Permutation-based inference for high-dimensional data.

Hemerik, J.

Permutation-based inference for high-dimensional data

Jesse Hemerik

Printing: Off Page, the Netherlands

ISBN: 978-94-6182-885-9

Permutation-based inference for high-dimensional data

Proefschrift

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van de Rector Magnificus prof.mr. C.J.J.M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op dinsdag 1 mei 2018
klokke 16.15 uur

door

Jesse Hemerik

geboren te Leiden in 1989

Promotores:	Prof. dr. J. J. Goeman
	Dr. A. Solari
	· *University of Milano-Bicocca, Milaan*
Leden promotiecommissie:	Prof. dr. S. le Cessie
	Prof. dr. A. W. van der Vaart
	· *Universiteit Leiden en Universiteit van Amsterdam*
	Prof. dr. M. A. van de Wiel
	· *Vrije Universiteit Medisch Centrum en Vrije Universiteit, Amsterdam*

Table of Contents

1 Introduction

This thesis is about statistical methods based on permutations or other transformations of data. Throughout much of this work, emphasis is on testing many hypotheses simultaneously, which is called *multiple testing*. A key message of this thesis is that by using permutation-based multiple testing, the dependence structure of the data can be taken into account, which leads to powerful methods.

In this introduction, we will first give some typical examples of *high-dimensional data*. We will then explain the key statistical issues when testing hypotheses about such data. Finally, we will introduce the new methods in this thesis, which address those issues.

1.1 High-dimensional data: examples

In many biomedical and psychological experiments, a large number of measurements are collected per individual, for example, measurements of many protein concentrations in the blood. Such measurements are referred to as *high-dimensional data*. We now give two examples of high-dimensional data that are frequently studied in respectively biomedical research and neuroscience.

Our DNA contains about 20,000 genes. A gene is a piece of DNA that codes for a molecule that has a function in the body. A gene contains a sequence of nucleotides. There are four types of nucleotides: adenine, thymine, guanine and cytosine (A, T, G and C). A gene is used to produce a string of RNA. Usually this is in turn used to code for the synthesis of a protein. Proteins are involved in various processes in our body, including our immune system and other processes related to disease. The speed at which a gene produces RNA strings to make proteins varies over time and per person and is called the *expression level* of that gene. By measuring expression levels and comparing these between different individuals, cells or tumors, we can learn whether these levels are associated with certain phenotypical traits, such as disease status. We can then build a model that predicts the disease type based on the gene expression profile. This can help to allow cheap, early and accurate diagnosis of diseases. Likewise, we can build a model that chooses the optimal treatment for a particular patient, based on his or her gene expression profile.

There are various technologies for measuring gene expression, each with their own pros and cons, related for example to financial costs and measurement accuracy. Until a decade ago the main method for measuring gene expression was to use *DNA microarrays*. A DNA microarray is a collection of microscopic DNA spots attached to a surface. A piece of RNA only binds to such a spot if it contains the complementary sequence of nucleotides. Thus, measuring how much RNA binds to a specific spot, leads to an estimate of the amount of the corresponding type of RNA in the sample. This is used to estimate the expression level of the gene that coded for this piece of RNA.

In the last decade more advanced techniques for measuring gene expression have appeared under the banner of *next-generation sequencing*. These techniques provide improvements such as lower costs and better quantification of lowly and highly expressed genes. Often the expression levels of thousands of genes are measured, leading to very high-dimensional data.

A different field of research where high-dimensional data occur, is the analysis of brain images. Researchers can measure the activity of various brain regions of a person, while that person performs some task. In this way it can be inferred which brain regions are involved in which types of human behavior.

Measurement of brain activity is often performed using *functional mag-*

netic resonance imaging (*fMRI*). In brain regions with high activity there is more deoxygenated hemoglobin in the blood, which interferes with a magnetic resonance signal, so that the active regions can be distinguished.

Often brain activity is measured of many thousands of small disjoint areas in the brain (referred to as *voxels*), which cover the whole brain. This is a good example of high-dimensional data. A large number of brain regions are investigated simultaneously, which usually means that the researcher tries to answer a lot of research questions at the same time. The more questions are asked, the more wrong answers could be given. Thus, great care should be taken in posing and addressing these questions, to limit the number of false research findings. (See (Bennett et al., 2009) for an example of how this can go wrong.)

1.2 Multiple testing and the false discovery proportion

Often many hypotheses are tested simultaneously. For example, for 20,000 genes one could test the hypothesis that their expression rate is not (substantially) associated with some phenotype. When multiple hypotheses are tested, the most classical way to avoid type-I errors, is to make sure that with high (e.g. 95%) probability, no true hypotheses are rejected. This is called strongly controlling the *Family-Wise Error Rate* (FWER). Controlling the FWER means that a hypothesis is only rejected if there is extremely strong proof to do this. However, often the proof against a single hypothesis is not extremely strong. The consequence may be that no rejections are made, even if there is (clinically relevant) signal in the data.

This way of avoiding type-I errors, i.e. controlling the probability of any type-I errors, is sometimes too strict. Indeed, it is not only important to avoid type-I errors, but also to not miss out on important scientific findings. Hence, researchers have been looking for less strict statistical approaches, which allow researchers better to detect the presence of false hypotheses. Note that such approaches allow more false findings than FWER controlling methods, so that the results have a different meaning.

Several less strict approaches exist, one of which is to select the rejected hypotheses in such a way that the *false discovery rate* (FDR) is controlled. The FDR is the expected value of the *false discovery proportion* (FDP). This is the fraction of false findings (type-I errors) among the rejected

hypotheses. It is defined to be 0 when there are no rejections. The most well-known FDR controlling method is by Benjamini and Hochberg (1995). Controlling the FDR means to keep the FDR below a certain value, e.g. 0.05. This guarantees that on average, the FDP is at most 0.05. (This does not mean that a rejected hypothesis is false with probability at least 0.95. Indeed, there can be a positive correlation between the FDP and the number of rejections.)

Compared to FWER control, controlling the FDR usually enables the user to reject more false hypotheses. On the other hand, with FDR controlling methods there are also more rejected true hypotheses. (Rather than 'rejected', we may use the term 'selected', since we are not extremely sure that a rejected hypothesis is false.) Correspondingly, controlling the FDR represents a trade-off between keeping the number of true positives large and keeping the number of false positives small.

When a researcher tests several hypotheses, he or she may want to be confident that the FDP is smaller than a certain value. However, when the FDR is smaller than 0.05, this does not mean that the FDP is likely smaller than 0.05. Thus, the researcher has an estimate of the FDP (namely, the nominal FDR), but not a confidence interval around the estimate. In many cases, it would be relevant to provide such a confidence interval. For example, it indicates how accurate the estimate is.

Therefore, statistical methods are being developed that provide a confidence interval for the FDP. Focus is usually on confidence upper bounds for the FDP. Similarly, methodology has appeared that allows controlling the FDP with confidence. This means to keep the FDP below some given, small value with some given, large probability. The present thesis provides several contributions in this field (chapters 3 and 4), as we will discuss in section 1.4.

1.3 Dependence and permutation methods

Most multiple testing methods perform many hypothesis tests separately and then combine the results to decide on which hypotheses to reject. The result of an individual test is usually a test statistic. In particular this could be a p-value. When m hypotheses are tested, m test statistics are obtained. These need to be used somehow to decide which hypotheses are rejected.

Often there are dependencies in the data. For example, the expression levels of different genes are correlated. The consequence is that the m test statistics are also dependent. The dependence structure influences the behavior of a multiple testing method. Hence, when choosing such a method, the dependence structure in the data should be taken into account.

In many situations, however, there is limited knowledge about the dependence structure in the data. Consequently, it is not known which multiple testing methods will control the considered error rate (e.g. the FWER or FDR) and which methods will not. This problem is often remedied by choosing the most conservative multiple testing method, to be sure that the error rate is controlled. The downside of this is that the rejection criterion is then likely to be more conservative than the user wanted.

A different approach to the problem of an *a priori* unknown dependence structure, is to use permutation-based multiple testing methods. Some of these methods have known, exact, finite-sample properties in several simple but important scenarios, such as randomized clinical trials. In multiple testing, permutation-based methods take into account the correlation structure in the data while making few assumptions on that structure. The only assumption made usually comes down to the following: the joint distribution of the part of the data under the null hypothesis should be invariant under permutation, e.g. shuffling cases and controls. (This usually implies that the hypotheses are point hypotheses, but extensions to interval hypotheses are sometimes possible.)

As a first example of a permutation-based multiple testing procedure, we consider the $maxT$ method by Westfall and Young (1993). This method strongly controls the FWER. It is the permutation-based analog of (the *closed testing*-based improvement of) Bonferroni's method. At confidence level 0.05, Bonferroni's method rejects all hypotheses with p-values smaller than $0.05/m$. The reason is that for any dependence structure of the p-values, $0.05/m$ does not exceed the 0.05-quantile of the minimum of the null p-values. The method by Westfall and Young (1993) however, makes use of the data to determine the rejection threshold. It permutes the data many times, each time calculating all m p-values and recording the smallest p-value. (Rather than p-values, other test statistics could be used, hence the name $\max T$.) The 0.05-quantile of these minimums is used as the rejection threshold. (This threshold, just as Bonferroni's threshold, can be further improved by using closed testing.) Especially when the p-values

are strongly positively correlated, this threshold can be much higher than $0.05/m$.

Another popular permutation-based multiple testing method is *Significance Analysis of Microarrays* (*SAM*) by Tusher et al. (2001). This method estimates the FDP. It is a general method, which is not at all limited to microarray data. The method works as follows. A rejection cut-off is chosen, such that each hypothesis is rejected only if its test statistic lies above the rejection cut-off. The data are permuted many times. Each time it is recorded how many hypotheses would have been rejected if the permuted data had been observed. The mean or median number of rejections for the permuted data, is an estimate of the number of false positives. Dividing this by the number rejections, gives an estimate of the FDP. In section 1.4 it is discussed how this method is improved in this thesis.

1.4 This thesis

Chapter 2 of this thesis is an introduction to the fundamentals of permutation testing. Although we will usually write 'permutation methods', most methods in this thesis are not restricted to permutations, but can use any group of transformations, such as rotations or sign-flipping.

Chapters 3 and 4 make use of some of the theory in chapter 2. Chapter 3 constructs confidence bounds for the FDP. Chapter 4 generalizes chapter 3 to simultaneous confidence bounds, which in particular allows for FDP control with confidence. Chapter 5 is quite different from chapters 2-4. It lets go of the fundamental assumption of distributional invariance under transformations. Rather, it provides tests which are only asymptotically exact. It presents methods for robust testing in generalized linear models (GLMs). Apart from being important on its own, chapter 5 broadens the applicability of the earlier chapters, allowing those methods to be used in the context of GLMs.

1.4.1 Permutation-based confidence on the false discovery proportion

Chapter 2 of this thesis lays the foundations for permutation testing based on randomly sampled permutations. Chapter 3 builds further on chapter 2 by improving the permutation-based multiple testing method SAM, which

was discussed above. The orginal SAM procedure only provided an estimate of the FDP, with unknown properties. We extend the method by providing a confidence upper bound for the FDP. We show that for $\alpha \in [0, 1)$, the $(1 - \alpha)$-quantile of the numbers of rejections for the permuted verisons of the data is a $(1 - \alpha)100\%$-confidence upper bound for the number of false positives. In particular, if we take $\alpha = 0.5$, we obtain a median unbiased estimate of the number of false positives.

Moreover, by using a method related to closed testing (Goeman and Solari, 2011), we obtain an improved confidence bound, which is at least as small as the basic bound. Although we derive a computational shortcut to compute the improved bound, when there are many hypotheses, the calculations are still computationally infeasible. Hence we provide a method which approximates the improved bound. The approximation method can be used when there are many thousands of hypotheses. In simulations, this method provided valid $(1 - \alpha)100\%$-confidence upper bounds.

Since our method is based on permutations, it tends to provide better (lower) FDP bounds than parametric methods. Another important advantage of the permutation approach is that the p-values (or test statistics) that the methods uses, need not be exact. The reason is that any test statistics can be used; the only assumption made by the method, is permutation-invariance of the part of the data under the null.

Chapter 4 generalizes the method of chapter 3 to simultaneous confidence bounds. For a range of rejection thresholds of interest, it provides confidence upper bounds for the FDP. These bounds are simultaneous: with probability at least $1-\alpha$, these bounds are all valid. This allows for *post hoc* selection of the rejection threshold: if the threshold is chosen based on the data, then still a valid confidence bound is obtained. Thus, based on the data, the user could choose a threshold that leads to a desired confidence bound for the FDP.

A similar method was already published by Meinshausen (2006). This method served as an inspiration for chapters 3 and 4 of this thesis. In chapter 4 we improve the method of Meinshausen by 1. modifying it to be exact, 2. generalizing it and 3. improving its power. In particular, we construct an iterative method, which has more power than the single-step method by Meinshausen (2006) and is still exact. We also provide a fast approximation of this method, which provided valid confidence bounds in our simulations.

1.4.2 Robust testing in generalized linear models

It is often of interest to test if a coefficient in a GLM equals 0 or some other value, or to construct a confidence interval around an estimated coefficient. When the specified model is correct, this can be done with classical parametric tests such as the likelihood ratio test or score test. In many situations however, the assumed model is not even approximately exact, due to e.g. overdispersion, heteroscedasticity or ignored nuisance parameters. The traditional parametric tests then lose their properties.

In chapter 5 we propose a novel type of test, based on sign-flipping score contributions. (Note that the score, the derivative of the log-likelihood, is a sum of n individual contributions.) The basic idea underlying this test is the following. Under a point null hypothesis of the form $H_0 : \beta = 0$, the score contributions have mean 0. By recalculting the score many times after randomly multiplying the summands by -1 with probability 0.5, we obtain a reference distribution, that we compare with the original score. We reject the null hypothesis if the original score lies in the 5%-tails of this distribution. This method is closely related to the permutation test. It is often robust against the mentioned forms of model misspecification. For example, if there is constant overdispersion, then the variance of the score is misspecified, so that the classical parametric tests fail. Our test remains asymptotically exact, however, since both the observed score and flipped scores are misspecified by the same factor.

A key assumption underlying this test, is that the n score contributions are independent. When nuisance parameters are estimated, this is no longer the case. As a consequence, the variance of the original score is shrunk and our test becomes conservative. To solve this issue, we consider the *effective score* (Marohn, 2002). This is the part of the score that is orthogonal to the nuisance score. It is asymptotically unaffected by the nuisance estimation. As a consequence, we again obtain an asymptotically exact test.

If the scores were independent and symmetric, then our sign-flipping test would be a special case of the exact test in chapter 2. Due to its close relationship to permutation tests, our sign-flipping test can be combined with the multiple testing methods from chapters 3 and 4. This allows using those methods in the context of GLMs.

2

Exact testing with random permutations

Abstract

When permutation methods are used in practice, often a limited number of random permutations are used to decrease the computational burden. However, most theoretical literature assumes that the whole permutation group is used, and methods based on random permutations tend to be seen as approximate. There exists a very limited amount of literature on exact testing with random permutations and only recently a thorough proof of exactness was given. In this paper we provide an alternative proof, viewing the test as a "conditional Monte Carlo test" as it has been called in the literature. We also provide extensions of the result. Importantly, our results can be used to prove properties of various multiple testing procedures based on random permutations.

This chapter has been published as: Jesse Hemerik and Jelle Goeman (2017). Exact testing with random permutations. *TEST* (Online First version)

2.1 Introduction

Permutation tests are nonparametric tests that are used in particular when the null hypothesis implies distributional invariance under certain transformations (Fisher, 1936; Lehmann and Romano, 2005; Ernst et al., 2004). Apart from permutations, other groups of transformations can be used, such as rotations (Langsrud, 2005).

When the set of transformations used is not a group, a permutation test can be very conservative or anti-conservative. The first author who explicitly assumed a group structure is Hoeffding (1952). The role of the group structure has recently been emphasized (Southworth et al., 2009; Goeman and Solari, 2010). Southworth et al. (2009) note that in particular the set of 'balanced permutations' cannot be used, since it is not a group.

Often it is computationally infeasible to use the whole group of permutations, due to its large cardinality. In that case random permutations are used, as was first proposed by Dwass (1957). Often a permutation p-value based on random permutations is simply seen as an estimate of the permutation p-value.

It is known that naively using random permutations instead of all possible permutations can lead to extreme anti-conservativeness (Phipson and Smyth, 2010), especially when combined with multiple testing procedures. Therefore sometimes the identity permutation, which corresponds to the original observation, is included with the random permutations (Ge et al., 2003; Lehmann and Romano, 2005). Lehmann and Romano (2005) (p.636) state that when the identity is added, the estimated p-value is stochastically larger than the uniform distribution on $[0, 1]$ under the null. Phipson and Smyth (2010) note that adding the identity can make the permutation test exact, i.e. of level α exactly. They do not mention the role of the underlying group structure. Instead they view the permutation test as a Monte Carlo test, which is known to be exact in some situations if the original observation is added.

Referring to Monte Carlo is not sufficient, because despite being related, a Monte Carlo test is very different from a permutation test. Monte Carlo samples are draws from the null distribution. In the permutation context, the random permutations of the data are instead drawn from a conditional null distribution, i.e. the permutation distribution. Hence the proof by Phipson and Smyth (2010) is incomplete and it remained unclear what assumptions (e.g. a group structure) are essential for the validity of random

permutation tests. For example, it is unclear from Phipson and Smyth that random sampling from balanced permutations would lead to invalid tests.

In Hemerik and Goeman (2018) a test is given based on random transformations. In the present paper we extend this work, investigating fundamental properties of random permutation tests. Our main focus is on the level of tests. Other properties, e.g. consistency, do not generally hold but can be established for more specific scenarios (Lehmann and Romano, 2005; Pesarin, 2015; Pesarin and Salmaso, 2013) by using results presented here. Our results are general and can be used to prove properties of various multiple testing methods based on random permutations, such as Westfall and Young (1993), Tusher et al. (2001), Meinshausen and Bühlmann (2005) and Meinshausen (2006). In the literature there are two approaches to proving permutation tests with fixed permutations: a conditioning-based approach (Pesarin, 2015) and a more direct approach (Hoeffding, 1952; Lehmann and Romano, 2005). We will give proofs with both approaches.

The structure of the paper is as follows. In section 2.2 we review known results on the level of a permutation test based on a fixed group of transformations. The concepts and definitions from section 2.2 are used throughout the paper. Testing with random permutations is covered in Section 2.3. In section 2.3.1 permutation tests are contrasted with Monte Carlo tests. Estimation of p-values is discussed in section 2.3.2. Exact tests and p-values based on random transformation are given in sections 2.3.3 and 2.3.4. In section 2.4 some additional applications of these results are mentioned.

2.2 Fixed transformations

Here we discuss tests that use the full group of transformations.

2.2.1 Basic permutation test

Let X be data taking values in a sample space $\mathcal{X}$. Let G be a finite set of transformations $g : \mathcal{X} \to \mathcal{X}$, such that G is a group with respect to the operation of composition of transformations. This means that G satisfies the following three properties: G contains an identity element (the map $x \mapsto x$); every element of G has an inverse in G; for all $a_1, a_2 \in G$, $a_1 \circ a_2 \in G$. This assumption of a group structure for G is fundamental throughout the paper, since it ensures that $Gg = G$ for all $g \in G$, i.e. that

the set G is permutation-invariant.

Considering a general group of transformations rather than only permutations is useful, since in many practical situations the group consists of e.g. rotations (Langsrud, 2005; Solari et al., 2014) or maps that multiply part of the data by -1 (Pesarin and Salmaso (2010), pp. 54 and 168). We write $g(X)$ as gX. Consider any test statistic $T : \mathcal{X} \to \mathbb{R}$. Throughout this paper we are concerned with testing the following null hypothesis of permutation-invariance.

Definition 2.2.1. Let H_{p} be any null hypothesis which implies that the joint distribution of the test statistics $T(gX)$, $g \in G$, is invariant under all transformations in G of X. That is, writing $G = \{a_1, ..., a_{\#G}\}$, under H_{p}

$$\big(T(a_1X), ..., T(a_{\#G}X)\big) \stackrel{d}{=} \big(T(a_1gX), ..., T(a_{\#G}gX)\big) \tag{2.1}$$

for all $g \in G$.

Note that (2.1) holds in particular when for all $g \in G$

$$X \stackrel{d}{=} gX.$$

Composite null hypotheses are usually not of the form H_{p}, but for specific scenarios, properties of tests of such hypotheses can be established using results in this paper.

The most basic permutation test rejects H_{p} when $T(X) > T^{(k)}(X)$, where

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

are the sorted test statistics $T(gX)$, $g \in G$, and $k = \lceil (1-\alpha)\#G \rceil$ with $\alpha \in [0, 1)$. As is known and stated in the following theorem, this test has level at most α.

Theorem 2.2.1. *Under H_p, $\mathbb{P}\big\{T(X) > T^{(k)}(X)\big\} \leq \alpha$.*

We now give two proofs: a conditioning-based approach and an approach without conditioning. Both approaches are more or less known. The conditioning-based proof is similar to that in Pesarin (2015) but the setting is more general. For each $x \in \mathcal{X}$, define O_x to be the orbit of x, which is the set $\{gx : g \in G\} \subseteq \mathcal{X}$.

Proof. Let $A = \{x \in \mathcal{X} : T(x) > T^{(k)}(x)\}$ be the set of elements of the sample space that lead to rejection. Suppose H_{p} holds. By the group structure, $Gg = G$ for all $g \in G$. Consequently, $T^{(k)}(gX) = T^{(k)}(X)$ for all $g \in G$. Thus $\#\{g \in G : gX \in A\} =$

$$\#\{g \in G : T(gX) > T^{(k)}(gX)\} = \#\{g \in G : T(gX) > T^{(k)}(X)\} \leq \alpha \#G.$$

Endow the space of orbits with the σ-algebra that it inherits from the σ-algebra on $\mathcal{X}$. Analogously to the proof of theorem 15.2.2 in Lehmann and Romano (2005), we obtain

$$\mathbb{P}(X \in A \mid O_X) = \frac{1}{\#G} \#\{g \in G : gX \in A\}.$$

By the argument above, this is bounded by α. Hence

$$\mathbb{P}(X \in A) = E\big\{\mathbb{P}(X \in A \mid O_X)\big\} \leq \alpha$$

as was to be shown. □

We now state a different proof without conditioning. A similar proof can be found in Hoeffding (1952) and Lehmann and Romano (2005) (p. 634).

Proof. By the group structure, $Gg = G$ for all $g \in G$. Hence $T^{(k)}(gX) = T^{(k)}(X)$ for all $g \in G$. Let h have the uniform distribution on G. Then under H_{p}, the rejection probability is

$$\begin{aligned}
&\mathbb{P}\big\{T(X) > T^{(k)}(X)\big\} = \\
&\mathbb{P}\big\{T(hX) > T^{(k)}(hX)\big\} = \\
&\mathbb{P}\big\{T(hX) > T^{(k)}(X)\big\}.
\end{aligned}$$

The first equality follows from the null hypothesis and the second equality holds since $T^{(k)}(X) = T^{(k)}(hX)$. Since h is uniform on G, the above probability equals

$$\mathbb{E}\Big[(\#G)^{-1} \cdot \#\big\{g \in G : T(gX) > T^{(k)}(X)\big\}\Big] \leq \alpha,$$

as was to be shown.

□

The test of theorem 2.2.1 is not always exact. When the data are discrete then the basic permutation test is often slightly conservative, due to a non-zero probability of tied values in X. Under the following condition, which is often satisfied for continuous data, but usually not for discrete data, the test is exact for certain values of α.

Condition 2.2.1. There is a partition $\{G_1, ..., G_m\}$ of G with $id \in G_1$ and $\#G_1 = ... = \#G_m$, such that under H_p with probability 1 for all $g, g' \in G$, $T(gX) = T(g'X)$ if and only if g and g' are in the same set G_i.

Proposition 2.2.1. *Under condition 2.2.1, the test of theorem 2.2.1 is exact if and only if* $\alpha \in \{0, 1/m, ..., (m-1)/m\}$.

The proof of this result is analogous to the proof of theorem 2.2.1. As an example where condition 2.2.1 holds, consider a randomized trial where $X \in \mathbb{R}^{2n}$ and the test statistic is

$$T(X) = \sum_{i=1}^{n} X_i - \sum_{i=n+1}^{2n} X_i, \tag{2.2}$$

where $X_1, ..., X_n$ are cases and $X_{n+1}, ..., X_{2n}$ are controls and all X_i are independent and identically distributed under the null. Let

$$m = \binom{2n}{n}.$$

If the observations are continuous then the set of α for which the test is exact is $\big\{0, 1/m, ..., (m-1)/m\big\}$, reflecting the fact that there are m equivalence classes of size $n!n!$ of permutations that always give the same test statistic.

The test of theorem 2.2.1 is often conservative when the data are discrete, since then condition 2.2.1 is usually not satisfied. Moreover, in many cases, the value 0.05 is not in the set mentioned in proposition 2.2.1 and hence the permutation test for $\alpha = 0.05$ is conservative, even if condition 2.2.1 is satisfied. The test can be adapted to be exact by randomizing it, i.e. by rejecting H_p with a suitable probability a in the boundary case that $T(X) = T^{(k)}$ (Hoeffding, 1952). Here

$$a = a(X) = \frac{\alpha \#G - M^+(X)}{M^0(X)}, \tag{2.3}$$

where

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

This adaptation has the advantage that it is always exact. Even if condition 2.2.1 is satisfied, the adaptation can be useful to guarantee that the level of the test is exactly the nominal level α. On the other hand, this test is less reproducible than the test of theorem 2.2.1, since its outcome may depend on a random decision. Which test is to be preferred, would depend on the context.

When the set G is not a group, the test can be highly anti-conservative or conservative. For example, the set of *balanced permutations* is a subset of the set of all permutations, but is not a subgroup. These permutations have been used in various papers since they can have an intuitive appeal. They are discussed in Southworth et al. (2009), who warn against their use since they lead to anti-conservative tests. The fact that permutations have been used incorrectly illustrates that more emphasis should be put on the assumption of a group structure.

Intuitively, the reason why a group structure is needed for theorem 2.2.1 is the following. Suppose for simplicity that H_p implies that $X \stackrel{d}{=} gX$ for all $g \in G$. The permutation test works since under H_p, for every permutation $g \in G$ the probability $\mathbb{P}\{T(gX) > T^{(k)}(X)\}$ is the same. The reason is that under H_p, for every $g \in G$, the joint distribution of gX and the set GX, i.e. of (gX, GX), is the same. Indeed, since $G = Gg$ (group structure), the set GX is a function of gX, namely $GX = f(gX)$, with f given by $f(x) = Gx$. Thus, for $g, g' \in G$, $(gX, GX) = (gX, f(gX)) \stackrel{d}{=} (g'X, f(g'X)) = (g'X, GX)$. When G is not a group, the joint distribution of gX and the set GX is not generally independent of g.

2.2.2 Permutation *p*-values

Permutation p-values are p-value based on permutations of the data. Here we will discuss permutation p-values based on the full permutation group. p-values based on random permutations are considered in section 2.3.4.

It is essential to note that there is often no unique null distribution of $T(X)$, since H_p often does not specify a unique null distribution of the data. Correspondingly, $T^{(k)}(X)$ should not be seen as the $(1-\alpha)$-quantile of *the* null distribution.

When a test statistic t is a function (which is not random) of the data and has a unique distribution under a hypothesis H, then a p-value in the strict sense, $\mathbb{P}_H(t \geq t_{obs})$, is defined where t_{obs} is the observed value of t. Since under H_p $T(X)$ does not always have a unique null distribution, often there exists no p-value in the strict sense based on this test statistic. However, under condition 2.2.1 the statistic

$$D = \#\big\{g \in G : T(gX) \geq T(X)\big\}$$

does have a unique null distribution. Thus a p-value in the strict sense based on $-D$ is then defined. Denoting by d the observed value of D, we have under H_{p}

$$\mathbb{P}(-D \geq -d) = \mathbb{P}(D \leq d) = \mathbb{P}\big\{T(X) > T^{(\#G-d)}(X)\big\} = \frac{d}{\#G}.$$

This is indeed what is usually considered to be the permutation p-value. This equality holds under condition 2.2.1. In other cases, such as when the observations are discrete, the null hypothesis often does not specify a unique null distribution of D. Thus there is not always a p-value in the strict sense based on D.

When H_p does not specify a unique null distribution of any sensible test statistic, as a resolution a 'worst-case' p-value could be defined. However sometimes better solutions are possible, e.g. the randomized p-value p' in section 2.3.4. In general, a p-value in the weak sense can be considered, i.e. any random variable p satisfying $\mathbb{P}(p \leq c) \leq c$ for all $c \in [0, 1]$ for every distribution under the null hypothesis. For H_{p}, $D/\#G$ is always a p-value in the weak sense.

2.3 Random transformations

In section 2.3 we extend the results of the previous section to tests based on random transformations. Since permutation testing with random permutations is often confused with Monte Carlo testing, in Section 2.3.1 the differences between the two are made explicit. Since random permutations are often used for estimation (rather than exact computation) of p-values, estimation of permutation p-values is discussed in section 2.3.2. Exact tests and p-values are given in sections 2.3.3 and 2.3.4 respectively. These two sections contain most of the novel results of the paper.

2.3.1 Comparison of Monte Carlo and permutation tests

In a basic Monte Carlo experiment, the null hypothesis H_0 is that X follows a specific distribution. A Monte Carlo test is used when there is no analytical expression for the $(1-\alpha)$-quantile of the null distribution of $T(X)$, such that the observed value of $T(X)$ cannot simply be compared to this quantile. To test H_0, independent realizations $X_2, ..., X_w$ are drawn from the null distribution of X. Assume that $T(X), T(X_2), ..., T(X_w)$ are continuous. Writing $X_1 = X$, let

$$B' = \#\{1 \leq j \leq w : T(X_j) \geq T(X)\}$$

and let b' denote its observed value. It is easily verified that under H_0, B' has the uniform distribution on $\{1, ..., w\}$.

The Monte Carlo test rejects H_0 when $T(X) > T^{(k')}$, where $k' = \lceil (1-\alpha)w \rceil$ and $T^{(1)} \leq ... \leq T^{(w)}$ are the sorted test statistics $T(X_j)$, $1 \leq j \leq w$. Note that $T^{(k')}$ is not the exact $(1-\alpha)$-quantile of the null distribution of $T(X)$, but nevertheless the test is exact. The reason is that the null distribution of B' is known. The test rejects H_0 if and only if $-B'$ exceeds the $(1-\alpha)$-quantile of its null distribution. Equivalently, it rejects when the Monte Carlo p-value

$$\mathbb{P}_{H_0}(B' \leq b') = b'/w,$$

where b' is the observed value of B', is at most α.

The validity of a random permutation test is not as obvious. Let $g_2, ..., g_w$ be random permutations from G. (There are various ways of sampling them, which we discuss later.) One permutation, g_1, is fixed to be $id \in G$, reflecting the original observation. Then, similarly to a Monte Carlo test, the permutation test rejects H_p if and only if $T(X) > T^{(k')}(X)$, where now $T^{(1)} \leq ... \leq T^{(w)}$ are the sorted test statistics $T(g_j X)$, $1 \leq j \leq w$.

Note that contrary to the Monte Carlo sample $X_1, ..., X_w$, the permutations $g_1 X, ..., g_w X$ are not independent under the null. Thus the random permutation test is not analogous to the Monte Carlo test. To prove the validity of the test based on random permutations, we must use that $g_1 X, ..., g_w X$ are independent and identically distributed conditionally on the orbit O_X. It is however not obvious what properties G should have in order that $g_1 X = X$ can be seen as a random draw from O_X conditionally on O_X. It will be seen that it suffices that G is a group. In that case, the test can be said to be a 'conditional Monte Carlo test'.

2.3.2 Estimated p-values

In practice it is often computationally infeasible to calculate the permutation p-value based on the whole permutation group, $D/\#G$. To work around this problem, there are two approaches in the literature. In both approaches, random permutations are used. The first approach is *calculating* (rather than estimating) a p-value based on the random permutations. This is discussed in section 2.3.4. The second approach is *estimating* the p-value $D/\#G$, which we discuss now.

In practice the p-value $p = D/\#G$ is often estimated using random permutations. The random permutations are typically all taken to be uniform on G and can be drawn with or without replacement. The estimate of p is often taken to be $\hat{p} = B/w$, with B as defined above. This is an unbiased estimate of p, i.e. $E\hat{p} = p$, and usually $\lim_{w\to\infty} \hat{p} = p$.

A more conservative estimate $\tilde{p} = (B + 1)/(w + 1)$ is sometimes also used. This formula is discussed in section 2.3.4.

