeman and Solari's method and Meinshausen's method.*

realization	$\#R(t)$	$\underline{S}(t)$ for G&S	$\underline{S}(t)$ for M.
1	4	4	3
2	4	4	3
3	3	3	3
4	3	3	3
5	4	4	3
6	3	3	2
7	3	3	2
8	4	4	3
9	4	4	3
10	3	3	3

Table 5: *For $t = 0.01$ and ten realizations $\hat{X}$, this table shows the amount of rejections $\#R(t)$ and the lower bounds $\underline{S}(t)$ found with Goeman and Solari's method and Meinshausen's method.*

We think that the reason is that the R code, that was used for Meinshausen's paper, did not permute the columns of the intial p-value matrix $\mathbf{P}$ (in step 2), contrary to as is written. Indeed, the line of code that permutes the columns, is commented in Meinshausen's R package "howmany" (this is line 22 in the R function "get.boundingfunction.dependent"):

```
#for (p in 1:m) Quantile[,p] <- sample(Quantile[,p],n.permutation)
for (perm in 1:n.permutation) Quantile[perm,] <- sort(Quantile[perm,])
for (p in 1:m) Quantile[,p] <- sort(Quantile[,p])
```

When we do not permute the columns, we indeed get non-trivial results as well.

We will now give the results of some simulations where we compared Goe-

realization	$\#R(t)$	$\underline{S}(t)$ for G&S	$\underline{S}(t)$ for M.
1	2	2	2
2	1	1	1
3	3	3	3
4	0	0	0
5	1	1	1
6	3	3	2
7	1	1	1
8	3	3	3
9	3	3	3
10	3	3	3

Table 6: *For $t = 0.0005$ and ten realizations $\hat{X}$, this table shows the amount of rejections $\#R(t)$ and the lower bounds $\underline{S}(t)$ found with Goeman and Solari's method and Meinshausen's method.*

man and Solari's method with Meinshausen's method without shuffling[14] of the columns in step 2.[15] We haven't added the R code here, since it is similar to that in Appendix A. We took the data and the test (a Wilcoxon test) exactly as in Section 4.1 of Meinshausen's paper. That is, we let $m = 1000$ and let the vector of class variables Y be a vector of length n of Bernoulli variables (with $p = 0.5$). As $\hat{X}$ we took an n-by-m matrix, where the rows are independent and each row $(\hat{X}_{i1}, ..., \hat{X}_{im}) \sim \mathcal{N}(\mu, \Sigma)$, where the covariance matrix Σ is given by $\Sigma_{ii} = 1$ and $\Sigma_{ij} = \rho$ for $i \neq j$, where ρ is to be specified. For all $1 \leq i \leq n$, $1 \leq j \leq m$, the mean of X_{ij} is given by $\mu_{ij} = 0$ if $j \leq 600$ and $\mu_{ij} = Y_i$ for $j > 600$. For each $1 \leq j \leq 1000$ we tested the null hypothesis H_j that the j-th column of $\hat{X}$ is independent of Y. Hence there are 600 true null-hypotheses. We defined P_j to be the p-value resulting from a Wilcoxon test of the j-th column of $\hat{X}$ and Y as the binary variables. We used $w = 500$ random permutations (where we took the first permutation to be the identity).

We defined J in Goeman and Solari's method to be

$$J := \{1 \leq i \leq w : q_{i,100} \geq v\},$$

where

$$v(X, \Pi) = \max\{v' \in \{q_{1,100}, ...q_{w,100}\} : \#\{1 \leq i \leq w : q_{i,100} \geq v'\} \geq (1 - \alpha)w\},$$

where $q_{i,j}$ is the (i, j)-th element of the matrix $\mathbf{Q}$.

We calculated lower bounds for $\rho = 0$ (Table 7) and $\rho = 0.4$ (Table 8), for different values of t. When $\rho = 0.4$, the elements of each row of $\hat{X}$ are positively correlated. As we can see from comparing these tables, the correlations in $\hat{X}$ seem make the variance of $\#R$ and the lower bounds larger. Meinshausen also

[14]The rows and columns are still sorted, however.

