Universität Bremen

Institut für Statistik

Statistische Methoden zur Replizierbarkeitsbewertung im Rahmen mehrstufiger Studien

vorgelegt von

Anh-Tuan Hoang

aus Münster

Dissertation zur Erlangung des

naturwissenschaftlichen Doktorgrades

der Universität Bremen

Bremen 2022

Advisor: Prof. Dr. Thorsten Dickhaus, University of Bremen

First reviewer: Prof. Dr. Thorsten Dickhaus, University of Bremen

Second reviewer: Prof. Dr. Ruth Heller, Tel Aviv University

Date of the defence: 21 April 2022

Contents

Chapter 1

Introduction

A scientific result may be found in only a single study, owing to its specific laboratory methods, statistical methods or due to bad analysis. To strengthen a result, one can look for certain strengthening characteristics, for example repeatability, reproducability and replicability. There are different definitions of these; we adopt the one introduced by Miłkowski et al. (2018), where our understanding of replicability lies between their definitions of direct and conceptual replicability. If the scientific result can be found again under the same setup, conducted by the same scientists, it is considered repeatable, and if it is possible to find the same result under merely the same setup, it is called reproducible. Moreover, if the result can be found in another, independently conducted study, it is called replicable / replicated. This work deals with the latter, i.e. results that can be found in at least two studies that are independent of each other. The search for such results is called replicability analysis and requires appropriate statistical methods, which are developed in this work.

The problem of finding replicated findings has been widely acknowledged in the scientific community. An unspoken rule among venture capitalists states that "at least fifty percent of published studies, even those in top-tier academic journals, can't be repeated with the same conclusions by an industrial lab", cf. Osherovich (2011). Furthermore, Prinz et al. (2011) found that in only about twenty to twenty-five percent of the sixty-seven studies that they investigated, the reported data were in line with their own findings. In another example, the biotechnology firm Amgen only managed to scientifically confirm six findings out of fifty-three investigated papers in pre-clinical cancer research, cf. Begley and Ellis (2012). Ioannidis (2005) and Moonesinghe et al. (2007) illustrated how several factors can decrease the positive predictive value, the post-study probability that a formally statistically significant finding is indeed true. For a comprehensive overview of the challenges in research with regard to replicability, and for reasons and evidence of the "replicability crisis", we refer to Romero (2019). Note, that our goal is not to improve the quality of studies but to give a statistical framework with which replicable findings are determined more easily.

Given s independent studies that examine related research hypotheses but may differ in their setups, such as populations or laboratory methods, we are interested in the question of whether in at least γ out of these s studies findings can be made, where γ is chosen pre-analysis. For example, the aforementioned investigations by Prinz et al. (2011) and Begley and Ellis (2012) used $\gamma = s = 2$, and performed their replicability analysis on multiple research findings simultaneously. Now, assume that each study is associated with one hypothesis testing problem, for which a rejection means that a finding has been made. The γ-out-of-s hypothesis is denoted $H^{\gamma/s}$, its rejection means that at least γ out of s hypotheses are false, is also called partial conjunction (PC) null hypothesis. For $\gamma = 1$, it is also referred to as the global null and for $\gamma = s$ as the conjunction null hypothesis cf. Friston et al. (2005); Benjamini and Heller (2008). To that end, Benjamini et al. (2009) presented a procedure that identifies the largest γ for which the PC null hypothesis can still be rejected. Of course, the higher this γ, the stronger the statement, with $\gamma = s$ implying that the result has been replicated in all studies. The difference between replicability analysis and meta-analysis lies in γ, where $\gamma \geq 2$ is being used for the former and $\gamma = 1$ for the latter. Hence, the goal of a meta-analysis is to find at least one result from multiple, independent studies with common hypotheses, which possibly would not find significance alone, cf. for example Olkin (1995). Some of the oldest resources for meta-analyses include Fisher (1973); Tippett (1952); Stouffer et al. (1949), the latter of which proposed their method in a footnote.

Our analysis will focus on p-values, combining the p-values from the s studies into one so called PC p-value that evaluates the PC null hypothesis. If $s = \gamma$, the maximum of the s p-values is usually utilized, which performs badly. In case of $s = 2$, Bogomolov and Heller (2013) proposed a more powerful method that splits the significance levels between the two studies. Further, Heller et al. (2014) generalized their approach for the case of a high proportion of hypotheses that are true in both studies. A similar method

that only tests the hypotheses that are selected in both studies is provided by Bogomolov and Heller (2018), and a Bayesian approach was given by Heller and Yekutieli (2014).

We call p-values that correspond to true null hypotheses, null p-values, and p-values that correspond to false null hypotheses, non-null p-values. Null p-values are called valid if they are stochastically at least as large as the uniform distribution between 0 and 1, and usually, they are exactly uniformly distributed, cf. Lemma 3.3.1 in Lehmann and Romano (2005); if a valid null p-value is not uniform, it is called conservative. For rejections that are based on valid p-values the type I error probability is controlled while the significance level is not fully exhausted if the p-value is conservative. Additionally, conservative p-values can be disadvantageous if uniformity, or close to uniformity of the null p-value, is expected or assumed under the null. These conservative p-values can for example emerge if the corresponding null hypothesis comprises more than one parameter, that determine the distribution of the p-value. Such null hypotheses are called composite as opposed to simple null hypotheses, which only contain one parameter; p-values are usually defined such that they are uniformly distributed under the least favorable parameter configuration (LFC) in the corresponding null hypothesis, cf. Def. 2.1 in Dickhaus (2014). This means that the p-value is uniform under the LFC and stochastically larger, or conservative, under the other parameter configurations in the null. Furthermore, if p-values are discretely distributed, they can not be uniformly distributed, even under the LFC. In the replicability context, null PC p-values can be conservative, even if all null p-values before combining are uniformly distributed, cf. Benjamini and Heller (2008).

To combine the s p-values into one PC p-value that is valid for the PC null hypothesis, combination functions are needed. The PC null hypothesis $H^{\gamma/s}$ is false if and only if among all sets of $s-\gamma+1$ null hypotheses at least one of these null hypotheses is false, thus, one can derive PC p-values by taking the maximum of the $\binom{s}{s-\gamma+1}$ meta-analysis p-values. If for each of these p-values the same combination function f is used, the maximum of these is equivalent to combining only the $s-\gamma+1$ largest p-values with f (assuming f is increasing in its arguments), cf. Benjamini and Heller (2008); taking different combination functions was discussed by Wang and Owen (2019). It is important to examine which combination functions work well in what type of settings. Birnbaum (1954) showed that every non-decreasing p-value combination function of independent p-values works best for some alternative hypothesis, and that all such optimal combination functions must be non-decreasing. A common way of combining p-values for the global null hypotheses, $\gamma = 1$, is by averaging the p-values, see for example Vovk and Wang (2020) for a recent overview. Popular examples include the Fisher method, cf. Fisher (1973), and the Stouffer method (also called Z-test or inverse Normal test), cf. Stouffer et al. (1949), which are both derived by deriving the distribution of the sum of transformed uniformly distributed random variables. An important difference between the two is that the Fisher method places more weight on the smaller over the larger p-values, whereas the Stouffer method does not have this asymmetry, cf. Whitlock (2005). Furthermore, it is of interest to consider whether the combination function should be symmetric, i.e. if it does not change after permuting the p-values; many combination functions have weighted versions that usually perform better than their non-weighted versions but require additional information, for example the sample sizes of the studies. Won et al. (2009) recommended using a weighted version of the Stouffer method over the Fisher method if analyzing s of the same hypothesis. However, the performance is similar if considering the generalized Fisher combination by Lancaster (1961) while the weighted Stouffer method can be further improved by weighting according to the (unknown) effect sizes, cf. Zaykin (2011). A generalized Fisher test that utilizes estimated effects was found to be especially powerful given few studies with large effects, cf. Chen (2011). More on generalized versions of the Fisher method and weighted versions of the Stouffer method can be found in Lancaster (1961); Chen (2011); Zaykin (2011); Chen and Nadarajah (2014). Further examples of averaging methods are the arithmetic and the harmonic mean, which require adjustment factors, cf. for example Rüschendorf (1982); Won et al. (2009); Wilson (2019); Vovk and Wang (2020). Edgington (1972) and Kocak (2017) considered the sum of the p-values, similar to the arithmetic mean, but derive a combination p-value from the exact distribution of the sum of independent, uniformly distributed random variables. Another common type of combination function is one that only considers the smallest p-value, i.e. only the smallest p-value needs to be significant for the combination p-value to be significant. Examples are $1-(1-\min_i\{p_i\})^s$ as proposed by Tippett (1952), its generalization (that considers the k-th smallest p-value) by Wilkinson (1951), and the Bonferroni method, s times the minimum p-value, motivated by the Bonferroni inequality. Another combination function can be derived from the fact that the minimum of s independent uniformly distributed random variables on $[0,1]$ is Beta-distributed with parameters s and 1. As a mixture between the averaging and the minimum approach, Futschik et al. (2019) proposed the most significant partial (transformed) average of the i smallest p-values, $i = 1,\ldots,s$.

Under the PC null $H^{\gamma/s}$, combination functions are preferred whose distribution come closer to uniformity. The settings can differ with regard to the distribution of the null p-values, uniform versus conservative, as well as the number of null p-values. Under the alternative, i.e. there are at least γ false

null hypotheses, combination functions are preferred that have the highest power / are stochastically smallest. Evidence can be spread more evenly among the non-null p-values (spread out evidence patterns) or more focused among few of the false p-values (focused evidence patterns). To this end, Loughin (2004) investigated the behavior of several combination functions under different evidence patterns. Won et al. (2009) proposed a weighted method that is most powerful if alternatives are simple and each test has known expected effect sizes, or known ratios of thereof. Furthermore, Kocak (2017) generated Beta-distributed p-values and investigated the performance of some combination functions in different subsets of the parameter space, while Heard and Rubin-Delanchy (2018) calculated p-value combinations as likelihood ratio tests with simple alternatives that specify the p-value distribution, and also gave recommendations in their Table 1 for more general settings. The aforementioned references only dealt with the combination of independent p-values, which is, in the context of replicability analysis, what we are interested in. Hartung (1999) presented a version of the weighted Stouffer method for the case of dependence between the statistics, and Demetrescu et al. (2006) showed its validity for more general correlation matrices. Alves and Yu (2014) found that combination functions that assume independent p-values exaggerate the evidence if used with correlated p-values, and discussed introducing weights as a possible solution.

Suppose now, instead of testing one PC null hypothesis, we are faced with a set of multiple PC null hypotheses. This means that we are looking at multiple scientific results, and determine simultaneously which ones are replicated. Simultaneous inference of multiple hypotheses can be problematic if not accounting for some type I error measure. One such is the probability of at least one false positive, called the family-wise error rate (FWER), cf. for example Hochberg and Tamhane (1987). Famously, testing each p-value at α divided by the number of tested p-values (Bonferroni correction) controls the FWER at level α under each configuration, however, this can become very conservative if the number of tested hypotheses is high. Instead, Benjamini and Hochberg (1995) proposed controlling the expected proportion of false positives to total positives, called the false discovery rate (FDR), which allows for more than one false rejection as long as their number is low proportional to the total number of rejections. In both cases, an adjustment to the set of PC p-values is necessary. So-called multiple testing procedures that control a multiple testing Type I error are usually more conservative than simply applying individual tests; we refer to Shaffer (1995) for an overview of common multiple testing procedures as well as an overview of multiple testing in general. An improvement to the Bonferroni correction, given the p-values are independent, is provided in an early paper by Simes (1986), and an established procedure that controls the FDR in case of independent p-values is the linear step-up procedure proposed by Benjamini and Hochberg (1995), commonly called BH procedure. Storey et al. (2004) introduced a class of multiple testing procedures that generalize the BH procedure and prove the FDR control under the setting of independent null p-values. A certain type of positive dependency structure on the test statistics, called PRDS, was given by Benjamini and Yekutieli (2001) under which the BH procedure controls the FDR. For arbitrary dependency, they provided an adjustment to the BH procedure and proved that it controls the FDR. Storey (2002) provided conservatively biased point estimates of the FDR when rejecting all p-values less than a t, and Storey and Tibshirani (2003) defined the so-called q-value that denotes for a hypothesis the FDR if the corresponding p-value and all smaller p-values are deemed significant. On the other hand, Storey (2003) defined a type I error measure called the positive false discovery rate, similar to the FDR, that conditions on at least one rejection. For some further type I error measures in the Bayesian setting we refer to Efron (2008).

In this work, we investigate two solutions to the problem of conservative p-values, which are both related to selective inference, i.e. simultaneous inference after the selection of the strongest results, cf. Fithian et al. (2014); Taylor and Tibshirani (2015). In case of p-values these are the smallest ones; the non-selected p-values are thrown out and an adjustment to the selected p-values needs to be made to account for the selection (in this case to ensure validity). The non-selected p-values can for example either be randomized (between 0 and 1) or discarded; we call the first randomized p-values and the latter conditional p-values. Similar examples in the literature include Zaykin et al. (2002), who recognized the problem of combining p-values where there are a few large p-values, i.e. conservative p-values, and suggested combining only the p-values below a threshold in a product, which they called truncated product method. Alternatively, Dudbridge and Koeleman (2003) suggested the rank truncated product by only multiplying the K smallest p-values. Zhao et al. (2019) gave a condition for the validity of conditional p-values which they called uniform validity; see Figure 2 therein for a graphical illustration. On the other hand, randomized p-values for composite null hypotheses have been discussed by Dickhaus (2013), and for discrete models by Finner and Strassburger (2007); Habiger and Peña (2011); Dickhaus et al. (2012); Habiger (2015) among others. Instead of randomizing or conditioning the PC p-values, i.e. after combining, one can also do so on the base p-values before combining them to PC p-values. This makes sense if the base p-values are conservative themselves, cf. Zhao et al. (2019). A different approach is provided by Wang et al. (2021), that filters for the worst case under the null; endpoints j for which

$H_j^{(\gamma-1)/s}$ is false but $H_j^{\gamma/s}$ is true, dominate in the estimation of the FDR, however, the true proportion of these can be low, thus, making the estimation of the FDR unnecessarily conservative.

An important difference between conditional p-values and randomized p-values is that the first can be reproduced given the original p-values, whereas the latter depend on additional random numbers, cf. Dickhaus (2013). Randomized p-values should therefore not be used in decision making but can still be useful if estimating certain functionals like the proportion π_0 of true null hypotheses. An important resource for the estimation of π_0 is provided by Schweder and Spjøtvoll (1982), who assumed for their estimator that the non-null p-values are (almost surely) below a parameter λ, and that an expected proportion of $1-\lambda$ of the null p-values are above λ. This latter proportion of null p-values above λ is much higher if the null p-values are conservative, which leads to an unnecessary upward bias in the estimation of π_0. Langaas et al. (2005) gave an approach to estimating π_0 in a mixture model with decreasing and convex, decreasing density of the p-values under the alternative. Similarly, Kumar Patra and Sen (2016) provided in a two-component mixture model a non-parametric estimator of the distribution of the p-values under the alternative and an estimator for π_0 without the need for a tuning parameter. For a recent overview of π_0 estimators, see Chen (2019). Estimating π_0 can be very beneficial, since adaptive procedures that utilize estimates of π_0 can be significantly more powerful than their non-adaptive versions, which, usually, conservatively assume $\pi_0 = 1$, i.e. the latter are too conservative if some hypotheses are false. Benjamini and Hochberg (2000) showed that the adaptive BH procedure controls the FDR if the p-values are independent and certain conditions for the π_0-estimator are met, and provide a dynamic version of Schweder-Spjotvoll's estimator for π_0, i.e. an estimator that uses the data to select λ. Storey (2002) slightly modified Schweder and Spjotvoll's estimator and showed that the BH procedure with the π_0 estimate as plug-in controls the FDR asymptotically if certain assumptions are fulfilled. Mosig et al. (2001) introduced an iterative algorithm based on a histogram of the p-values that counts the excess p-values on the left side of the $[0, 1]$-interval, and Nettleton et al. (2006) showed how to calculate this algorithm without iteration. A proposed two-stage procedure that utilizes the BH procedure twice, first to estimate the number of true null hypotheses based on the number of rejections, controls the FDR if the test statistics are independent, cf. Benjamini et al. (2006). Furthermore, Blanchard and Roquain (2009) presented a categorization of adaptive procedures, and Finner and Gontscharuk (2009) investigated the FWER control of adaptive single-step and step-down procedures with plug-in estimators for π_0. Liang and Nettleton (2012) gave a condition under which dynamic adaptive procedures, those are adaptive procedures that use dynamic estimators for π_0, derived conservative π_0- and FDR estimators, and MacDonald et al. (2019) proved FDR control for a large class of dynamic adaptive procedures. Heesen and Janssen (2015) provided guarantees for the (asymptotic) FDR control for the adaptive BH procedure for certain types of dependencies among the p-values, and Heesen and Janssen (2016) considered a weighted approach in the estimation of π_0 and show that the corresponding adaptive step-down tests control the FDR.

We identify three overarching problems with regard to replicability analysis, which are

(i) conservative null p-values inherent in replicability analysis,

(ii) determining appropriate p-value combination functions for PC null hypotheses, and

(iii) the multiple testing of partial conjunction null hypotheses.

The first problem is being addressed in Chapters 2, 3 and partly in Chapter 5. We employ randomized PC p-values from minimum approaches, randomized PC p-values for the Schweder-Spjøtvoll estimator, and conditional PC p-values, respectively. In Chapter 4 we investigate different combination functions and their performance in different evidence patterns, and in Chapter 5 we investigate the Type I error control of the BH procedure with conditional PC p-values. To round out the thesis, we provide a short summary of the main chapters in the following section, and conclude the thesis in Chapter 6.

Summary of the main results

English version

This work is a compilation thesis whose main results consist of slight modifications of the author's articles. These can be found in Chapters 2 – 5; here we provide short summaries. More detailed summaries are given at the beginning of the respective chapters.

- In Chapter 2 we formulate a replicability model, combining the base p-values with a minimum type of combination. We randomize these according to Dickhaus (2013), provide an alternative way of calculating these randomized p-values, and give conditions for their validity.
- In Chapter 3 we generalize the definition of randomized p-values from Chapter 2 and investigate their benefits for the estimation of the proportion π_0 of true null hypotheses, especially if the non-randomized PC p-values are conservative.
- In Chapter 4 we examine several combination functions for different settings of true and false PC null hypotheses. We mainly vary parameter configurations by the number of true null hypotheses, by signal strengths, and by the choice of γ in the PC null hypothesis.
- In Chapter 5 we generalize the condition for validity to any combination function that is increasing in its arguments. The focus lies in conditional p-values, but results can similarly be made with randomized p-values from Chapter 3. Also, we investigate the power and the FDR of the BH procedure when used with these conditional p-values.

We conclude the thesis in Chapter 6 by naming some persistent issues and giving some ideas for future research that builds on our contributions.

German version

Diese Arbeit ist eine kumulative Dissertation, deren Hauptergebnisse die Artikel des Autoren umfasst. Diese befinden sich in Kapitel 2 – 5; wir geben hier kurze Zusammenfassungen. Detailliertere Zusammenfassungen befinden sich jeweils am Anfang jedes Kapitels.

- In Kapitel 2 formulieren wir ein Replizierbarkeitsmodell, wo die Basis p-Werte mit einer Minimum-Methode kombiniert werden. Diese randomisieren wir gemäß Dickhaus (2013), geben einen alternativen Berechnungsweg für diese randomisierten p-Werte an, und geben Konditionen für deren Validität.
- In Kapitel 3 verallgemeinern wir die Definition aus Kapitel 2 und untersuchen deren Nutzen in der Schätzung des Anteils π_0 der wahren Nullhypothesen, insbesondere wenn die nicht-randomisierten PC p-Werte konservativ sind.
- In Kapitel 4 untersuchen wir mehrere Kombinationsfunktionen in unterschiedlichen Konfigurationen wahrer und falscher Nullhypothesen. Wir variieren Parameterkonfigurationen hauptsächlich in der Anzahl wahrer Nullhypothesen, in der Signalstärke, und in der Wahl von γ in der PC Nullhypothese.
- In Kapitel 5 verallgemeinern wir die Validitätskonditionen für jegliche Kombinationsfunktionen, die wachsend in ihren Argumenten sind. Der Fokus ist auf bedingten p-Werten, aber Ergebnisse gelten ebenfalls in ähnlicher Weise für randomisierte p-Werte wie in Kapitel 3. Außerdem prüfen wir die Teststärke und die FDR der BH Prozedur mit bedingten p-Werten.

Wir beenden die Dissertation in Kapitel 6 mit einer Zusammenfassung bestehender Probleme und nennen mögliche Ideen für zukünftige Forschung, die auf unseren Beiträgen aufbaut.

Chapter 2

Randomized p-values for multiple testing and their application in replicability analysis

This chapter is a slightly modified version of Hoang and Dickhaus (2022) published in Biometrical Journal and has been reproduced here with the permission of the copyright holder. Appendix A contains the original appendix to this paper.

Authors

Anh-Tuan Hoang, Institute for Statistics, University of Bremen, Bremen, Germany

Prof. Dr. Thorsten Dickhaus, Institute for Statistics, University of Bremen, Bremen, Germany

Abstract We are concerned with testing replicability hypotheses for many endpoints simultaneously. This constitutes a multiple test problem with composite null hypotheses. Traditional p-values, which are computed under least favourable parameter configurations (LFCs), are over-conservative in the case of composite null hypotheses. As demonstrated in prior work, this poses severe challenges in the multiple testing context, especially when one goal of the statistical analysis is to estimate the proportion π_0 of true null hypotheses. Randomized p-values have been proposed to remedy this issue. In the present work, we discuss the application of randomized p-values in replicability analysis. In particular, we introduce a general class of statistical models for which valid, randomized p-values can be calculated easily. By means of computer simulations, we demonstrate that their usage typically leads to a much more accurate estimation of π_0 than the LFC-based approach. Finally, we apply our proposed methodology to a real data example from genomics.

Summary This chapter deals with the problem of conservative null PC p-values by applying randomized p-values.

We generalize the definition of randomized p-values for composite null hypotheses from Dickhaus (2013) for a special class of models, see Section 2.2. In the definition found in Dickhaus (2013), an estimator for the derived parameter is utilized to decide whether to randomize or not. In many models, this decision can equivalently be made based on the size of the p-value. To this end, we provide instructions for the calculation of a constant c, that depends on the model, such that the decision to randomize if and only if $p > c$ is equivalent to the definition given by Dickhaus (2013), see Theorem 2.1. If randomized, we replace the original p-value by a uniformly distributed random variable U. If not randomized, we adjust the p-value p such that it is uniformly distributed under LFCs given $p \leq c$, more specifically we multiply p by $1/c$. In Theorem 2.4 we show that these randomized p-values come closer to uniformity under both the null and the alternative hypothesis.

We consider a minimum approach of combining p-values based on the Beta-distribution to obtain p-values that are valid for PC null hypotheses; the results can be generalized to any minimum approach of combination. We consider the randomized version of the combined p-value and provide in Theorems 2.2, 2.3 and 2.5 conditions for the validity of these randomized p-values. In Section 2.5, we investigate the benefits of the use of randomized p-values for the Schweder-Spjøtvoll estimator $\hat{\pi}_0$. In Figure 2.2, we show graphically with the aid of empirical cumulative distribution functions how in one realization $\hat{\pi}_0$ may improve when the null LFC-based p-values are conservative. In simulations, we find that using randomized p-values decreases the bias of $\hat{\pi}_0$ especially if the conservativity of the

LFC-based p-values increases, for example from a higher γ or from more conservative base null p-values. In Section 2.6, we apply a replicability analysis on data from eight genome-wide association studies with the goal of identifying susceptibility loci for Crohn's disease (taken from Franke et al. (2010)). We use one study to select the most promising features and calculate $\hat{\pi}_0$ on the remaining features, and find improvements in all but one setting when using randomized p-values instead of LFC-based p-values.

Declaration of individual contributions Co-author and supervisor Prof. Dr. Thorsten Dickhaus came up with the idea for the paper, and I developed the theoretical results including their proofs. Simulations and evaluations including the figures and tables pertaining to the simulations were done by me. The final text was written and proof-read by both authors.

2.1 Introduction

The replication of scientific results is essential for their acceptance by the scientific community. In order to judge whether a scientific result has been replicated in an independent study, appropriate scientific methods are needed. We are concerned with developing such methods by formalizing the replication as a statistical hypothesis which has to be tested with an appropriate procedure. In particular, a simultaneous replicability analysis for many endpoints or markers, respectively, requires specialized multiple test procedures. We propose the usage of randomized p-values, as introduced by Dickhaus (2013), in this context.

For a single hypothesis test based on a test statistic $T(X)$, where X is the observable random variable, mathematically representing the data set, a (non-randomized) p-value $p(X)$ is a (deterministic) transformation of $T(X)$ onto $[0,1]$. Small values of $p(X)$ indicate incompatibility of the observed data with the null hypothesis H of interest. When basing test decisions on the p-value, type I error control at any pre-defined significance level $\alpha \in (0,1)$ is then equivalent to

$$\mathrm{P}_\vartheta(p(X) \leq \alpha) \leq \alpha \text{ for all } \vartheta \in H, \tag{2.1}$$

where P_ϑ denotes the probability measure under the parameter ϑ of the statistical model under consideration. A p-value fulfilling (2.1) for every $\alpha \in (0,1)$ is called a valid p-value.

In the case of a composite null hypothesis H, valid p-values have to satisfy $\mathrm{P}_\vartheta(p(X) \leq \alpha) \leq \alpha$ simultaneously for all parameter values $\vartheta \in H$. Hence, it is of interest to determine parameter values in H which maximize the probability in (2.1). These are called least favourable parameter configurations (LFCs). Under continuity assumptions, the p-value will usually be uniformly distributed under LFCs. However, if $\vartheta \in H$ is not an LFC, we typically have a strict inequality in condition (2.1) for many values of $\alpha \in (0,1)$. The p-value is then called conservative for these latter values of α under ϑ.

In the context of simultaneous testing of multiple null hypotheses, this deviation from the uniform distribution is problematic when utilizing data-adaptive multiple tests that rely on a pre-estimation of the proportion π_0 of the true null hypotheses. Non-uniformity can for example be caused by the presence of composite null hypotheses, as described before, or by the discreteness of the model. Randomized p-values resulting from a data-dependent mixing of the original p-value and an additional, on $[0,1]$ uniformly distributed random variable U, that is stochastically independent of the data X, are then often considered in the literature. The distribution of the randomized p-values under the null is typically much closer to uniformity than that of the non-randomized ones. In case of discrete models randomized p-values for simple null hypotheses $H = \{\vartheta^*\}$ have been discussed, among others, by Finner and Strassburger (2007); Habiger and Peña (2011); Dickhaus et al. (2012); Habiger (2015). These randomized p-values are closely related to well-known randomized hypothesis tests in discrete models, and they are exactly uniformly distributed under ϑ^*. For composite null hypotheses H, even in non-discrete models, it is generally not possible to achieve exact uniformity without abandoning the data completely. Dickhaus (2013) proposed one set of data-dependent weights for the mixing of X and U, that works well for composite, one-sided null hypotheses at least in certain location parameter models.

Due to the irreproducibility of the values of the random variable U, randomized p-values are not suitable for the final decision making. However, as demonstrated by Dickhaus et al. (2012), Dickhaus (2013) and others, they are very useful in the context of estimating the proportion π_0 of true null hypotheses. One popular estimator for π_0 has been proposed by Schweder and Spjøtvoll (1982). We will denote this estimator by $\hat{\pi}_0 \equiv \hat{\pi}_0(\lambda)$, where $\lambda \in [0,1)$ is a tuning parameter, and will refer to $\hat{\pi}_0$ as the Schweder-Spjøtvoll estimator. The proposal is to utilize the randomized p-values as defined in Dickhaus (2013) in $\hat{\pi}_0$. Since validity of the p-values utilized in $\hat{\pi}_0$ is essential for (mean) conservativeness of $\hat{\pi}_0$ (see Lemma 1 in Dickhaus et al. (2012)), we will provide some sufficient conditions for the validity of these randomized p-values in the sequel.

We will be particularly interested in calculating valid randomized p-values and in estimating π_0 in the context of a replicability analysis, where one aims at identifying discoveries made across more than one of $s \geq 2$ given independent studies. The null hypothesis of no replication is a special type of a composite null hypothesis. While a typical meta-analysis (see, e. g., Kulinskaya et al. (2008)) pools the available data across the studies, replicability analysis requires findings to hold in at least γ studies, where $2 \leq \gamma \leq s$ is a pre-defined parameter. This is an important distinction, since in a meta-analysis, one extremely small p-value may suffice to produce a small combined p-value, regardless of the evidence contributed by the other studies. Instead of combining all (endpoint-specific) p-values from the s studies, replicability analysis will usually apply a combination of all but the $\gamma - 1$ smallest of these p-values. In the context of bio-marker identification we consider $s \geq 2$ independent studies that examine $m \geq 2$ endpoints as possible bio-markers for a given disease. Whether one endpoint constitutes a bio-marker may differ between the studies, since the latter are (usually) conducted under different settings like different (sub-)populations. It is of interest to find bio-markers that are associated with the disease in at least γ different settings to rule out findings that can only be ascribed to one specific study setup. With our proposed methodology, it is possible to accurately estimate the number of replicated bio-markers. This is of interest in itself, but can also be used to increase the power of a multiple test for replicability. More details are provided in Sections 2.4 and 2.5.

Simultaneous testing of multiple replicability statements has also been the focus in prior literature. Benjamini et al. (2009) made use of partial conjunction nulls, meaning that at least a pre-specified number of the (study-specific) null hypotheses for a given endpoint are true, see also Benjamini and Heller (2008). They propose combining the $s - \gamma + 1$ largest p-values for each endpoint in an appropriate manner, and then using a false discovery rate (Benjamini and Hochberg, 1995) controlling procedure on these partial conjunction p-values. Bogomolov and Heller (2013) presented algorithms that separate $s = 2$ studies into primary and follow-up study. An empirical Bayesian approach has been proposed by Heller and Yekutieli (2014). Heller et al. (2014) introduced the r-value for each hypothesis, which indicates the lowest significance level with respect to the false discovery rate at which the corresponding hypothesis can be rejected. This allows for a ranking among the examined features. Bogomolov and Heller (2018) proposed to first select the promising features from each study separately and then to test the selected features.

2.2 Model setup

In the following, we introduce a general model for which randomized p-values are easily computable. The parameter s will be the number of studies, and the parameter $m \geq 2$ the number of endpoints (potential bio-markers) which also equals the number of null hypotheses. In the examples in Sections 2.2 and 2.3 we only consider the case of $s = 1$, which can be interpreted as bio-marker identification without replicability requirements. In Section 2.4, where we introduce replicability analysis, we only consider $s \geq 2$.

Consider a statistical model $(\Omega, \mathcal{F}, (\mathrm{P}_\vartheta)_{\vartheta \in \Theta})$ and let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m) : \Theta \to \Theta'$ denote a derived parameter, in which $\Theta' = \Theta'_1 \times \cdots \times \Theta'_m$ is a subset of $\mathbb{R}^{sm} = \mathbb{R}^s \times \cdots \times \mathbb{R}^s$, $s \geq 1$, where $\mathbb{R}$ denotes the set of real numbers. We assume that a consistent and, at least asymptotically, unbiased estimator $\hat{\boldsymbol{\theta}}_j = (\hat{\theta}_{1,j}, \ldots, \hat{\theta}_{s,j}) : \Omega \to \mathbb{R}^s$, for $\boldsymbol{\theta}_j(\vartheta) = (\theta_{1,j}(\vartheta), \ldots, \theta_{s,j}(\vartheta))$ is available $(j = 1, \ldots, m)$.

We consider m null hypotheses and their corresponding alternatives given by $\boldsymbol{\theta}_j(\vartheta) \in H_j$ versus $\boldsymbol{\theta}_j(\vartheta) \in K_j = \Theta'_j \setminus H_j$, where H_j and K_j are non-empty subsets of Θ'_j and Borel sets of $\mathbb{R}^s$ $(j = 1, \ldots, m)$.

Furthermore, we assume that a marginal test φ_j for testing H_j against K_j is constructed as $\varphi_j(x) = \mathbf{1}\{T_j(x) \in \Gamma_j(\alpha)\}$, where $\Gamma_j(\alpha)$ denotes a rejection region, $\alpha \in (0, 1)$ denotes a fixed, local significance level, $x \in \Omega$ an observation, and $T_j : \Omega \to \mathbb{R}$ a measurable mapping such that the test statistic $T_j(X)$ has a continuous cumulative distribution function under any $\vartheta \in \Theta$ $(j = 1, \ldots, m)$. We often write $\hat{\boldsymbol{\theta}}_j$ or T_j instead of $\hat{\boldsymbol{\theta}}_j(X)$ and $T_j(X)$, respectively $(j = 1, \ldots, m)$.

The following general assumptions are made:

($GA1$) For all $j = 1, \ldots, m$, there exists a constant $c_j \in [0, 1]$, such that $\{x \in \Omega : T_j(x) \in \Gamma_j(c_j)\}$ $= \{x \in \Omega : \hat{\boldsymbol{\theta}}_j(x) \in K_j\}$ holds.

($GA2$) Nested rejection regions: for every $j = 1, \ldots, m$ and $\alpha' < \alpha$, it holds $\Gamma_j(\alpha') \subseteq \Gamma_j(\alpha)$.

($GA3$) For every $j = 1, \ldots, m$ and $\alpha \in (0, 1)$, it holds $\sup_{\vartheta : \boldsymbol{\theta}_j(\vartheta) \in H_j} \mathrm{P}_\vartheta(T_j \in \Gamma_j(\alpha)) = \alpha$.

($GA4$) For every $j = 1, \ldots, m$, the set of LFCs for φ_j, i. e., the set of parameter values that yield the supremum in ($GA3$), does not depend on α.

The conditions $(GA2)-(GA4)$ are the same as required for the models in Dickhaus (2013), whereas for assumption $(GA1)$ in Dickhaus (2013) only the condition $\{x \in \Omega:\ T_j(x) \in \Gamma_j(\alpha)\} \subseteq \{x \in \Omega:\ \hat{\boldsymbol{\theta}}_j(x) \in K_j\}$ for α small enough has to be met.

Assumption $(GA1)$ serves as a connection between the test statistic $T_j(X)$ and the estimator $\hat{\boldsymbol{\theta}}_j(X)$, for $1 \leq j \leq m$. It requires, that $\{\hat{\boldsymbol{\theta}}_j \in K_j\}$ is in itself a rejection event at level c_j. Furthermore, assumption $(GA2)$ together with $(GA1)$ implies

$$\begin{array}{ll} \{x \in \Omega:\ T_j(x) \in \Gamma_j(\alpha)\} \subseteq \{x \in \Omega:\ \hat{\boldsymbol{\theta}}_j(x) \in K_j\} & \text{for } \alpha < c_j \quad \text{as well as} \\ \{x \in \Omega:\ T_j(x) \in \Gamma_j(\alpha)\} \supseteq \{x \in \Omega:\ \hat{\boldsymbol{\theta}}_j(x) \in K_j\} & \text{for } \alpha > c_j \end{array} \tag{2.2}$$

for all $j = 1, \ldots, m$. Assumption $(GA3)$ means that under any LFC for φ_j the rejection probability is exactly α.

LFC-based p-values for the marginal tests φ_j are formally defined as

$$p_j^{LFC}(X) = \inf_{\{\tilde{\alpha} \in (0,1):\, T_j(x) \in \Gamma_j(\tilde{\alpha})\}} \ \sup_{\{\vartheta:\, \boldsymbol{\theta}_j(\vartheta) \in H_j\}} \mathrm{P}_\vartheta(T_j(X) \in \Gamma_j(\tilde{\alpha})).$$

Under assumptions $(GA2)$ – $(GA4)$, we obtain that

$$p_j^{LFC}(X) = \inf\{\tilde{\alpha} \in (0,1) : T_j(X) \in \Gamma_j(\tilde{\alpha})\} \ (j = 1, \ldots, m).$$

Such LFC-based p-values $p_j^{LFC}(X)$ are uniformly distributed on $[0,1]$ under any LFC for φ_j (Lehmann and Romano, 2005, Lemma 3.3.1). If $\Gamma_j(\alpha) = (F_j^{-1}(1-\alpha), \infty)$, where F_j is the cumulative distribution function of $T_j(X)$ under an LFC for φ_j, the above definition leads to $p_j^{LFC}(X) = 1 - F_j(T_j(X))$.

Example 2.1. *Models* 1 *and* 2 *in Dickhaus (2013) are one-sided normal means models that fulfil the general assumptions* $(GA1)$ – $(GA4)$. *Notice that the index* i *in Dickhaus (2013) corresponds to the index* j *in our notation, and that the dimension* s *of the derived parameters is one in both models.*

Dickhaus (2013) showed that the general assumptions $(GA2)$ – $(GA4)$ *hold in these models. Our stricter assumption* $(GA1)$ *follows in both models from the fact that the estimator* $\hat{\boldsymbol{\theta}}_j(X)$ *is positive, i.e. inside the alternative, if and only if the test statistic* $T_j(X)$ *is positive, which is equivalent to* $T_j(X) \in \Gamma_j(1/2)$, *such that* $c_j = 1/2$ *in* $(GA1)$ $(j = 1, \ldots, m)$.

2.3 The randomized p-values

2.3.1 General properties

Let $U_1, \ldots, U_m$ be stochastically independent and identically, on $[0,1]$ uniformly distributed random variables, defined on the same probability space as X, such that each U_j is stochastically independent of X. We obtain the randomized p-value $p_j^{rand}(X, U_j)$ by mixing U_j and p_j^{LFC} in a data-dependent manner, specifically

$$p_j^{rand}(X, U_j) = w_j(X)\ U_j + (1 - w_j(X))\ G_j\big(p_j^{LFC}(X)\big), \tag{2.3}$$

where G_j is a suitable function necessary for the validity of that randomized p-value and $0 \leq w_j(X) \leq 1$ is a data-dependent weight $(j = 1, \ldots, m)$.

We consider the choice $w_j(X) = \mathbf{1}_{H_j}\{\hat{\boldsymbol{\theta}}_j(X)\}$ $(j = 1, \ldots, m)$, which follows the definition of the randomized p-values as introduced in Dickhaus (2013).

Definition 2.1.
We define randomized p-values as follows

$$p_j^{rand}(X, U_j) = U_j\ \mathbf{1}_{H_j}\{\hat{\boldsymbol{\theta}}_j(X)\} + G_j\big(p_j^{LFC}(X)\big)\ \mathbf{1}_{K_j}\{\hat{\boldsymbol{\theta}}_j(X)\},$$

where G_j *denotes the conditional cumulative distribution function of* $p_j^{LFC}(X)$, *given the event* $\{\hat{\boldsymbol{\theta}}_j \in K_j\}$, *under any LFC for* φ_j $(j = 1, \ldots, m)$.

The reasoning behind Definition 2.1 is the general setup of selective inference as described, for instance, by Fithian et al. (2014): First, promising features are selected and then inference is performed conditionally to the selection event. In our case, the selection event for endpoint j is $\{\hat{\boldsymbol{\theta}}_j(X) \in K_j\}$, because otherwise we abandon the data for that endpoint in the subsequent data analysis. Moreover, the conditional cumulative distribution function (given the selection event) for endpoint j is G_j, with which $p_j^{LFC}(X)$ has to be transformed once endpoint j has been selected. Ideally, we want p-values to be uniformly distributed on $[0,1]$ under null hypotheses. For a fixed $j = 1, \ldots, m$, we therefore set

$p_j^{rand}(X) = U_j$ if $\hat{\boldsymbol{\theta}}_j(X) \in H_j$ holds. Due to (2.2), $\varphi_j(x) = 0$ whenever $\hat{\boldsymbol{\theta}}_j(x) \in H_j$ holds, when applying a local significance level $\alpha < c_j$. This means that in case of $\hat{\boldsymbol{\theta}}_j(x) \in H_j$ we cannot reject H_j at a significance level lower than c_j. Since c_j can be very large, e. g. $1/2$ in Example 2.1 and even larger in our models for replicability analysis in Section 2.4, we can, in practice, assume that H_j is true in case of $\hat{\boldsymbol{\theta}}_j(X) \in H_j$, and switch to a uniform variate U_j that has the desired properties for a p-value under H_j.

In the following theorem we give formulas for the calculation of the function G_j and the randomized p-value p_j^{rand} $(j = 1, \ldots, m)$.

Theorem 2.1.
Let $j \in \{1, \ldots, m\}$ be fixed and $\vartheta_0 \in \Theta$ with $\boldsymbol{\theta}_j(\vartheta_0) \in H_j$ be any LFC for φ_j. Under assumptions $(GA1)$ – $(GA4)$ we obtain the following.

(i) It holds that $c_j = \mathrm{P}_{\vartheta_0}(\hat{\boldsymbol{\theta}}_j(X) \in K_j)$.

(ii) The conditional cumulative distribution function G_j of $p_j^{LFC}(X)$, given $\hat{\boldsymbol{\theta}}_j \in K_j$, is a piecewise linear function in $t \in [0,1]$, more precisely it holds $G_j(t) = t\,\boldsymbol{1}_{[0,c_j]}(t)/c_j + \boldsymbol{1}_{(c_j,1]}(t)$.

(iii) The randomized p-values, as defined in Definition 2.1, are of the form

$$p_j^{rand}(X, U_j) = U_j \;\boldsymbol{1}_{H_j}\{\hat{\boldsymbol{\theta}}_j(X)\} + p_j^{LFC}(X)\, c_j^{-1}\, \boldsymbol{1}_{K_j}\{\hat{\boldsymbol{\theta}}_j(X)\}.$$

Since $p_j^{LFC}(x) < c_j$ implies $\hat{\boldsymbol{\theta}}_j(x) \in K_j$, and $p_j^{LFC}(x) > c_j$ implies $\hat{\boldsymbol{\theta}}_j(x) \in H_j$, for all $x \in \Omega$, we have

$$p_j^{rand}(x, u_j) = u_j\, \mathbf{1}_{(c_j,1]}\{p_j^{LFC}(x)\} + p_j^{LFC}(x)\, c_j^{-1}\, \mathbf{1}_{[0,c_j)}\{p_j^{LFC}(x)\}$$

for any $x \in \Omega$ and $u_j \in [0,1]$, when disregarding the case $p_j^{LFC}(x) = c_j$, for which $p_j^{rand}(x, u_j)$ is either 1 or u_j $(j = 1, \ldots, m)$.

Example 2.2. *We apply Theorem 2.1 to both models in Example 2.1. In both models it holds, that $p_j^{LFC}(x) < t$ is equivalent to $T_j(x) \in \Gamma_j(t)$ for all $x \in \Omega$ and $t \in [0,1]$. In particular, $p_j^{LFC}(x) < c_j$ is equivalent to $\hat{\boldsymbol{\theta}}_j(x) \in K_j$, $x \in \Omega$, such that $\boldsymbol{1}_{H_j}\{\hat{\boldsymbol{\theta}}_j(X)\}$ and $\boldsymbol{1}_{K_j}\{\hat{\boldsymbol{\theta}}_j(X)\}$ in Part 3 of Theorem 2.1 can be replaced by $\boldsymbol{1}_{[c_j,1]}\{p_j^{LFC}(X)\}$ and $\boldsymbol{1}_{[0,c_j)}\{p_j^{LFC}(X)\}$, respectively. Let $j = 1, \ldots, m$ be fixed.*

(i) (Multiple Z-tests model) From Theorem 2.1 it follows that $c_j = \mathrm{P}_{\vartheta_0}(\hat{\boldsymbol{\theta}}_j \in K_j) = 1/2$, and that

$$G_j(t) = 2t\,\boldsymbol{1}_{[0,\frac{1}{2}]}(t) + \boldsymbol{1}_{(\frac{1}{2},1]}(t),\ t \in [0,1], \tag{2.4}$$

$$p_j^{rand}(x, u_j) = u_j\,\boldsymbol{1}_{(\frac{1}{2},1]}\big\{p_j^{LFC}(x)\big\} + 2\,p_j^{LFC}(x)\,\boldsymbol{1}_{[0,\frac{1}{2}]}\big\{p_j^{LFC}(x)\big\} \tag{2.5}$$

for $x \in \Omega$ and $u_j \in [0,1]$.

(ii) (Multiple t-tests model) Analogously to the multiple Z-tests model, it follows directly from Theorem 2.1, that $c_j = 1/2$ and that the expressions for $G_j(t)$ and $p_j^{rand}(x, u_j)$, respectively, are as in (2.4) and (2.5).

These results agree with the calculations in (Dickhaus, 2013, pp.1971, 1973).

2.3.2 Conditions for the validity of the randomized p-values

As mentioned before, valid p-values are usually required for a conservative (non-negatively biased) estimation of the proportion π_0 of true null hypotheses, particularly if the Schweder-Spjøtvoll estimator $\hat{\pi}_0$ is applied. This section provides some conditions for the validity of the randomized p-values as defined in Definition 2.1 for our model setup.

Theorem 2.2.
Let $j \in \{1, \ldots, m\}$ be fixed. Under the general assumptions $(GA1)$ – $(GA4)$, assume that $p_j^{LFC}(X)$ has a continuous and strictly increasing cumulative distribution function under any $\vartheta \in \Theta$. Then, the randomized p-value p_j^{rand}, as defined in Definition 2.1, is a valid p-value if and only if

$$\mathrm{P}_\vartheta\big(T_j(X) \in \Gamma_j(z)\big) \le z\,\frac{\mathrm{P}_\vartheta(\hat{\boldsymbol{\theta}}_j \in K_j)}{\mathrm{P}_{\vartheta_0}(\hat{\boldsymbol{\theta}}_j \in K_j)}, \quad 0 \le z \le \mathrm{P}_{\vartheta_0}(\hat{\boldsymbol{\theta}}_j \in K_j),$$

for any $\vartheta \in \Theta$ with $\boldsymbol{\theta}_j(\vartheta) \in H_j$ and for any LFC ϑ_0 for φ_j.

In many applications, a rejection of H_j after observing $T_j(x)$ implies a rejection of H_j if we observe larger test values $T_j(y) \geq T_j(x)$, $x, y \in \Omega$. More specifically, the rejection regions are often of the form $\Gamma_j(\alpha) = (b(\alpha), \infty)$ for some non-decreasing boundary function $b : [0,1] \to \mathbb{R}$. Usually, $b(\alpha) = F^{-1}(1-\alpha)$, $\alpha \in [0,1]$, where F is the cumulative distribution function of $T_j(X)$ under an LFC for φ_j, such that $(GA3)$ holds. Among others, the models from Example 2.1 fulfil this condition, under which the validity of the randomized p-value p_j^{rand} follows from $T_j(X)$ being smaller in the hazard rate order under any $\vartheta \in \Theta$ with $\boldsymbol{\theta}_j(\vartheta) \in H_j$ than under an LFC for φ_j.

We denote the hazard rate order and the likelihood ratio order with "$\leq_{\mathrm{hr}}$" and "$\leq_{\mathrm{lr}}$", respectively. We use the notation $T_j(X)^{(\vartheta)}$ to refer to the distribution of $T_j(X)$ under $\vartheta \in \Theta$. For a more detailed introduction to our notations we refer to Appendix A.

Theorem 2.3. *Let a model as in Section 2.2 be given and $j = 1, \ldots, m$ be fixed. We assume that the rejection regions are of the form $\Gamma_j(\alpha) = (F^{-1}(1-\alpha), \infty)$, $\alpha \in [0,1]$, where F is the cumulative distribution function of $T_j(X)$ under any LFC $\vartheta_0 \in \Theta$ for φ_j.*

Then the randomized p-value p_j^{rand} as defined in Definition 2.1 is valid if it holds $T_j(X)^{(\vartheta)} \leq_{\mathrm{hr}} T_j(X)^{(\vartheta_0)}$ for all $\vartheta \in \Theta$ with $\boldsymbol{\theta}_j(\vartheta) \in H_j$ and any LFC ϑ_0 for φ_j.

Corollary 2.1.
By Theorem 1.C.2 in Shaked and Shanthikumar (2007), replacing the hazard rate order by the likelihood ratio order in Theorem 2.3 is also sufficient for the validity of p_j^{rand}.

Example 2.3. *We show via Theorem 2.3 that the randomized p-values as calculated in Example 2.2 are valid. Let $j \in \{1, \ldots, m\}$ be fixed.*

(*i*) *(Multiple Z-tests Model) Let $\vartheta_0 \in \Theta$ with $\boldsymbol{\theta}_j(\vartheta_0) = 0$ be an LFC for φ_j, and $\vartheta \in \Theta$ with $\boldsymbol{\theta}_j(\vartheta) = \mu_j < 0$. Recall that $T_j(X) = \bar{X}_j$ is normally distributed on $\mathbb{R}$ with variance $1/n_j$ and expected values μ_j and 0 under ϑ and ϑ_0, respectively. It is easy to show that $f_{\vartheta_0}(t)/f_{\vartheta}(t)$ is non-decreasing in t and therefore $T_j(X)^{(\vartheta)} \leq_{\mathrm{lr}} T_j(X)^{(\vartheta_0)}$ holds, where f_{ϑ} and f_{ϑ_0} denote the Lebesgue densities of $N(\mu_j, 1/n_j)$ or $N(0, 1/n_j)$, respectively. According to Corollary 2.1 our randomized p-values p_j^{rand} are valid in this model.*

(*ii*) *(Multiple t-tests Model) Now we have that $T_j(X) = n_j^{1/2}\bar{X}_j/S_j$ possesses a non-central t-distribution with non-centrality parameter $\tau_j(\vartheta)$ and $n_j - 1$ degrees of freedom, $\tau_j(\vartheta) = n_j^{1/2}\mu_j/\sigma_j$, and $\mu_j = \boldsymbol{\theta}_j(\vartheta)$.*

According to (Karlin and Rubin, 1956a, p. 639) and (Karlin, 1956, p. 126), non-central t-distributions $\left(t_{\mu,\nu}\right)_{\mu\in\mathbb{R}}$ have monotone likelihood ratio, i.e. $t_{\mu_1,\nu} \leq_{\mathrm{lr}} t_{\mu_2,\nu}$ if and only if $\mu_1 \leq \mu_2$. For $\vartheta, \vartheta_0 \in \Theta$ with $\boldsymbol{\theta}_j(\vartheta) \leq 0$ and $\boldsymbol{\theta}_j(\vartheta_0) = 0$, it is $\tau_j(\vartheta) = n_j^{1/2}\boldsymbol{\theta}_j(\vartheta)/\sigma_j \leq 0 = \tau_j(\vartheta_0)$, and therefore $T_j(X)^{(\vartheta)} \leq_{\mathrm{lr}} T_j(X)^{(\vartheta_0)}$. According to Corollary 2.1 our randomized p-values p_j^{rand} in this model are valid.

Under certain conditions randomized p-values p_j^{rand} as defined in Definition 2.1 are closer to Uni$[0,1]$ than their LFC-based counterparts p_j^{LFC} under the null hypothesis H_j, that is,

$$\mathrm{Uni}[0,1] \leq_{\mathrm{st}} p_j^{rand}(X, U_j)^{(\vartheta)} \leq_{\mathrm{st}} p_j^{LFC}(X)^{(\vartheta)}$$

or, equivalently,

$$\mathrm{P}_\vartheta(p_j^{LFC}(X) \leq t) \leq \mathrm{P}_\vartheta(p_j^{rand}(X, U_j) \leq t) \leq t$$

for all $t \in [0,1]$ and $\vartheta \in \Theta$ with $\boldsymbol{\theta}_j(\vartheta) \in H_j$.

Theorem 2.4. *Let a model as in Section 2.2 be given and $j \in \{1, \ldots, m\}$ be fixed. If the cumulative distribution function of $p_j^{LFC}(X)$ is convex under a fixed $\vartheta \in \Theta$, then it holds*

$$\mathrm{Uni}[0,1] \leq_{\mathrm{st}} p_j^{rand}(X, U_j)^{(\vartheta)} \leq_{\mathrm{st}} p_j^{LFC}(X)^{(\vartheta)}.$$

On the other hand, if the cumulative distribution function of $p_j^{LFC}(X)$ is concave under a fixed $\vartheta \in \Theta$, then it holds

$$p_j^{LFC}(X)^{(\vartheta)} \leq_{\mathrm{st}} p_j^{rand}(X, U_j)^{(\vartheta)} \leq_{\mathrm{st}} \mathrm{Uni}[0,1].$$

Remark 2.1. (*i*) *If $p_j^{LFC}(X)$ is a valid p-value, its cumulative distribution function can never be strictly concave under the null hypothesis H_j ($j = 1, \ldots, m$). This can be seen as follows: The points $(0,0)$ and $(1,1)$ necessarily lie on the graph of the cumulative distribution function of $p_j^{LFC}(X)$, because the support of $p_j^{LFC}(X)$ is (a subset of) $[0,1]$. Now, if the cumulative distribution function of $p_j^{LFC}(X)$ would be strictly concave under (some parameter value in) H_j, we would find a value $t^* \in (0,1)$ such that the value of the cumulative distribution function of $p_j^{LFC}(X)$ evaluated at t^* exceeds t^*. This is a contradiction to the assumed validity of $p_j^{LFC}(X)$.*

(*ii*) *From Theorem* 2.4*, if the cumulative distribution function of* $p_j^{LFC}(X)$ *is convex under all* $\vartheta \in \Theta$ *with* $\boldsymbol{\theta}_j(\vartheta) \in H_j$*, the randomized p-value* p_j^{rand} *is a valid p-value* $(j = 1, \ldots, m)$.

(*iii*) *The condition mentioned in the second remark is stronger than the condition* $T_j(X)^{(\vartheta)} \leq_{\text{hr}} T_j(X)^{(\vartheta_0)}$ *in Theorem* 2.3*. Namely, the convexity of the cumulative distribution function of* $p_j^{LFC}(X)$ *under* ϑ *is equivalent to* $T_j(X)^{(\vartheta)} \leq_{\text{lr}} T_j(X)^{(\vartheta_0)}$ *whenever* ϑ_0 *is an LFC for* φ_j $(j = 1, \ldots, m)$*. It is known (see the theorem in Appendix A.2.3) that the likelihood ratio order is stronger than the hazard rate order.*

Example 2.4. *For both models from our ongoing examples the cumulative distribution function of* p_j^{LFC} *is convex under* H_j *and concave under* K_j*. Therefore, for each* $\vartheta \in \Theta$ *one of the two conditions in Theorem* 2.4 *is satisfied. Thus,* p_j^{rand} *is always closer to* $\text{Uni}[0,1]$ *than* p_j^{LFC} *(in the sense of stochastic order) in both models.*

2.4 Randomized p-values in replicability analysis

2.4.1 Model setup

We come back to the framework of bio-marker identification. We want to find bio-markers that have been verified in at least γ studies, where the parameter $\gamma \in \{2, \ldots, s\}$ is pre-defined and fixed. For $\gamma = s$, we declare discoveries replicated, only if they have been made in each considered study. It is clear, that the set of true null hypotheses is increasing in γ.

For each endpoint j and study i, we denote the true effect on the considered disease state by a parameter $\theta_{i,j}$, where $\theta_{i,j} > 0$ means a positive effect. We consider an endpoint to be a bio-marker only if it exhibits a positive effect on the disease. This can be replaced by testing for any one fixed, directional association between the endpoint and the disease. The parameters $\theta_{i,j}$ may differ inherently in i due to the different settings across the studies like different populations or different laboratory / statistical methods.

We consider the model from Section 2.2 for $s \geq 2$. Unless stated otherwise, we only consider $\Theta' = \mathbb{R}^{sm}$, i.e. each derived parameter $\theta_{i,j}$ may take any value in $\mathbb{R}$ $(i = 1, \ldots, s;\ j = 1, \ldots, m)$.