Using the unbiased estimate $\hat{p} = B/w$ can be very dangerous, as Phipson and Smyth (2010) thoroughly explain. The reason is that $\hat{p}$ is almost never stochastically larger than the uniform distribution on $[0, 1]$ under H_{p}. This is immediately clear from the fact that $\hat{p}$ usually has a strictly positive probability of being zero. Consequently, if H_{p} is rejected if $\hat{p} \leq c$ for some cut-off c, then the type-I error rate can be larger than c. Often this difference will be small for large w. However, when c is itself small due to e.g. Bonferroni's multiple testing correction, then $\mathbb{P}(\hat{p} \leq c)$ can become many times larger than c under H_{p}. This is because this probability does not converge to zero as $c \downarrow 0$ for fixed w. Thus, as Phipson and Smyth (2010) note, using $\hat{p}$ in combination with e.g. Bonferroni can lead to completely faulty inference. Appreciable anti-conservativeness also occurs if very few (e.g. 25–100) random permutations are used (as in e.g. Byrne et al. (2013) and Schimanski et al. (2013)).

When possible, computing exact p-values is always to be preferred over estimating p-values. Exact p-values based on random permutations are given in section 2.3.4.

2.3.3 Random permutation tests

Here we discuss exact tests based on random transformations. Apart from theorem 2.3.1 (Hemerik and Goeman, 2018), the results in this section are

novel.

Phipson and Smyth (2010) also consider exact p-values based on random permutations. The proofs in Phipson and Smyth (2010) are incomplete, since they do not show the role of the group structure of the set of all permutations. Lehmann and Romano (2005) (p.636) remark without proof that if G is a group, then under H_{p} the p-value $(B+1)/(w+1)$ is always stochastically larger than uniform on $[0,1]$, but they state no other properties. In Hemerik and Goeman (2018) for the first time a theoretical foundation is given for the random permutation test, using the group structure of the set G. Here this work is extended with additional results.

Theorem 2.3.1 states that the permutation test with random permutations has level at most α if the identity map is added. This was remarked several times in the literature and proved in Hemerik and Goeman (2018). We first define the vector of random transformations.

Definition 2.3.1. Let G' be the vector $(id, g_2, ..., g_w)$, where id is the identity in G and $g_2, ..., g_w$ are random elements from G. Write $g_1 = id$. The transformations can be drawn either with or without replacement: the statements in this paper hold for both cases. If we draw $g_2, ...g_w$ *without* replacement, then we take them to be uniformly distributed on $G \setminus \{id\}$, otherwise uniform on G. In the former case, $w \leq \#G$.

Theorem 2.3.1. *Let G' be as in Definition 3.2.1. Let $T^{(1)}(X, G') \leq ... \leq T^{(w)}(X, G')$ be the ordered test statistics $T(g_j X)$, $1 \leq j \leq w$. Let $\alpha \in [0, 1)$ and recall that $k' = \lceil (1-\alpha)w \rceil$.*

Reject H_p when $T(X, G') > T^{(k')}(X, G')$. Then the rejection probability under H_p is at most α.

A proof of theorem 2.3.1 is in Hemerik and Goeman (2018) and we recall it here.

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

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

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

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

Since $h = g_j$ with j uniform, this equals

$$\mathbb{E}\Big[w^{-1}\#\big\{1 \leq j \leq w : T^{(j)}(X, G') > T^{(k')}(X, G')\big\}\Big] \leq \alpha,$$

as was to be shown. □

We now prove theorem 2.3.1 with a conditioning-based approach, viewing the test as a "conditional Monte Carlo" test as it has been called in the literature.

Proof. We prove the result for the case of drawing with replacement. The proof for drawing without replacement is analogous. Note that (X, G') takes values in $\mathcal{X} \times \{id\} \times G^{w-1}$. Let $A \subset \mathcal{X} \times \{id\} \times G^{w-1}$ be such that the test rejects if and only if $(X, G') \in A$.

Endow the space of orbits with the σ-algebra that it inherits from the σ-algebra on $\mathcal{X}$. Suppose H_{p} holds. Assume that almost surely O_X contains $\#G$ distinct elements. In case not, the proof is analogous. Analogously to the proof of theorem 15.2.2 in Lehmann and Romano (2005), we obtain

$$\mathbb{P}\big\{(X, G') \in A \mid O_X\big\} = \frac{\#\big(O_X \times \{id\} \times G^{w-1}\big) \cap A}{\#O_X \times \{id\} \times G^{w-1}}. \tag{2.4}$$

We now argue that this is at most α. Fix X. Let $\tilde{X}$ have the uniform distribution on O_X. It follows from the group structure of G that the entries of $G'\tilde{X}$ are just independent uniform draws from O_X. Thus from the Monte Carlo testing principle it follows that $\mathbb{P}\{(\tilde{X}, G') \in A\big\} \leq \alpha$. Since $(\tilde{X}, G')$ was uniformly distributed on $O_X \times \{id\} \times G^{w-1}$, it follows that (2.4) is at most α. We conclude that

$$\mathbb{P}\big\{(X, G') \in A\big\} = E\Big[\mathbb{P}\big\{(X, G') \in A \mid O_X\big\}\Big] \leq \alpha,$$

as was to be shown. □

Theorem 2.3.1 implies that $(B+1)/(w+1)$ is always a p-value in the weak sense if all random permutations (including g_1) are uniform draws with replacement from G or without replacement from $G\backslash\{g_1\}$. Under more specific assumptions, theorem 2.3.1 can be extended to certain composite null hypotheses. Proposition 2.3.1 states that under condition 2.2.1 and suitable sampling, the test with random permutations is exact. The formula in Section 2.3.4 for the p-value under sampling without replacement is equivalent to the last part of this result.

Proposition 2.3.1. *Suppose condition 2.2.1 holds. Let $h_1 \in G_1, ..., h_m \in G_m$. Then the result of theorem 2.3.1 still holds if $g_2, ..., g_w$ are drawn with replacement from $\{h_1, ..., h_m\}$ or without replacement from $\{h_2, ..., h_m\}$. Moreover, in the latter case, the test of theorem 2.3.1 is exact for all $\alpha \in \{0/w, 1/w, ..., (w-1)/w\}$.*

Proof. We consider the case that $g_2, ..., g_w$ are drawn without replacement from $\{h_2, ..., h_m\}$ and show that the test is exact for $\alpha \in \{0/w, ..., (w-1)/w\}$. Write $G' = (g_1, ..., g_w)$. Let h have the uniform distribution on $\{g_1, ..., g_w\}$. For each $g \in G$ let $i(g) \in \{1, ..., m\}$ be such that $g \in G_{i(g)}$. Suppose H_p holds. From the group structure of G it follows that the sets $\big\{i(g_1), ..., i(g_w)\big\}$ and $\big\{i(g_1h^{-1}), ..., i(g_wh^{-1})\big\}$ have the same distribution. Consequently

$$\mathbb{P}\big\{T(X) > T^{(k')}(X, G')\big\} =$$
$$\mathbb{P}\big\{T(X) > T^{(k')}(X, G'h^{-1})\big\}.$$

As in the above proof of theorem 2.3.1 we find that this equals $\mathbb{P}\big\{T(hX) > T^{(k')}(X, G')\big\}$.

Since $\alpha \in \{0/w, ..., (w-1)/w\}$ and $T(g_1X), ..., T(g_wX)$ are distinct, it holds with probability one that

$$\#\big\{1 \le j \le w : T(g_jX) > T^{(k')}\big\} = \alpha w.$$

Since h is uniform, it follows that $\mathbb{P}\big\{T(hX) > T^{(k')}(X, G')\big\} = \alpha$. □

Using this result it can be shown that specific tests with random permutations are unbiased. The test of theorem 2.3.1 can be slightly conservative if α is not chosen suitably or due to the possibility of ties. Recall that the same holds for the basic permutation test that uses all transformations in G. The adaptation by Hoeffding at (2.3) then guarantees exactness. The

following is a generalization of Hoeffding's result to random transformations.

Proposition 2.3.2. *Consider the setting of theorem 2.3.1. Let*

$$a = a(X, G') = \frac{w\alpha - M^+(X, G')}{M^0(X, G')}, \tag{2.5}$$

where

$$M^+(X, G') := \#\{1 \le j \le w : T(g_j X) > T^{(k')}(X, G')\},$$

$$M^0(X, G') := \#\{1 \le j \le w : T(g_j X) = T^{(k')}(X, G')\}.$$

Reject if $T(X) > T^{(k')}(X, G')$ and reject with probability a if $T(X) = T^{(k')}(X, G')$. Then the rejection probability is exactly α under H_p.

Proof. Assume H_{p} holds. Note that

$$\mathbb{P}(\text{reject}) = E\big\{\mathbb{1}_{\{T(X)>T^{(k')}(X,G')\}} + a(X, G')\mathbb{1}_{\{T(X)=T^{(k')}(X,G')\}}\big\}.$$

Write $M^+ = M^+(X, G')$ and $M^0 = M^0(X, G')$. Analogously to the first four steps of the first proof of theorem 2.3.1, it follows for h as defined there that the above equals

$$\begin{aligned}
&E\big\{\mathbb{1}_{\{T(hX)>T^{(k')}(X,G')\}} + a(X, G')\mathbb{1}_{\{T(hX)=T^{(k')}(X,G')\}}\big\} = \\
&E\big\{\mathbb{1}_{\{T(hX)>T^{(k')}(X,G')\}}\big\} + E\big\{a(X, G')\mathbb{1}_{\{T(hX)=T^{(k')}(X,G')\}}\big\} = \\
&E\big\{M^+ w^{-1}\big\} + E\Big[E\Big\{\frac{w\alpha - M^+}{M^0}\mathbb{1}_{\{T(hX)=T^{(k')}(X,G')\}} \mid M^+, M^0\Big\}\Big] = \\
&E\big\{M^+ w^{-1}\big\} + E\Big[\frac{w\alpha - M^+}{M^0}E\Big\{\mathbb{1}_{\{T(hX)=T^{(k')}(X,G')\}} \mid M^+, M^0\Big\}\Big] = \\
&E\big\{M^+ w^{-1}\big\} + E\Big[\frac{w\alpha - M^+}{M^0}M^0 w^{-1}\Big] = \alpha,
\end{aligned}$$

as was to be shown. □

The test of proposition 2.3.2 entails a randomized decision: in case $T(X) = T^{(k')}$, the test randomly rejects with probability a. This is in itself not objectionable, since the test is randomized anyway due to the random transformations. Note that in the situation of proposition 2.3.1

under drawing without replacement the test is already exact, such that proposition 2.3.2 is not needed to obtain an exact test.

In theorem 2.3.1, the requirement of using the whole group is replaced by suitable random sampling from the group. Interestingly, the following sampling scheme is also possible. Let $G^* \subseteq G$ be any finite subset of G, where we now allow G to be an infinite group as well. Write $k^* = \lceil (1-\alpha)\#G^* \rceil$. Let h be uniformly distributed on G^* and independent. Reject H_p if and only if

$$T(X) > T^{(k^*)}(X, G^*h^{-1}),$$

i.e. if $T(X)$ exceeds the $(1-\alpha)$-quantile of the values $T(gh^{-1})$, $g \in G^*$. This is a randomized rejection rule, since it depends on h, which is randomly drawn each time the test is executed. The rejection probability is at most α, which follows from an argument analogous to the last five steps of the first proof of theorem 2.3.1. Note that if G^* is a group itself, then $G^*h^{-1} = G^*$ and this test becomes non-random, coinciding with the basic permutation test. Thus it is a generalization thereof. This result allows using a permutation test when G is an infinite group of transformations, from which it may not be obvious how to sample uniformly. One simply uses any finite subset G^* of the infinite group.

2.3.4 *p*-values based on random transformations

Phipson and Smyth (2010) give formulas for p-values, when permutations are randomly drawn. Here we provide the required assumptions and proofs, which follow from section 2.3.3. We then provide some additional results.

Write

$$B = \#\big\{1 \leq j \leq w : T(g_j X) \geq T(X)\big\}, \tag{2.6}$$

where $g_1, ..., g_w$ are random permutations with distribution to be specified. Let b be the observed value of B. Under condition 2.2.1, Phipson and Smyth's p-values are exactly equal to $\mathbb{P}_{H_\mathrm{p}}(-B \geq -b)$. Under condition 2.2.1, if $g_1, ..., g_w$ are drawn such that they are from distinct elements G_i of the partition and not from G_1, the p-value $\mathbb{P}_{H_\mathrm{p}}(-B \geq -b)$ is exactly

$$\frac{b+1}{w+1}.$$

The validity of this formula follows from proposition 2.3.1. For the case that permutations are drawn with replacement, where $g_1, ..., g_w$ are independent

and uniform on G, Phipson and Smyth also provide a formula for $\mathbb{P}_{H_\mathrm{p}}(-B \geq -b)$, under condition 2.2.1.

The formula $(B+1)/(w+1)$ simplifies to the formula B/w if the identity map is added to the random permutations. It follows that the permutation test based on random permutations becomes exact for certain α if the identity is added. Note that this only holds if condition 2.2.1 is satisfied and all permutations are from distinct equivalence classes G_i.

We now state some additional results that follow from section 2.3.3. Corresponding to the randomized test of proposition 2.3.2, a randomized p-value can be defined as follows. The advantage of this p-value is that it is always uniform on $[0,1]$ under H_p without requirement of additional assumptions, and it is easy to compute. Consider the randomized test of proposition 2.3.2 (hence with G' as in definition 3.2.1). Suppose without loss of generality that when $T(X) = T^{(k')}$, the test rejects if and only if $a > u$, where u is uniform on $[0,1]$ and independent. Define the randomized p-value by

$$p' = \frac{\#\{1 \leq j \leq w : T(g_j X) > T(X)\}}{w} + u\frac{\#\{1 \leq j \leq w : T(g_j X) = T(X)\}}{w}.$$

This p-value has the property that $p' \leq \alpha$ if and only if the randomized test rejects. This implies in particular that p' is exactly uniform on $[0,1]$ under H_p. The fact that p' is randomized is in itself not objectionable, since it is randomized anyway due to the random transformations.

A simple upper bound to p' is

$$\frac{\#\{1 \leq j \leq w : T(g_j X) \geq T(X)\}}{w},$$

a p-value in the weak sense, which translates to $(B+1)/(w+1)$ when $g_1, ..., g_w$ are for example all independent uniform draws from G. It is not exactly uniform on $[0,1]$ under H_p. However, when w is large and there are few ties among the test statistics, it tends to closely approximate p', so that it may be used for simplicity.

2.4 Applications

We briefly mention some applications where our results are particularly useful. We have considered data X that lie in an arbitrary space $\mathcal{X}$ and

an arbitrary group of transformations G. For example, we allow X to be a vector of functions, which is the type of data investigated by functional data analysis (FDA) (Cuevas, 2014; Goia and Vieu, 2016). Cox and Lee (2008) consider permutation testing with such functional data. To formulate an exact random permutation test in such a setting, the present paper is useful.

In Hemerik and Goeman (2018), properties are proven of the popular method SAM ("Significance Analysis of Microarrays", Tusher et al., 2001). This is a permutation-based multiple testing method which provides an estimate of the false discovery proportion, the fraction of false positives among the rejected hypotheses. Using theorem 2.3.1, Hemerik and Goeman (2018) showed for the first time how a confidence interval can be constructed around this estimate.

In a basic permutation test, the observed statistic $T(X)$ is compared to $T^{(k)} \in \mathbb{R}$, a quantile of the permutation distribution. The permutation-based multiple testing method by Meinshausen (2006), which provides simultaneous confidence bounds for the false discovery proportion, also constructs a quantile based on the permutation distribution. There, however, $l \in \mathbb{N}$ hypotheses and hence l statistics $T_1(X), ..., T_l(X)$, are considered. (They consider p-values as test statistics.) Correspondingly, the quantile which Meinshausen constructs is l-dimensional. It turns out that the crucial step of the proof (the second last line of the proof, p. 231) relies on the principle behind the basic permutation test. The present article can be used to make this method exact. (For example, in Meinshausen (2006), id should be added to the random permutations.)

In Goeman and Solari (2011), it is suggested to combine the method by Meinshausen (2006) with closed testing, which leads to a very computationally intensive method. Hence preferably only a limited number of permutations (e.g. 100) would be used. The present paper allows using such a limited number of transformations, while still obtaining an exact method.

2.5 Discussion

This paper proves properties of tests with random permutations in a very general setting. Properties such as unbiasedness of tests of composite null hypotheses and consistency do not hold in general but may be proved for more specific scenarios. For fixed permutations, there are many results

regarding such properties (Hoeffding, 1952; Lehmann and Romano, 2005; Pesarin and Salmaso, 2010, 2013) which may be extended to random permutations.

Aside from the permutation test, there are many multiple testing methods which employ permutations, some of which are mentioned in section 2.4. Another example is the $\max T$ method by Westfall and Young (1993). These methods are precisely based on the principle behind the permutation test. This paper can provide better insight into these procedures, when random permutations are used.

3

False discovery proportion estimation by permutations: confidence for SAM

Abstract

SAM ("Significance Analysis of Microarrays") is a highly popular permutation-based multiple testing method that estimates the false discovery proportion (FDP), the fraction of false positives among all rejected hypotheses. Perhaps surprisingly, until now this method had no known properties. This paper extends SAM by providing $(1-\alpha)$-confidence upper bounds for the FDP, so that exact confidence statements can be made. As a special case, an estimate of the FDP is obtained that underestimates the FDP with probability at most 0.5. Moreover, using a closed testing procedure, this paper decreases the upper bounds and estimates in such a way

This chapter has been published as: Jesse Hemerik and Jelle Goeman (2017). False discovery proportion estimation by permutations: confidence for significance analysis of microarrays. *Journal of the Royal Statistical Society, series B (Statistical Methodology)* 80(1), 137-155

that the confidence level is maintained. We base our methods on a general result on exact testing with random permutations.

3.1 Introduction

When multiple hypotheses are tested, interest is often in estimating the false discovery proportion (FDP), the number of false positives divided by the total number of rejections. When there are no rejections, the FDP is defined to be zero. When there is unknown dependence in the data, a challenge is to find methods that are powerful but also require few assumptions on the dependence structure (van der Laan et al., 2004a; Meinshausen, 2006; Genovese and Wasserman, 2006). A highly popular method for estimation of the FDP is Significance Analysis of Microarrays (SAM) (Tusher et al., 2001). SAM is a very general method that is not at all limited to microarray data. It requires no parametric assumpions and almost no assumptions on the dependence structure in the data. Instead, it adapts to the dependence stucture by using permutations. Consequently Tusher et al. (2001) have been cited more than 10,000 times.

The SAM procedure in Tusher et al. is based on two ideas. The first is estimation of the FDP based on permutations of the data. The second idea is a specific choice of the test statistic, involving a small fudge factor. In this paper, focus is on the first idea. We do not consider a specific type of test statistics, but allow any test statistics to be used.

The rationale of SAM is the following. SAM rejects all hypotheses with test statistics lying in the user-defined rejection region. The number of false positives is estimated by considering permuted versions of the data. The median of the numbers of rejections for the permuted versions of the data is the estimate of the number of false positives. This value divided by the number of rejections is an estimate of the FDP. No properties of the estimate have been proven (although Dudoit et al. (2002, 2003) note that SAM estimates the Per-Family Error Rate). Until now SAM has been a very sensible, but quite heuristic method.

Based on Storey (2002; 2004) an adaptation of SAM was suggested to decrease the estimate. It is based on the idea that when there are relatively many false hypotheses, the original SAM method tends to overestimate the number of false positives. The reason is that the many false hypotheses cannot lead to false positives and SAM did not take this into account. Hence

it was suggested to multiply the basic SAM estimate by an estimate $\widehat{\pi}_0$ of the fraction of true hypotheses π_0. Usually this leads to a lower estimate of the FDP. This multiplication by the plug-in $\widehat{\pi}_0$ has been implemented in the *samr* *R*-package, the main software for SAM (Chu et al., 2001). Like the original SAM method, this newer procedure has no known properties.

The first aim of this paper is to construct a $(1-\alpha)$-confidence upper bound for the FDP for $\alpha \in [0, 1)$. That is, we extend SAM by providing a confidence interval around the estimate of the FDP. Thus, for small α, we obtain a high-confidence upper bound for the FDP. For $\alpha = 0.5$, we obtain an estimate of the FDP, which underestimates the FDP with probability at most 0.5. This estimate coincides with the estimate of the *samr* package without multiplication by $\widehat{\pi}_0$. Our $(1-\alpha)$-upper bound is the $(1-\alpha)$-quantile of the permutation distribution of the number of rejections. It was inspired by a work by Meinshausen (2006), who provides upper bounds for the FDP that are uniformly valid over multiple rejection regions.

The further contributions of this paper are procedures that decrease the basic $(1-\alpha)$-bound, in such a way that the exact properties are maintained. In particular, for $\alpha = 0.5$ the estimate is improved. These uniform improvements do not require additional assumptions. As with the plug-in method based on $\widehat{\pi}_0$, the gain is largest when there are relatively many false hypotheses. The improvements are derived using a result by Goeman and Solari (2011), who provide uniform FDP bounds by using a closed testing procedure. Our derivation reveals surprising connections between SAM and closed testing.

All our methods are based on a general result on exact testing with randomly sampled permutations, which extends the work of Phipson and Smyth (2010). This result also allows proving properties of other existing methods based on random permutations. All methods in this paper have been implemented in the *R* package *confSAM*.

This paper is built up as follows. In Section 3.2.1 the basic $(1-\alpha)$-bound for the FDP is discussed. A closed-testing based improvement of this method is presented in Section 3.3, including a fast approximation of this improvement. In Section 3.4 a conservative shortcut is constructed for the method from Section 3.3. The proposed methods are applied to simulated data in in Section 3.5. Section 3.6 contains an analysis of real data.

3.2 Basic upper bound

In this section the basic $(1-\alpha)$-bound for the FDP is discussed.

3.2.1 Setting and notation

Througout the paper, we consider the following setting. Let X be data, taking values in a sample space $\mathcal{X}$. Consider hypotheses $H_1, ..., H_m$ and test statistics $T_i : \mathcal{X} \to \mathbb{R}$, $1 \leq i \leq m$. For each $1 \leq i \leq m$, let $D_i \subseteq \mathbb{R}$ be a rejection region associated with hypothesis H_i and test statistic T_i. That is, H_i is rejected if and only if $T_i(X) \in D_i$, so that

$$\mathcal{R} = \{1 \leq i \leq m : T_i(X) \in D_i\}$$

is the set of indices of rejected hypotheses. We simply call $\mathcal{R}$ the set of rejected hypotheses. We write $\mathcal{R}^c = \{1, ..., m\} \setminus \mathcal{R}$. Let

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

be the set of true hypotheses. Let

$$N = \#\mathcal{N}, \quad R = \#\mathcal{R}$$

and

$$V = \#\mathcal{N} \cap \mathcal{R},$$

the number of false positives. Since sets and numbers such as $\mathcal{R}$, R and V depend on the data, we denote them as functions on $\mathcal{X}$. For example, for $x \in \mathcal{X}$, $\mathcal{R}(x)$ denotes the set of rejected hypotheses for data x. The set $\mathcal{N}$ does not depend on the data, since the hypotheses are fixed. Thus $V(x) = \#(\mathcal{R}(x) \cap \mathcal{N})$. When no argument is written, this means that the argument is X. For example, R is short for $R(X)$. The false discovery proportion is

$$FDP = \frac{V}{R}$$

if $R > 0$ and 0 otherwise.

All methods in this paper are based on permutations or other transformations of the data. Let G be a finite set of transformations $g : \mathcal{X} \to \mathcal{X}$, such that G is a group under composition of maps. Throughout the paper we use the word 'group' in the strict algebraic sense, rather than loosely

in the meaning of 'set' as is often done in the statistical literature. We write $g(x)$ as gx. In practice G is often a group of permutation maps. Sometimes other groups of transformations will be used, such as rotations (Langsrud, 2005; Solari et al., 2014) and multiplication of part of the data by -1 (Pesarin and Salmaso (2010), pp. 54 and 168).

The following assumption, made throughout this paper, underlies many permutation-based multiple testing methods, e.g. Westfall and Young's max$T$ method (1993), Meinshausen and Bühlmann (2005) and Meinshausen (2006).

Assumption 3.2.1. The joint distribution of the test statistics $T_i(gX)$ with $i \in \mathcal{N}$, $g \in G$, is invariant under all transformations in G of X.

In applications, an argument needs to be given for this distributional assumption. As a mathematical example where the assumption is satisfied, consider a basic randomized trial where H_i implies that the distribution of the expression level of gene i is the same for cases and controls. Typically each T_i only depends on the expression levels measured for gene i. Then Assumption 3.2.1 is satisfied if the multivariate distribution of the expression levels corresponding to $\mathcal{N}$ is the same for cases and controls.

It is allowed to define $\mathcal{N}$ simply as the largest set of hypotheses for which Assumption 3.2.1 is satisfied, as in Meinshausen and Bühlmann (2005). This is a less usual definition of $\mathcal{N}$, but Assumption 3.2.1 is then guaranteed to hold.

Throughout the paper, random transformations from G are used. The vector of random transformations is defined as follows.

Definition 3.2.1. Let G' be the vector $(id, g_2, ..., g_w)$, where id is the identity in G and $g_2, ..., g_w$ are random elements from G. Write $g_1 = id$. The random transformations can be drawn either with or without replacement: the statements in this paper hold for both cases. In the latter case, $w \leq \#G$. If we draw $g_2, ...g_w$ without replacement, then we take them to be uniformly distributed on $G \setminus \{id\}$, otherwise uniform on G.

For $\mathcal{I} \subseteq \{1, ..., m\}$ and $x \in \mathcal{X}$, write

$$R_{\mathcal{I}}(x) = \#\mathcal{R}(x) \cap \mathcal{I}.$$

Let

$$R_{\mathcal{I}}^{(1)} \leq ... \leq R_{\mathcal{I}}^{(w)}$$

be the sorted values $R_{\mathcal{I}}(g_j X)$, $1 \leq j \leq w$. We have $R_{\{1,\ldots,m\}} = R$, so write $R^{(j)} := R^{(j)}_{\{1,\ldots,m\}}$, $1 \leq j \leq w$.

Throughout the paper, $\alpha \in [0,1)$ and $k = \lceil (1-\alpha)w \rceil$, the smallest integer at least as large as $(1-\alpha)w$. The minimum of two numbers a and b is denoted by $a \wedge b$.

3.2.2 Upper bound and median unbiased estimate

Here the upper bound and estimate for the FDP are constructed. We first prove the permutation principle that underlies our methods. It is known that the permutation test is exact when the set of transformations (e.g. permutations) has a group structure (Hoeffding, 1952). For example, the set of all possible permutation maps is a group. In recent decades, permutation methods have become popular. Often random permutations are used, to limit the computation time. Usually a p-value based on random permutations is seen as an estimate of the true permutation p-value. However, it is also possible to compute an exact p-value based on random permutations, if they are suitably sampled from a group. Phipson and Smyth (2010) provide formulas for exact p-values based on random permutations under some assumptions. Their results imply that to obtain a valid test, the original observation should be included with the random permutations. However, they ignore the role of the group structure, which is fundamental to permutation methods. Moreover, it has not been clear how results on testing with random permutations generalise to other permutation methods (such as SAM and Meinshausen, 2006). We now state a general result on testing with random transformations. This result can be used to prove properties of various permutation-based multiple testing methods. We will illustrate this in Theorem 3.2.2, where we apply Theorem 3.2.1 in the SAM context.

Theorem 3.2.1. *Let $S : \mathcal{X} \to \mathbb{R}$ be a test statistic. Let $S^{(1)}(X, G') \leq \ldots \leq S^{(w)}(X, G')$ be the ordered test statistics $S(g_j X)$, $1 \leq j \leq w$.*

Consider a null hypothesis H_0 which implies that the joint distribution of the test statistics $S(gX)$, $g \in G$, is invariant under all transformations in G of X. Then under H_0, $\mathbb{P}(S(X, G') > S^{(k)}(X, G')) \leq \alpha$.

Proof. From the group structure of G, it follows that for all $1 \leq j \leq w$, $G' g_j^{-1}$ and G' have the same distribution, if we disregard the order of the

elements. Let j have the uniform distribution on $\{1, ..., w\}$ and write $h = g_j$. Under H_0,

$$\begin{aligned}&\mathbb{P}\big\{S(X) > S^{(k)}(X, G')\big\} = \\ &\mathbb{P}\big\{S(X) > S^{(k)}(X, G'h^{-1})\big\} = \\ &\mathbb{P}\big\{S(hX) > S^{(k)}(hX, G'h^{-1})\big\}.\end{aligned}$$

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

$$\begin{aligned}&\mathbb{P}\big\{S(hX) > S^{(k)}(h^{-1}hX, G')\big\} = \\ &\mathbb{P}\big\{S(hX) > S^{(k)}(X, G')\big\}\end{aligned}$$

Since $h = g_j$ with j uniform, this equals

$$\mathbb{E}\Big[w^{-1}\#\big\{1 \le j \le w : S^{(j)}(X, G') > S^{(k)}(X, G')\big\}\Big] \le \alpha,$$

as was to be shown. □

The value $R^{(k)}$ is the $(1-\alpha)$-quantile of the numbers of rejections for the permuted versions of the data. The following theorem states that this simple quantile is a $(1-\alpha)$-upper bound for the number of false positives V.

Theorem 3.2.2. *The number $\overline{V} := R^{(k)} \wedge R$ is a $(1-\alpha)$-upper bound for V, i.e.*

$$\mathbb{P}\big(V \le \overline{V}\big) \ge 1 - \alpha.$$

Proof. Let

$$V^{(1)} \le ... \le V^{(w)}$$

be the sorted values $V(g_jX) = \#\big(\mathcal{R}(g_jX) \cap \mathcal{N}\big)$, $1 \le j \le w$. With Theorem 3.2.1 it follows that

$$\mathbb{P}\big(V > V^{(k)}\big) \le \alpha.$$

Since $V^{(k)} \le R^{(k)}$,

$$\mathbb{P}\big(V \le R^{(k)}\big) \ge 1 - \alpha$$

and the result follows. □

Note that $V \leq \overline{V}$ holds if and only if $V/R \leq \overline{V}/R$, provided $R > 0$. Thus $\overline{V}/R$, which is interpreted as 0 when $R = 0$, is a $(1-\alpha)$-upper bound for the FDP. Note that taking $\alpha = 0.5$ in the above theorem provides an estimate $\overline{V}$ of V with the property that $\mathbb{P}\big(V \leq \overline{V}\big) \geq 0.5$. We will call such an estimate *median unbiased*, in line with the existing notion of a median unbiased estimator of a parameter.

By assumption, the dependence structure of the test statistics $T_i(X)$, $i \in \mathcal{N}$, is maintained by their permutation distribution. The quantile $\overline{V}$ is based on the permutation distribution of the number of rejections R, which is based on the permutation distribution of the test statistics. Hence $\overline{V}$ is adapted to the joint distribution of the test statistics $T_i(X)$, $i \in \mathcal{N}$. Therefore the method does not need to take into account a worst-case scenario for their dependence structure. Thus, by using permutations, relatively tight bounds for the FDP tend to be obtained.

3.2.3 Choice of rejection regions

The rejection regions D_i can be freely chosen, provided that they are not based on the data. When they are based on the data, this may introduce some selection bias, especially when the regions are cherry-picked in such a way that the number of rejections is large compared the estimate of V.

When the rejection regions D_i do not depend on the data, it is not always possible to choose sensible rejection regions when little is known about the distribution of the test statistics. We can use permutation p-values as test statistics however, such that we can always choose a sensible rejection region, for instance $(0, 0.01]$. This leads to about $0.01N$ false positives on average. In the setting of SAM, such p-values based on permutations (or other transformations) can nearly always be calculated. These p-values can be based on random permutations, as in Theorem 3.2.1. Moreover, it is allowed to base all m p-values on a single collection of random permutations (independent from $g_1, ..., g_w$), since the test statistics are essentially assumption-free. Note that permutation p-values are never smaller than one divided by the number of permutations. Thus, when the rejection region is $(0, c)$, more than c^{-1} permutations should be used.

When the cutoff c is very small, the number of random permutations needed is very large, which may make using permutation p-values time-consuming. A possible practical solution is the following. Often the tail of the permutation distribution of each T_i can be modeled by e.g. a general-

ized Pareto distribution (Knijnenburg et al., 2009). In that case, draw a small number of random permutations, compute the corresponding values of the test statistic T_i and fit such a distribution to these values. Then use $D_i = (q_i, \infty)$ as the rejection region for T_i, where q_i is the $(1-c)$-quantile of the distribution determined for T_i. Note that this means that the rejection regions are data-dependent. However, since they depend on permutations of the data and not on cherry-picking the regions that give the strongest results, the selection bias tends to be very limited or absent. Since this paper focuses on proving exact properties however, we will keep the assumption that the rejection regions D_i are fully independent.