[15]The proof of Meinshausen's method that we have given, is still correct for this adaptation of the method.

found that the lower bounds vary a lot for $\rho = 0.4$. He obtained better lower bounds for larger n. The tables suggest that Goeman and Solari's lower bound is best when t is not very small, for the data we used. For $t = 0.002$, it is a close call. However, usually t will be chosen larger, considering that these methods will be used for exploratory research. Thus, the tables suggest that Goeman and Solari's method is often the best choice, for these data. Of course, the performance of Goeman and Solari's method depends on the choice of J.

For more general statements, more comparisons for more types of data must be performed. More elaborate comparisons of both methods can be subject of future research.

	t=0.002			t=0.01			t=0.05			t=0.2		
	$\#R$	G&S	M.	$\#R$	G&S	M.	$\#R$	G&S	M.	$\#R$	G&S	M.
1	50	44	44	113	90	87	211	129	109	415	154	124
2	45	39	38	123	99	94	243	168	137	443	179	145
3	45	38	39	102	81	77	222	145	124	426	194	153
4	35	28	28	108	83	79	220	147	116	425	195	152
5	44	38	39	115	92	89	228	154	130	427	174	142

Table 7: For $\rho = 0$ and five different realizations of the data, this table shows $\#R(t)$ and the lowerbounds $\underline{S}(t)$ found with Goeman and Solari's method ("G&S") and Meinshausen's method without random shuffling of the columns ("M."), for different values of t.

	t=0.002			t=0.01			t=0.05			t=0.2		
	$\#R$	G&S	M.	$\#R$	G&S	M.	$\#R$	G&S	M.	$\#R$	G&S	M.
1	12	0	1	64	11	1	184	14	1	376	14	1
2	183	169	168	316	262	222	466	294	228	685	294	228
3	147	135	135	300	251	243	456	288	249	647	288	249
4	10	10	0	53	10	0	181	10	0	432	10	0
5	90	77	77	223	181	159	364	212	170	527	212	170
6	135	116	121	260	213	186	399	246	197	562	246	197
7	5	0	0	36	0	0	132	0	0	412	0	0
8	44	32	30	126	76	69	260	126	88	429	126	88

Table 8: For $\rho = 0.4$ and eight different realizations of the data, this table shows $\#R(t)$ and the lowerbounds $\underline{S}(t)$ found with Goeman and Solari's method ("G&S") and Meinshausen's method without random shuffling of the columns ("M."), for different values of t.

9 Optimization of Goeman and Solari's method

The boundary curve $\mathbf{k}$ in Goeman and Solari's method depends on the index set J chosen. That is, the lower bound that this method gives, depends on J. Write $\underline{S}_J(t)$ or $\underline{S}_{\mathbf{k}}(t)$ for the lowerbound of this method, and write $S'(t)$ for the lowerbound that Meinshausen's method (with or without random shuffling of the columns) gives. We will look at how we can choose J to optimize the lower bound $\underline{S}^J(t)$.

Write

$$\Gamma := \{A \subseteq \{1, ..., m\} : \#A \geq \lceil (1-\alpha)w \rceil\}.$$

If J is such that

$$\underline{S}_J(t) = \mu := \max\{\underline{S}'_J(t) : \ J' \in \Gamma\},$$

then J is 'optimal', in the sense that there is no $J' \in \Gamma$ that gives a better lower bound (in the point t).

However, what we cannot do in Goeman and Solari's method, is – once we have the matrix $\mathbf{Q}$ – calculate μ and then pick a J for which $\underline{S}_J(t) = \mu$. The reason is that calculating $\underline{S}_{J'}(t)$ for any $J' \in \Gamma$ requires knowing the original 'unpermuted' p-value curve. But we are only allowed to choose $\mathbf{k}$ on the basis of $\mathbf{Q}$, and only in such a way that $\mathbf{k}$ is invariant under interchanging of the rows of $\mathbf{Q}$. That is, we are allowed to use the set of p-value curves when choosing $\mathbf{k}$, but we are not allowed to use the knowledge which curve is the original curve.