Before we get to constructing the test statistic $T_j(X)$ and the rejection region $\Gamma_j(\alpha)$, we first make some requirements about the marginal model setup $(j = 1, \ldots, m)$. This will make it easier to present sufficient conditions for the general assumptions $(GA1)$ – $(GA4)$ from Section 2.2. We do not require the data for different endpoints in the same study to be independent.

For every study $i = 1, \ldots, s$ and marker $j = 1, \ldots, m$ we test $H_{i,j} = \{\theta_{i,j} \leq 0\}$ vs. $K_{i,j} = \{\theta_{i,j} > 0\}$. We assume that a consistent and, at least asymptotically, unbiased estimator $\hat{\theta}_{i,j} : \Omega \to \mathbb{R}$ for $\theta_{i,j}(\vartheta)$ is available. Furthermore, the marginal test $\varphi_{i,j}$ for testing $H_{i,j}$ against $K_{i,j}$ is based on a test statistic $T_{i,j}(X)$ and rejection region $\Gamma_{i,j}(\alpha)$, where $\alpha \in (0,1)$ denotes the (local) significance level, $x \in \Omega$ an observation, and $T_{i,j} : \Omega \to \mathbb{R}$ a measurable mapping such that the test statistic $T_{i,j}(X)$ has a continuous cumulative distribution function under any $\vartheta \in \Theta$. The corresponding LFC-based p-values are then denoted by $p_{i,j}(X)$.

For every $i = 1, \ldots, s$ and $j = 1, \ldots, m$ we make the following assumptions:

$(RA1)$ It holds $\Gamma_{i,j}(\alpha) = (F_{i,j}^{-1}(1-\alpha), \infty)$ and $p_{i,j}(X) = 1 - F_{i,j}\big(T_{i,j}(X)\big)$, the set of LFCs for $\varphi_{i,j}$ is $\{\vartheta \in \Theta : \theta_{i,j}(\vartheta) = 0\}$, and $F_{i,j}$ denotes the cumulative distribution function of $T_{i,j}(X)$ under an LFC for $\varphi_{i,j}$.

$(RA2)$ The assumptions $(GA1) - (GA4)$ are fulfilled with respect to j for any fixed $i \in \{1, \ldots, s\}$. We denote with $c_{i,j}$ the value, that satisfies $\{x \in \Omega : T_{i,j}(x) \in \Gamma_{i,j}(c_{i,j})\} = \{x \in \Omega : \hat{\theta}_{i,j}(x) \in K_{i,j}\} = \{x \in \Omega : \hat{\theta}_{i,j}(x) > 0\}$ for assumption $(GA1)$.

$(RA3)$ There exists a $d_j \in (0,1)$ such that $p_{i,j}(x) < d_j$ if and only if $\hat{\theta}_{i,j}(x) > 0$, for all $x \in \Omega$ and $1 \leq i \leq s$.

$(RA4)$ It holds $\lim\limits_{\theta_{i,j}(\vartheta) \to \infty} \mathrm{P}_\vartheta\big(F_{i,j}(T_{i,j}(X)) = 1\big) = 1$.

In one-sided problems, assumption $(RA1)$ is usually fulfilled. Due to $(RA1)$ it now holds

$$p_{i,j}(x) < t \iff T_{i,j}(x) \in \Gamma_{i,j}(t),\ x \in \Omega. \tag{2.6}$$

Assumption $(RA3)$ is akin to assumption $(GA1)$ from Section 2.2, and follows from $(RA2)$ if and only if $c_{1,j} = \cdots = c_{s,j}$ holds $(j = 1, \ldots, m)$. In the latter case, we have that $d_j = c_{1,j}$.

For convenience we write

$$\lim_{\theta_{i,j}(\vartheta)\to\infty} \mathrm{P}_\vartheta(T_{i,j}(X) \le t) = \mathrm{P}_{\vartheta_1}(T_{i,j}(X) \le t),\ t \in \mathbb{R},$$

where ϑ_1 is such that $\theta_{i,j}(\vartheta_1) = \infty$ although ϑ_1 is technically not a parameter. Assumption $(RA4)$ is equivalent to $p_{i,j}(X)$ being zero almost surely under any such ϑ_1 $(i = 1, \dots, s;\ j = 1, \dots, m)$.

For any endpoint we define replicability of a bio-marker finding as the evidence of a positive effect size in at least γ out of the s studies. Let $H_1, \dots, H_m$ be the *non-replicability* null hypotheses and $K_1, \dots, K_m$ be the respective alternative hypotheses. Formally, we define

$$\begin{aligned} H_j &= \{(\theta_{1,j}, \dots, \theta_{s,j}) \in \Theta_j' \mid \theta_{i,j} \le 0 \text{ for at least } s - \gamma + 1 \text{ indices } i \in \{1, \dots, s\}\}, \\ K_j &= \{(\theta_{1,j}, \dots, \theta_{s,j}) \in \Theta_j' \mid \theta_{i,j} > 0 \text{ for at least } \gamma \text{ indices } i \in \{1, \dots, s\}\} \end{aligned}$$

for $j = 1, \dots, m$. Furthermore, write $\boldsymbol{\theta}_j(\vartheta) = (\theta_{1,j}(\vartheta), \dots, \theta_{s,j}(\vartheta))$ and $\hat{\boldsymbol{\theta}}_j = (\hat{\theta}_{1,j}, \dots, \hat{\theta}_{s,j}) : \Omega \to \mathbb{R}^s$ $(j = 1, \dots, m)$.

To make a decision about the replicability of an effect for marker j we consider the ordered p-values $p_{(1),j} < \dots < p_{(s),j}$ for the hypotheses $H_{i,j}$ $(i = 1, \dots, s)$, in the s studies. One plausible approach is to look at the γ smallest p-values and reject H_j if these are all below a suitable threshold. We therefore define $T_j(X) = 1 - p_{(\gamma),j}(X)$ and $\Gamma_j(\alpha) = (F^{-1}_{\mathrm{Beta(s-\gamma+1,1)}}(1-\alpha), 1]$, thus rejecting H_j if the γ smallest p-values are all below $1 - F^{-1}_{\mathrm{Beta}(s-\gamma+1,1)}(1-\alpha)$, where $F_{\mathrm{Beta(s-\gamma+1,1)}}$ denotes the cumulative distribution function of the $\mathrm{Beta}(s - \gamma + 1, 1)$ distribution. For the LFC-based p-value we then have $p_j^{LFC}(x) = 1 - F_{\mathrm{Beta}(s-\gamma+1,1)}\big(T_j(x)\big)$ $(j = 1, \dots, m)$.

Let G_j be the conditional cumulative distribution function of $p_j^{LFC}(X)$ given $\hat{\boldsymbol{\theta}}_j(X) \in K_j$ under any LFC $\vartheta_0 \in \Theta$ for φ_j, cf. Definition 2.1. According to Theorem 2.1 it holds that $c_j = 1 - (1 - d_j)^{n-\gamma+1}$,

$$G_j(t) = \frac{t}{1 - (1 - d_j)^{n-\gamma+1}} \mathbf{1}_{[0,1-(1-d_j)^{n-\gamma+1}]}(t) + \mathbf{1}_{(1-(1-d_j)^{n-\gamma+1},1]}(t), \quad 0 \le t \le 1,$$

and

$$p_j^{rand}(x, u_j) = u_j \mathbf{1}_{[1-(1-d_j)^{n-\gamma+1},1]}\big\{p_j^{LFC}(x)\big\} + \frac{p_j^{LFC}(x)}{1 - (1 - d_j)^{n-\gamma+1}} \mathbf{1}_{[0,1-(1-d_j)^{n-\gamma+1})}\big\{p_j^{LFC}(x)\big\}$$

for $x \in \Omega$ and $0 \le u_j \le 1$; see the proof of Lemma 2.1 in Appendix A.3.

Lemma 2.1. *If assumptions $(RA1)$ – $(RA4)$ are fulfilled, the model in Section 2.4.1 satisfies the general assumptions $(GA1)$ – $(GA4)$ from Section 2.2.*

Lemma 2.1 allows us to check the general assumptions $(GA1)$ – $(GA4)$ of the overall model by looking at the single studies. As such, it is not difficult to provide models that fulfil $(GA1)$ – $(GA4)$.

Example 2.5. *In the following, we consider models, in which we utilize either a Z-test or a t-test for each study i and endpoint j.*

(*i*) *Model 1: In each study $i = 1, \dots, s$ we consider a multiple Z-tests model. For fixed sample sizes $n_{i,j}$, $(i = 1, \dots, s;\ j = 1, \dots, m)$, we consider the observations $x \in \Omega = \mathbb{R}^{\sum_{i,j} n_{i,j}}$ as realizations of $X = \{X_k^{(i,j)} : i = 1, \dots, s, j = 1, \dots, m, k = 1, \dots, n_{i,j}\}$.*

For each study i and marker j the observations $X_1^{(i,j)}, \dots, X_{n_{i,j}}^{(i,j)}$ are stochastically independent and identically, normally distributed on $\mathbb{R}$ with expected value $\theta_{i,j}(\vartheta)$ and variance 1, where $\vartheta \in \Theta$ is the underlying parameter. It is $\Theta = \mathbb{R}^{sm}$, where we denote the parameters by $\vartheta = (\mu_{i,j} : 1 \le i \le s, 1 \le j \le m)$, such that $\theta_{i,j}(\vartheta) = \mu_{i,j}$ $(i = 1, \dots, s;\ j = 1, \dots, m)$.

As before, we test the null hypothesis $H_{i,j} = \{\mu_{i,j} \le 0\}$ against the alternative $K_{i,j} = \{\mu_{i,j} > 0\}$ $(i = 1, \dots, s;\ j = 1, \dots, m)$. A consistent and unbiased estimator for $\mu_{i,j}$ is $\hat{\theta}_{i,j}(X) = \bar{X}_{i,j} = n_{i,j}^{-1} \sum_{k=1}^{n_{i,j}} X_k^{(i,j)}$, which is normally distributed on $\mathbb{R}$ with expected value $\mu_{i,j}$ and variance $1/n_{i,j}$ $(i = 1, \dots, s;\ j = 1, \dots, m)$.

Furthermore, we choose the test statistic $T_{i,j}(X) = \hat{\theta}_{i,j}(X)$ and rejection region $\Gamma_{i,j}(\alpha) = \big(\Phi^{-1}_{(0,1/n_{i,j})}(1-\alpha), \infty\big)$, where $\Phi_{(\mu,\sigma^2)}$ is the cumulative distribution function of the normal distribution on $\mathbb{R}$ with expected value μ and variance σ^2 $(i = 1, \dots, s;\ j = 1, \dots, m)$

Assumptions $(RA1)$ – $(RA3)$ have already been discussed before, with $c_{i,j} = d_j = 1/2$ for all i, j, and $(RA4)$ is clear. Under this model, due to Lemma 2.1, assumptions $(GA1)$ – $(GA4)$ are fulfilled.

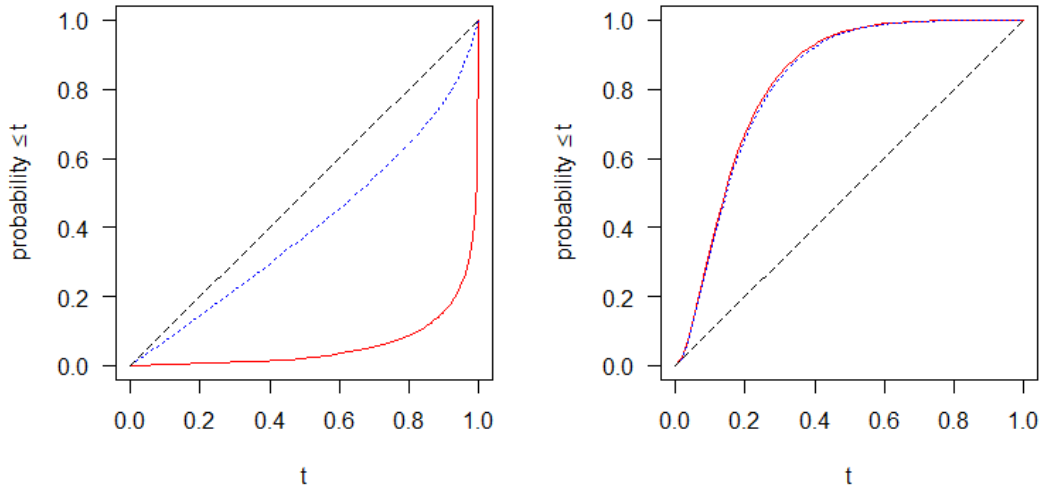


Figure 2.1: A comparison of the cumulative distribution functions of $p_j^{LFC}(X)$ (solid) and $p_j^{rand}(X, U_j)$ (dotted) under Model 1 for $s = 10, \gamma = 6$, and $n_{1,j} = \cdots = n_{s,j} = 50$. The true parameters are $\boldsymbol{\theta}_j(\vartheta) = (-1.5\, n_{1,j}^{-1/2}, \ldots, -1.5\, n_{s-\gamma+1,j}^{-1/2}, 1, \ldots, 1)$ on the left and $\boldsymbol{\theta}_j(\vartheta) = (2\, n_{1,j}^{-1/2}, \ldots, 2\, n_{s,j}^{-1/2})$ on the right side. For comparison, the dashed line depicts the cumulative distribution function of the standard uniform distribution, which is given by $t \mapsto t$ for $t \in [0, 1]$.

(*ii*) *Model* 2*: For multiple t-tests instead of Z-tests, where the observations have unknown variance (cf. Model* 2 *in Dickhaus (2013)), assumptions* $(RA1)$ – $(RA4)$ *are analogous to verify, which again results in an overall model that fulfils assumptions* $(GA1)$ – $(GA4)$.

We give a sufficient condition for the validity of the randomized p-values p_j^{rand} that result from our model setup based on Theorem 2.3.

Theorem 2.5. *Let a model as above be given, such that assumptions* $(RA1)$ – $(RA4)$ *are fulfilled, and let* $j \in \{1, \ldots, m\}$ *be fixed.*

If, for all $i = 1, \ldots, s$ *and* $\vartheta, \vartheta_0 \in \Theta$ *with* $\boldsymbol{\theta}_j(\vartheta), \boldsymbol{\theta}_j(\vartheta_0) \in H_j$ *and* $\theta_{i,j}(\vartheta) \leq 0 = \theta_{i,j}(\vartheta_0)$*, it holds* $T_{i,j}(X)^{(\vartheta)} \leq_{\mathrm{hr}} T_{i,j}(X)^{(\vartheta_0)}$*, then* p_j^{rand} *is a valid p-value.*

Remark 2.2. (*i*) *Theorem* 2.5 *still holds if we replace the hazard rate order* $\leq_{\mathrm{hr}}$ *by the likelihood ratio order* $\leq_{\mathrm{lr}}$.

(*ii*) *Under a model that fulfils Theorem 2.5 the randomized p-value* $p_{i,j}^{rand}$ *resulting from study* i *and marker* j *is valid as well, cf. Theorem 2.3.*

Example 2.6.
The randomized p-values p_j^{rand} $(j = 1, \ldots, m)$*, in Models* 1 *and* 2*, as introduced in Example 2.5 are valid. Here, we show that for Model* 1.

Recall that $T_{i,j}(X) = \hat{\theta}_{i,j}(X)$ *is normally distributed on* $\mathbb{R}$ *with expected value* $\theta_{i,j}(\vartheta)$ *and variance* $1/n_{i,j}$ *under* $\vartheta \in \Theta$*, where* $n_{i,j}$ *is the fixed sample size* $(i = 1, \ldots, s;\ j = 1, \ldots, m)$*. For* $i \in \{1, \ldots, s\}$ *and* $\vartheta, \vartheta_0 \in \Theta$*, such that* $\boldsymbol{\theta}_j(\vartheta), \boldsymbol{\theta}_j(\vartheta_0) \in H_j$ *and* $\theta_{i,j}(\vartheta) \leq 0$*,* $\theta_{i,j}(\vartheta_0) = 0$*, it holds* $T_{i,j}(X)^{(\vartheta)} \leq_{\mathrm{lr}} T_{i,j}(X)^{(\vartheta_0)}$*, cf. Example* 2.2*. It follows from Theorem* 2.5*, that* p_j^{rand} *is a valid p-value* $(j = 1, \ldots, m)$.

In Fig. 2.1*, for a fixed* $j \in \{1, \ldots, m\}$*, we compare the cumulative distribution functions of* p_j^{LFC} *and* p_j^{rand} *for* $\boldsymbol{\theta}_j(\vartheta) \in H_j$*,* $\boldsymbol{\theta}_j(\vartheta) = (-1.5\, n_{1,j}^{-1/2}, \ldots, -1.5\, n_{s-\gamma+1,j}^{-1/2}, 1, \ldots, 1)$*, and* $\boldsymbol{\theta}_j(\vartheta) \in K_j$*,* $\boldsymbol{\theta}_j(\vartheta) = (2\, n_{1,j}^{-1/2}, \ldots, 2\, n_{s,j}^{-1/2})$*, in the first and second graph, respectively, where we set* $s = 10, \gamma = 6$*, and the sample sizes to* $n_{1,j} = \cdots = n_{s,j} = 50$.

The left graph shows that the randomized p-value $p_j^{rand}(X, U_j)$ *is stochastically not larger than the LFC-based p-value* $p_j^{LFC}(X)$ *but remains valid, i.e. not smaller than a uniform distribution on* $[0, 1]$*. It is apparent that* $p_j^{rand}(X, U_j)$ *comes much closer to the uniform distribution on* $[0, 1]$*. The right graph, however, illustrates that the randomized p-value* $p_j^{rand}(X, U_j)$ *is stochastically (slightly) larger than the LFC-based p-value* $p_j^{LFC}(X)$*, under a parameter* $\vartheta \in \Theta$ *with* $\boldsymbol{\theta}_j(\vartheta) \in K_j$.

2.5 Estimation of the proportion of true null hypotheses

2.5.1 Motivation

In this section we demonstrate how randomized p-values generally lead to a more precise estimation of the proportion π_0 of true null hypotheses than LFC-based ones, when used in the Schweder-Spjøtvoll estimator. This is useful for data-adaptive multiple test procedures, but knowing $m_0 = m \cdot \pi_0$ can also be valuable in itself. In bio-marker identification, for instance, the size of $m_1 = m - m_0$ can be an indicator for the complexity of the examined disease.

The Schweder-Spjøtvoll estimator is given by $\hat{\pi}_0 \equiv \hat{\pi}_0(\lambda) = \{1-\hat{F}_m(\lambda)\}/(1-\lambda)$, where $\hat{F}_m$ denotes the empirical cumulative distribution function of the m marginal p-values, and $\lambda \in [0,1)$ is a tuning parameter (Schweder and Spjøtvoll, 1982). The estimator $\hat{\pi}_0(\lambda)$ represents the proportion of p-values exceeding λ divided by the expected proportion of the latter given uniformly distributed p-values. Assuming that the p-values corresponding to the false null hypotheses are (almost surely) below λ, and the ones corresponding to the true null hypotheses are uniformly distributed on $[0,1]$, the term $1-\hat{F}_m(\lambda)$ is then, in expectation equal to $(1-\lambda)\pi_0$, leading to an unbiased estimator $\hat{\pi}_0(\lambda)$ for π_0. Graphically, the estimator $\hat{\pi}_0(\lambda)$ equals one minus the intercept of the straight line connecting $(\lambda, \hat{F}_m(\lambda))$ with $(1,1)$. We sometimes write $\hat{\pi}_0^{LFC}$ and $\hat{\pi}_0^{rand}$ to emphasize the usage of the LFC-based or the randomized p-values in the estimator $\hat{\pi}_0$, respectively. In the present work, we demonstrate the usefulness of utilizing randomized p-values in $\hat{\pi}_0$ mainly by means of computer simulations and real data analysis. Mathematical investigations regarding the bias and the mean squared error of $\hat{\pi}_0$, when used with randomized p-values, can be found in Hoang and Dickhaus (2022).

2.5.2 Simulations

First, we simulated one realization of the empirical cumulative distribution functions of $(p_j^{LFC})_{j=1,\ldots,m}$ and $(p_j^{rand})_{j=1,\ldots,m}$, computed on the same data, where we chose $m = 500$, $s = 10$, $\gamma = 6$, and $\pi_0 = 0.7$. Hence, we consider 10 studies, each examining the same 500 endpoints, where $m_1 = 150$ of these have a positive effect in at least $\gamma = 6$ and the other $m_0 = 350$ have a positive effect in less than 6 studies. We call these true and false endpoints, respectively, according to whether their respective null hypotheses are true or false. For each true and false endpoint we drew the number of studies with positive effects binomially from $\{0,\ldots,\gamma-1\}$ and $\{\gamma,\ldots,s\}$ with success probabilities $p_0 = 0.8$ and $p_1 = 0.8$, respectively. For each study i and endpoint j we set the sample size to $n_{i,j} = 50$ and drew the non-positive effect $\theta_{i,j}(\vartheta)$ uniformly from $(\mu_{\min}\ 50^{-1/2}, 0]$ and the positive effects uniformly from $(0, \mu_{\max}]$, where we chose $\mu_{\min} = -2.5$ and $\mu_{\max} = 1.5$ $(i = 1,\ldots,s;\ j = 1,\ldots,m)$.

Figure 2.2 displays one realization of the empirical cumulative distribution functions of the marginal, LFC-based and the marginal, randomized p-values, respectively. The estimation $\hat{\pi}_0(\lambda)$ is more accurate if the empirical cumulative distribution function of the utilized marginal p-values at point $t = \lambda$ is closer to the thick line connecting $(0, 1-\pi_0)$ with $(1,1)$, also at $t = \lambda$. Clearly, $\hat{\pi}_0^{rand}(\lambda)$ is more accurate than $\hat{\pi}_0^{LFC}(\lambda)$ for $0.1 < \lambda < 1$. Also, $\hat{\pi}_0^{rand}(\lambda)$ is more stable with respect to λ, as the lower curvature of the respective empirical cumulative distribution function suggests.

Next, we calculated the expected values of $\hat{\pi}_0^{LFC}(\lambda)$ and $\hat{\pi}_0^{rand}(\lambda)$ for different values of $\pi_0, (\mu_{\min}, \mu_{\max})$, and γ, where we set $s = 10, m = 100$, and $\lambda = 1/2$. Apart from that, we drew everything else as before.

We looked at each combination of $\pi_0 \in \{0.6, 0.7, 0.8, 0.9\}, (\mu_{\min}, \mu_{\max}) \in \{(0,2), (-0.5,3), (-1,4), (-1.5,5)\}$, and $\gamma \in \{2,4,6,10\}$. Each pair $(\mu_{\min}, \mu_{\max})$ was chosen such that $|\mu_{\min}|$ and $\mu_{\max}$ increase simultaneously, and thus, model uncertainty increases in both directions.

Figure 2.3 illustrates the effect of γ on the expected value of $\hat{\pi}_0(1/2)$ in each setting when utilizing LFC-based p-values (crosses) or randomized p-values (circles), respectively. For the exact numbers we refer to Table 2.1 and Table 2.2, respectively. All values have been double-checked by Monte Carlo simulations.

According to Lemma 1 in Dickhaus et al. (2012), the Schweder-Spjøtvoll estimator $\hat{\pi}_0(\lambda)$ applied to either of the p-values has a non-negative bias. In each setting we observe lower expected values and therefore lower bias for $\hat{\pi}_0^{rand}(1/2)$ than for $\hat{\pi}_0^{LFC}(1/2)$. The difference between the expectations tend to be more emphasized for higher γ and higher model uncertainty, i.e. for larger $\mu_{\max}$ and $|\mu_{\min}|$.

As mentioned before, we expect a more stable estimation $\hat{\pi}_0(\lambda)$ of π_0 with respect to λ when utilizing the randomized p-values. For the parameter settings $\pi_0 = 0.6, \gamma = 8, \mu_{\min} = -2$, and $\mu_{\max} = 4$, Fig. 2.4 compares the expected values of $\hat{\pi}_0(\lambda)$ for $\lambda = 0.1, 0.2, \ldots, 0.9$ and either p-values. The figure confirms the expected behavior. We checked many other configurations, too. They lead to similar results, although not always so pronounced.

Finally, we examined the higher variance of $\hat{\pi}_0^{rand}(1/2)$ when utilizing the randomized p-values $(p_j^{rand})_j$, due to the additional randomization by U_j $(j = 1,\ldots,m)$. We calculated the standard de-

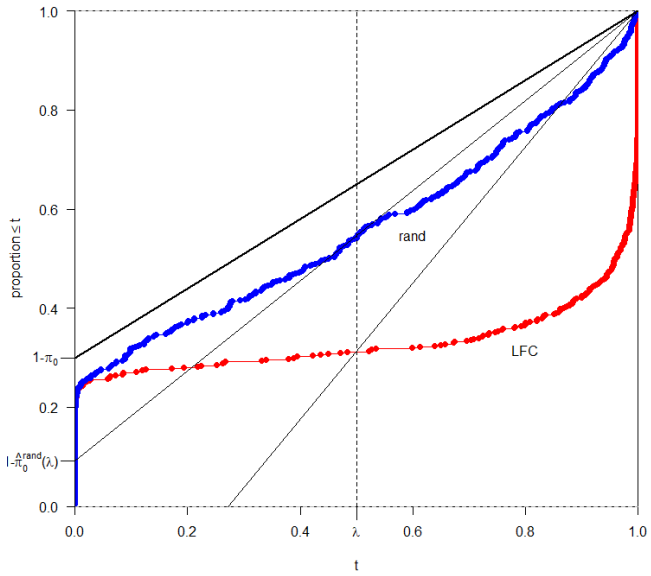


Figure 2.2: One realization of the empirical cumulative distribution functions of the marginal p-values $(p_j^{LFC})_j$ and $(p_j^{rand})_j$, respectively, under Model 1 for $m = 500, s = 10, \gamma = 6$, and $\pi_0 = 0.7$. The thick, straight line connects the points $(1, 1)$ and $(0, 1 - \pi_0)$. The two thinner, straight lines connect the points $(1, 1)$ and $(\lambda, \hat{F}_m(\lambda))$ and intersect the vertical axis at $(0, 1 - \hat{\pi}_0^{rand}(\lambda))$ or $(0, 1 - \hat{\pi}_0^{LFC}(\lambda))$ for the respective p-values.

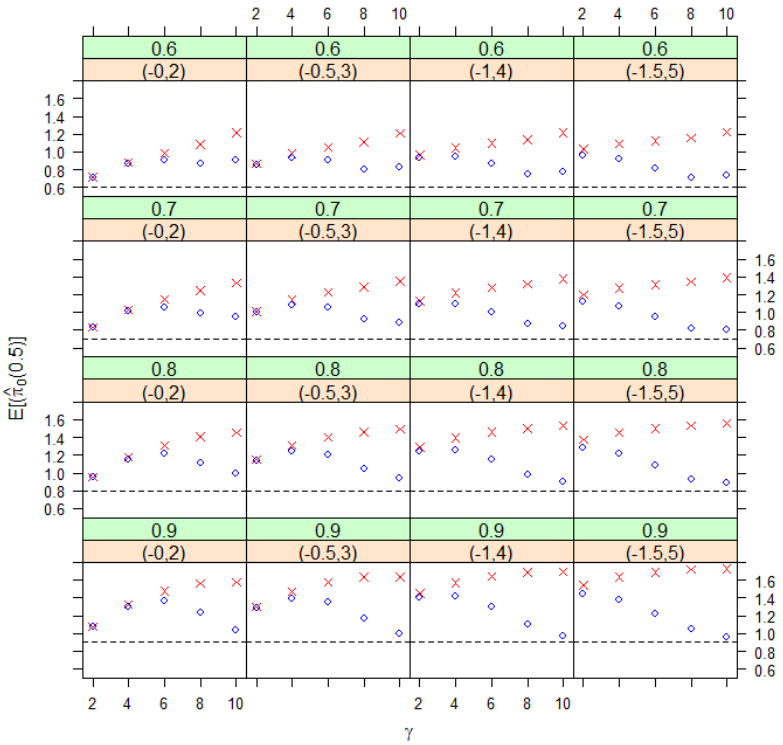


Figure 2.3: A comparison of the expected values of $\hat{\pi}_0(1/2)$ utilizing either $(p_j^{LFC})_j$ (crosses) or $(p_j^{rand})_j$ (circles) in all considered settings. In each graph the horizontal axis displays the parameter γ. The graphs differ in their choice of $(\mu_{\min}, \mu_{\max})$ (columns) and π_0 (rows). Dashed lines represent the true values of the proportion π_0 of true null hypotheses.

Table 2.1: Expected values of $\hat{\pi}_0^{LFC}(1/2)$ using the LFC-based p-values $(p_j^{LFC})_{j=1,\dots,m}$ in Model 1 with $s = 10$

$\gamma = 2$	π_0	0.6	0.7	0.8	0.9
$(\mu_{\min}, \mu_{\max})$					
(0,2)		0.7162	0.8356	0.9550	1.0743
(-0.5,3)		0.8654	1.0097	1.1539	1.2981
(-1,4)		0.9668	1.1280	1.2891	1.4503
(-1.5,5)		1.0295	1.2010	1.3726	1.5442

$\gamma = 4$	π_0	0.6	0.7	0.8	0.9
$(\mu_{\min}, \mu_{\max})$					
(0,2)		0.8820	1.0288	1.1756	1.3225
(-0.5,3)		0.9800	1.1433	1.3066	1.4699
(-1,4)		1.0471	1.2216	1.3961	1.5706
(-1.5,5)		1.0884	1.2698	1.4512	1.6326

$\gamma = 6$	π_0	0.6	0.7	0.8	0.9
$(\mu_{\min}, \mu_{\max})$					
(0,2)		0.9883	1.1506	1.3129	1.4752
(-0.5,3)		1.0510	1.2251	1.3993	1.5735
(-1,4)		1.0964	1.2786	1.4608	1.6429
(-1.5,5)		1.1247	1.3118	1.4989	1.6859

$\gamma = 8$	π_0	0.6	0.7	0.8	0.9
$(\mu_{\min}, \mu_{\max})$					
(0,2)		1.0829	1.2440	1.4051	1.5663
(-0.5,3)		1.1093	1.2841	1.4589	1.6336
(-1,4)		1.1366	1.3194	1.5022	1.6850
(-1.5,5)		1.1547	1.3423	1.5299	1.7175

$\gamma = 10$	π_0	0.6	0.7	0.8	0.9
$(\mu_{\min}, \mu_{\max})$					
(0,2)		1.2141	1.3335	1.4529	1.5723
(-0.5,3)		1.2069	1.3504	1.4939	1.6374
(-1,4)		1.2165	1.3751	1.5337	1.6923
(-1.5,5)		1.2241	1.3923	1.5606	1.7288

viation of $\hat{\pi}_0(1/2)$ utilizing either the LFC-based p-values or the randomized p-values, for the same settings as we did for Fig. 2.3 via Monte Carlo simulations. For the results we refer to Tables 2.3 and 2.4.

Using $(p_j^{rand})_{j=1,\dots,m}$, we observe higher standard deviations of $\hat{\pi}_0^{rand}(1/2)$ in each setting short of one. However, the largest standard deviation when using the randomized p-values across all considered settings was below 0.1. We also compared the mean squared errors of $\hat{\pi}_0^{rand}$ and $\hat{\pi}_0^{LFC}$ in all considered parameter settings. In each setting the mean squared error was higher when using the LFC-based p-values.

2.6 An application to multiple Crohn's disease genome-wide association studies

We looked at the data from multiple genome-wide association studies with the goal of identifying susceptibility loci for Crohn's disease (Franke et al., 2010). The authors looked at six distinct genome-wide association studies, further dividing two of these resulting in a total of eight distinct studies, which comprised 6,333 disease cases and 15,056 healthy controls altogether. In their discovery panel, they combined these eight studies in a meta-analysis and looked at the most promising features in a further replication panel. For lack of data on the latter part we only looked at the data stemming from the original eight studies.

In their work, the authors applied multiple Z-tests for the logarithmic odds ratios in each scan and combined them to test for two-sided associations of phenotype and genotype at each of m loci. For more details on the statistical framework for such type of studies we refer to Chapter 9 of Dickhaus (2014) and to Sections 2 and 3 of Dickhaus et al. (2015). For these tests, randomized p-values in the sense of

Table 2.2: Expected values of $\hat{\pi}_0^{rand}(1/2)$ using $(p_j^{rand})_{j=1,\ldots,m}$ in Model 1 with $s = 10$

$\gamma = 2$	π_0	0.6	0.7	0.8	0.9
$(\mu_{\min}, \mu_{\max})$					
(0,2)		0.7149	0.8340	0.9532	1.0723
(-0.5,3)		0.8557	0.9983	1.1410	1.2836
(-1,4)		0.9356	1.0915	1.2475	1.4034
(-1.5,5)		0.9609	1.1210	1.2811	1.4413
$\gamma = 4$	π_0	0.6	0.7	0.8	0.9
$(\mu_{\min}, \mu_{\max})$					
(0,2)		0.8660	1.0101	1.1543	1.2985
(-0.5,3)		0.9312	1.0863	1.2415	1.3966
(-1,4)		0.9432	1.1004	1.2576	1.4148
(-1.5,5)		0.9169	1.0697	1.2225	1.3752
$\gamma = 6$	π_0	0.6	0.7	0.8	0.9
$(\mu_{\min}, \mu_{\max})$					
(0,2)		0.9121	1.0615	1.2110	1.3604
(-0.5,3)		0.9033	1.0528	1.2023	1.3519
(-1,4)		0.8645	1.0080	1.1516	1.2951
(-1.5,5)		0.8144	0.9498	1.0852	1.2206
$\gamma = 8$	π_0	0.6	0.7	0.8	0.9
$(\mu_{\min}, \mu_{\max})$					
(0,2)		0.8689	0.9914	1.1139	1.2365
(-0.5,3)		0.8027	0.9251	1.0474	1.1698
(-1,4)		0.7494	0.8670	0.9846	1.1022
(-1.5,5)		0.7081	0.8209	0.9337	1.0465
$\gamma = 10$	π_0	0.6	0.7	0.8	0.9
$(\mu_{\min}, \mu_{\max})$					
(0,2)		0.9121	0.9552	0.9982	1.0413
(-0.5,3)		0.8275	0.8852	0.9430	1.0007
(-1,4)		0.7749	0.8409	0.9068	0.9727
(-1.5,5)		0.7404	0.8117	0.8830	0.9543

Dickhaus (2013) can also be defined. However, in such a two-sided setting each parameter in the null hypotheses $H_j = \{(\theta_{1,j}, \ldots, \theta_{s,j}) \in \mathbb{R}^s : \theta_{k,j} = 0 \text{ for at least } s - \gamma + 1 \text{ indices } k\}$ $(j = 1, \ldots, m)$, would lie next to the respective alternative $K_j = \mathbb{R}^s \setminus H_j$ making each one an LFC for their respective null hypothesis. In spite of the composite nature of the null hypotheses, the LFC-based p-values would then hold a uniform distribution under any parameter in the null hypothesis and using randomized p-values would be unnecessary.

Instead, we looked at the original Z-scores for associations in one fixed direction between the investigated single-nucleotide polymorphisms and Crohn's disease. Each of the eight studies investigated the effect of 953,241 single-nucleotide polymorphisms on Crohn's disease. We designated one of the studies as a primary study and selected the most promising features with the Benjamini-Hochberg step-up procedure at false discovery rate (Benjamini and Hochberg, 1995) levels $q = 0.2$ or $q = 0.5$. After selection we ended up with $m = 630$ and $m = 2{,}257$ single-nucleotide polymorphisms, respectively, and tested their associations' replicability based on the remaining $s = 7$ studies. For both false discovery rate levels q, we looked at the choices $\gamma = 2$ and $\gamma = 4$, and calculated the LFC-based and randomized p-values as in the model described in Section 2.4.1. For these values of γ, we have $c_j = 2^{-(7-2+1)} = 2^{-6}$ and $c_j = 2^{-(7-4+1)} = 2^{-4}$, respectively, where $d_j = 1/2$ results from the model $(j = 1, \ldots, m)$.

We then calculated the Schweder-Spjøtvoll estimator $\hat{\pi}_0(\lambda)$ with $\lambda = 1/2$ for the four parameter settings. Figure 2.5 illustrates the empirical cumulative distribution functions of the LFC-based and the randomized p-values, respectively, after selection. The values for the settings $(q, \gamma) = (0.2, 2), (0.2, 4), (0.5, 2), (0.5, 4)$ are, in order,

$$\begin{aligned}(\hat{\pi}_0^{LFC}(\lambda), \mathrm{E}(\hat{\pi}_0^{rand}(\lambda))) \quad = \quad & (0.4603, 0.4651), (0.8857, 0.7572), \\ & (0.9880, 0.9668), (1.5498, 1.3013),\end{aligned}$$

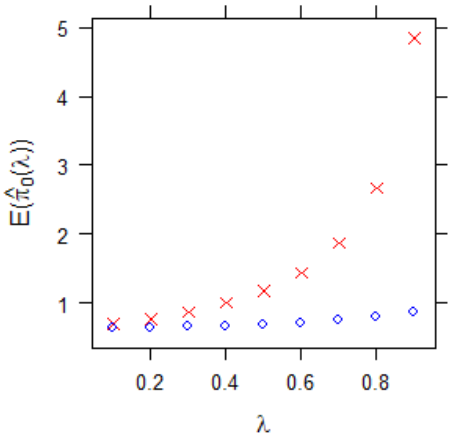


Figure 2.4: The expected values of $\hat{\pi}_0(\lambda)$ for different tuning parameters λ under Model 1 for $m = 100, s = 10, n_{i,j} = 50, \gamma = 8, \pi_0 = 0.6, \mu_{\min} = -2, \mu_{\max} = 4$, and $p_0 = p_1 = 0.8$ when using either the LFC-based p-values (crosses) or the randomized p-values (circles).

where E refers to the randomness of $(U_j : 1 \leq j \leq m)$. These are also displayed above their corresponding graphs. The standard deviation for the estimation using the randomized p-values are $\text{var}^{1/2}(\hat{\pi}_0^{rand}(\lambda)) = 0.00276, 0.01542, 0.00377, 0.01109$ for the respective settings in the same order. The values corresponding to the use of the randomized p-values are a result of Monte Carlo simulations with 100,000 repetitions in each setting.

Let us discuss these results. An increase in the false discovery rate level q, used in the selection process, increases the proportion π_0 which favours the use of the randomized p-values. A higher γ increases the proportion π_0 and reduces the constant $c_j = 2^{-(7-\gamma+1)}$ $(j = 1, \ldots, m)$, both benefiting the performance of the estimator $\hat{\pi}_0^{rand}(\lambda)$ in terms of bias. Choosing q and γ both too high can lead to a too large π_0 making it difficult to estimate the latter as the example with $q = 0.5$ and $\gamma = 4$ demonstrates. On the other hand, choosing both q and γ too low results in a low proportion of true null hypotheses, of which the remaining do not offer high enough deviation from the alternative to facilitate the usage of randomized p-values as the example with $q = 0.2$ and $\gamma = 2$ demonstrates.

2.7 Discussion

In the context of simultaneous testing of composite null hypotheses, we have demonstrated that the usage of randomized p-values leads to a more accurate estimation of π_0 when compared with the usage of LFC-based p-values. We have explicitly demonstrated this for the Schweder-Spjøtvoll estimator $\hat{\pi}_0$. The higher estimation variances induced by the uniform random variates used for randomization are in most cases negligible, so that the mean squared error is lower for $\hat{\pi}_0^{rand}$ than for $\hat{\pi}_0^{LFC}$.

Our theory applies to any choice of the parameter $\gamma = 2, \ldots, s$. We have not further discussed the choice of γ nor do we make recommendations in this work. Choosing γ close to s results in strong replicability statements, but potentially only few rejections. On the other hand, in the presence of a very large number of studies s, replicability statements may not be suitable when choosing $\gamma = 2$. Thus, one could make γ dependent on s, like $\gamma = \beta s$ for $\beta \in (0, 1)$. Alternatively, instead of pre-defining γ, we could for each $j = 1, \ldots, m$ determine the largest $\gamma = \gamma(j)$, for which we would still reject H_j. It is then possible to declare replicability for endpoint j if $\gamma(j)/s > \beta$ holds, where $\beta \in (0, 1)$ is pre-defined.

Furthermore, we have not discussed the incorporation of the estimated proportion of true null hypotheses in so-called adaptive multiple tests. Blanchard and Roquain (2009) presented a categorization of adaptive procedures that divide between plug-in, two-stage and one-stage procedures, and provided adaptive procedures that control the false discovery rate. Finner and Gontscharuk (2009) investigated the problem of controlling the family-wise error rate when using an estimator of π_0 as a plug-in estimator in single-step or step-down procedures. Bogomolov and Heller (2018) gave an adaptive procedure that incorporates estimations of the proportion of true null hypotheses among the selected features and controls the false discovery rate for replicability analysis with two studies. It remains to be investigated to what extent the usage of randomized p-values can improve the power of such adaptive procedures. In the case of $s = 1$, some results in this direction can be found in Dickhaus (2013). These results indicate, that the power gain can be substantial.

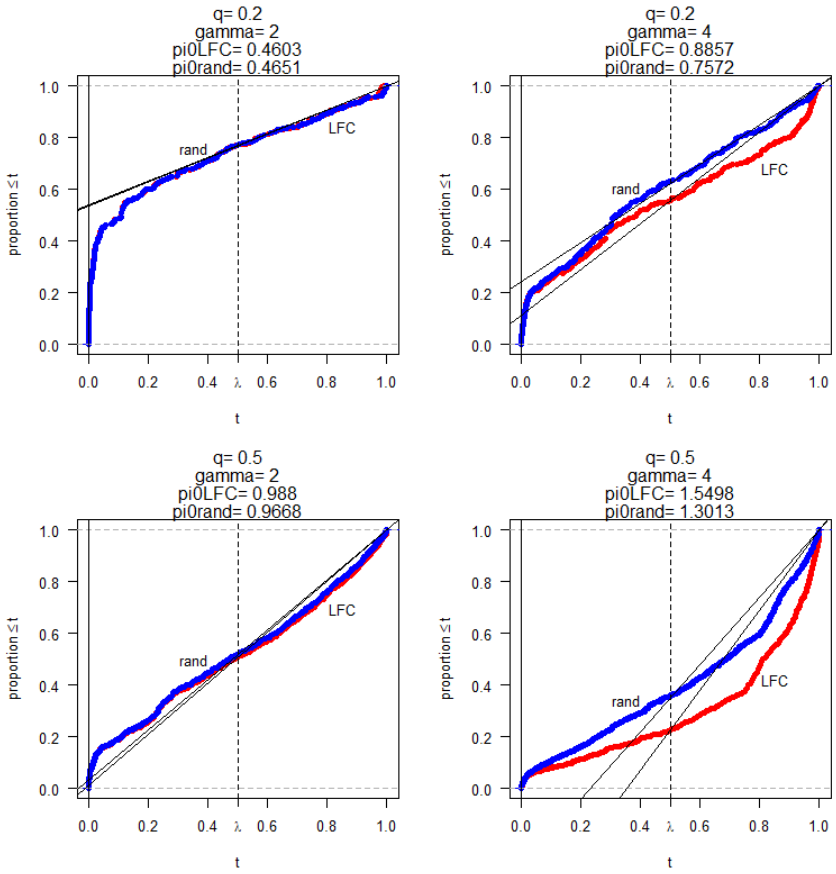


Figure 2.5: The empirical cumulative distribution functions of the LFC-based and the randomized p-values, respectively, in the multiple Crohn's disease genome-wide associations studies example after selection. Selection has been conducted with the Benjamini–Hochberg step-up procedure with false discovery rates $q = 0.2, 0.5$, and the p-values are calculated according to the model as described in Section 2.4 with $\gamma = 2, 4$. The straight lines connect the points $(1, 1)$ and $(\lambda, \hat{F}_m(\lambda))$, and intersect the vertical axis in the point $(0, 1 - \hat{\pi}_0(\lambda))$, where $\lambda = 1/2$. The values $\hat{\pi}_0^{LFC}(\lambda)$ and $\mathrm{E}(\hat{\pi}_0^{rand}(\lambda))$ are displayed above their respective graphs as pi0LFC and pi0rand, respectively.

Table 2.3: Empirical standard deviations for $\hat{\pi}_0(1/2)$ using $(p_j^{LFC})_{j=1,\dots,m}$ in Model 1 with $s = 10$, resulting from a Monte Carlo simulation with 10,000 repetitions

$\gamma = 2$	π_0	0.6	0.7	0.8	0.9
$(\mu_{\min}, \mu_{\max})$					
(0,2)		0.0758	0.0830	0.0879	0.0932
(-0.5,3)		0.0696	0.0747	0.0809	0.0856
(-1,4)		0.0614	0.0662	0.0708	0.0745
(-1.5,5)		0.0545	0.0585	0.0629	0.0666

$\gamma = 4$	π_0	0.6	0.7	0.8	0.9
$(\mu_{\min}, \mu_{\max})$					
(0,2)		0.0685	0.0740	0.0786	0.0840
(-0.5,3)		0.0598	0.0639	0.0686	0.0739
(-1,4)		0.0513	0.0548	0.0594	0.0630
(-1.5,5)		0.0450	0.0484	0.0518	0.0554

$\gamma = 6$	π_0	0.6	0.7	0.8	0.9
$(\mu_{\min}, \mu_{\max})$					
(0,2)		0.0606	0.0655	0.0689	0.0743
(-0.5,3)		0.0517	0.0561	0.0597	0.0626
(-1,4)		0.0439	0.0470	0.0502	0.0535
(-1.5,5)		0.0379	0.0416	0.0437	0.0464

$\gamma = 8$	π_0	0.6	0.7	0.8	0.9
$(\mu_{\min}, \mu_{\max})$					
(0,2)		0.0610	0.0627	0.0647	0.0660
(-0.5,3)		0.0500	0.0527	0.0547	0.0560
(-1,4)		0.0425	0.0447	0.0460	0.0481
(-1.5,5)		0.0366	0.0376	0.0391	0.0412

$\gamma = 10$	π_0	0.6	0.7	0.8	0.9
$(\mu_{\min}, \mu_{\max})$					
(0,2)		0.0754	0.0759	0.0752	0.0751
(-0.5,3)		0.0653	0.0649	0.0655	0.0650
(-1,4)		0.0560	0.0556	0.0560	0.0557
(-1.5,5)		0.0485	0.0482	0.0487	0.0475

Another advantage of using randomized p-values in $\hat{\pi}_0$ (and potentially other statistical procedures which operate on marginal p-values and rely on concentration properties of their empirical cumulative distribution function around its expectation) is that this robustifies $\hat{\pi}_0$ to a certain extent against dependencies among the p-values. Namely, when calculating the randomized p-values, the U_j's are generated independently of each other. Therefore, the strength of dependency among the randomized p-values is often much less pronounced than that among the LFC-based p-values. This has beneficial consequences, as demonstrated for instance by Hoang and Dickhaus (2022). For a similar recent investigation, see Neumann et al. (2021).

In future work, we will compare our proposed methodology with other recent approaches to dealing with conservative p-values in the context of multiple testing, in particular the approaches by Tian and Ramdas (2019) and by Zhao et al. (2019). Furthermore, it may also be of interest to study more general randomized p-values of the form given in (2.3) for a potentially smooth function w_j of the form $w_j(X) = f_j(p_j^{LFC}(X))$, where $f_j : [0,1] \to [0,1]$ is an increasing function with $f_j(0) = 0$ and $f_j(1) = 1$.

Finally, one challenging extension of our proposed methodology is to investigate randomized p-values for other types of summary statistics, in particular combination test statistics of Fisher- or Stouffer-Liptak-type; see, e. g., van Zwet and Oosterhoff (1967), Kim et al. (2013) and the references therein. In Appendix A.4 we compare their (non-randomized) use in $\hat{\pi}_0$ with the use of our proposed randomized p-values that result from our summary statistics. Under the same model and considering the same parameter settings as in Appendix 2.5.2 the use of the randomized p-values in the Schweder-Spjøtvoll estimator is still more accurate in most cases. Another possibility in this direction is to consider statistics derived from Bayesian models, for instance local false discovery rates or Bayes factors, as in Yekutieli (2015) and Dickhaus (2015), respectively.

Table 2.4: Empirical standard deviations for $\hat{\pi}_0(1/2)$ using $(p_j^{rand})_{j=1,\dots,m}$ in Model 1 with $s = 10$, resulting from a Monte Carlo simulation with 10,000 repetitions

$\gamma = 2$	π_0	0.6	0.7	0.8	0.9
$(\mu_{\min}, \mu_{\max})$					
(0,2)		0.0758	0.0830	0.0881	0.0933
(-0.5,3)		0.0703	0.0754	0.0815	0.0864
(-1,4)		0.0640	0.0691	0.0741	0.0782
(-1.5,5)		0.0622	0.0662	0.0720	0.0762

$\gamma = 4$	π_0	0.6	0.7	0.8	0.9
$(\mu_{\min}, \mu_{\max})$					
(0,2)		0.0696	0.0748	0.0797	0.0850
(-0.5,3)		0.0640	0.0693	0.0745	0.0795
(-1,4)		0.0627	0.0677	0.0735	0.0776
(-1.5,5)		0.0660	0.0715	0.0762	0.0808

$\gamma = 6$	π_0	0.6	0.7	0.8	0.9
$(\mu_{\min}, \mu_{\max})$					
(0,2)		0.0677	0.0729	0.0777	0.0822
(-0.5,3)		0.0667	0.0722	0.0776	0.0819
(-1,4)		0.0701	0.0750	0.0801	0.0862
(-1.5,5)		0.0733	0.0779	0.0838	0.0891

$\gamma = 8$	π_0	0.6	0.7	0.8	0.9
$(\mu_{\min}, \mu_{\max})$					
(0,2)		0.0794	0.0819	0.0861	0.0897
(-0.5,3)		0.0774	0.0832	0.0871	0.0917
(-1,4)		0.0779	0.0841	0.0882	0.0935
(-1.5,5)		0.0783	0.0836	0.0892	0.0933

$\gamma = 10$	π_0	0.6	0.7	0.8	0.9
$(\mu_{\min}, \mu_{\max})$					
(0,2)		0.0987	0.0983	0.0975	0.0985
(-0.5,3)		0.0967	0.0955	0.0971	0.0976
(-1,4)		0.0949	0.0951	0.0968	0.0960
(-1.5,5)		0.0963	0.0947	0.0947	0.0949

Chapter 3

On the usage of randomized p-values in the Schweder–Spjøtvoll estimator

This chapter is a slightly modified version of Hoang and Dickhaus (2021b) published in Annals of the Institute of Statistical Mathematics and has been reproduced here with the permission of the copyright holder. Appendix B.1 contains the original appendix and Appendix B.2 the original supplementary material to this paper.

Authors

Anh-Tuan Hoang, Institute for Statistics, University of Bremen, Bremen, Germany

Prof. Dr. Thorsten Dickhaus, Institute for Statistics, University of Bremen, Bremen, Germany

Abstract We consider multiple test problems with composite null hypotheses and the estimation of the proportion π_0 of true null hypotheses. The Schweder-Spjøtvoll estimator $\hat{\pi}_0$ utilizes marginal p-values and relies on the assumption that p-values corresponding to true nulls are uniformly distributed on $[0, 1]$. In the case of composite null hypotheses, marginal p-values are usually computed under least favorable parameter configurations (LFCs). Thus, they are stochastically larger than uniform under non-LFCs in the null hypotheses. When using these LFC-based p-values, $\hat{\pi}_0$ tends to overestimate π_0. We introduce a new way of randomizing p-values that depends on a tuning parameter $c \in [0, 1]$. For a certain value $c = c^\star$ the resulting bias of $\hat{\pi}_0$ is minimized. This often also entails a smaller mean squared error of the estimator as compared to the usage of LFC-based p-values. We analyze these points theoretically, and we demonstrate them numerically in simulations.

Summary In this chapter, we generalize the definition of the randomized p-values from Chapter 2, and investigate in more detail their benefits for the Schweder–Spjøtvoll estimator.

We consider randomized p-values now with arbitrary threshold parameter $c \in [0, 1]$. The choices $c = 0$ and $c = 1$ correspond to a uniformly distributed random variable U and the original p-value, i.e. the results of always and never randomizing, respectively. In Theorem 3.1 we give some conditions for the validity of these randomized p-values, equivalent to the definition of uniform validity in Zhao et al. (2019). Under a stricter version of the uniform validity, we show in Theorem 3.2 that randomized p-values can either decrease or increase stochastically in the parameter c.

In Section 3.4, we consider multiple PC null hypotheses, and investigate the estimation of π_0 with the Schweder-Spjøtvoll estimator $\hat{\pi}_0(\lambda)$. If each combined p-value is being randomized with the same parameter c, we show that there exists a parameter $c^\star \in [0, 1]$ that minimizes the bias of $\hat{\pi}_0(\lambda)$. In Section 3.4.2, we investigate the mean squared error of $\hat{\pi}_0(\lambda)$. The variance of $\hat{\pi}_0(\lambda)$ for $\lambda = 1/2$ decreases with c, if using independent randomized p-values with common threshold parameter c, but the MSE-minimizing parameter c^{MSE} converges to the bias-minimizing parameter $c^\star$ for $m \to \infty$. In a model in which the non-randomized p-values are positively dependent, the variance of $\hat{\pi}_0(\lambda)$ for $\lambda = 1/2$ increases monotonically in c which means that the variance decreases with the amount of randomization. Finally, we provide an estimator for $c^\star$ in Section 3.4.3.

In Section 3.5 we consider the more conservative Storey estimator $\hat{\pi}_0^+(\lambda)$ in the data-adaptive BH procedure, cf. Storey (2002). The use of valid, independent p-values in $\hat{\pi}_0^+(\lambda)$ is all that is required for the data-adaptive BH procedure to control the FDR. Furthermore, the bias-minimizing parameter $c^\star$ for $\hat{\pi}_0(\lambda)$ also minimizes the bias of $\hat{\pi}_0^+(\lambda)$. In simulations we find, that this adaptive procedure performs better with randomized p-values than with non-randomized p-values across

several degrees of dependence. Finally, we discuss the benefits of randomized p-values in models of previous literature. Meinshausen and Rice (2006) considered the problem of estimating a lower bound for $1 - \pi_0$ which gets smaller if the p-values are conservative, and Ghosal and Roy (2011) utilized a mixture model for the estimation of π_0 that suffers a similar problem if conservative p-values are employed. Finally, we compare our randomized p-values to conditional p-values in the context of global null hypotheses and online testing.

Declaration of individual contributions Co-author and supervisor Prof. Dr. Thorsten Dickhaus came up with the idea for the paper, and I developed the theoretical results including their proofs. Simulations and evaluations including the figures and tables pertaining to the simulations were done by me. The final text was written and proof-read by both authors.

3.1 Introduction

In multiple test problems with composite null hypotheses, to account for type I errors, marginal tests are usually calibrated with respect to least favorable parameter configurations (LFCs). These are parameter values in (or on the boundary of) the corresponding null hypotheses under which the marginal tests are most likely to reject. Under certain assumptions, the resulting marginal LFC-based p-values are then uniformly distributed on $[0,1]$ (Uni$[0,1]$-distributed) under LFCs, but stochastically larger than Uni$[0,1]$ under non-LFCs in the null hypothesis. Under the alternative, LFC p-values usually tend to be stochastically smaller than Uni$[0,1]$.

While the latter property is desirable in terms of protecting against type II errors, the deviation from uniformity under null hypotheses is problematic for some estimators of the proportion π_0 of true null hypotheses that use the empirical cumulative distribution function (ecdf) of all marginal p-values. We will denote the latter ecdf by $\hat{F}_m$ throughout the remainder, where m is the number of all null hypotheses. One ecdf-based estimator for π_0 was introduced by Schweder and Spjøtvoll (1982), and it is given by

$$\hat{\pi}_0 \equiv \hat{\pi}_0(\lambda) = \frac{1 - \hat{F}_m(\lambda)}{1 - \lambda}, \tag{3.1}$$

where $\lambda \in (0,1)$ is a tuning parameter.