3.3 Closed testing for improved bounds

Especially when there are many false hypotheses, the basic bound $\overline{V}$ does not exhaust α. The reason is that $\overline{V}$ then tends to be substantially larger than the bound $V^{(k)}$, as can be seen from their definitions. The bound $V^{(k)}$ cannot be computed in practice but has been shown to be a $(1-\alpha)$-upper bound in the proof of Theorem 3.2.2. The closed testing principle (Marcus et al., 1976) is a powerful method for familywise error rate control. Goeman and Solari (2011) show how closed testing can be used to obtain upper bounds for the FDP. By relating SAM to closed testing, we will be able to derive a potentially smaller upper bound for V than the basic bound $\overline{V}$. The improved bound is still valid with probability $1-\alpha$.

3.3.1 Closed testing

We recall the closed testing principle and how it can be used to obtain uniform upper bounds for the FDP. For each nonempty $\mathcal{I} \subseteq \{1, ..., m\}$, denote by $H_\mathcal{I}$ the intersection hypothesis $\bigcap_{i\in\mathcal{I}} H_i$, the hypotheses that all hypotheses H_i, $i \in I$, are true. Suppose that for each nonempty $\mathcal{I} \subseteq \{1, ..., m\}$ an α-level test for $H_\mathcal{I}$ is defined. These $2^m - 1$ tests are called local tests. The closed testing procedure rejects all $H_\mathcal{I}$ for which all $H_\mathcal{J}$ with $\mathcal{I} \subseteq \mathcal{J} \subseteq \{1, ..., m\}$ are rejected by their local tests. By Marcus et al. (1976), the probability that the closed testing procedure rejects at least one true intersection hypothesis is at most α. Thus the procedure strongly controls the familywise error rate at level α.

The FDP upper bounds are derived as follows. Let

$$\mathcal{C} = \{\mathcal{I} \subseteq \{1, ..., m\} : \quad H_{\mathcal{I}} \text{ is rejected by the closed testing procedure}\}.$$

For each $\mathcal{K} \subseteq \{1, ..., m\}$ define

$$\overline{V}_{\text{ct}}(\mathcal{K}) = \max\{\#\mathcal{I} : \ \mathcal{I} \subseteq \mathcal{K}, \ \mathcal{I} \notin \mathcal{C}\},$$

where the maximum is defined to be zero if the set is empty. By Goeman and Solari (2011) the following holds.

Theorem 3.3.1. *Uniformly over all $\mathcal{K} \subseteq \{1, ..., m\}$, $\overline{V}_{ct}(\mathcal{K})$ is a $(1-\alpha)$-upper bound for $\#\mathcal{K} \cap \mathcal{N}$, i.e.*

$$P\left[\bigcap_{\mathcal{K} \subseteq \{1,...,m\}} \left\{\#\mathcal{K} \cap \mathcal{N} \leq \overline{V}_{ct}(\mathcal{K})\right\}\right] \geq 1 - \alpha.$$

The proof is in Goeman and Solari (2011). Note that $\#\mathcal{K} \cap \mathcal{N}$ is the number of false positives if $\mathcal{K}$ is the rejected set. Thus the theorem provides bounds for the numbers of false positives that are uniform over all possible rejected sets. Thus, if the rejected set is chosen based on the data, then the corresponding upper bound is still valid with probability at least $1 - \alpha$.

A closed testing procedure depends on its local tests. For different local tests, different closed testing procedures are obtained. The more power the closed testing procedure has, the lower the resulting FDP bound tends to be. One particular closed testing procedure leads directly to the basic bound $\overline{V}$. The reader can check that this is the closed testing procedure based on the local tests that reject $H_{\mathcal{I}}$ if and only if $R_{\mathcal{I}} > R^{(k)}$. In Section 3.3.2 a more powerful closed testing procedure is considered, which allows improvement of the basic bound $\overline{V}$.

3.3.2 Improved FDP bounds

To obtain an improvement of the bound $\overline{V}$ for the number of false positives, we use a more sophisticated closed testing procedure. For each nonempty $\mathcal{I} \subseteq \{1, ..., m\}$ consider the local test that rejects $H_{\mathcal{I}}$ if and only if $R_{\mathcal{I}} > R_{\mathcal{I}}^{(k)}$. (See the notation defined in Section 3.2.1.) This test has level at most α by theorem 3.2.1. Throughout the rest of this paper, let $\overline{V}_{\text{ct}}(\mathcal{K})$ refer to the closed testing procedure based on these local tests. Note that $\overline{V}_{\text{ct}}(\mathcal{R})$

provides an upper bound for $V = \#\mathcal{R} \cap \mathcal{N}$, the number of true hypotheses in $\mathcal{R}$. We write $\overline{V}_{\mathrm{ct}} := \overline{V}_{\mathrm{ct}}(\mathcal{R})$.

The bound $\overline{V}_{\mathrm{ct}}(\mathcal{R})$ is ideal in the sense that no smaller $(1-\alpha)$-bound for V is given in this paper or elsewhere in the literature, under our assumptions. In practice however, it is often computationally infeasible to calculate this bound without the use of shortcuts. Indeed, when naively computing $\overline{V}_{\mathrm{ct}}$, a huge number of local tests needs to performed unless m is small. This section is devoted to deriving an exact shortcut for calculating $\overline{V}_{\mathrm{ct}}$. In Section 3.4 a conservative shortcut will be derived, i.e. an efficient method for finding an upper bound for $\overline{V}_{\mathrm{ct}}$.

The following lemma offers a shortcut for determining whether $H_{\mathcal{I}}$ is rejected by our closed testing procedure.

Lemma 3.3.1. *For $\mathcal{I} \subseteq \{1, ..., m\}$, $\mathcal{I} \in \mathcal{C}$ if and only if $R_{\mathcal{I}} > R^{(k)}_{\mathcal{I} \cup \mathcal{R}^c}$.*

Proof. To prove the first implication, note that

$$\begin{aligned}
&\mathcal{I} \in \mathcal{C} \Leftrightarrow \\
&\text{For all } \mathcal{I} \subseteq \mathcal{J} \subseteq \{1, ..., m\},\ R_{\mathcal{J}} > R^{(k)}_{\mathcal{J}} \Rightarrow \\
&R_{\mathcal{I} \cup \mathcal{R}^c} > R^{(k)}_{\mathcal{I} \cup \mathcal{R}^c} \Leftrightarrow \\
&R_{\mathcal{I}} > R^{(k)}_{\mathcal{I} \cup \mathcal{R}^c}.
\end{aligned}$$

For the reverse implication, suppose

$$\#\mathcal{I} \cap \mathcal{R} > R^{(k)}_{\mathcal{I} \cup \mathcal{R}^c} \tag{3.1}$$

and let $\mathcal{I} \subseteq \mathcal{J} \subseteq \{1, ..., m\}$. Then

$$R_{\mathcal{J}} = \#\mathcal{I} \cap \mathcal{R} + \#(\mathcal{J} \setminus \mathcal{I}) \cap \mathcal{R} \tag{3.2}$$

and, because obviously $\#\mathcal{A} \geq R^{(k)}_{\mathcal{A} \cup \mathcal{B}} - R^{(k)}_{\mathcal{B}}$ for $\mathcal{A}, \mathcal{B} \subseteq \{1, ..., m\}$,

$$\#(\mathcal{J} \setminus \mathcal{I}) \cap \mathcal{R} \geq R^{(k)}_{((\mathcal{J} \setminus \mathcal{I}) \cap \mathcal{R}) \cup (\mathcal{I} \cup \mathcal{R}^c)} - R^{(k)}_{\mathcal{I} \cup \mathcal{R}^c} \geq R^{(k)}_{\mathcal{J}} - R^{(k)}_{\mathcal{I} \cup \mathcal{R}^c}. \tag{3.3}$$

Combining (3.1), (3.2) and (3.3) yields

$$R_{\mathcal{J}} > R^{(k)}_{\mathcal{I} \cup \mathcal{R}^c} + R^{(k)}_{\mathcal{J}} - R^{(k)}_{\mathcal{I} \cup \mathcal{R}^c} = R^{(k)}_{\mathcal{J}}.$$

Thus all $H_{\mathcal{J}}$ with $\mathcal{I} \subseteq \mathcal{J} \subseteq \{1, ..., m\}$ are rejected by their local tests, so that $\mathcal{I} \in \mathcal{C}$. □

Due to this shortcut, te determine whether $\mathcal{I} \in \mathcal{C}$ it is not necessary to perform all local tests for the hypotheses $H_{\mathcal{J}}$ with $\mathcal{I} \subseteq \mathcal{J}$. Instead, it suffices to check if $R_{\mathcal{I}} > R^{(k)}_{\mathcal{I}\cup\mathcal{R}^c}$.

Using this fact and additional observations, the following exact shortcut is obtained for determining $\overline{V}_{\mathrm{ct}}$.

Proposition 3.3.1. *The bound* $\overline{V}_{ct}$ *equals*

$$R \wedge \left[\min \left\{ 1 \leq M \leq R : \text{ for all } \mathcal{I} \subseteq \mathcal{R} \text{ with } \#\mathcal{I} = M,\ M > R^{(k)}_{\mathcal{I}\cup\mathcal{R}^c} \right\} - 1 \right]. \tag{3.4}$$

Proof. By Lemma 3.3.1, $\overline{V}_{\mathrm{ct}}(\mathcal{R}) =$

$$\max\{\#\mathcal{I} : \mathcal{I} \subseteq \mathcal{R}, R_{\mathcal{I}} \leq R^{(k)}_{\mathcal{I}\cup\mathcal{R}^c}\} =$$
$$\max\{\#\mathcal{I} : \mathcal{I} \subseteq \mathcal{R}, \#\mathcal{I} \leq R^{(k)}_{\mathcal{I}\cup\mathcal{R}^c}\} =$$
$$R \wedge \left[\min \left\{ 1 \leq M \leq R : \text{ for all } \mathcal{I} \subseteq \mathcal{R} \text{ with } \#\mathcal{I} \geq M,\ \#\mathcal{I} > R^{(k)}_{\mathcal{I}\cup\mathcal{R}^c} \right\} - 1 \right].$$

For any $\mathcal{I} \subseteq \mathcal{R}$, if $\#\mathcal{I} > R^{(k)}_{\mathcal{I}\cup\mathcal{R}^c}$ then for all $\mathcal{I} \subseteq \mathcal{J} \subseteq \mathcal{R}$, $\#\mathcal{J} > R^{(k)}_{\mathcal{J}\cup\mathcal{R}^c}$. Hence the above equals (3.4). □

Using Proposition 3.3.1, $\overline{V}_{\mathrm{ct}}$ can be calculated much faster than by naive computation based on the definition of $\overline{V}_{\mathrm{ct}}$. When R or $\overline{V}_{\mathrm{ct}}$ is large however, calculating $\overline{V}_{\mathrm{ct}}$ is often still infeasible; the computation time is roughly proportional to $\binom{R}{\overline{V}_{\mathrm{ct}}+1}$ if $\overline{V}_{\mathrm{ct}} < R$. Hence in Section 3.3.3 a method for approximating $\overline{V}_{\mathrm{ct}}$ is defined. Moreover, in Section 3.4 a conservative shortcut is derived, which calculates an upper bound to $\overline{V}_{\mathrm{ct}}$ in relatively little time. The performance of these methods is illustrated with simulations in Sections 3.5.5 and 3.5.6.

3.3.3 Approximation method

We propose a method for approximating the bound $\overline{V}_{\mathrm{ct}}$, for cases where computing $\overline{V}_{\mathrm{ct}}$ is infeasible. Proposition 3.3.1 states that

$$\overline{V}_{\mathrm{ct}} = R \wedge \left[\min \left\{ 1 \leq M \leq R : \ M > \mu(M) \right\} - 1 \right], \tag{3.5}$$

where

$$\mu(M) := \max\{R^{(k)}_{\mathcal{I}\cup\mathcal{R}^c} : \quad \mathcal{I} \subseteq \mathcal{R} \text{ and } \#\mathcal{I} = M\}.$$

Determining $\mu(M)$ can be computationally infeasible, since the number of subsets $\mathcal{I} \subseteq \mathcal{R}$ with $\#\mathcal{I} = M$ can be huge. Hence we propose to draw a big number of random sets $\mathcal{I} \subseteq \mathcal{R}$ with $\#\mathcal{I} = M$ and calculate the maximum for this collection of subsets. That is, we consider an approximation

$$\mu^*(M) = \max\{R^{(k)}_{\mathcal{I}\cup\mathcal{R}^c} : \quad \mathcal{I} \in \mathcal{S}\},$$

where $\mathcal{S}$ is some large random subcollection of $\{\mathcal{I} \subseteq \mathcal{R} : \ \#\mathcal{I} = M\}$. This leads to the estimate

$$\overline{V}^*_{\mathrm{ct}} = R \wedge \left[\min\left\{1 \le M \le R : \ M > \mu^*(M)\right\} - 1\right]. \tag{3.6}$$

Note that $\mu^*(M) \le \mu(M)$ and consequently $\overline{V}^*_{\mathrm{ct}} \le \overline{V}_{\mathrm{ct}}$. Hence $\overline{V}^*_{\mathrm{ct}}$ is not guaranteed to be a $(1-\alpha)$-confidence upper bound. However, in many cases, including the simulation scenarios considered in Section 3.5.6 , $\overline{V}^*_{\mathrm{ct}}$ is still a $(1-\alpha)$-confidence bound for V. The reason is that $\overline{V}^*_{\mathrm{ct}}$ converges to $\overline{V}_{\mathrm{ct}}$ for $\#\mathcal{S} \to \infty$. For details see Section 3.5.6.

3.4 Conservative shortcut

Here we will construct a conservative shortcut for the closed testing-based method that provides $\overline{V}_{\mathrm{ct}}$. The shortcut is much more computationally efficient and can often be used when there are thousands of rejections. The upper bound $\overline{V}_{\mathrm{sc}}$ that the shortcut provides is always less than or equal to the basic bound $\overline{V}$. On the other hand, it only improves $\overline{V}$ in specific settings, some of which are considered in Section 3.5.5. The bound $\overline{V}_{\mathrm{sc}}$ is often larger than $\overline{V}_{\mathrm{ct}}$. It is never smaller than $\overline{V}_{\mathrm{ct}}$, which guarantees its validity as a $(1-\alpha)$-confidence bound. Thus the following ordering holds:

$$\overline{V} \ge \overline{V}_{\mathrm{sc}} \ge \overline{V}_{\mathrm{ct}} \ge \overline{V}^*_{\mathrm{ct}}.$$

The bound $\overline{V}_{\mathrm{ct}}$ depends on $\mu(M)$, the computation of which can be computationally infeasible. Lemma 3.4.1 provides an upper bound $U(M)$ for this maximum which can often be computed in a limited amount of time even when there are thousands of rejections. This will lead to the conservative shortcut for the full closed testing-based method.

Note that the upper bounds $\overline{V}$ and $\overline{V}_{\mathrm{ct}}$ are functions of $\mathcal{R}(g_1X), ..., \mathcal{R}(g_wX)$, the rejected sets for the transformed versions of the data. Reindex the transformations such that $R(g_1X) \leq ... \leq R(g_wX)$. Hence $R(g_jX) = R^{(j)}$ for all $1 \leq j \leq w$. The collection of rejected sets can be represented by a binary $m \times w$-matrix $\mathbf{M}$, where

$$\mathbf{M}_{(i,j)} = \begin{cases} 1 & \text{if } i \in \mathcal{R}(g_jX), \\ 0 & \text{otherwise.} \end{cases}$$

Since the upper bound $\overline{V}_{\mathrm{ct}}$ can be viewed as a function of this matrix, the problem of finding shortcuts for the closed testing-based method is essentially a combinatorial one.

To have an intuitive understanding of Lemma 3.4.1 below, it is useful to view the quantities considered in it as functions of the matrix $\mathbf{M}$. For each $1 \leq j \leq w$, we define

$$S_j = R_{\mathcal{R}}(g_jX).$$

Note that this is the sum of the elements of $\mathbf{M}$ that are both in the j-th column and in the rows corresponding to the rejected set $\mathcal{R} = \mathcal{R}(X)$. Next, define for each $i \in \mathcal{R}$

$$\Sigma_i = \sum_{1 \leq j \leq w} R_{\{i\}}(g_jX).$$

This is simply the sum of the elements of the i-th row of $\mathbf{M}$. Let $\Sigma_{(1)} \leq ... \leq \Sigma_{(R)}$ be the sorted values Σ_i and for $1 \leq M \leq R$ let

$$\Sigma = \Sigma(M) := \Sigma_{(1)} + ... + \Sigma_{(R-M)}.$$

We now state the main result on which the conservative shortcut is based.

Lemma 3.4.1. *For each $s \in \mathbb{N}$ define*

$$N_s = \#\{1 \leq j < k : \ R^{(j)} < R^{(k)} - s\},$$
$$M_s = k - 1 - N_s.$$

For each $s \in \mathbb{N}$ and $N_s < j \leq w$, let

$$K_j^s = \max\left\{0, S_j - (R^{(j)} - R^{(k)} + s)\right\}$$

and let $K_{(1)}^s \geq ... \geq K_{(w-N_s)}^s$ be these values sorted from large to small.

For $1 \leq M \leq R$ *let*

$$U(M) = R^{(k)} - 1 - \max A,$$

where $A =$

$$\left\{ s \in \mathbb{N} : \Sigma(M) > \sum_{1 \leq j \leq N_s} S_j + \sum_{N_s < j \leq w} \min\{S_j, R^{(j)} - R^{(k)} + s\} + \sum_{1 \leq j \leq M_s} K^s_{(j)} \right\}.$$

Then

$$U(M) \geq \mu(M).$$

Proof. Let $1 \leq M \leq R$. Write

$$\mathcal{B} = \left\{ \mathcal{R}^{\mathsf{c}} \subseteq \mathcal{I} \subseteq \{1, ..., m\} : \ \#\mathcal{I} = \#\mathcal{R}^{\mathsf{c}} + M \right\}.$$

Note that

$$\mu(M) = \max\{R^{(k)}_{\mathcal{I}} : \mathcal{I} \in \mathcal{B}\}.$$

The proof consists of three parts. In part 1 we show that $0 \in A$ implies $R^{(k)} > \mu(M)$. In part 2 we note that for all $s \in \mathbb{N}$, $s \in A$ implies $R^{(k)} - s > \mu(M)$. We then conclude that by definition $U(M) \geq \mu(M)$.
–*Part 1.*
Suppose $0 \in A$. Consider any $\mathcal{I} \in \mathcal{B}$. Write $\mathcal{I}^{\mathsf{c}} = \{1, ..., m\} \setminus \mathcal{I}$. Note that $\#\mathcal{I}^{\mathsf{c}} = R - M$. Hence, by choice of Σ,

$$\sum_{1 \leq j \leq w} R_{\mathcal{I}^{\mathsf{c}}}(g_j X) \geq \Sigma. \tag{3.7}$$

Since $0 \in A$,

$$\sum_{1 \leq j \leq w} R_{\mathcal{I}^{\mathsf{c}}}(g_j X) > \sum_{1 \leq j \leq N_0} S_j + \sum_{N_0 < j \leq w} \min\{S_j, R^{(j)} - R^{(k)}\} + \sum_{1 \leq j \leq M_0} K^0_{(j)}. \tag{3.8}$$

First note that

$$\sum_{1 \leq j \leq N_0} R_{\mathcal{I}^{\mathsf{c}}}(g_j X) \leq \sum_{1 \leq j \leq N_0} S_j,$$

since $R_{\mathcal{I}^{\mathsf{c}}}(g_j X) \leq S_j$ for all j. Hence with (3.8) it follows that

$$\sum_{N_0 < j \leq w} R_{\mathcal{I}^{\mathsf{c}}}(g_j X) > \sum_{N_0 < j \leq w} \min\{S_j, R^{(j)} - R^{(k)}\} + \sum_{1 \leq j \leq M_0} K^0_{(j)}. \tag{3.9}$$

Suppose that $R_{\mathcal{I}}^{(k)} = R^{(k)}$. This implies that

$$\#\big\{N_0 < j \le w : R_{\mathcal{I}}(g_j X) < R^{(k)}\big\} \le M_0, \tag{3.10}$$

and equivalently

$$\#\big\{N_0 < j \le w : R_{\mathcal{I}}(g_j X) \ge R^{(k)}\big\} > (w - N_0) - M_0. \tag{3.11}$$

For the indices j in the set at (3.10),

$$R_{\mathcal{I}^c}(g_j X) \le S_j.$$

Moreover, for the indices j in the set at (3.11),

$$R_{\mathcal{I}^c}(g_j X) = R^{(j)} - R_{\mathcal{I}}(g_j X) \le \min\{S_j, R^{(j)} - R^{(k)}\}.$$

These observations imply that $\sum_{N_0<j\le w} R_{\mathcal{I}^c}(g_j X)$ is at most the right side of (3.9), which contradicts (3.9). Hence $R^{(k)} \ne R_{\mathcal{I}}^{(k)}$, i.e. $R^{(k)} > R_{\mathcal{I}}^{(k)}$. Since this holds for all $\mathcal{I} \in \mathcal{B}$, by definition $R^{(k)} > \mu(M)$.
–Part 2.
Consider any $\mathcal{I} \in B$. Let $s \in \mathbb{N}$. In part 1 we supposed that $0 \in A$; we now more generally suppose that $s \in A$. It follows like in part 1 that $R^{(k)} - s > R_{\mathcal{I}}^{(k)}$ and consequently $R^{(k)} - s > \mu(M)$.
–Part 3.
Since this holds for all $s \in A$, we have $R^{(k)} - \max A > \mu(M)$, i.e.

$$R^{(k)} - 1 - \max A = U(M) \ge \mu(M).$$

□

By (3.5), Lemma 3.4.1 and the fact that $\overline{V}_{\mathrm{ct}} \le \overline{V}$,

$$\overline{V}_{\mathrm{sc}} := \overline{V} \wedge \left[\min\Big\{1 \le M \le R : \ M > U(M)\Big\} - 1 \right]$$

is an upper bound for $\overline{V}_{\mathrm{ct}}$. Recall that the calculation of $\overline{V}_{\mathrm{sc}}$ is usually feasible when there are many thousands of rejections, but this shortcut only provides an improvement over $\overline{V}$ in some situations with many false hypotheses, as is illustrated with simulations in Section 3.5.5.

3.5 Simulations

We investigate the performance of the discussed methods on simple simulated data. In sections 3.5.2 and 3.5.3 variants of the basic SAM method are investigated as upper bounds (for $\alpha = 0.05$) and as estimates (for $\alpha = 0.5$). Some of the variants considered are based on plug-in estimates of the fraction of true hypotheses π_0 as in the *samr* package. In section 3.5.4 the closed testing-based bound $\overline{V}_{\text{ct}}$ is compared to the basic bound $\overline{V}$. The performance of the shortcut is illustrated in Section 3.5.5. All simulations were performed with *R*.

3.5.1 Simulated data and tests used

Here we describe the simulated data and tests used for all simulations. The simulated data matrix was the $2n \times m$-matrix

$$\mathbf{X} = \mathbf{X}' + \mathbf{Z}.$$

It can be seen as representing m gene expression levels of $2n$ persons. Here $\mathbf{X}'$ is a $2n \times m$-matrix of independent normally distributed variables with variance 1. For some $0 \leq F \leq m$, in the first F columns of $\mathbf{X}$ the first n entries had mean 1 and all other entries had mean 0. The matrix $\mathbf{Z}$ was used to make the entries in each row of $\mathbf{X}$ correlated. It is defined by $\mathbf{Z}_{ji} := s_i Z_j$, where $s_i = 1 - 2\mathbb{1}_{\{i>m/2\}}$ and each Z_j is independent and normally distributed with mean 0 and standard deviation σ_Z. For $1 \leq j \leq 2n$ and $1 \leq i < i' \leq m$ we have $Cov(\mathbf{X}_{ji}, \mathbf{X}_{ji'}) = E(\pm Z_j^2) = \pm {\sigma_Z}^2$, hence the correlation is

$$\rho(\mathbf{X}_{ji}, \mathbf{X}_{ji'}) = \frac{\pm {\sigma_Z}^2}{1 + {\sigma_Z}^2}.$$

For each $1 \leq i \leq m$, let H_i be the null hypothesis that $\mathbf{X}_{1,i}..., \mathbf{X}_{2n,i}$ are independent and standard normally distributed. Thus the first F hypotheses were false, such that the fraction of true hypotheses was $\pi_0 = (m - F)/m$.

Under H_i, the test statistic

$$T_i := \sum_{j=1}^{n} \mathbf{X}_{j,i} - \sum_{j=n+1}^{2n} \mathbf{X}_{j,i}$$

is normally distributed with variance $2n \cdot (1+{\sigma_Z}^2)$, so that we can efficiently calculate the corresponding two-sided p-value, i.e. the probability under H_i of a larger value of $|T_i|$ than observed. The test statistics used in the simulations were these p-values. For each hypothesis we used the same rejection region D of the form $(0, c)$, where $c \in (0, 1)$ is some cutoff.

As the group of transformations we used the $(2n)!$ maps that shuffle the rows of $\mathbf{X}$, leaving each individual row intact. These can e.g. be seen as permutations of cases and controls. In all the simulations except those of Section 3.5.5 we used $w = 100$, i.e. each time we drew 99 random permutations and added the identity. For larger w similar results are obtained (see also Marriott, 1979). The values of m, π_0, the cutoff c, α and $|\rho|$ are specified per case below.

3.5.2 Performance of variants of SAM as bounds

For $m = 1000$, $\alpha = 0.05$ and rejection region $D = (0, 0.01)$, we investigate the performance of variants of SAM as $(1-\alpha)100\%$-confidence upper bounds of the FDP. Some of the variants of SAM discussed here are based on an estimate $\widehat{\pi}_0$ of $\pi_0 = N/m$. Like the *samr* package, we calculated $\widehat{\pi}_0$ as

$$\frac{\#\{1 \leq i \leq m : P_i > q_i\}}{0.5 \cdot m},$$

where q_i is the 0.5-quantile of the permutation distribution of P_i. We write $\overline{FDP} := R^{(k)}/R$, where $\overline{FDP} = 0$ for $R = 0$. Note that $\overline{FDP}$ is potentially larger than 1. We also write $\widehat{\pi}_0' = \widehat{\pi}_0 \wedge 1$ and $\overline{FDP}' = \overline{FDP} \wedge 1$.

Table 3.1 shows 95%-confidence intervals for the probabilities that the bounds were smaller than the real FDP, for different values of π_0 and $|\rho|$. The value $|\rho| = 0.5$ corresponds to $\sigma_Z = 1$. From Table 3.1 it can be seen that $\overline{FDP}'$ is the only bound with the desired property $\mathbb{P}(\text{upper bound} < FDP) \leq \alpha$. For the other bounds, this probability is much larger than α for many settings (see also Korn et al., 2007), especially under dependence. This is related to the known fact that the estimate $\widehat{\pi}_0$ often has low accuracy under dependence (Qiu et al., 2005; Qiu and Yakovlev, 2006; Kim and van de Wiel, 2008; Schwartzman and Lin, 2011). For $\alpha = 0.1$ we got similar results.

The tables in Sections 3.5.2 and 3.5.3 are based on 5000 simulations per setting, which took about half an hour per setting on a good PC.

Table 3.1: 95%-confidence intervals for $\mathbb{P}(\text{upper bound} < FDP)$, for $\alpha = 0.05$. Probabilities larger than 0.05 are shown in boldface.

π_0	$\lvert\rho\rvert$	$\overline{FDP}'$	$\widehat{\pi}_0 \cdot \overline{FDP}$	$\widehat{\pi}_0' \cdot \overline{FDP}$	$\widehat{\pi}_0' \cdot \overline{FDP}'$
1	0	$.038 \pm .006$	$\mathbf{.063} \pm .007$	$\mathbf{.063} \pm .007$	$\mathbf{.505} \pm .014$
1	0.5	$.047 \pm .006$	$\mathbf{.114} \pm .010$	$\mathbf{.114} \pm .010$	$\mathbf{.381} \pm .013$
0.95	0	$.012 \pm .004$	$.028 \pm .005$	$.028 \pm .005$	$.028 \pm .005$
0.95	0.5	$.043 \pm .006$	$\mathbf{.106} \pm .009$	$\mathbf{.106} \pm .009$	$\mathbf{.238} \pm .012$
0.8	0	$.000 \pm .001$	$.002 \pm .002$	$.002 \pm .002$	$.002 \pm .002$
0.8	0.5	$.029 \pm .005$	$\mathbf{.088} \pm .009$	$\mathbf{.088} \pm .009$	$\mathbf{.181} \pm .011$
0.5	0	$.000 \pm .001$	$.000 \pm .001$	$.000 \pm .001$	$.000 \pm .001$
0.5	0.5	$.016 \pm .004$	$.055 \pm .007$	$.055 \pm .007$	$\mathbf{.116} \pm .009$

3.5.3 Performance of variants of SAM as estimators

We performed the same simulations as in Section 3.5.2, but with $\alpha = 0.5$. For $\alpha = 0.5$ we write $\widehat{FDP} := \overline{FDP}$ and we let $\widehat{FDP}' = \widehat{FDP} \wedge 1$. Table 3.2 shows for the different estimates the probability of underestimating the FDP, for different values of π_0 and of the correlation.

The simulations confirm that, as we have proven, $\mathbb{P}(\widehat{FDP}' \leq FDP) \leq 0.5 = \alpha$, i.e. it is a median unbiased estimator. Note also that for the estimate $\widehat{\pi}_0' \cdot \widehat{FDP}'$, this does not hold in all situations. For many of the simulated settings however, all estimates were median unbiased.

In Table 3.3 95%-confidence intervals are shown (assuming normality) for the expected absolute errors of the different estimators. Note that for large π_0, $\widehat{FDP}'$ was a more accurate estimator than the estimators that use $\widehat{\pi}_0$ or $\widehat{\pi}_0'$. For small π_0 and no correlation, the other estimates were more accurate. When there was correlation, the estimates based on $\widehat{\pi}_0$ were often less accurate than $\widehat{FDP}'$. The reason for this may be that $\widehat{\pi}_0$ and $\widehat{\pi}_0'$ were less accurate estimators of π_0 under dependence than under independence. The low accuracy of $\widehat{\pi}_0$ under dependence is a known issue (Qiu et al., 2005; Qiu and Yakovlev, 2006; Kim and van de Wiel, 2008; Schwartzman and Lin, 2011).

Apart from recording the absolute errors we also recorded the relative

Table 3.2: Estimates of $\mathbb{P}(\text{estimate} < FDP)$ for $\alpha = 0.5$. Each estimate is based on 5000 simulations, such that it differs less than 0.015 from the real value with probability at least 95%.

π_0	$\lvert\rho\rvert$	$\widehat{FDP}'$	$\widehat{\pi}_0 \cdot \widehat{FDP}$	$\widehat{\pi}'_0 \cdot \widehat{FDP}$	$\widehat{\pi}'_0 \cdot \widehat{FDP}'$
1	0	0.44	0.51	0.51	0.69
1	0.5	0.45	0.47	0.47	0.47
0.95	0	0.32	0.43	0.43	0.43
0.95	0.5	0.44	0.46	0.47	0.47
0.8	0	0.08	0.29	0.29	0.29
0.8	0.5	0.37	0.45	0.45	0.45
0.5	0	0.00	0.16	0.16	0.16
0.5	0.5	0.28	0.40	0.40	0.40

Table 3.3: 95%-confidence intervals for $\mathbb{E}|\text{estimate} - FDP|$ for $\alpha = 0.5$. In each row, the smallest average error is shown in boldface.

π_0	$\lvert\rho\rvert$	$\widehat{FDP}'$	$\widehat{\pi}_0 \cdot \widehat{FDP}$	$\widehat{\pi}'_0 \cdot \widehat{FDP}$	$\widehat{\pi}'_0 \cdot \widehat{FDP}'$
1	0	**.091** $\pm$ .004	.305 $\pm$ .012	.298 $\pm$ .012	.104 $\pm$ .004
1	0.5	**.332** $\pm$ .011	.543 $\pm$ .018	.469 $\pm$ .014	.348 $\pm$ .011
0.95	0	.099 $\pm$ .002	**.096** $\pm$ .002	**.096** $\pm$ .002	**.096** $\pm$ .002
0.95	0.5	**.387** $\pm$ .009	.520 $\pm$ .017	.456 $\pm$ .014	.399 $\pm$ .009
0.8	0	.050 $\pm$ .001	**.034** $\pm$ .001	**.034** $\pm$ .001	**.034** $\pm$ .001
0.8	0.5	**.236** $\pm$ .008	.268 $\pm$ .010	.256 $\pm$ .009	.244 $\pm$ .008
0.5	0	.044 $\pm$ .000	**.015** $\pm$ .000	**.015** $\pm$ .000	**.015** $\pm$ .000
0.5	0.5	.145 $\pm$ .006	.144 $\pm$ .007	.143 $\pm$ .007	**.142** $\pm$ .007

differences

$$\left|\frac{\text{estimate}}{FDP} - 1\right|.$$

For this error measure we got similar results. In particular, $\widehat{FDP}'$ was the best estimator of the FDP for large π_0.