If we were allowed to calculate $\underline{S}_{J'}(t)$ for all $J' \in \Gamma$ before choosing J in Goeman and Solari's method, then we could always choose J such that $\underline{S}_J(t) \geq S'(t)$, i.e. Goeman and Solari's bound would always be at least as good as Meinshausen's bound, uniformly over all $t \in [0, 1]$. This is because there are at least $\lceil 1-\alpha)w \rceil$ p-value curves that lie above $Q^{l(\alpha)}$, and as the boundary curve in Goeman and Solari's method we can take the minimum of these curves.

Proposition 9.1. *There exists a $J' \in \Gamma$ such that $\underline{S}'(t) \leq \underline{S}_{J'}(t)$ for all $t \in [0, 1]$.*

Proof. By construction of Meinshausen'n boundary curve $Q^{l(\alpha)}$, we can (and do) choose $J' \in \Gamma$ such that

$$(P_{(1)}(g_j X), ..., P_{(m)}(g_j X)) \geq Q^{l(\alpha)}$$

for all $j \in J'$. In Goeman and Solari's method, the boundary curve $\mathbf{k} = (k_1, ..., k_m)$ depends on the chosen $J \in \Gamma$, so write $\mathbf{k} = \mathbf{k}^J = (k_1^J, ..., k_m^J)$. Let $1 \leq i \leq m$. Choose $j \in J'$ with $P_{(i)}(g_j X) = k_i^{J'}$. (That this is possible follows directly from the definition of $k_i^{J'}$.)

Since it holds that

$$Q_i^{l(\alpha)}(X, \Pi) \leq P_{(i)}(g_j X),$$

it follows that $Q_i^{l(\alpha)} \leq k_i^{J'}$. This holds for all $1 \leq i \leq m$, so

$$Q^{l(\alpha)} \leq \mathbf{k}^{J'}.$$

But this means that for all $\tau \in [0,t]$, $B(\tau)$ for Meinshausen's method (see Section 6.2) is larger than or equal to $B(\tau)$ for Goeman and Solari's method. Hence $\underline{S}'(t) \leq \underline{S}_{J'}(t)$ for all $t \in [0,1]$. □

Suggestion for future research

For the following we can take $\Gamma := \{A \subseteq \{1,...,m\} : \#A = \lceil (1-\alpha)w \rceil\}$. We now suggest a method to find a J that gives a good lower bound (at one point t), without 'breaking the rules' of Goeman and Solari's Method (i.e. such that Assumptions 1 and 2 still hold).

We want to optimize

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

Suppose we have calculated $\mathbf{Q}$. We will first define for each $0 \leq \tau \leq 1$ an estimator $\hat{R}(\tau)$ of $\#R(\tau)$, in such a way that $\mathbf{k} = \mathbf{k}^J$ is allowed to depend on $\hat{R}$. For each p-value curve $(P_{(1)}(g_i X), ..., P_{(m)}(g_i X))$, $1 \leq i \leq w$, calculate

$$N_i(\tau) := \#\{1 \leq j \leq m : P_j(g_i X) \leq \tau\}.$$

Define the estimator $\hat{R}(\tau)$ by $\hat{R}(\tau) = \max_{1 \leq i \leq w} N_i(\tau)$. This is an estimator of $\#R(\tau)$, since the original p-value curve contains relatively many of points with vertical coordinate below t (assuming at least quite a few null hypotheses are false and this is reflected in the p-values).

$B(\tau)$ depends on the boundary curve $\mathbf{k}$ chosen, so write $B(\tau) = B^{\mathbf{k}}(\tau)$. For each $\mathbf{k} \in \mathbf{K} := \{\mathbf{k}^{J'} : J' \in \Gamma\}$, define the estimator of $\underline{S}_{\mathbf{k}}(t)$ by

$$\underline{Z}_{\mathbf{k}}(t) := \max_{0 \leq \tau \leq t} \hat{R}(\tau) - B^{\mathbf{k}}(\tau).$$

Let $\hat{\mu} := \max\{\underline{Z}_{\mathbf{k}}(t) : \ \mathbf{k} \in \mathbf{K}\}$. Define $\mathbf{k}_0$ to be a random element from

$$V := \{\mathbf{k} \in \mathbf{K} : \ \underline{Z}_{\mathbf{k}}(t) = \hat{\mu}\}.$$

Use this boundary curve $\mathbf{k}_0$ for the calculation of the lower bound $\underline{S}(t)$ in Goeman and Solari's method.