In this work, we will investigate the bias, given by $\mathrm{bias}_\vartheta(\hat{\pi}_0) = \mathbb{E}_\vartheta[\hat{\pi}_0] - \pi_0$, and the mean squared error (MSE), given by $\mathrm{MSE}_\vartheta[\hat{\pi}_0] = \mathbb{E}_\vartheta[(\hat{\pi}_0 - \pi_0)^2]$, of $\hat{\pi}_0$ under various statistical models, where ϑ denotes the model parameter. Notice that $\mathrm{MSE}_\vartheta[\hat{\pi}_0] = \mathrm{Var}_\vartheta(\hat{\pi}_0) + \mathrm{bias}^2_\vartheta(\hat{\pi}_0)$. In the case that $\mathrm{bias}_\vartheta(\hat{\pi}_0) = 0$, $\hat{\pi}_0$ is called unbiased. Under the restriction of valid p-values (i. e., p-values that are stochastically not smaller than Uni$[0,1]$ under null hypotheses), $\hat{\pi}_0(\lambda)$ can only be unbiased, if the marginal p-values that correspond to the true null hypotheses are Uni$[0,1]$-distributed. The Schweder-Spjøtvoll estimator is an unbiased estimator if, in addition, all p-values that correspond to the false null hypotheses are smaller than λ with probability one. In general, $\hat{\pi}_0(\lambda)$ is non-negatively biased if used with valid p-values. The aforementioned properties of $\hat{\pi}_0$ follow, for example, from the calculations in Appendix I of Dickhaus et al. (2012). It is also known for a longer time (cf., e. g., the discussion by Storey et al. (2004) after their Eq. (4)), that the variance of $\hat{\pi}_0(\lambda)$ increases with increasing λ in most cases.

Non-uniformity of p-values under null hypotheses happens for instance in case of discrete models, which has been, among others, investigated by Finner and Strassburger (2007), Habiger and Peña (2011), Dickhaus et al. (2012), and Habiger (2015). The randomization approach proposed by Dickhaus et al. (2012) results in uniformly distributed p-values under simple (i. e., one-elementary) nulls. In case of composite null hypotheses, the deviation of p-values from uniformity occurs, when marginal test statistics do not have a unique distribution under the null hypotheses and the marginal tests hence cannot be calibrated precisely with respect to their type I error probabilities. To provide more uniform p-values under composite null hypotheses Dickhaus (2013) proposed randomized p-values that result from a data-dependent mixing of the LFC-based p-values and additional Uni$[0,1]$-distributed random variables that are (stochastically) independent of the data. In certain models, these randomized p-values can be simplified to have a linear structure (cf. Hoang and Dickhaus (2022)).

While accurate estimations of π_0 are valuable in themselves, they can also improve the power of existing multiple test procedures. Namely, many of such procedures are (implicitly) calibrated to control the family-wise error rate (FWER) or the false discovery rate (FDR), respectively, for the case that every null hypothesis is true, that is, in case of $\pi_0 = 1$, which is often the worst case. If some null hypotheses are false, these procedures become over-conservative. Adjusting them according to a pre-estimate of π_0 can improve the overall power of these tests. Benjamini and Hochberg (2000) discuss these so-called adaptive procedures where the original procedure is the linear step-up test from Benjamini and Hochberg (1995). Storey (2003) proved that applying the linear step-up test by Benjamini and Hochberg (1995)

at an adjusted level controls the FDR if the p-values are independent. Finner and Gontscharuk (2009) investigated the use of estimators of π_0 as plug-in estimators in single-step or step-down procedures and proved that the Bonferroni procedure at an adjusted level controls the FWER if the marginal p-values are independent. Further results and references on adaptive multiple tests (for FDR control) can be found in Heesen and Janssen (2015, 2016), and MacDonald et al. (2019).

We focus on the case of composite null hypotheses and present a new way of randomizing LFC-based p-values. To this end, we utilize a set of stochastically independent and identically Uni$[0,1]$-distributed random variables $U_1, \ldots, U_m$, which are (stochastically) independent of the data X, as well as a set of constants $c_1, \ldots, c_m$, where $c_j \in [0,1]$ for all $1 \leq j \leq m$. For a (continuously distributed) LFC-based p-value $p_j^{LFC}(X)$ we propose randomized p-values defined as

$$p_j^{rand}(X, U_j, c_j) = U_j \mathbf{1}\{p_j^{LFC}(X) \geq c_j\} + p_j^{LFC}(X) c_j^{-1} \mathbf{1}\{p_j^{LFC}(X) < c_j\}, \tag{3.2}$$

$j = 1 \ldots, m$.

In many models this definition comprises the one of Dickhaus (2013) for certain values of $c_j \in [0,1]$ (cf. Hoang and Dickhaus (2022)). It is clear that c_j determines how close p_j^{rand} is to either U_j or p_j^{LFC}. The choices $c_j = 0$ and $c_j = 1$ lead to $p_j^{rand} = U_j$ (by convention) or $p_j^{rand} = p_j^{LFC}$ (with probability one), respectively. Under certain conditions, it holds $U_j \leq_{\text{st}} p_j^{rand} \leq_{\text{st}} p_j^{LFC}$ under the j-th null hypothesis and $p_j^{LFC} \leq_{\text{st}} p_j^{rand} \leq_{\text{st}} U_j$ under the j-th alternative for all $c_j \in [0,1]$, where $\leq_{\text{st}}$ denotes the stochastic order (see, e. g., Theorem 3.2 below). For definitions and notations of the stochastic order $\leq_{\text{st}}$ and further ones, we refer to Appendix B. While Uni$[0,1]$-distributed p-values are desirable under null hypotheses, we want to keep them small under alternatives. When using $p_1^{rand}(X, U_1, c_1), \ldots, p_m^{rand}(X, U_m, c_m)$ in $\hat{\pi}_0$, we discuss how the choice of the constants $c_1, \ldots, c_m$ affects the bias of $\hat{\pi}_0$. Under the restriction of identical c_j's, we find that there exists a $c^\star \in [0,1]$ for which $\hat{\pi}_0$ has minimal bias when using $p_1^{rand}(X, U_1, c^\star), \ldots, p_m^{rand}(X, U_m, c^\star)$. We mainly focus on the bias instead of the MSE, since it turns out that $c^\star$ is close to the MSE-minimizing value of c, especially if m is large. Furthermore, if the LFC-based p-values are positively dependent, the variance of the Schweder-Spjøtvoll estimator is much higher when using the LFC-based p-values instead of the randomized p-values. The problem of minimizing the MSE may in this case lead to the trivial choice of $c = 0$.

The rest of the work is organized as follows. In Section 3.2 we provide the model framework. In Section 3.3 we analyze properties of our proposed randomized p-values, and compare them to the LFC-based ones. Section 3.4 presents theoretical and numerical results regarding the bias and the MSE of $\hat{\pi}_0$ when used with the proposed randomized p-values. Section 3.5 illustrates the performance of resulting data-adaptive multiple tests for control of the FDR. In Section 3.6, we compare our proposed methodology with other approaches from the literature. We conclude with a discussion in Section 3.7.

3.2 Model Setup

We consider a statistical model $(\Omega, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$, where ϑ denotes the parameter of the model and Θ the corresponding parameter space. In the context of multiple testing we define a derived parameter $\theta = \theta(\vartheta) = (\theta_1(\vartheta), \ldots, \theta_m(\vartheta))^\top$ with values in $\mathbb{R}^m$, $m \geq 2$. The j-th component $\theta_j(\vartheta)$ of this derived parameter is assumed to be the object of interest in the j-th null hypothesis H_j, $j = 1, \ldots, m$, where the family of m null hypotheses $H_1, \ldots, H_m$ and the family of their corresponding alternatives $K_1, \ldots, K_m$ consist of non-empty Borel sets of $\mathbb{R}$. For each $j = 1, \ldots, m$ we test $\theta_j(\vartheta) \in H_j$ against $\theta_j(\vartheta) \in K_j = \mathbb{R} \setminus H_j$.

We assume that for each $j = 1, \ldots, m$ a test statistic $T_j : \Omega \to \mathbb{R}$ and a rejection region $\Gamma_j(\alpha) \subset \mathbb{R}$ are given, where $\alpha \in (0,1)$ denotes a fixed, local significance level. We denote by $x \in \Omega$ the realization of X. The test statistics $\{T_j(X)\}_{1 \leq j \leq m}$ are assumed to have absolutely continuous distributions with respect to the Lebesgue measure under any $\vartheta \in \Theta$. The marginal tests φ_j for testing H_j versus K_j are given by $\varphi_j(X) = \mathbf{1}\{T_j(X) \in \Gamma_j(\alpha)\}$, where $\varphi_j(x) = 1$ means rejection of H_j in favor of K_j and $\varphi_j(x) = 0$ means that H_j is retained, for observed data x and $1 \leq j \leq m$. Note, that we do not make any (general) assumptions about the dependence structure among the different test statistics at this point.

Furthermore, we make the following additional general assumptions:

($A1$) Nested rejection regions: For every $j = 1, \ldots, m$ and $\alpha' < \alpha$, it holds that $\Gamma_j(\alpha') \subseteq \Gamma_j(\alpha)$.

($A2$) For every $j = 1, \ldots, m$, it holds $\sup_{\vartheta : \theta_j(\vartheta) \in H_j} \mathbb{P}_\vartheta(T_j(X) \in \Gamma_j(\alpha)) = \alpha$.

($A3$) The set of LFCs for φ_j, i. e., the set of parameter values that yield the supremum in ($A2$), does not depend on α.

Under assumption $(A1)$, rejections at significance levels α' always imply rejections at larger significance levels $\alpha > \alpha'$. Assumption $(A2)$ means that under any LFC for φ_j the rejection probability is exactly α.

LFC-based p-values for the marginal tests $\{\varphi_j\}_{1\leq j\leq m}$ are formally defined as

$$p_j^{LFC}(X) = \inf_{\{\tilde{\alpha}\in(0,1):T_j(x)\in\Gamma_j(\tilde{\alpha})\}} \sup_{\{\vartheta:\theta_j(\vartheta)\in H_j\}} \mathbb{P}_\vartheta(T_j(X)\in\Gamma_j(\tilde{\alpha})).$$

Under assumptions $(A1)$ – $(A3)$, we obtain that

$$p_j^{LFC}(X) = \inf\{\tilde{\alpha}\in(0,1) : T_j(X)\in\Gamma_j(\tilde{\alpha})\},\ j=1,\ldots,m. \tag{3.3}$$

With assumption $(A2)$, any such LFC-based p-value $p_j^{LFC}(X)$ is uniformly distributed on $[0,1]$ under any LFC for φ_j; cf. Lemma 3.3.1 of Lehmann and Romano (2005). Let F_ϑ be the cumulative distribution function (cdf) of $T_j(X)$ under $\vartheta\in\Theta$. If the rejection region $\Gamma_j(\alpha)$ is given by $(F_{\vartheta_0}^{-1}(1-\alpha),\infty)$, where ϑ_0 is an LFC for φ_j, then the definition in (3.3) simplifies to $p_j^{LFC}(X) = 1 - F_{\vartheta_0}(T_j(X))$. Rejection regions of that type are typical if test statistics tend to larger values under alternatives, which is often the case.

As examples, we give two models that fulfil the general assumptions $(A1)$ – $(A3)$.

Example 3.1 (Multiple Z-tests model)**.** *We consider $X = (X_{i,j} : i = 1,\ldots,n_j,\ j = 1,\ldots,m)$, where $(n_j)_{j=1,\ldots,m}$ are fixed sample sizes. For all j the random variables $X_{1,j},\ldots,X_{n_j,j}$ are assumed to be stochastically independent and identically normally distributed as $N(\theta_j(\vartheta),1)$, where $\vartheta = (\vartheta_1,\ldots,\vartheta_m)^\top \in \Theta = \mathbb{R}^m$ is the (main) parameter of the model and $\theta(\vartheta)$, given by $\theta_j(\vartheta) = \vartheta_j$ for $1\leq j\leq m$, is the derived parameter. For each $1\leq j\leq m$, we are interested in the null hypothesis $H_j : \vartheta_j \leq 0$ against its alternative $K_j : \vartheta_j > 0$, and consider the test statistic $T_j(X) = n_j^{-1}\sum_{i=1}^{n_j} X_{i,j} \sim N(\vartheta_j, n_j^{-1})$. Furthermore, we let $\Gamma_j(\alpha) = (\Phi^{-1}_{(0,n_j^{-1})}(1-\alpha),\infty)$, leading to the LFC-based p-value $p_j^{LFC}(X) = 1 - \Phi_{(0,n_j^{-1})}(T_j(X))$, where $\Phi_{(\mu,\sigma^2)}$ denotes the cdf of the normal distribution on $\mathbb{R}$ with parameters μ and σ^2. For each $j = 1,\ldots,m$, the set of LFCs for φ_j is $\{\vartheta\in\Theta : \vartheta_j = 0\}$, independently of α. As mentioned before, we do not specify the dependence structure of $T_{j_1}(X)$ and $T_{j_2}(X)$ for $1\leq j_1\neq j_2\leq m$. The latter dependence structure may be regarded as a further (nuisance) parameter of the model.*

Example 3.2 (Two-sample means comparison model)**.** *Let $j = 1,\ldots,m$ be fixed. For given sample sizes $n_{1,j}$ and $n_{2,j}$, let $X_{1,j},\ldots,X_{n_{1,j},j}$ and $Y_{1,j},\ldots,Y_{n_{2,j},j}$ be jointly stochastically independent, observable random variables. Assume that $X_{1,j},\ldots,X_{n_{1,j},j}$ are identically distributed with $X_{1,j}\sim N(\theta_{1,j}(\vartheta),\sigma_j^2)$, and that $Y_{1,j},\ldots,Y_{n_{2,j},j}$ are identically distributed with $Y_{1,j}\sim N(\theta_{2,j}(\vartheta),\sigma_j^2)$, where $\sigma_j^2 > 0$ is unknown. Similarly as in Example 3.1, the parameter vector ϑ consists of all unknown means and all unknown variances of the model. For each $1\leq j\leq m$, we compare the means of the two samples. To this end, we let $\theta_j(\vartheta) = \theta_{1,j}(\vartheta) - \theta_{2,j}(\vartheta)$ and assume that $H_j : \theta_j(\vartheta)\leq 0$ versus $K_j : \theta_j(\vartheta) > 0$ is the marginal test problem of interest. Let $\bar{X}_j = n_{1,j}^{-1}\sum_{i=1}^{n_{1,j}} X_{i,j}$, $\bar{Y}_j = n_{2,j}^{-1}\sum_{i=1}^{n_{2,j}} Y_{i,j}$, and*

$$S_j(X) = \frac{1}{n_{1,j}+n_{2,j}-2}\Big[\sum_{i=1}^{n_{1,j}}(X_{i,j}-\bar{X}_j)^2 + \sum_{i=1}^{n_{2,j}}(Y_{i,j}-\bar{Y}_j)^2\Big].$$

Under an LFC for φ_j, that is, any $\vartheta\in\Theta$ with $\theta_j(\vartheta) = 0$, the test statistic

$$T_j(X) = \sqrt{\frac{n_{1,j}n_{2,j}}{n_{1,j}+n_{2,j}}}(\bar{X}_j - \bar{Y}_j)/S_j$$

follows Student's t-distribution with $n_{1,j}+n_{2,j}-2$ degrees of freedom, denoted by $t_{n_{1,j}+n_{2,j}-2}$. The corresponding rejection region is $\Gamma_j(\alpha) = (F^{-1}_{t_{n_{1,j}+n_{2,j}-2}}(1-\alpha),\infty)$ and the LFC-based p-value is given by $p_j^{LFC}(X) = 1 - F_{t_{n_{1,j}+n_{2,j}-2}}(T_j(X))$, where $F_{t_{n_{1,j}+n_{2,j}-2}}$ denotes the cdf of $t_{n_{1,j}+n_{2,j}-2}$. Again, the aforementioned set of LFCs for φ_j does not depend on α, for each $1\leq j\leq m$. For the dependence structure among different coordinates $j_1\neq j_2$, we argue as in Example 3.1.

3.3 The randomized p-values

3.3.1 General properties

Definition 3.1. *Let a model as in Section* 3.2 *and a set of random variables $U_1,\ldots,U_m$, that are defined on the same probability space as X, jointly stochastically independent, identically* Uni$[0,1]$*-distributed (under any $\vartheta\in\Theta$), and stochastically independent of the data X, be given. For each $j = 1,\ldots,m$ and given constants $c_1,\ldots,c_m$ with $c_j\in[0,1]$ for all $1\leq j\leq m$, we define our randomized p-values as in Equation* (3.2)*, where $p_j^{rand}(X,U_j,0) = U_j$ by convention.*

For a more general definition of these p-values, we refer to Appendix B. Before we discuss the properties of these randomized p-values and compare them to LFC-based ones, we give a few remarks.

Remark 3.1.

(*a.*) *If $p_j^{LFC}(X)$ is stochastically large, then it is likely that $p_j^{rand}(X, U_j, c_j) = U_j$ holds. This means that under the null hypothesis H_j, the distribution of p_j^{rand} will typically be close to a* Uni$[0,1]$*-distribution. On the other hand, if K_j is true and $p_j^{LFC}(X)$ is stochastically small, the randomized p-value $p_j^{rand}(X, U_j, c_j)$ is more likely to be equal to $p_j^{LFC}(X)/c_j \geq p_j^{LFC}(X)$ than it is to be equal to U_j.*

(*b.*) *Under an LFC ϑ_0 for φ_j the randomized p-value $p_j^{rand}(X, U_j, c_j)$ is uniformly distributed on $[0,1]$ for any $1 \leq j \leq m$. Namely, it holds that*

$$\begin{aligned}\mathbb{P}_{\vartheta_0}(p_j^{rand}(X, U_j, c_j) \leq t) &= \mathbb{P}_{\vartheta_0}(U_j \leq t)\, \mathbb{P}_{\vartheta_0}(p_j^{LFC}(X) \geq c_j) + \mathbb{P}_{\vartheta_0}(p_j^{LFC}(X) < t c_j) \\ &= t(1 - c_j) + t\, c_j = t,\end{aligned}$$

where we have used that $p_j^{LFC}(X)$ is Uni$[0,1]$*-distributed under any LFC ϑ_0 for φ_j, due to assumptions $(A1)$ – $(A2)$, and that U_j is always* Uni$[0,1]$*-distributed, no matter the value of ϑ.*

As mentioned in Section 3.1, the use of valid p-values in the Schweder-Spjøtvoll estimator ensures that the latter has a non-negative bias; cf. Lemma 1 of Dickhaus et al. (2012). Therefore it is of interest to give some conditions for the validity of our randomized p-values.

Theorem 3.1. *Let a model as in Section* 3.2 *be given and $j \in \{1, \ldots, m\}$ be fixed. Then, $p_j^{rand}(X, U_j, c_j)$ is a valid p-value for a given $c_j \in [0,1]$ if and only if the following condition (1.) is fulfilled. Furthermore, either of the following conditions (2.) and (3.) is a sufficient condition for the validity of $p_j^{rand}(X, U_j, c_j)$ for any $c_j \in [0,1]$.*

(1.) *For every $\vartheta \in \Theta$ with $\theta_j(\vartheta) \in H_j$, it holds*

$$\mathbb{P}_\vartheta(p_j^{LFC}(X) \leq t\, c_j) \leq t\, \mathbb{P}_\vartheta(p_j^{LFC}(X) \leq c_j)$$

for all $t \in [0,1]$.

(2.) *For every $\vartheta \in \Theta$ with $\theta_j(\vartheta) \in H_j$, $\mathbb{P}_\vartheta(p_j^{LFC}(X) \leq t)/t$ is non-decreasing in t.*

(3.) *The cdf of $p_j^{LFC}(X)$ is convex under any parameter $\vartheta \in \Theta$ with $\theta_j(\vartheta) \in H_j$.*

If the LFC-based p-value is given by $p_j^{LFC}(X) = 1 - F_{\vartheta_0}(T_j(X))$, where $\vartheta_0 \in \Theta$ is an LFC for φ_j, then the following condition (4.) *is equivalent to condition* (2.)*, while condition* (5.) *is equivalent to condition* (3.)*.*

(4.) *For every $\vartheta \in \Theta$ with $\theta_j(\vartheta) \in H_j$, it holds $T_j(X)^{(\vartheta)} \leq_{\mathrm{hr}} T_j(X)^{(\vartheta_0)}$.*

(5.) *For every $\vartheta \in \Theta$ with $\theta_j(\vartheta) \in H_j$, it holds $T_j(X)^{(\vartheta)} \leq_{\mathrm{lr}} T_j(X)^{(\vartheta_0)}$.*

With $\leq_{\mathrm{hr}}$ and $\leq_{\mathrm{lr}}$ we mean the hazard rate order and the likelihood ratio order, respectively. The notation $T_j(X)^{(\vartheta)}$ refers to the distribution of $T_j(X)$ under $\vartheta \in \Theta$. The relationship $T_j(X)^{(\vartheta)} \leq_{\mathrm{hr}} T_j(X)^{(\vartheta_0)}$ is equivalent to $(1 - F_{\vartheta_0}(t))/(1 - F_\vartheta(t))$ being non-decreasing in t, and $T_j(X)^{(\vartheta)} \leq_{\mathrm{lr}} T_j(X)^{(\vartheta_0)}$ is equivalent to $f_{\vartheta_0}(t)/f_\vartheta(t)$ being non-decreasing in t, where f_ϑ denotes the Lebesgue density of $T_j(X)$ under $\vartheta \in \Theta$.

The proof of Theorem 3.1 is given in Appendix B.

Corollary 3.1. *Under the models from Examples* 3.1 *and* 3.2*, the randomized p-values $(p_j^{rand}(X, U_j, c_j))_{1 \leq j \leq m}$ are valid for any $(c_1, \ldots, c_m)^\top \in [0,1]^m$.*

Proof. The multiple Z-tests model from Example 3.1 fulfils the general assumptions $(A1)$ – $(A3)$ from Section 3.2. Let $j \in \{1, \ldots, m\}$ be arbitrarily chosen. For a parameter value $\vartheta \in \Theta$ with $\theta_j(\vartheta) = \vartheta_j \in H_j$, i. e., $\vartheta_j \leq 0$, it is easy to show that $f_0(t)/f_{\vartheta_j}(t)$ is non-decreasing in t, where f_z denotes the Lebesgue density of the $N(z, n_j^{-1})$-distribution. Following Theorem 3.1, $p_j^{rand}(X, U_j, c_j)$ is valid for any constant $c_j \in [0,1]$. The choice of $c_j = 1/2$ for all $1 \leq j \leq m$ results in the randomized p-values from Dickhaus (2013) for this model.

The two-sample means comparison model from Example 3.2 fulfils the general assumptions $(A1)$ – $(A3)$, too. Again, let $j \in \{1, \ldots, m\}$ be arbitrarily chosen. Under any parameter value $\vartheta \in \Theta$ it holds that

$T_j(X) \sim t_{\tau_j, n_{1,j}+n_{2,j}-2}$, where $\tau_j = \sqrt{\frac{n_{1,j}n_{2,j}}{n_{1,j}+n_{2,j}}}\theta_j(\vartheta)/\sigma_j$, and $t_{\tau,\nu}$ denotes the non-central t-distribution with non-centrality parameter τ and ν degrees of freedom. The family $(t_{\tau, n_{1,j}+n_{2,j}-2})_{\tau\in\mathbb{R}}$ of distributions possesses the monotone likelihood ratio (MLR) property, i. e., it holds $t_{\tau_1, n_{1,j}+n_{2,j}-2} \leq_{\mathrm{lr}} t_{\tau_2, n_{1,j}+n_{2,j}-2}$ if and only if $\tau_1 \leq \tau_2$; cf. Karlin (1956) and Karlin and Rubin (1956a). For a parameter value $\vartheta \in \Theta$ with $\theta_j(\vartheta) \in H_j$, i. e., $\theta_{1,j}(\vartheta) \leq \theta_{2,j}(\vartheta)$, it holds that $\tau_j = \sqrt{\frac{n_{1,j}n_{2,j}}{n_{1,j}+n_{2,j}}}\theta_j(\vartheta)/\sigma_j \leq 0$ and therefore $T_j(X)^{(\vartheta)} \leq_{\mathrm{lr}} T_j(X)^{(\vartheta_0)}$, where ϑ_0 is an LFC for φ_j, i. e., $\theta_{1,j}(\vartheta_0) = \theta_{2,j}(\vartheta_0)$. According to Theorem 3.1, $p_j^{rand}(X, U_j, c_j)$ is valid for any choice of the constant $c_j \in [0,1]$ in this model. □ □

3.3.2 A comparison between the LFC-based and the randomized p-values

For any $1 \leq j \leq m$, we want to compare the cdfs of $p_j^{LFC}(X)$ and $p_j^{rand}(X, U_j, c_j)$. Due to the discussion below (3.2), this comparison is trivial for $c_j = 0$ and for $c_j = 1$, respectively. Therefore, let us assume here that c_j is bounded away from zero and from one. For example, one may for the moment assume that $c_j = 0.5$ is chosen, for concreteness.

We first note that

$$\begin{aligned}\mathbb{P}_\vartheta(p_j^{LFC}(X) \leq t) =& \mathbb{P}_\vartheta(p_j^{LFC}(X) \leq t \mid p_j^{LFC}(X) > c_j)\, \mathbb{P}_\vartheta(p_j^{LFC}(X) > c_j) \\ &+ \mathbb{P}_\vartheta(p_j^{LFC}(X) \leq t,\, p_j^{LFC}(X) \leq c_j), \qquad (3.4)\\ \mathbb{P}_\vartheta(p_j^{rand}(X, U_j, c_j) \leq t) =& \mathbb{P}_\vartheta(U_j \leq t)\, \mathbb{P}_\vartheta(p_j^{LFC}(X) > c_j) + \mathbb{P}_\vartheta(p_j^{LFC}(X) \leq t c_j). \qquad (3.5)\end{aligned}$$

Now, if the value of the derived parameter $\theta_j(\vartheta)$ is so "deep inside" H_j that $\mathbb{P}_\vartheta(p_j^{LFC}(X) > c_j)$ is large, then the first summands in (3.4) and (3.5) dominate the second ones, and we see that

$$\mathbb{P}_\vartheta(p_j^{LFC}(X) \leq t \mid p_j^{LFC}(X) > c_j) \leq \mathbb{P}_\vartheta(p_j^{LFC}(X) \leq t) \leq t = \mathbb{P}_\vartheta(U_j \leq t).$$

Thus, provided that $p_j^{rand}(X, U_j, c_j)$ is a valid p-value, its distribution under H_j will typically be closer to Uni$[0,1]$ than that of $p_j^{LFC}(X)$.

However, if ϑ is such that K_j is true instead and that $\mathbb{P}_\vartheta(p_j^{LFC}(X) \leq c_j)$ is large, it holds that

$$\begin{aligned}\mathbb{P}_\vartheta(p_j^{LFC}(X) \leq t,\, p_j^{LFC}(X) \leq c_j) &= \mathbb{P}_\vartheta(p_j^{LFC}(X) \leq \min(t, c_j)) \\ &\geq \mathbb{P}_\vartheta(p_j^{LFC}(X) \leq t c_j).\end{aligned}$$

Thus, under K_j the cdf of $p_j^{LFC}(X)$ will typically be pointwise larger than the cdf of $p_j^{rand}(X, U_j, c_j)$.

The former heuristic argumentation cannot be made mathematically rigorous in general. However, if condition (3.) in Theorem 3.1 is fulfilled, p_j^{rand} does indeed always lie between U_j and p_j^{LFC} under the null hypothesis H_j, in the sense of the stochastic order. The same holds under the alternative K_j, if a condition similar to (3.) is fulfilled in the case of $\theta_j(\vartheta) \in K_j$.

Theorem 3.2. *Let a model as in Section 3.2 be given and $j \in \{1, \ldots, m\}$ be fixed.*

If the cdf of $p_j^{LFC}(X)$ is convex under a fixed $\vartheta \in \Theta$, then

$$p_j^{rand}(X, U_j, c_j)^{(\vartheta)} \leq_{\mathrm{st}} p_j^{rand}(X, U_j, \tilde{c}_j)^{(\vartheta)}$$

for any $0 \leq c_j \leq \tilde{c}_j \leq 1$.

If the cdf of $p_j^{LFC}(X)$ is concave under a fixed $\vartheta \in \Theta$, then it holds that

$$p_j^{rand}(X, U_j, \tilde{c}_j)^{(\vartheta)} \leq_{\mathrm{st}} p_j^{rand}(X, U_j, c_j)^{(\vartheta)}$$

for any $0 \leq c_j \leq \tilde{c}_j \leq 1$.

We give the proof of Theorem 3.2 in Appendix B.

Remark 3.2. *Let $j \in \{1, \ldots, m\}$ be fixed.*

1. *If the j-th LFC-based p-value is given by $p_j^{LFC}(X) = 1 - F_{\vartheta_0}(T_j(X))$, where ϑ_0 is an LFC for φ_j, then $p_j^{LFC}(X)$ has a convex cdf under $\vartheta \in \Theta$ if and only if $T_j(X)^{(\vartheta)} \leq_{\mathrm{lr}} T_j(X)^{(\vartheta_0)}$, and a concave cdf under $\vartheta \in \Theta$ if and only if $T_j(X)^{(\vartheta_0)} \leq_{\mathrm{lr}} T_j(X)^{(\vartheta)}$ (cf. the proof of Theorem 3.1 in the Appendix B).*

2. *If condition (3.) from Theorem 3.1 is fulfilled, then Theorem 3.2 implies*

$$U_j \leq_{\mathrm{st}} p_j^{rand}(X, U_j, c_j)^{(\vartheta)} \leq_{\mathrm{st}} p_j^{LFC}(X)^{(\vartheta)}$$

for all $\vartheta \in \Theta$ with $\theta_j(\vartheta) \in H_j$ and any $c_j \in [0,1]$. This also implies the validity of $p_j^{rand}(X, U_j, c_j)$, as it was claimed in Theorem 3.1.

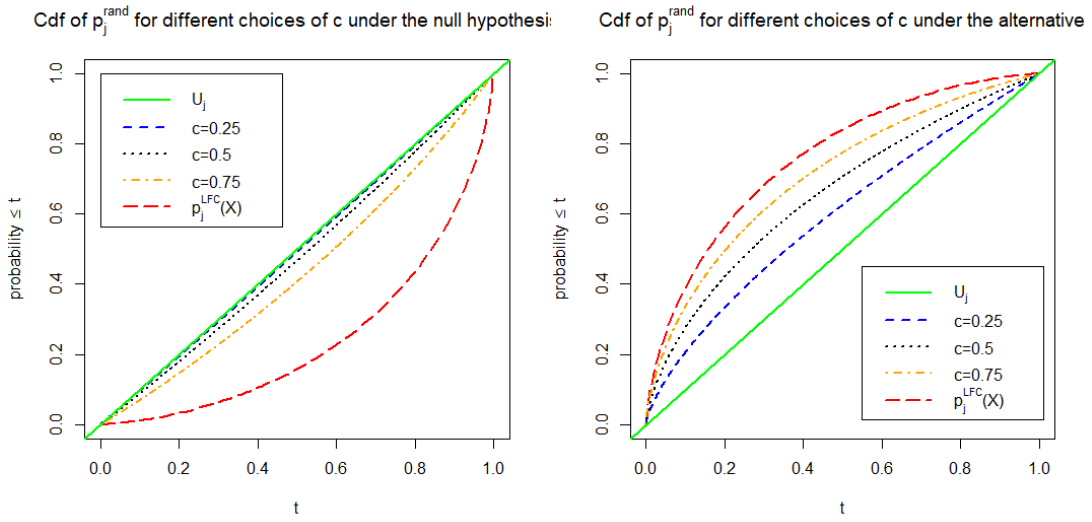


Figure 3.1: A comparison of the cdfs of $p_j^{rand}(X, U_j, c)$ for $c \in \{0, 0.25, 0.5, 0.75, 1\}$ under the multiple Z-tests model. In the left graph, $\theta_j(\vartheta) = -1/\sqrt{n_j}$ and in the right graph, $\theta_j(\vartheta) = 1/\sqrt{n_j}$, where $n_j = 50$. The value of $j \in \{1, \ldots, m\}$ is arbitrary.

3. *The cdf of $p_j^{LFC}(X)$ can never be concave under H_j.*

Corollary 3.2. *For the multiple Z-tests model and the two-sample means comparison model from Examples* 3.1 *and* 3.2*, respectively, it holds for any $1 \leq j \leq m$ and any $c_j \in [0, 1]$, that*

$$U_j \leq_{\text{st}} p_j^{rand}(X, U_j, c_j)^{(\vartheta)} \leq_{\text{st}} p_j^{LFC}(X)^{(\vartheta)}$$

under any $\vartheta \in \Theta$ with $\theta_j(\vartheta) \in H_j$, as well as

$$p_j^{LFC}(X)^{(\vartheta)} \leq_{\text{st}} p_j^{rand}(X, U_j, c_j)^{(\vartheta)} \leq_{\text{st}} U_j$$

under any ϑ with $\theta_j(\vartheta) \in K_j$.

We conclude this section by illustrating the assertions of Theorem 3.2 and Corollary 3.2 under the multiple Z-tests model. In Figure 3.1 we compare the cdfs of $p_j^{rand}(X, U_j, c)$ for an arbitrary $j \in \{1, \ldots, m\}$ with $c = 0, 0.25, 0.5, 0.75$, and 1 under $\vartheta \in \Theta$, where we set $\theta_j(\vartheta) = -1/\sqrt{n_j}$ or $\theta_j(\vartheta) = 1/\sqrt{n_j}$ for $n_j = 50$, respectively. It is apparent that the cdfs move from that of the Uni$[0, 1]$-distribution to the one of $p_j^{LFC}(X)$ with increasing c.

3.4 Estimation of the proportion of true null hypotheses

3.4.1 The expected value of the Schweder-Spjøtvoll estimator

We consider the usage of $\{p_j^{rand}(X, U_j, c_j)\}_{1 \leq j \leq m}$ in the Schweder-Spjøtvoll estimator $\hat{\pi}_0 \equiv \hat{\pi}_0(\lambda)$ defined in (3.1). It can easily be seen from the representation on the right-hand side of (3.1), that the bias of $\hat{\pi}_0(\lambda)$ decreases if $\mathbb{E}_\vartheta[\hat{F}_m(\lambda)]$ increases, under any $\vartheta \in \Theta$. Thus, in terms of bias reduction of $\hat{\pi}_0(\lambda)$ (for a fixed, given value of λ) stochastically small (randomized) p-values (with pointwise large cdfs) are most suitable. In order to avoid a negative bias of $\hat{\pi}_0(\lambda)$, we furthermore have to ensure validity of the p-values utilized in $\hat{\pi}_0(\lambda)$. Hence, if the cdfs of the LFC-based p-values are convex under null hypotheses and concave under alternatives, the optimal ("oracle") value of c_j is zero whenever H_j is true and one whenever K_j is true; cf. Theorem 3.2. This is also in line with Remark 6 of Dickhaus et al. (2012), who showed that $\hat{\pi}_0(\lambda)$ is unbiased if the p-values utilized in $\hat{\pi}_0(\lambda)$ are Uni$[0, 1]$-distributed under true null hypotheses and almost surely smaller than λ under false null hypotheses. Under the restriction of identical c_j's, i. e., $c_1 = c_2 = \ldots = c_m \equiv c$, one may expect that an optimal ("oracle") value of c (leading to a small, but non-negative bias of $\hat{\pi}_0(\lambda)$) should be close to $1 - \pi_0$. The restriction $c_1 = c_2 = \ldots = c_m \equiv c$ will be made throughout the remainder for computational convenience and feasibility.

Definition 3.2. *The Schweder-Spjøtvoll estimator $\hat{\pi}_0(\lambda)$, if used with $p_1^{rand}(X, U_1, c), \ldots, p_m^{rand}(X, U_m, c)$, will be denoted by $\hat{\pi}_0(\lambda, c)$ throughout the remainder. Notice, that in the estimators $\hat{\pi}_0(\lambda, 0)$ and $\hat{\pi}_0(\lambda, 1)$,*

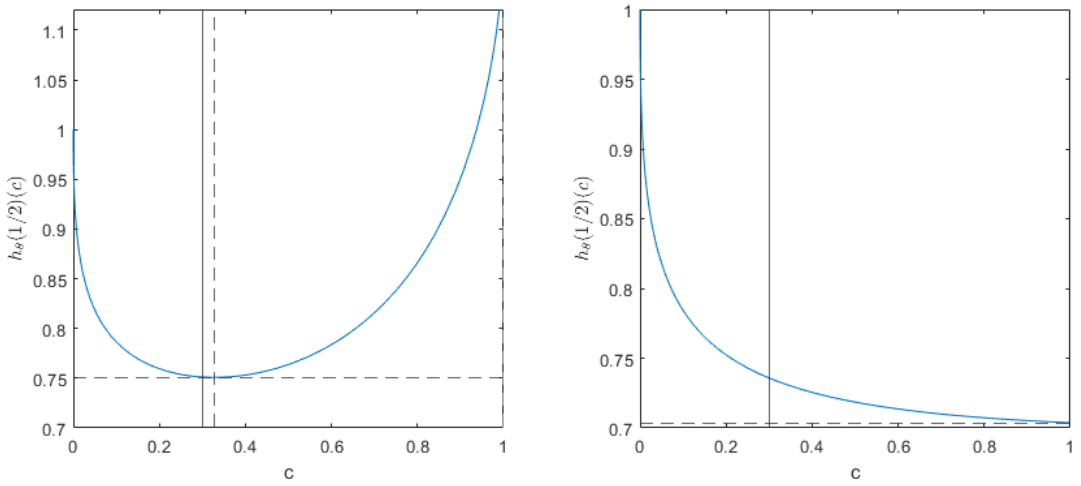


Figure 3.2: Two plots of $c \mapsto h_\vartheta(1/2, c)$ for $c \in [0,1]$ under the multiple Z-tests model. We set $\pi_0 = 0.7$, and $\vartheta \in \Theta$ such that $\theta_j(\vartheta) = 2.5/\sqrt{50}$ if K_j is true, $j = 1, \ldots, m = 1{,}000$. The parameter value under each null is $\theta_j(\vartheta) = -1/\sqrt{50}$ in the left graph and $\theta_j(\vartheta) = 0$ (leading to uniform null p-values) in the right graph. The solid vertical line indicates $c = 1 - \pi_0$, while the dashed one indicates the minimizing argument $c^\star$ of $c \mapsto h_\vartheta(1/2, c)$. The dashed horizontal line indicates $h_\vartheta(1/2, c^\star)$.

respectively, we use $U_1, \ldots, U_m$ (as the marginal p-values) and $p_1^{LFC}(X), \ldots, p_m^{LFC}(X)$, respectively. Furthermore, we consider the function $h_\vartheta : [0,1) \times [0,1] \to \mathbb{R}$, given by $h_\vartheta(\lambda, c) = \mathbb{E}_\vartheta[\hat{\pi}_0(\lambda, c)]$, where $\vartheta \in \Theta$ is the underlying parameter value.

Lemma 3.1. *For every $\lambda \in [0,1)$ and under any $\vartheta \in \Theta$, $h_\vartheta(\lambda, 0) = 1$. If the cdfs of the $p_1^{LFC}(X), \ldots, p_m^{LFC}(X)$ are continuous under ϑ, then there exists a minimizing argument $c^\star \in [0,1]$ of $h_\vartheta(\lambda, \cdot)$.*

Proof. In the case of $c = 0$, $p_j^{rand}(X, U_j, 0) = U_j$ for each $j \in \{1, \ldots, m\}$, and $\mathbb{E}_\vartheta[\hat{\pi}_0(\lambda, 0)] = (1-\lambda)/(1-\lambda) = 1$, proving the first assertion.

In order to show the second assertion, we note that under any $\vartheta \in \Theta$

$$\mathbb{E}_\vartheta[\hat{F}_m(\lambda)] = \sum_{j=1}^{m} \left[\lambda\, \mathbb{P}_\vartheta(p_j^{LFC}(X) \geq c) + \mathbb{P}_\vartheta(p_j^{LFC}(X) \leq c\lambda)\right]. \tag{3.6}$$

The right-hand side of (3.6) is continuous in c if the cdfs of the p-values $p_1^{LFC}(X), \ldots, p_m^{LFC}(X)$ are continuous under ϑ. Since $[0,1]$ is a compact set, the function $h_\vartheta(\lambda, \cdot)$ attains a minimum on $[0,1]$, by the extreme value theorem. □ □

For an illustration, let us consider the multiple Z-tests model from Example 3.1, where we set the total number of null hypotheses to $m = 1{,}000$ and the sample sizes to $n_j = 50$ for all $j = 1, \ldots, m$. As mentioned before, the choice of $c = 1/2$ leads to the randomized p-values as defined in Dickhaus (2013) for this model. Figure 3.2 displays the graphs of the function $c \mapsto h_\vartheta(1/2, c)$ for two different parameter values $\vartheta \in \Theta$ under this model. In both cases, $\pi_0 = 0.7$ (meaning that 700 null hypotheses are true and 300 are false) and $\theta_j(\vartheta) = 2.5/\sqrt{50}$ whenever H_j is false.

On the left graph of Figure 3.2, $\theta_j(\vartheta) = -1/\sqrt{50}$ whenever H_j is true. The minimum of $c \mapsto h_\vartheta(1/2, c)$ is attained at $c^\star = 0.3276$ and yields $\mathbb{E}_\vartheta[\hat{\pi}_0(1/2, c^\star)] = 0.7508$. It is apparent that $h_\vartheta(1/2, c)$ is largest for $c = 1$, that is, when utilizing the LFC-based p-values $\{p_j^{LFC}(X)\}_{1 \leq j \leq m}$. Finally, we see that the optimal bias of $\hat{\pi}_0(1/2)$ when using the same $c_j \equiv c$ for all $1 \leq j \leq m$ is larger than zero (compare the dashed and the dotted horizontal lines).

On the right graph of Figure 3.2, $\theta_j(\theta) = 0$ whenever H_j is true. In this case, the estimator $\hat{\pi}_0(1/2, 1)$ has the lowest bias among all estimators $\{\hat{\pi}_0(1/2, c) : c \in [0,1]\}$, meaning that $c^\star = 1$. This is because for every j with $\theta_j(\vartheta) \in H_j$, ϑ is an LFC for φ_j and thus $p_j^{LFC}(X)$ is Uni$[0,1]$-distributed under ϑ. In such cases, $p_j^{rand}(X, U_j, c)$ is Uni$[0,1]$-distributed for any c under H_j, while $p_j^{LFC}(X)^{(\vartheta)} \leq_{\text{st}} p_j^{rand}(X, U_j, c)^{(\vartheta)}$ if K_j is true, due to Theorem 3.2.

3.4.2 Minimizing the MSE

From a decision-theoretic perspective, the bias alone is not enough to judge the estimation quality of $\hat{\pi}_0$. A more commonly used criterion for the quality of an estimator is its MSE. Therefore, we investigate the MSE of $\hat{\pi}_0$, when using the randomized p-values, in this section. The bias of $\hat{\pi}_0(\lambda)$ does not

depend on the dependence structure of the utilized marginal p-values $p_1, \dots, p_m$. To see this, notice that $\mathbb{E}_\vartheta[\hat{F}_m(\lambda)] = m^{-1} \sum_{j=1}^m \mathbb{P}_\vartheta(p_j \leq \lambda)$. Calculating the variance of $\hat{\pi}_0(\lambda)$, however, requires the knowledge of the dependence structure of the p-values, because

$$\mathrm{Var}_\vartheta\left(\hat{F}_m(\lambda)\right) = \frac{1}{m^2}\left[\sum_{j=1}^m \mathrm{Var}_\vartheta\left(\mathbf{1}\{p_j \leq \lambda\}\right) + \sum_{i \neq j} \mathrm{Cov}_\vartheta\left(\mathbf{1}\{p_i \leq \lambda\}, \mathbf{1}\{p_j \leq \lambda\}\right)\right].$$

The independent case

Here, we present some results regarding the variance of $\hat{\pi}_0$ when the marginal LFC-based p-values are stochastically independent. In this, we assume that the cdf of p_j^{LFC} is convex under H_j and concave under K_j. This assumption is often fulfilled, especially in the models studied before.

Lemma 3.2. *Assume that $p_1^{LFC}(X), \dots, p_m^{LFC}(X)$ are stochastically independent. Then, the variance of $\hat{\pi}_0(1/2)$ is monotonically decreasing in $c \in [0,1]$ if used with $p_1^{rand}(X, U_1, c), \dots, p_m^{LFC}(X, U_m, c)$. Furthermore, the maximum variance of $\hat{\pi}_0(1/2)$, among all $c \in [0,1]$, equals $1/m$.*

Proof. Since $\mathrm{Var}_\vartheta\left(\hat{\pi}_0(\lambda)\right) = (1-\lambda)^{-2}\mathrm{Var}_\vartheta\left(\hat{F}_m(\lambda)\right)$, we have to show that the latter is decreasing in c, where $\hat{F}_m(\lambda)$ is the ecdf of the randomized p-values $p_1^{rand}(X, U_1, c), \dots, p_m^{LFC}(X, U_m, c)$ at point λ.

If the p-values are independent, then so are the randomized p-values, thus

$$\mathrm{Var}_\vartheta\left(\hat{F}_m(\lambda)\right) = \frac{1}{m^2}\sum_{j=1}^m \mathrm{Var}_\vartheta\left(\mathbf{1}\{p_j^{rand}(X, U_j, c) \leq \lambda\}\right).$$

We can show that each summand $\mathrm{Var}_\vartheta\left(\mathbf{1}\{p_j^{rand}(X, U_j, c) \leq \lambda\}\right)$ is decreasing in c, $j = 1, \dots, m$.

For a fixed j, it holds $\mathrm{Var}_\vartheta\left(\mathbf{1}\{p_j^{rand}(X, U_j, c) \leq \lambda\}\right) = f(c) - f(c)^2$, where f is the cdf of $p_j^{rand}(X, U_j, c)$ at point λ. Furthermore, it holds

$$\frac{d}{dc}\left(f(c) - f(c)^2\right) = f'(c)\left(1 - 2f(c)\right).$$

Due to Theorem 2 $f'(c)$ is non-positive under H_j and non-negative under K_j for all $c \in [0,1]$. More particularly, since $f(0) = \mathbb{P}\left(U_j \leq \lambda = 1/2\right) = 1/2$ the term $1 - 2f(c)$ is non-positive under H_j and non-negative under K_j. In total, it holds

$$\frac{d}{dc}\left(f(c) - f(c)^2\right) \leq 0.$$

For the maximum variance, we have to plug $c = 0$ into the variance formula, for which the randomized p-values are the uniformly distributed $U_1, \dots, U_m$. For these, it holds $\mathrm{Var}_\vartheta\left(\hat{F}_m(\lambda)\right) = \lambda(1-\lambda)/m$ and $\mathrm{Var}_\vartheta\left(\hat{\pi}_0(\lambda)\right) = \lambda/((1-\lambda)m) = 1/m$ if $\lambda = 1/2$. □ □

Lemma 3.2 implies that in case of stochastically independent LFC-based p-values, randomization increases the variance of $\hat{\pi}_0(1/2)$. However, if m is large, the variance of $\hat{\pi}_0(1/2)$ has a small impact on the MSE, because the bias of $\hat{\pi}_0(1/2)$ does not explicitly depend on m. We demonstrate this with an example. As in Section 3.3.1, we consider the multiple Z-tests model from Example 3.1 for the choices of $\theta_j(\vartheta) = -0.5/\sqrt{50}$ if H_j is true and $\theta_j(\vartheta) = 1.5/\sqrt{50}$ otherwise. We set $\pi_0 = 0.8$, $\lambda = 1/2$, and $m = 100$.

Figure 3.3 displays the bias, the variance, and the MSE of $\hat{\pi}_0(\lambda)$ when used with the randomized p-values $p_1^{rand}(X, U_1, c), \dots, p_m^{LFC}(X, U_m, c)$ as functions of $c \in [0,1]$. The bias curve starts at $1 - \pi_0$ for $c = 0$ and has its minimum 0.1171 at $c = c^\star = 0.3626$. For $c > c^\star$, it monotonically increases to 0.3331 at $c = 1$. The variance curve starts at $1/m = 0.01$ for $c = 0$ and decreases with increasing c, as expected. The MSE curve is mostly affected by the (squared) bias curve. The MSE-minimizing value c^{MSE} of c equals 0.3742, which is slightly larger than $c^\star$. For our choice of the parameters, the MSE has its minimum at $c = 1$ only for $m = 1$; the MSE curve has a u-shape for all $m \geq 2$.

Since the variance of $\hat{\pi}_0(1/2)$ decreases with c and is upper-bounded by $1/m$, the MSE-minimizing value c^{MSE} of c converges to $c^\star$ from above. We calculated c^{MSE} for some values of m, where we used the same model and settings as before, cf. Table 3.1.

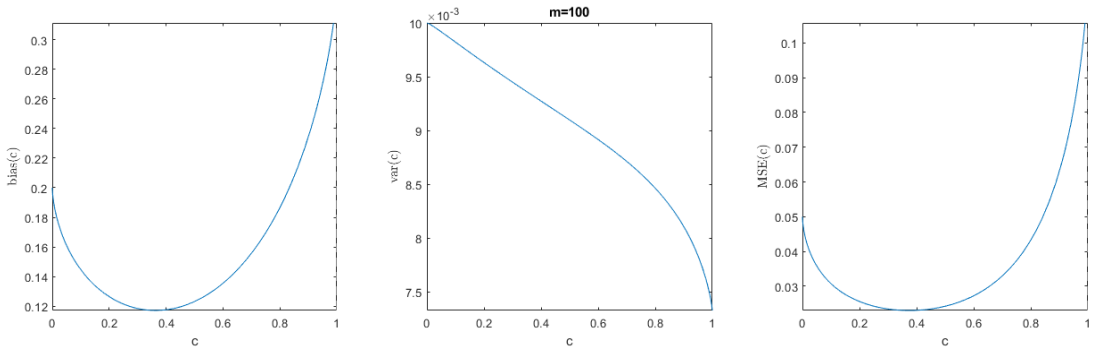


Figure 3.3: Plots of the bias, the variance and the MSE of $\hat{\pi}_0(1/2)$, respectively, as functions of c, when used with the p-values $p_1^{rand}(X, U_1, c), \ldots, p_m^{LFC}(X, U_m, c)$ under the multiple Z-tests model as described in Section 3.4.2 with $m = 100$. Here, we assume stochastically independent test statistics.

m	c^{MSE}
1	1
10	0.4759
50	0.3858
100	0.3742
500	0.3649
1,000	0.3638
10,000	0.3627
∞	$c^\star = 0.3626$

Table 3.1: The MSE minimizing value c^{MSE} of c for different numbers of hypotheses m under the model and the parameter setting described in Section 3.4.2. Independently of m, the bias minimizing parameter $c^\star$ equals 0.3626 here.

The positively dependent case

In this section we consider positively dependent p-values $p_1^{LFC}(X), \ldots, p_m^{LFC}(X)$. We calculate the variance and MSE curves for varying degrees of positive dependence. Again, we consider the multiple Z-tests model from Example 3.1, where the test statistics $T_1(X), \ldots, T_m(X)$ now have a pairwise correlation coefficient $\mathrm{Corr}(T_i(X), T_j(X)) = \rho \in [0, 1]$, for all $i \neq j$. One further model class, referring to Gumbel-Hougaard copula dependency structures, is considered in Appendix B.2.

Apart from ρ, we choose the same model settings as in the previous section. As mentioned before, the bias curve as well as $c^\star$ do not depend on the dependence structure of the p-values. However, we included it in the following figure to facilitate the comparison of MSE and (squared) bias. We considered $\rho = 0.5$ in Figure 3.4. The independent case, which we have considered in the previous section, corresponds to $\rho = 0$.

As displayed in Figure 3.4, the variance of $\hat{\pi}_0(\lambda)$ increases monotonically in c here, which means that the variance of $\hat{\pi}_0(\lambda)$ decreases with the amount of randomization. The overall magnitude of the

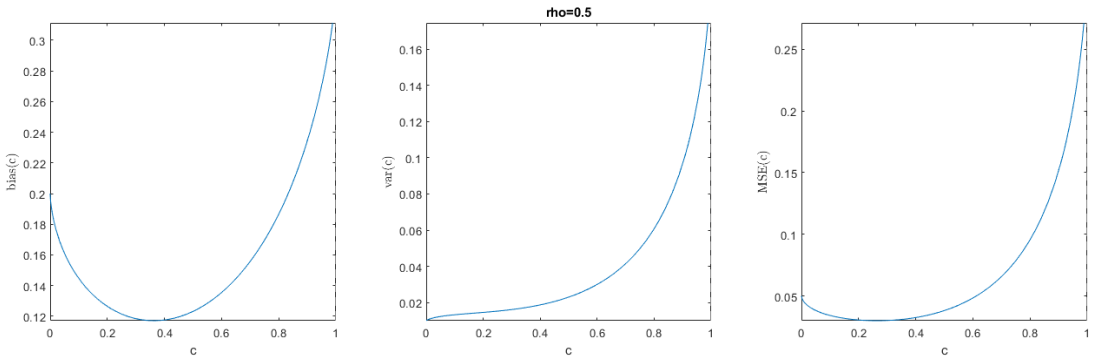


Figure 3.4: Plots of the bias, the variance and the MSE of $\hat{\pi}_0(1/2)$, respectively, as functions of c, when used with the p-values $p_1^{rand}(X, U_1, c), \ldots, p_m^{LFC}(X, U_m, c)$ under the multiple Z-tests model as described in Section 3.4.2 with $m = 100$. Here, the test statistics have a pairwise correlation of $\rho = 0.5$.

variances is also much higher now when compared with the case of $\rho = 0$. The MSE-minimizing value of c is 0.2642 here, meaning that the optimal amount of randomization is higher for $\rho = 0.5$ than for $\rho = 0$.

This example illustrates that the presence of positive dependence among the marginal p-values can deteriorate the performance of the Schweder-Spjøtvoll estimator. Similar results in this direction have recently been obtained by Neumann et al. (2021). Since the randomized p-values are a mix of the LFC-based p-values and the stochastically independent random variables $U_1, \ldots, U_m$, the degree of dependence is lower among the randomized p-values, and thus also the variance of the Schweder-Spjøtvoll estimator when used with the latter.

3.4.3 Estimating π_0 in practice

The expected value in $h_\vartheta(\lambda, c) = \mathbb{E}_\vartheta[\hat{\pi}_0(\lambda, c)]$ discussed in Section 3.4.1 refers to the joint distribution of $\{U_j\}_{1 \leq j \leq m}$ and the data X under ϑ. In practice, the distribution of X under ϑ is unknown, but we have a realized data sample $X = x \in \Omega$ at hand, from which $p_1^{LFC}(x), \ldots, p_m^{LFC}(x)$ can be computed. Throughout this section, let us assume a statistical model such that any of the conditions (2.) – (5.) from Theorem 3.1 is fulfilled, so that $p_1^{rand}(X, U_1, c), \ldots, p_m^{rand}(X, U_m, c)$ are valid p-values for any $c \in [0, 1]$.

In analogy to (3.6), we obtain that the conditional expected value (with respect to the U_j's) of $\hat{\pi}_0(\lambda, c)$ under the condition $X = x$ is given by

$$\mathbb{E}[\hat{\pi}_0(\lambda, c) \mid X = x] = \frac{1}{1-\lambda} \left[1 - \frac{1}{m} \sum_{j=1}^{m} \left[\lambda \, \mathbf{1}\{p_j^{LFC}(x) \geq c\} + \mathbf{1}\{p_j^{LFC}(x) \leq \lambda c\} \right] \right]. \tag{3.7}$$

Our proposal for practical purposes is to minimize (3.7) with respect to $c \in [0, 1]$, for fixed $\lambda \in [0, 1)$. Thus, this approach focuses on minimizing the conditional bias of $\hat{\pi}_0$, given the data. Denoting the solution of this minimization problem by c_0, we then propose to utilize $p_1^{rand}(x, U_1, c_0), \ldots, p_m^{rand}(x, U_m, c_0)$ in $\hat{\pi}_0(\lambda)$.