The closed testing-based estimate $\overline{V}_{\text{ct}}/R$ (for $\alpha = 0.5$) and its approximation $\overline{V}^*_{\text{ct}}/R$ are often more accurate than $\widehat{FDP}'$ (results not shown). For $|\rho| = 0$ and $\pi_0 \leq 0.8$ however, the estimates based on $\hat{\pi}_0$ still performed

best.

3.5.4 Performance of the closed testing-based bound

Here we illustrate that the bound $\overline{V}_{\mathrm{ct}}$ based on the full closed testing procedure often improves the basic bound $\overline{V}$. We computed $\overline{V}_{\mathrm{ct}}$ using the shortcut in Proposition 3.3.1. Recall that calculating $\overline{V}_{\mathrm{ct}}$ is often computationally infeasible when R or $\overline{V}_{\mathrm{ct}}$ is large, hence we took $m = 100$. Further, we took $\alpha = 0.1$ and $D = (0, 0.01)$ as the rejection region. We calculated 95%-confidence intervals (assuming normality) for the expected values of the $(1-\alpha)$-upper bounds. The results are shown in Table 3.4. Recall that $\overline{V}/R$ and $\overline{V}_{\mathrm{ct}}/R$ are the $(1-\alpha)$-confidence FDP bounds corresponding to $\overline{V}$ and $\overline{V}_{\mathrm{ct}}$.

Table 3.4: 95%-confidence intervals for $\mathbb{E}$(upper bound). The column below "R" shows the average number of rejections. The value $|\rho| = 0.2$ corresponds to $\sigma_Z = 0.5$.

π_0	$\lvert\rho\rvert$	R	$\overline{V}/R$	$\overline{V}_{\mathrm{CT}}/R$
0.9	0	8.8 ± 0.1	0.35 ± 0.01	0.33 ± 0.01
0.9	0.2	7.6 ± 0.2	0.47 ± 0.01	0.46 ± 0.01
0.7	0	23.9 ± 0.5	0.18 ± 0.01	0.12 ± 0.00
0.7	0.2	20.2 ± 1.1	0.23 ± 0.02	0.18 ± 0.02
0.5	0	39.1 ± 0.5	0.15 ± 0.01	0.08 ± 0.00
0.5	0.2	40.0 ± 1.8	0.17 ± 0.01	0.11 ± 0.01

The table shows that if π_0 is near 1, the basic bound $\overline{V}$ and the closed testing-based bound $\overline{V}_{\mathrm{ct}}$ are close, but when there are many false hypotheses, closed testing provides a substantial improvement. The same holds for $\alpha = 0.5$.

The simulations were computationally intensive, especially for $\pi_0 = 0.5$ when there were many rejections. For this value of π_0, 100 simulations took about 40 hours on a good PC.

3.5.5 Performance of the conservative shortcut

We illustrate that in some settings for $\alpha = 0.5$ the estimate $\overline{V}_{sc}$ obtained with the conservative shortcut defined in Section 3.4 is lower than the basic estimate $\overline{V}$. In these simulations $m = 2000$. We also took $w = 2000$

and $D = (0, 0.1)$, because the shortcut usually does not improve $\overline{V}$ if the cutoff and w are small. For different values of π_0 and the correlation, we calculated confidence intervals (assuming normality) for the expected absolute difference from the real FDP, for $\widehat{FDP}'$ and $\widehat{FDP}_{sc} := \overline{V}_{sc}/R$. The results are shown in Table 3.5. The computation time was a few minutes per 100 simulations.

As expected, the shortcut only improved $\widehat{FDP}'$ when π_0 was far from 1. The shortcut provides less small bounds than the full closed testing procedure, but is computationally feasible for larger datasets. For such datasets, it is the best computationally feasible bound that has been proven to be a $(1-\alpha)$-confidence bound.

Table 3.5: 95%-confidence intervals for $\mathbb{E}|\text{estimator} - FDP|$. The value $|\rho| = 0.2$ corresponds to $\sigma_Z = 0.5$.

π_0	$\lvert\rho\rvert$	$\widehat{FDP}'$	$\widehat{FDP}_{sc}$
0.8	0	0.117 ± 0.006	0.117 ± 0.006
0.8	0.2	0.145 ± 0.010	0.145 ± 0.010
0.5	0	0.157 ± 0.002	0.150 ± 0.002
0.5	0.2	0.140 ± 0.006	0.140 ± 0.006
0.1	0	0.177 ± 0.001	0.105 ± 0.001
0.1	0.2	0.171 ± 0.002	0.157 ± 0.003

3.5.6 Performance of the approximation method

We now investigate the approximation method (Section 3.3.3), which provides smaller bounds than the conservative shortcut. Its validity as a $(1-\alpha)$-confidence bound has not been proven (for finite $\#\mathcal{S}$), hence we use simulations to investigate its validity.

Firstly we show that in the simulation settings where computation of $\overline{V}_{\text{ct}}$ was feasible, the estimate $\overline{V}^*_{\text{ct}}$ is on average close $\overline{V}_{\text{ct}}$. In the settings of Section 3.5.4 ($\alpha = 0.1$), we constructed $\mathcal{S}$ as a collection of 1000 independent, uniformly distributed random subsets from $\{\mathcal{I} \subseteq \mathcal{R} : \ \#\mathcal{I} = M\}$ (duplicates were allowed). In table 3.6 it can be seen that $\overline{V}^*_{\text{ct}}$ was on average close to $\overline{V}_{\text{ct}}$. This means that they were usually equal and sometimes $\overline{V}^*_{\text{ct}}$ was equal to $\overline{V}_{\text{ct}} - 1$ or $\overline{V}_{\text{ct}} - 2$. Taking $\#\mathcal{S}$ smaller (larger) than 1000

naturally resulted in a larger (smaller) average estimation error (result not shown).

Table 3.6: The last column shows the average absolute difference between the two upper bounds of the FDP. The second-last column shows confidence intervals for the expected values of $\overline{V}_{\mathrm{ct}}/R$. The value $|\rho| = 0.2$ corresponds to $\sigma_Z = 0.5$.

π_0	$\lvert\rho\rvert$	$\overline{V}_{\mathrm{ct}}/R$	$\lvert\overline{V}_{\mathrm{ct}}/R - \overline{V}^*_{\mathrm{ct}}/R\rvert$
0.9	0	0.33 ± 0.01	0.000
0.9	0.2	0.46 ± 0.01	0.000
0.7	0	0.12 ± 0.00	0.005
0.7	0.2	0.18 ± 0.02	0.006
0.5	0	0.08 ± 0.00	0.005
0.5	0.2	0.11 ± 0.01	0.006

The estimate $\overline{V}^*_{\mathrm{ct}}$ of $\overline{V}_{\mathrm{ct}}$ is good, but not perfect. This is irrelevant for our purposes however, as long as $\overline{V}^*_{\mathrm{ct}}$ has the desired property of being a $(1-\alpha)$-confidence bound. In the last column of Table 3.7 it can be seen that this is the case for the simulation setting of Sections 3.5.2 and 3.5.3. (Note that we took $m = 1000$ again, since the approximation method is feasible for large m.) Here $\#\mathcal{S}$ was again taken to be 1000.

It was also interesting to compare $\overline{V}^*_{\mathrm{ct}}$ to the bound $V^{(k)}$. The bound $V^{(k)}$, which is unknown in practice, was shown to be a $(1-\alpha)$-confidence bound in the proof of Theorem 3.2.2. Table 3.7 shows that the probability that $\overline{V}^*_{\mathrm{ct}} < V^{(k)}$ was very small in the simulation settings of Sections 3.5.2 and 3.5.3, with $\overline{V}^*_{\mathrm{ct}} > V^{(k)}$ being much more likely. Since $V^{(k)}$ is a $(1-\alpha)$-upper bound, it is then not surprising that $\overline{V}^*_{\mathrm{ct}}$ is also a $(1-\alpha)$-upper bound in these settings.

For other values of α (and for p-values based on a t-statistic), we similarly found that $\overline{V}^*_{\mathrm{ct}}$ was a $(1-\alpha)$-confidence bound. Based on these findings, it seems reasonable to use $\overline{V}^*_{\mathrm{ct}}$ as a $(1-\alpha)$-confidence upper bound in practice, given that the test statistics are p-values as was the case in our simulation settings. We recommend taking $\#\mathcal{S}$ as large as possible in practice (try $\#\mathcal{S} \geq 10^4$).

Table 3.7: The same simulation setting as in Sections 3.5.2 and 3.5.3 was taken. The second last column shows the estimated probability that $\overline{V}^*_{\mathrm{ct}}$ was smaller than the bound $V^{(k)}$. The last column shows the error rate. The estimates are based on 5000 simulations per setting (taking up to 5 hours).

π_0	$\lvert\rho\rvert$	$\mathbb{E}R$	$\mathbb{P}(\overline{V}^*_{\mathrm{ct}} < V^{(k)})$	$\mathbb{P}(\overline{V}^*_{\mathrm{ct}} < V)$
1	0	10.0	0.000	0.039
1	0.5	10.0	0.018	0.050
0.95	0	27.9	0.000	0.014
0.95	0.5	18.2	0.013	0.048
0.8	0	81.6	0.000	0.002
0.8	0.5	40.0	0.007	0.035
0.5	0	188.3	0.000	0.000
0.5	0.5	87.8	0.002	0.024

3.6 Application to data

We illustrate the performance of the $(1-\alpha)$-upper bound $\overline{V}/R$ on real data. We analyse the freely available dataset that was used for the original SAM paper by Tusher et al. (2001). The dataset contains gene expression levels of about 7000 genes measured with a microarray. For each gene there are eight observations, of which four from unirradiated cells and four from irradiated cells. In each of these two groups there are two observations from one cell line and two observations from another cell line (making four observations per cell line). More details are in Tusher et al. (2001).

We performed the same analysis as Tusher et al., with the addition that we calculated $(1-\alpha)$-confidence upper bounds for the FDP. Not all 8! permutations were used but only the permutation maps that permuted within the two cell lines. There are $4!4! = 576$ such permutation maps. Note that this set of permutation maps has a group structure. This group consists of 36 classes of 16 equivalent permutations that always give the same test statistic. Using one permutation from each class leads to the same analysis as with 576 permutations, so we only use 36 distinct permutations. The same permutations are used in Tusher et al. (2001).

For gene i, H_i is defined as the hypothesis that the distribution of the expression level of gene i is the same for all cells. Note that Assumption 3.2.1 is satisfied if the joint distribution of the gene expression levels

corresponding to $\mathcal{N}$ is the same for cases and controls. As a biological argument for this exhangeability, note that it seems unlikely that the treatment would affect the joint distribution of the gene expressions corresponding to $\mathcal{N}$, while leaving the marginal distributions unchanged.

In Tusher et al. and here, the user chooses a threshold $\Delta \geq 0$. Based on Δ and the data, the rejection region D is calculated. This region is of the form $(-\infty, c_1) \cup (c_2, \infty)$, with c_1, $c_2 \in \mathbb{R}$. Details on how the cut-offs c_1 and c_2 are based on Δ and the data are in Tusher et al.. The larger Δ is, the fewer hypotheses are rejected and the smaller the FDP tends to be. The dependence of the cut-offs on the data might lead to bias. The bias is minor or absent however, as long as Δ is not cherry-picked after looking at the data. In the analysis here and in Tusher et al. no plug-in estimate of π_0 was used.

Considering the same values of the threshold Δ as Tusher et al. and some larger values, we calculated the corresponding estimates of the FDP as well as the basic $(1-\alpha)$-confidence upper bound for the FDP. The results are shown in Table 3.8. Here $\overline{FDP}_\gamma$ stands for $\overline{V}/R$ for $1-\alpha = \gamma$, so that e.g. $\overline{FDP}_{0.9}$ is a 90%-confidence upper bound for the FDP. $\widehat{\overline{FDP}}_{\text{mean}}$ stands for $\widehat{V}_{\text{mean}}/R$ where $\widehat{V}_{\text{mean}}$ is the mean of the values $R(gX)$, $g \in H$, where H is the set of 36 permutations . This is the estimate that is reported in Table 1 in Tusher et al. (2001). Keep in mind that the bounds are not uniform over Δ or α.

Table 3.8: For different values of the threshold Δ, estimators and bounds for the FDP are shown. R is the number of rejected hypotheses. The value $\overline{FDP}_{0.5}$ is a median unbiased estimator of the FDP and $\overline{FDP}_{0.95}$ is a 95%-confidence upper bound for the FDP.

Δ	R	$\widehat{\overline{FDP}}_{\text{mean}}$	$\overline{FDP}_{0.5}$	$\overline{FDP}_{0.9}$	$\overline{FDP}_{0.95}$
0.3	571	0.56	0.45	0.97	1
0.4	282	0.46	0.34	0.99	1
0.6	162	0.35	0.25	0.98	1
0.9	80	0.24	0.13	0.88	0.98
1.2	46	0.18	0.09	0.67	0.98
1.8	26	0.14	0.08	0.46	0.85
2.5	12	0.12	0.08	0.42	0.75
3	10	0.12	0.10	0.30	0.70
3.5	3	0.06	0	0.33	0.33

Some of our results are slightly different from those in Table 1 in Tusher et al. (2001), which may be due to a minor difference in the code or the data used. Note that for every Δ the estimate $\overline{FDP}_{0.5}$ based on the median is smaller the estimate $\widehat{FDP}_{\text{mean}}$ based on the mean. This is because the permutation distribution of R tended to be skewed to the right. Note that for $\alpha = 0.05$ and smaller values of Δ, we obtain trivial 95%-confidence bounds. For example, for $\Delta = 0.6$ we do not have 95% confidence that at least one of the 162 rejected hypotheses is false. For larger values of Δ the cut-offs are stricter and we do get useful 95%-confidence bounds.

Note that since there are only 36 permutations, the 95%-confidence bound for V is the second largest value among $R(gX)$, $g \in H$. Thus it is in fact a $(35/36)100\% \approx 97.2\%$ confidence bound. For $\Delta = 3.5$ there are 3 rejections and we know with 97.2% confidence that at least two of these are true findings. We also know with 50% confidence that all three rejections are true findings. For $\Delta = 3$ there are 10 rejections and we know with 90% confidence (and indeed $(33/36)100\% \approx 91.7\%$ confidence) that at least 7 of these are true findings, although we cannot generally pinpoint which of the rejected hypotheses are false.

Calculating $\overline{V}_{\text{ct}}$ was only feasible for $\Delta \geq 2.5$ and sometimes offered an improvement over $\overline{V}$. For example, for $\Delta = 3$ and $\alpha = 0.05$, the bound was 0.6 instead of 0.7. Usually the basic bound was not improved for $\Delta \geq 2.5$, due to the relatively small number of rejections for such Δ and the discreteness of the already small bound $\overline{V}_{\text{ct}}$.

For $\Delta < 2.5$, when computing $\overline{V}_{\text{ct}}$ was not feasible, we performed the approximation method (with $\#\mathcal{S} = 10^4$). The results are shown in Table 3.9. The improvements are relatively small in this situation, since there is no proof that π_0 is far away from 1 for these data.

In many practical situations FDP bounds (and the FDP itself) tend to decrease with R, but this is only a tendency. Examples of exceptions can be seen in both Table 3.8 and Table 3.9. Hence a user might find that decreasing Δ post hoc would both increase R and decrease the bound, which would be very tempting. This could lead to selection bias however; Δ should be chosen before looking at the data.

Table 3.9: For different values of the threshold Δ, estimators and bounds for the FDP, derived with the approximation method, are shown. Here $\overline{FDP}^*_\gamma$ stands for $\overline{V}^*_{\mathrm{ct}}/R$ for $1-\alpha=\gamma$. Where it improved the basic bound (see Table 3.8), the result is shown in boldface.

Δ	R	$\overline{FDP}^*_{0.5}$	$\overline{FDP}^*_{0.9}$	$\overline{FDP}^*_{0.95}$
0.3	571	**0.43**	**0.81**	1
0.4	282	0.34	**0.82**	1
0.6	162	0.25	**0.88**	1
0.9	80	0.13	0.88	**0.94**
1.2	46	0.09	0.67	**0.96**
1.8	26	0.08	0.46	**0.81**

3.7 Discussion

SAM is a widely applied method, since it requires few assumptions on the dependence structure of the data and nevertheless adapts to this structure. Until now SAM had no known properties. In this paper the assumptions underlying SAM have been made explicit. Moreover it has been shown how SAM can be extended to provide a $(1-\alpha)$-confidence upper bound for the FDP. For $\alpha=0.5$ a median unbiased estimate of the FDP is obtained. The *samr* *R*-package multiplies this estimate by an estimate of the fraction of true hypotheses π_0 to obtain a lower estimate of the FDP. We have shown using simulations that this often still results in a median unbiased estimate of the FDP, although in many cases the estimate becomes less accurate. For $\alpha=0.05$ and $\alpha=0.1$, multiplying the $(1-\alpha)$-confidence bound by the estimate of π_0 often does not result in a $(1-\alpha)$-confidence bound.

We have shown that by using a closed testing procedure the basic bound can be decreased, in such a way that the confidence level is maintained. The improvement over the basic bound can be appreciable, as simulations illustrate. The improved bound only depends on rejected sets for permuted versions of the data. Once these are known, the computation time is not influenced by the complexity of the test statistics. Hence the choice of test statistics typically does not determine the computational feasibility of the method.

When there are many rejected hypotheses, the closed testing-based method is often computationally infeasible. Therefore we have included

a fast approximation of this method, which still provides confidence for our simulation settings. We have also constructed a conservative shortcut, which provides larger bounds but has proven validity. This shortcut only improves the basic bound in specific settings. Both these fast alternatives to the closed testing-based method are feasible when there are many thousands of rejections.

Our methods provide an FDP bound for the prespecified rejection region. The region cannot generally be picked after looking at the data, since the bounds are not uniform over multiple rejection regions. There exists a limited amount of literature on uniformly valid FDP bounds (Meinshausen, 2006; Goeman and Solari, 2011). An example is the method by Meinshausen (2006), which is closely related to SAM. There are opportunities to improve some of these methods, similarly to the way SAM has been improved here. This may be the subject of future research.

Theorem 3.2.1 provides a general permutation principle which can be used to prove properties of methods based on random permutations (SAM; Meinshausen, 2006; Westfall and Young, 1993). This result is related to Phipson and Smyth (2010) but is more generally useful. We have used it to prove the validity of the methods in this paper. It may be used to prove properties of other permutation-based procedures in the future.

4

Permutation-based simultaneous confidence bounds for the false discovery proportion

Abstract

When many hypotheses are tested, interest is often in ensuring that the proportion of false discoveries (FDP) is small with high confidence. In this paper, confidence upper bounds for the FDP are constructed, which are simultaneous over all rejection cut-offs. In particular this allows the user to select a set of hypotheses such that the FDP lies below some constant with high confidence. Our methods use permutations to account for the dependence structure in the data. So far only Meinshausen provided an exact, permutation-based and computationally feasible method for simultaneous FDP bounds. We improve this procedure by embedding it within a closed testing framework. Further, we provide a generalisation of the method. It lets the user specify a set from which the envelope of confidence bounds is selected. This gives the user more freedom in determining the properties of the method. Interestingly, several existing permutation methods, such as SAM and Westfall and Young's $\max T$ method, are obtained as spe-

cial cases. The different procedures in this paper are compared using both simulated and real data.

4.1 Introduction

The goal of many multiple testing methods is to reject as many hypotheses as possible while incurring few type-I errors. The resulting proportion of type-I errors among the rejections is called the False Discovery Proportion (FDP). The FDP has received much attention in recent years since under strong dependence, it represents a more relevant measure than the False Discovery Rate (FDR) (e.g. Benjamini and Hochberg, 1995), the expected value of the FDP (Schwartzman and Lin, 2011; Schwartzman, 2012; Guo et al., 2014). Under strong dependence the FDR can be far from the true FDP.

In practical applications, when rejecting all hypotheses with p-values less than an certain threshold, one would like to know a $(1-\alpha)100\%$-confidence upper bound for the FDP. The goal of this paper is to provide confidence bounds for the FDP which are simultaneous over multiple thresholds. This allows the user to freely select the threshold *post hoc*, i.e. after looking at the data, and still obtain a valid confidence bound.

There exist several methods that control the probability that the FDP exceeds a prespecified constant ('exceedance control', e.g. van der Laan et al., 2004b; Farcomeni, 2009; Lehmann and Romano, 2012; Guo et al., 2014). The number of published exact methods allowing post hoc selection however is limited (Meinshausen, 2006). Most methods (including those in the present paper) are special cases or shortcuts for the general methods in Genovese and Wasserman (2006) and Goeman and Solari (2011), which we show to be equivalent. These methods are based on closed testing (Marcus et al., 1976), which means that many *local tests* need to be performed. These procedures are not always computationally feasible, but in some cases fast computational shortcuts exist, making them feasible. This is e.g. the case (Meijer et al., 2017) when the local tests are based on Simes' probability inequality (Simes, 1986). However, these and other parameric local tests are conservative for many dependence structures of the p-values, making the resulting FDP bounds conservative as well.

In multiple testing, when a permutation method can be used, this often offers an improvement in power over parametric procedures. The reason is

that permutation methods take into account the dependence structure of the test statistics (e.g. p-values), instead of having to account for worst-case scenarios as do parametric methods. For FDP confidence, existing permutation methods are Korn et al. (2004, 2007), Meinshausen and Bühlmann (2005) and Hemerik and Goeman (2018), but only Meinshausen (2006) offers simultaneous FDP bounds and hence post hoc selection. It was also the only permutation method that provides exceedance control of the FDP. Meinshausen's procedure often outperforms parametric methods.

In the present paper, Meinshausen (2006) is generalised and improved. Interestingly, the generalisation has other well-known permutation methods as special cases, for example the $\max T$ method by Westfall and Young (1993) and the method in Hemerik and Goeman (2018) (an extension of SAM by Tusher et al., 2001).

We improve Meinshausen (2006) in the following ways. First, we show how random permutations can be employed in such a way that the method is exact (i.e. all bounds are valid with probability at least $1-\alpha$). Second, as will be explained we allow many families of candidate confidence envelopes, which allows more freedom in constructing the simultaneous upper bounds. Moreover, unlike in Meinshausen, these candidate envelopes do not depend on the data, so that potential bias is avoided. Further, we show how the power of Meinshausen can be uniformly increased using closed testing (see also Goeman and Solari, 2011). We also provide an exact iterative method (not always equivalent to closed testing) that improves Meinshausen as well, but is more computationally feasible. For cases with many hypotheses, where the iterative method is infeasible, we suggest an approximation of it, which can be used even when there are many thousands of rejected hypotheses. The approximation method maintained the nominal error rate in all our simulation scenarios.

This paper is built up as follows. Section 4.2 introduces notation and assumptions. In Section 4.3, single-step procedures, including Meinshausen (2006), are explained. Next, in Section 4.4, we show how these upper bounds can be improved with a closed-testing based method. In Section 4.5 the iterative method is presented. The various methods are compared using simulations and real data in Sections 4.6 and 4.7 respectively.

4.2 Setting and notation

Let X be random data, taking values in a sample space Ω. Hypotheses $H_1, ..., H_m$ are considered with corresponding p-values $P_i : \Omega \to [0,1]$, $1 \leq i \leq m$. We will often suppress the dependence on X in the notation, e.g. P_i is short for $P_i(X)$. Without loss of generality we assume that $P_1 \leq ... \leq P_m$. Write $\mathcal{N} = \{1 \leq i \leq m : H_i \text{ is true}\}$, $n = \#\mathcal{N}$ and let $\mathbf{Q}$ be the sorted vector $(P_i : i \in \mathcal{N})$ (assume $\mathcal{N} \neq \emptyset$ for convenience).

Let $\alpha \in [0,1)$ and $\mathbb{T} \subseteq [0,1]$ be independent of the data. For $t \in \mathbb{T}$ define $\mathcal{R} = \mathcal{R}(t) = \{1 \leq i \leq m : P_i \leq t\}$. This is the set of indices of the rejected hypotheses if each hypothesis H_i is rejected when $P_i \leq t$. Write $R = \#\mathcal{R}$ and $V = \#(\mathcal{R} \cap \mathcal{N})$, the number of false positives. Note that $\mathcal{R}$ and V depend on the data, but $\mathcal{N}$ does not. Further, we have $FDP(t) = V(t)/R(t)$, which is interpreted as 0 when $R(t) = 0$.

All nonparametric methods in this paper are based on permutations or other transformations of the data. Let G be a finite set of transformations $g : \Omega \to \Omega$, such that G is a group (in the algebraic sense) with respect to the operation of composition of transformations. In practice G is often a group of permutation maps. Sometimes other groups of transformations can be used, such as rotations (Langsrud, 2005; Solari et al., 2014) and multiplication of part of the data by -1 (Pesarin and Salmaso (2010), pp. 54 and 168).

All permutation-based procedures in the paper rely on the following assumption.

Assumption 4.2.1. The joint distribution of the p-values $P_i(g(X))$ with $i \in \mathcal{N}$, $g \in G$, is invariant under all transformations in G of X.

This assumption underlies many permutation-based multiple testing methods, e.g. Westfall and Young's $\max T$ method (1993), Tusher et al. (2001), Hemerik and Goeman (2018), Meinshausen and Bühlmann (2005) and Meinshausen (2006). Usually this assumption means that the joint distribution of the part of the data corresponding to $\mathcal{N}$ should be invariant under permutation.

In this paper random transformations from G are used, which are defined as follows.

Definition 4.2.1. Let $g_1 := id$ be the identity in G and $g_2, ..., g_w$ random elements from G. The random transformations can be drawn either

with or without replacement: the statements in this paper hold for both cases. If $g_2, ..., g_w$ are drawn without replacement, then they are taken to be uniformly distributed on $G \setminus \{id\}$, otherwise uniform on G.

For $\mathcal{I} \subseteq \{1, ..., m\}$ and $1 \leq j \leq w$, write $R_j^{\mathcal{I}}(t) := \#\{i \in \mathcal{I} : P_i(g_j(X)) \leq t\}$, $R_j := R_j^{\{1,...,m\}}$ and $R^{\mathcal{I}} := R_1^{\mathcal{I}}$.

4.3 Single-step procedures

4.3.1 Confidence envelopes

The aim of this paper is to derive as small as possible *confidence envelopes*, which we define similarly to Genovese and Wasserman (2006). In Meinshausen and Bühlmann (2005) these are referred to as bounding functions.

Definition 4.3.1. A confidence envelope is a (possibly random) function $B : \mathbb{T} \to \mathbb{N}$ satisfying

$$\mathbb{P}\Big(\bigcap_{t \in \mathbb{T}} \big\{V(t) \leq B(t)\big\}\Big) \geq 1 - \alpha.$$

Note that with probability at least $1 - \alpha$, simultaneously for all $t \in \mathbb{T}$, the numbers $B(t)$ are upper bounds for the numbers of false positives $V(t)$. Note that if $B(t) \geq V(t)$ and $R(t) > 0$, then $B(t)/R(t) \geq FDP(t)$. Hence, from simultaneous upper bounds for $V(t)$ simultaneous upper bounds for $FDP(t)$ immediately follow.

Let $(\cdot)^+$ denote the positive part function. Note that given a confidence envelope B, the bounds $(R(t) - B(t))^+$, $t \in \mathbb{T}$, are simultaneous $(1 - \alpha)$-lower bounds for the numbers of true findings $R(t) - V(t)$, $t \in \mathbb{T}$. Since these bounds are simultaneous,

$$R(t) - \max\{(R(s) - B(s))^+ : s \in \mathbb{T}, s \leq t)\} \tag{4.1}$$

is a potentially improved confidence envelope. In this way, the simultaneity is exploited to improve the envelope when $(R - B)^+$ is not non-decreasing.

We will often consider a general collection of functions of the following form.

Definition 4.3.2. For each nonempty $\mathcal{I} \subseteq \{1, ..., m\}$ consider $B_{\mathcal{I}} : \mathbb{T} \to \mathbb{N}$, such that $B_{\mathcal{I}} \leq B_{\mathcal{J}}$ whenever $\mathcal{I} \subseteq \mathcal{J} \subseteq \{1, ..., m\}$ and such that $B_{\mathcal{N}}$ is a confidence envelope.

Suppose the $B_{\mathcal{I}}$, $\mathcal{I} \subseteq \{1, ..., m\}$, are given. The function $B_{\mathcal{N}}$ is a confidence envelope, but is in practice unknown, since $\mathcal{N}$ is unknown. The larger function $B^{\max} := \max_{\mathcal{I}} B_{\mathcal{I}} = B_{\{1,...,m\}}$ however is a known confidence envelope. The envelope G_α of Theorem 2 in Meinshausen and Bühlmann (2005) is a special case of this. All confidence envelopes B derived in this paper are uniform improvements of $B^{\max}$, i.e. $B(t) \leq B^{\max}(t)$ for all $t \in \mathbb{T}$.

Confidence envelopes can de derived from *critical vectors*.

Definition 4.3.3. A vector $\mathbf{c} = (c_1, ..., c_{\#\mathbf{c}})$ with $\#\mathbf{c} \geq n$ is a *critical vector* if

$$\mathbb{P}\big(\bigcap_{i=1}^{n} \{Q_i \geq c_i\}\big) \geq 1 - \alpha. \tag{4.2}$$

Proposition 4.3.1. *If $\mathbf{c} \in \mathbb{R}^m$ is a critical vector then the map $B : \mathbb{T} \to \{1, ..., m\}$ defined by*

$$B(t) = \#\big\{1 \leq i \leq \#\mathbf{c} : c_i \leq t\big\}$$

is a confidence envelope. If $\mathbb{T} = [0, 1]$, then the reverse implication also holds.

Proof. With probability at least $1 - \alpha$, $\mathbf{Q} \geq \mathbf{c}$, and then for each $t \in [0, 1]$,

$$V(t) = \#\big\{1 \leq i \leq n : Q_i \leq t\big\} \leq \#\big\{1 \leq i \leq \#\mathbf{c} : c_i \leq t\big\} = B(t).$$

Thus B is a confidence envelope. For $\mathbb{T} = [0, 1]$ the reader can check the claimed equivalence. □

Observe that the larger $\mathbf{c}$ is, the smaller the confidence envelope obtained with Proposition 4.3.1 is. Hence it is of interest to find as large as possible $\mathbf{c}$. The existing literature provides various critical vectors and we can use these to construct confidence envelopes. An example is given in the following.

4.3.2 Parametric confidence envelopes

In many practical situations, the distribution of $\mathbf{Q}$ is such that a well-known probability inequality by Simes (1986) holds (Rødland, 2006):

$$\mathbb{P}\Big(\bigcap_{i=1}^{n} \{Q_i \geq i\alpha/n\}\Big) \geq 1 - \alpha. \tag{4.3}$$

This probability equality provides a critical vector, which can be used to obtain a confidence envelope $B : [0, 1] \to \{1, ..., n\}$ with Proposition 4.3.1:

$$B(t) = \#\{1 \leq i \leq n : i\alpha/n \leq t\}.$$

However, n is not known, so that this envelope is unknown in practice. One can instead note that $n \leq m$ and use the confidence envelope given by

$$B(t) = \#\{1 \leq i \leq m : i\alpha/m \leq t\} = \#\{1 \leq i \leq m : i \leq mt/\alpha\} = \lfloor mt/\alpha \rfloor \wedge m. \tag{4.4}$$

This confidence envelope can be improved using closed testing, as in Meijer et al. (2017), or using the nonparametric methods in this paper.

Simes' probability inequality is not valid for all possible dependence structures of $\mathbf{Q}$, so that the above confidence envelope cannot always be used. Even if Simes' probability inequality holds, the critical vector based on it can be very conservative, because the probability at (4.3) can be larger than $1-\alpha$. This often happens when the Q_i are positively correlated. Other parametric critical vectors are also often conservative or require much stronger assumptions (Cai and Sarkar, 2008; Gou and Tamhane, 2014). We discuss alternative, nonparametric methods in the following.

4.3.3 Meinshausen's nonparametric confidence envelope

When Assumption 4.2.1 is satisfied, often a better confidence envelope can be constructed by using the permutation distribution of the p-values $\mathbf{Q}$. Since by assumption this permutation distribution retains the dependence structure of these p-values, it can be used to construct an envelope which is adapted to this structure. Until now this was only done by Meinshausen (2006). We now recall his method, before uniformly improving it in Sections 4.4 and 4.5.