To see that Assumptions 1 and 2 in Section 7.2 are satisfied for this definition of the boundary curve, note the following. For each $1 \leq i \leq w$, let P^i be the p-value curve $(P_{(1)}(g_i X), ..., P_{(m)}(g_i X))$. Note that given the *set* $\{P^1, ..., P^w\}$, $\hat{R}(\tau)$ and $\mathbf{K}$ are known, and if $\mathbf{k}$ is also given, then $B^{\mathbf{k}}(\tau)$ is know as well. Hence, given $\{P^1, ..., P^w\}$, $\{\underline{Z}_{\mathbf{k}}(t) : \ \mathbf{k} \in \mathbf{K}\}$ is known, hence μ is known. Hence V is known. Naturally, given $\mathbf{Q}$, the set $\{P^1, ..., P^w\}$ is known. Moreover, this set is invariant under interchanging of the rows of $\mathbf{Q}$. Thus the same holds for V, hence the same holds for $\mathbf{k}_0$. We conclude that Assumptions 1 and 2 are satisifed for this definition of the boundary curve.

Clearly, if $\underline{Z}_{\mathbf{k}}(t)$ is a good estimator for $\underline{S}_{\mathbf{k}}(t)$ for all $\mathbf{k} \in \mathbf{K}$, then $\hat{\mu}$ is a good estimator for $\mu := \max\{\underline{S}'_J(t) : \ J' \in \Gamma\}$. If we then use $\mathbf{k}_0$ in Goeman and

Solari's method, the obtained lower bound is near μ. Note that the larger the number of null hypotheses is and the clearer this is reflected in the p-values, the closer $\hat{R}(\tau)$ is likely to be to $\#R(\tau)$. Future research can tell how well this method performs.

Note that it might take long to compute $\hat{\mu}$, since it is defined as a maximum over sets which are themselves already defined as maximums. In that case ways should be sought to approximate $\hat{\mu}$ in less time. A simplistic solution would be to not calculate the maximum over all elements of Γ, but only over a random subset of Γ.

10 Discussion

When performing a permutation test (or a test using other transformations), it is often computationally infeasible to use all transformations. Fortunately, there are often ways to use only a small part of the given group of transformations, while still maintaining the same type I error rate. Firstly, one could of course use a subgroup of the given group, or one could use a completely different group of transformations. Secondly, there are sometimes cosets of equivalent transformations, and it suffices to use one element from each coset. Finally, random transformations from the group can be drawn. This has been noted in the literature. However, as we have shown, an additional identity transformation should always be added to the set of random transformations. (When drawing without replacement, the identity should be used exactly once.)

We have expanded two multiple testing methods: Meinshausen's method and Goeman and Solari's method. Both are methods for finding a uniform lower bound for the amount of correctly rejected hypotheses. We have shown that for both methods random transformations can be used, but the identity permutation should always be added. The proof of Meinshausen's method and Goeman and Solari's method with random permutations, is partly analogous to the proof of the basic permutation test using random transformations.

We have come to the conclusion that Meinshausen's method very often gives no non-trivial result when the columns of $\mathbf{P}$ are randomly shuffled in step 2 of this method. Moreover, Goeman and Solari's method gave better results for the data we used. We also compared Meinshausen's procedure *without* random shuffling of the columns, to Goeman and Solari's method. We used data that Meinshausen used in his paper. Goeman and Solari's method gave better bounds $\underline{S}(t)$ for most rejected sets, for the data we used. For more general statements, the methods can be compared for more datasets (and choices of J) in the future. Goeman and Solari's method is quite promising.