Minimizing (3.7) with respect to $c \in [0, 1]$ is equivalent to maximizing the function $c \mapsto g_x(\lambda, c)$, given by

$$g_x(\lambda, c) = \sum_{j=1}^{m} \big(\lambda \, \mathbf{1}\{p_j^{LFC}(x) \geq c\} + \mathbf{1}\{p_j^{LFC}(x) \leq \lambda c\} \big), \tag{3.8}$$

with respect to $c \in [0, 1]$. Hence, the solution c_0 is such, that most of the (realized) LFC-based p-values are outside of the interval $(\lambda c_0, c_0)$. An optimal choice c_0 can be determined numerically by either evaluating $g_x(\lambda, \cdot)$ on a given grid $0 = c_0 < \cdots < c_N = 1$ or on the set $\{p_1^{LFC}(x), \ldots, p_m^{LFC}(x), p_1^{LFC}(x)/\lambda, \ldots, p_m^{LFC}(x)/\lambda\}$ (excluding values larger than 1). Notice, that $g_x(\lambda, \cdot)$ can only change its values at points from the second set.

We demonstrate this procedure with an example. Again, consider the multiple Z-tests model and the same parameter setting as for deriving the left graph in Figure 3.2. Under these settings, we randomly drew one sample $x \in \Omega$ and applied the proposed procedure with $\lambda = 1/2$. After the removal of elements exceeding one from the set $\{p_1^{LFC}(x), \ldots, p_m^{LFC}(x), 2p_1^{LFC}(x), \ldots, 2p_m^{LFC}(x)\}$, 1,406 relevant points remained for the evaluation of $g_x(1/2, \cdot)$. As displayed in Figure 3.5, the maximum of $g_x(1/2, \cdot)$ is for the observed x attained at $c_0 = 0.3286$. This is an optimal c given the realized values $p_1^{LFC}(x), \ldots, p_m^{LFC}(x)$. For comparison, recall that we have seen in Section 3.4.1 that $c^\star = 0.3276$ minimizes the bias of $\hat{\pi}_0(1/2, c)$ on average over $X \sim \mathbb{P}_\vartheta$.

Figure 3.6 displays the ecdfs pertaining to $p_1^{LFC}(x), \ldots, p_m^{LFC}(x)$ and $p_1^{rand}(x, u_1, c_0), \ldots, p_m^{rand}(x, u_m, c_0)$, respectively, where $\{u_1, \ldots, u_m\}$ is one particular set of realizations of the random variables $U_1, \ldots, U_m$. Furthermore, the two dotted vertical lines in Figure 3.6 indicate the interval $[c_0/2, c_0]$. Recall that c_0 is chosen such, that most of the (realized) LFC-based p-values are outside of the latter interval. This can visually be confirmed, since the ecdf pertaining to $p_1^{LFC}(x), \ldots, p_m^{LFC}(x)$ is rather flat on $[c_0/2, c_0]$.

For any ecdf $t \mapsto \hat{F}_m(t)$ utilized in $\hat{\pi}_0(\lambda)$, the offset at $t = 0$ of the straight line connecting the points $(1, 1)$ and $(\lambda, \hat{F}_m(\lambda))$ equals $1 - \hat{\pi}_0(\lambda)$; cf., e. g., Figure 3.2.(b) in Dickhaus (2014). We therefore obtain an accurate estimate of π_0 if the ecdf $t \mapsto \hat{F}_m(t)$ utilized in $\hat{\pi}_0(\lambda)$ is at $t = \lambda$ close to the straight line connecting the points $(1, 1)$ and $(0, 1 - \pi_0)$. The latter "optimal" line is the expected ecdf of marginal p-values that are Uni$[0, 1]$-distributed under the null and almost surely equal to zero under the alternative. In Figure 3.6, the ecdf pertaining to $p_1^{rand}(x, u_1, c_0), \ldots, p_m^{rand}(x, u_m, c_0)$ is much closer to that optimal line than the ecdf pertaining to $p_1^{LFC}(x), \ldots, p_m^{LFC}(x)$. Consequently, for this particular dataset the estimation approach based on $p_1^{rand}(x, u_1, c_0), \ldots, p_m^{rand}(x, u_m, c_0)$ leads to a much more precise estimate of $\pi_0 = 0.7$ then the one based on $p_1^{LFC}(x), \ldots, p_m^{LFC}(x)$. The estimate based on $p_1^{LFC}(x), \ldots, p_m^{LFC}(x)$ even exceeds one in this example. We have repeated this simulation several times (results not included here) and the conclusions have always been rather similar.

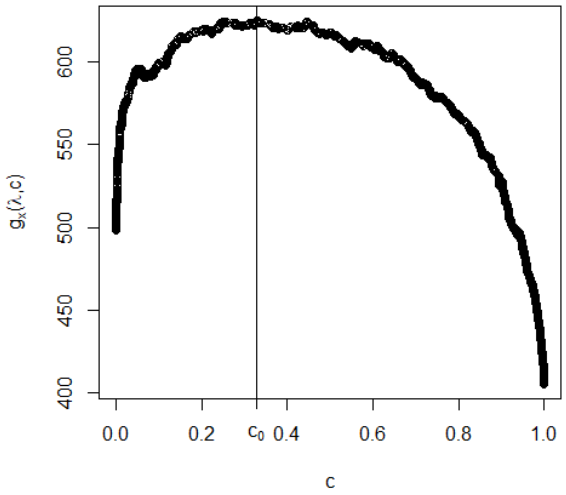


Figure 3.5: A plot of the function $c \mapsto g_x(\lambda, c)$, for $\lambda = 1/2$, evaluated on those 1,406 elements of the set $\{p_1^{LFC}(x), \ldots, p_m^{LFC}(x), \frac{p_1^{LFC}(x)}{\lambda}, \ldots, \frac{p_m^{LFC}(x)}{\lambda}\}$ which are not larger than one. Here, $g_x(\lambda, \cdot)$ attains its maximum at $c_0 = 0.3286$. The underlying data x have randomly been drawn under the multiple Z-tests model.

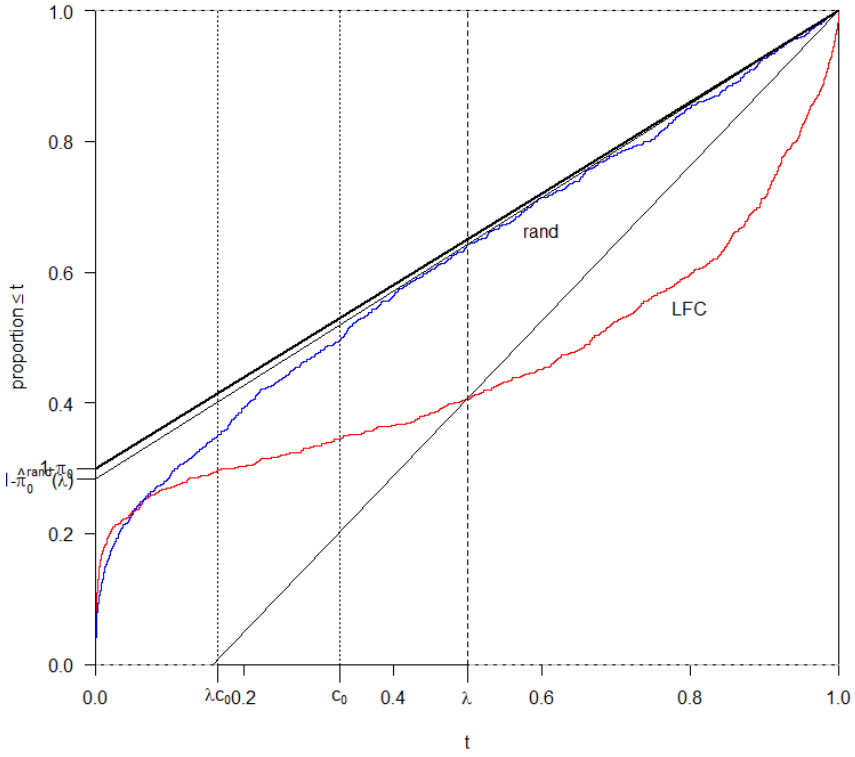


Figure 3.6: The ecdfs $\hat{F}_m$ of $(p_j^{LFC}(x))_{j=1,\ldots,m}$ and $(p_j^{rand}(x, u_j, c_0))_{j=1,\ldots,m}$, respectively, under the multiple Z-tests model for $\pi_0 = 0.7$. The underlying data x are the same as in Figure 3.5. The thicker straight line connects the points $(0, 1 - \pi_0)$ and $(1, 1)$, while the two thinner straight lines connect $(\lambda, \hat{F}_m(\lambda))$ with $(1, 1)$ for the two aforementioned ecdfs. The offset of each of the two thinner lines at $t = 0$ equals $1 - \hat{\pi}_0(\lambda)$ for the respective ecdf, where $\lambda = 1/2$. The two dotted vertical lines indicate the interval $[\lambda c_0, c_0]$, where c_0 is as in Figure 3.5.

3.5 Impact on data-adaptive multiple tests

A multiple testing procedure $\varphi : \Omega \to \{0,1\}^m$ is a measurable mapping, such that, for each j, $\varphi_j(x) = 1$ means that we reject H_j on the basis of the observed data x and $\varphi_j(x) = 0$ means that H_j is retained. The false discovery rate (FDR) of a given multiple testing procedure φ under ϑ is given by

$$\mathrm{FDR}_\vartheta(\varphi) = \mathbb{E}_\vartheta\left[\frac{V}{\max(R,1)}\right],$$

where V is the (random) number of false rejections and R is the total (random) number of rejections. We say that φ controls the FDR at level $\alpha \in (0,1)$, if $\sup_{\vartheta\in\Theta}\mathrm{FDR}_\vartheta(\varphi) \leq \alpha$ holds true.

Throughout this section, we consider the so-called linear step-up (LSU) procedure by Benjamini and Hochberg (1995), which works as follows. Given the ordered marginal p-values $p_{(1)} \leq \cdots \leq p_{(m)}$, the correspondingly ordered null hypotheses $H_{(1)}, \ldots, H_{(m)}$, and thresholds $\delta_j = \alpha j/m$, $j = 1, \ldots, m$, let $k = \max\{1 \leq j \leq m : p_{(j)} \leq \delta_j\}$. If there is no such k, we set $k = 0$. Then, we reject all null hypotheses $H_{(j)}$ for which $j \leq k$.

If the marginal p-values are jointly stochastically independent, the LSU test controls the FDR at level $\alpha\pi_0$. Thus, it is possible to replace the thresholds δ_j by δ_j/π_0, $j = 1, \ldots, m$, while still controlling the FDR at level α. Note, that this leads to larger thresholds for each j. Index-wise larger thresholds increase the power of the resulting multiple test. However, as π_0 is unknown, we instead use $\tilde{\delta}_j = \delta_j G$, where G is an estimator for $1/\pi_0$ based on the marginal p-values. A step-up procedure with these kinds of thresholds is called a data-adaptive multiple test procedure.

An important result is the Theorem 11 in Blanchard and Roquain (2009). It provides a condition under which a given measurable, coordinate-wise non-increasing function $G : [0,1]^m \to (0,\infty)$ and jointly independent p-values leads to FDR control at level α of the resulting data-adaptive LSU test. This condition is given by

$$\mathbb{E}_\vartheta[G_{j\to 0}] \leq \pi_0^{-1} \tag{3.9}$$

for all $j \in \{1, \ldots, m\}$ and all parameter values $\vartheta \in H_j$. The notation $G_{j\to 0}$ means that the j-th p-value is replaced by zero when G is applied.

The estimator $G = 1/\hat{\pi}_0(\lambda)$ based on the Schweder-Spjøtvoll estimator does not fulfil the condition (3.9), but $G = 1/\hat{\pi}_0^+(\lambda)$ based on the modified, more conservative Storey estimator

$$\hat{\pi}_0^+(\lambda) = \frac{1 - \hat{F}_m(\lambda) + 1/m}{1-\lambda} = \hat{\pi}_0(\lambda) + \frac{1}{(1-\lambda)m}$$

does, cf. Storey (2002). Since this result applies to any set of valid p-values, we can use our randomized p-values in the data-adaptive LSU test with $G = 1/\hat{\pi}_0^+(\lambda)$.

Remark 3.3. *In the previous sections, we focused on the Schweder-Spjøtvoll estimator. However, it is clear that* $\mathrm{Var}(\hat{\pi}_0^+(\lambda)) = \mathrm{Var}(\hat{\pi}_0(\lambda))$ *and* $\mathrm{bias}(\hat{\pi}_0^+(\lambda)) = \mathrm{bias}(\hat{\pi}_0(\lambda)) + [(1-\lambda)m]^{-1}$. *Consequently, the same bias minimizing value* $c^\star$ *of* c *applies to both estimators.*

In the remainder of this section, we assess how well the adaptive LSU test with $G = 1/\hat{\pi}_0^+(\lambda)$ performs when used with our randomized p-values with $c = c^\star$, and when used with the LFC-based p-values. We consider both the independent and the positively dependent case, although FDR control is not guaranteed in the latter case. To this end, we employ the same model as in Section 3.4.2, and in Section 3.4 of Blanchard and Roquain (2009). We set $\pi_0 = 0.75$, and consider pairwise correlations $\rho \in \{0, 0.25, 0.5, 0.75\}$ of the test statistics. The underlying parameter values have been set to $\theta_j(\vartheta) = -0.2r$ if H_j is true, and $\theta_j(\vartheta) = 1 + 0.25r$ if K_j is true, $j = 1, \ldots, m$, where $r \in \{1, \ldots, 10\}$ denotes a signal strength which is expressed in units of the standard deviation of the test statistics.

Figures 3.7 – 3.9 display the averaged π_0 estimations (i. e., the values of $1/G = \hat{\pi}_0^+$), the FDR, and the 'performances', respectively. Under the performance of a multiple test, we mean a function that increases in the power of the test and decreases in its FDR. We chose

$$\mathrm{Performance}_\vartheta(\varphi) = \mathrm{Power}_\vartheta(\varphi) - \frac{\mathrm{FDR}_\vartheta(\varphi)}{\alpha}, \tag{3.10}$$

where power denotes the proportion of correctly rejected null hypotheses. All values displayed in Figures 3.7 – 3.9 have been calculated via Monte-Carlo simulations with 100,000 repetitions. The subplots in each figure correspond to different values for ρ, and we varied the signal strength parameter r on the horizontal axis. In each simulation, the randomized p-values with $c = c^\star$ have only been employed in G, while the LFC-based p-values have been used in the comparison with $(\tilde{\delta}_j)_{1\leq j\leq m}$. The reason for this

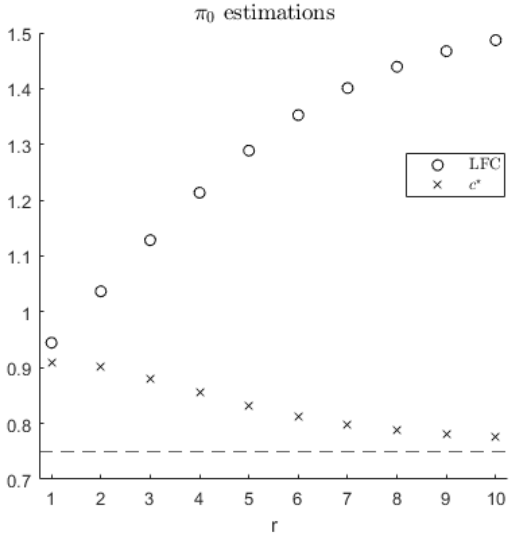


Figure 3.7: Monte Carlo averages (referring to 100,000 simulation runs) of the π_0 estimations that are utilized in the adaptive LSU tests described in Section 3.5. Considered is the modified Storey estimator with $\lambda = 1/2$, used with the LFC-based p-values (circles) or the randomized p-values with $c = c^\star$ (crosses), respectively. The signal strength is denoted by r. The horizontal dashed line indicates π_0.

procedure is, that randomized p-values are irreproducible and should therefore not be used for making test decisions.

As for the Schweder-Spjøtvoll estimator, the expected value of the modified Storey estimator does not depend on the particular dependence structure of the utilized marginal p-values. Hence, Figure 3.7 is representative for any ρ. For all considered values of r, the LFC-based p-values lead to uninformative π_0 estimations. For $r = 1$ the estimations corresponding to the two different sets of p-values are closest, since for small r the null p-values are close to being uniform. Using $c = c^\star$ leads to the lowest π_0 estimations, on average, among all $c \in [0, 1]$, because a non-negative bias is guaranteed here. With increasing r, the π_0 estimations get worse when using the LFC-based p-values and better when using the randomized p-values.

In Figure 3.8 we display the realized values of the FDR of the procedures described above. As expected, the FDR-values in cases with independent LFC-based p-values ($\rho = 0$) are all smaller than $\alpha = 0.05$. The two approaches perform similarly in this case. Under positive dependence ($\rho > 0$), the FDR of the multiple test using the LFC-based p-values is higher than that of the multiple test using randomized p-values. Especially for low r and high ρ, the FDR is not always controlled when using the LFC-based p-values. Furthermore, we notice that the FDR decreases with increasing signal strength r in all approaches. Even though the expected π_0 estimations are the same regardless of the dependence structure of the LFC-based p-values, we notice that the FDR increases with increasing correlation ρ. This is due to the increase in the variance of $1/G$, cf. also Section 3.4.2.

Figure 3.9 displays the performances in the sense of (3.10) of the adaptive LSU tests utilizing the different sets of p-values. Under independence, the performances of the different approaches are close to each other. Under positive dependence, however, the performance when using the LFC-based p-values is inferior to that of the randomized p-values, especially for low values of r.

3.6 Relationships to other approaches

There are several ways to draw connections between our proposed methodology and existing literature. One way is to consider statistical methods presented in previous literature that assume uniformly distributed p-values under null hypotheses, and to discuss the advantages of our randomized p-values compared to conservative p-values. Another way is to compare our approach with further strategies that help make conservative p-values more uniformly distributed under null hypotheses.

3.6.1 Implementation of our proposed approach into existing procedures

Our proposed methodology of randomizing p-values can be used in connection with any estimator of π_0 which relies on the ecdf of marginal p-values. Such estimators constitute a major class of estimators of π_0; cf., e. g., Table 1 in Chen (2019).

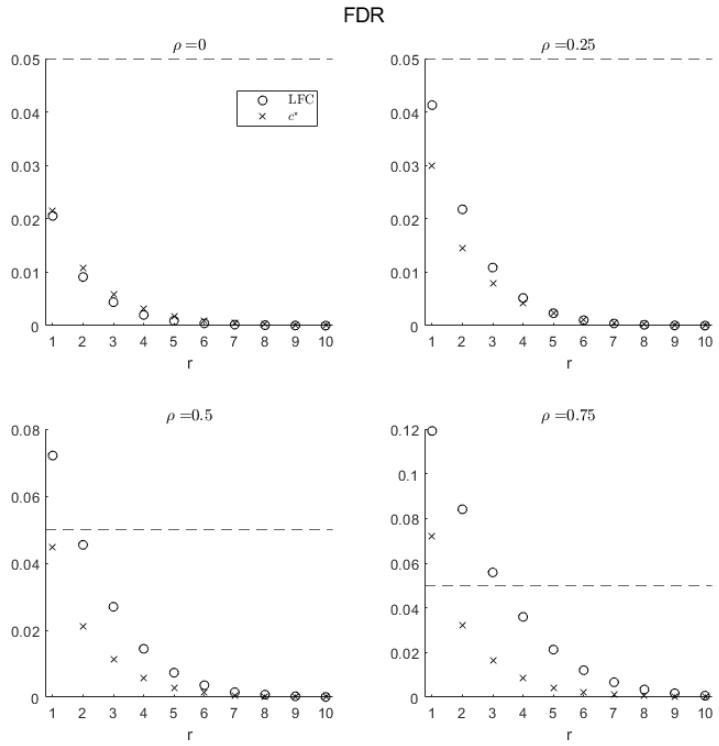


Figure 3.8: The FDR of the adaptive LSU tests. The subplots differ in the value of the correlation parameter ρ. Considered is the modified Storey estimator with $\lambda = 1/2$, used with the LFC-based p-values (circles), or the randomized p-values with $c = c^\star$ (crosses), respectively. The signal strength is denoted by r. The horizontal dashed lines are at $\alpha = 0.05$.

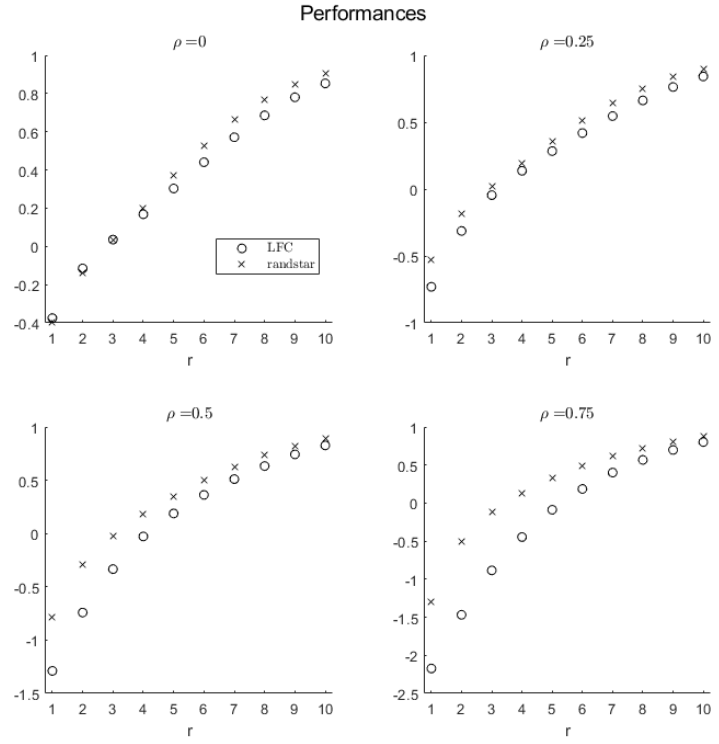


Figure 3.9: The performances of the adaptive LSU tests. The subplots differ in the value of the correlation parameter ρ. Considered is the modified Storey estimator with $\lambda = 1/2$, used with the LFC-based p-values (circles), or the randomized p-values with $c = c^\star$ (crosses), respectively. The signal strength is denoted by r.

For example, Meinshausen and Rice (2006) consider the problem of estimating a lower bound $\hat{\pi}_1$ such that

$$\mathbb{P}_\vartheta\left(\hat{\pi}_1 \leq \pi_1\right) \geq 1-\alpha, \tag{3.11}$$

where $\pi_1 = 1-\pi_0$ is the proportion of false null hypotheses. They propose the estimator

$$\hat{\pi}_1 = \sup_{\lambda\in(0,1)}\left[\frac{\hat{F}_m(\lambda)-\lambda-c_m(\lambda)}{1-\lambda}\right] = \sup_{\lambda\in(0,1)}\left[1-\hat{\pi}_0(\lambda)-\frac{c_m(\lambda)}{1-\lambda}\right],$$

where $c_m(\lambda)$ is a constant which is independent of the model. For valid p-values, $1-\hat{\pi}_0(\lambda)$ is a non-positively biased estimator for π_1, and the constant $c_m(\lambda)/(1-\lambda)$ is to make sure that (3.11) is satisfied. However, if the null p-values are conservative rather than uniformly distributed, we face an analogous problem as with the Schweder-Spjøtvoll (point) estimator. For stochastically increasing p-values, the expected value of the estimator $\hat{\pi}_1$ gets smaller and therefore provides a less informative lower bound. Using randomized p-values that come closer to uniformity under nulls instead of conservative LFC-based p-values can therefore help to minimize the size of the confidence region.

Uniformly distributed p-values under null hypotheses are also beneficial in a different, but related context. Ghosal and Roy (2011) consider a mixture model for the p-values. For a parameter θ_0 and a parameter space Θ_1 where $\theta_0 \notin \Theta_1$, let $\{h_\theta : \theta \in \Theta = \{\theta_0\} \cup \Theta_1\}$, be a parametrized family of Lebesgue-densities. Assume that each (transformed) p-value in the multiple testing problem has an overall density

$$h(t) = \pi_0 h_{\theta_0}(t) + (1-\pi_0)\int_{\theta\in\Theta_1} h_\theta(t)\mathrm{d}G(\theta),$$

where G is a distribution over the parameter space Θ_1. Thus, each p-value has a probability of π_0 of being null and the density h_{θ_0} is such that the p-values are uniformly distributed under the (simple) nulls. Among other conditions, if it holds $h_\theta(t)/h_{\theta_0}(t) \to 0$ for $|t| \to \infty$ and each parameter $\theta \in \Theta_1$, then h uniquely identifies π_0. An analogous result holds for the characteristic functions of the densities h_θ, $\theta \in \Theta$. In practice, assuming conservative p-values under the null hypotheses instead of uniformly distributed ones, leads to an overestimation of $h(t)$ for large t and therefore to overestimations of π_0. Switching to our more uniformly distributed p-values can therefore result in more accurate estimations for π_0 in this context.

3.6.2 Comparison with other approaches that address conservativity of p-values

In previous literature, there have been other approaches that deal with the problem of conservative p-values. For example, Zhao et al. (2019) propose discarding p-values that are larger than a constant $\tau \in (0,1)$, and multiplying the remaining p-values with $1/\tau$. This allows them to use a smaller number of p-values in the test. They show that these transformed p-values $\{p_j^{LFC}/\tau \mid p_j^{LFC} \leq \tau\}$ are valid if the p-value p_j^{LFC} is uniformly conservative, which is equivalent to condition (1.) in Theorem 1, where $\tau = c_j$, $j = 1,\ldots,m$. More particularly, if any of the conditions in Theorem 1 hold, both the transformed p-value p_j^{LFC}/τ, given $p_j^{LFC} \leq \tau$, and our randomized p-value $p_j^{rand}(X, U_j, \tau)$ are valid, $j = 1,\ldots,m$. The difference is that instead of discarding the p-values that are larger than τ, we replace them with the uniformly distributed random variables $U_1,\ldots,U_m$.

Among other things, Zhao et al. (2019) consider the global null hypothesis $H_0 = \bigcap_{j=1}^m H_j$ that all individual null hypotheses are true, and present several examples of global p-values for testing H_0. One method is the so-called Bonferroni-correction, in which we reject H_0 if any of the marginal p-values is smaller than α/m, where $\alpha \in (0,1)$ is a significance level. In other words, we reject H_0 if

$$\min\{p_j^{LFC} \mid j = 1,\ldots,m\} \cdot m \leq \alpha.$$

Let $S_\tau = \{j \mid p_j^{LFC} \leq \tau\}$ be the set of non-discarded indices, then the conditional, global Bonferroni p-value, based on the adjusted p-values by Zhao et al. (2019), is $\min\{p_j^{LFC}/\tau \mid j = 1,\ldots,m\}|S_\tau|$. Using our randomized p-values with $c_j = \tau$ for each j, yields

$$\min\left\{p_j^{LFC}\mathbf{1}\{p_j^{LFC} \leq \tau\}/\tau + U_j\mathbf{1}\{p_j^{LFC} > \tau\} \mid j = 1,\ldots,m\right\} \cdot m$$

instead. Thus, using our randomized p-values leads to the smaller (not larger) min-term, however, $|S_\tau| \leq m$, so none of the two methods is strictly better than the other in every setting. Similarly, if applying Fisher's combination, the adjusted p-values by Zhao et al. (2019) replace the discarded p-values by 1, where our randomized p-values replace them by $U_j \leq 1$, thus making the statistic that tests

the global null H_0 larger when using our randomized p-values. However, the discard method gains an advantage by replacing m with $|S_\tau|$.

In Tian and Ramdas (2019), a similar method is employed in the context of online testing based on the concept of α consumption. Consider a sequence of null hypotheses $H_1, H_2, \ldots$, and denote I_0 the set of indices j for which H_j is true. Furthermore, denote the false discovery proportion (FDP) at point $t > 0$ by $\text{FDP}(t) = |R(t) \cap I_0| / \max\{|R(t)|, 1\}$, where $R(t)$ is the set of indices $j \leq t$ for which H_j has been rejected. For some sequence $(\alpha_j)_{j=1}^{\infty}$ in the range $[0, 1]$, they first consider the oracle estimate

$$\text{FDP}^\star(t) = \frac{\sum_{j=1}^{t} \alpha_j \mathbf{1}\{j \in I_0\}}{\max\{|R(t)|, 1\}} \tag{3.12}$$

for the FDP.

Notice, however, that the set I_0 is unknown. A conservative approach would be to replace $\mathbf{1}\{j \in I_0\}$ by one in (3.12). In a certain sense, this may be interpreted as estimating π_0 as one. This conservative approach has then been improved by utilizing the estimate

$$\widehat{\text{FDP}}_{\text{SAFFRON}}(t) = \frac{\sum_{j \leq t} \alpha_j \frac{\mathbf{1}\{p_j > \lambda\}}{1-\lambda}}{\max\{|R(t)|, 1\}}. \tag{3.13}$$

Comparing the numerators in (3.12) and (3.13), we conclude that the SAFFRON approach is similar to estimating π_0 by the Schweder-Spjøtvoll estimator $\hat{\pi}_0(\lambda)$, when $t = m$. However, as pointed out in the previous sections, $\hat{\pi}_0(\lambda)$ overestimates π_0 if the null p-values are conservative. Therefore, Tian and Ramdas (2019) propose to use $\mathbf{1}\{\lambda\tau < p_j \leq \tau\}$ instead of $\mathbf{1}\{p_j > \lambda\}$, where $\tau \in (0, 1)$, resulting in

$$\widehat{\text{FDP}}_{\text{ADDIS}}(t) = \frac{\sum_{j \leq t} \alpha_j \frac{\mathbf{1}\{\lambda\tau < p_j \leq \tau\}}{\tau - \lambda\tau}}{\max\{|R(t)|, 1\}}.$$

Again, translating this to a π_0 estimator, when $t = m$, we get

$$\hat{\pi}_0^{discard}(\lambda, \tau) = \frac{1}{m} \sum_{j=1}^{m} \frac{\mathbf{1}\{\lambda\tau < p_j \leq \tau\}}{\tau(1-\lambda)} = \frac{1}{\tau m} \sum_{j=1}^{m} \mathbf{1}\{p_j \leq \tau\} \frac{\mathbf{1}\{\lambda < p_j/\tau\}}{1-\lambda}.$$

The estimator $\hat{\pi}_0^{discard}(\lambda, \tau)$ is an unbiased estimator for π_0 if the non-null p-values are almost surely smaller than $\lambda\tau$, and the null p-values are uniformly distributed. However, $\hat{\pi}_0^{discard}(\lambda, \tau)$ is generally not non-negatively biased if the p-values are merely valid. This means that this estimator does not satisfy (3.9).

Furthermore, the estimator $\hat{\pi}_0^{discard}(\lambda, \tau)$ is similar to using the adjusted p-values from Zhao et al. (2019) in the Schweder-Spjøtvoll estimator $\hat{\pi}_0(\lambda)$, where instead of $|S_\tau|$, the number of non-discarded p-values, it employs τm. Albeit, Tian and Ramdas (2019) show that

$$\mathbb{E}_\vartheta\left[\frac{\mathbf{1}\{\lambda\tau < p_j \leq \tau\}}{\tau(1-\lambda)}\right] \leq \mathbb{E}_\vartheta\left[\frac{\mathbf{1}\{\lambda < p_j\}}{1-\lambda}\right],$$

holds, if the cdf of p_j is convex under ϑ, and p_j is thus uniformly conservative, cf. condition (3.) in Theorem 3.1. Therefore, if π_0 is large and the null p-values are uniformly conservative, $\hat{\pi}_0^{discard}(\lambda, \tau)$ has a lower bias than $\hat{\pi}_0(\lambda)$ (but possibly negative), and $\widehat{\text{FDP}}_{\text{ADDIS}}(t) \leq \widehat{\text{FDP}}_{\text{SAFFRON}}(t)$.

3.7 Discussion

We have demonstrated how randomized p-values can be utilized in the Schweder-Spjøtvoll estimator $\hat{\pi}_0$. Whenever composite null hypotheses are under consideration, our proposed approach leads to a reduction of the bias and of the MSE of $\hat{\pi}_0$, when compared to the usage of LFC-based p-values, at least in our simulations. Furthermore, our approach also robustifies $\hat{\pi}_0$ against dependencies among $p_1^{LFC}(X), \ldots, p_m^{LFC}(X)$. The latter property is important in modern high-dimensional applications, where the biological and/or technological mechanisms involved in the data-generating process virtually always lead to dependencies (cf. Stange et al. (2016)), especially in studies with multiple endpoints which are all measured for the same observational units. Furthermore, we have explained in detail how the proposed methodology can be applied in practice. Worksheets in `R`, with which all results of the present work can be reproduced, are available from the first author upon request.

Statistical models that fulfil any of the conditions (2.) – (5.) from Theorem 3.1 admit valid randomized p-values $\{p_j^{rand}(X, U_j, c_j)\}_{1 \leq j \leq m}$ for any choice of the constants $(c_j)_{1 \leq j \leq m} \in [0, 1]^m$. We gave two such

models in Examples 3.1 and 3.2. These models have a variety of applications, for instance in the life sciences; cf., e. g., Part II of Dickhaus (2014). Closely related examples are the replicability models considered in Hoang and Dickhaus (2022). Identifying additional model classes that have that property is a topic for future research. Furthermore, in models for which the j-th LFC-based p-value is of the form $p_j^{LFC}(X) = 1 - F_{\vartheta_0}(T_j(X))$ for $1 \leq j \leq m$ and in which $(T_j(X)^{(\vartheta)})_{\theta_j(\vartheta)}$ is an MLR family, the cdf of $p_j^{rand}(X, U_j, R_j)$ is always between those of Uni$[0,1]$ and $p_j^{LFC}(X)$. The latter follows from point 1 in Remark 3.2 and Theorem 3.2. Distributions with the MLR property include exponential families, for example the family of univariate normal distributions with fixed variance and the family of Gamma distributions (cf. Karlin and Rubin (1956a)). Also, the family of non-central t-distributions and the family of non-central F-distributions have the MLR property with respect to their non-centrality parameters (cf. Karlin (1956)). It is of interest to deeper investigate properties of our randomized p-values in such models.

There are several further possible extensions of the present work. First, in Section 3.4 we only considered the usage of $p_1^{rand}(X, U_1, c_1), \ldots, p_m^{rand}(X, U_m, c_m)$ in $\hat{\pi}_0$ for identical constants $c_1 = \cdots = c_m \equiv c$. In future work, it may be of interest to develop a method for choosing each c_j individually, for instance depending on the size of the j-th LFC-based p-value. Second, we have chosen c_0 in Section 3.4.3 such, that the conditional (to the observed data $X = x$) bias of $\hat{\pi}_0(\lambda)$ is minimized. Another approach, which can be pursued in future research, is to choose a c_0 that minimizes the MSE of $\hat{\pi}_0(\lambda)$ instead. Third, we restricted our attention to the Schweder-Spjøtvoll estimator $\hat{\pi}_0(\lambda)$. However, there exists a wide variety of other ecdf-based estimators in the literature (see, for instance, Table 1 in Chen (2019) for a recent overview), which are prone to suffer from the same issues as $\hat{\pi}_0(\lambda)$ when used with LFC-based p-values in the context of composite null hypotheses. One other ecdf-based estimator for π_0 is the more conservative estimator $\hat{\pi}_0^+(\lambda) = \hat{\pi}_0(\lambda) + 1/(m(1-\lambda))$ proposed by Storey (2002). The bias of $\hat{\pi}_0^+$ when used with the randomized p-values $p_1^{rand}(X, U_1, c), \ldots, p_m^{rand}(X, U_m, c)$ is minimized for the same $c = c^\star$ from Section 3.4. Thus, the same algorithm as outlined in Section 3.4.3 can be applied to $\hat{\pi}_0^+$ in practice. In future research, randomization approaches for other ecdf-based estimators can be investigated.

We have not elaborated on the choice of λ in the present work. The standard choice of $\lambda = 1/2$ seemed to work reasonably well in connection with our proposed randomized p-values. We have also performed some preliminary sensitivity analyses (not included here) with respect to λ, which indicated that the sensitivity of $\hat{\pi}_0$ with respect to λ is less pronounced for the case of randomized p-values than for the case of LFC-based p-values. Investigating this phenomenon deeper, both from the theoretical and from the numerical perspective, is also a worthwhile topic for future research.

Finally, in case of composite null hypotheses under discrete models, one may apply two stages of randomization: In the first stage, the discreteness of the model is addressed by applying a randomization procedure as proposed by Dickhaus et al. (2012). Under LFCs, this leads to exactly uniformly distributed randomized p-values. Then, in a second stage of randomization, the approach proposed in this work helps to alleviate the problem of conservative p-values (under non-LFCs) resulting from the composite nature of the null hypotheses.

Chapter 4

Combining independent p-values in replicability analysis: A comparative study

This chapter is derived in part from an article published in Journal of Statistical Computation and Simulation (copyright Taylor & Francis), available online: `http://www.tandfonline.com/10.1080/00949655.2021.2022678`. Appendix C contains the original appendix of the paper.

Authors

Anh-Tuan Hoang, Institute for Statistics, University of Bremen, Bremen, Germany

Prof. Dr. Thorsten Dickhaus, Institute for Statistics, University of Bremen, Bremen, Germany

Abstract Given a family of null hypotheses $H_1, \ldots, H_s$, we are interested in the hypothesis H_s^γ that at most $\gamma - 1$ of these null hypotheses are false. Assuming that the corresponding p-values are independent, we are investigating combined p-values that are valid for testing H_s^γ. In various settings in which H_s^γ is false, we determine which combined p-value works well in which setting. Via simulations, we find that the Stouffer method works well if the null p-values are uniformly distributed and the signal strength is low, and the Fisher method works better if the null p-values are conservative, i.e. stochastically larger than the uniform distribution. The minimum method works well if the evidence for the rejection of H_s^γ is focused on only a few non-null p-values, especially if the null p-values are conservative. Methods that incorporate the combination of e-values work well if the null hypotheses $H_1, \ldots, H_s$ are simple.

Summary In this chapter we investigate which combination functions for PC null hypotheses work well in what type of settings.

We derive valid p-values for PC null hypotheses $H^{\gamma/s}$ from global null hypothesis combination functions by essentially combining only the $s - \gamma + 1$ largest base p-values. Apart from some averaging and minimum approaches of combining p-values we also consider e-values whose relation to p-values is roughly inverse, where higher e-values entail stronger evidence against the null, cf. Grünwald et al. (2020); Vovk and Wang (2021). The latter we calculate as Bayes-factors, see Section 4.4.2.

We generate the base p-values from two models, a Beta-Model and a Normal-Model. In the latter model, the p-values result from a Gaussian shift model with known variance. We assume that the base p-values are stochastically independent, and that they are stochastically larger / smaller inside the null / alternative away from the LFC. For $s = 6$, the evidence patterns with different numbers of false null hypotheses and signal strengths, that we consider in the simulations, are outlined in Table 4.1. Furthermore, we distinguish between the non-conservative versions of the parameter configurations, where null p-values are uniform, and their conservative versions, if applicable.

In simulations, investigating the power of various combination functions, we find that utilizing e-values works better in non-conservative parameter configurations. Averaging methods work better than minimum methods if the evidence is more spread, and vice versa if the evidence is more focused, and the Stouffer p-value works better than the Fisher p-value if the null p-values are uniform. Furthermore, the minimum approach works better if γ is large. We find similar results under the (PC) null, where the minimum approach works (relatively) better if the number of conservative null p-values is high.

Declaration of individual contributions Co-author and supervisor Prof. Dr. Thorsten Dickhaus came up with the idea for the paper, and I developed the theoretical results including their proofs. Simulations and evaluations including the figures and tables pertaining to the simulations were done by me. The final text was written and proof-read by both authors.

4.1 Introduction

Given a set of studies, which are examining related research hypotheses under different conditions, it is often of interest to assess whether findings can be made in at least $\gamma \geq 2$ of the considered studies. Studies may, for example, differ in their population or laboratory methods. The search for results in at least two studies is called replicability analysis. It alleviates the possibility that a positive outcome depends on the specific settings of a single study.

Formally, we consider a family of $s \geq 2$ null hypotheses $H_1, \ldots, H_s$ and their corresponding alternative hypotheses $K_1, \ldots, K_s$. For each pair of hypotheses H_i and K_i we assume that a p-value p_i is available and that these p-values $p_1, \ldots, p_s$ are jointly stochastically independent. For $\gamma \leq s$, we are interested in the partial conjunction/replicability null hypothesis

$$H_s^\gamma = \{\text{at least } s - \gamma + 1 \text{ of the null hypotheses } H_1, \ldots, H_s \text{ are true}\}, \tag{4.1}$$

thus its alternative is that at least γ null hypotheses are false. Our goal is to compare different p-value combinations for $p_1, \ldots, p_s$ that are valid for H_s^γ. A p-value is called valid for a null hypothesis H if it is stochastically not smaller than the uniform distribution on $[0, 1]$ (Uni$[0, 1]$) under all parameter values that entail validity of H.

Since it holds $H_s^1 \subseteq H_s^2 \subseteq \cdots \subseteq H_s^s$, valid p-values for H_s^1 need not be valid for H_s^γ, $\gamma \geq 2$. Benjamini and Heller (2008) investigated the theory of testing the partial conjunction null hypotheses H_s^γ. They show that valid p-values for H_s^γ can be derived from combination p-values for $H_{s-\gamma+1}^1$ by essentially combining the $s - \gamma + 1$ largest p-values.

There are several ways to combine the independent p-values $p_1, \ldots, p_s$ of a set of null hypotheses $H_1, \ldots, H_s$ to test for the null hypothesis H_s^1. Birnbaum (1954) showed that for each p-value combination that is non-decreasing in each p-value, there exists an alternative hypothesis for which the combination is best. Nevertheless, we consider different null hypothesis setups, and identify which of our considered combinations work best in which general situation.

A common way of combining p-values for H_s^1 is via averaging, see also Vovk and Wang (2020). This approach evaluates the sum of the transformed p-values $f(p_1, \ldots, p_s) = \sum_i \varphi(p_i)$. If the distribution of $f(U_1, \ldots, U_s)$ is known with cdf F_f, where $U_i \sim$ Uni$[0, 1]$, $i = 1, \ldots, s$, then defining $F_f(f(p_1, \ldots, p_s))$ leads to a valid p-value under H_s^1. Two well known examples of such are Fisher's method which uses $\varphi = \log$, and Stouffer's method, which uses $\varphi(x) = -\Phi^{-1}(1 - x)$, where Φ^{-1} is the quantile function of the standard normal distribution on $\mathbb{R}$. If the distribution of $f(U_1, \ldots, U_s)$ is unknown, one can instead modify $p = G_0(f(p_1, \ldots, p_s))$ with a suitable function G_0 so that p is at least valid under H_s^1. This includes for example the arithmetic mean and the harmonic mean, cf. Rüschendorf (1982), Vovk and Wang (2020), Wilson (2019). Either way, large p-values can overshadow small p-values in averaging methods, which can be problematic if the null p-values are conservative under nulls, that is, if they are stochastically larger than Uni$[0, 1]$. On the other hand, none of the p-values need to be smaller than a significance level $\alpha \in (0, 1)$ for the combined p-value to be smaller than α. Thus, averaging methods can be powerful if the evidence for a rejection of the global null hypothesis is spread out between the p-values. Pearson's method of averaging via a product of the p-values is of a similar nature, cf. Pearson (1938).

On the other hand, there are p-value combination functions that do not take the size of all p-values fully into account. For example the combined p-value $s \cdot \min\{p_1, \ldots, p_s\}$ resulting from the Bonferroni method is relatively unaffected by conservative p-values, but the p-value cannot be smaller than α if none of the marginal p-values are. Similarly, since the minimum $\min\{p_1, \ldots, p_s\}$ of stochastically independent Uni$[0, 1]$-distributed p-values is Beta-distributed with parameters 1 and s, Beta$(1, s)$, we can also consider $p = F_{\text{Beta}(1,s)}(\min\{p_1, \ldots, p_s\})$ as a valid p-value for H_s^1. Another example is the maximum of the p-values, which is valid for H_s^1 if the p-values are independent, cf. Vovk and Wang (2020).

Similarly to the work of Loughin (2004), we differentiate between alternative hypotheses that have minimally spread evidence and ones that have spread out evidence among all the false null hypotheses. If $s = 6$ and $\gamma = 2$, the null hypothesis H_s^γ is, for example, false if only two null hypotheses are false or if all null hypotheses are false. However, Loughin (2004) only considered the global null hypotheses H_s^1, that every null hypothesis is true. We extend his work by taking the more general partial conjunction hypothesis H_s^γ into account. Birnbaum (1954) already noted that in case of $s = 2$ studies, the Wilkinson p-value (case 1) is more sensitive to evidence in one study than the Fisher p-value, cf. Wilkinson (1951). Furthermore, Loughin (2004) only considered Uni$[0, 1]$-distributed null p-values. However, it is known

that for example in case of composite null hypotheses, conservative null p-values are more common, cf. Hoang and Dickhaus (2021b), Hoang and Dickhaus (2022). In simulations, we investigate how well the different p-value combination functions deal with conservative null p-values.

Not covered in this paper is the kind of meta analysis that tests H_s^1 against the alternative of H_s^s, which is a proper subset of the alternative of H_s^1 if $s > 1$, i.e. each of the null hypotheses $H_1, \ldots, H_s$ are either all true or all false (for example repetition of an experiment). Kocak (2017) modeled the marginal p-values under alternatives as Beta(α, β)-distributed, Beta-distributed with parameters α and β, and determined which p-value combinations work well for this question in which subsets of $(0, \infty)^2$ for the parameters (α, β). Under the assumption that the marginal p-values are Uni$[0, 1]$-distributed under H_s^1, Heard and Rubin-Delanchy (2018) calculate p-value combinations as likelihood ratio tests and therefore uniformly most powerful test statistics for the above kind of meta analysis under several models.

This work is structured as follows. In Section 4.2 we introduce our model and notations, and in Section 4.3 we present the p-value combinations that we consider. Section 4.4 contains our comparisons of the p-value combination functions via simulations. Finally, we conclude with a discussion in Section 4.5.

4.2 Model Setup

Let $(\Omega, \mathcal{F}, (\mathbb{P}_{\boldsymbol{\theta}})_{\boldsymbol{\theta} \in \Theta})$ be a statistical model, and let $\boldsymbol{X}$ be the data and $\boldsymbol{\theta} \in \Theta$ the parameter of the model and Θ the corresponding parameter space. We consider a set of null hypotheses $(H_i)_{i=1,\ldots,s}$ and their corresponding alternatives $(K_i)_{i=1,\ldots,s}$ such that H_i and K_i are non-empty subsets of the parameter space $\Theta \subseteq \mathbb{R}^s$. We assume that $H_i = \{\boldsymbol{\theta} : \theta_i \leq 0\}$ and $K_i = \{\boldsymbol{\theta} : \theta_i > 0\}$ for each $i = 1, \ldots, s$. Thus, each hypothesis pair H_i and K_i only depends on the i-th component of the parameter value $\boldsymbol{\theta}$, $i = 1, \ldots, s$.

Let a set of corresponding p-values $(p_i)_{i=1,\ldots,s}$ be given, such that for any parameter value $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_s)^T \in \Theta$

$$p_i(\boldsymbol{X}) \sim f_{\theta_i},\ i = 1, \ldots, s,$$

where f_{θ_i} is a Lebesgue density with support on $[0, 1]$. More particularly, we assume that the density function f_{θ_i} only depends on the i-th component θ_i of $\boldsymbol{\theta}$. Throughout this paper, we use $p_i \equiv p_i(\boldsymbol{X})$ by abuse of notation, and say 'p-value' and 'p-variable' interchangeably.

We make the following general assumptions to our model:

($A1$) The p-values $p_1, \ldots, p_s$ are jointly stochastically independent under each parameter value $\boldsymbol{\theta} \in \Theta$.

($A2$) Under any $\boldsymbol{\theta} \in \Theta$ such that $\theta_i = 0$, we assume that $f_0(t) = \mathbf{1}\{0 \leq t \leq 1\}$, i.e. that $p_i(\boldsymbol{X})$ is Uni$[0, 1]$-distributed under $\theta_i = 0$, $i = 1, \ldots, s$.

($A3$) For each i, we assume that the p-value $p_i(\boldsymbol{X})$ is stochastically decreasing in θ_i, i.e. $p_i(\boldsymbol{X})^{(\theta_i)} \leq_{\text{st}} p_i(\boldsymbol{X})^{(\tilde{\theta}_i)}$ if and only if $\theta_i \geq \tilde{\theta}_i$.

Assumption ($A1$) is for example fulfilled if the null hypotheses $H_1, \ldots, H_s$ are from a set of independent studies. If the parameter space Θ contains no parameter values $\boldsymbol{\theta}$ with negative i-th components θ_i, p_i is Uni$[0, 1]$-distributed under each $\boldsymbol{\theta} \in H_i$. Otherwise, p_i may be conservative if $\theta_i < 0$, $i = 1, \ldots, s$.

The relation $\leq_{\text{st}}$ in assumption ($A3$) denotes the usual stochastic order between two random variables, cf. for example (Shaked and Shanthikumar, 2007, Chapter 1.A). The notation $p_i(\boldsymbol{X})^{(\theta_i)}$ refers to the distribution of $p_i(\boldsymbol{X})$ under θ_i. Under assumptions ($A2$) and ($A3$), p_i is a valid p-value for H_i and parameters $\boldsymbol{\theta}$ with $\theta_i = 0$ are the least favorable parameter configurations (LFC parameters). See for example Section 2 in Hoang and Dickhaus (2022) for a definition of LFC-based p-values.

Remark 4.1. *Assumption* ($A3$) *is for example fulfilled if p_i is an antitone transformation of a test statistic $T_i(\boldsymbol{X})$ such that $(T_i(\boldsymbol{X})^{(\theta_i)})_{\theta_i}$ is likelihood ratio ordered, that is, if the distribution of $T_i(\boldsymbol{X})$ under $\boldsymbol{\theta}$ is smaller under the likelihood ratio order than under $\tilde{\boldsymbol{\theta}}$ if and only if $\theta_i \leq \tilde{\theta}_i$ holds for their i-th components (cf. for example Chapter* 1.C *in Shaked and Shanthikumar (2007) for a definition of the likelihood ratio order).*

We are interested in the partial conjunction null hypothesis H_s^γ from (4.1), where $1 \leq \gamma \leq s$ is a given constant. The goal of this work is to compare p-value combination maps $f : [0, 1]^s \to [0, 1]$ for which $f(p_1, \ldots, p_s)$ is a valid p-value for H_s^γ.

4.3 Combination functions for p-values

In this section, we introduce the p-value combinations $f(p_1, \ldots, p_s)$ that we investigate for the null hypothesis H_s^γ. Let $U_1, \ldots, U_s$ be stochastically independent and identically Uni$[0, 1]$-distributed random variables.

We first assume the existence of a p-value combination function $g : [0,1]^{s-\gamma+1} \to [0,1]$, that is non-decreasing in each argument and valid for the null hypothesis $H^1_{s-\gamma+1}$, i.e.

$$\text{Uni}[0,1] \leq_{\text{st}} g\left(U_1, \ldots, U_{s-\gamma+1}\right). \tag{4.2}$$

Let $p_{(1)} \leq \ldots \leq p_{(s)}$ be the ordered p-values. According to Lemma 1 in Benjamini and Heller (2008), this combination function applied to the $s-\gamma+1$ largest p-values $p_{(\gamma)}, \ldots, p_{(s)}$ among $p_1, \ldots, p_s$ is valid for H^γ_s, i.e. $f\left(p_1, \ldots, p_s\right)$ is valid for H^γ_s, where $f : [0,1]^s \to [0,1]$ is a combination function with

$$f\left(p_1, \ldots, p_s\right) = g\left(p_{(\gamma)}, \ldots, p_{(s)}\right). \tag{4.3}$$

Hence, in order to find p-value combination functions for the partial conjunction null hypothesis H^γ_s, we only have to consider p-value combination functions $g : [0,1]^{s-\gamma+1} \to [0,1]$ for $H^1_{s-\gamma+1}$. The functions g, that we use in this paper, can be divided into two classes.

1. We have a component-wise non-increasing function $g_0 : [0,1]^{s-\gamma+1} \to \mathbb{R}$, such that the distribution of $g_0\left(U_1, \ldots, U_{s-\gamma+1}\right)$ is known with continuous cdf F_{g_0}. We then define the p-value $g\left(p_1, \ldots, p_{s-\gamma+1}\right) = 1 - F_{g_0}\left(g_0\left(p_1, \ldots, p_{s-\gamma+1}\right)\right)$. Note, that it holds $g\left(U_1, \ldots, U_{s-\gamma+1}\right) \sim \text{Uni}\left[0,1\right]$ by the principle of probability integral transform.

2. For a component-wise non-decreasing function $g_0 : [0,1]^{s-\gamma+1} \to \mathbb{R}$ we consider $g_0\left(p_1, \ldots, p_{s-\gamma+1}\right)$ and find a constant $c \in \mathbb{R}$ such that $\text{Uni}\left[0,1\right] \leq_{\text{st}} c \cdot g_0\left(U_1, \ldots, U_{s-\gamma+1}\right)$. We then define the p-value $g\left(p_1, \ldots, p_{s-\gamma+1}\right) = c \cdot g_0\left(p_1, \ldots, p_{s-\gamma+1}\right)$.

In the following, we take some well known p-value combination functions g for $H^1_{s-\gamma+1}$ from previous literature. Firstly, we consider the Fisher and the Stouffer combination. The combined p-value by Fisher for $H^1_{s-\gamma+1}$ applied to $p_{(\gamma)}, \ldots, p_{(s)}$ is defined as

$$g(p_{(\gamma)}, \ldots, p_{(s)}) = 1 - F_{\chi^2_{2(s-\gamma+1)}}\left(-2\sum_{i=\gamma}^{s} \log\left(p_{(i)}\right)\right),$$

where $F_{\chi^2_{2(s-\gamma+1)}}$ is the cdf of the χ^2-distribution with $2(s-\gamma+1)$ degrees of freedom (cf. (Fisher, 1973, Section 21.1)). This combination function uses the fact that $f_0(U_1, \ldots, U_{s-\gamma+1}) = -2\sum_{i=1}^{s-\gamma+1} \log\left(U_i\right)$ is chi-square distributed with $2(s-\gamma+1)$ degrees of freedom.

The combined p-value by Stouffer for $H^1_{s-\gamma+1}$ applied to $p_{(\gamma)}, \ldots, p_{(s)}$ is defined as

$$g(p_{(\gamma)}, \ldots, p_{(s)}) = 1 - \Phi\left(\frac{1}{\sqrt{s-\gamma+1}}\sum_{i=\gamma}^{s} \Phi^{-1}\left(1-p_{(i)}\right)\right).$$

where Φ is the cdf of the standard normal distribution on $\mathbb{R}$ (cf. (Stouffer et al., 1949, Footnote 14 in Section V of Chapter 4)). This combination function uses the fact that $g(U_1, \ldots, U_{s-\gamma+1}) = (s-\gamma+1)^{-1/2}\sum_{i=1}^{s-\gamma+1} \Phi^{-1}\left(1-U_i\right)$ is standard normally distributed. Both combined p-values require the p-values $p_1, \ldots, p_s$ to be stochastically independent.