Central to the method is a family of *candidate envelopes*, explained below. In Meinshausen (2006) these depend on p-values corresponding to false hypotheses, so that the joint distribution of $\mathbf{Q}$ and the candidate envelope picked in Meinshausen is not generally permutation invariant (Blanchard et al. 2017, p. 25, also note this). Hence we consider candidate envelopes that are independent of the data. An additional difference is that we include the original observation with the random permutations, since otherwise the method is not always exact. Otherwise, the method provided here is the same as the procedure in Meinshausen (2006).

Let $\mathbb{B}$ be a set of maps $\mathbb{T} \to \mathbb{N}$, independent of the data. $\mathbb{B}$ is the family of *candidate envelopes*. Suppose that for all B, $B' \in \mathbb{B}$, either $B \geq B'$ or $B' \geq B$. Examples of such $\mathbb{B}$ are in Section 4.3.4.

Meinshausen's confidence envelope (with the above adaptations) is defined as follows.

Theorem 4.3.1. *Let*

$$B^{\mathrm{m}} = \min\left\{B \in \mathbb{B} :\ w^{-1}\#\Big\{1 \leq j \leq w : \bigcap_{t\in\mathbb{T}} \big\{R_j(t) \leq B(t)\big\}\Big\} \geq 1-\alpha\right\},$$

where we assume that $\mathbb{B}$ *is such that this minimum exists. Then* B^{m} *is a confidence envelope.*

Proof. Let

$$B_{\mathcal{N}} = \min\left\{B \in \mathbb{B} :\ w^{-1}\#\Big\{1 \leq j \leq w : \bigcap_{t\in\mathbb{T}} \big\{R^{\mathcal{N}}_j(t) \leq B(t)\big\}\Big\} \geq 1-\alpha\right\}.$$

By the permutation principle (Hemerik and Goeman, 2017, 2018),

$$\mathbb{P}\Big(\bigcap_{t\in\mathbb{T}} \big\{R^{\mathcal{N}}(t) \leq B_{\mathcal{N}}(t)\big\}\Big) \geq 1-\alpha.$$

Since $R^{\mathcal{N}} = V$, this means that $B_{\mathcal{N}}$ is a confidence envelope. Hence the larger function B^{m} is also a confidence envelope.

□

Note that B^{m} is always an element of the family $\mathbb{B}$. Hence the choice of $\mathbb{B}$ has a crucial influence on B^{m}. It is an important assumption that for all B, $B' \in \mathbb{B}$, either $B \geq B'$ or $B' \geq B$. This guarantees that $B_{\mathcal{N}}(t) \leq B^{\mathrm{m}}(t)$ for all $t \in \mathbb{T}$.

The relation of B^{m} to Definition 4.3.2 is the following. For each nonempty $\mathcal{I} \subseteq \{1, ..., m\}$, let

$$B_{\mathcal{I}} = \min\left\{B \in \mathbb{B} :\ w^{-1}\#\Big\{1 \leq j \leq w : \bigcap_{t\in\mathbb{T}} \big\{R^{\mathcal{I}}_j(t) \leq B(t)\big\}\Big\} \geq 1-\alpha\right\}. \tag{4.5}$$

These $B_{\mathcal{I}}$ satisfy Definition 4.3.2. For this choice of $B_{\mathcal{I}}$, $B^{\max}$ is precisely B^{m}. We improve this envelope in Sections 4.4 and 4.5.

4.3.4 Examples of candidate envelopes

We will now give some examples of families $\mathbb{B}$. Consider $\mathbb{B} = \{B^\lambda : \lambda \in [0, \infty)\}$, where $B^\lambda : \mathbb{T} \to \{1, ..., m\}$ is defined by

$$B^\lambda(t) = \#\{1 \leq i \leq m : i\lambda \leq t\}. \tag{4.6}$$

Note that by Proposition 4.3.1, B^λ is a confidence envelope if the vector $(\lambda, 2\lambda, ..., m\lambda)$ is a critical vector. This vector is simply Simes' vector multiplied by a constant. As another example, instead of considering the candidate envelopes (4.6), one could translate (shift) them by replacing $i\lambda$ by e.g. $i\lambda - 0.001$. This often results in better bounds for the larger cut-offs in $\mathbb{T}$, as illustrated in Figure 4.1 and Section 4.7.

If variables $U_1, ..., U_m$ are independent and uniformly distributed on $[0, 1]$, and $U_{(1)} \leq ... \leq U_{(m)}$ are the sorted values of these variables, then it is well known that $U_{(i)}$ has a beta distribution:

$$U_{(i)} \sim \text{Beta}(i, m + 1 - i).$$

For each $\lambda \in [0, 1]$ consider the function $B^\lambda : \mathbb{T} \to \{1, ..., m\}$ given by

$$B^\lambda(t) = \#\{1 \leq i \leq m : q_i^\lambda \leq t\},$$

where q_i^λ is the λ-quantile of the Beta$(i, m + 1 - i)$ distribution. In Section 4.7 we will consider $\{B^\lambda : \lambda \in (0, 1)\}$ as one of the sets of candidate envelopes. A heuristic reason for considering this set of candidate functions is that some of them can be similar in shape to some of the functions $t \mapsto R_j(t)$, $2 \leq j \leq w$. Consequently, the resulting confidence envelopes tend to be relatively tight. We applied the proposed families $\mathbb{B}$ to the data of section 4.7, see Figure 4.1. More examples of candidate critical vectors (and hence candidate envelopes) are in Blanchard et al. (2008).

We will now show that two existing multiple testing methods, "Significance Analysis of Microarrays" (SAM) (Tusher et al., 2001; Hemerik and Goeman, 2018) and the single-step max$T$ method by Westfall and Young (1993), are special cases of the general method at Theorem 4.3.1. These methods essentially only differ with respect to the family $\mathbb{B}$ of candidate envelopes on which they are based.

Let $c \in \mathbb{T} \subseteq [0, 1]$. SAM is obtained with the family $\mathbb{B}$ of candidate envelopes $\{B^0, B^1, ..., B^m\}$, where $B^\lambda : \mathbb{T} \to \{0, ..., m\}$ is given by

$$B^\lambda(t) = \begin{cases} \lambda, & \text{if } t \leq c \\ m, & \text{otherwise.} \end{cases}$$

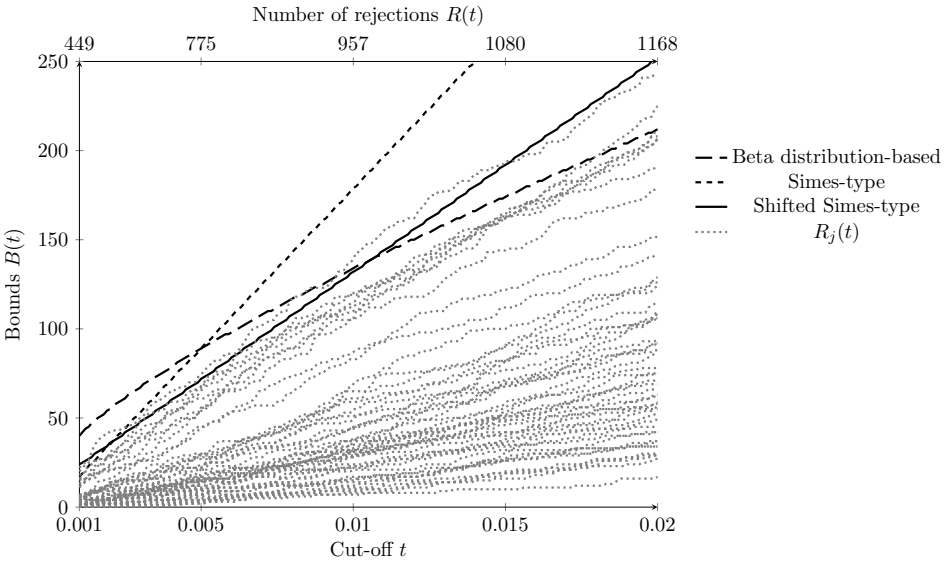


Figure 4.1: For three different families $\mathbb{B}$, resulting 90%-confidence envelopes are shown for cut-offs in $\mathbb{T} = [0.001, 0.02]$ (van de Vijver data). Moreover, for some of the permuted versions of the data, the corresponding numbers of rejections $R_j(t)$ are shown. Each confidence envelope lies above 90% of these curves.

Note that if Meinshausen's method is applied based on these candidate functions, then the resulting upper bound $B^{\mathrm{m}}(c)$ for $V(c)$ is simply the $\lceil (1-\alpha)w \rceil w^{-1}$-quantile of the values $R_j(c)$, $1 \leq j \leq w$. This is precisely the (most basic) upper bound in Hemerik and Goeman (2018).

Consider the family $\mathbb{B} = \{B^\lambda : \lambda \in [0,1]\}$, where $B^\lambda : [0,1] \to \{0, ..., m\}$ is defined by

$$B^\lambda(t) = \begin{cases} 0, & \text{if } t < \lambda \\ m, & \text{otherwise.} \end{cases}$$

Applying Theorem 4.3.1 to these candidate envelopes results in the upper bound $B^{\mathrm{m}} = B^{\lambda'}$, where λ' is the α-quantile of the values $\min_{1 \leq i \leq m} P_i(g_j X)$, $1 \leq j \leq w$. The bound $B^{\lambda'}(t)$ equals zero for $t < \lambda'$, which means that the family-wise error rate is strongly controlled if the hypotheses $\{1 \leq i \leq m : P_i < \lambda'\}$ are rejected. This is exactly the set of hypotheses that the single-step $\max T$ method rejects (Westfall and Young, 1993). Moreover, using the iterative method in Section 4.5, the full $\max T$ method can be obtained.

4.4 Improved bounds by closed testing

Goeman and Solari (2011) show how closed testing (Marcus et al., 1976) can be used to obtain upper bounds for the FDP. As will be seen, this result is equivalent to that in Genovese and Wasserman (2006). By relating Meinshausen's procedure to this method, we will derive uniformly improved upper bounds for the FDP.

For each nonempty $\mathcal{I} \subseteq \{1, ..., m\}$, denote by $H_{\mathcal{I}}$ the intersection hypothesis $\bigcap_{i \in \mathcal{I}} H_i$. Suppose that for each nonempty $\mathcal{I} \subseteq \{1, ..., m\}$ a test for $H_{\mathcal{I}}$ is defined and suppose $H_{\mathcal{N}}$ is rejected with probability at most α. These $2^m - 1$ tests are called *local tests*. The *closed testing procedure* rejects all $H_{\mathcal{I}}$ for which all $H_{\mathcal{J}}$ with $\mathcal{J} \supseteq \mathcal{I}$ are rejected.

Genovese and Wasserman (2004, 2006) formulate the FDP bounds as follows. We slightly generalise their setup, since we consider any level-α local tests. Let $\mathcal{U}$ be the set of $\mathcal{B} \subseteq \{1, ..., m\}$ for which $H_{\mathcal{B}}$ is not rejected by its local test. For $\mathcal{K} \subseteq \{1, ..., m\}$, Genovese and Wasserman (2006) consider the bound

$$\overline{V}_{ct}(\mathcal{K}) = \max\{\#\mathcal{B} \cap \mathcal{K} : \mathcal{B} \in \mathcal{U}\}, \tag{4.7}$$

where the maximum is defined to be zero if the set is empty. The following holds.

Theorem 4.4.1. *Uniformly over all $\mathcal{K} \subseteq \{1, ..., m\}$, $\overline{V}_{ct}(\mathcal{K})$ is a $(1-\alpha)$-upper bound for $\#\mathcal{N} \cap \mathcal{K}$, i.e.*

$$\mathbb{P}\left[\bigcap_{\mathcal{K} \subseteq \{1,...,m\}} \left\{\#\mathcal{N} \cap \mathcal{K} \leq \overline{V}_{ct}(\mathcal{K})\right\}\right] \geq 1 - \alpha.$$

Proof. With probability at least $1 - \alpha$, $H_{\mathcal{N}}$ is not rejected by its local test, and then $\#\mathcal{N} \cap \mathcal{K} \leq \overline{V}_{\mathrm{ct}}(\mathcal{K})$ for all $\mathcal{K} \subseteq \{1, ..., m\}$. □

Note that $\#\mathcal{N} \cap \mathcal{K}$ is the number of false positives if $\mathcal{K}$ is the rejected set. Thus the theorem provides bounds for the numbers of false positives that are uniform over all possible rejected sets.

It turns out that the bounds $\overline{V}_{ct}(\mathcal{K})$ are equal to the bounds constructed in Goeman and Solari (2011). They consider

$$\mathcal{C} := \{\mathcal{I} \subseteq \{1, ..., m\} : \quad H_{\mathcal{I}} \text{ is rejected by the closed testing procedure}\}.$$

For each $\mathcal{K} \subseteq \{1, ..., m\}$ they define the bound as

$$\max\{\#\mathcal{I} : \mathcal{I} \subseteq \mathcal{K},\ \mathcal{I} \notin \mathcal{C}\}, \tag{4.8}$$

Uniformly over all $\mathcal{K} \subseteq \{1, ..., m\}$, (4.8) is a $(1-\alpha)$-upper bound for $\#\mathcal{N} \cap \mathcal{K}$. To prove this, note that with probability at least $1 - \alpha$, $H_{\mathcal{N}}$ is not rejected by its local test, and then $\mathcal{N} \cap \mathcal{K} \notin \mathcal{C}$ for all $\mathcal{K} \subseteq \{1, ..., m\}$.

We now show that the bounds (4.7) and (4.8) are equal, which has never been noted to our knowledge.

Theorem 4.4.2. *The bounds* (4.7) *and* (4.8) *are equal for every* $\mathcal{K} \subseteq \{1, ..., m\}$.

Proof. We are done if we show that

$$\begin{aligned}
&\max\{\#\mathcal{B} \cap \mathcal{K} : \mathcal{B} \in \mathcal{U}\} = \\
&\max\{\#\mathcal{B} \cap \mathcal{K} : \mathcal{B} \in \mathcal{U} \text{ and } \mathcal{B} \cap \mathcal{K} \notin \mathcal{C}\} = && (4.9)\\
&\max\{\#\mathcal{B} \cap \mathcal{K} : \mathcal{B} \subseteq \{1, ..., m\} \text{ and } \mathcal{B} \cap \mathcal{K} \notin \mathcal{C}\} = && (4.10)\\
&\max\{\#\mathcal{I} : \mathcal{I} \subseteq \mathcal{K} \text{ and } \mathcal{I} \notin \mathcal{C}\}.
\end{aligned}$$

The first and the last equality clearly hold. It is also clear that (4.9) $\leq$ (4.10), so it is left to show that (4.10) $\leq$ (4.9), which we now do. Note that if $\mathcal{B} \subseteq \{1, ..., m\}$ and $\mathcal{B} \cap \mathcal{K} \notin \mathcal{C}$, then there is a $\mathcal{B}' \in \mathcal{U}$ with $\mathcal{B}' \supseteq \mathcal{B} \cap \mathcal{K}$ and $\mathcal{B}' \cap \mathcal{K} \notin \mathcal{C}$. Then obviously $\#\mathcal{B} \cap \mathcal{K} \leq \#\mathcal{B}' \cap \mathcal{K} \leq$(4.9). It follows that (4.10) $\leq$ (4.9). □

The equivalent formulations (4.8) and (4.7) are closely related, since in both cases the maximum is taken over all subsets of $\mathcal{K}$ that are not rejected by the closed testing procedure. Nevertheless the two formulations suggest different algorithms for computing the upper bound. If a shortcut exists for the closed testing procedure, then an algorithm based on (4.8) may be faster than one based on (4.7).

As an example of a local test, consider the one which rejects $H_{\mathcal{I}}$ when

$$\bigcup_{t \in \mathbb{T}} \{R^{\mathcal{I}}(t) > B_{\mathcal{I}}(t)\}, \tag{4.11}$$

where $B_{\mathcal{I}}$ is as in Definition 4.3.2. For example, $B_{\mathcal{I}}$ can be defined as (4.5). Using these local tests in (4.7) we obtain simultaneous bounds $\overline{V}_{\mathrm{ct}}(\mathcal{K})$ for

all $\mathcal{K} \subseteq \{1, ..., m\}$. Note that the function $B^{\text{ct}} : \mathbb{T} \to \{1, ..., m\}$ given by $B^{\text{ct}}(t) = \overline{V}_{\text{ct}}(\mathcal{R}(t))$ is then a confidence envelope. It can be shown that $B^{\text{ct}}(t) \leq B^{\max}(t)$ for all $t \in \mathbb{T}$, i.e. it is a uniform improvement. (This follows from Goeman and Solari (2011), equation (7).) Thus an improvement of Meinshausen's method (Theorem 4.3.1) is obtained if $B_{\mathcal{I}}$ is taken to be (4.5).

In practice calculation of $\overline{V}_{\text{ct}}(\mathcal{K})$ is computationally infeasible for large m, unless shortcuts are available. This is e.g. the case when the local tests are based on Simes' probability inequality (Meijer et al., 2017), i.e. when $B_{\mathcal{I}}(t) = \#\{1 \leq i \leq \#\mathcal{I} : i\alpha/\#\mathcal{I} \leq t\}$. This parametric method is considered in Sections 4.6 and 4.7 for comparison with our nonparametric methods. When $B_{\mathcal{I}}$ is permutation-based, fast exact shortcuts for computing $\overline{V}_{\text{ct}}$ are often not available.

In the following, an iterative method is considered which is faster than closed testing. It is a conservative shortcut for the full closed testing-based method, since it provides upper bounds that are at least as large.

4.5 Iterative method

We now derive a general, iterative method for improvement of the basic confidence envelope $B^{\max}$. In each iteration step, the method uses an FDP upper bound obtained in the previous step. Some sequential family-wise error rate controlling methods, where in each step the rejections from the previous steps are used (e.g. Holm, 1979; Westfall and Young, 1993), are special cases of the procedure.

4.5.1 Exact method

Consider the setting of Definition 4.3.2. Fix $s \in \mathbb{T}$. Define the event

$$E := \bigcap_{t \in \mathbb{T}} \Big\{ V(t) \leq B_{\mathcal{N}}(t) \Big\}.$$

Assume E holds. Then $V(s) \leq B^{\max}(s)$. Consequently, there is a set $\mathcal{K} \subseteq \mathcal{R}(s)$ with $\#\mathcal{K} = R(s) - B^{\max}(s)$ such that $\mathcal{N} \subseteq \mathcal{K}^c := \{1, ..., m\} \setminus \mathcal{K}$. Thus $B_{\mathcal{K}^c} \geq B_{\mathcal{N}}$. In practice it is not known for which set $\mathcal{K}$ this holds, but we know that

$$B^1(t) := \max\{B_{\mathcal{K}^c}(t) : \mathcal{K} \in \mathcal{R}(s), \#\mathcal{K} = R(s) - B^{\max}(s)\}$$

satisfies $V(t) \leq B^1(t)$ for all $t \in \mathbb{T}$. Hence there is a set $\mathcal{K} \subseteq \mathcal{R}(s)$ with $\#\mathcal{K} = R(s) - B^1(s)$ such that $\mathcal{N} \subseteq \mathcal{K}^c$. Thus, like before

$$B^2(t) = \max\{B_{\mathcal{K}^c}(t) : \mathcal{K} \in \mathcal{R}(s), \#\mathcal{K} = R(s) - B^1(s)\}$$

satisfies $V(t) \leq B^2(t)$ for all $t \in \mathbb{T}$ under E. We similarly define B^3, B^4,... and note that $B^{\max} \geq B^1 \geq B^2 \geq ...$ and from a certain $i \in \mathbb{N}$, $B^i = B^{i+1} = ...$. In many practical situations convergence is reached after only a few steps. We write $B^{\text{it}} = \min_{i \in \mathbb{N}} B^i$. We refer to the procedure which derives B^{it} as the *iterative method*. Under E, $V(t) \leq B^{\text{it}}(t)$ for all $t \in \mathbb{T}$. Since $\mathbb{P}(E) \geq 1 - \alpha$, it follows that $B^{\text{it}}(t)$ is a confidence envelope. It can be shown that for all $t \in \mathbb{T}$

$$B^{\text{ct}}(t) \leq B^{\text{it}}(t) \leq B^{\max}(t).$$

This iterative procedure can be modified in several ways. Above, in the i-th step B^i is computed using one cut-off s. A better bound could be obtained by doing this for many $s \in \mathbb{T}$ and letting B^i be the pointwise minimum of all the improved bounds obtained. The resulting bound B^i would still be valid under E. Including such steps however may substantially increase the computational burden, so it may be better to use the method based on one cut-off s as described above.

When $B_{\mathcal{I}}$ is be defined as (4.5), we will refer to the iterative method as the *nonparametric iterative method*. This method is a uniform improvement of Meinshausen's envelope B^{m} in Section 4.3.3, if the same family $\mathbb{B}$ is used.

The nonparametric iterative procedure is much faster than the corresponding procedure based on closed testing (with the same $B_{\mathcal{I}}$, see (4.11)). However, it is often still computationally infeasible, since performing one step of this procedure requires calculating a maximum of a possibly very large set. The method is only feasible if this set is not too large, which happens for instance if $\mathcal{R}(s)$ or $B^{\max}(s)$ is small. This consideration may be used to guide the choice of s.

4.5.2 Approximation method

We suggest a method for approximating the confidence envelope B^{it}, for cases where calculating it is computationally infeasible. In the iterative method, computing any $B^i(s)$ requires determining a maximum of a potentially very large set. The approximation method computes the maximum over a smaller, random subset, to limit the computation time.

Write $\widehat{B}^0 := B^{\max}$ and for $i = 1, 2, ...$ iteratively compute $\widehat{B}^i(s) := \max\{B_{\mathcal{K}^c}(s) : \mathcal{K} \in \mathbb{K}^i\}$, where $\mathbb{K}^i$ is some large random subcollection of $\{\mathcal{K} \in \mathcal{R}(s) : \#\mathcal{K} = R(s) - \widehat{B}^{i-1}(s)\}$. Recall that if $B^i(s) = B^{i+1}(s)$, then $B^{i+1} = B^{\text{it}}$. Hence if $\widehat{B}^i(s) = \widehat{B}^{i+1}(s)$, then $\widehat{B}^{i+1}(t) = \max\{B_{\mathcal{K}^c}(t) : \mathcal{K} \in \mathbb{K}^{i+1}\}$ can be seen as an estimate of $B^{\text{it}}(t)$.

Observe that for $\#\mathbb{K}^1 \to \infty$, almost surely $\widehat{B}^1(s) \to B^1(s)$ (assuming $\mathbb{K}^1$ is uniformly sampled). Similarly, if $\#\mathbb{K}^1, ..., \#\mathbb{K}^i \to \infty$, then $\widehat{B}^i(s) \to B^i(s)$ and hence $\widehat{B}^{i+1} \to B^{i+1}$ uniformly. For finite $\#\mathbb{K}^i$, the approximation method may potentially be anti-conservative, but this was not the case in our simulation settings.

4.6 Simulations

To compare the methods of this paper, we applied them to simple simulated data. In Section 4.6.1 the performance of the iterative method as compared to the single-step method is investigated. In Section 4.6.2 the validity of the approximation method is discussed. See the data analysis in Section 4.7 for a comparison of our nonparametric methods with the parametric variants.

The simulated data matrix was the $20 \times m$-matrix $\mathbf{X} = \mathbf{X}' + \mathbf{Z}$. It can be seen as representing m measurements for 20 persons. Here $\mathbf{X}'$ is a $20 \times m$-matrix of independent normally distributed variables with variance 1. For some $0 \leq F \leq m$, in the first F columns of $\mathbf{X}$ the first 10 entries had mean 1.5 and all other entries had mean 0. The matrix $\mathbf{Z}$, which determined the correlation structure of $\mathbf{X}$, is defined by $\mathbf{Z}_{ji} := s_i Z_j$, where $s_i = 1$ for i odd and $s_i = -1$ for i even. Here each Z_j is independent and normally distributed with mean 0 and standard deviation σ_Z. For $1 \leq j \leq 20$ and $1 \leq i < i' \leq m$ note that the correlation is $\rho(\mathbf{X}_{ji}, \mathbf{X}_{ji'}) = \pm(\sigma_Z^2)/(1 + {\sigma_Z}^2)$.

For each $1 \leq i \leq m$, let H_i be the null hypothesis that $\mathbf{X}_{1,i}..., \mathbf{X}_{20,i}$ are independent and standard normally distributed. Thus the fraction of true hypotheses was $\pi_0 := (m - F)/m$. For each H_i, P_i was defined as the p-value from a two sided t-test comparing the first 10 individuals with the last 10.

As G we took all 20! permutations of cases and controls. In all the simulations we used $w = 100$, i.e. each time we drew 99 random permutations (with replacement) and added the identity. For larger w similar results are obtained (see also Marriott, 1979). We took $\alpha = 0.1$. The values of m, π_0

and $|\rho|$ are specified per case below.

4.6.1 Performance of the iterative method

We now illustrate that the nonparametric single-step method of Section 4.3.3 is improved by the corresponding iterative procedure (Section 4.5.1). We took $m = 50$ since the iterative method is not always feasible for large numbers of hypotheses. We took $\mathbb{T} = [0.001, 0.01]$. As candidate envelopes we took $B^\lambda(t) = \#\{1 \leq i \leq m : i\lambda - 0.001 \leq t\}$, $\lambda \in [0, \infty)$. In the iterative method s was taken to be 0.005. The iterative method was always terminated after three steps, when it had usually converged.

We estimated for different values of π_0 and $|\rho|$ the expected values of the FDP bounds. Above the columns the cut-offs that were used, are shown. For example, a cut-off of 0.01 means that all hypotheses with p-values smaller than 0.01 were rejected.

The results are shown in Table 4.1. For the cut-off 0.001, the upper bounds were usually zero. The improvement with the iterative method was largest when π_0 was small, i.e. when there were many false hypotheses. We expect that the relative improvement with the iterative method is larger when m is larger. The full iterative method can then be infeasible, but instead the approximation method can be used, see Sections 4.6.2 and 4.7.

A closed testing-based method using nonparametric tests would have been infeasible for $m = 50$. Hence the iterative method is the best available option in these settings.

4.6.2 Performance of the approximation method

We first compare the approximation method (Section 4.5.2) with the iterative method. This is done in the settings of Section 4.6.1 with $m = 50$. Write $\overline{FDP}_{\text{it}}(t) = B^{\text{it}}(t)/R(t)$ and let $\overline{FDP}_{\text{ap}}$ be the estimate of $\overline{FDP}_{\text{it}}$ obtained with the approximation method. Again three iteration steps were used.

We recorded the average difference between the iterative and approximate bound, $|\overline{FDP}_{\text{it}} - \overline{FDP}_{\text{ap}}|$. In each step of the approximation method 100 random combinations were used (uniformly drawn with replacement), i.e. $\#\mathbb{K}^1 = \#\mathbb{K}^2 = 100$. Despite this limited number of random combinations, the approximations were already rather good: in all settings the mean value of $|\overline{FDP}_{\text{it}} - \overline{FDP}_{\text{ap}}|$ was at most 0.0008 (results not shown).

Table 4.1: Comparison of the single-step method with the **iterative** method (in boldface). The values shown are the estimated expected values of the bounds. The values above the columns indicate the cut-offs. Each estimate is based on 1000 simulations, so that for each setting and cut-off the standard error of the mean difference between the two bounds is smaller than $9 \cdot 10^{-4}$.

		Cut-off					
		0.001		0.005		0.01	
π_0	$\lvert\rho\rvert$						
0.8	0	0.000	**0.000**	0.172	**0.170**	0.306	**0.302**
0.8	0.5	0.000	**0.000**	0.216	**0.216**	0.438	**0.435**
0.6	0	0.000	**0.000**	0.104	**0.101**	0.187	**0.179**
0.6	0.5	0.002	**0.002**	0.201	**0.198**	0.323	**0.318**
0.4	0	0.000	**0.000**	0.073	**0.067**	0.131	**0.117**
0.4	0.5	0.001	**0.001**	0.148	**0.144**	0.233	**0.228**

This means that the difference $\overline{FDP}_{\text{it}} - \overline{FDP}_{\text{ap}}$ was usually 0 and sometimes slightly larger. Naturally, when $\#\mathbb{K}^1$ and $\#\mathbb{K}^2$ were taken larger, the approximations were even better.

Note that whether $\overline{FDP}_{\text{ap}}$ closely approximates $\overline{FDP}_{\text{it}}$ is irrelevant for our purposes, as long as

$$\mathbb{P}\Big(\bigcap_{t\in\mathbb{T}}\{FDP(t) \leq \overline{FDP}_{\text{ap}}(t)\}\Big) \geq 1-\alpha.$$

This was always the case in the settings of sections 4.6.1 and in the analogous setting with $m = 1000$ (results not shown). Table 4.2 shows the improvement with the approximation method relative to the single-step method in the settings with $m = 1000$. The improvement is largest for small π_0 and $|\rho|$. It can be seen that the bounds do not always increase with the cut-off, which is due to the choice of $\mathbb{B}$ and the fact that $R(t)$ increases with $t \in \mathbb{T}$.

4.7 Data analysis

To illustrate and compare the methods in this paper, we apply them to a dataset by van de Vijver, available in the R package *cancerdata*. The dataset contains survival data on 295 cancer patients. For each individual,

Table 4.2: Comparison of the single-step method with the **approximation** method (in boldface). The values shown are the estimated expected values of the bounds. Each estimate is based on 1000 simulations, so that for each setting and cut-off the standard error of the mean difference between the two bounds is smaller than $5 \cdot 10^{-4}$.

		Cut-off					
π_0	$\lvert\rho\rvert$	0.001		0.005		0.01	
0.8	0	0.045	**0.045**	0.086	**0.082**	0.132	**0.127**
0.8	0.5	0.346	**0.344**	0.346	**0.343**	0.418	**0.414**
0.6	0	0.025	**0.022**	0.048	**0.041**	0.075	**0.064**
0.6	0.5	0.194	**0.189**	0.188	**0.182**	0.227	**0.219**
0.4	0	0.020	**0.014**	0.037	**0.026**	0.058	**0.041**
0.4	0.5	0.144	**0.137**	0.132	**0.124**	0.160	**0.150**

time to metastasis (if any), survival and the follow-up time are known. Moreover, for each individual the expression rates of 4928 genes are known (we excluded 20 genes with missing values).

We consider hypotheses H_i, $1 \leq i \leq 4928$, where H_i is the hypothesis that metastasis-free survival is not associated with the expression rate of gene i. The set G of transformations used was the collection of all 295! maps that permute (as pairs) the follow-up times and metastasis-free survival indicators of the 295 individuals. Here we took $w = 100$, i.e. we used 99 random permutations and included the original data. We have proven that our methods are valid regardless of the number of random permutations. Taking w larger leads to similar results (see also Marriott, 1979).

For each gene separately, we fitted a Cox proportional hazards model with this gene as the only covariate. We then computed a p-value for association with metastasis-free survival. The validity of the following non-parametric methods does not rely on the validity of the assumptions of the Cox model. Indeed, the p-values need not be exact as long as for each permutation they are defined in the same way.

Note that the following results are only valid under Assumption 4.2.1. This assumption says that the joint distribution of the gene expression rates corresponding to $\mathcal{N}$ (rather than just the marginals) should be independent of metastasis-free survival. This biological assumption seems reasonable, since it seems unlikely that this joint distribution is associated with metastasis-free survival if the marginals are not.

We applied eight different (variants of) methods to the data. With each method we obtained simultaneous FDP bounds. The set $\mathbb{T}$ of cut-offs is specified per case. We took $\alpha = 0.1$ such that the simultaneous bounds are valid with probability at least 90%. For three cut-offs, the bounds are shown in Table 4.3. Here the rows correspond to the methods. The first two methods are parametric and the other methods are based on permutations. We will now discuss the methods in the order of the rows of Table 4.3 and compare the results.

1. The first method used (see the first row of Table 4.3) is the parametric closed testing-based method with local tests based on Simes' probability inequality (see Section 4.4). The bounds were again obtained using the *pickSimes* function. Note that Simes' probability inequality is an assumption, which cannot be guaranteed to hold.

2. The second method is the same as the first, except that the local tests are not based on Simes' probability inequality, but on a different probability inequality (by Hommel, 1983) that always holds. Since this method uses no assumption on the dependence structure of the p-values, the bounds obtained are much larger than those from the first method.