Goeman and Solari's method involves choosing an index set J, corresponding to the p-value curves that are selected to construct the boundary curve $\mathbf{k}$. There is always a J among the allowed index sets, for which Goeman and Solari's lower bound $\underline{S}(t)$ is uniformly at least as good as Meinshausen's bound. However, the process of choosing J should satisfy certain assumptions. At the end of Section 9 we have suggested a method to find a good J, such that these assumptions

are satisfied. Finding this J requires calculating a maximum of values $\underline{Z}_{\mathbf{k}}(t)$, which are themselves already defined as maximums. It is imaginable that this may become too time-consuming. In that case ways can be sought to reduce the computation time. Future research can show how well this method works.

A R script

We used the following *R* script to calculate lower bounds with Goeman and Solari's method and Meinshausen's method (with random column-shuffling). For more explanation, see Section 8.3, for which the code was used. The code used in the other sections is similar.

```
#Compares lower bounds. Each row of the matrix 'lowerb' gives
#three lowerbounds. The first two are the
#lower bounds found with Goeman and Solari's method without
#and with added identity permutation respectively. The third one is
#the lower bound found with Meinshausen's method (with added id).

nloops <-5 #the number of times that data are generated and
     #lower bounds are calculated
lowerb <- mat.or.vec(nloops,3)
nrejections <- mat.or.vec(nloops,1)
m <- 8
difr <-mat.or.vec(nloops,m)
betacheck<-matrix(1,nloops)


n <- 20
alpha <- 0.1
t <- 0.01
w <- 600    #the number of perms

for(f in 1:nloops){
Z <- matrix(rnorm(20, sd=0.5),1,n)
y <- rep(0:1,n/2)
X <- matrix(rnorm(n*m), n, m)

#make the first 4 X_i dependent of y
for(i in 1:4){
X[,i]<-X[,i]+2*y
}

#Make part of the values in each row of X correlated. (optional)
for (i in 1:n){
for (j in 3:6){
X[i,j] <- X[i,j] + Z[i]
}}


#make initial matrix of p-values
```

```
pvmatr.unsorted <- matrix(nrow=w,ncol=m)
pvmatr <- matrix(nrow=w,ncol=m)

#-----------------------------------------------------------------
#calculate the G&S-lower bound for when the identity is NOT added
#-----------------------------------------------------------------

for (i in 1:w){
yperm <- sample(y, size=n, replace=FALSE)
pvmatr.unsorted[i,] <- apply(X,2, function(h) t.test(h~yperm, var.equal=TRUE)$p.val)
pvmatr[i,]<-sort(pvmatr.unsorted[i,])
}

#make a vector with for each permutation the smallest p-value
sm.pvals <- mat.or.vec(1,w)
for (i in 1:w){
sm.pvals[i] <- min(pvmatr[i,])
}

#select the (alpha*w)-th smallest value among the w smallest p-values
v <- sort(pvmatr[,1])[floor(alpha*w)]

#let the critical curve K be the minimum of the (1-alpha)w curves for which
#the smallest p-value is larger than v

K <- mat.or.vec(1,m)+1
for(j in 1:w){
if(pvmatr[j,1]>=v){
for(i in 1:m){
K[i]<- min(K[i], pvmatr[j,i])
}}}

P <- sort(apply(X,2, function(h) t.test(h~y, var.equal=TRUE)$p.val))


#calculate the lower bound
for(j in 1:m){
if(P[j]<t){
R <- sum(P<=P[j])
B <- sum(K<=P[j])
RminB<-R-B
lowerb[f,1]<-max(lowerb[f,1],RminB)
}}

nrejections[f] <- sum(P<=t)
```

```
#-----------------------------------------------------------------
#calculate the G&S-lower bound for when the identity IS added
#-----------------------------------------------------------------

#'add' the identity permutation
pvmatr.unsorted[1,] <- apply(X,2, function(h) t.test(h~y, var.equal=TRUE)$p.val)
pvmatr[1,] <- sort(pvmatr.unsorted[1,])

#make a vector with for each permutation the smallest p-value
sm.pvals <- mat.or.vec(1,w)
for (i in 1:w){
sm.pvals[i] <- min(pvmatr[i,])
}

#select the (alpha*w)-th smallest value among the w smallest p-values
v <- sort(pvmatr[,1])[floor(alpha*w)]

#let the critical curve be the minimum of the (1-alpha)w curves for which
#the smallest p-value is larger than v

K <- mat.or.vec(1,m)+1
for(j in 1:w){
if(pvmatr[j,1]>=v){
for(i in 1:m){
K[i]<- min(K[i], pvmatr[j,i])
}}}


#calculate the lower bound

for(j in 1:m){
if(P[j]<t){
R <- sum(P<=P[j])
B <- sum(K<=P[j])
RminB<-R-B
lowerb[f,2]<-max(lowerb[f,2],RminB)
}}

#-----------------------------------------------------------------------
#calculate the Meinshausen-lower bound for when the identity is added
#-----------------------------------------------------------------------

#permute columns
permcol <- matrix(nrow=w, ncol=m)
for (i in 1:m){
permcol[,i] <- sample(pvmatr.unsorted[,i], size=w, replace=FALSE)
```

```
}

#sort each row
sortedrows <- matrix(nrow=w, ncol=m)
for (i in 1:w){
sortedrows[i,] <- sort(permcol[i,])
}

#sort each column (ending up with the matrix Q)
Q <- matrix(nrow=w, ncol=m)
for (i in 1:m){
Q[,i] <- sort(sortedrows[,i])
}

#sort the p-values of each permutation
psort <- matrix(nr=w,nc=m)
for (i in 1:w){
psort[i,] <- sort(pvmatr[i,])
}

#define beta (see Meinsh)
beta <- mat.or.vec(1,w)
for (i in 1:w){ for (j in 1:w){
if( all(Q[i,] <= psort[j,])){
beta[i] <- beta[i]+1
}}}
beta<-beta/w

#find l(alpha)
lalpha <- 0
for (i in 1:w){
if( beta[i] >= 1-alpha){
lalpha <- lalpha+1
}
}

betacheck[f]<-beta[1]

#calculate the lower bound

for(j in 1:m){
if(P[j]<t){
R <- sum(P<=P[j])
B <- sum(Q[lalpha,]<=P[j])
RminB<-R-B
lowerb[f,3]<-max(lowerb[f,3],RminB)
```

```
}}

difr[f,] <- K-Q[lalpha,] #difference between the two boundary curves

}
nrejections
lowerb
#difr
```