The next two combined p-values evaluate only the smallest p-value. The combined p-value using the minimum is defined by

$$g(p_{(\gamma)}, \ldots, p_{(s)}) = F_{\text{Beta}(1,s-\gamma+1)}\left(p_{(\gamma)}\right),$$

where $F_{\text{Beta}(1,s-\gamma+1)}$ is the cdf of the Beta-distribution with parameters 1 and $s-\gamma+1$. It requires that the p-values are stochastically independent, and is motivated by the fact that $g(U_1, \ldots, U_{s-\gamma+1}) = \min\left\{U_1, \ldots, U_{s-\gamma+1}\right\}$ is Beta$(1, s-\gamma+1)$-distributed. The Bonferroni method, which utilizes the Bonferroni inequality, leads to

$$g(p_{(\gamma)}, \ldots, p_{(s)}) = (s-\gamma+1)p_{(\gamma)}.$$

It also evaluates the minimum but does not require independent p-values.

Some further p-value combination functions that we consider make use of so-called e-values (see Grünwald et al. (2020); Vovk and Wang (2019)). Their relation to p-values is roughly inverse, where higher e-values entail stronger evidence against the null. In our simulations in Section 4.4, we calculate a Bayes factor e_j for each null hypothesis H_j, $j = 1, \ldots, s$. These are in some cases e-values, i.e. random variables with expected values not greater than one under H_j. More details on this problem are provided in Section 4.4.4 and in Appendix A.1 of Vovk and Wang (2020).

Analogously to the problem of p-values, we define a combination function h for H^γ_s by

$$h(e_1, \ldots, e_s) = h_0(e_{(1)}, \ldots, e_{(s-\gamma+1)}),$$

where h_0 is a valid combination function for $H^1_{s-\gamma+1}$, i.e. $h_0(e_1, \ldots, e_{s-\gamma+1})$ is a valid e-value for $H^1_{s-\gamma+1} = \bigcap_{i=1}^{s-\gamma+1} H_i$ if $e_1, \ldots, e_{s-\gamma+1}$ are valid e-values for $H_1, \ldots, H_{s-\gamma+1}$, respectively. We explain in Appendix C why h is a valid combination function for H^γ_s. Finally, to compare the e-value approaches to the ones utilizing p-values, we transform the e-value $h(e_1, \ldots, e_s)$ to a p-value, $\max\{h(e_1, \ldots, e_s)^{-1}, 1\}$, for H^γ_s (where $0^{-1} = 1$ and $\infty^{-1} = 0$).

Some examples of e-value combination functions h_0 for $H^1_{s-\gamma+1}$ include the arithmetic mean given by

$$h_0\left(e_1, \ldots, e_{s-\gamma+1}\right) = \frac{1}{s-\gamma+1} \sum_{i=1}^{s-\gamma+1} e_i,$$

and the product given by

$$h_0\left(e_1, \ldots, e_{s-\gamma+1}\right) = \prod_{i=1}^{s-\gamma+1} e_i,$$

(cf. Vovk and Wang (2019)). Some reasoning on why we chose these functions for h_0 is given in Propositions 3.1 and 4.2 in Vovk and Wang (2020).

4.4 Simulations

In this section we compare the the p-values for H^γ_s from Section 4.3 in simulations. The marginal p-values $p_1, \ldots, p_s$ in our simulations are given by two different models.

4.4.1 Models for p-value generation

We consider Beta-distributed p-values, which has also been used for example by Loughin (2004). For a parameter value $\boldsymbol{\theta} \in \Theta = \mathbb{R}^s$, we define the density function f_{θ_i} of the i-th p-value as

$$\begin{cases} f_{\theta_i} \sim \text{Beta}\left(1-\theta_i, 1\right), & \theta_i \leq 0, \\ f_{\theta_i} \sim \text{Beta}\left(1, 1+\theta_i\right), & \theta_i > 0, \end{cases}$$

where $\text{Beta}(\alpha, \beta)$ denotes the Beta-distribution with parameters α and β.

As the second model, we consider the Normal-Model, where the p-values result from a Gaussian shift model with known variance $\sigma^2 > 0$. Here, we define the density function of the i-th p-value as

$$f_{\theta_i}(t) = \frac{\varphi_{(\theta_i, \sigma^2)}\left(\Phi^{-1}_{(0,\sigma^2)}(1-t)\right)}{\varphi_{(0,\sigma^2)}\left(\Phi^{-1}_{(0,\sigma^2)}(1-t)\right)}, \; t \in [0,1],$$

where $\varphi_{(\mu,\sigma^2)}$ is the density function, and $\Phi^{-1}_{(\mu,\sigma^2)}$ the quantile function of the normal distribution with expected value μ and variance σ^2.

Lemma 4.1. *Both models satisfy Assumptions* $(A1) - (A3)$ *from Section 4.2.*

Proof: Assumption $(A1)$ is clear. Regarding assumption $(A2)$, the Beta-distribution $\text{Beta}(1,1)$ with parameters $\alpha = \beta = 1$ is the $\text{Uni}[0,1]$ distribution, and therefore $f_0 = \mathbf{1}_{[0,1]}(t)$. Analogously this is also the case for the Normal-Model.

For assumption $(A3)$, we analyze the cdf of the i-th p-value. In the Beta-Model, if $\theta_i \leq 0$, the p-value $p_i(\boldsymbol{X})$ is $\text{Beta}(1-\theta_i, 1)$-distributed with cdf $F_{\theta_i}(t) = t^{-\theta_i+1}\mathbf{1}_{[0,1]}(t) + \mathbf{1}_{(1,\infty)}(t)$, which is decreasing for decreasing θ_i and each fixed t. If $\theta_i > 0$, the cdf of $p_i(\boldsymbol{X})$ is $F_{\theta_i}(t) = \left(1-(1-t)^{\theta_i+1}\right)\mathbf{1}_{[0,1]}(t) + \mathbf{1}_{(1,\infty)}(t)$, which is increasing in θ_i and each fixed t. Thus assumption $(A3)$ is fulfilled in the Beta-Model.

In the Normal-Model, we refer to Remark 4.1, where the test statistic $T_i(\boldsymbol{X})$ is normally distributed with expected value θ_i and (known) variance σ^2.

This concludes the proof of Lemma 4.1.

In our simulations below, we draw the true parameter value θ_i uniformly from intervals $\left[\theta_i^b, 0\right]$ and $\left(0, \theta_i^b\right]$ if $\theta_i^b \leq 0$ or $\theta_i^b > 0$, respectively. Similarly to Loughin (2004), we write $\theta_i^b = r\mu_i^b$, $r > 0$, $i = 1, \ldots, s$. Holding each μ_i^b constant, we can vary the potential "signal strength" of each p-value with r, i.e. with increasing r the i-th p-value p_i gets stochastically larger / more conservative under H_i (assuming $\mu_i^b \neq 0$) and stochastically smaller under K_i under θ_i^b.

Table 4.1 summarizes the different patterns $(\mu_1^b, \ldots, \mu_s^b)^T$ that we use in our simulations, cf. also Table 3 in Loughin (2004). We set the number of studies to $s = 6$. The patterns are first ordered in their amount of false null hypotheses, i.e. the amount of indices i with $\mu_i^b > 0$. Patterns with the same amount

Pattern	μ_1^b	μ_2^b	μ_3^b	μ_4^b	μ_5^b	μ_6^b	$\sum_i(\mu_i^b)^2$
1	0	0	0	0	1	5	26
2	0	0	0	0	3	3	18
3	0	0	0	1	1	4	18
4	0	0	0	2	2	2	12
5	0	0	1	1	1	3	12
6	0	0	1.5	1.5	1.5	1.5	9
7	0	0.5	0.5	0.5	0.5	4	17
8	0	1	1	1	1	2	8
9	0	1.2	1.2	1.2	1.2	1.2	7.2
10	0.2	0.2	0.2	0.2	0.2	5	25.2
11	0.5	0.5	0.5	0.5	2	2	9
12	0.5	0.5	1.25	1.25	1.25	1.25	6.75
13	1	1	1	1	1	1	6

Table 4.1: The evidence patterns with uniformly distributed p-values under nulls

of false null hypotheses are then ordered decreasingly in their order of dispersion $\sum_i(\mu_i^b)^2$. Furthermore, we denote by pattern jc the conservative version of pattern j, where we replace each $\mu_i^b = 0$ by $\mu_i^b = -2$. Patterns 10 – 13 have no conservative versions.

4.4.2 Calculation of Bayes factors

We calculate the marginal Bayes factors for the two approaches that utilize e-value combinations under the same models as in Section 4.4.1. For this, we need to make some assumptions about the prior distributions of the parameter values under the null hypotheses and under the alternatives.

We assume that it is known beforehand whether the marginal null hypotheses H_j, $j = 1, \ldots, s$, are simple (Patterns 1 – 13) or composite (Patterns 1c – 9c). In both cases we calculate the Bayes factors under the assumption that all parameter values $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_s)^T \in K_i$ are such that the i-th component θ_i is drawn uniformly from the interval $(0, 5r]$. Under simple null hypotheses the resulting Bayes factors are e-values, i.e. they have expected values not larger than one, cf. Vovk and Wang (2020). Under composite null hypotheses, we calculate the Bayes factors under the assumption that θ_i is uniformly distributed on $[-3r, 0]$ if $\boldsymbol{\theta} \in H_i$. The numbers 5 and -3 were chosen such that the true underlying parameter values θ_i drawn from any of the patterns in Table 4.1 are included in $(0, 5r]$ or $[-3r, 0]$.

In the latter case, the resulting Bayes factors are not valid e-values for the marginal null hypotheses, i.e. their expected value is larger than one for some parameters under the null. More specifically, the i-th Bayes factor has an increasing expected value under increasing $\theta_i \in [-3r, 0]$. Therefore, under all parameters $\boldsymbol{\theta} \in H_i$, it has its largest expected value when $\theta_i = 0$. See Appendix C for a proof of this. To create valid e-values we therefore divide the Bayes factors in Patterns 1c – 9c by this expected value. Note, that computing this constant requires no extra information beyond the information necessary for calculating the Bayes factors.

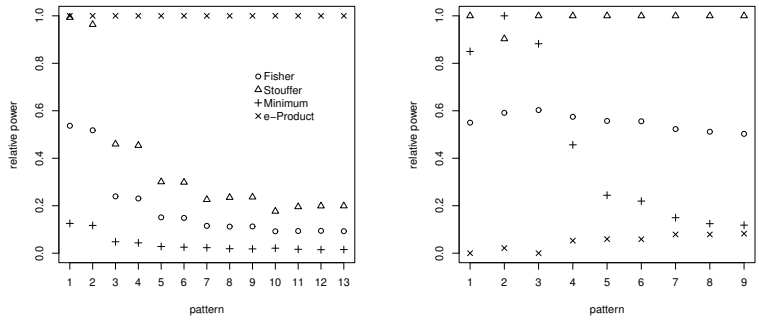
4.4.3 Power Simulations

The power of a p-value p under a parameter value $\boldsymbol{\theta}$ in the alternative given a significance level $\alpha \in (0, 1)$ is defined as $\mathbb{P}_{\boldsymbol{\theta}}(p \leq \alpha)$. Under various parameter settings, where H_s^γ is false, we approximate the relative power (relative to the best performing one in each setting, where we set the significance level to $\alpha = 0.05$) of each p-value combination via a Monte-Carlo simulation with 100,000 repetitions.

First, we look at different evidence structures in Table 4.1. For a pattern where H_s^γ is false, the evidence for its rejection can be focused in few false p-values or it can be more evenly spread between the false p-values, compare for example Pattern 3 versus Pattern 4. Furthermore, we want to investigate how the choice of γ affects the performance of the p-value combination functions for different types of evidence structures.

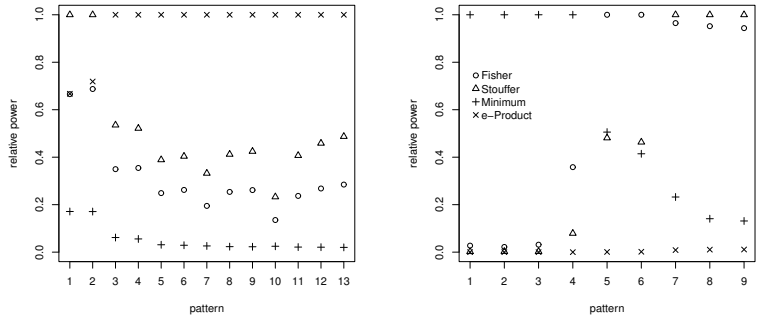
Evidence Structures

For the sake of clearness of the graphical displays, we decided to only display the simulation results for the Stouffer, Fisher and minimum p-value as well as the product of the e-values (called e-product). The harmonic mean and the arithmetic mean of the e-values (not displayed) performed badly to mediocrely throughout.



(a) Under uniform patterns (b) Under conservative patterns

Figure 4.1: Relative power of the combined p-values under the Beta-Model with $\gamma = 2$ and $r = 1$.



(a) Under uniform patterns (b) Under conservative patterns

Figure 4.2: Relative power of the combined p-values under the Beta-Model with $\gamma = 2$ and $r = 5$.

Figures 4.1 – 4.2 have been derived under the Beta-Model of generating the marginal p-values. We set $\gamma = 2$ and the significance strengths to $r = 1, 5$ in Figures 4.1 and 4.2, respectively. In Figures 4.3 – 4.4, we generated the marginal p-values according to the Normal-Model with $\sigma = 1/\sqrt{50}$. We set $\gamma = 2$ and the significance strengths in the figures to $r = 0.5\sigma, 1.5\sigma$, respectively. The distribution of the p-values under nulls is indicated above the graphics.

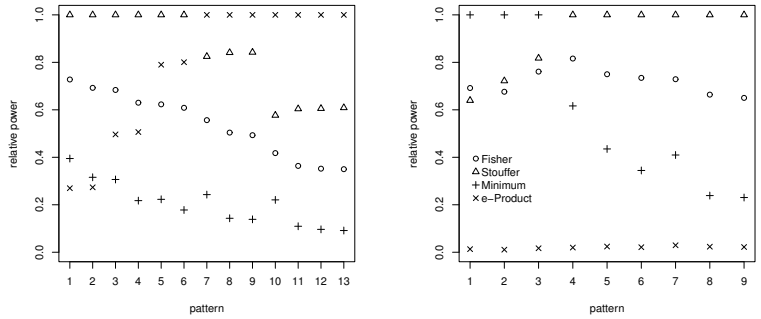
At first, we summarize the observations of the two figures with lower signal strength r, Figures 4.1 and 4.3. If the null p-values are Uni$[0, 1]$, the Stouffer p-value has the highest power to reject H_6^2 if the evidence is more focused (lower pattern number). If the evidence is more spread out the e-product has the highest power. The power of the Fisher p-value is slightly below that of the Stouffer p-value and the minimum p-value performs badly.

If the null p-values are conservative in Figures 4.1 and 4.3, the minimum p-value performs best if the evidence is focused. If the evidence is more spread out, both the Stouffer and, to a lesser extent, the Fisher p-value have the highest power. The e-product performs badly in this case.

In Figures 4.2 and 4.4 we used a higher signal strength r. If the null hypotheses are simple and the evidence is focused, the Stouffer p-value has the highest power under the Beta-Model. If the evidence is more spread out the e-product has the highest power. Under the Normal-Model the minimum p-value and the Fisher p-value are most powerful if the evidence is focused and the Stouffer p-value if the evidence is less focused. The e-product is most powerful if the evidence is spread out.

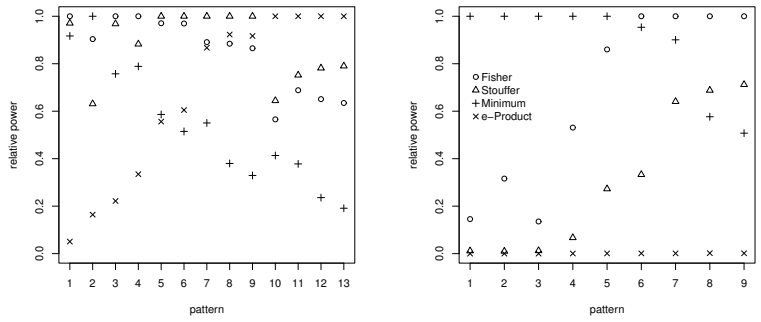
If the null p-values are conservative in Figures 4.2 and 4.4, the minimum p-value has the highest powers in the first patterns. The Stouffer has highest power under the Beta-Model and the Fisher p-value has highest power under the Normal-Model in the latter half of the patterns. The e-product performs badly throughout all the patterns.

To summarize, the e-product work best if we consider the non-conservative versions of the patterns, especially if the null p-values are uniformly distributed and the evidence is spread. If the null p-values are uniformly distributed and the evidence is focused the Stouffer p-value has the highest power. In the conservative patterns, the minimum p-value works best for lower pattern numbers. For the higher pattern numbers the Stouffer p-value works well if the signal strength is lower, and the Fisher p-value works well



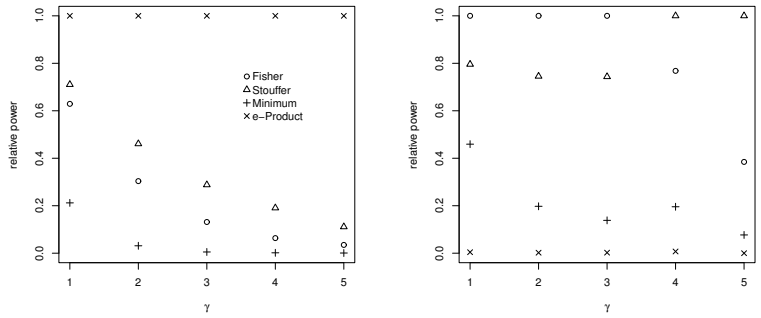
(a) Under uniform patterns (b) Under conservative patterns

Figure 4.3: Relative power of the combined p-values under the Normal Model with $\gamma = 2$ and $r = 0.5\sigma$.



(a) Under uniform patterns (b) Under conservative patterns

Figure 4.4: Relative power of the combined p-values under the Normal Model with $\gamma = 2$ and $r = 1.5\sigma$.



(a) Pattern 7 (b) Pattern 7c

Figure 4.5: Relative power of the combined p-values under the Beta-Model in Patterns 7 and 7c, with $\gamma = 1, \ldots, 5$ and $r = 10$.

if the signal strength is higher. The results between the two p-value generating models are mostly similar.

The parameter γ

In this section, we investigate the influence of γ on the relative performances of the p-value combination functions. More specifically, we chose Patterns 7 and 7c, in which five of the null hypotheses are false, and thus H_s^γ is false for $\gamma = 1, \ldots, 5$. Again, we look at the relative powers of the p-value combination functions, relative to the best performing combination function in each setting.

In Figure 4.5 we employed the Beta-Model. The p-value obtained from the e-product has the highest power throughout all values of $\gamma = 1, \ldots, 5$ in Pattern 7. With increasing γ, the power of the other combined p-values fall off faster than the power of the e-product. Under Pattern 7c the Fisher and the Stouffer p-value have the highest power, the former if $\gamma = 1, 2, 3$ and the latter if $\gamma = 4, 5$. The e-value approach, which is adjusted in this case, performs much worse.

In Figure 4.6 we used the Normal-Model. The p-values are close in power for $\gamma = 1$, their power is essentially 1 in absolute values. For $\gamma > 1$ in Pattern 7, the power of all the p-values fall relative to the power of the Stouffer p-value. The Fisher p-value performs relatively well and its power only falls off after $\gamma = 3$. In Pattern 7c, the Fisher p-value has the highest power if γ is between 2 and 4. For $\gamma = 5$, the minimum p-value has the highest power.

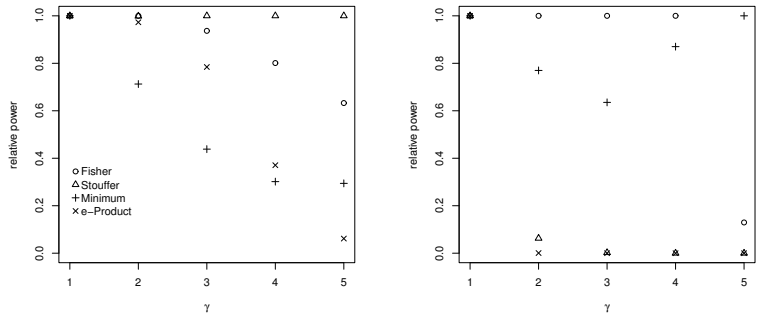
While the results under the Beta-Model suggest the superiority of the approach using e-values in Pattern 7, the results under the Normal-Model are more diverse. In both models, the Fisher p-value has higher power than the Stouffer p-value if the null p-values are conservative, and vice versa if the null p-values are uniformly distributed. Furthermore, the minimum p-value works (relatively) well if γ is large, especially if γ is the true number of false null hypotheses, which is five in Patterns 7 and 7c.

We illustrate this with a short example under the assumption that H_s^γ is false, that is, at least γ of the null hypotheses $H_1, \ldots, H_s$ are false. In terms of power, the worst case scenario for a monotonic combination function occurs if the p-values are as large as possible, which is the case if γ null hypotheses are false with corresponding p-values that are uniformly distributed, and $s - \gamma$ true null hypotheses with corresponding p-values that are almost surely 1. Note, that the distribution of false p-values is lower bounded by Uni$[0, 1]$ due to Assumption $(A3)$. Under this worst case scenario, the ordered, marginal p-values are $U_{(1)}, \ldots, U_{(\gamma)}, 1, \ldots, 1$, therefore the $s-\gamma+1$ largest p-values are $U_{(\gamma)}, 1, \ldots, 1$. Thus, testing for H_s^γ, the minimum p-value only directly evaluates $U_{(\gamma)}$, while averaging methods for instance by Fisher and Stouffer evaluate $U_{(\gamma)}, 1, \ldots, 1$, in this extreme case. Testing for $H_s^{\gamma-1}$ (which is also false if H_s^γ is false), the minimum p-value evaluates $U_{(\gamma-1)}$ whereas Fisher and Stouffer now consider $U_{(\gamma-1)}, U_{(\gamma)}, 1, \ldots, 1$. The ratio of non-one to one p-values increases with decreasing γ, which favors averaging methods more than the minimum p-value.

4.4.4 Null p-value simulations

In the previous simulations we only considered the case of false null hypotheses H_s^γ. In this section we investigate the behavior of the p-value combination functions under the null hypothesis H_s^γ.

Each of the presented p-value combination functions in Section 4.3 is valid for the null hypothesis H_s^γ, i.e. they are stochastically at least as large as Uni$[0, 1]$. Conservative p-values are common under composite null hypotheses, where the p-value is only calibrated with respect to the LFC parameter under



(a) Pattern 7 (b) Pattern 7c

Figure 4.6: Relative power of the combined p-values under the Normal-Model in Patterns 7 and 7c, with $\gamma = 1, \ldots, 5$ and $r = 3\sigma$.

the null. While still maintaining the type I error control, conservative null p-values can be problematic in several multiple testing setups that require uniformly distributed null p-values.

In simulations, we approximate the cdf $F_{\boldsymbol{\theta}}(\alpha)$ at point α of the combined p-values for H_s^γ under different parameter values $\boldsymbol{\theta}$ for which the null hypothesis H_s^γ hold. If α is the significance level, the value $F_{\boldsymbol{\theta}}(\alpha)$ for such $\boldsymbol{\theta}$ is the probability of a false rejection of H_s^γ. Since we only consider valid p-values, it holds $F_{\boldsymbol{\theta}}(\alpha) = \alpha$ if the p-value is Uni$[0, 1]$-distributed and $F_{\boldsymbol{\theta}}(\alpha) \leq \alpha$ if the p-value is conservative. It is of interest that $F_{\boldsymbol{\theta}}(\alpha)$ is as close to α as possible. One example is the problem of estimating the proportion π_0 of true null hypotheses in a multiple testing setup with the Schweder-Spjøtvoll estimator $\hat{\pi}_0(\alpha)$, cf. Schweder and Spjøtvoll (1982). The estimator $\hat{\pi}_0(\alpha)$ utilizes the marginal p-values, and its bias $\mathbb{E}_{\boldsymbol{\theta}}\left[\hat{\pi}_0(\alpha)\right] - \pi_0 \geq 0$ increases with decreasing $F_{\boldsymbol{\theta}}(\alpha)$ for any of the marginal p-values, cf. Hoang and Dickhaus (2022).

The choice $\lambda = \alpha$ in the Schweder-Spjøtvoll estimator $\hat{\pi}_0(\lambda)$ was proposed by Blanchard and Roquain (2009). For arbitrary parameter values $\lambda \in [0, 1)$ in the Schweder-Spjøtvoll estimator we have to look at the entire cdf $F_{\boldsymbol{\theta}}$. If the p-value is Uni$[0, 1]$-distributed, its cdf is a straight line between $(0, 0)$ and $(1, 1)$, and more conservative p-values have a cdf below that line. For select parameters values $\boldsymbol{\theta}$, we approximate the cdf of some of the p-values.

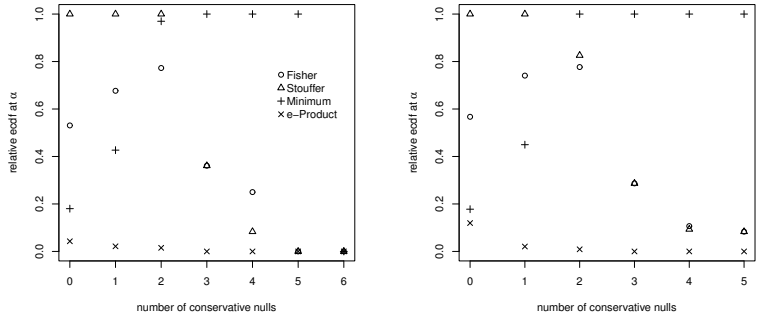
Figures 4.7 and 4.8 plot the empirical cumulative distribution functions (ecdfs) of the p-value combinations at point α, relative to the largest one in each setting, generated by a Monte-Carlo simulation with 100,000 repetitions, where we test for the rejection of H_6^2, i.e. that at least two null hypotheses are false. We use $(\mu_1^b, \ldots, \mu_6^b)^T = (0, \ldots, 0)^T$ on the left and $(\mu_1^b, \ldots, \mu_6^b)^T = (2, 0, \ldots, 0)^T$ on the right graphs. Furthermore, we replace 0 by -1 in $(\mu_1^b, \ldots, \mu_6^b)^T$ if the respective null is conservative. The number of times we do this is indicated on the horizontal axis.

In Figure 4.7 we used the Beta-Model and in Figure 4.8 we used the Normal-Model. The results are similar. The Stouffer p-value has the highest ecdf at α if the number of conservative nulls is low (below two or three), the minimum p-value has the highest ecdf at α if that number is higher. The Fisher p-value has mediocre performances and comes closer to the best p-values on the right graphs. The e-product has the lowest ecdf values at α.

Additionally, we display the ecdfs of the Stouffer, the Fisher and the minimum p-value in the cases of 1 and 4 conservative null p-values under the Beta-Model like in the right plot of Figure 4.7. The values for $(\mu_1^b, \ldots, \mu_6^b)^T$ are displayed above the plots, r is 5. First, we notice that the ecdfs are closer to the identical line on the left plot than they are on the right. On the left plot the ecdfs are close to each other, whereas on the right one the ecdf of the minimum p-value is noticeably closer to the identical line compared to the other two ecdfs. Another difference is that the ecdfs are not ordered consistently at each point $t \in [0, 1]$ on the left plot, which implies that the corresponding p-values are not stochastically ordered. On the right plot, however, the ecdf of the minimum p-value seems to be the largest at each point $t \in [0, 1]$, and therefore the minimum p-value is stochastically closest to Uni$[0, 1]$ in this more conservative setting.

4.5 Discussion

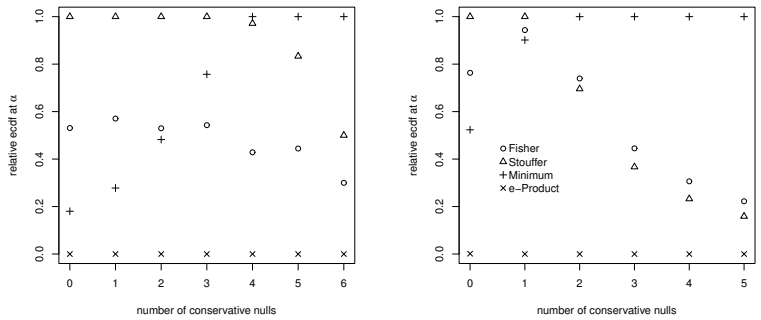
We compared a number of p-value combination functions for independent p-values and compared their power under the alternative hypothesis and their degree of conservativity under the null hypothesis



(a) The underlying pattern is $(\mu_1^b, \ldots, \mu_s^b)^T = (0, \ldots, 0)^T$.

(b) The underlying pattern is $(\mu_1^b, \ldots, \mu_s^b)^T = (2, 0, \ldots, 0)^T$.

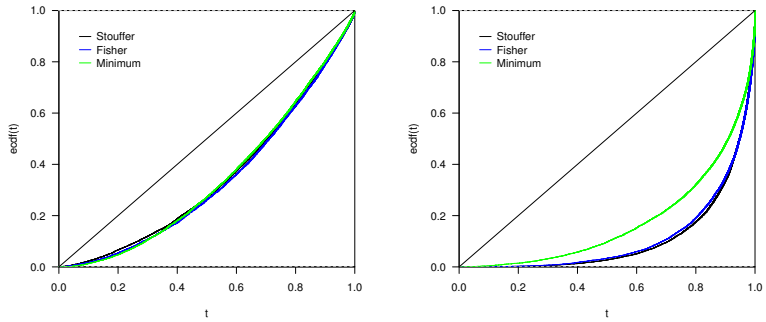
Figure 4.7: The p-values are generated under the Beta-Model. The graphs display the approximations of the cdf at α, relative to the (average) maximum estimation in the respective simulation, of the p-value combination functions testing the null hypothesis H_6^2, via Monte Carlo simulation with 100,000 repetitions. The signal strength r is 5. We replace 0 by -1 if conservative, the number of times we do this varies on the horizontal axis.



(a) The underlying pattern is $(\mu_1^b, \ldots, \mu_s^b)^T = (0, \ldots, 0)^T\sigma$.

(b) The underlying pattern is $(\mu_1^b, \ldots, \mu_s^b)^T = (2, 0, \ldots, 0)^T\sigma$.

Figure 4.8: The p-values are generated under the Normal-Model with standard deviation $\sigma = 1/\sqrt{50}$. The graphs display the approximations of the cdf at α, relative to the (average) maximum estimation in the respective simulation, of the p-value combination functions testing the null hypothesis H_6^2, via Monte Carlo simulation with 100,000 repetitions. The signal strength r is 1.5σ. We replace 0 by -1 if conservative, the number of times we do this varies on the horizontal axis.



(a) The underlying pattern is $(\mu_1^b, \ldots, \mu_6^b)^T = (2, -1, 0, \ldots, 0)^T$.

(b) The underlying pattern is $(\mu_1^b, \ldots, \mu_6^b)^T = (2, -1, \ldots, -1, 0)^T$.

Figure 4.9: The ecdfs from a Monte-Carlo simulation with 10,000 repetitions of the Fisher, Stouffer and minimum p-values in H_6^2. The marginal p-values have been generated under the Beta-Model with signal strength $r = 5$.

when testing the partial conjunction hypothesis H_s^γ. Among the p-value combination functions that we considered in this paper, we can roughly distinguish between two classes. One are the p-values that rely on a weighted average of the $s - \gamma + 1$ largest p-values and one that only evaluates the $(s - \gamma + 1)$-th largest p-value. They mainly differ in how they deal with spread out evidence versus focused evidence, and conservative versus uniformly distributed null p-values.

Among the considered p-value combination functions, the approaches that utilize Bayes factors as e-values, work best if the null hypotheses $H_1, \ldots, H_s$ are simple, and the Stouffer, the Fisher and the minimum p-values have the best results, if the null hypotheses are composite. The three latter p-values excel in different situations under the alternative and the null hypothesis. Under the alternative, the Stouffer p-value appears to have the highest power if the marginal p-values are uniformly distributed and the signal strength is low. The Fisher p-value works better than the other combination functions if the marginal null p-values are conservative and the evidence is spread out between several false null hypotheses. The minimum method works best if the evidence is focused on few false null hypotheses, especially when the null p-values are conservative. Under the null hypothesis H_s^γ, the Stouffer p-value is closest to uniformity if the marginal null p-values are Uni$[0, 1]$-distributed. However, if there is at least one conservative null p-value, the minimum method works better than the other combination methods. The Fisher p-value is stochastically closer to Uni$[0, 1]$ than the Stouffer p-value if some of the marginal null p-values are conservative. One major difference between the Stouffer and the Fisher p-value is that the latter emphasizes the smallest p-values more and is thus less affected by conservative p-values, cf. Owen (2009), which coincides with our results. While our selection of combination functions for H_s^γ is not exhaustive, the overall conclusions in this paper can be generalized to p-values that evaluate an average of all p-values versus p-values that place more weight on the smallest p-values.

Since this paper is concerned about replicability analyses, we limited our research to p-values from independent tests. It is interesting to see how the results differ, if the p-values are dependent. Methods that are designed for combining independent p-values, like the Fisher, Stouffer and minimum p-value, tend to fail to work properly if the marginal p-values are positively correlated, cf. Alves and Yu (2014). Introducing weights can be helpful in this matter. Furthermore, we only consider two models for the generation of the marginal p-values. The Beta-Model was also used in Loughin (2004) and our results are similar when $\gamma = 1$. The Normal-Model had similar results as well. In both models the p-values have non-decreasing densities under the null hypothesis and non-increasing under the alternative. Deviation from this assumption is not considered here and is an attractive topic for future research.

Chapter 5

Conditional combination test p-values

This chapter is a slightly modified version of Dickhaus et al. (2021) previously published on arXiv. Appendix D contains the original supplementary material of the paper.

Authors

Prof. Dr. Thorsten Dickhaus, Institute for Statistics, University of Bremen, Bremen, Germany

Prof. Dr. Ruth Heller, Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv-Yafo, Israel

Anh-Tuan Hoang, Institute for Statistics, University of Bremen, Bremen, Germany

Abstract The partial conjunction null hypothesis is tested in order to discover a signal that is present in multiple studies. We propose methods for multiple testing of partial conjunction null hypotheses which make use of conditional p-values based on combination test statistics. Specific examples comprise the Fisher combination function and the Stouffer combination function. The conditional validity of the corresponding p-values is proved for certain classes of one-parametric statistical models, including one-parameter natural exponential families. The standard approach of carrying out a multiple test procedure on the (unconditional) partial conjunction p-values can be extremely conservative. We suggest alleviating this conservativeness, by eliminating many of the conservative partial conjunction p-values prior to the application of a multiple test procedure. This leads to the following two step procedure: first, select the set with partial conjunction p-values below a selection threshold; second, within the selected set only, apply a family-wise error rate or false discovery rate controlling procedure on the conditional partial conjunction p-values. By means of computer simulations and real data analyses, we compare the proposed methodology with other recent approaches.

Summary In this chapter we deal with the multiple testing of conditional PC p-values.

We define as conditional p-values, p/τ, given $p \leq \tau$. The goal is to remove the conservative p-values before advancing with the conditional p-values in further analysis. In Theorem 5.1 we pose conditions for the validity of conditional PC p-values, for all choices of $\tau \in (0, 1]$, also called uniform validity. The result holds for example for uniformly valid base p-values and any combination p-value $P^{\gamma/s}$ that is monotonically increasing in each argument.

In Section 5.3, we consider the multiple testing of a set of conditional PC p-values. For simplicity we consider the same γ, s and τ for each testing problem; the algorithm is described in Algorithm 5.1. Instead of choosing τ beforehand we also consider the adaptive approach of choosing τ based on the to be discarded p-values as suggested by Zhao et al. (2019). FDR and asymptotic FDP guarantees for the BH procedure when using conditional p-values under different conditions are given in Propositions 5.1 – 5.3.

In Section 5.4, we give examples of models where PC p-values are uniformly valid. The second model has discretely distributed PC p-values, for which using the prior definition of conditional p-values Theorem 5.1 does not apply. In fact, one can show that discrete p-values with finite support can never be uniformly valid. A more general definition is provided in Definition 5.1 and a similar result to Theorem 5.1 is given in Lemma 5.1.

Simulation results are presented in Section 5.5. At level $\alpha = 0.05$, we apply adaFilter (Wang et al. (2021)), or the BH procedure and the adaptive BH procedure with the Schweder-Spjøtvoll estimator

as plug-in estimator using either unconditional p-values, the conditional p-values with pre-chosen τ, or the conditional p-values with different adaptive approaches for choosing τ. Using the conditional p-values increases the number of correct rejections in each setting, both with a pre-chosen τ and adaptive τ. Furthermore, in settings where for many endpoints $H^{\gamma/s}$ is true but $H^{(\gamma-1)/s}$ is false, adaFilter tends to work better than our approach. Power increases using the adaptive BH procedure compared to the non-adaptive BH procedure are only noticeable if π_0 is not too close to 1.

A replicability analysis on meta-analysis data taken from Franke et al. (2010) in Section 5.6 confirms that using the conditional PC p-values in the BH procedure consistently over all choices of τ leads to a higher number of rejections over the use of unconditional p-values, see Figure 5.2. Compared with the adaFilter approach, however, the choice of τ is more important, especially if $\gamma = 5$, which suggests that, in the given data, many endpoints that can be rejected at $\gamma = 5$ can also be rejected at $\gamma = 4$.

In Appendix D, we present additional simulations for example on the cdfs of the PC p-values before and after conditioning, on the dependence of PC p-values if endpoints within studies are dependent. Moreover, we complement the simulations in Section 5 by considering further combination functions, stronger dependence among the PC p-values, and randomized p-values in the estimation of π_0 for the adaptive BH procedure. Furthermore, we provide theoretical results concerning the asymptotic (FDP) FDR control in the BH procedure for pre-chosen parameters τ, or adaptive parameters $\hat{\tau}$. Finally, we also provide an alternative adaptive approach of choosing τ, we discuss conditioning before combining the base p-values for the PC null hypotheses, and we provide additional information on the conditions for the combination functions for Theorem 5.1.

Declaration of individual contributions Individual contributions are not as clearly distinguishable in this chapter, as the authors met up regularly to discuss and contribute to every part of the chapter. Nevertheless, I will attempt to give an account of my contributions: I was the main author of Section 5.2, not including Remark 5.1. The proof of Theorem 5.1 was proposed by Prof. Dr. Ruth Heller, and I wrote the final version of it. Furthermore, I wrote the summary of approaches for choosing τ at the beginning of Section 5.3, Definition 5.1 and Lemma 5.1 in Section 5.4, the former being Prof. Dr. Ruth Heller's idea, and Appendices D.5, D.6 and D.7. Apart from these, I contributed to the simulation code for Section 5.6 (implementation of the adaptive τ approach).

5.1 Introduction

"Replicability is widely taken to ground the epistemic authority of science." (Romero (2019)). In fact, the replication of scientific results is essential for their acceptance by the scientific community. In order to assess whether a scientific result has indeed been replicated in an independent study, appropriate scientific methods are needed. During recent years, statistical methods have been developed in this context; see, e.g., Bogomolov and Heller (2013), Heller et al. (2014), Heller and Yekutieli (2014), Bogomolov and Heller (2018), Wang and Owen (2019), Hung and Fithian (2020), and Hoang and Dickhaus (2022). In the aforementioned articles, it has been proposed to formalize the replication of a certain finding as a statistical hypothesis which can be tested on the basis of a data sample by employing an appropriate statistical test procedure. Especially in the context of modern high-throughput technologies, the simultaneous testing of many (say $m >> 1$) non-replicability null hypotheses (corresponding to m different features or endpoints, respectively) is of considerable interest, connecting the replicability assessment with the theory of multiple hypothesis testing.

The no-replicability null hypothesis is a specific instance of a partial conjunction (PC) null hypothesis (Benjamini and Heller, 2008), defined as follows: given s individual null hypotheses and $\gamma \in \{2, \ldots, s\}$, at most $\gamma - 1$ individual null hypotheses are false. Thus, a rejected PC null hypothesis leads to the conclusion that at least γ individual null hypotheses are false. In replicability analysis, this means that the result is replicated in at least γ studies. PC null hypotheses are also used for other types of inference: for example, in Benjamini and Heller (2008), in order to identify the brain voxels in which at least γ out of s covariates are associated with the outcome; in Sun and Wei (2011), in order to identify the genes expressed in at least γ out of s time points; in Karmakar and Small (2020), in order to discover the outcomes with at least γ out of s evidence factors; in Li et al. (2021), in order to identify the genetic segments containing distinct association with the phenotype in at least γ out of s diverse environments.

Regarded as a subset of the parameter space of a statistical model, a PC null hypothesis is a composite null hypothesis, such that standard methods for computing a corresponding p-value can be very conservative; cf. Dickhaus (2013). One approach to overcome this conservativity is to exploit concepts from selective inference (see, among others, Fithian et al. (2014) and Zhao et al. (2019)).

In the present work, we elaborate on selective inference methods for multiple testing of PC null hypotheses. A two-stage multiple test procedure is proposed which first selects promising features (or endpoints) by means of their (conventional) p-values arising from a combination test for replicability. The p-values of the so-selected features get adjusted for the selection event by conditioning on the latter. In the second stage of testing, the conditional p-values are used in a (standard) multiple test. A key mathematical result will be that the proposed conditional p-values are valid (in the sense of Equation (1) in Hoang and Dickhaus (2022)). This property will imply type I error control of the proposed two-stage multiple test. We will illustrate these theoretical points with prototypical statistical models and by analyzing simulated as well as real data.

Wang et al. (2021) provide another interesting approach to address the conservativeness of PC p-values. They use a clever filtration and the Bonferroni combining method for testing multiple PC hypotheses. We compare and contrast our method with theirs, as well as with the direct approach of applying a multiple test procedure on the PC p-values, while highlighting the potential advantages of our conditional approach.

The remainder of the work is structured as follows. In Sections 5.2 and 5.3 we present our proposed statistical methodology. Section 5.2 describes the proposed conditional p-value for one single PC null hypothesis and discusses its uniform validity. Section 5.3 explains how a family of such conditional p-values can be used for multiple testing of a family of PC null hypotheses. Section 5.4 is devoted to exemplary statistical models to which our considerations apply. In Section 5.5, computer simulations are presented, and Section 5.6 deals with an application in the context of genome-wide association studies. We conclude with a discussion in Section 5.7. A variety of additional results is presented in Appendix D.

5.2 Proposed conditional p-value and its validity

Given a set of null hypotheses $H_1, \ldots, H_s$ together with their corresponding stochastically independent random p-values $P_1, \ldots, P_s$, and given a constant $1 \leq \gamma \leq s$, we are interested in testing the partial conjunction null hypothesis

$$H^{\gamma/s} = \{\text{at most } \gamma - 1 \text{ null hypotheses are false}\}.$$

We assume that the individual p-values are valid, i.e., $\mathbb{P}(P_i \leq \alpha) \leq \alpha$ for all $\alpha \in (0,1)$ if H_i is true. We also assume that a combination p-value $P^{\gamma/s} \equiv P^{\gamma/s}(P_1, \ldots, P_s)$, that is increasing in $P_i, i = 1, \ldots, s$, is valid for $H^{\gamma/s}$ (see examples in Section 5.4). If $H^{\gamma/s}$ is true, the least favorable parameter configuration (LFC) is the one that maximizes the probability of rejection, $\mathbb{P}(P^{\gamma/s} \leq \alpha)$. The LFC typically leads to $\gamma-1$ p-values that are zero almost surely, and the remaining $s-\gamma+1$ p-values are uniform. We assume that $P^{\gamma/s}$ is uniform under any LFC $\pi(0, \ldots, 0, U_1, \ldots, U_{s-\gamma+1})$ in $H^{\gamma/s}$, where π is any permutation vector of s elements and $U_1, \ldots, U_{s-\gamma+1}$ are stochastically independent and identically Uni$[0,1]$-distributed, where Uni$[0,1]$ denotes the (continuous) uniform distribution on the interval $[0,1]$. Among those that fulfil these assumptions are for example the Fisher, Stouffer and Simes combination functions. For more information about PC null hypotheses and suitable combination p-values $P^{\gamma/s}$ we refer to Benjamini and Heller (2008) and Hoang and Dickhaus (2021a).

Our goal is to find conditions for $P^{\gamma/s}$ and $P_1, \ldots, P_s$ such that, for a $\tau \in (0,1]$, the conditional p-value $P^{\gamma/s}/\tau$ given $P^{\gamma/s} \leq \tau$ is also valid for $H^{\gamma/s}$. This turns out to be very useful when we consider a family of PC null hypotheses in the following sections, since we expect that far less than τ of the true PC null hypotheses will have PC p-values at most τ, due to their conservativeness. Thus by selecting all PC p-values at most τ, we greatly reduce the multiplicity problem. But in order to use the conditional PC p-values on the reduced family of selected PC hypotheses, we need to prove they are indeed valid (given selection). We define sufficient conditions for validity in Section 5.2.1 and prove that the conditional PC p-values are indeed valid in Section 5.2.2.

Throughout, we use the following notation. For any random variable X and event A, let $[X|A]$ denote any random variable whose distribution is the conditional distribution of X given A. Furthermore, we let $\leq_{\text{st}}$ denote the usual stochastic order, $\leq_{\text{rh}}$ the reversed hazard rate (rh) order, $\leq_{\text{hr}}$ the hazard rate (hr) order, and $\leq_{\text{lr}}$ the likelihood ratio (lr) order, cf. Sections $1.A. - 1.C.$ in Shaked and Shanthikumar (2007).

5.2.1 The set-up

We denote by $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_s)^\top$ the parameter (vector) of the statistical model under consideration for all data ascertained in all s studies. For each $i \in \{1, \ldots, s\}$, the (marginal) test problem is $H_i : \theta_i \geq \theta_i^*$ versus $K_i : \theta_i < \theta_i^*$, where $(\theta_i^* : 1 \leq i \leq s)$ are given constants. We assume that the distribution of

P_i is independent of $\theta_{i'}$, for all $i' \neq i$. For each $i \in \{1, \ldots, s\}$ we assume the p-value to be valid, i.e., $\mathbb{P}_{\theta_i}(P_i \leq \alpha) \leq \alpha$ for all $\alpha \in [0,1]$ and $\theta_i \geq \theta_i^*$. Furthermore, we require that it holds

$$\mathbb{P}_{\theta_i^*}(P_i \leq \alpha) = \alpha \tag{5.1}$$

for all α, i.e., that P_i is uniformly distributed under the parameter θ_i^*. This requirement can be removed, but it simplifies the exposition (it typically holds for continuous test statistics). We address the more general setting without requirement (5.1) in Section 5.4 following the example of multiple binomial tests. We need P_i to satisfy conditional validity as well, i.e., $\mathbb{P}_{\theta_i}(P_i/\tau \leq \alpha \mid P_i \leq \tau) \leq \alpha$ for all $\alpha \in [0,1]$ and $\theta_i \geq \theta_i^*$. This is equivalent to requiring that (Zhao et al., 2019)

$$\forall t \leq \tau,\ \theta_i \geq \theta_i^* : \mathbb{P}_{\theta_i}(P_i \leq t \mid P_i \leq \tau) \leq \frac{t}{\tau}. \tag{5.2}$$

Let $P_i^{(\theta_i)}$ denote a random variable with the same distribution as P_i, when the data is generated from the distribution indexed by the parameter θ_i. The usual validity can also be written as $\mathrm{Uni}[0,1] \leq_{\mathrm{st}} P_i^{(\theta_i)}$ for all $\theta_i \in H_i$, and the equation (5.2) holding for all $\tau \in [0,1]$, i.e. conditional validity for all τ, is equivalent to $\mathrm{Uni}[0,1] \leq_{\mathrm{rh}} P_i^{(\theta_i)}$ for all $\theta_i \in H_i$, which is stronger.

Thus, we assume that one of the following two conditions hold:

$(A1)$ It holds that $\mathrm{Uni}[0,1] \leq_{\mathrm{rh}} P_i^{(\theta_i)}$, for all $\theta_i \in H_i$, $i = 1, \ldots, s$.

$(A2)$ For all $\theta_i, \tilde{\theta}_i \in \mathbb{R}$ with $\theta_i \leq \tilde{\theta}_i$, it holds that $P_i^{(\theta_i)} \leq_{\mathrm{rh}} P_i^{(\tilde{\theta}_i)}$, $i = 1, \ldots, s$.

Condition $(A2)$ is stronger than $(A1)$. For testing multiple PC null hypotheses, assuming $(A1)$ is enough. We provide the result in Section 5.2.2 assuming $(A2)$ since it may be of independent interest, see Remark 5.1.

A known result is that the likelihood ratio order implies the reversed hazard rate order, cf. Theorem $1.C.1$. in Shaked and Shanthikumar (2007). Thus, p-values P_i that are isotone or antitone transformations of test statistics that are likelihood ratio ordered in the parameter θ_i fulfil condition $(A2)$. The latter property is also frequently referred to as monotone likelihood ratio (MLR) of the considered test statistics. For example any one-parametric, linear exponential family fulfils this MLR property with respect to its sufficient statistic, cf. Karlin and Rubin (1956b).

5.2.2 Main result

The conditional p-value $P^{\gamma/s}/\tau$ given $P^{\gamma/s} \leq \tau$ is valid for $H^{\gamma/s}$ if and only if

$$\frac{\mathbb{P}_{\boldsymbol{\theta}}\left(P^{\gamma/s} \leq t\right)}{\mathbb{P}_{\boldsymbol{\theta}}\left(P^{\gamma/s} \leq \tau\right)} \leq \frac{t}{\tau}, \tag{5.3}$$

for all $t \in [0, \tau]$, and $\boldsymbol{\theta} \in H^{\gamma/s}$, see for example Zhao et al. (2019).

The condition in (5.3) is equivalent to $[U \mid U < \tau] \leq_{\mathrm{st}} \left[P^{\gamma/s} \mid P^{\gamma/s} < \tau\right]^{(\boldsymbol{\theta})}$, where U is a $\mathrm{Uni}[0,1]$-distributed random variable. Therefore, if $U \leq_{\mathrm{rh}} (P^{\gamma/s})^{(\boldsymbol{\theta})}$, we get (5.3) for all τ, and therefore validity of the conditional p-value for all τ (cf. Shaked and Shanthikumar 2007, Section $1.B.6$). We call $P^{\gamma/s}$ *uniformly valid*, if its conditional p-value is valid for all τ.

Theorem 5.1 (Uniform conditional validity)**.**

(i) If $(A1)$ *holds, then* $\mathrm{Uni}[0,1] \leq_{\mathrm{rh}} (P^{\gamma/s})^{(\boldsymbol{\theta})}$*, for all* $\boldsymbol{\theta} \in H^{\gamma/s}$*.*

(ii) If $(A2)$ *holds, then* $(P^{\gamma/s})^{(\boldsymbol{\theta})} \leq_{\mathrm{rh}} (P^{\gamma/s})^{(\tilde{\boldsymbol{\theta}})}$*, for all* $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \in \mathbb{R}^s$*, where* $\boldsymbol{\theta} \leq \tilde{\boldsymbol{\theta}}$*, component-wise.*

In particular, $(A1)$ and $(A2)$ imply that $P^{\gamma/s}(P_1, \ldots, P_s)$ is a uniformly valid p-value for $H^{\gamma/s}$. Furthermore, $(A2)$ implies that the distributions $\left((P^{\gamma/s})^{(\boldsymbol{\theta})}\right)_{\boldsymbol{\theta}}$ are rh-ordered with respect to every component of $\boldsymbol{\theta}$.

Proof. We show the second statement first. To this end, let $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \in \mathbb{R}^s$, where only the first components $\theta_1, \tilde{\theta}_1$ of the two vectors are different, with $\theta_1 \leq \tilde{\theta}_1$. Under Condition $(A2)$, it holds $P_1^{(\boldsymbol{\theta})} \leq_{\mathrm{rh}} P_1^{(\tilde{\boldsymbol{\theta}})}$. Since $p_1 \mapsto P^{\gamma/s}(p_1, p_2, \ldots, p_s)$ is non-decreasing for all given values $p_2, \ldots, p_s$, it follows

$$P^{\gamma/s}(P_1, p_2, \ldots, p_s)^{(\boldsymbol{\theta})} \leq_{\mathrm{rh}} P^{\gamma/s}(P_1, p_2, \ldots, p_s)^{(\tilde{\boldsymbol{\theta}})} \tag{5.4}$$

from Theorem 1.B.43 in Shaked and Shanthikumar (2007) for all given $p_2, \ldots, p_s$. Since we assumed that P_1 and P_2 are stochastically independent, $[P^{\gamma/s}(P_1, p_2, \ldots, p_s) \mid P^{\gamma/s}(P_1, p_2, \ldots, p_s) \leq \tau]$ and $\big[[P^{\gamma/s}(P_1, P_2, p_3 \ldots, p_s) \mid P^{\gamma/s}(P_1, P_2, p_3, \ldots, p_s) \leq \tau] \mid P_2 = p_2\big]$ have the same distribution, for any given $p_2, \ldots, p_s$, and any τ. Therefore,

$$\mathbb{P}_{\boldsymbol{\theta}}(P^{\gamma/s}(P_1, P_2, p_3 \ldots, p_s) \leq t \mid P^{\gamma/s}(P_1, P_2, p_3, \ldots, p_s) \leq \tau) = \int F_{\boldsymbol{\theta}|p_2,\ldots,p_s,\tau}(t) d\mathbb{P}_{\boldsymbol{\theta}}^{P_2}(p_2), \tag{5.5}$$

where $F_{\boldsymbol{\theta}|p_2,\ldots,p_s,\tau}(t) = \mathbb{P}_{\boldsymbol{\theta}}(P^{\gamma/s}(P_1, p_2, \ldots, p_s) \leq t \mid P^{\gamma/s}(P_1, p_2, \ldots, p_s) \leq \tau)$ and $\mathbb{P}_{\boldsymbol{\theta}}^{P_2}$ is the distribution of P_2 under $\boldsymbol{\theta}$. With analogous notations for $\tilde{\boldsymbol{\theta}}$, we have that $\mathbb{P}_{\boldsymbol{\theta}}^{P_2} = \mathbb{P}_{\tilde{\boldsymbol{\theta}}}^{P_2}$, because $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}$ do not differ in their second components and P_2 does not depend on the other components $\theta_{i'}, i' \neq 2$.

For any fixed τ, it holds $F_{\boldsymbol{\theta}|p_2,\ldots,p_s,\tau}(t) \geq F_{\tilde{\boldsymbol{\theta}}|p_2,\ldots,p_s,\tau}(t)$ from (5.4), so that (5.5) and $\mathbb{P}_{\boldsymbol{\theta}}^{P_2} = \mathbb{P}_{\tilde{\boldsymbol{\theta}}}^{P_2}$ imply

$$\begin{aligned} & \mathbb{P}_{\boldsymbol{\theta}}(P^{\gamma/s}(P_1, P_2, p_3 \ldots, p_s) \leq t \mid P^{\gamma/s}(P_1, P_2, p_3, \ldots, p_s) \leq \tau) \\ \geq\ & \mathbb{P}_{\tilde{\boldsymbol{\theta}}}(P^{\gamma/s}(P_1, P_2, p_3 \ldots, p_s) \leq t \mid P^{\gamma/s}(P_1, P_2, p_3, \ldots, p_s) \leq \tau) \end{aligned}$$

for all $p_3, \ldots, p_s$ and any τ. But this means

$$P^{\gamma/s}(P_1, P_2, p_3 \ldots, p_s)^{(\boldsymbol{\theta})} \leq_{\mathrm{rh}} P^{\gamma/s}(P_1, P_2, p_3 \ldots, p_s)^{(\tilde{\boldsymbol{\theta}})}.$$

Analogously, since P_1, P_2, P_3 are jointly stochastically independent,

$$[P^{\gamma/s}(P_1, P_2, p_3, p_4 \ldots, p_s) \mid P^{\gamma/s}(P_1, P_2, p_3, p_4 \ldots, p_s) \leq \tau]$$

and

$$\Big[[P^{\gamma/s}(P_1, P_2, P_3, p_4 \ldots, p_s) \mid P^{\gamma/s}(P_1, P_2, P_3, p_4 \ldots, p_s) \leq \tau] \mid P_3 = p_3\Big]$$

have the same distribution. We can (iteratively) argue as before, and at the end, we have

$$P^{\gamma/s}(P_1, \ldots, P_s)^{(\boldsymbol{\theta})} \leq_{\mathrm{rh}} P^{\gamma/s}(P_1, \ldots, P_s)^{(\tilde{\boldsymbol{\theta}})}.$$

The proof for any $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \in \mathbb{R}^s$, that only differ in their i-th components (instead of their first ones) is analogous. Since the relation $\leq_{\mathrm{rh}}$ is transitive, we obtain the statement (ii) of the theorem. Namely, in the case of arbitrary parameter vectors $\boldsymbol{\theta} \leq \tilde{\boldsymbol{\theta}}$, we can apply the above reasoning successively to each coordinate i in which $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\theta}}$ differ.