3. Thirdly, we applied the nonparametric single-step method (Section 4.3.3), where the family $\mathbb{B}$ of candidate envelopes was based on the beta distribution as explained in Section 4.3.3. We took $\mathbb{T} = [0.001, 0.01]$. Note that these bounds are better than the bounds obtained with the two parametric methods. The reason for this is twofold. First, permutations were used such that the method took into account the dependence structure of the data. Second, bounds were not computed for all possible sets of hypotheses, but only for cut-offs in $\mathbb{T}$, causing the bounds for these cut-offs to be smaller. The nonparametric method effortlessly adapts to $\mathbb{T}$, while there is no known parametric method that does this.

4. Methods 3 and 4 are the same, except that in method 4 $\mathbb{B}$ was taken to be the family of Simes-type candidate envelopes given at (4.6). These candidate envelopes $B^\lambda(t)$ are relatively small for small cut-offs t, compared to the family based on the beta distribution. Consequently it is seen in the table that the bound for method 4 is better than that

for method 3 when the cut-off is small (0.001). When the cut-off is larger (0.01) it is the other way around.

5. Methods 4 and 5 are the same, except that in method 5 $\mathbb{T} = [0, 1]$ was taken. Since the bounds are now uniform over a larger set, naturally they are larger than those obtained with method 4 for all cut-offs in $[0.001, 0.01]$.

6. Method 6 is the same as method 5, except that in the definition of the candidate envelopes $B^\lambda(t)$ at (4.6), λi is replaced by $\lambda i - 0.001$. By comparing rows 5 and 6 in the table, it can be seen that this leads to much better (i.e. smaller) upper bounds for many cut-offs.

7. Methods 7 and 8 are variants of the approximation of the iterative method as defined in Section 4.5.2. The first step of method 7 coincides with method 4, and then additional iterative steps were performed as in Section 4.5.2 (with $s = 0.005$ and $\#\mathbb{K}^i = 1000$). Note the uniform improvement in comparison to method 4.

8. Method 8 coincides with method 7, except that the family $\mathbb{B}$ was shifted as in method 6. Compared to method 7, this improves the upper bounds for the larger cut-offs, as before.

The first conclusion to be drawn from these results, is that the a priori chosen family $\mathbb{B}$ of candidate envelopes has a large impact on the resulting confidence envelope. When Simes-type candidate envelopes are used, it can be useful to use shifted versions relative to (4.6). The second conclusion is that when $\mathbb{T}$ becomes smaller than $[0, 1]$, the bounds from the nonparametric method can improve substantially, while there is no known parametric method that adapts to $\mathbb{T}$.

Although the performance of the methods strongly relies on the family $\mathbb{B}$, it should be noted that one family of candidate envelopes cannot be uniformly better than any other. For example, for very small cut-offs (not shown) method 6 was outperformed by method 5.

Precisely because the family $\mathbb{B}$ has a large impact on the results, it should be emphasized that this set must be chosen before looking at the data. In the opposite case, the family $\mathbb{B}$ would be constructed based on the data in such a way that the results are as attractive as possible, which could induce selection bias.

Table 4.3: Comparison of eight methods. For three cut-off values, simultaneous 90%-confidence upper bounds for the FDP are shown.

		Cut-off		
Method	$\mathbb{T}$	0.001	0.005	0.01
1: Parametric (Simes)	-	0.096	0.280	0.409
2: Parametric (no Simes)	-	0.552	0.741	0.790
3: Beta	$[0.001, 0.01]$	0.076	0.101	0.125
4: Simes-type	$[0.001, 0.01]$	0.038	0.115	0.186
5: Simes-type	$[0, 1]$	0.143	0.397	0.512
6: Simes-type (shift)	$[0, 1]$	0.053	0.093	0.137
7: Iterative	$[0.001, 0.01]$	0.033	0.098	0.158
8: Iterative (shift)	$[0.001, 0.01]$	0.047	0.085	0.125
Number of rejections		449	775	957

4.8 Discussion

The multiple testing procedure by Meinshausen (2006) is a good example of an 'exploratory' method (Goeman and Solari, 2011). It offers the researcher freedom to select, based on the data, a set of hypotheses of interest and to obtain a confidence statement on these post hoc selected hypotheses. Until now it was the only permutation-based method that provides simultaneous confidence bounds for the FDP or exceedance control of the FDP.

Meinshausen's upper bounds are obtained based on the permutation distribution of all m p-values. However, it would suffice to only use the p-values under the null. Since the set of true hypotheses is not a priori known however, this cannot directly be done, and consequently Meinshausen's bounds are conservative as soon as there are any false hypotheses.

In this paper, we have shown how Meinshausen can be uniformly improved. Our methods require no additional assumptions. The largest improvement in power can be achieved with closed testing, but this approach is often computationally infeasible when permutations are used, unless there are few hypotheses. We also provide a more feasible approach in the form of an iterative method, which also improves Meinshausen. For cases where the iterative method is infeasible, we suggest an approximation of this method. The approximation method has no proven validity but performed well in

simulations. It is feasible when there are many thousands of rejected hypotheses.

In this work we discuss only p-values as test statistics, but many of the results can in principle be generalised to arbitrary test statistics. Moreover, when p-values are used, these are not required to be correct, as long as they are defined in the same way for all permuted versions of the data.

When permutations are used, our most powerful method (based on closed testing) is infeasible unless the number of hypotheses is around a dozen. The reason is that a large number of local tests need to be taken into account. When parametric local tests based on Simes' probability inequality are used, a shortcut is available, so that the method based on closed testing is often feasible. However, this method is often outperformed by the computationally feasible permutation methods, especially under dependence in the data. Thus, when permuting is an option, this is often to be preferred. In the future, shortcuts might be found for our most computationally intensive permutation methods.

5

Robust testing in generalized linear models by sign-flipping score contributions

Abstract

Generalized linear models are often misspecified due to overdispersion, heteroscedasticity and ignored nuisance variables. Existing quasi-likelihood methods for testing in misspecified models often do not provide satisfactory type-I error rate control. We provide a novel semi-parametric test, based on sign-flipping individual score contributions. This test is often robust against the mentioned forms of misspecification and provides better type-I error control than its competitors. When nuisance parameters are estimated, our basic test becomes conservative. We show how to take nuisance estimation into account to obtain an asymptotically exact test. Our proposed test is asymptotically equivalent to its parametric counterpart.

5.1 Introduction

We consider the problem of hypothesis testing in a generalized linear model (GLM) with a correct link function, but potentially misspecified distribu-

tion. When the model is misspecified, the traditional parametric tests tend to lose their properties, since they estimate the Fisher information under incorrect assumptions.

When a parametric model to be tested is potentially misspecified, the most obvious approach is to extend the model with more parameters, e.g. to add an overdispersion parameter. However, such approaches still require assumptions, for example that the overdispersion is constant. Hence a fully parametric approach is not always the best option.

Another well-known approach to testing in possibly misspecified GLMs is to use a Wald-type test, where a sandwich estimate of the variance of the coefficient estimate is used. The sandwich estimate corrects for the potentially misspecified variance. As long as the linear predictor and link are correct, such a test is asymptotically exact under mild assumptions. For small samples, however, sandwich estimates often perform poorly and the test can be very liberal (Boos, 1992; Freedman, 2006; Maas and Hox, 2004; Kauermann and Carroll, 2000).

Recent decades have seen an increase in the use of permutation approaches for various testing problems (Westfall and Young, 1993; Pesarin, 2001; Chung et al., 2013; Hemerik and Goeman, 2018). These methods are useful since they require few parametric assumptions. Especially when multiple hypotheses are tested, permutation methods are powerful since they can often take into account the dependence structure in the data. In the past, permutation methods have already been used to test in linear models (Winkler et al., 2014, and references therein). Rather than permutations, sometimes other transformations are used, such as rotations (Solari et al., 2014) and sign-flipping of residuals (Winkler et al., 2014). The existing permutation tests for GLMs, however, are limited to models with identity link function.

Like some existing methods for testing in linear models, this paper presents a sign-flipping approach. Our approach is new however, since rather than flipping residuals, we flip individual score contributions (note that the score, the derivative of the log-likelihood, is a sum of n individual score contributions). Moreover, we allow testing in a wide range of models, not only regression models with identity link. Under mild assumptions, the only requirement for the test to be asymptotically exact, is that the individual score contributions have mean 0. Consequently, if the link function is correct, our method is robust against several types of model

specification, such as arbitrary overdispersion, heteroscedasticity and, in some cases, ignored nuisance parameters.

The main reason for this robustness is that we do not require to estimate the variance of the score, the Fisher information. Rather, we perform a permutation-type test based on the score contributions, where rather than permutation, we use sign-flipping. Intuitively, the advantage of this approach over explicitly estimating the variance, is the following: if the scores are perfectly symmetric around zero, then our test is exact for small n, even if the score contributions have misspecified variances (Pesarin and Salmaso, 2010). A parametric test, on the other hand, is then usually not exact.

In case nuisance effects are estimated, the individual score contributions become dependent and our test is no longer asymptotically exact. To deal with this problem, we consider the *effective score*, which is less dependent on the nuisance estimate than the basic score (Marohn, 2002). In this case we need slightly more assumptions: the variance misspecification is not always allowed to depend on the covariates. The resulting test is asymptotically exact.

The methods in this paper have been implemented in the R package *flipscores*, available on CRAN.

In Section 5.2 we consider the scenario that no nuisance effects need to be estimated. In Section 5.3 we show how the estimation of nuisance effects can be taken into account. Section 5.4 contains simulations and Section 5.5 an analysis of real data.

5.2 Models with known nuisance parameters

Consider random variables $\nu_1, ..., \nu_n$, which satisfy Assumption 5.2.1 below. These will often be individual score contributions (see Section 5.3, Rao, 1948, or Hall and Mathiason, 1990, p. 86), but the results in Sections 5.2.1-5.2.3 are satisfied for any random variables satisfying this assumption.

Assumption 5.2.1. Assume that ν_i, $i \in \mathbb{N}$, are random variables which are independent of each other and have finite third moment. Suppose that for some $c < 1$, for every $i \in \mathbb{N}$, $\mathbb{P}(\nu_i = 0) \leq c$. Let $\sigma_i^2 = Var(\nu_i)$. Assume that as $n \to \infty$, $s_n^2 := \frac{1}{n}\sum_{i=1}^n \sigma_i^2 \to c'$ for some constant $c' \geq 0$ and $\lim_{n\to\infty}(ns_n^2)^{-1} = 0$.

Throughout Section 5.2, we consider any hypothesis H_0 which implies that $\mathbb{E}\nu_i = 0$ for all $1 \leq i \leq n$. This is in particular the case if H_0 is a point hypothesis and $\nu_1, ..., \nu_n$ are the corresponding score contributions. Generalizations to hypotheses of the form $\nu_i \geq C$, $i = 1, ..., n$, are straightforward, as well as extensions to two-sided tests.

A key assumption throughout Section 5.2 is that $\nu_1, ..., \nu_n$ are independent. As soon as nuisance parameters need to be estimated, however, score contributions become dependent. This issue is the topic of Section 5.3.

5.2.1 Basic sign-flipping test

Let $w \in \mathbb{N}_{>1}$ be the number of random sign-flippings to be used and $\alpha \in [0, 1)$. Define $g_1 = (1, ..., 1) \in \mathbb{R}^n$ and for every $2 \leq j \leq w$ let $g_j = (g_{j1}, ..., g_{jn})$ be independent and uniformly distributed on $\{-1, 1\}^n$.

Given values $T^1, ..., T^w \in \mathbb{R}$, we let $T^{(1)} \leq ... \leq T^{(w)}$ be the sorted values and write $T^{[1-\alpha]} = T^{(\lceil (1-\alpha)w \rceil)}$. We will need the following lemma.

Lemma 5.2.1. *Suppose that for $n \to \infty$, a vector $\boldsymbol{T}^n = (T^1, ..., T^w)$ converges in distribution to a vector $\boldsymbol{T}$ of i.i.d. continuous variables. Then $\mathbb{P}(T^1 > T^{[1-\alpha]}) \to \lfloor \alpha w \rfloor / w$.*

Proof. Note that $\mathbb{P}(T^1 > T^{[1-\alpha]}) = \mathbb{P}(\boldsymbol{T}^n \in A)$, where

$$A = \{(t_1, ..., t_w) \in \mathbb{R}^w : \#\{2 \leq j \leq w : t_j < t_1\} \geq \lceil (1-\alpha)w \rceil\}.$$

Note that if $t \in \partial A$, then $t_i = t_j$ for some $1 \leq i < j \leq w$. It follows that $\mathbb{P}(\boldsymbol{T} \in \partial A) = 0$. Since $\mathbb{1}_A$ is continuous on $(\partial A)^c$, it follows from the continuous mapping theorem (Van der Vaart, 1998, Theorem 2.3) that $\mathbb{1}_A(\boldsymbol{T}^n) \xrightarrow{d} \mathbb{1}_A(\boldsymbol{T})$.

The elements of $\boldsymbol{T}$ are i.i.d. draws from the same distribution. Hence it follows from the Monte Carlo testing principle (Lehmann and Romano, 2005) that under H_0, $\mathbb{P}(\mathbf{T} \in A) = \lfloor \alpha w \rfloor / w$. Thus $\mathbb{P}(\mathbf{T}^n \in A) \to \lfloor \alpha w \rfloor / w$. □

Throughout the rest of Section 5.2, for every $1 \leq j \leq w$, we let

$$T^j = n^{-1/2} \sum_{i=1}^{n} g_{ji} \nu_i.$$

We now show that the basic sign-flipping test is asymptotically exact.

Theorem 5.2.1. *Suppose that Assumption 5.2.1 holds. Consider the test that rejects H_0 if and only if $T^1 > T^{[1-\alpha]}$. Then, as $n \to \infty$, the level of this test converges to $\lfloor \alpha w \rfloor / w \leq \alpha$. Moreover, the statistics $T^1, ..., T^w$ are asymptotically normal and independent with mean 0 and common variance $\lim_{n\to\infty} s_n^2$.*

Proof. Suppose H_0 holds. We will show that $\boldsymbol{T}^n$ converges in distribution to a multivariate normal distribution with mean $\mathbf{0}$ and variance $\lim_{n\to\infty} s_n^2 \boldsymbol{I}$, where $\boldsymbol{I}$ is the $w \times w$ identity matrix. It then follows from Lemma 5.2.1 that $\mathbb{P}(T^1 > T^{[1-\alpha]}) \to \lfloor \alpha w \rfloor / w$.

Under H_0, for each $1 \leq j \leq w$, $\mathbb{E}(T^j) = 0$. For every $1 \leq j \leq w$, $var(T^j) = n^{-1} \sum_{i=1}^n var(\nu_i) = s_n^2$. Let $\boldsymbol{Q}_n$ be the covariance matrix of $\boldsymbol{T}^n$. $\boldsymbol{Q}_n$ has zeroes off the diagonal. Indeed, for $1 \leq j < k \leq w$,

$$cov(T^j, T^k) =$$

$$cov(n^{-1/2} \sum_{i=1}^n g_{ji}\nu_i, n^{-1/2} \sum_{i=1}^n g_{ki}\nu_i) = 0,$$

since the g_{ki}, $2 \leq k \leq w$, are independent with mean 0. Hence $\boldsymbol{Q}_n$ converges to $\lim_{n\to\infty} s_n \boldsymbol{I}$. Note that $\boldsymbol{T}^n$ is a sum of n vectors. By the multivariate Lindeberg-Feller central limit theorem (Greene (2012), p. 1123) $\boldsymbol{T}^n$ converges in distribution to a multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\lim_{n\to\infty} s_n^2 \boldsymbol{I}$.

We have shown that $\boldsymbol{T}^n$ converges in distribution to a vector $\boldsymbol{T}$, say, of i.i.d. normal random variables. It now follows from Lemma 5.2.1 that $\mathbb{P}(T^1 > T^{[1-\alpha]}) \to \lfloor \alpha w \rfloor / w$.

□

Note that our test does not rely on an approximate symmetry assumption (as e.g. Canay et al., 2017). Indeed, even if the scores are very skewed, asymptotically our test is exact. However, if the ν_i are symmetric, then even for small n the level is always at most α, as noted in the following proposition.

Proposition 5.2.1. *Suppose that $\nu_1, ..., \nu_n$ are independent and continuous and that under H_0, for each $i \in \mathbb{N}$, $\nu_i \overset{d}{=} -\nu_i$. Then the level of the test of Theorem 5.2.1 is at most α for any $n \in \mathbb{N}$. Moreover, if $g_2, ..., g_w$ are uniformly drawn from $\{1, -1\}^n \setminus (1, ..., 1)$ without replacement (so that only g_1 takes the value $(1, ..., 1)$), then the level is exactly $\lfloor \alpha w \rfloor / w$.*

Proof. Note that $(\nu_1, ..., \nu_n) \stackrel{d}{=} (g_{j1}\nu_1, ..., g_{jn}\nu_n)$ for every $1 \leq j \leq w$. This means that the test becomes a basic random transformation test and all results follow directly from Theorem 1 in Hemerik and Goeman (2018). □

If the g_j are drawn with replacement or the ν_i are discrete, then the level of the test of Proposition 5.2.1 is (slightly) smaller than $\lfloor \alpha w \rfloor / w$ for finite n, due to the possibility of ties among the test statistics T^j, $1 \leq j \leq w$. Otherwise the level is $\lfloor \alpha w \rfloor / w$.

Note that when the level is $\lfloor \alpha w \rfloor / w$, it can be advantageous to take w such that α is a multiple of $1/w$, to exhaust the nominal level.

5.2.2 Fixed transformations

In the test of Theorem 5.2.1, random transformations are used, rather than the whole set of $|\{-1, 1\}^n| = 2^n$ transformations. The first reason is that this will often be done in practice, especially when using all 2^n transformations would we computationally intensive (which can be the case in a multiple testing context or when n is large). Second, the fact that the number of random transformations w is fixed and $g_1, ..., g_w$ are independent (due to drawing with replacement), in fact simplified the proof.

It may be of interest to use every element of $G = \{-1, 1\}^n$ exactly once, for example to be able to obtain p-values as small as 2^{-n}. The following proposition states that the test that uses all 2^n transformations once, is asymptotically equivalent to the test of Theorem 5.2.1.

Proposition 5.2.2. *Suppose that Assumption 5.2.1 holds. Let T' be the $\lceil (1-\alpha)|G| \rceil$-th smallest value among T^g, $g \in G$, where $T^g = n^{-1/2} \sum_{i=1}^n g_i \nu_i$. Let $\phi_{n,w}$ and ϕ'_n be the rejection functions $\mathbf{1}_{\{T^1 > T^{\lceil 1-\alpha \rceil}\}}$ and $\mathbf{1}_{\{T^1 > T'\}}$, respectively. Assume that $\mathbb{P}(E_1) \to 0$ as $\epsilon^{-1}, n, w \to \infty$, where*

$$E_1 = \{T^{(\lceil (1-\alpha-\epsilon)w \rceil)} \leq T^1 \leq T^{(\lceil (1-\alpha+\epsilon)w \rceil)}\}.$$

Then, for w, $n \to \infty$, $\phi'_n - \phi_{n,w} \xrightarrow{d} 0$.

Proof. Let $0 < \epsilon < \min\{\alpha, 1-\alpha\}$. Consider the event

$$E_2 = \{T^{(\lceil (1-\alpha-\epsilon)w \rceil)} \leq T' \leq T^{(\lceil (1-\alpha+\epsilon)w \rceil)}\}.$$

It follows from the law of large numbers that $N_\epsilon, W_\epsilon \in \mathbb{N}$ exist such that for all $n > N_\epsilon, w > W_\epsilon$,

$$\mathbb{P}(E_2) > 1 - \epsilon.$$

Hence, for all $n > N_\epsilon, w > W_\epsilon$,

$$\mathbb{E}\big\{(\phi'_n - \phi_{n,w}) \cdot \mathbb{1}_{E_1^c}\big\} < \epsilon.$$

For $\epsilon^{-1}, w, n \to \infty$, $\mathbb{P}(E_1) \to 0$. It follows that for w, $n \to \infty$,

$$\mathbb{E}(\phi'_n - \phi_{n,w}) \to 0.$$

□

Note that in Theorem 5.2.1 and Proposition 5.2.2, we did not assume continuity of the observations ν_i. For finite n, continuity is needed to avoid ties among the test statistics to guarantee exactness. However, for $n \to \infty$, $\mathbb{P}(T^j = T^k) \to 0$ for any $1 \leq j < k \leq w$ regardless of the distribution of the ν_i (unless most ν_i are constant 0, which we rule out in Assumption 5.2.1). This allows using Theorem 5.2.1 for discrete GLMs.

5.2.3 Asymptotic equivalence with parametric test

Nonparametric tests such as permutation tests are often asymptotically equivalent to their corresponding parametric test (Wald and Wolfowitz, 1944; Noether, 1949). This is also the case for our sign-flipping test. This has not yet been proven under our specific assumptions, so we provide the proof here. This proof is related to the proof of Theorem 5.2.1, but here we let w increase to infinity.

Proposition 5.2.3. *Suppose that ν'_i, $i \in \mathbb{N}$, satisfy Assumption 5.2.1 and $\mathbb{E}(\nu'_i) = 0$, $i \in \mathbb{N}$. For every i, consider $\nu_i = \nu'_i + n^{-1/2}\kappa$, with $\kappa \in \mathbb{R}$. As usual, for $1 \leq j \leq w$ let*

$$T^j = n^{-1/2} \sum_{i=1}^{n} g_{ji}\nu_i.$$

Let $\phi_{n,w} = \mathbb{1}_{\{T^1 > T^{[1-\alpha]}\}}$ and $\phi'_n = \mathbb{1}_{\{T^1 > s_n\Phi(1-\alpha)\}}$, where $s_n^2 = n^{-1}\sum_{i=1}^n var(\nu_i) = var\{T^1\}$ and Φ is the cdf of the standard normal distribution. Then $\phi_{n,w} - \phi'_n \xrightarrow{d} 0$ for $w, n \to \infty$.

Proof. For $2 \leq j \leq w$, $\lim_{n\to\infty} var(T^j) =$

$$\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} var\{g_{ji}(\nu_i' + \kappa n^{-1/2})\} = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \mathbb{E}(\nu_i' + \kappa n^{-1/2})^2 =$$

$$\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \mathbb{E}(\nu_i'^2 + 2\nu_i'\kappa n^{-1/2} + \kappa^2 n^{-1}) = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} var(\nu_i) = \lim_{n\to\infty} var(T^1).$$

Hence, as in the proof in Theorem 5.2.1, $\{T^2, ..., T^w\}$ asymptotically has a multivariate normal distribution with mean $\mathbf{0}$ and variance $\lim_{n\to\infty} s_n^2 \boldsymbol{I}$, where $\boldsymbol{I}$ is the $(w-1) \times (w-1)$ identity matrix.

It is now straightforward to show that $T^{[1-\alpha]} - s_n \Phi(1-\alpha) \xrightarrow{d} 0$ for $n, w \to \infty$. Since T^1 converges to a continuous distribution, it follows that also $\phi_{n,w} - \phi_n' \xrightarrow{d} 0$ for $n, w \to \infty$. □

5.2.4 Robustness

As a main example we consider the exponential family, i.e. suppose independent variables $Y_1, ..., Y_n$ have densities of the form

$$f(y_i) = \exp\left\{\frac{y_i\eta_i - b(\eta_i)}{a_i} + c(y_i)\right\}, \tag{5.1}$$

where $\eta_i = x_i\beta + z_i\gamma$, $x_i, \beta \in \mathbb{R}$, $z_i, \gamma \in \mathbb{R}^{k-1}$. Here β is the coefficient of interest and γ (which usually includes the intercept) may need to estimated, as discussed in Section 5.3. The canonical link function g satisfies $\eta_i = g(\mu_i)$, where $g^{-1}(\eta_i) = \mu_i = \mathbb{E}(y_i) = b'(\eta_i)$ and $a_i = \text{var}(y_i)/b''(\eta_i)$ (Agresti, 2015). For $H_0 : \beta = \beta_0$, the score $\sum_{i=1}^{n} \nu_i = \frac{\partial}{\partial\beta} l(\beta|\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{y})|_{\beta=\beta_0}$ is

$$\sum_{i=1}^{n} \frac{x_i\big(y_i - b'(\eta_i)\big)}{a_i}\Big|_{\beta=\beta_0} = \sum_{i=1}^{n} \frac{x_i\big(y_i - \mathbb{E}(y_i)\big)}{a_i}\Big|_{\beta=\beta_0}. \tag{5.2}$$

For example, the Poisson model has link function $g = \log$, $b(\eta_i) = \exp(\eta_i)$, $a_i = 1$ and $c(y_i) = -\log(y_i!)$. Hence $\mathbb{E}(y_i) = b'(\eta_i) = \exp(\eta_i)$. Thus the score function is

$$\sum_{i=1}^{n} x_i\big(y_i - \mu_i\big)|_{\beta=\beta_0} = \sum_{i=1}^{n} x_i\big\{y_i - \exp(z_i\gamma)\big\}.$$

For the normal distribution, $a_i = \sigma^2$, so that the score is

$$\sum_{i=1}^{n} \frac{x_i(y_i - \eta_i)}{\sigma^2}\Big|_{\beta=\beta_0}. \tag{5.3}$$

Apart from some mild assumptions, the main assumption made in Theorem 5.2.1 is that $\mathbb{E}(\nu_i) = 0$, $i = 1, ..., n$. This is satisifed as soon as $\mu_i|_{\beta=\beta_0}$ is the true expected value of Y_i. Then the test is asymptotically exact even if the a_i are misspecified, i.e. if the variance or distributional shape of Y_i is misspecified. The a_i are even allowed to be misspecified by a factor which depends on the covariates, as long as Assumption 5.2.1 holds.

As a concrete example, consider the normal model with identity link function, which assumes that $var(Y_1) = ... = var(Y_n)$. If the real distribution is heteroscedastic, then the test will still be exact for finite n, since the ν_i are symmetric. The parametric test, however, loses its properties, since the estimated variance does not have the assumed chi-squared distribution. In Section 5.4 it is illustrated that our approach can be much more robust against heteroscedasticity than a parametric test.

Another example is the situation where the model is Poisson, i.e. $var(Y_i) = \mu_i$ is assumed, but in reality $var(Y_i) > \mu_i$, a form of overdispersion which occurs very often in practice. Then the parametric score test underestimates the Fisher information and is anti-conservative. To take the overdispersion into account it could be explicitly estimated. However, if the overdispersion factor is not constant, but depends on the covariates, then again the parametric test loses its properties. Theorem 5.2.1, however, often still applies, so that an asymptotically exact test is obtained.

Further, note that if $\mathbb{E}(Y_i)$ depends on a nuisance variable Z_i^l which is latent and ignored, where Z_i^l is independent of X_i, then the test may still be valid. The reason is that marginal over Z_i^l, $\mathbb{E}(Y_i)$ may still be computed correctly (see, for example, Section 5.4.2). Such latent nuisance variables will increase the variance of Y_i, however, which poses a problem for the parametric score test, which needs to compute the Fisher information. When the latent variable is not independent of X_i, this usually does pose a problem for our test (even as $n \to \infty$), since $\mathbb{E}(Y_i - \mu_i)$ becomes dependent on X_i under H_0.

5.3 Taking into account nuisance estimation

Consider independent and identically distributed pairs (Y_i, X_i), $i = 1, ..., n$ where $\boldsymbol{X}_i \in \mathbb{R}^d$, $d \in \mathbb{N}$, and $Y_i \in \mathbb{R}$ has distribution $\mathbb{P}_{\beta,\gamma_0,X_i}$, which depends on parameter of interest $\beta \in \mathbb{B} \subseteq \mathbb{R}$ and unknown nuisance parameter $\gamma_0 \in \mathbb{G} \subseteq \mathbb{R}^{k-1}$. We will discuss the issues arising from estimating γ_0 and propose a solution, which allows us to obtain an asymptotically exact test.

As above, we consider the null hypothesis $H_0 : \beta = \beta_0$. Generalizations to hypotheses of the form $\beta_0 \geq C$ are straightforward, as well as generalizations to two-sided tests.

Suppose that $\mathbb{P}_{\beta,\gamma,\boldsymbol{X}_i}$ has a density $f_{\beta,\gamma,\boldsymbol{X}_i}$ with respect to some dominating measure. Let γ denote an arbitrary element of $\mathbb{G}$. For $1 \leq i \leq n$ write

$$\nu_{\gamma,i} = \frac{\partial}{\partial\beta} \log f_{\beta,\gamma,\boldsymbol{X}_i}(Y_i)|_{\beta=\beta_0},$$

where we assume the derivative exists. The value ν_i is the score for the i-th observation. Under H_0, $\mathbb{E}(\nu_{\gamma_0,i}) = 0$, $i = 1, ..., n$. The score for all n observations simultaneously is $n^{1/2}S_\gamma$, where

$$S_\gamma = n^{-1/2} \sum_{i=1}^{n} \nu_{\gamma,i}.$$

Assume that $\hat{\gamma}$ is a $\sqrt{n}$-consistent estimate of γ_0. For every $1 \leq i \leq n$, let

$$\boldsymbol{\nu}_{\hat{\gamma},i}^{(k-1)} = \frac{\partial}{\partial\gamma} \log f_{\beta_0,\gamma,\boldsymbol{X}_i}(Y_i)\Big|_{\gamma=\hat{\gamma}} \in \mathbb{R}^{k-1}$$

denote the $(k-1)$-vector of score contributions for the nuisance parameters. Let

$$\boldsymbol{S}_\gamma^{(k-1)} = n^{-1/2} \sum_{i=1}^{n} \boldsymbol{\nu}_{\gamma,i}^{(k-1)} \in \mathbb{R}^{k-1}$$

be the vector of nuisance scores.

For $1 \leq j \leq w$, let the superscript j denote that g_j has been applied:

$$S_\gamma^j = n^{-1/2} \sum_{i=1}^{n} g_{ji}\nu_{\gamma,i},$$

$$\boldsymbol{S}_\gamma^{(k-1),j} = n^{-1/2} \sum_{i=1}^{n} g_{ji}\boldsymbol{\nu}_{\gamma,i}^{(k-1)}.$$

5.3.1 Asymptotically exact test

When the nuisance parameter γ_0 is unknown, it needs to be estimated, which is typically done by maximizing the likelihood of the data under the null hypothesis. Consequently, the distribution of $S_{\hat{\gamma}}$ can be substantially different than that of S_{γ_0}, the score based on the true nuisance parameters. Indeed, the asymptotic variance of $S_{\hat{\gamma}}$ is not the Fisher information, but the *effective Fisher information*, (Rippon and Rayner, 2010; Rayner, 1997; Hall and Mathiason, 1990; Marohn, 2002; Cox and Hinkley, 1979, section 9.3), which is the asymptotic variance of the *effective score*, defined below. The effective information is usually smaller than the information, given that the score for the parameter of interest and the nuisance score are correlated. Intuitively, the reason is that the nuisance variable will be used to explain part of the apparent effect of the variable of interest, also asymptotically.

The estimation of γ_0 makes the summands $\nu_{\hat{\gamma},1}, ..., \nu_{\hat{\gamma},n}$ underlying $S_{\hat{\gamma}}$ correlated, in such a way that $var(S_{\hat{\gamma}}) < var(S_{\gamma_0})$ (if the score is correlated with the nuisance score). Note however that after random flipping, the summands are not correlated anymore. This means that the variance of $S_{\hat{\gamma}}$ is asymptotically smaller than the variance of $S^j_{\hat{\gamma}}$, $2 \leq j \leq w$ (see the proof of Theorem 5.3.1). Hence, using $\nu_{\hat{\gamma},1}, ..., \nu_{\hat{\gamma},n}$ in the test of Theorem 5.2.1 can lead to a conservative test, even as $n \to \infty$.

To make the test asymptotically exact again, we would like to adapt the individual scores such that they are less dependent on the random variation of $\hat{\gamma}$. We do this by considering the so-called *effective score*, which "is 'less dependent' on the nuisance parameter than the usual score statistic" (Marohn, 2002, p. 344).