References

[1] Benjamini, Y., Hochberg, J. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289-300.

[2] Goeman, J., Solari, A. (2010). The sequential rejection principle of familywise error control. *The Annals of Statistics* **38**, 3782-3810.

[3] Goeman, J., Solari, A. (2011). Multiple Testing for Exploratory Research. *Stat. Science* **26**, 584-597.

[4] Goeman, J., Solari, A. (2012). Tutorial in biostatistics: multiple hypothesis testing in genomics. *Statist. Med.* **00**, 1-29.

[5] Goeman, J., Solari, A. (2013) Unpublished manuscript, which suggests a method similar to Meinshausen's method.

[6] Good, P. (2005) *Permutation, parametric, and bootstrap tests of hypotheses.* Springer, New York.

[7] Lehman, E., Romano, J. (2005) *Testing statistical hypotheses*, ch. 15. Springer, New York.

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

[9] Meinshausen, N. (2006). False Discovery Control for multiple Tests of Association Under General Dependence. *Scand. Journ. of Stat.* **33**, 227-237.

[10] Phipson, B., Smyth, G. (2010). Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology* **9**, article 39.

[11] Solari, A., Finos, L., Goeman, J. (2013). Unpublished preprint, "Rotation based multiple testing in the multivariate linear model".

[12] Southworth, L., Kim, S., Owen, A. (2009). Properties of balanced permutations. *Journal of Computational Biology* **16**, 625-638.

[13] Westfall, P., Troendle, J. (2008). Multiple testing with minimal assumptions. *Biometric Journal* **50**, 745-755.

[14] Westfall, P., Young, S. (1993) *Resampling-based multiple testing*, ch. 2. John Wiley & Sons, Canada.