For establishing statement (i) of the theorem, we assume $(A1)$ instead of $(A2)$. We assumed that, under an LFC of $P^{\gamma/s}$ under $H^{\gamma/s}$,

$$P^{\gamma/s}(\pi(0, \ldots, 0, U_1, \ldots, U_{s-\gamma+1}))$$

is uniformly distributed, where π is any permutation vector of s elements. For any $\boldsymbol{\theta} \in H^{\gamma/s}$, there are at least $s - \gamma + 1$ components θ_i with $\theta_i \in H_i$, and thus for at least $s - \gamma + 1$ indices i, it holds $\mathrm{Uni}[0,1] \leq_{\mathrm{rh}} P_i^{(\boldsymbol{\theta})}$. For the remaining $\gamma - 1$ marginal p-values it necessarily holds $0 \leq_{\mathrm{rh}} P_i^{(\boldsymbol{\theta})}$. Thus, there exists a permutation (vector) π such that $\pi(0, \ldots, 0, U_1, \ldots, U_{s-\gamma+1})$ is in the rh order component-wise smaller than $(P_1, \ldots, P_s)^{(\boldsymbol{\theta})}$. As in the proof of statement (ii), one can then show

$$P^{\gamma/s}(\pi(0, \ldots, 0, U_1, \ldots, U_{s-\gamma+1})) \leq_{\mathrm{rh}} P^{\gamma/s}(P_1, \ldots, P_s)^{(\boldsymbol{\theta})} \tag{5.6}$$

by sequentially replacing each component on the left-hand side by the (in rh order) larger p-value on the right-hand side. This means $\mathrm{Uni}[0,1] \leq_{\mathrm{rh}} P^{\gamma/s}(P_1, \ldots, P_s)^{(\boldsymbol{\theta})}$ and therefore uniform validity of $P^{\gamma/s}(P_1, \ldots, P_s)$ for $H^{\gamma/s}$. □

Remark 5.1. *Theorem 5.1 is closely related to the notion of the uniform conditional stochastic order (UCSO). This notion has been discussed, among others, by Whitt (1980, 1982), and by Lynch et al. (1987).*

Remark 5.2. *Instead of requiring that $P^{\gamma/s}$ is uniform under any LFC of the form $\pi(0, \ldots, 0, U_1, \ldots, U_{s-\gamma+1})$ in $H^{\gamma/s}$, where π is any permutation vector of s elements, the condition*

$$\mathrm{Uni}[0,1] \leq_{\mathrm{rh}} P^{\gamma/s}(\pi(0, \ldots, 0, U_1, \ldots, U_{s-\gamma+1}))$$

would have been sufficient for Theorem 5.1. For an example of a combination function $P^{\gamma/s}$ that does not fulfil the above condition, we refer to Appendix D.7.

5.3 Testing multiple partial conjunction null hypotheses

In this section, we assume that $m > 1$ (marginal) PC null hypotheses $H_1^{\gamma/s}, \ldots, H_m^{\gamma/s}$ are simultaneously under consideration under the scope of one and the same statistical model. In this, the assumption that γ and s are the same for all $1 \leq j \leq m$ is not necessary, and it is only made for notational convenience.

If all marginal test problems $H_j^{\gamma/s}$ versus $K_j^{\gamma/s}$, $1 \leq j \leq m$, are such that our Theorem 5.1 applies to each of them, m valid conditional (or randomized) p-values can readily be obtained. Many standard multiple test procedures like, for instance, the famous linear step-up test by Benjamini and Hochberg (1995) (the BH procedure) for control of the false discovery rate (FDR), have as their main assumption regarding the marginal p-values on which they operate that the latter marginal p-values are valid. (Oftentimes, some further conditions regarding the dependency structure among the marginal p-values have to be fulfilled, but transforming unconditional p-values into conditional p-values for each $j \in \{1, \ldots, m\}$ separately does not alter the latter dependency structure.)

On the basis of the aforementioned considerations, we propose for multiple testing of PC hypotheses the following workflow.

Algorithm 5.1.

(i) Compute for each $j \in \{1, \ldots, m\}$ an unconditional p-value $P_j^{\gamma/s}$ for testing $H_j^{\gamma/s}$ versus $K_j^{\gamma/s}$.

(ii) Choose cutoffs $\tau_1, \ldots, \tau_m$. For all those coordinates j for which $P_j^{\gamma/s} > \tau_j$, retain $H_j^{\gamma/s}$.

(iii) In the case that there exists at least one index j with $P_j^{\gamma/s} \leq \tau_j$: Transform, for each $j \in \{1, \ldots, m\}$ with $P_j^{\gamma/s} \leq \tau_j$ separately, the marginal unconditional p-value $P_j^{\gamma/s}$ into a marginal conditional p-value as described in Section 5.2. For continuous test statistics, the resulting conditional PC p-value is $P_j^{\gamma/s}/\tau_j$.

(iv) Utilize the conditional p-values obtained in step (iii) in a standard multiple test procedure $\boldsymbol{\varphi}$ (say), which merely requires validity of the marginal p-values on which it operates.

For additional power enhancement, we propose to employ a data-adaptive multiple test procedure in step (iv) which makes use of a pre-estimate of the proportion π_0 of true null hypotheses among the selected ones and incorporates the estimate in its decision rule. Adaptive procedures are especially useful when the fraction of null hypothesis is small. Even if in the family of m PC null hypotheses considered, the fraction of PC null hypotheses is close to one, following selection the fraction of true PC null hypotheses among the selected may be far smaller than one.

For the sake of simplicity, we focus on the case $\tau_1 = \cdots = \tau_m = \tau$ and provide a brief summary of approaches for selecting τ. Let S_τ be the set of those indices in $\{1, \ldots, m\}$ for which the unconditional p-values are not greater than τ.

1. **(No conditioning)** With $\tau = 1$ we use the unconditional p-values $P_1^{\gamma/s}, \ldots, P_m^{\gamma/s}$ in step (iv).

2. **(Pre-specified τ)** Choose a $\tau \in (0, 1)$ beforehand. Multiple testing in step (iv) is done on $\{p_j/\tau \mid j \in S_\tau\}$.

3. **(Adaptive choice of τ)** Proposition 4 in Zhao et al. (2019) provides an adaptive way, based on the p-values $p_1^{\gamma/s}, \ldots, p_m^{\gamma/s}$, of choosing τ that retains the validity of any valid *global* test, if $P_1^{\gamma/s}, \ldots, P_m^{\gamma/s}$ are jointly stochastically independent and uniformly valid. Given a sequence $0 \leq \tau_1 < \ldots \tau_K \leq 1$ of τ's between 0 and 1, we go from τ_k to τ_{k-1}, starting with τ_K, if the p-values $\{p_j^{\gamma/s} \mid j \notin S_{\tau_k}\}$ greater than τ_k fulfil certain conditions. The idea is to only use $\{p_j^{\gamma/s} \mid j \notin S_\tau\}$ to (adaptively) choose τ, and to use $\{p_j^{\gamma/s}/\tau \mid j \in S_\tau\}$ in steps (ii) – (iv). Zhao et al. (2019) give an example of a condition on the p-values $\{p_j^{\gamma/s} \mid j \notin S_{\tau_k}\}$ motivated by the Bonferroni test. The (Bonferroni-) adjusted p-values $p_j^{\gamma/s}|S_\tau|/\tau$, $j = 1, \ldots, m$, are minimized as functions of τ, if $|S_\tau|/\tau$ is minimized. For more details, we refer to Section 3.3 in Zhao et al. (2019). Since $|S_\tau| = m\hat{F}_m(\tau)$, where $\hat{F}_m$ is the empirical cumulative distribution function (ecdf) of the p-values $P_1^{\gamma/s}, \ldots, P_m^{\gamma/s}$, adaptive minimization of any function of $G(\tau) = H(|S_\tau|, \tau)$, where H is increasing in $|S_\tau|$ and decreasing in τ, as in Zhao et al. (2019), would retain Proposition 4 in Zhao et al. (2019). In Appendix D.5 we give another example of one such approach.

5.3.1 Theoretical properties of multiple test procedures targeted for FDR control

Applying the BH procedure on $\{p_j^{\gamma/s}, j \in S_\tau\}$, for a fixed pre-specified τ, guarantees control of the FDR for the entire family of PC hypotheses if all $m \times s$ p-values are stochastically independent, under the conditions of Theorem 5.1. This clearly follows since the BH procedure controls the FDR on the set of valid and independent p-values (Benjamini and Hochberg, 1995), and $\{p_j^{\gamma/s}, j \in S_\tau\}$ is such a set. Using the same reasoning, the adaptive BH procedure on $\{p_j^{\gamma/s}, j \in S_\tau\}$ also guarantees control of the FDR if we use Storey's estimator (Storey et al., 2004) for the fraction of null hypotheses in S_τ. In this section we provide theoretical guarantees for FDR control when the within-study p-values are dependent (which is the norm in high-dimensional studies), or when τ is chosen adaptively.

First, for the BH procedure using a pre-specified τ, we show that the the FDR is controlled at the nominal level if the original within study p-values are independent or positive regression dependent on the subset (PRDS) of true null hypotheses (Benjamini and Yekutieli, 2001).

Proposition 5.1. *Assume that the conditions of Theorem 5.1 are satisfied for each p-value. Moreover, assume that for each study, the p-values are PRDS on the subset of p-values corresponding to true null hypotheses, and the p-values across studies are independent. In addition, assume that the PRDS property is preserved for every subset of p-values under marginalization over the remaining p-values*[1]*. Then the FDR of the BH procedure at level α on $\{p_j^{\gamma/s}/\tau, j \in S_\tau\}$, for a fixed pre-specified τ, is at most α.*

Proof. According to Theorem 4.1 in Bogomolov (2021), the level α BH procedure on the PC p-values, $\{p_j^{\gamma/s}, j = 1, \ldots, m\}$, guarantees that the FDR on the PC null hypotheses is controlled at level α if the p-values are PRDS on the subset of p-values corresponding to true null hypotheses within each study, and the p-values across studies are independent. Since the dependence structure of any subset of p-values is unchanged, any subset taken is also PRDS within each study on the subset of p-values corresponding to true null hypotheses. In addition, Theorem 5.1 guarantees for the subset indexed by S_τ, that $\{p_j^{\gamma/s}/\tau, j \in S_\tau\}$ are valid p-values. Thus the level α BH procedure on $\{p_j^{\gamma/s}/\tau, j \in S_\tau\}$ controls the FDR for the family of null hypotheses $\{H_j^{\gamma/s}, j \in S_\tau\}$, as well as unconditionally for $\{H_j^{\gamma/s}, j = 1, \ldots, m\}$. □

The following proposition is a slight generalization of Proposition 4 of Zhao et al. (2019) and deals with the case of an adaptively chosen τ in the context of Algorithm 5.1.

Proposition 5.2. *Let $\boldsymbol{\theta} \in \Theta$ be arbitrary, but fixed. Assume that $(P_j^{\gamma/s} : 1 \leq j \leq m)$ are jointly stochastically independent. Assume that the multiple test $\boldsymbol{\varphi}$ employed in step (iv) of Algorithm 5.1 is such, that*

$$\mathbb{E}_{\boldsymbol{\theta}}\left[g(V_m, R_m)\right] \leq \alpha \tag{5.7}$$

holds true for any fixed value of $\tau \in (0, 1]$, where V_m is the (random) number of type I errors of $\boldsymbol{\varphi}$, R_m the (random) total number of rejections of $\boldsymbol{\varphi}$, g some measurable function taking values in $[0, 1]$, and $\alpha \in (0, 1)$ some given constant.

Now, let $\mathcal{F}_x = \sigma\left(\{P_j^{\gamma/s} : P_j^{\gamma/s} \geq x\}\right)$ for $x \in [0, 1]$, and assume that $\tilde{\tau}$ is a backward stopping time in the sense that the event $\{\tilde{\tau} \geq x\}$ is $\mathcal{F}_x$-measurable for any $x \in [0, 1]$. Then, (5.7) remains true if the fixed value of τ is replaced by the random value of $\tilde{\tau}$.

Proof. Proposition 4 of Zhao et al. (2019) yields the assertion for the special case of $g(V_m, R_m) = \mathbb{I}\{V_m > 0\}$ and for $\boldsymbol{\theta}$ in the global null hypothesis. (Actually, the test $\boldsymbol{\varphi}$ considered by Zhao et al. (2019) is just a single test for the global null hypothesis H_0 (say), such that $\mathbb{P}_{\boldsymbol{\theta}}(V_m > 0)$ reduces to the type I error probability of that single test under $\boldsymbol{\theta} \in H_0$.) However, as already indicated by Zhao et al. (2019) in their Section 3.4, the proof of their Proposition 4, which is presented in their Appendix A.3, does neither make use of the specific form of the function g nor of the fact that $\boldsymbol{\theta} \in H_0$ is assumed. Therefore, their proof applies to general (measurable and integrable) functions g and to arbitrary parameter values $\boldsymbol{\theta} \in \Theta$. □

The next proposition provides an asymptotic FDR control guarantee for the following greedy choice of τ: the value which leads to the largest number of rejections in step (iv) of Algorithm 5.1 when using the level α BH procedure. This value cannot be too small (since in this case too many false PC null hypotheses are not selected) nor too large (since in this case too many stochastically larger than uniform PC p-values are among the selected). The choice of this greedy τ can be written concisely as follows. Recall that the BH cutoff for any family of null hypotheses of size K is $x_{BH} = \max\{x : \frac{K \times x}{R(x)} \leq \alpha\}$, where

[1] This is satisfied, e.g., under the subset pivotality condition 2.1 in Westfall and Young (1993), or when the dependence structure of every subset of p-values is preserved under marginalization over the remaining p-values.

$R(x)$ is the number of p-values at most x, and all hypotheses with p-values at most x_{BH} are rejected. In step (iv) of Algorithm 5.1, the number of hypotheses is $K = |S_\tau|$, and the number of conditional PC p-values at most x equals the number of unconditional PC p-values at most τx, i.e., $|S_{\tau x}|$. More formally, for any $x \in [0,1]$,

$$R(x) = \sum_{i \in S_\tau} \mathbb{I}\left(\frac{p_i^{\gamma/s}}{\tau} \leq x\right) = \sum_{i \in S_\tau} \mathbb{I}\left(p_i^{\gamma/s} \leq \tau x\right) = \sum_{i=1}^{m} \mathbb{I}\left(p_i^{\gamma/s} \leq \tau x\right) = |S_{\tau x}|.$$

Denoting the BH threshold for a given τ by $\hat{x}(\tau)$:

$$\hat{x}(\tau) = \max\left\{x : \frac{|S_\tau| \times x}{|S_{\tau x}| \vee 1} \leq \alpha\right\},$$

we choose $\hat{\tau} = \arg\max_{\tau \in \{\tau_1, \ldots, \tau_K\}} S_{\tau \hat{x}(\tau)}$, where $0 < \tau_1 < \ldots < \tau_K \leq 1$ is a pre-defined finite set of K candidate values for the selection threshold. The false discovery proportion (FDP) of the rejections made with $(\hat{\tau}, \hat{x}(\hat{\tau}))$ is asymptotically almost surely (a.s.) at most α, assuming the following limits exist:

$$\forall x \in (0,1] : \lim_{m \to \infty} \frac{V_m(x)}{m_0} = G_0(x) \text{ and } \lim_{m \to \infty} \frac{R_m(x) - V_m(x)}{m - m_0} = G_1(x) \text{ a.s.}, \tag{5.8}$$

where m_0 is the number of true PC null hypotheses; $V_m(x)$ is the (random) number of true PC null hypotheses with p-values below x; $R_m(x)$ is the number of (random) p-values below x; and G_0 and G_1 are continuous functions such that

$$\forall x \in (0,1] : 0 < G_0(x) \leq x; \tag{5.9}$$

$$\lim_{m \to \infty} \frac{m_0}{m} = \pi_0 \text{ exists.} \tag{5.10}$$

Proposition 5.3. *Assume that condition (A1) or (A2) is satisfied for each p-value. Moreover, assume that the convergence assumptions of equations* (5.8)–(D.3) *hold for the PC p-values* $\{p_j^{\gamma/s}, j = 1, \ldots, m\}$. *Then, the FDP of the BH procedure at level* α *on* $\{p_j^{\gamma/s}/\hat{\tau}, j \in S_{\hat{\tau}}\}$ *is asymptotically at most* α.

See Appendix D.7 for the proof.

5.4 Illustrative example applications

In this section, we exemplify applications of our proposed methodology in the context of two widely used statistical model classes. Namely, we consider multiple Z-tests in Gaussian shift models and multiple binomial tests for success parameters of Bernoulli distributions.

Model 5.1 (Multiple Z-tests)**.** *For given sample sizes* $n_{i,j}$, *assume that we can observe* $\{X_k^{(i,j)} : i = 1, \ldots, s, j = 1, \ldots, m, k = 1, \ldots, n_{i,j}\}$, *and that for each study* i *and coordinate* j *the observables* $X_1^{(i,j)}, \ldots, X_{n_{i,j}}^{(i,j)}$ *are stochastically independent and identically normally distributed on* $\mathbb{R}$ *with expected value* $\theta_{i,j}$ *and a known variance, which may without loss of generality be assumed to be equal to one. For the study- and endpoint-specific test problem* $H_{i,j} = \{\theta_{i,j} \geq \theta_{i,j}^*\}$ *versus* $K_{i,j} = \{\theta_{i,j} < \theta_{i,j}^*\}$, *we consider the test statistic* $T_{i,j} = n_{i,j}^{-1/2} \sum_{k=1}^{n_{i,j}} \left(X_k^{(i,j)} - \theta_{i,j}^*\right)$, *which is normally distributed on* $\mathbb{R}$ *with expected value* $\sqrt{n_{i,j}} \cdot \left(\theta_{i,j} - \theta_{i,j}^*\right)$ *and variance one, for all* $i \in \{1, \ldots, s\}$ *and* $j \in \{1, \ldots, m\}$. *The corresponding p-variable is given by* $P_{i,j} = \Phi\left(T_{i,j}\right)$, *where* Φ *denotes the cumulative distribution function (cdf) of the standard normal distribution on* $\mathbb{R}$.

Remark 5.3. *If the variance of* $X_1^{(i,j)}$ *is unknown under Model 5.1, we can instead consider the t-distributed Studentized means as test statistics* $T_{i,j}$. *Then* (A2) *is still fulfilled, because the family of non-central t-distributions* $\left(T_{i,j}^{(\theta_{i,j})}\right)_{\theta_{i,j}}$ *is likelihood ratio ordered with respect to the non-centrality parameter, cf. Karlin and Rubin (1956a).*

Model 5.2 (Multiple binomial tests)**.** *For given sample sizes* $n_{i,j}$, *assume that we can observe* $\{X_k^{(i,j)} : i = 1, \ldots, s, j = 1, \ldots, m, k = 1, \ldots, n_{i,j}\}$, *and that for each study* i *and coordinate* j *the observables* $X_1^{(i,j)}, \ldots, X_{n_{i,j}}^{(i,j)}$ *are stochastically independent and identically Bernoulli-distributed indicator variables with success parameter* $\pi_{i,j}$, *which we assume to lie in the open interval* $(0,1)$, *to avoid pathologies. The*

corresponding canonical parameter of the resulting linear exponential family is given by $\theta_{i,j} = logit(\pi_{i,j}) = \log(\pi_{i,j}/(1-\pi_{i,j}))$. *The study- and endpoint-specific test problem* $H_{i,j} = \{\theta_{i,j} \geq \theta^*_{i,j}\}$ *versus* $K_{i,j} = \{\theta_{i,j} < \theta^*_{i,j}\}$ *is related to the original parameter values by noticing that* $\pi_{i,j} \mapsto \theta_{i,j} = logit(\pi_{i,j})$ *is a strictly increasing (thus one-to-one) transformation of* $\pi_{i,j} \in (0,1)$ *onto* $\mathbb{R}$. *We denote the value of the success parameter corresponding to the value* $\theta^*_{i,j}$ *of the canonical parameter by* $\pi^*_{i,j}$. *Moreover, we consider the test statistic* $T_{i,j} = \sum_{k=1}^{n_{i,j}} X_k^{(i,j)}$, *which is binomially distributed on* $\{0,\dots,n_{i,j}\}$ *with parameters* $n_{i,j}$ *and* $\pi_{i,j}$ *for all* $i \in \{1,\dots,s\}$ *and* $j \in \{1,\dots,m\}$. *The corresponding (random) p-value is given by* $P_{i,j} = F_{Bin(n_{i,j},\pi^*_{i,j})}(T_{i,j})$, *where* $F_{Bin(n_{i,j},\pi^*_{i,j})}$ *denotes the cdf of the binomial distribution with parameters* $n_{i,j}$ *and* $\pi^*_{i,j}$.

Under Model 5.2, the base p-values $P_{i,j}$ are discrete and do not fulfil our assumptions from Section 5.2. More particularly, one can show that discrete p-values with finite support can never be greater than Uni$[0,1]$ in rh order, and can therefore never be uniformly valid (unless the p-value is a.s. equal to one). Therefore, we introduce a more general definition of conditional p-values that coincide with the version from Section 5.2 if the p-value is continuously distributed. As before, we consider the discrete p-value $P_j^{\gamma/s}(P_{1,j},\dots,P_{s,j})$, $j = 1,\dots,m$.

Definition 5.1. *The conditional p-values we consider in Model 2, and more generally in discrete models, are* $\left(P_j^{\gamma/s}/\mathbb{P}_{\theta^*_j}(P_j^{\gamma/s} \leq \tau) \mid P_j^{\gamma/s} \leq \tau\right)_{j=1,\dots,m}$, *where* θ^*_j *is an LFC parameter for* $H_j^{\gamma/s}$.

Note that it holds $\mathbb{P}_{\theta^*_j}(P_j^{\gamma/s} \leq \tau) < \tau$, if τ is not in the support of $P_j^{\gamma/s}$, but it a.s. holds $P_j^{\gamma/s} \leq \mathbb{P}_{\theta^*_j}(P_j^{\gamma/s} \leq \tau)$ if $P_j^{\gamma/s} \leq \tau$. If $P_j^{\gamma/s}$ is uniformly distributed under LFCs, this definition coincides with the one from Section 5.2.

Lemma 5.1. *If it holds* $(P_j^{\gamma/s})^{(\boldsymbol{\theta}^*_j)} \leq_{\text{rh}} (P_j^{\gamma/s})^{(\boldsymbol{\theta}_j)}$ *for all* $\theta_j \in H_j^{\gamma/s}$, *then* $P_j^{\gamma/s}/\mathbb{P}_{\theta^*_j}(P_j^{\gamma/s} \leq \tau)$ *given* $P_j^{\gamma/s} \leq \tau$ *is valid for all* $\tau \in (0,1]$.

Proof. For $\theta_j \in H_j^{\gamma/s}$, we have to show that it holds

$$\mathbb{P}_{\theta_j}\left(P_j^{\gamma/s} \leq t\mathbb{P}_{\theta^*_j}(P_j^{\gamma/s} \leq \tau)\right) \leq t\mathbb{P}_{\theta_j}\left(P_j^{\gamma/s} \leq \mathbb{P}_{\theta^*_j}(P_j^{\gamma/s} \leq \tau)\right), \quad t \in [0,1].$$

We assumed that $\mathbb{P}_{\theta_j}(P_j^{\gamma/s} \leq t)/\mathbb{P}_{\theta^*_j}(P_j^{\gamma/s} \leq t)$ is increasing in t, thus, it holds

$$\frac{\mathbb{P}_{\theta_j}\left(P_j^{\gamma/s} \leq t\mathbb{P}_{\theta^*_j}(P_j^{\gamma/s} \leq \tau)\right)}{\mathbb{P}_{\theta^*_j}\left(P_j^{\gamma/s} \leq t\mathbb{P}_{\theta^*_j}(P_j^{\gamma/s} \leq \tau)\right)} \leq \frac{\mathbb{P}_{\theta_j}\left(P_j^{\gamma/s} \leq \mathbb{P}_{\theta^*_j}(P_j^{\gamma/s} \leq \tau)\right)}{\mathbb{P}_{\theta^*_j}\left(P_j^{\gamma/s} \leq \mathbb{P}_{\theta^*_j}(P_j^{\gamma/s} \leq \tau)\right)}.$$

Note, that the cdf of $P_j^{\gamma/s}$ under θ^*_j is the identity function on the set of the support points, and thus it holds $\mathbb{P}_{\theta^*_j}\left(P_j^{\gamma/s} \leq \mathbb{P}_{\theta^*_j}(P_j^{\gamma/s} \leq \tau)\right) = \mathbb{P}_{\theta^*_j}(P_j^{\gamma/s} \leq \tau)$. With $\mathbb{P}_{\theta^*_j}\left(P_j^{\gamma/s} \leq t\mathbb{P}_{\theta^*_j}(P_j^{\gamma/s} \leq \tau)\right) \leq t\mathbb{P}_{\theta^*_j}(P_j^{\gamma/s} \leq \tau)$, the proof is complete. □

That Model 2 fulfils the condition in Lemma 5.1 follows from the fact that the set of Bernoulli distributions is monotone likelihood ratio ordered with respect to the success parameter $\theta_{i,j}$, for all i,j. The rest is similar to Part (i) in Theorem 5.1, which does not require continuously distributed p-values.

Under Models 5.1 and 5.2, we may consider the following combination p-values.

(i) Fisher combination: For each $j \in \{1,\dots,m\}$, we let

$$P_j^{\gamma/s} = 1 - F_{\chi^2_{2(s-\gamma+1)}}\left(-2\sum_{i=\gamma}^{s}\log\left(P_{i:s,j}\right)\right),$$

where the notation $P_{i:s,j}$ refers to order statistics, and $F_{\chi^2_{2(s-\gamma+1)}}$ denotes the cdf of the chi-square distribution with $2(s-\gamma+1)$ degrees of freedom; cf. Section 21.1 in Fisher (1973).

(ii) Stouffer combination: For each $j \in \{1,\dots,m\}$, we let

$$P_j^{\gamma/s} = 1 - \Phi\left(\frac{1}{\sqrt{s-\gamma+1}}\sum_{i=\gamma}^{s}\Phi^{-1}\left(1-P_{i:s,j}\right)\right);$$

cf. Footnote 14 in Section V of Chapter 4 of Stouffer et al. (1949).

By making use of the aforementioned properties of the family of normal or Student's t distributions on $\mathbb{R}$ and of the family of Bernoulli distributions with a fixed number of trials and success parameter in $(0,1)$, respectively, it is straightforward to check that the assumptions of Theorem 5.1 are fulfilled under Models 5.1 and 5.2, both for the Fisher combination and for the Stouffer combination.

Remark 5.4. *Under Model 5.1, choosing the Stouffer combination method in a fixed coordinate j leads to the so-called "Gaussian selection differential" as the test statistic for testing $H_j^{\gamma/s}$ versus $K_j^{\gamma/s}$; cf., e. g., Nagaraja (1981) and Section 6.3 in Tong (1990). Hence, Theorem 5.1 implies the reverse hazard rate order for the selection differential in a Gaussian shift model with respect to shifts in the mean.*

5.5 Computer simulations

To assess and compare the performance of different replicability analysis approaches, we have carried out simulation studies. We have four specific aims: (1) to compare the power of our conditional approach with the unconditional approach, as well as with the "indirect" approach, adaFilter (Wang et al., 2021); (2) to study the effect of dependency within each study; (3) to examine how the choice of selection threshold matters, as well as the performance of an adaptive selection of this threshold for the conditional approach; (4) to examine the use of adaptive procedures in our conditional approach, which make use of an estimate of the fraction of null hypotheses on the selected hypotheses.

We consider the following data generation setting, which is partly inspired by the Crohn's disease example considered in Section 5.6, where we analyse $s = 8$ GWAS studies, and it is estimated that about 90% of the SNPs have no negative association with the disease in all the studies. We set the number of PC hypotheses based on $s = 8$ independent studies to $m = 20000$, $\{H_i^{\gamma/8}, i = 1, \ldots, 20000\}$, for $\gamma \in \{2, 3, 4, 5\}$. The fraction of true global null hypotheses (i.e., for which all s elementary hypotheses are true) is 0.9; the fraction of false null PC hypotheses is $\pi_1 \in \{0.002, 0.02\}$, where each non-null PC hypothesis has equal probability; each of the remaining true null PC hypotheses also has equal probability, adding up to an additional $1 - 0.9 - \pi_1$ true null PC hypotheses (but false global nulls). This type of data generation was considered in Wang et al. (2021) and made available in their function *GenPMat()* in their R package implementing AdaFilter available at `https://github.com/jingshuw/adaFilter`. We use their package both for generating data from the Gaussian shift model and for the adaFilter analysis (described in Section 5.1). The test statistics are generated from the Gaussian shift model, with non-null means sampled independently from $|\theta_{i,j}| \in \{3, 4\}$.

Within each study the test statistics are Gaussian, with a block correlation structure. The covariance within each block is symmetric with off diagonal entry $\rho = 0.9$ and diagonal entry one. The block size is 10 (see Appendix D.4 for results with block size 100).

We apply the replicability analysis detailed in (i)-(iv) of Section 5.3 for a selection threshold $\tau \in (0, 1)$. In step (i), the PC p-values are computed using the Fisher combination method (the use of Stouffer and Simes combination methods is evaluated in Appendix D.4).

In steps (ii)-(iii), τ_j is either fixed for all hypotheses at 0.1, or chosen adaptively as in Zhao et al. (2019). The adaptive approach of choosing τ by Zhao et al. (2019), as briefly described in Section 5.3, continues from τ_k to τ_{k-1} if we can reject $q > \hat{F}_m(\tau_k)/\tau_k$, where $m(\hat{F}_m(\tau_k + \omega) - \hat{F}_m(\tau_k)) \sim \text{Binomial}(m, q\omega)$ at a pre-specified significance level β, and $\omega \leq 1 - \tau_K$. Above, $\hat{F}_m$ denotes the empirical cumulative distribution function of the p-values $p_1^{\gamma/s}, \ldots, p_m^{\gamma/s}$. Note that this stopping condition only uses the p-values $\{p_j^{\gamma/s}/\tau \mid j \notin S_{\tau_k}\}$ larger than τ_k. Since their method assumes the test statistics are independent, we apply it on a subsample of independent test statistics (taking into consideration only statistics that are approximately 1.5 times the block size apart). Moreover, we examine several values of their parameter β: the larger the value of β, the smaller the estimate of the adaptive threshold. The window size is $w = 0.1$ as in Zhao et al. (2019).

In step (iv), the following two multiple testing procedures are applied to the conditional p-values for FDR control at the 0.05 level: the Benjamini-Hochberg (BH, Benjamini and Hochberg 1995) procedure at level 0.05; the adaptive BH procedure suggested in Storey et al. (2004) which applies the BH procedure at level $0.05/\hat{\pi}_0$, where $\hat{\pi}_0$ is (a slight variation on) the plug-in estimator of Schweder and Spjøtvoll (1982) for the fraction of true PC null hypotheses following selection (i.e., among all hypotheses with PC p-value at most τ).

We compare this analysis with the unconditional approach of applying the BH procedure at level 0.05 on the PC p-values. We also compare to the approach of Wang et al. (2021): first filter the potential hypotheses using the PC p-values based on the Bonferroni combination method, for the null PC hypotheses $\{H_i^{(\gamma-1)/8}, i = 1, \ldots, m\}$; then identify as discoveries the subset of filtered hypotheses for which the estimated FDP is at most 0.05, by using the PC p-values based on the Bonferroni combination method, for the null PC hypotheses of interest $\{H_i^{\gamma/8}, i = 1, \ldots, m\}$.

Table 5.1: In the symmetric block dependent setting, the average number of true discoveries for $\gamma = 2, 3, 4, 5$ for the following procedures at level 0.05: adaFilter by Wang et al. (2021), BH on PC p-values, BH and adaptive BH (denoted aBH) on conditional PC p-values using selection threshold $\tau = 0.1$, adaptive threshold at $\beta = 0.1$ and adaptive threshold at $\beta = 0.5$. Based on 5000 repetitions.

				conditional					
				τ=0.1		$\hat{\tau}$ with $\beta = 0.1$		$\hat{\tau}$ with $\beta = 0.5$	
π_1	γ	adaFilter	BH	BH	aBH	BH	aBH	BH	aBH
0.002	2	25.3	25.7	28.4	28.2	28.4	28.3	28.6	28.6
	3	17.0	14.5	20.1	20.1	19.8	19.6	20.1	20.3
	4	10.5	6.2	12.0	12.2	11.5	11.2	11.8	11.8
	5	7.0	2.1	6.4	6.6	5.9	5.7	6.2	6.1
0.02	2	338.5	302.4	319.9	327.9	320.5	321.8	320.4	325.5
	3	281.3	202.3	242.7	268.7	249.7	254.2	249.5	259.8
	4	217.8	110.1	160.1	195.9	176.1	182.7	176.1	189.2
	5	168.7	46.8	91.7	124.0	112.8	121.0	112.7	127.0

Tables 5.1 and 5.2 show the average number of true discoveries (our notion of power) and FDP for the novel procedures, using the selection threshold pre-specified as $\tau = 0.1$ or adaptively chosen as in Zhao et al. (2019) with $\beta = 0.1, 0.5$, versus competitors. Compared with the unconditional approach (which corresponds to $\tau = 1$) of applying BH on the PC p-values, we see that the power is greater with the novel approach, for every γ. This is expected, since the threshold for discovery for each selected hypothesis is lower using the conditional approach. An intuitive explanation is as follows. Figure 5.1 shows the distribution of the number of hypotheses selected for each γ. The number selected divided by the threshold for selection, $|S_\tau|/\tau$, is much smaller than $m = 20000$. Selection tends to eliminate many more PC null hypotheses than the expected number $\tau \times (1 - \pi_1) \times m$, because most p-values have a null distribution that is stochastically much larger than uniform, see Appendix D.7 for details. So had we considered doing the Bonferroni procedure following selection, each selected hypothesis would have been rejected if the PC p-value is at most $\alpha \times \tau/|S_\tau| > \alpha/m$. Therefore, if all the non-nulls (with enough power for detection) are selected, the conditional approach has more power than the unconditional approach. This reasoning carries over also to BH instead of Bonferroni, as Table 5.1 shows. The power advantage over the unconditional approach ranges from a power increase of (at least) 6% for $\gamma = 2$ to a power increase of more than 200% for $\gamma = 5$. From Table 5.2 it is clear that the conditional and unconditional procedures are below the nominal 0.05 level, and that the unconditional approach is the most conservative (i.e., with lowest FDR level). This conservatism is due to the fact that most PC p-values have a null distribution that is stochastically much larger than uniform, and the conditional p-values have a null distribution that is closer to uniform, but still conservative, see Appendix D.1 for details.

Compared with adaFilter, we see in Table 5.1 that the power is greater with the novel approach when $\pi_1 = 0.002$ but not when $\pi_1 = 0.02$. The setting with $\pi_1 = 0.02$ is more favorable to adaFilter (over $\pi_1 = 0.002$), since the ratio of false γ/s PC hypotheses to false $(\gamma - 1)/s$ PC hypotheses is larger (due to the fact that we keep the number of true global null hypotheses unchanged as π_1 increases), so the selection step of adaFilter is more efficient. In Table 5.2 we see that adaFilter controls the FDR but is less conservative than the other procedures.

Since the fraction of true PC hypotheses is close to one, the adaptive BH procedure on the PC p-values does not have a power advantage over BH so we do not compare this method. We see in Table 5.1 that when $\pi_1 = 0.002$, the adaptive approach also does not have a power advantage over BH on the conditional PC p-values, but when $\pi_1 = 0.02$, the adaptive approach makes more discoveries on average. This is so because the advantage of adaptivity increases as $|S_\tau|$ contains a smaller fraction of nulls. For example, for $\gamma = 2$, Figure 5.1 shows that hundreds of hypotheses are selected, so when $\pi_1 = 0.002$ the fraction of PC nulls is very close to one (since only $20000 \times 0.002 = 40$ PC hypotheses are non-null) but when $\pi_1 = 0.02$ the fraction of PC nulls can be far from one (since $20000 \times 0.02 = 400$ PC hypotheses are non-null and most of them are likely to be among the selected). The advantage of the adaptive procedure is greater as γ increases: for $\gamma = 2$ only few additional discoveries are made on average, but for $\gamma = 5$ it is 10% or more.

Finally, we see from Table 5.1 that all choices of τ provide similar power for $\pi_1 = 0.002$, as well as for $\pi_1 = 0.02$ when $\gamma = 2$. However, for $\pi_1 = 0.02$ and $\gamma > 2$, the BH procedure on the pre-specified $\tau = 0.1$ has lower power compared with BH on the adaptively selected τ. Moreover, the adaptive choice with $\beta = 0.5$ is at least as powerful as with $\beta = 0.1$ in all settings considered. Table 5.3 shows the average value of the adaptive selection threshold over 5000 repetitions for each simulations setting.

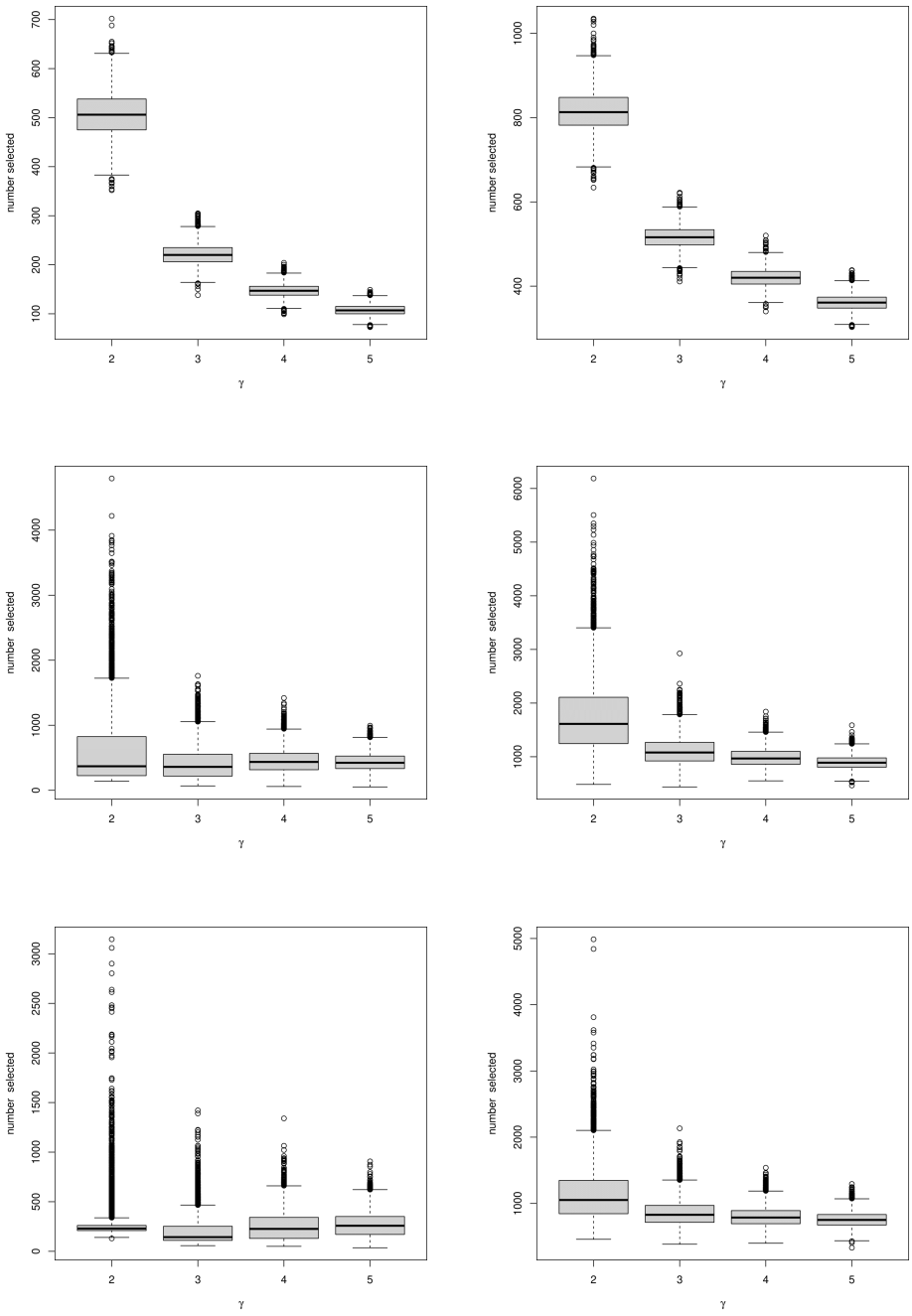


Figure 5.1: The distribution of the number of hypotheses selected, $|S_\tau|$, for every γ, in the simulation with $\pi_1 = 0.002$ (left column) and $\pi_1 = 0.02$ (right column), for τ selected as follows: pre-specified at $\tau = 0.1$ (top row), for τ selected as in Zhao et al. (2019) with $\beta = 0.1$ (middle row), and with $\beta = 0.5$ (bottom row). Based on 5000 repetitions.

Table 5.2: In the symmetric block dependent setting, the average FDP (estimated FDR) for $\gamma = 2, 3, 4, 5$ for the following procedures at level 0.05: adaFilter by Wang et al. (2021), BH on PC p-values, BH and adaptive BH (denoted aBH) on conditional PC p-values using selection threshold $\tau = 0.1$, adaptive threshold at $\beta = 0.1$ and adaptive threshold at $\beta = 0.5$. Based on 5000 repetitions.

				conditional					
				τ=0.1		$\hat{\tau}$ with $\beta = 0.1$		$\hat{\tau}$ with $\beta = 0.5$	
π_1	γ	adaFilter	BH	BH	aBH	BH	aBH	BH	aBH
0.002	2	0.046	0.004	0.016	0.014	0.015	0.015	0.017	0.018
	3	0.046	0.002	0.017	0.018	0.016	0.015	0.017	0.019
	4	0.045	0.001	0.015	0.017	0.013	0.011	0.014	0.015
	5	0.044	0.000	0.009	0.011	0.007	0.006	0.009	0.008
0.02	2	0.035	0.004	0.010	0.017	0.011	0.012	0.011	0.015
	3	0.032	0.002	0.008	0.019	0.010	0.012	0.010	0.014
	4	0.027	0.001	0.005	0.016	0.009	0.011	0.009	0.013
	5	0.022	0.000	0.003	0.009	0.007	0.008	0.007	0.010

Table 5.3: In the symmetric block dependent setting, the average adaptively selected τ using the method of Zhao et al. (2019) with $\beta = 0.1, 0.5$.

π_1	γ	$\hat{\tau}$ with $\beta = 0.1$	$\hat{\tau}$ with $\beta = 0.5$
0.002	2	0.11	0.06
	3	0.16	0.09
	4	0.25	0.16
	5	0.33	0.22
0.02	2	0.22	0.14
	3	0.27	0.21
	4	0.35	0.29
	5	0.42	0.36

5.6 An application to Crohn's disease genome-wide association studies

Identifying genomic regions with replicated association with Crohn's disease is important for better understanding the disease pathogenesis. Franke et al. (2010) carried out a meta-analysis of eight genome-wide association (GWA) studies of Crohn's disease in order to identify loci that are associated with the disease in at least one study. In this section, we illustrate our suggested replicability analysis in order to identify the SNPs associated with Crohn's disease in at least $\gamma \in \{2, 3, 4, 5\}$ out of the eight studies.

For every study, the data consists of z test statistics for the null hypothesis of no association between SNP and Crohn's disease, for m = 953,241 autosomal SNPs. For each SNP, replication of association across studies can be defined with or without regard to the direction of association. For illustration purposes, we consider here only one direction. So our starting point for the replicability analysis is a matrix of 953,241 $\times$ 8 left sided p-values.

For FDR control at the 0.05 level, we apply the BH procedure or the adaptive BH procedure on the conditional p-values. The PC p-values are computed using the Fisher combination method on the left sided p-values, and the threshold for selection is either fixed or estimated from the data using the method of Zhao et al. (2019). Since the SNPs are dependent, only one in (approximately) 100 SNPs is used for estimation of the data adaptive threshold (the lag chosen was 100 since for the eight studies, the autocorrelation graph at lag 100 was very small). For comparison, we also applied adaFilter (Wang et al., 2021), BH on the PC p-values, and adaptive BH on the PC p-values, at the 0.05 level. (The analysis using the right sided p-values is omitted, since it provided qualitatively similar results in terms of the relative performance of the various methods.)

Figure 5.2 shows the number of rejections by each method for each γ. Our novel approach makes more discoveries than the unconditional approach of applying the BH or adaptive BH procedure on the PC p-values. Moreover, for a wide range of selection thresholds, the novel approach makes more discoveries than adaFilter. In particular, more discoveries are made using the adaptive thresholds: for $\gamma = 2$ (the minimal replicability requirement) more than twice as many SNPs are discovered. The advantage of the conditional approach over adaFilter for $\gamma < 5$ is due to the fact that a large fraction of the hypotheses rejected with $H^{(\gamma-1)/s}$ cannot be rejected with $H^{\gamma/s}$, so following the filtering stage adaFilter still faces a large multiplicity problem but with a less efficient test statistic (which is a single order statistic, thus

Table 5.4: In the Crohn's disease dataset, the threshold for selection, number selected, and estimated fraction of null PC hypotheses among the selected for the following three methods for threshold selection: pre-specified at 0.1, and adaptively selected τ using the method of Zhao et al. (2019) with $\beta = 0.1, 0.5$.

γ	threshold selection method	τ	$\lvert S_\tau \rvert$	estimated fraction of null PC hypotheses
2	pre-specified at $\tau = 0.1$	0.10	32810	1.03
	Zhao et al. (2019) with $\beta = 0.1$	0.15	52575	1.09
	Zhao et al. (2019) with $\beta = 0.5$	0.01	3936	0.66
3	pre-specified at $\tau = 0.1$	0.10	8771	1.09
	Zhao et al. (2019) with $\beta = 0.1$	0.05	3972	0.85
	Zhao et al. (2019) with $\beta = 0.5$	0.04	3240	0.82
4	pre-specified at $\tau = 0.1$	0.10	2767	1.03
	Zhao et al. (2019) with $\beta = 0.1$	0.07	1860	0.94
	Zhao et al. (2019) with $\beta = 0.5$	0.06	1594	0.91
5	pre-specified at $\tau = 0.1$	0.10	1006	0.89
	Zhao et al. (2019) with $\beta = 0.1$	0.14	1570	1.09
	Zhao et al. (2019) with $\beta = 0.5$	0.08	799	0.84

it does not pool the information across studies by summation).

In order for the conditional approach with adaptive BH to make more discoveries than with BH, the estimated fraction of null PC hypotheses has to be below one. This occurs only when the selection threshold is small enough. Table 5.4 shows the number of selected PC hypotheses and the estimated fraction of null PC hypotheses among the selected when the selection threshold is pre-specified at $\tau = 0.1$, as well as when it is adaptively chosen by the method in Zhao et al. (2019) for $\beta = 0.1, 0.5$.

5.7 Discussion

We present a powerful approach for testing multiple PC hypotheses: first select the promising candidates, which are the PC hypotheses with PC p-values at most a certain threshold τ; then apply a valid multiple testing procedure on the conditional PC p-values within the selected set only. Results from simulations and data analysis highlight the potential usefulness of our approach for the discovery of consistent signals across multiple studies.

For high dimensional studies, the test statistics within each study are typically dependent. Moreover, FDR may be preferred over FWER for controlling false positive findings. We expect our approach with FDR control to be highly robust to dependencies within each study, for the following reasons. First, we have a theoretical guarantee that for PRDS dependency within each study the finite sample FDR is controlled, and for local dependency within each study the asymptotic FDR is controlled. Second, a vast amount of empirical evidence suggests that the BH procedure controls the FDR at the nominal level for most dependencies occurring in practice, and this robustness carries over to our novel approach which applies the BH procedure on conditional PC p-values. Third, the dependency among PC p-values is less severe than within individual studies. As γ increases, the PC p-values are less dependent, since the overlap between the identity of studies combined to form the PC p-values is reduced (and the studies are independent), see Appendix D.7 for details. The competitor AdaFilter (Wang et al., 2021) does not provide a finite sample FDR guarantee for any type of dependence among the test statistics, but it does provide an asymptotic guarantee under assumptions of weak dependence.

The choice of the selection threshold τ can have a large effect on the power to detect false PC null hypotheses. In our numerical experiments, we show that for a wide range of τ values our conditional approach leads to greater power than the unconditional approach. We also show that for an asymptotic FDR guarantee, when the dependence is local within each study, it is possible to greedily choose the value of τ that leads to the greatest number of rejections; and for a finite sample FDR guarantee when the PC p-values are independent, it is possible to use the approach in Zhao et al. (2019). The data adaptive choice of Zhao et al. (2019) works quite well in our numerical experiments. From Proposition 5.2 it follows that other data adaptive methods that only use the information $|S_\tau|$ for choosing τ from a series of decreasing cutoffs are equally valid. In Appendix D.5 we consider another data adaptive choice, which aims to balance the benefit of having a reduced selection set with the harm in inflating each PC p-value by the factor $1/\tau$ (in order to have a valid conditional PC p-value). The two data adaptive choices do not dominate each other, but it may be that for specific applications one choice is better than the other.

When the fraction of PC null hypotheses among the selected is small, the adaptive BH provides more

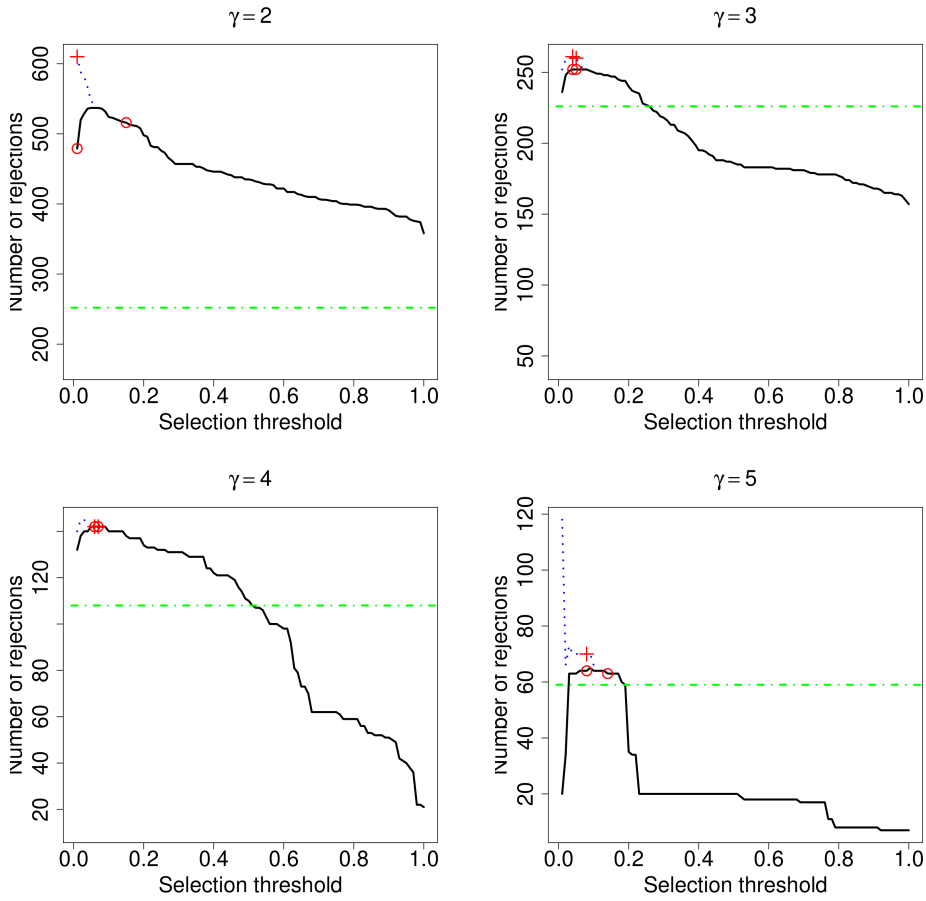


Figure 5.2: For every γ, the number of rejections at level 0.05, as a function of the selection threshold, using the conditional p-values for BH (solid black curve) and adaptive BH (dotted blue curve), as well as for adaFilter (dash-dotted green line). The number of rejections using the adaptive threshold with $\beta = 0.1$ and $\beta = 0.5$ for BH and for adaptive BH are in red circles and red pluses, respectively (with window size $w = 0.05$ in the algorithm of Zhao et al. 2019). The values at selection threshold '1.0' correspond to no selection, i.e., the BH and adaptive BH on the (unconditional) PC p-values. For adaptive BH, we only display results with estimated fraction of null PC hypotheses below one.

discoveries than the BH procedure on the conditional PC p-values. For simplicity, we used Storey's plug-in method for estimating the fraction of PC null hypotheses among the selected, as it is widely popular. However, other methods can be considered. In particular, methods of estimation of π_0 have been suggested in Hoang and Dickhaus (2022, 2021b) in settings where the p-values are conservative. Storey's plug-in method tends to overestimate the fraction of nulls in such settings. Hoang and Dickhaus (2022, 2021b) address this conservatism by suggesting to use randomized p-values. The mathematical conditions for validity of the randomized p-values (in the sense of Dickhaus 2013) are the same as for our proposed conditional p-values. Hence, the considerations of Hoang and Dickhaus (2022, 2021b) can directly be applied in our present context, too. In Section S4.3 we show the advantage of using the estimate of Hoang and Dickhaus (2022, 2021b) when the conditional PC p-values are conservative.

Our proposed workflow can be used with many p-value combining methods. Our results use the Fisher combining method, since it has excellent power properties for a wide range of signals (Benjamini and Heller, 2008; Hoang and Dickhaus, 2021a). We consider the Stouffer and Simes combining methods, which are also quite popular, in Section S4.1. However, in many applications, some of the null p-values may be very conservative, and then even the global null (i.e., when testing with $\gamma = 1$) p-values based on Fisher, Stouffer, and Simes combining methods are conservative, since they assume that the p-values to combine are uniformly distributed. This is the case when the p-values are discrete, or when the individual null hypotheses are composite. For example, in the Crohn's disease GWA studies, when testing for negative association, if the association is positive then the corresponding p-value will have a null distribution that is stochastically larger than uniform. Zhao et al. (2019) suggested for this purposes a test of the global null hypothesis that builds upon any combining function as follows: first, select for combining only the individual p-values that are at most a specified threshold; next, compute the conditional p-values on the selected set; finally, compute the global null conditional test p-value using the valid conditional p-values only. They demonstrated the power advantage of their conditional test over the unconditional global test (using, e.g., Fisher combining) when some p-values are conservative. The advantage is due to the fact that the unconditional test assumes that all p-values are uniformly distributed, but the conditional test only assumes this for the selected set. Thus, we expect our workflow in applications such as the Crohn's disease GWA studies to provide even more discoveries using a combining method adapted to conservative p-values as suggested by Zhao et al. (2019). In Section S6 we demonstrate that when the original p-values are conservative, using the combining method of Zhao et al. (2019) leads to conditional PC p-values that have a more uniform null distribution, compared with the Fisher combining method considered in this manuscript. Combining their method and ours yields even better results, especially if the proportion of true null hypotheses is high. We leave for future work the comprehensive examination of the benefits of using such state-of-the-art combining functions with our proposed methodology.