The effective score $S^*_{\hat{\gamma}}$ and the underlying summands $\nu^*_{\hat{\gamma},i}$, $i = 1, ..., n$ (which we assume have nonzero variance for $\hat{\gamma} = \gamma_0$) are defined as

$$S^*_{\hat{\gamma}} = S_{\hat{\gamma}} - \hat{\boldsymbol{\mathcal{I}}}'_{12}\hat{\boldsymbol{\mathcal{I}}}^{-1}_{22}\boldsymbol{S}^{(k-1)}_{\hat{\gamma}},$$
$$\nu^*_{\hat{\gamma},i} = \nu_{\hat{\gamma},i} - \hat{\boldsymbol{\mathcal{I}}}'_{12}\hat{\boldsymbol{\mathcal{I}}}^{-1}_{22}\boldsymbol{\nu}^{(k-1)}_{\hat{\gamma},i},$$

so that

$$S^*_{\hat{\gamma}} = n^{-1/2}\sum_{i=1}^{n}\nu^*_{\hat{\gamma},i}.$$

Here

$$\hat{\boldsymbol{\mathcal{I}}} = \begin{bmatrix} \hat{\mathcal{I}}_{11} & \hat{\boldsymbol{\mathcal{I}}}'_{12} \\ \hat{\boldsymbol{\mathcal{I}}}_{12} & \hat{\boldsymbol{\mathcal{I}}}_{22} \end{bmatrix},$$

with $\hat{\mathcal{I}}_{11} \in \mathbb{R}$ and $\hat{\boldsymbol{\mathcal{I}}}_{22} \in \mathbb{R}^{k-1}$ assumed invertible, is a consistent estimate of the population Fisher information $\boldsymbol{\mathcal{I}}$, which is the variance of $(\nu_{\gamma_0,i}, \boldsymbol{\nu}_{\gamma_0,i}^{(k-1)'})'$ marginal over $\boldsymbol{X}_i$, under H_0. In GLMs, typically $\hat{\boldsymbol{\mathcal{I}}} = n^{-1}\boldsymbol{X'}\hat{\boldsymbol{W}}\boldsymbol{X}$, where $\boldsymbol{X}$ is the design matrix and $\hat{\boldsymbol{W}}$ the estimated weight matrix (Agresti, 2015, p. 126). Further, for $1 \leq j \leq w$ we write

$$S_{\hat{\gamma}}^{*j} = S_{\hat{\gamma}}^{j} - \hat{\boldsymbol{\mathcal{I}}}'_{12}\hat{\boldsymbol{\mathcal{I}}}_{22}^{-1}\boldsymbol{S}_{\hat{\gamma}}^{(k-1),j}.$$

As discussed, $S_{\hat{\gamma}}$ is not generally asymptotically equivalent to S_{γ_0}. The effective score however is the residual from the projection of the score on the space spanned by the nuisance scores. Hence $S_{\gamma_0}^*$ is uncorrelated with the nuisance scores $\boldsymbol{S}_{\gamma_0}^{(k-1)}$ (Marohn, 2002). Consequently, as noted in the proof below, under mild regularity assumptions $S_{\hat{\gamma}}^* = S_{\gamma_0}^* + o(1)$, i.e. asymptotically the effective score it is not affected by $\hat{\gamma}$.

Note that if $\hat{\gamma}$ is the maximum likelihood estimate under H_0, then $\boldsymbol{S}_{\hat{\gamma}}^{(k-1)} = \boldsymbol{0}$, so that $S_{\hat{\gamma}}^* = S_{\hat{\gamma}}$. The summands $\nu_{\hat{\gamma},i}^*$ and $\nu_{\hat{\gamma},i}$ are different, however, and the key point is that $S_{\hat{\gamma}}^* = S_{\gamma_0}^* + o(1)$.

Like Marohn (2002), we assume that

$$S_{\hat{\gamma}} = S_{\gamma_0} - \boldsymbol{\mathcal{I}}'_{12}\sqrt{n}(\hat{\gamma} - \gamma_0) + o(1),$$

$$\boldsymbol{S}_{\hat{\gamma}}^{(k-1)} = \boldsymbol{S}_{\gamma_0}^{(k-1)} - \boldsymbol{\mathcal{I}}_{22}\sqrt{n}(\hat{\gamma} - \gamma_0) + o(1),$$

which is satisfied under mild assumptions such as continuous second order derivatives.

Theorem 5.3.1. *Consider the test of Theorem 5.2.1 with $T^j = S_{\hat{\gamma}}^{*j}$, $1 \leq j \leq w$. As $n \to \infty$, the level of this test converges to $\lfloor \alpha w \rfloor / w \leq \alpha$.*

Proof. Suppose that H_0 holds. Note that

$$S_{\hat{\gamma}}^* = S_{\hat{\gamma}} - \hat{\boldsymbol{\mathcal{I}}}'_{12}\hat{\boldsymbol{\mathcal{I}}}_{22}^{-1}\boldsymbol{S}_{\hat{\gamma}}^{(k-1)} = S_{\hat{\gamma}} - \boldsymbol{\mathcal{I}}'_{12}\boldsymbol{\mathcal{I}}_{22}^{-1}\boldsymbol{S}_{\hat{\gamma}}^{(k-1)} + o(1) =$$

$$S_{\gamma_0} - \boldsymbol{\mathcal{I}}'_{12}\sqrt{n}(\hat{\gamma} - \gamma_0) - \boldsymbol{\mathcal{I}}'_{12}\boldsymbol{\mathcal{I}}_{22}^{-1}\left\{\boldsymbol{S}_{\gamma_0}^{(k-1)} - \boldsymbol{\mathcal{I}}_{22}\sqrt{n}(\hat{\gamma} - \gamma_0)\right\} + o(1) =$$

$$S_{\gamma_0}^* + o(1).$$

Let $2 \leq j \leq w$ and

$$S_{\gamma}^{j+} = n^{-1/2} \sum_{i=1}^{n} \mathbf{1}_{\{g_{ji=1}\}} \nu_{\gamma,i},$$

$$S_{\gamma}^{j-} = n^{-1/2} \sum_{i=1}^{n} \mathbf{1}_{\{g_{ji=-1}\}} \nu_{\gamma,i}.$$

Note that

$$S_{\hat{\gamma}}^{j} = S_{\hat{\gamma}}^{j+} - S_{\hat{\gamma}}^{j-} = \left\{ S_{\gamma_0}^{j+} - \frac{1}{2}\sqrt{n}\boldsymbol{\mathcal{I}}'_{12}(\hat{\gamma}-\gamma_0) \right\} - \left\{ S_{\gamma_0}^{j-} - \frac{1}{2}\sqrt{n}\boldsymbol{\mathcal{I}}'_{12}(\hat{\gamma}-\gamma_0) \right\} + o(1) =$$

$$S_{\gamma_0}^{j+} - S_{\hat{\gamma}}^{j-} + o(1) = S_{\gamma_0}^{j} + o(1).$$

The intuitive reason why $S_{\hat{\gamma}}^{j} = S_{\gamma_0}^{j} + o(1)$, is that the estimation of $\hat{\gamma}$ does not cause the summands underlying $S_{\hat{\gamma}}^{j}$ to be correlated (as is the case for the original score $S_{\hat{\gamma}}$). Similarly we find that $\boldsymbol{S}_{\hat{\gamma}}^{(k-1),j} = \boldsymbol{S}_{\gamma_0}^{(k-1),j} + o(1)$ and conclude that $S_{\hat{\gamma}}^{*j} = S_{\gamma_0}^{*j} + o(1)$.

Let $\boldsymbol{T}^n$ be as in the proof of Theorem 5.2.1, with ν_i replaced by $\nu_{\gamma_0,i}^*$. Suppose H_0 holds and $\hat{\boldsymbol{\mathcal{I}}} = \boldsymbol{\mathcal{I}}$, so that the summands underlying $\boldsymbol{T}_j^n$ are independent. For every $1 \leq i \leq n$, $\mathbb{E}(\nu_{\gamma_0,i}^*) = 0$. The elements of $\boldsymbol{T}^n$ are uncorrelated and have common variance $var(\nu_{\gamma_0,1}^*)$. Hence, by the Lindeberg-Feller central limit theorem (Greene, 2012, p. 1123), $\boldsymbol{T}^n$ converges in distribution to $N(\mathbf{0}, var(\nu_{\gamma_0,1}^*)\boldsymbol{I})$. We supposed that $\hat{\boldsymbol{\mathcal{I}}} = \boldsymbol{\mathcal{I}}$ to use the central limite theorem, but the asymptotic distribution of $\boldsymbol{T}^n$ is the same if $\hat{\boldsymbol{\mathcal{I}}}$ is any consistent estimator of $\boldsymbol{\mathcal{I}}$.

Let $\hat{\boldsymbol{T}}^n$ be as in the proof of Theorem 5.2.1, with ν_i replaced by $\nu_{\hat{\gamma},i}^*$. For every $1 \leq j \leq w$, $S_{\hat{\gamma}}^{*j} = S_{\gamma_0}^{*j} + o(1)$. Thus $\hat{\boldsymbol{T}}^n$ and $\boldsymbol{T}^n$ are asymptotically equivalent. The result now follows from Lemma 5.2.1.

□

The test of Theorem 5.3.1 has a parametric counterpart, which uses that under H_0, $S_{\hat{\gamma}}^*$ is asymptotically normal with known variance, the effective information (Marohn, 2002; Hall and Mathiason, 1990). As $n, w \to \infty$, this test becomes equivalent to the test of Theorem 5.3.1, as the following proposition says.

Proposition 5.3.1. *Proposition 5.2.3 (asymptotic equivalence with the parametric test) still holds if for* $1 \leq j \leq w$ *we take*

$$T^j = n^{-1/2} \sum_{i=1}^{n} g_{ji} \nu^*_{\hat{\gamma},i}.$$

Proof. In case $\hat{\gamma} = \gamma_0$, the proof is analogous to that of Proposition 5.2.3. It is left to show that as $n, w \to \infty$, the test of Theorem 5.3.1 with $\hat{\gamma} = \gamma_0$ becomes equivalent to the test of Theorem 5.3.1 with $\hat{\gamma}$ any other $\sqrt{n}$-consistent estimator. This follows in a straightforward way since for w fixed, $\boldsymbol{T}^n$ and $\hat{\boldsymbol{T}}^n$ (as in the proof of Theorem 5.3.1) are asymptotically equivalent. □

5.3.2 Robustness

In Section 5.2.4 it was explained that the test of Theorem 5.2.1 is usually robust against misspecification of the variance of the score. The test of Theorem 5.3.1 is also robust against certain forms of variance misspecification. This is in particular the case if $S_{\hat{\gamma}}$ and $\boldsymbol{S}^{(k-1)}_{\hat{\gamma}}$ are misspecified by the same factor, see Proposition 5.3.2. This is for example the case if the variance misspecification factor is independent of the covariates.

Proposition 5.3.2. *Suppose that* $\hat{\boldsymbol{\mathcal{I}}} = n^{-1}\boldsymbol{X}'\hat{\boldsymbol{W}}\boldsymbol{X}$*, where* $\boldsymbol{X}$ *is an* $n \times k$ *design matrix with i.i.d. rows and* $\hat{\boldsymbol{W}}$ *a weight matrix. Consider a misspecification factor* $c_1 > 0$ *and misspecified scores*

$$\tilde{\nu}_{\hat{\gamma},i} = c_1 \nu_{\hat{\gamma},i}, \quad \tilde{\boldsymbol{\nu}}^{(k-1)}_{\hat{\gamma},i} = c_1 \boldsymbol{\nu}^{(k-1)}_{\hat{\gamma},i}, \quad i = 1, ..., n.$$

Further, for $c_2 > 0$ *consider the misspecified weight matrix* $\tilde{\boldsymbol{W}} = c_2 \hat{\boldsymbol{W}}$*. Let* $\tilde{\boldsymbol{\mathcal{I}}} = n^{-1}\boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{X}$ *be the misspecified average Fisher information. Let* $\tilde{\nu}^*_{\hat{\gamma},i} = \tilde{\nu}_{\hat{\gamma},i} - \tilde{\boldsymbol{\mathcal{I}}}'_{12}\tilde{\boldsymbol{\mathcal{I}}}^{-1}_{22}\tilde{\boldsymbol{\nu}}^{(k-1)}_{\hat{\gamma},i}$ *be the misspecified effective scores,* $i = 1, ..., n$*. Consider the test of Theorem 5.3.1, with* $S^{*j}_{\hat{\gamma}}$*,* $j = 1, ..., w$*, replaced by the misspecified effective score*

$$\tilde{S}^{*j}_{\hat{\gamma}} = n^{-1/2} \sum_{i=1}^{n} g_{ji} \tilde{\nu}^*_{\hat{\gamma},i}.$$

As $n \to \infty$*, the level of this test converges to* $\lfloor \alpha w \rfloor / w \leq \alpha$*.*

Proof. For every $1 \leq j \leq w$ we have

$$\tilde{S}^{*j}_{\hat{\gamma}} = c_1 S^j_{\hat{\gamma}} - c_2 \hat{\boldsymbol{\mathcal{I}}}'_{12} c_2^{-1} \hat{\boldsymbol{\mathcal{I}}}^{-1}_{22} c_1 \mathbf{S}^{(k-1),j}_{\hat{\gamma}} = c_1 S^{*j}_{\hat{\gamma}}.$$

Hence the test is identical to that of Theorem 5.3.1, since that test is unchanged if all T^j are multiplied by a constant. □

Proposition 5.3.2 is useful, since it tells us that if in a GLM $var(Y_i)$ is misspecified by a constant, such that $\hat{\boldsymbol{W}}$ and the scores are misspecified by a constant, the resulting test is still asymptotically exact. In Proposition 5.3.2 we assume that the misspecification factors of the weights and the scores are the same for all observations. This is satisfied for example when the model is binomial or Poisson, but the true distribution is respectively quasi-binomial or quasi-Poisson. We could slightly relax this assumption, allowing these factors to be random. Indeed, in practice the test can be very robust against heteroscedasticity (see Section 5.4). The variance misspecification is not generally allowed to depend on the covariates, since then $S_{\hat{\gamma}}$ and $\boldsymbol{S}^{(k-1)}_{\hat{\gamma}}$ can be misspecified by different factors asymptotically. There are exceptions however, see Sections 5.3.3 and 5.4.1.

The test based on the basic scores $\nu_{\hat{\gamma},i}$ is valid as soon as $\mathbb{E}(\nu_{\hat{\gamma},i}) = 0$, and hence it may sometimes be preferred over the test of Theorem 5.3.1. The test based on the basic score is asymptotically conservative if the score S_{γ_0} is correlated with the nuisance scores $\boldsymbol{S}^{(k-1)}_{\gamma_0}$, i.e. when $\boldsymbol{\mathcal{I}}_{12} \neq 0$. Hence it can be useful to redefine the covariate of interest such that $\boldsymbol{\mathcal{I}}_{12} = 0$. When $\hat{\boldsymbol{W}} = b\boldsymbol{I}$, $b > 0$, this means orthogonalizing the covariate of interest with respect to the nuisance covariates. When the model is potentially misspecified, then the weights and hence $\boldsymbol{\mathcal{I}}_{12}$ are not asymptotically known, but the user could substitute a best guess for the weights.

5.3.3 An example

As discussed, the test of Theorem 5.3.1 is often not asymptotically exact if the variance misspecification depends on the covariates. An important exception is the case where the model is

$$Y_i \sim N(\gamma_0 + \beta X'_i, \sigma^2) \quad i = 1, ..., n, \tag{5.4}$$

where γ_0 is the unknown intercept and $X'_i \in \mathbb{R}$. If the null hypotesis is $H_0 : \beta = \beta_0$, then γ_0 is a nuisance parameter that needs to be estimated.

(We do not need to know σ and can simply substitute 1 for it.) Hence, we compute the effective score. Note that for $1 \leq i \leq n$,

$$\nu_{\hat{\gamma},i} = x_i'(y_i - \hat{\mu}_i)/\sigma^2,$$
$$\boldsymbol{\nu}_{\hat{\gamma},i}^{(k-1)} = (y_i - \hat{\mu}_i)/\sigma^2.$$

Note that we can consistently estimate $\boldsymbol{\mathcal{I}}_{12}\boldsymbol{\mathcal{I}}_{22}^{-1}$ by $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i'$, so that the effective score contributions are

$$\nu_{\hat{\gamma},i}^{*} = (x_i' - \overline{x})(y_i - \hat{\mu}_i)/\sigma^2.$$

Thus, the effective score contributions are exactly the basic score contributions after centering $x_1', ..., x_n'$ around 0. Similarly, if $x_1', ..., x_n'$ are already centered, then $\nu_{\hat{\gamma},i}$ and $\nu_{\hat{\gamma},i}^{*}$ coincide, since then $\hat{\boldsymbol{\mathcal{I}}}_{12} = 0$.

The test of Theorem 5.3.1 is not always asymptotically exact if $S_{\hat{\gamma}}$ and $\boldsymbol{S}_{\hat{\gamma}}^{(k-1)}$ are misspecified by different factors. However, if $\hat{\boldsymbol{\mathcal{I}}}_{12} = 0$, then this does not apply anymore. The test of Theorem 5.3.1 then remains asymptotically exact and reduces to the test based on the basic score. For the model (5.4), this means that even if the misspecification of $var(Y_i)$ depends on X_i', we obtain an asymptotically exact test.

A particular case where this principle applies is the generalized Behrens-Fisher problem, where the aim is to test equality of the means μ^1 and μ^2 of two populations (or to test if $\mu^1 \leq \mu^2$ or $\mu^1 \geq \mu^2$). In this problem, it is only assumed that two independent samples from these populations are available, without making other assumptions such as equal variances. It is well-known that this problem has no exact solution under normality (Pesarin and Salmaso, 2010; Lehmann and Romano, 2005). Under mild assumptions, we obtain an asymptotically exact test for this problem. Pesarin and Salmaso (2010) already suggested sign-flipping residuals to solve this problem. This is equivalent to flipping scores in our linear model (5.4) with $|x_1'| = ... = |x_n'|$.

5.4 Simulations

To compare the tests in this paper with each other and existing tests, we applied them to simulated data. In particular we considered scenarios where the model was misspecified.

5.4.1 Overdispersion, heteroscedasticity and estimated nuisance

In Sections 5.4.1 and 5.4.2 the assumed model was Poisson, but in fact $Y_1, ..., Y_n$ were drawn from a negative binomial distribution.

The covariates $X, Z, Z^l \in \mathbb{R}$ were drawn from a multivariate normal distribution with zero mean and $var(X) = var(Z) = var(Z^l) = 1$. (For nonzero means, similar simulation results were obtained as below.) The response satisfied $\log(\mathbb{E}(Y_i)) = \log(\mu_i) = \eta_i =$

$$0 + \beta \cdot X_i + \gamma_0 \cdot Z_i + \gamma_0^l \cdot Z_i^l.$$

The null hypothesis was $H_0 : \beta = 0$. In Section 5.4.1 we took $\gamma^l = 0$. The coefficient γ_0 and the intercept 0 were nuisance parameters that were estimated by maximum likelihood under H_0. We took $\gamma_0 = 1$ and $\rho(X_i, Z_i) = 0.5$, $\rho(Z_i^l, Z_i) = 0$, $\rho(Z_i^l, X_i) = 0$. We took the dispersion parameter of the negative binomial distribution to be 1, so that $var(Y_i) = \mu_i + \mu_i^2$.

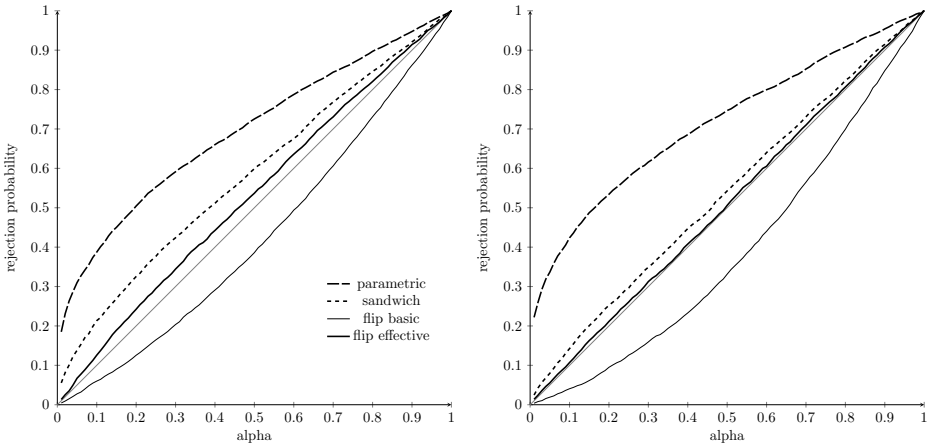
The assumed model, however, was Poisson, i.e. $var(Y_i) = \mu_i$ was assumed. Thus the true variance was larger than the assumed variance and the variance misspecification factor depended on μ_i, i.e. on the covariate Z_i. The assumed log link function was correct and in Section 5.4.1 the linear predictor was correct as well.

In Figure 5.4.1 the estimated rejection probabilites of four tests under $H_0 : \beta = 0$ are compared, based on 5000 repeated simulations. In all simulations the tests were two-sided.

One of the tests considered was the parametric score test. Since the assumed model was Poisson, the computed Fisher information was too small and the test was anti-conservative.

We also applied a Wald test, where we used a sandwich estimate (Agresti, 2015, p. 280) of the variance of $\hat{\beta}$, to correct for the misspecified variance function. We used the R package *gee* for this (available on CRAN), specifying blocks of size 1. As can be seen in Figure 5.4.1, this test was anti-conservative for small n. This was in particular due to the estimation error of the sandwich (Boos, 1992; Freedman, 2006; Maas and Hox, 2004; Kauermann and Carroll, 2000).

Further, we applied the sign-flipping test based on the basic scores $\nu_{\hat{\gamma},i}$. We took $w = 200$ (taking w larger gives similar results, see also Marriott,



(a) $\beta = 0$, $n = 50$ (b) $\beta = 0$, $n = 200$

Figure 5.1: Estimated rejection probabilities for four tests under misspecified variance and estimated nuisance. The null hypothesis was $H_0 : \beta = 0$.

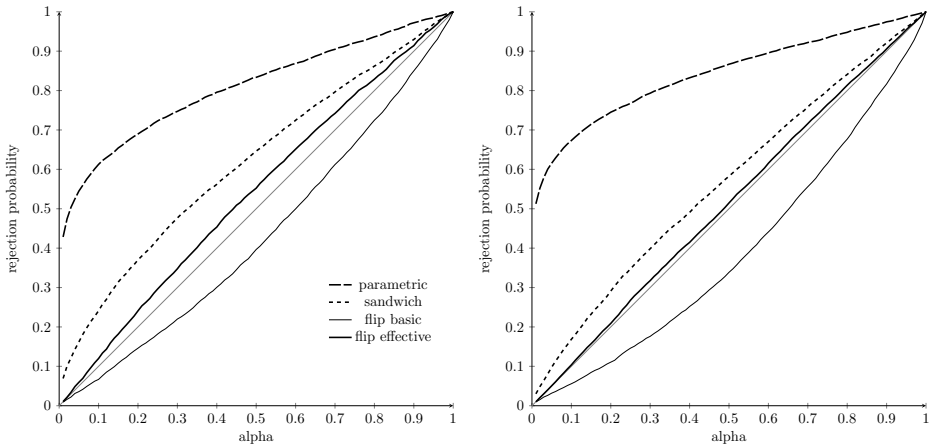
1979). Due to the estimation of γ_0, the variance of the score was shrunk and the test was conservative, as explained in Section 5.3.1.

Finally, we used the sign-flipping test of Theorem 5.3.1, which is based on the effective scores $\nu^*_{\hat{\gamma},i}$. In Section 5.3.2 it was already shown that this test is asymptotically exact under constant variance misspecification. Here, however, the variance misspecification factor was $1 + \mu_i$ (i.e. it depended on Z_i). Nevertheless the level of the test was approximately α. This illustrates that the test has some additional robustness, which we have not theoretically shown.

5.4.2 Ignored nuisance

The same simulations were performed as in Section 5.4.1, but with $\gamma^l_0 = 1$. Since $\gamma^l_0 = 0$ was assumed, Z^l_i represented an ignored, latent variable. Figure 5.4.2 shows similar results as Figure 5.4.1. The parametric test was even more anti-conservative than in Section 5.4.1. The reason is that the introduction of Z^l_i increased the variance Y_i, so that the variance of the score was even more misspecified than in Section 5.4.1.

The test of Theorem 5.3.1 was still nearly exact for $n = 200$, even though μ_i was misspecified. (Even marginally over Z^l_i, μ_i was misspecified.



(a) $n = 50, \beta = 0$ (b) $n = 200, \beta = 0$

Figure 5.2: Estimated rejection probabilities for four tests under misspecified variance, estimated nuisance and ignored nuisance. The null hypothesis was $H_0 : \beta = 0$.

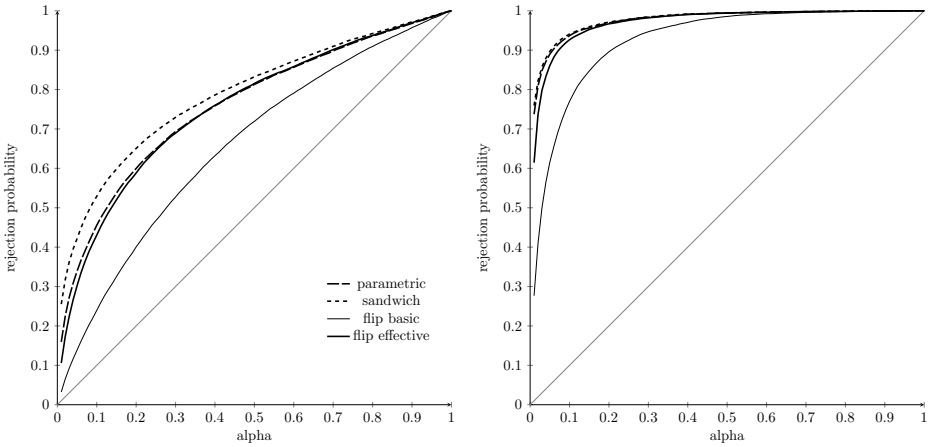
Possibly the estimation of the intercept corrected for the misspecification.)

A conclusion from the simulations of Sections 5.4.1 and 5.4.2, is that the sandwich-based approach should not always be seen as the most reliable way of testing models with misspecified variance functions. Indeed, in our simulations the test of Theorem 5.3.1 was substantially less anti-conservative (while having similar power, see Section 5.4.3).

5.4.3 Power

For a meaningful power comparison of the four tests, we considered the scenario where the assumed model was correct, i.e. the data distribution was Poisson and γ_0^l was 0. See figure 5.4.3. The estimated probabilities are based on 20,000 simulation loops.

Since the model was correct, asymptotically there was no better choice than the parametric test. The sign-flipping test of Theorem 5.3.1 had very similar power. The basic sign-flipping test was again conservative due to the estimation of γ_0. The sandwich-based test had the most power, but was anti-conservative (null behavior not shown).



(a) $\beta = 0.2$, $n = 50$ (b) $\beta = 0.2$, $n = 200$

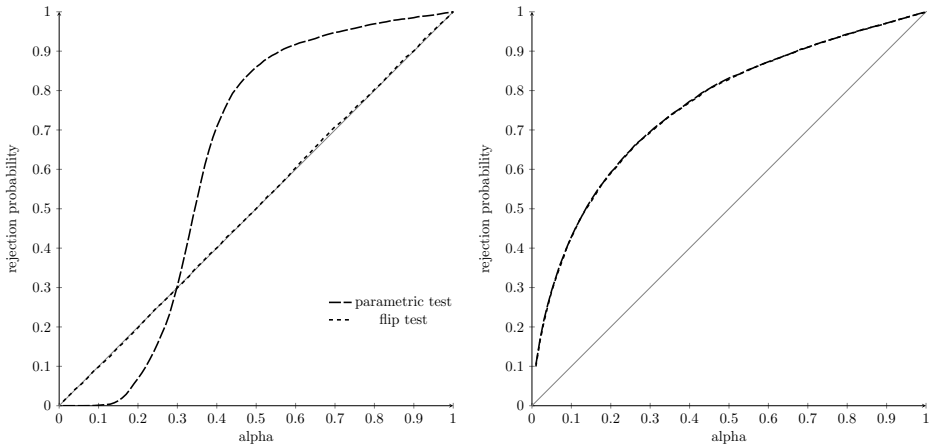
Figure 5.3: Power comparison of four two-sided tests under the correct model, with estimated nuisance. The null hypothesis was $H_0 : \beta = 0$.

5.4.4 Strong heteroscedasticity

When a Gaussian linear model is considered with $Y_i \sim N(\beta x_i, \sigma^2)$, $x_1 = ... = x_n = 1$ and $H_0 : \beta = 0$, the score contributions are $\nu_i = X_i(Y_i - 0)/\sigma^2 = Y_i/\sigma^2$. Thus the test of Theorem 5.2.1 simply flips the observations Y_i, $1 \leq i \leq n$. The parametric counterpart of this test is the one-sample t-test. The t-test needs to explicitly estimate the nuisance parameter σ^2; the sign-flipping test does not (simply substitute $\sigma = 1$).

We simulated strongly heteroscedastic data: we took $Y_i \sim N(\beta x_i, \sigma_i^2)$, with $\sigma_i = \exp(i)$, $1 \leq i \leq n = 10$. Consequently the t-statistic did not have the assumed distribution and the level of the t-test was far from nominal for most α, see Figure 5.4a. The sign-flipping test did not need to estimate the variance. In this setting the test has level $\lfloor \alpha w \rfloor / w$ exactly if the transformations $g_1, ..., g_w$ are drawn without replacement, since the observations are symmetric, see Proposition 5.2.1. (We drew $g_1, ..., g_w$ with replacement for convenience, but this gives almost the same test as drawing without replacement, due to the small probability of ties.)

For a meaningful power comparison, we considered the correct, homoscedastic model with $\sigma_1 = ... = \sigma_{10} = 1$. Figure 5.4b, based on 10^5 repeated simulations, shows that the tests had virtually the same power.



(a) $\mu = 0$, strong heteroscedasticity (b) $\mu = 0.5$, correct model

Figure 5.4: Comparison of the one-sample t-test and the sign-flipping test. The null hypothesis was $H_0 : \mu = 0$.

5.5 Data analysis

We analyzed the dataset *warpbreaks*. These data are used in the example code of the *gee* R package, available on CRAN. The dataset gives the number of warp breaks per loom, where a loom corresponds to a fixed length of yarn. There are 54 observations of 3 variables: the number of breaks, the type of wool (A or B) and the tension (low, medium or high). For each of the 6 possible combinations of wool and tension, there are 9 observations. Using various methods, we tested whether the number of breaks depends on the type of wool.

We first considered a basic Poisson model with

$$\log(\mu_i) = \gamma_1 + \beta \mathbf{1}_{\{\text{wool}=B\}} + \gamma_2 \mathbf{1}_{\{\text{tension}=\text{M}\}} + \gamma_3 \mathbf{1}_{\{\text{tension}=\text{H}\}}.$$

The γ_i, $1 \leq i \leq 3$, were nuisance parameters that were estimated using maximum likelihood. We first tested $H_0 : \beta = 0$ using the parametric score test, obtaining a p-value of $6.29 \cdot 10^{-5}$. (All tests performed were two-sided.)

However, the data were clearly overdispersed: for each combination of wool and tension, the empirical variance of the 9 observations was substantially larger than the empirical mean. Thus the p-value based on the parametric test had limited meaning. Fitting a quasi-Poisson model, which

assumes constant overdispersion, gave a p-value of 0.059.

As in Section 5.4, we also applied a Wald test, where we used a sandwich estimate (Agresti, 2015, p. 280) of the variance of $\hat{\beta}$, to correct for the misspecified variance function. This resulted in a p-value of 0.048.

Further, we used the sign-flipping test based on the basic scores $\nu_{\hat{\gamma},i}$, $i = 1, ..., 54$ (still using the basic Poisson model). We took $w = 10^6$. This resulted in a p-value of 0.113. This test is rather robust to model misspecification, but we know that it tends to be conservative when the score is correlated with the nuisance scores, as was the case here.

Finally, we performed the test of Theorem 5.3.1 based on the effective score. This test is asymptotically exact under the correct model and has been shown to be robust against several forms of variance misspecification. It provided a p-value of 0.065.

Based on this evidence, when maintaining a confidence level of 0.05, it seems that we cannot reject H_0. Indeed, only the sandwich-based test provided a p-value below 0.05, but this test is often anti-conservative, as discussed in Section 5.4.1.

5.6 Discussion

We have proposed a test which relies on the assumption that individual score distributions are independent and have mean 0 (in case of a point hypothesis) under the null. If the score contributions are misspecified due to overdispersion, heteroscedasticity or ignored nuisance covariates, then the traditional parametric tests lose their properties. The sign-flipping test is robust to these types of misspecification and is often still asymptotically exact. It often has similar power to the parametric score test, if the model is correct.

When nuisance parameters are estimated, the score contributions become dependent. If a nuisance score is correlated with the score of the parameter of interest, the estimation reduces the variance of the score, so that the sign-flipping test becomes conservative. As a solution we propose to use the effective score, which is the part of the score that is orthogonal to the nuisance score. The effective score is asymptotically unaffected by the nuisance estimation, so that we again obtain an asymptotically exact test. We have proven that this is still the case under constant overdispersion, and simulations illustrate additional robustness.

Despite its potential conservativeness, the test based on the basic score may be preferred over that based on the effective score, for its robustness. Moreover, the basic test does not require computing the Fisher information. This might provide an advantage in applications where this computation is time-consuming.