Chapter 6

Conclusion and outlook

This thesis deals with the statistical challenges of replicability analyses with applications in multiple testing. We often had to make simplifying assumptions, which leaves room for future research. Here, we mention some remaining issues and give some ideas for possible future projects.

Dealing with partial conjunction null hypotheses $H^{\gamma/s}$, conservative null p-values are one issue that is being dealt with in this work. As a solution, we either discard the PC p-value $p^{\gamma/s}$ or replace it with a uniformly distributed random variable U, if they are greater than a parameter $\tau \in [0, 1]$. Conditioning the PC p-values $p^{\gamma/s}(p_1, \ldots, p_s)$ like this, that is, after combining the p-values $p_1, \ldots, p_s$, reduces the conservativity inherent from the nature of PC null hypotheses but deals only indirectly with the conservativity that results from conservative base p-values $p_1, \ldots, p_s$. Conditioning before combining is briefly discussed in Appendix D.6. Doing so, it makes sense to retain $H^{\gamma/s}$, i.e. discard the endpoint, if fewer than γ p-values are selected. Alternatively, one could apply a rank truncation approach, for example, selecting the γ smallest p-values and test the conjunction null hypothesis $H^{\gamma/\gamma}$ on the (adjusted) selected p-values. If $\gamma = 2$, one can then use the approaches described in Bogomolov and Heller (2013) or Heller et al. (2014).

In the Schweder-Spjøtvoll estimator, the use of randomized p-values proved to be advantageous, especially if the non-randomized p-values are conservative if null. When using the same parameter c for the randomization of each p-value, the existence of a bias-minimizing parameter $c = c^\star$ has been proven, and an estimator has been provided in Chapter 2. However, the use of the latter in the randomized p-values does not guarantee their validity; a data-adaptive approach like described in Chapter 5 is needed. In the latter, motivated by Zhao et al. (2019), an approach that chooses τ based on the to be discarded p-values, $p_j^{\gamma/s} > \tau$, has been discussed. This was in the context of conditional p-values but also works for randomized p-values. Alternatively, one can estimate $c^\star$ with all but one p-value, and use $\hat{c}\left(p_1^{\gamma/s}, \ldots, p_{j-1}^{\gamma/s}, p_{j+1}^{\gamma/s}, \ldots, p_m^{\gamma/s}\right)$ to randomize $p_j^{\gamma/s}$. This might work if the p-values are independent, and m is sufficiently large.

Furthermore, our analysis of the use of randomized p-values in the Schweder-Spjøtvoll estimator $\hat{\pi}_0(\lambda)$ can be complemented by a more extensive analysis for the case of dependent LFC p-values. A perfunctory analysis was made in Chapter 3 in which we found that the use of randomized p-values can decrease the variance of $\hat{\pi}_0(\lambda)$ compared to the use positively dependent LFC p-values. Additionally, a more extended analysis of the impact λ has on the performance of $\hat{\pi}_0(\lambda)$ in connection with τ or c, i.e. how does λ change the bias-minimizing parameter $c^\star$, may be considered.

In this work, we solely focus on the case of using the same parameter $\tau_1 = \cdots = \tau_m = \tau$ for each p-value $p_1^{\gamma/s}, \ldots, p_m^{\gamma/s}$ when conditioning or randomizing. We showed in Chapter 3 that, under some assumptions, a smaller parameter τ_j would be better if H_j is true and vice versa if H_j is false, so a data-adaptive approach may be advantageous. In Appendix B we briefly discuss the use of random parameters R_j, however, we assumed that these are independent of the data. Sample splitting, albeit not very successful in previous literature, could be considered.

In our analysis of some commonly used combination functions in Chapter 4, we only considered independent p-values coming from the s studies, as is the scope of replicability analyses. An extension of the analysis to dependent p-values can prove informative; if the p-values are positively dependent, it may be easier to reject $H^{\gamma/s}$ compared to the independent case, and an adjustment to the statement of such may be needed. Furthermore, as mentioned in Chapter 4, we only considered p-values that have non-decreasing and non-increasing densities under the null and alternative, respectively, the former being a sufficient condition for uniform validity. More general p-value distributions may be considered in future research.

The choice of the parameter γ is an important question. Endpoints for which $H^{\gamma/s}$ is rightfully

rejected may still have anywhere between γ and s true discoveries. As proposed by Benjamini et al. (2009), one can test $H^{\gamma/s}$, $\gamma = 1, \ldots, s$, successively, obtaining a lower bound for the true number γ of false null hypotheses. Furthermore, when testing multiple PC null hypotheses, we have not considered different values of γ and s between different endpoints, as it is possible that some studies do not examine each endpoint.

We hope that our work has provided some insight into not only the challenges but also the potential that lies in the topic replicability analysis. There certainly are many more problems that did not make it into the thesis. Thank you for reading.

Acknowledgments

Firstly, I would like to thank my Ph.D. supervisor Prof. Dr. Thorsten Dickhaus for his constant support and advice throughout the past years. Among many things, he introduced me to the topic of multiple inference and randomized p-values, was always willing to help me with any problem, and helped me improve my scientific writing. My thanks also goes to Dr. André Neumann, who worked in the same office as me, before working from home became mandatory. He readily answered any questions I had, and later joined me in a project that unfortunately could not be finalized. My thanks goes to the working group, who were always helpful and friendly to me, easing me into the at first unfamiliar working environment of research. I would also like to thank Prof. Dr. Ruth Heller from the Tel Aviv University in Israel for always being helpful and encouraging. Without her, the paper in Chapter 5 would not have come into existence.

I gratefully acknowledge the financial support by the Deutsche Forschungsgemeinschaft (DFG), as well as the support by the University of Bremen, providing me with the office space and the tools necessary for my research. I would also like thank our secretary Martina Titze for helping me with all things on the administrative side of things. Finally, I would like to thank my family and friends for their presence and support throughout these last years.

Bibliography

Alves, G. and Yu, Y.-K. Accuracy evaluation of the unified p-value from combining correlated p-values. *PloS one*, 9(3):e91225, 2014.

Begley, C. G. and Ellis, L. M. Raise standards for preclinical cancer research. *Nature*, 483(7391):531–533, 2012.

Benjamini, Y. and Heller, R. Screening for partial conjunction hypotheses. *Biometrics*, 64(4):1215–1222, 2008. ISSN 0006-341X. doi: 10.1111/j.1541-0420.2007.00984.x. URL https://doi.org/10.1111/j.1541-0420.2007.00984.x.

Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, 57(1):289–300, 1995. ISSN 0035-9246. URL http://links.jstor.org/sici?sici=0035-9246(1995)57:1<289:CTFDRA>2.0.CO;2-E&origin=MSN.

Benjamini, Y. and Hochberg, Y. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Educ. Behav. Stat.*, 25(1):60–83, 2000.

Benjamini, Y. and Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, 29(4):1165–1188, 2001. ISSN 0090-5364. doi: 10.1214/aos/1013699998. URL https://doi.org/10.1214/aos/1013699998.

Benjamini, Y., Krieger, A. M., and Yekutieli, D. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3):491–507, 2006. ISSN 0006-3444. doi: 10.1093/biomet/93.3.491. URL https://doi.org/10.1093/biomet/93.3.491.

Benjamini, Y., Heller, R., and Yekutieli, D. Selective inference in complex research. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, 367(1906):4255–4271, 2009. ISSN 1364-503X. doi: 10.1098/rsta.2009.0127. URL https://doi.org/10.1098/rsta.2009.0127.

Birnbaum, A. Combining independent tests of significance. *J. Amer. Statist. Assoc.*, 49:559–574, 1954. ISSN 0162-1459. URL http://links.jstor.org/sici?sici=0162-1459(195409)49:267<559:CITOS>2.0.CO;2-3&origin=MSN.

Blanchard, G. and Roquain, E. Adaptive false discovery rate control under independence and dependence. *J. Mach. Learn. Res.*, 10:2837–2871, 2009. ISSN 1532-4435.

Bogomolov, M. Testing partial conjunction hypotheses under dependency, with applications to meta-analysis. *arXiv preprint arXiv:2105.09032*, 2021.

Bogomolov, M. and Heller, R. Discovering findings that replicate from a primary study of high dimension to a follow-up study. *J. Amer. Statist. Assoc.*, 108(504):1480–1492, 2013. ISSN 0162-1459. doi: 10.1080/01621459.2013.829002. URL https://doi.org/10.1080/01621459.2013.829002.

Bogomolov, M. and Heller, R. Assessing replicability of findings across two studies of multiple features. *Biometrika*, 105(3):505–516, 2018. ISSN 0006-3444. doi: 10.1093/biomet/asy029. URL https://doi.org/10.1093/biomet/asy029.

Chen, X. Uniformly consistently estimating the proportion of false null hypotheses via Lebesgue–Stieltjes integral equations. *J. Multivar. Anal.*, 173:724–744, 2019.

Chen, Z. Is the weighted z-test the best method for combining probabilities from independent tests? *J. Evol. Biol.*, 24(4):926–930, 2011.

Chen, Z. and Nadarajah, S. On the optimally weighted z-test for combining probabilities from independent studies. *Comput. Statist. Data Anal.*, 70:387–394, 2014. ISSN 0167-9473. doi: 10.1016/j.csda.2013.09.005. URL https://doi.org/10.1016/j.csda.2013.09.005.

Demetrescu, M., Hassler, U., and Tarcolea, A.-I. Combining significance of correlated statistics with application to panel data. *Oxf. Bull. Econ. Stat.*, 68(5):647–663, 2006.

Dickhaus, T. *Simultaneous Statistical Inference with Applications in the Life Sciences*. Berlin, Heidelberg: Springer, 2014.

Dickhaus, T. Randomized p-values for multiple testing of composite null hypotheses. *J. Statist. Plann. Inference*, 143(11):1968–1979, 2013. ISSN 0378-3758. doi: 10.1016/j.jspi.2013.06.011. URL `https://doi.org/10.1016/j.jspi.2013.06.011`.

Dickhaus, T. Simultaneous Bayesian analysis of contingency tables in genetic association studies. *Stat. Appl. Genet. Mol. Biol.*, 14(4):347–360, 2015. ISSN 2194-6302. doi: 10.1515/sagmb-2014-0052. URL `https://doi.org/10.1515/sagmb-2014-0052`.

Dickhaus, T., Straßburger, K., Schunk, D., Morcillo-Suarez, C., Illig, T., and Navarro, A. How to analyze many contingency tables simultaneously in genetic association studies. *Stat. Appl. Genet. Mol. Biol.*, 11(4):Art. 12, front matter+31, 2012. ISSN 2194-6302. doi: 10.1515/1544-6115.1776. URL `https://doi.org/10.1515/1544-6115.1776`.

Dickhaus, T., Stange, J., and Demirhan, H. On an extended interpretation of linkage disequilibrium in genetic case-control association studies. *Stat. Appl. Genet. Mol. Biol.*, 14(5):497–505, 2015. ISSN 2194-6302. doi: 10.1515/sagmb-2015-0024. URL `https://doi.org/10.1515/sagmb-2015-0024`.

Dickhaus, T., Heller, R., and Hoang, A.-T. Multiple testing of partial conjunction null hypotheses with conditional p-values based on combination test statistics. *arXiv preprint arXiv:2110.06692*, 2021.

Dudbridge, F. and Koeleman, B. P. Rank truncated product of P-values, with application to genomewide association scans. *Genet. Epidemiol.*, 25(4):360–366, 2003.

Edgington, E. S. An additive method for combining probability values from independent experiments. *J. Psychol.*, 80(2):351–363, 1972.

Efron, B. Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.*, 23(1):1–22, 2008. ISSN 0883-4237. doi: 10.1214/07-STS236. URL `https://doi.org/10.1214/07-STS236`.

Finner, H. and Gontscharuk, V. Controlling the familywise error rate with plug-in estimator for the proportion of true null hypotheses. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 71(5):1031–1048, 2009. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2009.00719.x. URL `https://doi.org/10.1111/j.1467-9868.2009.00719.x`.

Finner, H. and Strassburger, K. A note on P-values for two-sided tests. *Biom. J.*, 49(6):941–943, 2007. ISSN 0323-3847. doi: 10.1002/bimj.200710382. URL `https://doi.org/10.1002/bimj.200710382`.

Fisher, R. A. *Statistical methods for research workers*. Hafner Publishing Co., New York, 1973. Fourteenth edition—revised and enlarged.

Fithian, W., Sun, D., and Taylor, J. Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*, 2014.

Franke, A., McGovern, D. P., Barrett, J. C., Wang, K., Radford-Smith, G. L., Ahmad, T., Lees, C. W., Balschun, T., Lee, J., Roberts, R., et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.*, 42(12):1118–1125, 2010.

Friston, K. J., Penny, W. D., and Glaser, D. E. Conjunction revisited. *Neuroimage*, 25(3):661–667, 2005.

Futschik, A., Taus, T., and Zehetmayer, S. An omnibus test for the global null hypothesis. *Stat. Methods Med. Res.*, 28(8):2292–2304, 2019. ISSN 0962-2802. doi: 10.1177/0962280218768326. URL `https://doi.org/10.1177/0962280218768326`.

Ghosal, S. and Roy, A. Identifiability of the proportion of null hypotheses in skew-mixture models for the p-value distribution. *Electron. J. Stat.*, 5:329–341, 2011. doi: 10.1214/11-EJS609. URL `https://doi.org/10.1214/11-EJS609`.

Grünwald, P., de Heide, R., and Koolen, W. M. Safe testing. In *2020 Information Theory and Applications Workshop (ITA)*, pages 1–54. IEEE, 2020.

Habiger, J. D. Multiple test functions and adjusted p-values for test statistics with discrete distributions. *J. Statist. Plann. Inference*, 167:1–13, 2015. ISSN 0378-3758. doi: 10.1016/j.jspi.2015.06.003. URL https://doi.org/10.1016/j.jspi.2015.06.003.

Habiger, J. D. and Peña, E. A. Randomised P-values and nonparametric procedures in multiple testing. *J. Nonparametr. Stat.*, 23(3):583–604, 2011. ISSN 1048-5252. doi: 10.1080/10485252.2010.482154. URL https://doi.org/10.1080/10485252.2010.482154.

Hartung, J. A note on combining dependent tests of significance. *Biom. J.*, 41(7):849–855, 1999. ISSN 0323-3847. doi: 10.1002/(SICI)1521-4036(199911)41:7⟨849::AID-BIMJ849⟩3.0.CO;2-T. URL https://doi.org/10.1002/(SICI)1521-4036(199911)41:7<849::AID-BIMJ849>3.0.CO;2-T.

Heard, N. A. and Rubin-Delanchy, P. Choosing between methods of combining p-values. *Biometrika*, 105 (1):239–246, 2018. ISSN 0006-3444. doi: 10.1093/biomet/asx076. URL https://doi.org/10.1093/biomet/asx076.

Heesen, P. and Janssen, A. Inequalities for the false discovery rate (FDR) under dependence. *Electron. J. Stat.*, 9(1):679–716, 2015. doi: 10.1214/15-EJS1016. URL https://doi.org/10.1214/15-EJS1016.

Heesen, P. and Janssen, A. Dynamic adaptive multiple tests with finite sample FDR control. *J. Statist. Plann. Inference*, 168:38–51, 2016. ISSN 0378-3758. doi: 10.1016/j.jspi.2015.06.007. URL https://doi.org/10.1016/j.jspi.2015.06.007.

Heller, R., Bogomolov, M., and Benjamini, Y. Deciding whether follow-up studies have replicated findings in a preliminary large-scale omics study. *Proc. Natl. Acad. Sci. U.S.A.*, 111(46):16262–16267, 2014.

Heller, R. and Yekutieli, D. Replicability analysis for genome-wide association studies. *Ann. Appl. Stat.*, 8(1):481–498, 2014. ISSN 1932-6157. doi: 10.1214/13-AOAS697. URL https://doi.org/10.1214/13-AOAS697.

Hoang, A.-T. and Dickhaus, T. Combining independent p-values in replicability analysis: A comparative study. *J. Stat. Comput. Simul., accepted for publication*, 2021a.

Hoang, A.-T. and Dickhaus, T. On the usage of randomized p-values in the Schweder–Spjøtvoll estimator. *Ann. Inst. Stat. Math.*, pages 1–31, 2021b.

Hoang, A.-T. and Dickhaus, T. Randomized p-values for multiple testing and their application in replicability analysis. *Biom. J.*, 64(2):384–409, 2022.

Hochberg, Y. and Tamhane, A. C. *Multiple comparison procedures*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Inc., New York, 1987. ISBN 0-471-82222-1. doi: 10.1002/9780470316672. URL https://doi.org/10.1002/9780470316672.

Hung, K. and Fithian, W. Statistical methods for replicability assessment. *Ann. Appl. Stat.*, 14(3): 1063–1087, 2020.

Ioannidis, J. P. Why most published research findings are false. *PLoS Med.*, 2(8):e124, 2005.

Karlin, S. Decision theory for Pólya type distributions. Case of two actions, I. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. I*, pages 115–128. University of California Press, Berkeley and Los Angeles, 1956.

Karlin, S. and Rubin, H. Distributions possessing a monotone likelihood ratio. *J. Amer. Statist. Assoc.*, 51:637–643, 1956a. ISSN 0162-1459. URL http://links.jstor.org/sici?sici=0162-1459(195612)51:276<637:DPAMLR>2.0.CO;2-#&origin=MSN.

Karlin, S. and Rubin, H. The theory of decision procedures for distributions with monotone likelihood ratio. *Ann. Math. Statist.*, 27:272–299, 1956b. ISSN 0003-4851. doi: 10.1214/aoms/1177728259. URL https://doi.org/10.1214/aoms/1177728259.

Karmakar, B. and Small, D. S. Assessment of the extent of corroboration of an elaborate theory of a causal hypothesis using partial conjunctions of evidence factors. *The Annals of Statistics*, 48(6): 3283–3311, 2020.

Kim, S. C., Lee, S. J., Lee, W. J., Yum, Y. N., Kim, J. H., Sohn, S., Park, J. H., Lee, J., Lim, J., and Kwon, S. W. Stouffer's test in a large scale simultaneous hypothesis testing. *PLoS ONE*, 8(5):e63290, 2013.

Kocak, M. Meta-analysis of univariate P-values. *Comm. Statist. Simulation Comput.*, 46(2):1257–1265, 2017. ISSN 0361-0918. doi: 10.1080/03610918.2014.995818. URL https://doi.org/10.1080/03610918.2014.995818.

Kulinskaya, E., Morgenthaler, S., and Staudte, R. G. *Meta analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester, 2008. ISBN 978-0-470-02864-3. A guide to calibrating and combining statistical evidence.

Kumar Patra, R. and Sen, B. Estimation of a two-component mixture model with applications to multiple testing. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 78(4):869–893, 2016. ISSN 1369-7412. doi: 10.1111/rssb.12148. URL https://doi.org/10.1111/rssb.12148.

Lancaster, H. O. The combination of probabilities: an application of orthonormal functions. *Austral. J. Statist.*, 3:20–33, 1961. ISSN 0004-9581. doi: 10.1111/j.1467-842x.1961.tb00058.x. URL https://doi.org/10.1111/j.1467-842x.1961.tb00058.x.

Langaas, M., Lindqvist, B. H., and Ferkingstad, E. Estimating the proportion of true null hypotheses, with application to DNA microarray data. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(4):555–572, 2005. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2005.00515.x. URL https://doi.org/10.1111/j.1467-9868.2005.00515.x.

Lehmann, E. L. and Romano, J. P. *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, New York, third edition, 2005. ISBN 0-387-98864-5.

Li, S., Sesia, M., Romano, Y., Candès, E., and Sabatti, C. Searching for consistent associations with a multi-environment knockoff filter. *arXiv preprint arXiv:2106.04118*, 2021.

Liang, K. and Nettleton, D. Adaptive and dynamic adaptive procedures for false discovery rate control and estimation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 74(1):163–182, 2012. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2011.01001.x. URL https://doi.org/10.1111/j.1467-9868.2011.01001.x.

Loughin, T. M. A systematic comparison of methods for combining p-values from independent tests. *Comput. Statist. Data Anal.*, 47(3):467–485, 2004. ISSN 0167-9473. doi: 10.1016/j.csda.2003.11.020. URL https://doi.org/10.1016/j.csda.2003.11.020.

Lynch, J., Mimmack, G., and Proschan, F. Uniform stochastic orderings and total positivity. *Canad. J. Statist.*, 15(1):63–69, 1987. ISSN 0319-5724. doi: 10.2307/3314862. URL https://doi.org/10.2307/3314862.

MacDonald, P. W., Liang, K., and Janssen, A. Dynamic adaptive procedures that control the false discovery rate. *Electron. J. Stat.*, 13(2):3009–3024, 2019. doi: 10.1214/19-ejs1589. URL https://doi.org/10.1214/19-ejs1589.

Meinshausen, N. and Rice, J. Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Ann. Statist.*, 34(1):373–393, 2006. ISSN 0090-5364. doi: 10.1214/009053605000000741. URL https://doi.org/10.1214/009053605000000741.

Miłkowski, M., Hensel, W. M., and Hohol, M. Replicability or reproducibility? On the replication crisis in computational neuroscience and sharing only relevant detail. *J. Comput. Neurosci.*, 45(3):163–172, 2018.

Moonesinghe, R., Khoury, M. J., and Janssens, A. C. J. W. Most published research findings are false—but a little replication goes a long way. *PLoS Med.*, 4(2):e28, 2007.

Mosig, M. O., Lipkin, E., Khutoreskaya, G., Tchourzyna, E., Soller, M., and Friedmann, A. A whole genome scan for quantitative trait loci affecting milk protein percentage in israeli-holstein cattle, by means of selective milk dna pooling in a daughter design, using an adjusted false discovery rate criterion. *Genetics*, 157(4):1683–1698, 2001.

Nagaraja, H. N. Some finite sample results for the selection differential. *Ann. Inst. Statist. Math.*, 33(3):437–448, 1981. ISSN 0020-3157. doi: 10.1007/BF02480954. URL https://doi.org/10.1007/BF02480954.

Nettleton, D., Hwang, J. G., Caldo, R. A., and Wise, R. P. Estimating the number of true null hypotheses from a histogram of p values. *J. Agric. Biol. Environ. Stat.*, 11(3):337–356, 2006.

Neumann, A., Bodnar, T., and Dickhaus, T. Estimating the proportion of true null hypotheses under dependency: a marginal bootstrap approach. *J. Statist. Plann. Inference*, 210:76–86, 2021. ISSN 0378-3758. doi: 10.1016/j.jspi.2020.04.011. URL https://doi.org/10.1016/j.jspi.2020.04.011.

Olkin, I. Statistical and theoretical considerations in meta-analysis. *J. Clin. Epidemiol.*, 48(1):133–146, 1995.

Osherovich, L. Hedging against academic risk. *Science-Business eXchange*, 4(15):416–416, 2011.

Owen, A. B. Karl Pearson's meta-analysis revisited. *Ann. Statist.*, 37(6B):3867–3892, 2009. ISSN 0090-5364. doi: 10.1214/09-AOS697. URL https://doi.org/10.1214/09-AOS697.

Pearson, E. S. The probability integral transformation for testing goodness of fit and combining independent tests of significance. *Biometrika*, 30(1/2):134–148, 1938.

Prinz, F., Schlange, T., and Asadullah, K. Believe it or not: how much can we rely on published data on potential drug targets? *Nat. Rev. Drug Discov.*, 10(9):712–712, 2011.

Romero, F. Philosophy of science and the replicability crisis. *Philos. Compass*, 14(11):e12633, 2019.

Rüschendorf, L. Random variables with maximum sums. *Adv. in Appl. Probab.*, 14(3):623–632, 1982. ISSN 0001-8678. doi: 10.2307/1426677. URL https://doi.org/10.2307/1426677.

Schweder, T. and Spjøtvoll, E. Plots of P-values to evaluate many tests simultaneously. *Biometrika*, 69: 493–502, 1982.

Shaffer, J. P. Multiple hypothesis testing. *Annu. Rev. Psychol.*, 46(1):561–584, 1995.

Shaked, M. and Shanthikumar, J. G. *Stochastic orders*. Springer Series in Statistics. Springer, New York, 2007. ISBN 978-0-387-32915-4; 0-387-32915-3. doi: 10.1007/978-0-387-34675-5. URL https://doi.org/10.1007/978-0-387-34675-5.

Simes, R. J. An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3): 751–754, 1986. ISSN 0006-3444. doi: 10.1093/biomet/73.3.751. URL https://doi.org/10.1093/biomet/73.3.751.

Stange, J., Dickhaus, T., Navarro, A., and Schunk, D. Multiplicity- and dependency-adjusted p-values for control of the family-wise error rate. *Statist. Probab. Lett.*, 111:32–40, 2016. ISSN 0167-7152. doi: 10.1016/j.spl.2016.01.005. URL https://doi.org/10.1016/j.spl.2016.01.005.

Storey, J. D. A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 64(3): 479–498, 2002. ISSN 1369-7412. doi: 10.1111/1467-9868.00346. URL https://doi.org/10.1111/1467-9868.00346.

Storey, J. D. The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann. Statist.*, 31(6):2013–2035, 2003. ISSN 0090-5364. doi: 10.1214/aos/1074290335. URL https://doi.org/10.1214/aos/1074290335.

Storey, J. D. and Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA*, 100(16):9440–9445, 2003. ISSN 0027-8424. doi: 10.1073/pnas.1530509100. URL https://doi.org/10.1073/pnas.1530509100.

Storey, J. D., Taylor, J. E., and Siegmund, D. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 66(1):187–205, 2004. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2004.00439.x. URL https://doi.org/10.1111/j.1467-9868.2004.00439.x.

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., and Williams Jr, R. M. *The American soldier: Adjustment during army life.(Studies in social psychology in World War II), vol. 1*. Princeton Univ. Press, 1949.

Sun, W. and Wei, Z. Multiple testing for pattern identification, with applications to microarray time-course experiments. *J. Amer. Statist. Assoc.*, 106(493):73–88, 2011. ISSN 0162-1459. doi: 10.1198/jasa.2011.ap09587. URL https://doi.org/10.1198/jasa.2011.ap09587.

Taylor, J. and Tibshirani, R. J. Statistical learning and selective inference. *Proc. Natl. Acad. Sci. USA*, 112(25):7629–7634, 2015. ISSN 0027-8424. doi: 10.1073/pnas.1507583112. URL https://doi.org/10.1073/pnas.1507583112.

Tian, J. and Ramdas, A. ADDIS: an adaptive discarding algorithm for online FDR control with conservative nulls. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 9388–9396. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/9136-addis-an-adaptive-discarding-algorithm-for-online-fdr-control-with-conservative-nulls.pdf.

Tippett, L. H. C. The methods of statistics. page 395, 1952. 4th ed.

Tong, Y. L. *The multivariate normal distribution*. Springer Series in Statistics. Springer-Verlag, New York, 1990. ISBN 0-387-97062-2. doi: 10.1007/978-1-4613-9655-0. URL https://doi.org/10.1007/978-1-4613-9655-0.

van Zwet, W. R. and Oosterhoff, J. On the combination of independent test statistics. *Ann. Math. Statist.*, 38:659–680, 1967. ISSN 0003-4851. doi: 10.1214/aoms/1177698861. URL https://doi.org/10.1214/aoms/1177698861.

Vovk, V. and Wang, R. E-values: Calibration, combination, and applications. *Ann. Stat., acccepted for publication*, 2019.

Vovk, V. and Wang, R. Combining p-values via averaging. *Biometrika*, 107(4):791–808, 2020.

Vovk, V. and Wang, R. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021.

Wang, J., Gui, L., Su, W., Sabatti, C., and Owen, A. Detecting multiple replicating signals using adaptive filtering procedures. *Ann. Stat., accepted for publication.*, 2021.

Wang, J. and Owen, A. B. Admissibility in partial conjunction testing. *J. Amer. Statist. Assoc.*, 114 (525):158–168, 2019. ISSN 0162-1459. doi: 10.1080/01621459.2017.1385465. URL https://doi.org/10.1080/01621459.2017.1385465.

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

Whitlock, M. C. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *J. Evol. Biol.*, 18(5):1368–1373, 2005.

Whitt, W. Uniform conditional stochastic order. *J. Appl. Probab.*, 17(1):112–123, 1980. ISSN 0021-9002. doi: 10.2307/3212929. URL https://doi.org/10.2307/3212929.

Whitt, W. Multivariate monotone likelihood ratio and uniform conditional stochastic order. *J. Appl. Probab.*, 19(3):695–701, 1982. ISSN 0021-9002. doi: 10.1017/s0021900200037219. URL https://doi.org/10.1017/s0021900200037219.

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

Wilson, D. J. The harmonic mean p-value for combining dependent tests. *Proc. Natl. Acad. Sci. USA*, 116(4):1195–1200, 2019. ISSN 0027-8424. doi: 10.1073/pnas.1814092116. URL https://doi.org/10.1073/pnas.1814092116.

Won, S., Morris, N., Lu, Q., and Elston, R. C. Choosing an optimal method to combine P-values. *Stat. Med.*, 28(11):1537–1553, 2009. ISSN 0277-6715. doi: 10.1002/sim.3569. URL https://doi.org/10.1002/sim.3569.

Yekutieli, D. Bayesian tests for composite alternative hypotheses in cross-tabulated data. *TEST*, 24(2): 287–301, 2015. ISSN 1133-0686. doi: 10.1007/s11749-014-0407-1. URL https://doi.org/10.1007/s11749-014-0407-1.

Zaykin, D. V. Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis. *J. Evol. Biol.*, 24(8):1836–1841, 2011.

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

Zhao, Q., Small, D. S., and Su, W. Multiple testing when many p-values are uniformly conservative, with application to testing qualitative interaction in educational interventions. *J. Amer. Statist. Assoc.*, 114 (527):1291–1304, 2019. ISSN 0162-1459. doi: 10.1080/01621459.2018.1497499. URL https://doi.org/10.1080/01621459.2018.1497499.

Chapter A

Appendix for Chapter 2

A.1 Some concepts of stochastic ordering

We briefly introduce some concepts of stochastic ordering and notations. For some further results we refer to Appendix A.2.

A.1.1 Definition

Let X, Y be two random variables with cumulative distribution functions F, G, respectively.

(i) We say X *is smaller than* Y *in the usual stochastic order* or X is stochastically not larger than Y, denoted by $X \leq_{\text{st}} Y$, if and only if it holds $F(x) \geq G(x)$ for all $x \in (-\infty, \infty)$.

Intuitively, X is more likely than Y to take on small values.

(ii) We say X *is smaller than* Y *in the hazard rate order*, denoted by $X \leq_{\text{hr}} Y$, if and only if $(1 - G(t))/(1 - F(t))$ does not decrease in $t < \max\{u(X), u(Y)\}$, where $u(X), u(Y)$ denote the right endpoints of the supports of X, Y, respectively. We define $a/0 = \infty$, whenever $a > 0$.

Equivalently, if X and Y admit Lebesgue-density functions f, g, respectively, it holds $X \leq_{\text{hr}} Y$ if and only if $f(t)/(1-F(t)) \geq g(t)/(1-G(t))$ for all $t \in \mathbb{R}$, i.e. Y has a smaller hazard rate function.

(iii) If X, Y admit Lebesgue-density functions f, g, respectively, we say X *is smaller than* Y *in the likelihood ratio order*, denoted by $X \leq_{\text{lr}} Y$, if and only if $g(t)/f(t)$ is non-decreasing in t over the union of the supports of X and Y, where $a/0 = \infty$, whenever $a > 0$. Equivalently, it holds $X \leq_{\text{lr}} Y$ if and only if $f(y)g(x) \leq f(x)g(y)$, for all $x \leq y$

These three orders only depend on the distributions of X, Y, i.e. they only depend on F, G, f, g. Hence, we introduce the following notations.

A.1.2 Definition

Given a statistical model $\big(\Omega, \mathcal{F}, (\mathrm{P}_\vartheta)_{\vartheta \in \Theta}\big)$ and test statistics $T, S : \Omega \to \mathbb{R}$ with cumulative distribution functions F_ϑ, G_ϑ, respectively, and Lebesgue-density functions f_ϑ, g_ϑ, respectively, under $\vartheta \in \Theta$, we write $T^{(\vartheta_1)} \leq_{\text{st}} S^{(\vartheta_2)}$ for given parameters $\vartheta_1, \vartheta_2 \in \Theta$, if it holds $F_{\vartheta_1}(x) \geq G_{\vartheta_2}(x)$, for all x. Analogously, we denote $T^{(\vartheta_1)} \leq_{\text{hr}} S^{(\vartheta_2)}$, or $T^{(\vartheta_1)} \leq_{\text{lr}} S^{(\vartheta_2)}$, if $F_{\vartheta_1}, G_{\vartheta_2}, f_{\vartheta_1}, g_{\vartheta_2}$ satisfy the corresponding requirements for given parameters $\vartheta_1, \vartheta_2 \in \Theta$.

A.2 Some results regarding stochastic orders

We introduce some results regarding the hazard rate order. For a set of random variables $Z_1, \ldots, Z_n$, $n \geq 2$, we denote the order statistics of the first $m \leq n$ Z_i's by $Z_{(1:m)} \leq \cdots \leq Z_{(m:m)}$. For $m = n$ we usually write $Z_{(1)} \leq \cdots \leq Z_{(n)}$.

A.2.1 Theorem

Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$, be two sets of independent, not necessarily identically distributed, random variables.

(*i*) (Shaked and Shanthikumar, 2007, Theorem 1.B.28)
It holds $X_{(k:m)} \leq_{\text{hr}} X_{(k:m-1)}$ $(k = 1, \ldots, m-1)$.

(*ii*) (Shaked and Shanthikumar, 2007, Theorem 1.B.35)
If $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ all have the same support (a, b) for some $a < b$, and $X_i \leq_{\text{hr}} Y_j$ $(i = 1, \ldots, n;\ j = 1, \ldots, n)$, then $X_{(k:n)} \leq_{\text{hr}} Y_{(k:n)}$ $(k = 1, \ldots, n)$.

(*iii*) (Shaked and Shanthikumar, 2007, Theorem 1.B.2)
If $X \leq_{\text{hr}} Y$ and ψ is an increasing function, then $\psi(X) \leq_{\text{hr}} \psi(Y)$.

For proofs and further details, the reader may consult Chapter 1.*B*. and Chapter 1.*C*. in Shaked and Shanthikumar (2007).

Now, let $X_1, \ldots, X_n$ be independent random variables with support $(0, 1)$ and $U_1, \ldots, U_n$ be independent, uniformly distributed random variables on $[0, 1]$.

A.2.2 Lemma

For all fixed $n \geq 2$, $i \in \{1, \ldots, n\}$, if $X_k \leq_{\text{hr}} U_k$ holds for at least i indices $k \in \{1, \ldots, n\}$, then $X_{(i:n)} \leq_{\text{hr}} U_{(i:i)}$.

Proof. At first we consider the case $i = n$, that is, we assume $X_k \leq_{\text{hr}} U_k$ holds for all $k = 1, \ldots, n$. Then, we have $X_k \leq_{\text{hr}} U_l$ for all k, l, since the hazard rate order only depends on the distributions of X_k and U_l, and therefore $X_{(n:n)} \leq_{\text{hr}} U_{(n:n)}$ follows directly from Part 2 of Theorem A.2.1.

For $i = 1, \ldots, n-1$, we obtain from Part 1 of Theorem A.2.1, that $X_{(i:n)} \leq_{\text{hr}} X_{(i:n-1)} \leq_{\text{hr}} \cdots \leq_{\text{hr}} X_{(i:i)} \leq_{\text{hr}} U_{(i:i)}$, where the last inequality follows from the first part if $X_k \leq_{\text{hr}} U_k$ holds for $k = 1, \ldots, i$. Since $X_1, \ldots, X_n$ were assumed to have i such X_k, and prior calculations hold for any order of $X_1, \ldots, X_n$, we can assume $X_k \leq_{\text{hr}} U_k$ $(k = 1, \ldots, i)$, as desired. □

This lemma can be extended to any identically distributed $U_1, \ldots, U_k$ with support $(0, 1)$ or any support (a, b) shared with $X_1, \ldots, X_n$.

The following theorem is due to (Shaked and Shanthikumar, 2007, Theorem 1.C.2) and establishes a relationship between the three stochastic orders presented in Definition A.1.1.

A.2.3 Theorem

For two continuous random variables X, Y the likelihood ratio order $X \leq_{\text{lr}} Y$ implies the hazard rate order $X \leq_{\text{hr}} Y$. Both imply the stochastic order $X \leq_{\text{st}} Y$.

A.3 Proofs

A.3.1 Proof of Theorem 2.1

In order to show the first assertion, we notice that, due to assumption $(GA1)$, it holds $\{x \in \Omega :\ T_j(x) \in \Gamma_j(c_j)\} = \{x \in \Omega :\ \hat{\boldsymbol{\theta}}_j(x) \in K_j\}$. This implies

$$\mathrm{P}_{\vartheta_0}(\hat{\boldsymbol{\theta}}_j(X) \in K_j) = \mathrm{P}_{\vartheta_0}(T_j(X) \in \Gamma_j(c_j)) = \sup_{\vartheta : \boldsymbol{\theta}_j(\vartheta) \in H_j} \mathrm{P}_{\vartheta}(T_j(X) \in \Gamma_j(c_j)) = c_j.$$

Regarding the second assertion, we obtain that

$$G_j(t) = \mathrm{P}_{\vartheta_0}(p_j^{LFC}(X) \leq t \mid \hat{\boldsymbol{\theta}}_j(X) \in K_j) = \frac{\mathrm{P}_{\vartheta_0}(p_j^{LFC}(X) \leq t,\ \hat{\boldsymbol{\theta}}_j(X) \in K_j)}{\mathrm{P}_{\vartheta_0}(\hat{\boldsymbol{\theta}}_j(X) \in K_j)}. \tag{A.1}$$

Furthermore, it holds that

$$\begin{cases} \hat{\boldsymbol{\theta}}_j(x) \in K_j \Longrightarrow p_j^{LFC}(x) \leq c_j \Longrightarrow p_j^{LFC}(x) \leq t, & t \geq c_j, \\ p_j^{LFC}(x) \leq t \Longrightarrow p_j^{LFC}(x) < c_j \Longrightarrow \hat{\boldsymbol{\theta}}_j(x) \in K_j, & t < c_j, \end{cases} \tag{A.2}$$

for all $x \in \Omega$. Consequently, the numerator on the right hand side in (A.1) is either $\mathrm{P}_{\vartheta_0}(p_j^{LFC}(X) \le t) = t$ or $\mathrm{P}_{\vartheta_0}(\hat{\boldsymbol{\theta}}_j \in K_j)$ for $t < c_j$ and $t \ge c_j$, respectively. This leads to

$$\begin{aligned} G_j(t) &= \frac{t}{\mathrm{P}_{\vartheta_0}(\hat{\boldsymbol{\theta}}_j \in K_j)} \mathbf{1}_{[0,c_j)}(t) + \mathbf{1}_{[c_j,1]}(t) \\ &= \frac{t}{\mathrm{P}_{\vartheta_0}(\hat{\boldsymbol{\theta}}_j \in K_j)} \mathbf{1}_{[0,\mathrm{P}_{\vartheta_0}(\hat{\boldsymbol{\theta}}_j \in K_j))}(t) + \mathbf{1}_{[\mathrm{P}_{\vartheta_0}(\hat{\boldsymbol{\theta}}_j \in K_j),1]}(t). \end{aligned}$$

Finally, we show the third assertion. Using Part 2, we only have to show, that $\hat{\boldsymbol{\theta}}_j(x) \in K_j$ implies $p_j^{LFC}(x) \le c_j$ for all $x \in \Omega$, which is already shown in (A.2).

A.3.2 Proof of Theorem 2.2

We recall from Theorem 2.1 that it holds

$$p_j^{rand}(X, U_j) = U_j \ \mathbf{1}_{H_j}\{\hat{\boldsymbol{\theta}}_j(X)\} + \frac{p_j^{LFC}(X)}{c_j} \ \mathbf{1}_{K_j}\{\hat{\boldsymbol{\theta}}_j(X)\},$$

which implies

$$\begin{aligned} \mathrm{P}_\vartheta\big(p_j^{rand}(X, U_j) \le t\big) &= t \ \mathrm{P}_\vartheta\big(\hat{\boldsymbol{\theta}}_j(X) \in H_j\big) \\ &\quad + \mathrm{P}_\vartheta\left[\frac{p_j^{LFC}(X)}{c_j} \ \mathbf{1}_{K_j}\{\hat{\boldsymbol{\theta}}_j(X)\} \le t\right], \quad t \in [0,1]. \end{aligned} \tag{A.3}$$

Now, $\mathrm{P}_\vartheta\big(p_j^{rand}(X, U_j) \le t\big) \le t$ holds, if and only if for the second summand in $(A.3)$

$$\mathrm{P}_\vartheta\left[\frac{p_j^{LFC}(X)}{c_j} \mathbf{1}_{K_j}\{\hat{\boldsymbol{\theta}}_j(X)\} \le t\right] \le t \ \mathrm{P}_\vartheta\big(\hat{\boldsymbol{\theta}}_j(X) \in K_j\big) \tag{A.4}$$

is fulfilled. Note, that due to assumption $(GA1)$ the term $\mathbf{1}_{K_j}\{\hat{\boldsymbol{\theta}}_j(X)\}$ on the left-hand side in (A.4) can be omitted.

The statement in Theorem 2.2 was that

$$\mathrm{P}_\vartheta\big(T_j(X) \in \Gamma_j(z)\big) \le z \, \frac{\mathrm{P}_\vartheta(\hat{\boldsymbol{\theta}}_j \in K_j)}{\mathrm{P}_{\vartheta_0}(\hat{\boldsymbol{\theta}}_j \in K_j)}, \quad 0 \le z \le \mathrm{P}_{\vartheta_0}(\hat{\boldsymbol{\theta}}_j \in K_j),$$

is equivalent to the validity of p_j^{rand}.

This follows from (A.4) when substituting $z = t\, c_j = t\, \mathrm{P}_{\vartheta_0}(\hat{\boldsymbol{\theta}}_j \in K_j)$ and by seeing that $\mathrm{P}_\vartheta\big(p_j^{LFC}(X) \le t\big) = \mathrm{P}_\vartheta\big(T_j(X) \in \Gamma_j(t)\big)$, $t \in [0,1]$, holds.

A.3.3 Proof of Theorem 2.3

At first we show that

$$\mathrm{P}_\vartheta\big(T_j(X) > z\big) \le \mathrm{P}_{\vartheta_0}\big(T_j(X) > z\big) \ \frac{\mathrm{P}_\vartheta(\hat{\boldsymbol{\theta}}_j \in K_j)}{\mathrm{P}_{\vartheta_0}(\hat{\boldsymbol{\theta}}_j \in K_j)}, \ z \in [F^{-1}(1 - c_j), \infty], \tag{A.5}$$

holding for any $\vartheta \in \Theta$ with $\boldsymbol{\theta}_j(\vartheta) \in H_j$ is equivalent to the validity of p_j^{rand}.

We make use of the following auxiliary result.

Lemma. *Let $h_\vartheta : [0,1] \to [0,1]$ be defined as follows*

$$h_\vartheta(z) = \mathrm{P}_\vartheta\Big(T_j(X) \in \Gamma_j\big(z \ \mathrm{P}_{\vartheta_0}(\hat{\boldsymbol{\theta}}_j \in K_j)\big)\Big) - z \ \mathrm{P}_\vartheta(\hat{\boldsymbol{\theta}}_j \in K_j).$$

Then, for all $\vartheta \in \Theta$ with $\boldsymbol{\theta}_j(\vartheta) \in H_j$, it holds $h_\vartheta(0) = h_\vartheta(1) = 0$.

Proof. We see that $h_\vartheta(0) = \mathrm{P}_\vartheta\big(T_j(X) \in \Gamma_j(0)\big) = 0$. Due to assumption $(GA1)$ and Theorem 2.1 it holds

$$\big\{x \in \Omega : T_j(x) \in \Gamma_j\big(\underbrace{\mathrm{P}_{\vartheta_0}(\hat{\boldsymbol{\theta}}_j \in K_j)}_{c_j}\big)\big\} = \big\{x \in \Omega : \hat{\boldsymbol{\theta}}_j(x) \in K_j\big\},$$

which implies $h_\vartheta(1) = 0$. □

The condition of $h_\vartheta \leq 0$ holding for all ϑ with $\boldsymbol{\theta}_j(\vartheta) \in H_j$, is equivalent to the condition in Theorem 2.2, and hence equivalent to the validity of p_j^{rand}.

With our condition regarding the rejection region Γ_j and using the representation for h_ϑ from the previous lemma, it holds

$$\begin{aligned} h_\vartheta(t) &= \mathrm{P}_\vartheta\Big[T_j(X) \in \Gamma_j\big(t\ \mathrm{P}_{\vartheta_0}(\hat{\boldsymbol{\theta}}_j \in K_j)\big)\Big] - t\ \mathrm{P}_\vartheta(\hat{\boldsymbol{\theta}}_j \in K_j) \\ &= \mathrm{P}_\vartheta\Big[T_j(X) > F^{-1}\big(1 - t\ \mathrm{P}_{\vartheta_0}(\hat{\boldsymbol{\theta}}_j \in K_j)\big)\Big] - t\ \mathrm{P}_\vartheta(\hat{\boldsymbol{\theta}}_j \in K_j). \end{aligned} \tag{A.6}$$

Substituting $z = F^{-1}\big(1 - t\ \mathrm{P}_{\vartheta_0}(\hat{\boldsymbol{\theta}}_j \in K_j)\big)$ in (A.6), we obtain that

$$\begin{aligned} h_\vartheta(t) &= \mathrm{P}_\vartheta(T_j(X) > z) - (1 - F(z))\frac{\mathrm{P}_\vartheta(\hat{\boldsymbol{\theta}}_j \in K_j)}{\mathrm{P}_{\vartheta_0}(\hat{\boldsymbol{\theta}}_j \in K_j)} \\ &= \mathrm{P}_\vartheta(T_j(X) > z) - \mathrm{P}_{\vartheta_0}(T_j(X) > z)\frac{\mathrm{P}_\vartheta(\hat{\boldsymbol{\theta}}_j \in K_j)}{\mathrm{P}_{\vartheta_0}(\hat{\boldsymbol{\theta}}_j \in K_j)}, \end{aligned} \tag{A.7}$$

and thus $h_\vartheta(t) \leq 0$ for all $t \in [0,1]$ if and only if (A.5) holds. Furthermore, from assumption $(GA1)$ it holds $\big\{\hat{\boldsymbol{\theta}}_j \in K_j\big\} = \big\{T_j(X) \in \Gamma_j(c_j)\big\} = \big\{T_j(X) > F^{-1}(1 - c_j) =: a\big\}$, which implies, that (A.5) is equivalent to

$$\frac{\mathrm{P}_\vartheta\big(T_j(X) > a + b\big)}{\mathrm{P}_\vartheta\big(T_j(X) > a\big)} \leq \frac{\mathrm{P}_{\vartheta_0}\big(T_j(X) > a + b\big)}{\mathrm{P}_{\vartheta_0}\big(T_j(X) > a\big)}, \text{ for all } b > 0. \tag{A.8}$$

Now $T_j(X)^{(\vartheta)} \leq_{\mathrm{hr}} T_j(X)^{(\vartheta_0)}$ is equivalent to (A.8) holding for any a, and thus, it implies (A.5) and therefore the validity of p_j^{rand}.

A.3.4 Proof of Theorem 2.4

Let a model as in Section 2.2 be given and $j \in \{1, \ldots, m\}$ be fixed. We introduce the notation $p(X, U_j, c) = U_j \mathbf{1}\{p_j^{LFC} > c\} + p_j^{LFC}(X) c^{-1} \mathbf{1}\{p_j^{LFC}(X) \leq c\}$ for any $c \in [0,1]$. It is $p_j^{rand}(X, U_j) = p(X, U_j, c_j)$ almost surely. Notice, that $p(X, U_j, 0) = U_j$ and $p(X, U_j, 1) = p_j^{LFC}(X)$.

For given $t \in [0,1]$ and $\vartheta \in \Theta$ we look at the function $c \mapsto h(c) = \mathrm{P}_\vartheta(p(X, U_j, c) \leq t)$. We want to show that h is non-decreasing if the cumulative distribution function of $p_j^{LFC}(X)$ is convex and non-increasing if it is concave under ϑ. It holds

$$h(c) = t\,\mathrm{P}_\vartheta(p_j^{LFC}(X) > c) + \mathrm{P}_\vartheta(p_j^{LFC}(X) \leq c\,t)$$

and

$$h'(c) = -t\, f_\vartheta(c) + t\, f_\vartheta(c\,t),$$

where f_ϑ is the density of p_j^{LFC} under ϑ.

Now, if the cumulative distribution function of p_j^{LFC} is convex under ϑ, then f_ϑ is a non-decreasing function and $f_\vartheta(c\,t) \leq f_\vartheta(c)$ for all c, and analogously $f_\vartheta(c\,t) \geq f_\vartheta(c)$ for all c, if the cumulative distribution function of p_j^{LFC} is concave under ϑ.

A.3.5 Proof of Lemma 2.1

We start with assumption $(GA1)$. It holds $\hat{\boldsymbol{\theta}}_j(x) \in K_j$ if and only if $\hat{\theta}_{i,j}(x) > 0$ for at least γ indices $i \in \{1, \ldots, s\}$. Due to assumption $(RA2)$, the latter holds if and only if $p_{i,j}(x) < d_j$ for at least γ indices $i \in \{1, \ldots, s\}$, which is equivalent to $1 - p_{(\gamma),j}(x) > 1 - d_j$. Furthermore, $T_j(x) \in \Gamma_j(\alpha)$ is equivalent to $1 - p_{(\gamma),j}(x) > F^{-1}_{\mathrm{Beta}(s-\gamma+1,1)}(1-\alpha)$, such that for $c_j = 1 - F_{\mathrm{Beta}(s-\gamma+1,1)}(1 - d_j)$, assumption $(GA1)$ is satisfied, i.e. $\{x \in \Omega :\ T_j(x) \in \Gamma_j(c_j)\} = \{x \in \Omega :\ \hat{\boldsymbol{\theta}}_j(x) \in K_j\}$.

For the verification of $(GA2)$ (nested rejection regions), we see that for every $j \in \{1, \ldots, m\}$ and $\alpha' < \alpha$ it holds $F^{-1}_{\mathrm{Beta}(s-\gamma+1,1)}(1-\alpha') \geq F^{-1}_{\mathrm{Beta}(s-\gamma+1,1)}(1-\alpha)$ and therefore $\Gamma_j(\alpha') \subseteq \Gamma_j(\alpha)$.

To see that $(GA4)$ is fulfilled, let $j \in \{1, \ldots, m\}$ be fixed. We calculate the set of LFCs for φ_j, i.e. the set of parameters $\vartheta' \in \Theta$ that yield the supremum in

$$\sup_{\vartheta' \in \Theta : \boldsymbol{\theta}_j(\vartheta') \in H_j} \mathrm{P}_{\vartheta'}\big(T_j(X) \in \Gamma_j(\alpha)\big),$$

and show that it does not depend on α.

First, it holds $\mathrm{P}_\vartheta\big(T_j(X) \in \Gamma_j(\alpha)\big) = \mathrm{P}_\vartheta\big(1 - p_{(\gamma),j}(X) > F^{-1}_{\mathrm{Beta}(s-\gamma+1,1)}(1-\alpha)\big)$, which is larger the smaller the p-values $p_{1,j}(X), \ldots, p_{s,j}(X)$ (stochastically) are. For every $i = 1, \ldots, s$, due to $(RA4)$, there

exist parameters $\vartheta_i \in \Theta$, independent of α, such that $p_{i,j}(X) = 0$ almost surely under ϑ_i. Independently of α, this is satisfied for parameters with $\theta_{i,j}(\vartheta_i)$ large enough. It is clear, that for any LFC $\vartheta_0 \in \Theta$ for φ_j, it has to hold $\boldsymbol{\theta}_j(\vartheta_0) \in H_j$ and $\theta_{i,j}(\vartheta_0)$ large enough (without loss of generality equal to ∞) for $\gamma - 1$ indices i.

Without loss of generality, we consider a parameter $\vartheta_0 \in \Theta$ with $\boldsymbol{\theta}_j(\vartheta_0) \in H_j$ and $\theta_{1,j}(\vartheta_0) = \cdots = \theta_{\gamma-1,j}(\vartheta_0) = \infty$, leaving $\theta_{i,j}(\vartheta_0) \leq 0$ for the remaining indices $i = \gamma, \ldots, s$.

Due to assumption $(RA4)$, the p-values $p_{1,j}(X), \ldots, p_{\gamma-1,j}(X)$ are almost surely zero and $T_j(X) = 1 - p_{(\gamma),j}(X) = \max\{1 - p_{\gamma,j}(X), \ldots, 1 - p_{s,j}(X)\}$ almost surely under ϑ_0. We obtain that

$$\begin{aligned}
&\mathrm{P}_{\vartheta_0}\big(T_j(X) \in \Gamma_j(\alpha)\big) \\
=&\mathrm{P}_{\vartheta_0}\big(\max\{1 - p_{\gamma,j}(X), \ldots, 1 - p_{s,j}(X)\} > F^{-1}_{\mathrm{Beta}(s-\gamma+1,1)}(1-\alpha)\big) \\
=&1 - \mathrm{P}_{\vartheta_0}\big(\max\{1 - p_{\ell,j}(X) : \gamma \leq \ell \leq s\} \leq F^{-1}_{\mathrm{Beta}(s-\gamma+1,1)}(1-\alpha)\big). \qquad \text{(A.9)}
\end{aligned}$$

Since the studies are independent, (A.9) is equal to

$$\begin{aligned}
&1 - \prod_{i=\gamma}^{s} \mathrm{P}_{\vartheta_0}\big(1 - p_{i,j}(X) \leq F^{-1}_{\mathrm{Beta}(s-\gamma+1,1)}(1-\alpha)\big) \\
=&1 - \prod_{i=\gamma}^{s} \Big[1 - \mathrm{P}_{\vartheta_0}\big(p_{i,j}(X) < 1 - F^{-1}_{\mathrm{Beta}(s-\gamma+1,1)}(1-\alpha)\big)\Big]. \qquad \text{(A.10)}
\end{aligned}$$

Now, using the relation in (2.6), the term in (A.10) equals

$$1 - \prod_{i=u}^{s} \Big[1 - \mathrm{P}_{\vartheta_0}\big(T_{i,j}(X) \in \Gamma_{i,j}(\alpha_{i,j})\big)\Big],$$

where $\alpha_{i,j} = 1 - F^{-1}_{\mathrm{Beta}(s-u+1,1)}(1-\alpha)$, which is maximized if each term $\mathrm{P}_{\vartheta_0}\big(T_{i,j}(X) \in \Gamma_{i,j}(\alpha_{i,j})\big)$ is maximized over the set of all ϑ_0 with $\theta_{i,j}(\vartheta_0) \leq 0$ $(i = \gamma, \ldots, s)$. Due to assumption $(RA2)$, this is the case for any $\vartheta_0 \in \Theta$ with $\theta_{i,j}(\vartheta_0) = 0$ independently of $\alpha_{i,j}$ $(i = \gamma, \ldots, s)$, such that $\mathrm{P}_{\vartheta_0}\big(T_j(X) \in \Gamma_j(\alpha)\big)$ is being maximized by any parameter ϑ_0 with

$$\boldsymbol{\theta}_j(\vartheta_0) = (\underbrace{\infty, \ldots, \infty}_{\gamma-1}, \underbrace{0, \ldots, 0}_{s-\gamma+1})$$

independently of α.