A

Manual of the R package *confSAM*

Manual of the R package *confSAM*

Contents

Introduction

This is a manual for the R package *confSAM* by Jesse Hemerik and Jelle Goeman (2017), which is available on CRAN. This package provides confidence bounds for the false discovery proportion in the context of SAM (Tusher et al., 2001). If you use the *confSAM* package, please cite the paper Hemerik and Goeman (2018), which provides all underlying theory.

The false discovery proportion

Suppose hypotheses $H_1, ..., H_m$ are tested by calculating corresponding p-values $p_1, .., p_m$ and rejecting the hypotheses with small p-values. The number of false positive findings is then the number of hypotheses that are rejected even though they are true. The False Discovery Proportion (FDP) is this number divided by the total number of rejected hypotheses.

Instead of calculating a p-value for each hypothesis, it is also possible to calculate other test statistics, $T_1, ..., T_m$, say. One could then reject all hypotheses with test statistics e.g. exceeding some constant. (Possibly with a different constant for each hypothesis.)

In multiple testing it is often of interest to estimate how many of the rejected hypotheses are false findings. This is equivalent to estimating the FDP. The package *confSAM* allows estimation of this quantity.

As is usually the case with estimating quantities, providing a point estimate is not enough. What is also important is providing a confidence interval, so that one has e.g. 95% confidence that the quantity of interest lies in the interval. The package *confSAM* allows not only estimating the FDP, but also providing a confidence interval for it. More precisely, the

package provides an confidence upperbound for the FDP, so that the user has e.g. 95% confidence that the FDP is between zero and this bound.

The package *confSAM* incorporates different methods for providing estimates and upper bounds. The methods vary in complexity and computational intensity. In the following it is explained how these methods can be used with the function `confSAM`.

Use of permutations

The methods in this package can be used if the joint distribution of the part of the data that is under the null, is invariant under a group of permutations. For example, suppose that each test statistic T_i, $1 \leq i \leq m$, depends on some n-dimensional vector of obervations. Suppose for example that such a vector contains n gene expression level measurements: $n/2$ from cases and $n/2$ from controls. If the joint distribution of the gene expression levels corresponding to the true hypotheses is the same for cases and controls, then permuting the cases and controls does not change this joint distribution. In that case the methods in this package can be used.

Designs with more than two groups or other transformations than permutations are also possible. See Hemerik and Goeman (2018) for general theory.

Basic estimate and bound

For the function `confSAM`, essentially the only input required is a matrix of p-values (or other test statistics). Every row of the matrix should correspond to a (random) permutation of the data. For the assumption that these p-values (test statistics) should satisfy, see Hemerik and Goeman (2018), Assumption 1.

Obtaining the matrix of test statistics

The *samr* package contains a function `samr` that allows computation of the test statistics as defined in the original SAM paper (Tusher et al., 2001). More precisely, the object `tt` that `samr` returns, contains test statistics for the original data. Further, the object `ttstar0` contains a matrix of (unsorted) test statistics for the permuted version of the data. These objects can be used as input for our function `confSAM` (`ttstar0` should first be transposed).

Here we will not use *samr* to compute test statistics, but compute test statistics ourselves. As example data to work with, we consider the *nki70* dataset from the *penalized* package.

```
library(penalized)
data(nki70)
```

This survival data set concerns 144 lymph node positive breast cancer patients. For each patient there is a time variable and an event indicator variable (metastasis-free survival), as well as 70 gene expression level measurements. Using *confSAM* we will test the hypotheses

$H_1, ..., H_{70}$ where H_i is the hypothesis that the expression level of gene i is not associated with the survival curve.

To be able to use `confSAM`, we now construct the required matrix of p-values. We will use random permutations, i.e. random reshufflings of the 144 vectors of gene expression levels. Hence we first set the seed.

```
library(survival)
set.seed(21983)
w<-100 # number of random permutations
pvalues <- matrix(nr=w,nc=70)
survobj <- Surv(time=nki70$time, event=nki70$event)



#compute the 70 p-values for each random permutation
for(j in 1:w){
  if(j==1){
    permdata <- nki70 #original data
  }
  else{
    permdata <- nki70[sample(nrow(nki70)),] #randomly shuffle the rows
  }
  for (i in 1:70) {
    form <- as.formula(paste("survobj ~ ", names(nki70)[i+7] ))
    coxobj <- coxph(form, data=permdata)
    sumcoxobj <- summary(coxobj)
    pvalues[j,i] <- sumcoxobj$coefficients[,5]
  }
}
```

We took the first permutation to be the original group labelling. Hence pvalues[1,] contains the p-values for the original data. Note that it is possible that the model assumptions are not exactly satisfied and the p-values corresponding to true hypotheses are not exactly uniform. A good property of the methods in *confSAM* however is that it does not matter if the p-values are exact for the methods to be valid (as long as they are computed in the same way for each permutation). The reason is that any test statistics are allowed.

Now that we have our p-value matrix, we are ready to use `confSAM`. Let us say that we will reject all hypotheses with p-values smaller than 0.03. Recall that `pvalues[1,]` contains the p-values for the original data. Thus the number of rejected hypotheses is found as follows:

```
sum(pvalues[1,]<0.03)
```

```
## [1] 17
```

The number of rejections is also automatically given by the function `confSAM`, as we will see below.

Basic confidence bound

We now turn to computing a basic confidence bound for the number of false positives (or equivalently, the FDP). By Hemerik and Goeman (2018) a basic $(1-\alpha)$-confidence bound for the number of false positives V is the $(1-\alpha)$-quantile of the numbers of rejections for the permuted versions of the data. This basic upper bound is obtained as follows (for $\alpha = 0.1$):

```
library(confSAM)
confSAM(p=pvalues[1,], PM=pvalues, cutoff=0.03, alpha=0.1,
        method="simple")[3]
```

```
## Simple conf. bound for #fp:
##                           5
```

Here

- the argument *p* is the vector of *p*-values for the original unpermuted data;
- the argument *PM* is the matrix of *p*-values (or other test statistics);
- *cutoff* is the cut-off we have chosen. (*cutoff* is also allowed to be a vector of length `length(p)`, in which case the *i*-th *p*-value is compared to `cutoff[i]`.)
- *alpha* is such that $1-\alpha$ is the desired confidence level;
- *method="simple"* means that we want a simple upper bound (which is computationally fastest).

In our case, we made sure that the first row of the *p*-value matrix contained the original *p*-values. If we had not explicitly put the original *p*-values in the matrix (in which case they should be in the first row), then we should have included *includes.id=FALSE* in the function call.

Another argument that can be given to the function is *reject*, which can take the values *"small"*, *"large"* and *"absolute"*. This argument, together with the cutoff, determines which hypotheses are rejected. For example, the default is *"small"* and this means that all hypotheses with *p*-values (or test statistics) smaller than *cutoff* should be rejected. Setting *reject* to *"absolute"* means that all test statistics with absolute value greater than *cutoff* are rejected.

The above function call provides an upper bound for the number of false positives. Since we took $\alpha = 0.1$, we have 90% confidence that the number of false positives does not exceed this bound. To obtain a 90%-confidence upper bound for the false discovery proportion (FDP), we simply divide the above bound by the number of rejections.

Note that in the above we put '`[3]`' behind the function call. Leaving this out gives the full output of the function:

```
confSAM(p=pvalues[1,], PM=pvalues, cutoff=0.03, alpha=0.1, method="simple")
```

```
##                #rejections:     Simple estimate of #fp:
##                          17                           2
## Simple conf. bound for #fp:
##                           5
```

The first argument is the number of rejections, which we already computed above. The second argument is a basic median unbiased estimate of the number of false positives. (This is just the basic upper bound for $\alpha = 0.5$.) The function always produces these automatically since they require little additional computation time.

Improved upper bounds

Closed testing-based bound

Beside the basic upper bound discussed above, also more sophisticated upper bounds are derived in Hemerik and Goeman (2018). The first one we will discuss is related to the theory of closed testing. This method is the most computationally demanding, and is often infeasible when there are many hypotheses, but for the present example the method is feasible. The full closed testing-based method provides an upper bound that is often smaller than the basic upper bound discussed above, while still providing $(1 - \alpha)100\%$ confidence. Details on how the bound is defined are in Hemerik and Goeman (2018).

To use this method we change the *method* argument into *full*. Running the following code may take a while so it might be skipped.

```
confSAM(p=pvalues[1,], PM=pvalues, cutoff=0.03, alpha=0.1, method="full")
```

```
##                     #rejections:         Simple estimate of #fp:
##                               17                               2
## cl.testing-based bound for #fp:
##                                4
```

Note that the resulting upper bound is smaller than the bound obtained with method="simple". In cases where the closed-testing method is infeasible, in some cases with many false hypotheses the basic bound can still be improved by setting *method="csc"*. In that case an exact, but conservative shortcut for the full closed testing-based method is used. A method that is more likely to improve the basic bound however, is the approximation method discussed below.

Approximation method

Setting *method="full"* can lead to computational infeasibility. In that case a possible solution is to approximate this bound. This is done by setting *method="approx"*. This corresponds to the approximation method detailed in Hemerik and Goeman (2018).

For example, in our example increasing the cut-off leads to more rejections, and it is a property of the method that it then tends to take longer or can be infeasible. To limit the computational burden, the approximation method can be used. We now increase the cut-off to 0.2 and illustrate the approximation method. For comparison we also compute the simple bound explained above.

```
#simple method
confSAM(p=pvalues[1,], PM=pvalues, cutoff=0.2, alpha=0.1, method="simple")
```

```
##                #rejections:     Simple estimate of #fp:
##                          40                          12
## Simple conf. bound for #fp:
##                          22
```

```
#approximation method
confSAM(p=pvalues[1,], PM=pvalues, cutoff=0.2, alpha=0.1,
        method="approx", ncombs=1000)[3]
```

```
## Appr. cl.testing-based bound for #fp:
##                                    17
```

The last result again means that with 90% confidence, the number of false positives does not exceed the stated bound. Note that the bound obtained with the approximation method is smaller than the simple bound.

Note above that an additional argument *ncombs* appears (default is 1000). This is the number of combinations that the approximation method checks per step. Details are in Hemerik and Goeman (2018). The higher *ncombs* is, the more reliable the bound is as a $(1-\alpha)$-confidence bound, but the longer the computation takes. We recommend taking *ncombs* very high ($> 10^4$) if time allows it.

Bibliography

Agresti, A. (2015). *Foundations of linear and generalized linear models*. John Wiley & Sons.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, 289–300.

Bennett, C. M., M. Miller, and G. Wolford (2009). Neural correlates of interspecies perspective taking in the post-mortem atlantic salmon: An argument for multiple comparisons correction. *Neuroimage 47*(Suppl 1), S125.

Blanchard, G., P. Neuvial, and E. Roquain (2017). Post hoc inference via joint family-wise error rate control.

Blanchard, G., E. Roquain, et al. (2008). Two simple sufficient conditions for FDR control. *Electronic journal of Statistics 2*, 963–992.

Boos, D. D. (1992). On generalized score tests. *The American Statistician 46*, 327–333.

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

Cai, G. and S. K. Sarkar (2008). Modified Simes critical values under independence. *Statistics & Probability Letters 78*(12), 1362–1368.

Canay, I. A., J. P. Romano, and A. M. Shaikh (2017). Randomization tests under an approximate symmetry assumption. *Econometrica 85*(3), 1013–1030.

Chu, G., J. Li, B. Narasimhan, R. Tibshirani, and V. Tusher (2001). Significance analysis of microarrays users guide and technical document.

Chung, E., J. P. Romano, et al. (2013). Exact and asymptotically robust permutation tests. *The Annals of Statistics 41*(2), 484–507.

Cox, D. D. and J. S. Lee (2008). Pointwise testing with functional data using the westfall–young randomization method. *Biometrika 95*(3), 621–634.

Cox, D. R. and D. V. Hinkley (1979). *Theoretical statistics*. CRC Press.

Cuevas, A. (2014). A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference 147*, 1–23.

Dudoit, S., J. P. Shaffer, and J. C. Boldrick (2002). Multiple hypothesis testing in microarray experiments. Technical report. Available at http://www.bepress.com/ucbbiostat/paper110/.

Dudoit, S., J. P. Shaffer, and J. C. Boldrick (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, 71–103.

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

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

Farcomeni, A. (2009). Generalized augmentation to control the false discovery exceedance in multiple testing. *Scandinavian Journal of Statistics 36*(3), 501–517.

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

Freedman, D. A. (2006). On the so-called huber sandwich estimator and robust standard errors. *The American Statistician 60*(4), 299–302.

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

Genovese, C. and L. Wasserman (2004). A stochastic process approach to false discovery control. *Annals of Statistics*, 1035–1061.

Genovese, C. R. and L. Wasserman (2006). Exceedance control of the false discovery proportion. *Journal of the American Statistical Association 101*(476), 1408–1417.

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

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

Goia, A. and P. Vieu (2016). An introduction to recent advances in high/infinite dimensional statistics.

Gou, J. and A. C. Tamhane (2014). On generalized Simes critical constants. *Biometrical Journal 56*(6), 1035–1054.

Greene, W. H. (2012). Econometric analysis, harlow.

Guo, W., L. He, S. K. Sarkar, et al. (2014). Further results on controlling the false discovery proportion. *The Annals of Statistics 42*(3), 1070–1101.

Hall, W. and D. J. Mathiason (1990). On large-sample estimation and testing in parametric models. *International Statistical Review/Revue Internationale de Statistique*, 77–97.

Hemerik, J. and J. Goeman (2017). Exact testing with random permutations. *TEST (Online First version)*.

Hemerik, J. and J. J. Goeman (2018). False discovery proportion estimation by permutations: confidence for significance analysis of microarrays. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 80*(1), 137–155.

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

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, 65–70.

Hommel, G. (1983). Tests of the overall hypothesis for arbitrary dependence structures. *Biometrische Zeitschrift 25*(5), 423–430.

Kauermann, G. and R. J. Carroll (2000). The sandwich variance estimator: Efficiency properties and coverage probability of confidence intervals.

Kim, K. I. and M. A. van de Wiel (2008). Effects of dependence in high-dimensional multiple testing problems. *BMC bioinformatics 9*(1), 114.

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

Korn, E. L., M.-C. Li, L. M. McShane, and R. Simon (2007). An investigation of two multivariate permutation methods for controlling the false discovery proportion. *Statistics in medicine 26*(24), 4428–4440.

Korn, E. L., J. F. Troendle, L. M. McShane, and R. Simon (2004). Controlling the number of false discoveries: application to high-dimensional genomic data. *Journal of Statistical Planning and Inference 124*(2), 379–398.

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

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

Lehmann, E. L. and J. P. Romano (2012). Generalizations of the familywise error rate. In *Selected Works of EL Lehmann*, pp. 719–735. Springer.

Maas, C. J. and J. J. Hox (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica 58*(2), 127–137.

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

Marohn, F. (2002). A comment on locally most powerful tests in the presence of nuisance parameters. *Communications in Statistics-Theory and Methods 31*(3), 337–349.

Marriott, F. (1979). Barnard's Monte Carlo tests: How many simulations? *Applied Statistics*, 75–77.

Meijer, R., T. Krebs, A. Solari, and J. Goeman (2017). Simultaneous control of all false discovery proportions by an extension of Hommel's method. *arXiv preprint arXiv:1611.06739v2*.

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

Meinshausen, N. and P. Bühlmann (2005). Lower bounds for the number of false null hypotheses for multiple testing of associations under general dependence structures. *Biometrika 92*(4), 893–907.

Noether, G. E. (1949). On a theorem by Wald and Wolfowitz. *The Annals of Mathematical Statistics 20*(3), 455–458.

Pesarin, F. (2001). *Multivariate permutation tests: with applications in biostatistics*, Volume 240. Wiley Chichester.

Pesarin, F. (2015). Some elementary theory of permutation tests. *Communications in Statistics-Theory and Methods 44*(22), 4880–4892.

Pesarin, F. and L. Salmaso (2010). *Permutation tests for complex data: theory, applications and software*. John Wiley & Sons.

Pesarin, F. and L. Salmaso (2013). On the weak consistency of permutation tests. *Communications in Statistics-Simulation and Computation 42*(6), 1368–1379.

Phipson, B. and G. K. Smyth (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*(1), 39.

Qiu, X., L. Klebanov, and A. Yakovlev (2005). Correlation between gene expression levels and limitations of the empirical bayes methodology for finding differentially expressed genes. *Statistical Applications in Genetics and Molecular Biology 4*(1).

Qiu, X. and A. Yakovlev (2006). Some comments on instability of false discovery rate estimation. *Journal of bioinformatics and computational biology 4*(05), 1057–1068.

Rao, C. R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society*, Volume 44, pp. 50–57. Cambridge Univ Press.

Rayner, J. (1997). The asymptotically optimal tests. *Journal of the Royal Statistical Society: Series D (The Statistician) 46*(3), 337–345.

Rippon, P. and J. C. Rayner (2010). Generalised score and Wald tests. *Advances in Decision Sciences 2010*.

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

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

Schwartzman, A. (2012). Comment: Fdp vs fdr and the effect of conditioning. *Journal of the American Statistical Association 107*(499), 1039–1041.

Schwartzman, A. and X. Lin (2011). The effect of correlation in false discovery rate estimation. *Biometrika 98*(1), 199–214.

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

Solari, A., L. Finos, and J. J. Goeman (2014). Rotation-based multiple testing in the multivariate linear model. *Biometrics 70*(4), 954–961.

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

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64*(3), 479–498.

Storey, J. D., J. E. Taylor, and D. Siegmund (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66*(1), 187–205.

Tusher, V. G., R. Tibshirani, and G. Chu (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences 98*(9), 5116–5121.

van der Laan, M. J., S. Dudoit, and K. S. Pollard (2004a). Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statistical applications in genetics and molecular biology 3*(1), 15.

van der Laan, M. J., S. Dudoit, and K. S. Pollard (2004b). Multiple testing. Part III. procedures for control of the generalized family-wise error rate and proportion of false positives.

Van der Vaart, A. W. (1998). *Asymptotic statistics*, Volume 3. Cambridge university press.

Wald, A. and J. Wolfowitz (1944). Statistical tests based on permutations of the observations. *The Annals of Mathematical Statistics 15*(4), 358–372.

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

Winkler, A. M., G. R. Ridgway, M. A. Webster, S. M. Smith, and T. E. Nichols (2014). Permutation inference for the general linear model. *Neuroimage 92*, 381–397.

List of Publications

J. Hemerik and J. Goeman (2017). Exact testing with random permutations. *TEST* (Online First version).

J. Hemerik and J. Goeman (2017). False discovery proportion estimation by permutations: confidence for significance analysis of microarrays. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80(1), 137-155.

J. Hemerik, A. Solari and J. Goeman. Permutation-based simultaneous confidence bounds for the false discovery proportion. Submitted for publication.

J. Hemerik, J. Goeman and L. Finos. Robust testing in generalized linear models by sign-flipping score contributions. Submitted for publication.

P. Spitali, K. Hettne, R. Tsonaka, E. Sabir, A. Seyer, **J. Hemerik**, J. Goeman, E. Picillo, M. Ergoli, L. Politano and A. Aartsma-Rus (2018). Cross sectional serum metabolomic study of multiple forms of muscular dystrophy. *Journal of Cellular and Molecular Medicine* (Early View version).

Samenvatting

Wetenschappelijk onderzoekers zijn vaak geïnteresseerd in een nulhypothese over een populatie. Een nulhypothese is een uitspraak zoals: "Rokers hebben (gemiddeld) dezelfde bloeddruk als niet-rokers."Een ander voorbeeld is de hypothese dat bij Europese vrouwen met een bepaalde ziekte, een nieuw medicijn A de ziekte (gemiddeld) niet beter verhelpt dan het gangbare medicijn B. In het eerste voorbeeld is de populatie die onderzocht wordt 'alle mensen', in het tweede geval 'alle Europese vrouwen met de ziekte.'

Om de eerste hypothese te toetsen, is het geen doen om van alle Nederlanders de bloedruk te meten. Dit zou bijvoorbeeld te veel tijd en geld kosten. In het algemeen meten onderzoekers dan ook niet de hele populatie waarin ze geïnteresseerd zijn, maar een representatieve steekproef. Bij het tweede voorbeeld kan bijvoorbeeld een groep van 20 vrouwen verzameld worden. Hiervan krijgen 10 willekeurig gekozen vrouwen medicijn A en de 10 andere vrouwen medicijn B. De vrouwen krijgen niet te horen welk medicijn ze hebben gekregen. Als bij deze vrouwen medicijn A steeds veel beter werkt dan medicijn B, is er bewijs dat A beter is dan B. We kunnen dan overwegen om de hypothese te verwerpen, dat A niet beter werkt dan B.

Wat het analyseren van zulke steekproeven zo interessant maakt, is dat we nooit met zekerheid kunnen zeggen dat A beter is dan B. Het kan namelijk zo zijn dat medicijn A door puur toeval aan mensen werd gegeven die een betere weerstand hadden. In het algemeen is er altijd een risico dat een hypothese onterecht verworpen wordt, doordat een steekproef door puur toeval een misleidend beeld geeft van de populatie. Er is dus altijd sprake van onzekerheid als conclusies over populaties worden gedaan op basis van steekproeven. De statistiek houdt zich bezig met het kwantificeren van die onzekerheid.

Het is belangrijk om hypotheses zodanig te toetsen, dat de kans op het verwerpen van een ware hypothese klein is. Dit zouden we kunnen we doen

door een hypothese alleen te verwerpen, als de kans dat deze waar is heel klein is, op basis van de steekproef. Het berekenen van die kans blijkt echter nogal problematisch te zijn, en in de (frequentische) statistiek probeert men in plaats daarvan te bepalen hoe onwaarschijnlijk de data zouden zijn, als de hypothese waar was.

De onwaarschijnlijkheid van de data onder de nulhypothese wordt gemeten met een getal, de *toetsstatistiek*. Hoe onwaarschijnlijker de data zijn onder de nulhypothese, hoe groter de toetsstatistiek is. Als de toetsstatistiek onwaarschijnlijk groot is onder de nulhypothese, wordt de nulhypothese verworpen. De onwaarschijnlijkheid van de toetsstatistiek wordt gemeten met de p-waarde, een getal tussen 0 en 1. Hoe groter de toetsstatistiek is, hoe kleiner de p-waarde is. Als de nulhypothese waar is, is de p-waarde (bij benadering) uniform verdeeld tussen 0 en 1. Vaak verwerpt men de nulhypothese als de p-waarde 0.05 of lager is. Is de p-waarde groter dan 0.05, dan zegt men dat er niet voldoende bewijs is om de hypothese te verwerpen.

Om te begrijpen waarom dit een goede manier van toetsen is, kunnen we ons realiseren dat er twee soorten fouten kunnen worden gemaakt. Ten eerste kunnen we een hypothese verwerpen, terwijl deze in werkelijkheid waar is. Dit heet een type-I fout. Ten tweede kunnen we een hypothese niet-verwerpen als deze in werkelijkheid onwaar is. Dit heet een type-II fout. Meestal worden type-I fouten erger gevonden dan type-II fouten.

Dat een type-I fout ernstige gevolgen kan hebben, is duidelijk. Stel bijvoorbeeld dat we concluderen dat medicijn A beter werkt dan medicijn B, terwijl medicijn A in werkelijkheid minder goed werkt. Dan zal, op basis van dit resultaat, in de toekomst misschien medicijn A – het verkeerde medicijn – aan patiënten gegeven worden.

Een type-II fout wordt vaak minder erg gevonden. Eén van de redenen is dat bij het niet-verwerpen van een hypothese eigenlijk geen uitspraak wordt gedaan, dus ook geen foutieve. Een hiermee samenhangende reden is dat de hypothese meestal zegt dat ‘er niks interessants aan de hand is’ (bijvoorbeeld: ‘A werkt niet beter dan het gebruikelijke medicijn B’), waardoor er geen nadruk wordt gelegd op niet-verworpen hypotheses in wetenschappelijke publicaties. Sterker nog, niet-verworpen hypotheses worden vaak überhaupt niet vermeld in publicaties (zogeheten ‘publication bias’). Type-II fouten zijn dus jammer, maar vervuilen de wetenschappelijke literatuur niet zo erg als type-I fouten.

In het voorbeeld van medicijnen bijvoorbeeld, betekent een type-II fout dat medicijn A beter werkt, maar dit niet geconcludeerd wordt. Dit is jammer, maar er wordt geen (foute) conclusie getrokken. Bovendien krijgt het niet-verwerpen van de hypothese weinig aandacht in de wetenschappelijke gemeenschap. Daardoor zal A in de toekomst waarschijnlijk nog een keer met B vergeleken worden, wat dan wel tot verwerping van de hypothese kan leiden.

Aangezien men de nulhypothese verwerpt als de data onwaarschijnlijk zijn onder de nulhypothese, is de kans op verwerpen klein als de hypothese waar is. Er wordt dus expliciet voor gezorgd dat de kans op type-I fouten klein blijft: precies wat men wil.

Dit proefschrift legt voor een belangrijk deel de nadruk op *multiple testing*, oftewel meervoudig toetsen. Dit is het toetsen van veel (strict gesproken: twee of meer) hypotheses tegelijk. We zouden bijvoorbeeld voor 20,000 genen de hypothese kunnen toetsen, dat de activiteit van het gen geen verband houdt met een bepaald fenotype (bijvoorbeeld de variant van kanker, die iemand heeft). Bij zoveel hypotheses moet de statisticus, op basis van de hem beschikbaar gestelde data, voor elke hypothese bepalen of deze verworpen wordt.

Als er veel hypotheses getoetst worden, kunnen er veel type-I fouten gemaakt worden. Als de onderzoeker er vrij zeker van wil zijn dat er geen enkele type-I fout gemaakt wordt, moet hij ervoor zorgen dat een hypothese alleen verworpen wordt, als er extreem veel bewijs tegen is. Het gevolg van zulke strenge eisen is vaak dat er maar weinig hypotheses verworpen worden, zelfs als in werkelijkheid veel van de hypotheses onwaar zijn.

Soms gaat het er niet om, ervoor te waken dat er helemaal geen type-I fouten zijn, maar wil de onderzoeker dat met grote kans bijvoorbeeld 95% van de verworpen hypotheses onwaar zijn. Hoofdstuk 3 en 4 van dit proefschrift bevatten nieuwe statistische methoden die dit kunnen garanderen.

In hoodstuk 3 wordt een bestaande *multiple testing* methode verbeterd, getiteld SAM ("Significance Analysis of Mircorarrays"). Deze methode werd in eerste instantie gebruikt voor genetische data, maar is breder toepasbaar. De oorspronkelijke SAM methodologie schat de fractie van type-I fouten onder alle verworpen hypothesen. Als de door SAM geschatte fractie type-I fouten laag is, suggereert dit dat veel van de verworpen hypothesen onwaar zijn.

De door SAM geschatte fractie van type-I fouten kan ver van de

werkelijke fractie af liggen. In hoofdstuk 3 wordt een methode ontwikkeld om een bovengrens voor de werkelijke fractie te geven. De gebruiker kan zelf kiezen hoe betrouwbaar deze bovengrens moet zijn. Op deze manier kan de gebruiker bijvoorbeeld een bovengrens berekenen die boven de ware fractie type-I fouten ligt met een kans van tenminste 95%. Als de gevonden bovengrens bijvoorbeeld 0.15 is, dan weet de gebruiker met 95% zekerheid dat de fractie type-I fouten niet groter is dan 0.15.

De methode van hoodstuk 3 is geïmplementeerd in een klein software pakket, dat online vrij beschikbaar is. Hierdoor kan de methode gemakkelijk door statistici wereldwijd gebruikt worden. De appendix van dit proefschrift bevat een gebruiksaanwijzing voor deze software.

Hoofdstuk 4 is gerelateerd aan hoofdstuk 3. Ook hier worden bovengrenzen gegeven voor de fractie van type-I fouten in een collectie verworpen hypothesen. In tegenstelling tot hoofdstuk 3 echter, zijn de bovengrenzen in hoofdstuk 4 nog steeds betrouwbaar als het verwerpingscriterium wordt bepaald op basis van de data. Hierdoor krijgt de gebruiker meer vrijheid in het aanpassen van het verwerpingscriterium na het kijken naar de data.

Statistiek gaat niet alleen over toetsen, maar ook over het maken van voorspellingsmodellen. (Hierbij staat het kwantificeren van onzekerheid nog steeds heel centraal.) Het kan bijvoorbeeld zo zijn dat het per persoon verschilt welk medicijn waarschijnlijk het beste zal werken. Dat kan afhangen van variabelen zoals geslacht en leeftijd. Het kan dan nuttig zijn om een statistisch model te maken, dat voorspelt welk medicijn het beste bij iemand werkt, gegeven eigenschappen als leeftijd, geslacht en genen.

Hoofdstuk 5 van dit proefschrift gaat over het toetsen binnen zulke modellen. Het gaat er hierbij om aan te tonen of bepaalde variabele, bijvoorbeeld leeftijd, een statistisch aantoonbare invloed op de uitkomstvariabele (bijvoorbeeld het optimale medicijn) heeft.

Zulke statistische toetsen worden vaak uitgevoerd onder bepaalde aannames, bijvoorbeeld dat de variantie van de uitkomstvariabele op een bepaalde manier afhangt van de verwachtingswaarde. In hoofdstuk 5 wordt een nieuwe toets geconstrueerd, die vaak nog goed werkt als niet aan de aannames voldaan is. Daardoor kan de gebruiker er meer op vertrouwen dat de kans op type-I fouten echt onder (zeg) 5% ligt, zelfs als bepaalde aannames niet helemaal kloppen. Net zoals de methode van hoofdstuk 3, is de methode van hoofdstuk 5 geïmplementeerd in software die online vrij beschikbaar is.

De toets in hoofdstuk 5 is gerelateerd aan de *multiple testing* methodes in de eerdere hoofstukken. Daardoor wordt het mogelijk om die methodes toe te passen in de modellen waar hoofdstuk 5 over gaat.

Dankwoord

De volgende personen ben ik dankbaar voor een bijzonder goede tijd.

Jelle, je bent bijzonder vakkundig en prettig om mee samen te werken. Bovendien heb ik erg genoten van alle reizen die we hebben gemaakt naar conferenties.

Tijdens de eerste twee jaar van mijn promotieperiode werkte ik bij het Radboudumc in Nijmegen. Ik wil alle collega's daar bedanken voor de goede sfeer. In het bijzonder wil ik de biostatistici danken voor de prettige en leerzame bijeenkomsten. Monika, Jakub and Rosa, thanks for the good company. Verder wil ik mijn bijzonder gezellige kamergenoten Hans, Erik en Ralph bedanken.

At the LUMC I was very lucky with my office mates Giorgos, Carlo, Alexia, Markus en Ningning. Thank you for the really good time. Ik ben ook de rest van de afdeling dankbaar. Ik heb veel van jullie geleerd.

Some of the most memorable months of my life, were those that I spent in Italy. Livio and Aldo, thank you for being so pleasant to collaborate with. You have been very welcoming and I enjoyed the guided tours in Venice and Milan. I would also like to thank Florian and Fabricio for the good company.

Ook buiten het werk om zijn een aantal mensen heel belangrijk geweest voor de totstandkoming van dit proefschrift, vanwege hun gezelschap en hulp. In het bijzonder wil ik mijn moeder bedanken, Adriana, Rob, mijn Oma's, Monique, Jeannette, Peter, Sergio, Bas, Walter, Anousch, Danielle en alle vrienden van Eetplan. Vooral dankzij jullie is mijn leven vol humor, warmte en wijze raad.

Curriculum Vitae

Jesse Hemerik werd op 6 april 1989 geboren in Leiden. Vanaf 2001 bezocht hij daar het Stedelijk Gymnasium. In 2007 begon hij met de studie wiskunde aan de Universiteit Leiden. Vanaf 2011 volgde hij daar de masteropleiding Applied Mathematics. Hij sloot deze in 2013 af met het schrijven van een scriptie aan het Leids Universitair Medisch Centrum.

Vanaf 2014 was hij werkzaam als onderzoeker in opleiding bij de Department for Health Evidence van het Radboudumc te Nijmegen. In 2016 zette hij zijn onderzoek voort bij de afdeling Medische Statistiek en Bioinformatica (nu Biomedical Data Sciences) van het Leids Universitair Medisch Centrum. Tussen 2014 en 2017 bracht hij tevens in totaal vijf maanden door als onderzoeker aan de universiteit van Padua. Zijn onderzoek tijdens deze vier jaar leidde tot dit proefschrift.

Sinds 12 februari 2018 is hij werkzaam als postdoctoraal onderzoeker bij de Department of Biostatistics van het Institute of Basic Medical Sciences in Oslo.