Altogether, the set of LFCs for φ_j is

$$\{\vartheta \in \Theta : \boldsymbol{\theta}_j(\vartheta) \text{ is any permutation of } (\underbrace{\infty, \ldots, \infty}_{\gamma-1}, \underbrace{0, \ldots, 0}_{s-\gamma+1})\},$$

hence, obviously independent of α.

Finally, we verify $(GA3)$ as follows: For every $j \in \{1, \ldots, m\}$ and $\alpha \in (0, 1)$, it holds

$$\begin{aligned}
&\sup_{\vartheta \in \Theta : \boldsymbol{\theta}_j(\vartheta) \in H_j} \mathrm{P}_{\vartheta}(T_j(X) \in \Gamma_j(\alpha)) = \mathrm{P}_{\vartheta_0}\big(T_j(X) \in \Gamma_j(\alpha)\big) \qquad \text{(A.11)} \\
=&\mathrm{P}_{\vartheta_0}\big[\max\{1 - p_{\ell,j}(X) : \gamma \leq \ell \leq s\} \geq F^{-1}_{\mathrm{Beta}(s-\gamma+1,1)}(1-\alpha)\big], \qquad \text{(A.12)}
\end{aligned}$$

due to $(RA4)$, where $\vartheta_0 \in \Theta$ with $\boldsymbol{\theta}_j(\vartheta_0) = (\infty, \ldots, \infty, 0, \ldots, 0)$ is an LFC for φ_j.

Furthermore, $1 - p_{i,j}(X)$ is uniformly distributed on $[0, 1]$ under an LFC $\vartheta_0 \in \Theta$ with $\theta_{i,j}(\vartheta_0) = 0$ $(i = \gamma, \ldots, s)$. Since $\max(U_1, \ldots, U_k)$ is $\mathrm{Beta}(s - \gamma + 1, 1)$-distributed, for $U_1, \ldots, U_k$, that are stochastically independent and identically, uniformly distributed on $[0, 1]$, we obtain that (A.12) equals $1 - F_{\mathrm{Beta}(s-\gamma+1,1)}\big(F^{-1}_{\mathrm{Beta}(s-\gamma+1,1)}(1-\alpha)\big) = \alpha$, as desired.

A.3.6 Proof of Theorem 2.5

We want to show, that

$$T_j(X)^{(\vartheta)} \leq_{\mathrm{hr}} T_j(X)^{(\vartheta_0)} \qquad \text{(A.13)}$$

holds for any parameters $\vartheta \in \Theta$ with $\boldsymbol{\theta}_j(\vartheta) \in H_j$ and ϑ_0 an LFC for φ_j. Let $\vartheta \in \Theta$ with $\boldsymbol{\theta}_j(\vartheta) \in H_j$, i.e. $\theta_{i,j}(\vartheta) \leq 0$ for at least $s - \gamma + 1$ indices i, be given. Since the distribution of $T_j(X)$ does not depend

on the particular form of the LFC ϑ_0, we choose an LFC that fulfils $\theta_{i,j}(\vartheta) \leq 0 = \theta_{i,j}(\vartheta_0)$ for at least $s-\gamma+1$ indices i. Without loss of generality, let $\theta_{i,j}(\vartheta) \leq 0$ $(i = 1, \ldots, s-\gamma+1)$, and

$$\boldsymbol{\theta}_j(\vartheta_0) = (\underbrace{0, \ldots, 0}_{s-\gamma+1}, \underbrace{\infty, \ldots, \infty}_{\gamma-1}).$$

For $i = 1, \ldots, s-\gamma+1$, it is $\theta_{i,j}(\vartheta) \leq 0 = \theta_{i,j}(\vartheta_0)$, and therefore $T_{i,j}(X)^{(\vartheta)} \leq_{\text{hr}} T_{i,j}(X)^{(\vartheta_0)}$. Let $F_{i,j}$ be the cumulative distribution function of $T_{i,j}(X)$ under an LFC for $\varphi_{i,j}$, i.e. under a $\tilde{\vartheta} \in \Theta$ with $\theta_{i,j}(\tilde{\vartheta}) = 0$. For $i = 1, \ldots, s-\gamma+1$, it holds $\theta_{i,j}(\vartheta_0) = 0$, i.e. the parameter ϑ_0 is an LFC for $\varphi_{i,j}$ $(i = 1, \ldots, s-\gamma+1)$. From Part 3 in Theorem A.2.1, it follows that

$$1 - p_{i,j}(X) = F_{i,j}\big(T_{i,j}(X)\big)^{(\vartheta)} \leq_{\text{hr}} F_{i,j}\big(T_{i,j}(X)\big)^{(\vartheta_0)}. \tag{A.14}$$

Note that $F_{i,j}\big(T_{i,j}(X)\big)$ is uniformly distributed on $[0,1]$ under ϑ_0, $(i = 1, \ldots, s-u+1)$.

For ease of notation, we write $P_i = 1-p_{i,j}$ and $T_j(X) = 1-p_{(\gamma),j}(X) = P_{(s-\gamma+1)}(X)$. Under ϑ_0 it then holds $T_j(X)$ and $\max\{U_1, \ldots, U_{s-\gamma+1}\}$ are identically distributed, where $U_1, \ldots, U_{s-\gamma+1}$ are stochastically independent and identically, uniformly distributed on $[0,1]$, since $P_{s-\gamma+2}(X) = \cdots = P_s(X) = 1$ almost surely due to $(RA4)$.

Now, (A.13) is equivalent to $P_{(s-\gamma+1:n)}(X)^{(\vartheta)} \leq_{\text{hr}} P_{(s-\gamma+1:s)}(X)^{(\vartheta_0)} \sim U_{(s-\gamma+1:s-\gamma+1)}$, which follows directly from Lemma A.2.2, since, from (A.14), it holds $P_i(X)^{(\vartheta)} \leq_{\text{hr}} U_i$ for at least $s-\gamma+1$ indices $i \in \{1, \ldots, s\}$.

A.4 Further simulation results

The results of our Monte Carlo simulation with regard to the standard deviations, cf. the end of Section 2.5.2, are listed in Table 2.3 and Table 2.4 for the utilization of the LFC-based and the randomized p-values, respectively.

Furthermore, we looked at two different approaches for defining the LFC-based p-values. The test statistics $T_j(X) = 1 - p_{(\gamma),j}$ do not regard the size of the $s-\gamma$ larger p-values $p_{(\gamma+1)}, \ldots, p_{(s)}$ explicitly. Instead, one could consider

$$T_j^{(S)}(X) = (s-\gamma+1)^{-1/2} \sum_{i=\gamma}^{s} \Phi^{-1}\big(1 - p_{(i),j}(X)\big), \quad \Gamma_j^{(S)}(\alpha) = \big(\Phi^{-1}(1-\alpha), \infty\big),$$

or

$$T_j^{(F)}(X) = -2 \sum_{i=\gamma}^{s} \log(p_{(i),j}(X)), \quad \Gamma_j^{(F)}(\alpha) = \big(F^{-1}_{\chi^2_{2\ (s-\gamma+1)}}(1-\alpha), \infty\big),$$

motivated by the Stouffer method and the Fisher method for combining p-values, respectively, where Φ is the cumulative distribution function of the standard normal distribution in $\mathbb{R}$, and $F_{\chi^2_{2\ (s-\gamma+1)}}$ is the cumulative distribution function of a χ^2-distribution with $2\ (s-\gamma+1)$ degrees of freedom (Benjamini and Heller, 2008, Sec. 2.2). Benjamini and Heller (2008) showed that applying the Benjamini–Hochberg linear step up test from Benjamini and Hochberg (1995) on the LFC-based p-values $p_1^{LFC}, \ldots, p_m^{LFC}$ controls the false discovery rate even if the p-values within each study admit a positive dependence. For more details see Theorem 3 in Benjamini and Heller (2008).

Models based on these test statistics, however, do not fulfil assumption $(GA1)$ from Section 2.2, such that Theorem 2.1 does not apply, and calculating the randomized p-values $p_1^{rand}, \ldots, p_m^{rand}$ as in Definition 2.1 becomes more difficult.

We simulated the expected values of the estimator $\hat{\pi}_0(1/2)$ when utilizing the LFC-based p-values under these alternative test statistics. The results of the Monte Carlo simulations with 10 000 repetitions can be found in Table A.1 for the Stouffer-based and Table A.2 for the Fisher-based p-values. More accurate estimations as compared to $\hat{\pi}_0^{rand}$ are written in bold. Compared to the expected values when utilizing our randomized p-values $(p_j^{rand})_j$ both alternatives only perform better in case of $\mu_{\min} = 0$ and low γ ($\gamma = 2, 4, 6$ for Stouffer, and $\gamma = 2, 4$ for Fisher).

Table A.1: Expected values of $\hat{\pi}_0(1/2)$ using $(p_j^{(S)})_{j=1,\dots,m}$ under Model 1 with $s = 10$. Values result from Monte Carlo simulations with 10,000 repetitions. Values that come closer to the true proportion π_0 than under the use of our randomized p-values are written in bold.

$\gamma = 2$	π_0	0.6	0.7	0.8	0.9
$(\mu_{\min}, \mu_{\max})$					
(-0,2)		**0.6646**	**0.7746**	**0.8859**	**0.9968**
(-0.5,3)		0.9636	1.1246	1.2855	1.4458
(-1,4)		1.1254	1.3129	1.5002	1.6878
(-1.5,5)		1.1809	1.3777	1.5746	1.7714

$\gamma = 4$	π_0	0.6	0.7	0.8	0.9
$(\mu_{\min}, \mu_{\max})$					
(-0,2)		**0.7806**	**0.9095**	**1.0402**	**1.1699**
(-0.5,3)		1.005	1.1721	1.3395	1.5072
(-1,4)		1.1287	1.3166	1.5047	1.6928
(-1.5,5)		1.1775	1.3737	1.5697	1.7661

$\gamma = 6$	π_0	0.6	0.7	0.8	0.9
$(\mu_{\min}, \mu_{\max})$					
(-0,2)		**0.8837**	**1.0281**	**1.1737**	**1.3187**
(-0.5,3)		1.0401	1.2114	1.3836	1.5556
(-1,4)		1.1322	1.3196	1.5077	1.6947
(-1.5,5)		1.1742	1.3692	1.5636	1.7582

$\gamma = 8$	π_0	0.6	0.7	0.8	0.9
$(\mu_{\min}, \mu_{\max})$					
(-0,2)		1.0014	1.1511	1.301	1.4509
(-0.5,3)		1.0872	1.2572	1.4286	1.5989
(-1,4)		1.1484	1.3305	1.5131	1.6956
(-1.5,5)		1.1815	1.3704	1.5593	1.7484

$\gamma = 10$	π_0	0.6	0.7	0.8	0.9
$(\mu_{\min}, \mu_{\max})$					
(-0,2)		1.2141	1.3332	1.4525	1.5724
(-0.5,3)		1.2064	1.3501	1.4938	1.6377
(-1,4)		1.2167	1.3752	1.5335	1.6919
(-1.5,5)		1.2243	1.3924	1.5608	1.7285

Table A.2: Expected values of $\hat{\pi}_0(1/2)$ using $(p_j^{(F)})_{j=1,\dots,m}$ under Model 1 with $s = 10$. Values result from Monte Carlo simulations with 10,000 repetitions. Values that come closer to the true proportion π_0 than under the use of our randomized p-values are in bold.

$\gamma = 2$	π_0	0.6	0.7	0.8	0.9
$(\mu_{\min}, \mu_{\max})$					
(-0,2)		**0.6895**	**0.8036**	**0.9193**	**1.0347**
(-0.5,3)		0.9489	1.1068	1.2652	1.4232
(-1,4)		1.0946	1.2773	1.4599	1.6418
(-1.5,5)		1.1567	1.3497	1.5426	1.7353

$\gamma = 4$	π_0	0.6	0.7	0.8	0.9
$(\mu_{\min}, \mu_{\max})$					
(-0,2)		**0.8337**	**0.9716**	**1.1106**	**1.2498**
(-0.5,3)		1.0134	1.182	1.3509	1.5201
(-1,4)		1.1143	1.2997	1.4859	1.6715
(-1.5,5)		1.1606	1.3536	1.547	1.7408

$\gamma = 6$	π_0	0.6	0.7	0.8	0.9
$(\mu_{\min}, \mu_{\max})$					
(-0,2)		0.9427	1.0971	1.2528	1.4084
(-0.5,3)		1.0593	1.2345	1.4104	1.5858
(-1,4)		1.1282	1.3158	1.5037	1.6907
(-1.5,5)		1.1626	1.3561	1.5492	1.7426

$\gamma = 8$	π_0	0.6	0.7	0.8	0.9
$(\mu_{\min}, \mu_{\max})$					
(-0,2)		1.0495	1.2077	1.3659	1.5243
(-0.5,3)		1.1063	1.2811	1.4565	1.6315
(-1,4)		1.1486	1.3328	1.5175	1.7021
(-1.5,5)		1.1716	1.3616	1.5514	1.7416

$\gamma = 10$	π_0	0.6	0.7	0.8	0.9
$(\mu_{\min}, \mu_{\max})$					
(-0,2)		1.2141	1.3332	1.4525	1.5724
(-0.5,3)		1.2064	1.3501	1.4938	1.6377
(-1,4)		1.2167	1.3752	1.5335	1.6919
(-1.5,5)		1.2243	1.3924	1.5608	1.7285

Chapter B

Appendix for Chapter 3

B.1 The more general randomized p-values

B.1.1 Definition

Let $U_1, \ldots, U_m$ and X be as before. For a set of stochastically independent (not necessarily identically distributed) random variables $R_1, \ldots, R_m$ with values in $[0,1]$, that are defined on the same probability space as X, stochastically independent of the U_j's and the data X, and whose distributions do not depend on ϑ, we define

$$p_j^{rand}(X, U_j, R_j) = U_j \mathbf{1}\{p_j^{LFC}(X) \geq R_j\} + \frac{p_j^{LFC}(X)}{R_j} \mathbf{1}\{p_j^{LFC}(X) < R_j\}, \tag{B.1}$$

$j = 1, \ldots, m$. This definition includes the case $R_j \equiv c_j$ from Definition 3.1 for any constant $c_j \in [0,1]$, $j = 1, \ldots, m$. We generalize and prove Theorems 3.1 and 3.2 for the randomized p-values $\{p_j^{rand}(X, U_j, R_j)\}_{1 \leq j \leq m}$.

B.1.2 Theorem 3.1′

Let a model as in Section 3.2 be given and $j \in \{1, \ldots, m\}$ be fixed. Then, the j-th randomized p-value $p_j^{rand}(X, U_j, R_j)$ as in (B.1) is a valid p-value for a given random variable R_j with values in $[0,1]$ if and only if condition (0.) is fulfilled. Furthermore, either of the following conditions (1.′), (2.), and (3.) is a sufficient condition for the validity of $p_j^{rand}(X, U_j, R_j)$ for any random variable R_j with values in $[0,1]$.

(0.) For every $\vartheta \in \Theta$ with $\theta_j(\vartheta) \in H_j$, it holds

$$\mathbb{P}_\vartheta(p_j^{LFC}(X) \leq tR_j) \leq t\mathbb{P}_\vartheta(p_j^{LFC}(X) \leq R_j)$$

for all $t \in [0,1]$.

(1.′) For every $\vartheta \in \Theta$ with $\theta_j(\vartheta) \in H_j$, it holds

$$\mathbb{P}_\vartheta(p_j^{LFC}(X) \leq tu) \leq t\mathbb{P}_\vartheta(p_j^{LFC}(X) \leq u)$$

for all $u, t \in [0,1]$.

(2.) For every $\vartheta \in \Theta$ with $\theta_j(\vartheta) \in H_j$, $\mathbb{P}_\vartheta(p_j^{LFC}(X) \leq t)/t$ is non-decreasing in t.

(3.) The cdf of $p_j^{LFC}(X)$ is convex under any $\vartheta \in \Theta$ with $\theta_j(\vartheta) \in H_j$.

Let F_ϑ be the cdf of $T_j(X)$ under $\vartheta \in \Theta$. If the LFC-based p-value is given by $p_j^{LFC}(X) = 1 - F_{\vartheta_0}(T_j(X))$, where $\vartheta_0 \in \Theta$ is an LFC for φ_j, then the following condition (4.) is equivalent to condition (2.), while condition (5.) is equivalent to condition (3.).

(4.) For every $\vartheta \in \Theta$ with $\theta_j(\vartheta) \in H_j$, it holds $T_j(X)^{(\vartheta)} \leq_{\mathrm{hr}} T_j(X)^{(\vartheta_0)}$.

(5.) For every $\vartheta \in \Theta$ with $\theta_j(\vartheta) \in H_j$, it holds $T_j(X)^{(\vartheta)} \leq_{\mathrm{lr}} T_j(X)^{(\vartheta_0)}$.

Proof. First, we show that condition (0.) is equivalent to $p_j^{rand}(X, U_j, R_j)$ being valid. For $p_j^{rand}(X, U_j, R_j)$ to be valid it has to hold that

$$\mathbb{P}_\vartheta(p_j^{rand}(X, U_j, R_j) \leq t) \leq t,$$

for all $t \in [0, 1]$ and all $\vartheta \in \Theta$ with $\theta_j(\vartheta) \in H_j$. It holds that

$$\begin{aligned} \mathbb{P}_\vartheta(p_j^{rand}(X, U_j, R_j) \leq t) &= \mathbb{P}_\vartheta(U_j \leq t)\mathbb{P}_\vartheta(p_j^{LFC}(X) > R_j) + \mathbb{P}_\vartheta(p_j^{LFC}(X) \leq tR_j) \\ &= t\mathbb{P}_\vartheta(p_j^{LFC}(X) > R_j) + \mathbb{P}_\vartheta(p_j^{LFC}(X) \leq tR_j). \end{aligned} \tag{B.2}$$

Now, the term in (B.2) is not larger than t if and only if it holds

$$\mathbb{P}_\vartheta(p_j^{LFC}(X) \leq tR_j) \leq t\big[1 - \mathbb{P}_\vartheta(p_j^{LFC}(X) > R_j)\big] = t\mathbb{P}_\vartheta(p_j^{LFC}(X) \leq R_j),$$

which is condition (0.).

Let G be the cdf of R_j. From condition (1.′) it follows

$$\int_0^1 \mathbb{P}_\vartheta(p_j^{LFC}(X) \leq tu)\mathrm{d}G(u) \leq t \int_0^1 \mathbb{P}_\vartheta(p_j^{LFC}(X) \leq u)\mathrm{d}G(u),$$

for every $t \in [0, 1]$, thus condition (1.′) implies (0.).

Substituting $z = tu$ in condition (1.′) leads to

$$\mathbb{P}_\vartheta(p_j^{LFC}(X) \leq z) \leq z\frac{\mathbb{P}_\vartheta(p_j^{LFC}(X) \leq u)}{u}$$

for all $0 \leq z < u \leq 1$ and all $\vartheta \in \Theta$ with $\theta_j(\vartheta) \in H_j$, which is equivalent to condition (2.).

Now, we show that condition (3.) implies condition (1.′). Let $u \in [0, 1]$ be fixed. The inequality in (1.′) is always satisfied for $t = 0$ and $t = 1$. Since $t \mapsto t\mathbb{P}_\vartheta(p_j^{LFC}(X) \leq u)$ is a linear function and $t \mapsto \mathbb{P}_\vartheta(p_j^{LFC}(X) \leq tu)$ is a convex function, if (3.) is fulfilled, it holds

$$\mathbb{P}_\vartheta(p_j^{LFC}(X) \leq tu) \leq t\mathbb{P}_\vartheta(p_j^{LFC}(X) \leq u)$$

for all $t \in [0, 1]$.

Now we assume that $p_j^{LFC}(X) = 1 - F_{\vartheta_0}(T_j(X))$. At first we show that conditions (2.) and (4.) are equivalent. To this end, notice that the term

$$\frac{\mathbb{P}_\vartheta(p_j^{LFC}(X) \leq t)}{t} = \frac{\mathbb{P}_\vartheta(p_j^{LFC}(X) \leq t)}{\mathbb{P}_{\vartheta_0}(p_j^{LFC}(X) \leq t)} = \frac{\mathbb{P}_\vartheta(T_j(X) \geq F_{\vartheta_0}^{-1}(1-t))}{\mathbb{P}_{\vartheta_0}(T_j(X) \geq F_{\vartheta_0}^{-1}(1-t))}$$

is non-decreasing in t if and only if $\mathbb{P}_\vartheta(T_j(X) \geq z)/\mathbb{P}_{\vartheta_0}(T_j(X) \geq z) = (1 - F_\vartheta(z))/(1 - F_{\vartheta_0}(z))$ is non-increasing in z.

Lastly, we show that conditions (3.) and (5.) are equivalent. Let f_ϑ be the Lebesgue density of $T_j(X)$ under $\vartheta \in \Theta$. Let $\vartheta \in \Theta$ be such that $\theta_j(\vartheta) \in H_j$ holds. The convexity of $t \mapsto \mathbb{P}_\vartheta(p_j^{LFC}(X) \leq t)$ is equivalent to

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{P}_\vartheta(T_j(X) \geq F_{\vartheta_0}^{-1}(1-t)) &= \frac{\mathrm{d}}{\mathrm{d}t}\big[1 - F_\vartheta(F_{\vartheta_0}^{-1}(1-t))\big] \\ &= \frac{f_\vartheta(F_{\vartheta_0}^{-1}(1-t))}{f_{\vartheta_0}(F_{\vartheta_0}^{-1}(1-t))} \end{aligned}$$

being non-decreasing in t, or $f_\vartheta(z)/f_{\vartheta_0}(z)$ being non-increasing in z, which is equivalent to condition (5.); cf. the remarks after Theorem 3.1. □

In Theorem 3.1′, the conditions (2.) – (5.) are the same as in Theorem 3.1. Condition (1.′) is equivalent to condition (1.) in Theorem 3.1 holding for all $c_j \in [0, 1]$. Thus, $p_j^{rand}(X, U_j, c_j)$ being valid for all $c_j \in [0, 1]$ implies the validity of $p_j^{rand}(X, U_j, R_j)$ for any random variable R_j on $[0, 1]$, $j = 1, \ldots, m$. The reverse is also true, thus, the randomized p-value $p_j^{rand}(X, U_j, R_j)$ is valid for any random variable R_j on $[0, 1]$ if and only if it is valid for $R_j \equiv c_j$, for all $c_j \in [0, 1]$, $j = 1, \ldots, m$.

In the following, we show that Theorem 3.2 still holds if we replace the constants $c_j \leq \tilde{c}_j$ by the random variables $R_j \leq_{\mathrm{st}} \tilde{R}_j$.

B.1.3 Theorem 3.2′

Let a model as in Section 3.2 be given and $j \in \{1, \dots, m\}$ be fixed. If the cdf of $p_j^{LFC}(X)$ is convex under a fixed $\vartheta \in \Theta$, then it is

$$p_j^{rand}(X, U_j, R_j)^{(\vartheta)} \leq_{\text{st}} p_j^{rand}(X, U_j, \tilde{R}_j)^{(\vartheta)}$$

for any random variables $R_j, \tilde{R}_j$ on $[0,1]$, with $R_j \leq_{\text{st}} \tilde{R}_j$.

If the cdf of $p_j^{LFC}(X)$ is concave under a fixed $\vartheta \in \Theta$, then it holds that

$$p_j^{rand}(X, U_j, \tilde{R}_j)^{(\vartheta)} \leq_{\text{st}} p_j^{rand}(X, U_j, R_j)^{(\vartheta)}$$

for any random variables R_j and $\tilde{R}_j$ with values in $[0,1]$ and with $R_j \leq_{\text{st}} \tilde{R}_j$.

Proof. We first show both statements in Theorem 3.2′ for constants $0 \leq c_j \leq \tilde{c}_j \leq 1$ instead of random variables R_j and $\tilde{R}_j$, which amounts to the statements in Theorem 3.2.

For every fixed $t \in [0,1]$ and fixed $\vartheta \in \Theta$ we define the function $q : [0,1] \to [0,1]$ by

$$q(c) = \mathbb{P}_\vartheta(p_j^{rand}(X, U_j, c) \leq t) = t\mathbb{P}_\vartheta(p_j^{LFC}(X) > c) + \mathbb{P}_\vartheta(p_j^{LFC}(X) \leq ct).$$

Furthermore, we denote by f_ϑ the Lebesgue density of $p_j^{LFC}(X)$ under ϑ, such that it holds $q'(c) = -tf_\vartheta(c) + tf_\vartheta(ct)$, which is not positive if f_ϑ is non-decreasing and not negative if f_ϑ is non-increasing.

Let R_j and $\tilde{R}_j$ be random variables fulfilling the assumptions of the theorem. If q is non-decreasing, then it holds that $\mathbb{E}[q(R_j)] \leq \mathbb{E}[q(\tilde{R}_j)]$, and if q is non-increasing it holds that $\mathbb{E}[q(R_j)] \geq \mathbb{E}[q(\tilde{R}_j)]$, where $\mathbb{E}$ refers to the joint distribution of R_j and $\tilde{R}_j$. Since $\mathbb{E}[q(R_j)] = \mathbb{P}_\vartheta(p_j^{rand}(X, U_j, R_j) \leq t)$ and $\mathbb{E}[q(\tilde{R}_j)] = \mathbb{P}_\vartheta(p_j^{rand}(X, U_j, \tilde{R}_j) \leq t)$, the proof is completed. □

B.2 Additional simulations

In Section 3.4.2 we analyze the MSE of the Schweder-Spjøtvoll estimator in case the LFC p-values are independent or positively dependent. We employed the multiple Z-tests model and applied a pairwise correlation coefficient $\rho = 0$ or $\rho > 0$ on the test statistics. We can also determine the dependency structure among the LFC p-values directly by defining their copula. In case of independent LFC p-values, their joint copula is the product copula. In case of positively dependent LFC p-values we consider the Gumbel-Hougaard copula, defined as

$$C_\nu(x_1, \dots, x_m) = \exp\left[-\left(\sum_{j=1}^{m} -\ln(x_j)\right)^{1/\nu}\right],$$

where $\nu \geq 1$. For increasing ν the degree of dependence increases.

We employed the same model as in the left graph of Figure 2 in the paper, i.e. the multiple Z-tests model with $\pi_0 = 0.7$ and $\theta_j(\vartheta) = 2.5/\sqrt{n_j}$ if H_j is false and $\theta_j(\vartheta) = -1/\sqrt{n_j}$ if H_j is true, $m = 500$ and $n_j = 50$.

Figures B.1 and B.2 illustrate the effect of the copula of the p-values utilized in $\hat{\pi}_0$ on its variance and its MSE, respectively, in our context. On the left we assumed independent LFC p-values and on the right we assumed that the LFC p-values had the Gumbel-Hougaard copula as a joint copula with copula parameter $\nu = 2$. The values were calculated via Monte-Carlo Simulation with 100,000 repetitions.

In the left graph of Figure B.1, the variance of $\hat{\pi}_0(1/2, c)$ is decreasing in c, cf. also Lemma 1. Furthermore, the variance on the left graph is always below $1/m = 1/500$. So, we may conclude here that taking into account $U_1, \dots, U_m$ increases the variance of $\hat{\pi}_0$, but only to a magnitude which is in essentially all considered cases smaller than that of the bias reduction achieved by randomization. This is also in line with the findings of Dickhaus (2013); see the discussion around Table 2 in that paper.

In the right graph of Figure B.1, the behavior of the variance of $\hat{\pi}_0(1/2, c)$ is different. Here, the randomization reduces the variance of $\hat{\pi}_0$, often by a considerable amount. This can be explained by the fact, that in the dependence structure among $p_1^{rand}(X, U_1, c), \dots, p_m^{rand}(X, U_m, c)$ the Gumbel-Hougaard copula of $p_1^{LFC}(X), \dots, p_m^{LFC}(X)$ and the product copula of $U_1, \dots, U_m$ are "mixed", meaning that the degree of dependency among $p_1^{rand}(X, U_1, c), \dots, p_m^{rand}(X, U_m, c)$ is smaller than that among $p_1^{LFC}(X), \dots, p_m^{LFC}(X)$.

Furthermore, comparing the scalings of the vertical axes in the two graphs of Figure B.1, we can confirm the previous findings by Neumann et al. (2021) (and other authors), that (positively) dependent p-values lead to an increased variance of $\hat{\pi}_0$ when compared with the case of jointly stochastically independent p-values. These results are similar to our results in Section 4.2 and are intended to provide an additional example.

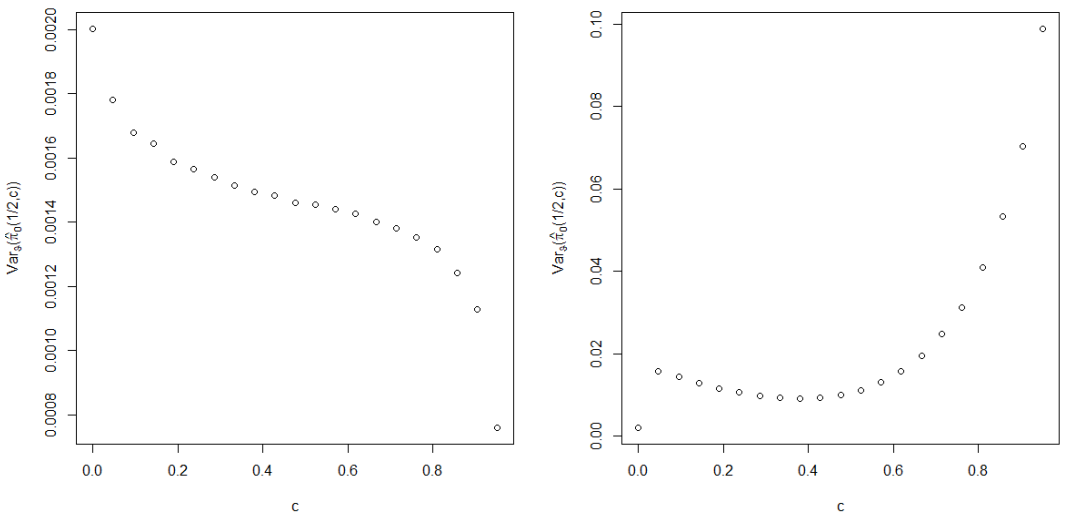


Figure B.1: The variance $\mathrm{Var}_\vartheta(\hat{\pi}_0(1/2,c))$ for $c = 0, 0.05, ..., 1$ in the multiple Z-tests model for $\pi_0 = 0.7$, and $\vartheta \in \Theta$ such that $\theta_j(\vartheta) = -1/\sqrt{50}$ if H_j is true and $\theta_j(\vartheta) = 2.5/\sqrt{50}$ if K_j is true, $j = 1, ..., m = 1,000$. The LFC-based p-values are jointly stochastically independent in the left graph and have the Gumbel-Hougaard copula with copula parameter $\nu = 2$ in the right graph.

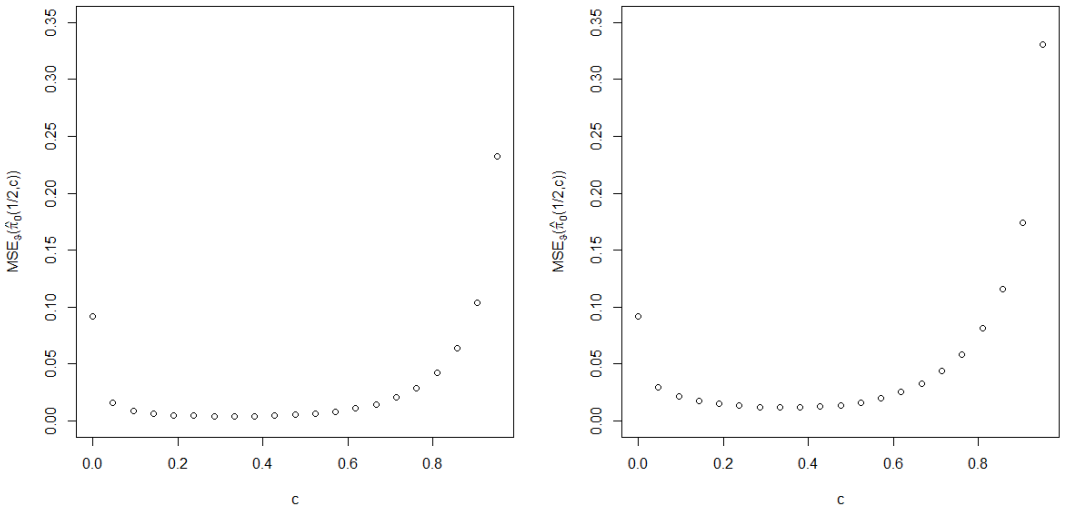


Figure B.2: The mean squared error $\mathrm{MSE}_\vartheta(\hat{\pi}_0(1/2,c))$ for $c = 0, 0.05, ..., 1$ in the multiple Z-tests model for $\pi_0 = 0.7$, and $\vartheta \in \Theta$ such that $\theta_j(\vartheta) = -1/\sqrt{50}$ if H_j is true and $\theta_j(\vartheta) = 2.5/\sqrt{50}$ if K_j is true, $j = 1, ..., m = 1,000$. The LFC-based p-values are jointly stochastically independent in the left graph and have the Gumbel-Hougaard copula with copula parameter $\nu = 2$ in the right graph.

Chapter C

Appendix for Chapter 4

C.1 Some results regarding e-variables and Bayes factors in our models

The following lemma is analogous to Lemma 1 in Benjamini and Heller (2008), for e-variables instead of p-values.

Lemma. *Let $e_1, \ldots, e_s$ be valid e-variables for $H_1, \ldots, H_s$, respectively. If $h_0 : [0, \infty]^{s-\gamma+1} \to [0, \infty]$ is a valid, symmetric and monotonically non-decreasing e-variable combination function for $H^1_{s-\gamma+1}$, then*

$$h\left(e_1, \ldots, e_s\right) = h_0\left(e_{(1)}, \ldots, e_{(s-\gamma+1)}\right)$$

is a valid e-variable for H^γ_s.

Proof. We have to show that

$$\mathbb{E}_\theta\left[h\left(e_1, \ldots, e_s\right)\right] \leq 1 \tag{C.1}$$

holds for all $\theta \in H^\gamma_s$. We determine the parameters $\theta \in H^\gamma_s$ for which $\mathbb{E}_\theta\left[h\left(e_1, \ldots, e_s\right)\right]$ is largest and show that (C.1) holds. Since h_0 is monotonically non-decreasing in all its arguments, so is h. Under H^γ_s, where at least $s-\gamma+1$ of the null hypotheses $H_1, \ldots, H_s$ are true, the worst case for (C.1) occurs, when exactly $s-\gamma+1$ of the null hypotheses $H_1, \ldots, H_s$ are true, since e-variables are unbounded under alternatives. The $\gamma - 1$ e-variables that correspond to the false null hypotheses are as large as possible in the worst case, for simplicity they are ∞.

Without loss of generality, assume that the null hypotheses $H_1, \ldots, H_{s-\gamma+1}$ are the true ones. The ordered e-values

$$(e_{(1)}, \ldots, e_{(s)}) = (\tau(e_1, \ldots, e_{s-\gamma+1}), \infty, \ldots, \infty),$$

where τ is a permutation map, are such that the $s-\gamma+1$ smallest e-variables correspond to the true null hypotheses. Therefore, in the worst case, the combination of these e-variables

$$h\left(e_{(1)}, \ldots, e_{(s-\gamma+1)}, \infty, \ldots, \infty\right) = h_0\left(e_1, \ldots, e_{s-\gamma+1}\right)$$

has an expected value not greater than 1, according to the definition of h_0.

□

The next result helps us construct valid e-variables under composite null hypotheses. We assume that a p-value model as in Section 4.2 is given. Furthermore, we assume that for all i the Lebesgue density f_{θ_i} of p_i is monotonically decreasing if $\theta_i \leq 0$ and monotonically increasing if $\theta_i > 0$. This latter assumption is for example fulfilled under the conditions in Remark 4.1. To see this, we use Assumption $(A2)$ and note that if the distributions $(p_i(\boldsymbol{X})^{(-\theta_i)})_{\theta_i}$ are likelihood ratio ordered, then $f_{\theta_i}(t)/f_{\tilde{\theta}_i}(t)$ is non-decreasing in t, if $\theta_i \leq \tilde{\theta}_i$.

Let i be fixed. For given Bayes marginal probability distributions $\mathbb{P}_0$ under H_i and $\mathbb{P}_1$ under K_i for the parameter values, we define the Bayes factor as

$$\mathrm{BF}(p) := \frac{\int f_{\theta_i}(p) d\mathbb{P}_1(\theta_i)}{\int f_{\theta_i}(p) d\mathbb{P}_0(\theta_i)}. \tag{C.2}$$

Its expected value under θ_i is

$$\mathbb{E}_{\theta_i}\left[\mathrm{BF}\right] = \int \mathrm{BF}(p) f_{\theta_i}(p) dp,$$

which is required to not be larger than 1 under each $\theta_i \in H_i$ for BF to be a valid e-variable for H_i. The following result helps determine whether BF is an e-variable.

Lemma. *Under the assumptions from above and for the Bayes factor* BF *as in* (C.2)*, the expected value* $\mathbb{E}_{\theta_i}[\mathrm{BF}]$ *of* BF *is non-decreasing in* θ_i*.*

Proof. With the assumptions we made, the Bayes factor $\mathrm{BF}(p)$ is non-decreasing in p. Thus the expected value of (the distribution) $[\mathrm{BF}(p_i(\boldsymbol{X}))]^{(\theta_i)}$ decreases with stochastically increasing $p_i(\boldsymbol{X})^{(\theta_i)}$ and therefore with decreasing θ_i.

□

Both the Beta-Model and the Normal-Model fulfil the assumptions for this lemma. With this lemma, the Bayes factor BF has its largest expected value under H_i if $\theta_i = 0$, regardless of the priors. Therefore, $\mathrm{BF}/\mathbb{E}_0[\mathrm{BF}]$ is an e-variable for H_i with expected value 1 under $\theta_i = 0$.

Under the same assumptions, it is also easy to see that the Bayes factor

$$\mathrm{BF}_0(p) := \frac{\int f_{\theta_i}(p) d\mathbb{P}_1(\theta_i)}{f_0(p)} = \int f_{\theta_i}(p) d\mathbb{P}_1(\theta_i)$$

that assumes a simple null hypothesis is an e-variable with expected value 1 under $\theta_i = 0$ even if H_i is composite.

Chapter D

Appendix for Chapter 5

D.1 The empirical distribution of the null PC p-values before and after conditioning on selection

When testing a family of PC hypotheses, it is common that the majority of the PC null hypotheses p-values may be very conservative (i.e., they have a distribution that is stochastically larger than uniform). This follows since in practice many of the true $H^{\gamma/s}$ hypotheses have signal in less than $\gamma - 1$ studies. Even if the p-value of each true individual null hypothesis is uniformly distributed, the p-value for a true PC hypothesis $H^{\gamma/s}$ can be uniform only if there are $\gamma - 1$ non-null individual hypotheses and their p-values are zero (i.e., there is overwhelming evidence that $\gamma - 1$ studies have effect).

In the setting considered in Section 5, the top row of Figure D.1 shows that the distribution of PC null p-values is conservative for $\gamma \geq 2$, and the conservativeness is more severe for larger values of γ: the cumulative distribution function (CDF) is further below the 45-degree diagonal line as γ increases. The bottom row of Figure D.1 shows that the distribution of the conditional PC null p-values is far less conservative, but still conservative.

D.2 Reduced dependence across PC p-values

Since each PC p-value is formed by combining s independent studies, we expect the dependence across PC p-values to be smaller than within individual studies with dependent test statistics. Moreover, since only the top $s - \gamma + 1$ p-values are combined, we expect the dependence to be weaker as γ increases. This is confirmed for the symmetric block correlation setting in Section 5 in Figure D.2. To generate this figure, we restricted ourselves to combining standard normal z-scores within each study into PC p-values, and then transformed the PC p-values into z-scores using the Gaussian quantile function. For two features that have correlation value $\rho \in \{0.1, \ldots, 0.9\}$ within each study (i.e., they are within the same block), we computed the correlation of the PC z-scores. We see that the PC z-scores are less correlated than the Z-scores within each study, and that the correlation decreases as γ increases.

D.3 Asymptotic FDR control when $m \to \infty$

In high dimensional applications, p-values are typically not independent. Storey et al. (2004) assumed the following for asymptotic FDR control on a family of m hypotheses:

$$\forall t \in (0,1] : \lim_{m\to\infty} \frac{V_m(t)}{m_0} = G_0(t) \text{ and } \lim_{m\to\infty} \frac{R_m(t) - V_m(t)}{m - m_0} = G_1(t) \text{ a.s.}, \tag{D.1}$$

where m_0 is the number of true null hypotheses; $V_m(t)$ is the (random) number of true null hypotheses with p-values below t; $R_m(t)$ is the number of (random) p-values below t; and G_0 and G_1 are continuous functions such that

$$\forall t \in (0,1] : 0 < G_0(t) \leq t; \tag{D.2}$$

$$\lim_{m\to\infty} \frac{m_0}{m} = \pi_0 \text{ exists.} \tag{D.3}$$

In their Theorem 6, they prove that if the convergence assumptions in (D.1)–(D.3) hold, then for each $\delta > 0$,

$$\lim_{m\to\infty} \inf_{t\geq\delta} \left\{ \frac{\hat{\pi}_0(\lambda)t}{\{R_m(t) \vee 1\}/m} - \frac{V_m(t)}{R_m(t) \vee 1} \right\} \geq 0$$

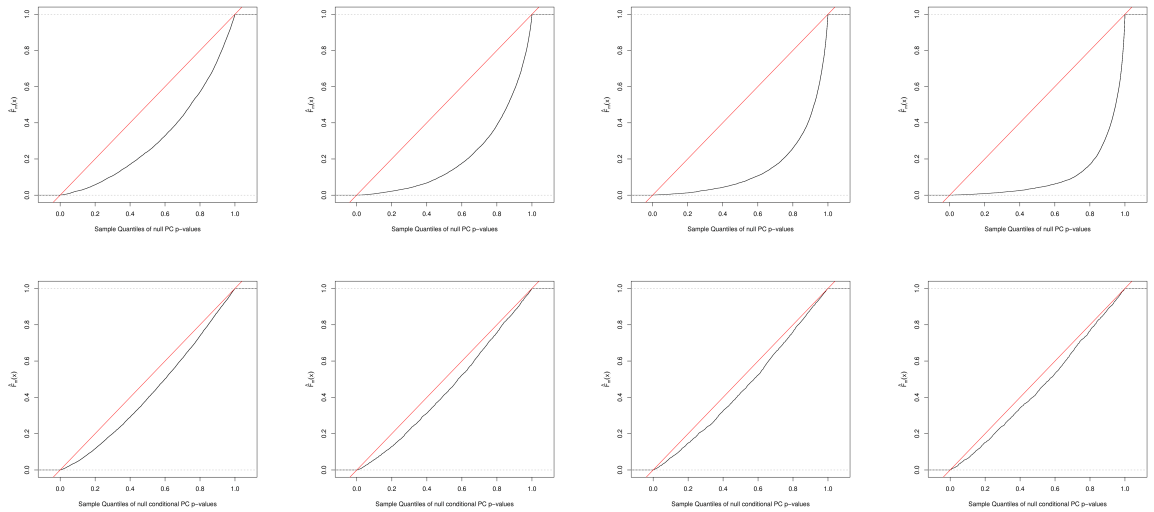


Figure D.1: In the data generation described in Section 4 (symmetric block dependent setting, with block correlation 0.9 in blocks of size 10, with 0.9 true global null hypotheses and 1-0.9-0.002= 0.098 additional true PC null hypotheses distributed evenly in all possible configurations), the estimated cumulative distribution function of the PC null p-values (row 1) and conditional PC null p-values following selection with a pre-specified threshold $\tau = 0.1$. Columns 1 to 4 correspond to γ values 2 to 5, respectively.

holds almost surely, where $\hat{\pi}_0(\lambda) = \frac{m - R_m(\lambda)}{(1-\lambda)m}$ is the plug-in estimate for the number of true null hypotheses. The implication of this result is that asymptotically, the false discovery proportion (FDP) is almost surely at most α if the adaptive BH procedure at level α is applied on the m p-values (so the FDR, which is the expected FDP, is also controlled at level α asymptotically.)

In many high dimensional applications, the dependence within the study is indeed local, thus it is reasonable to assume the convergence assumptions. This is the case in chief genomics applications: for GWAS or eQTLs, the covariance structure is that of a banded matrix; for microarrays or RNA-seq experiments, gene-gene networks are typically sparse (Wang et al., 2021).

We provide similar guarantees of our methodology, when inferring on multiple studies with local dependence. For the family of PC hypotheses, using the adaptive BH procedure on the PC p-values provides asymptotic FDP (and FDR) control if the convergence assumptions (D.1)–(D.3) are satisfied for the data generating the PC p-values. This is guaranteed, for example, if within each study there is a single null distribution from which the p-values from true null hypotheses are generated, and a single nonnull distribution form which the p-values from false null hypotheses are generated, and the limit for the fraction of each of the 2^s combinations of null and non-null hypotheses exists. More specifically, let $\vec{h}$ be the vector of hypotheses status indicators for a feature (so $h_i = 1$ if the ith null hypothesis is true, and zero otherwise, for $i = 1, \ldots, s$). Then if for every $\vec{h}$, the limit for the fraction of features with hypothesis states $\vec{h}$ exists, then convergence assumptions (D.1)–(D.3) are satisfied.

Our algorithm, in which we first select the PC hypotheses which have PC p-values at most the selection threshold τ, and then applies the BH procedure on the conditional PC p-values, also provides asymptotic FDP (and FDR) control. This result follows from the following proposition.

Proposition D.1. *Assume that condition (A1) or (A2) is satisfied for each p-value. Moreover, assume that the convergence assumptions of equations* (D.1)–(D.3) *hold for the PC p-values* $\{p_j^{\gamma/s}, j = 1, \ldots, m\}$. *Then, the FDP of the BH procedure at level* α *on* $\{p_j^{\gamma/s}/\tau, j \in S_\tau\}$, *for a fixed pre-specified* $\tau > 0$, *is asymptotically almost surely at most* α.

Proof. The (random) BH threshold, computed on the set of PC p-values which are considered in step (iv) of our Algorithm 2.1, is

$$\hat{x} = \max\left\{x : \frac{|S_\tau| \times x}{|S_{\tau x}| \vee 1} \leq \alpha\right\}, \tag{D.4}$$

where $|S_{\tau x}| = \sum_{j \in S_\tau} \mathbb{I}\left(\frac{p_j^{\gamma/s}}{\tau} \leq x\right) = \sum_{j=1}^m \mathbb{I}\left(p_j^{\gamma/s} \leq \tau x\right)$ is the number of rejected hypotheses when

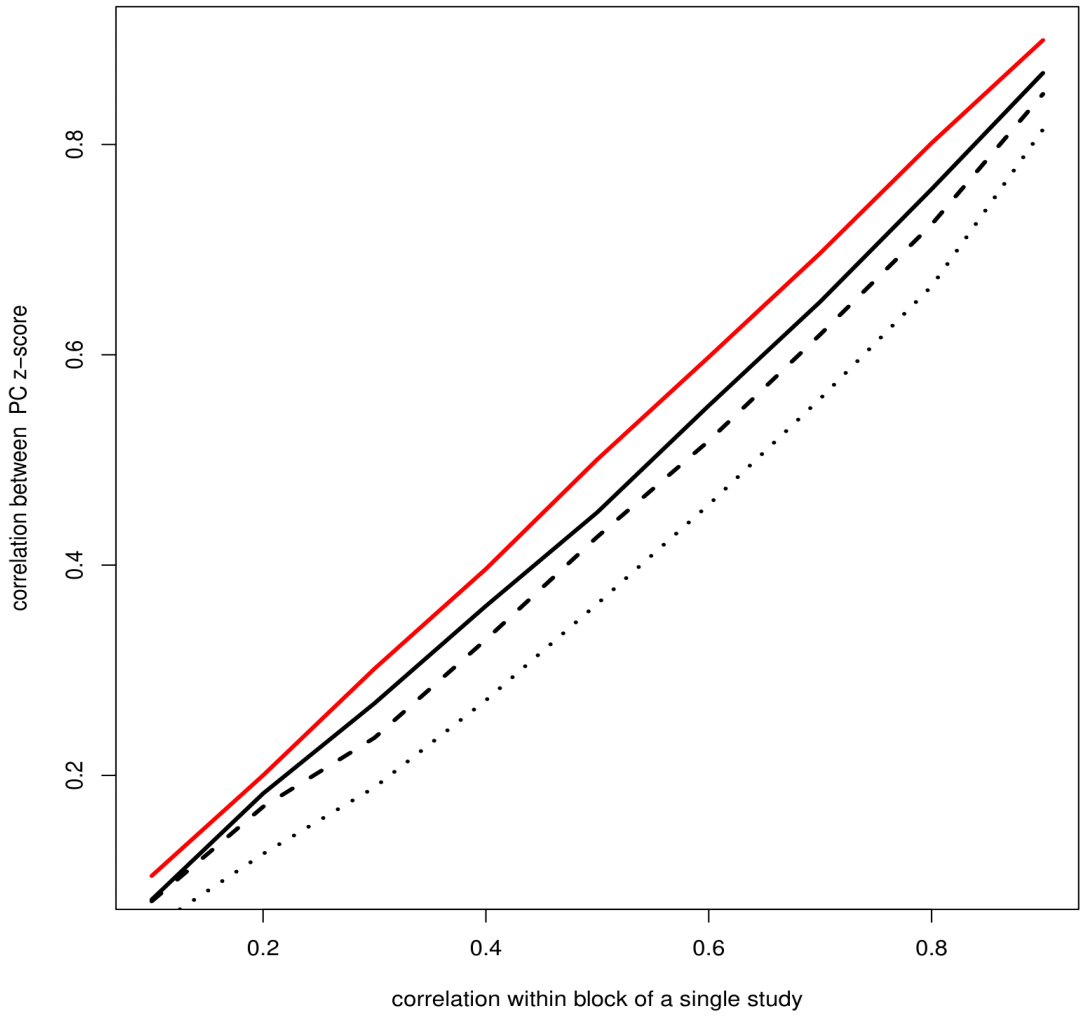


Figure D.2: In the data generation described in Section 4 (symmetric block dependent setting, with blocks of size 10) the correlation between two PC Z-scores from the same block versus the correlation between two PC Z-scores from the same block in the same study for $\gamma = 3$ (solid black), $\gamma = 5$ (dashed black), and $\gamma = 7$ (dotted black). The 45 degrees diagonal line is in red.

thresholding the conditional PC p-values at level x. So it is enough to show that

$$\lim_{m\to\infty} \sup_{0\le x\le 1} \left\{ FDP(\tau x) - \frac{|S_\tau| \times x}{|S_{\tau x}| \vee 1} \right\} \le 0$$

almost surely, where $FDP(\tau x)$ is the false discovery proportion when the hypotheses with PC p-values at most τx are rejected.

Let $\mathcal{H}^{\gamma/s}$ be the set of indices of the true PC null hypotheses (so $|\mathcal{H}^{\gamma/s}| = m_0$). From Theorem 1, it follows that if $H_j^{\gamma/s}$ is true, then $\mathbb{P}(P_j^{\gamma/s}/\tau \le x) \le x\mathbb{P}(P_j^{\gamma/s} \le \tau)$ for all $x \in [0,1]$. Therefore,

$$\begin{aligned}
G_0(\tau x) &= \lim_{m\to\infty} \frac{\sum_{j\in\mathcal{H}^{\gamma/s}} \mathbb{I}(p_j^{\gamma/s} \le \tau x)}{|\mathcal{H}^{\gamma/s}|} \\
&= \mathbb{E}\left(\lim_{m\to\infty} \frac{\sum_{j\in\mathcal{H}^{\gamma/s}} \mathbb{I}(p_j^{\gamma/s} \le \tau x)}{|\mathcal{H}^{\gamma/s}|} \right) \\
&= \lim_{m\to\infty} \mathbb{E}\left(\frac{\sum_{j\in\mathcal{H}^{\gamma/s}} \mathbb{I}(p_j^{\gamma/s} \le \tau x)}{|\mathcal{H}^{\gamma/s}|} \right) = \lim_{m\to\infty} \frac{\sum_{j\in\mathcal{H}^{\gamma/s}} \mathbb{P}(p_j^{\gamma/s} \le \tau x)}{|\mathcal{H}^{\gamma/s}|} \\
&\le \lim_{m\to\infty} \frac{\sum_{j\in\mathcal{H}^{\gamma/s}} \mathbb{P}(p_j^{\gamma/s} \le \tau)x}{|\mathcal{H}^{\gamma/s}|} = G_0(\tau)x. \qquad \text{(D.5)}
\end{aligned}$$

From (a slight modification) of the Glivenko-Cantelli theorem (Storey et al., 2004),

$$\lim_{m\to\infty} \sup_{0\le t\le 1} \left| \frac{|S_t|}{m} - \pi_0 G_0(t) - (1-\pi_0) G_1(t) \right| = 0$$

almost surely. Therefore,

$$\lim_{m\to\infty} \sup_{0\le x\le 1} \left| \frac{|S_\tau| \times x}{|S_{\tau x}| \vee 1} - \frac{\pi_0 G_0(\tau)x + (1-\pi_0) G_1(\tau)x}{(\pi_0 G_0(\tau x) + (1-\pi_0) G_1(\tau x)) \vee 1} \right| = 0 \qquad \text{(D.6)}$$

almost surely.

We use the above results in order to show that the difference between the FDP and its BH point estimate is at most zero. We start by expressing this difference as three differences, and then argue that each of these differences is a.s. upper bounded by zero:

$$\begin{aligned}
&FDP(\tau x) - \frac{|S_\tau| x}{|S_{\tau x}| \vee 1} = \frac{\sum_{j\in\mathcal{H}^{\gamma/s}} \mathbb{I}(p_j^{\gamma/s} \le \tau x) \vee 1}{\sum_{j=1}^m \mathbb{I}(p_j^{\gamma/s} \le \tau x)} - \frac{|S_\tau| x}{|S_{\tau x}| \vee 1} \\
&= \frac{\sum_{j\in\mathcal{H}^{\gamma/s}} \mathbb{I}(p_j^{\gamma/s} \le \tau x)}{\sum_{j=1}^m \mathbb{I}(p_j^{\gamma/s} \le \tau x)} - \frac{\pi_0 G_0(\tau x)}{\pi_0 G_0(\tau x) + (1-\pi_0) G_1(\tau x)} \\
&+ \frac{\pi_0 G_0(\tau x)}{\pi_0 G_0(\tau x) + (1-\pi_0) G_1(\tau x)} - \frac{\pi_0 G_0(\tau) x}{\pi_0 G_0(\tau x) + (1-\pi_0) G_1(\tau x)} \\
&+ \frac{\pi_0 G_0(\tau) x}{\pi_0 G_0(\tau x) + (1-\pi_0) G_1(\tau x)} - \frac{|S_\tau| x}{|S_{\tau x}| \vee 1}
\end{aligned} \qquad \text{(D.7)}$$

Result follows since: from assumptions (D.1)–(D.3), as in Storey et al. (2004), almost surely

$$\lim_{m\to\infty} \sup_{0\le x\le 1} \left| \frac{\sum_{j\in\mathcal{H}^{\gamma/s}} \mathbb{I}(p_j^{\gamma/s} \le \tau x)}{(\sum_{j=1}^m \mathbb{I}(p_j^{\gamma/s} \le \tau x)) \vee 1} - \frac{\pi_0 G_0(\tau x)}{(\pi_0 G_0(\tau x) + (1-\pi_0) G_1(\tau x)) \vee 1} \right| = 0;$$

from (D.5),

$$\frac{\pi_0 G_0(\tau x)}{(\pi_0 G_0(\tau x) + (1-\pi_0) G_1(\tau x)) \vee 1} - \frac{\pi_0 G_0(\tau) x}{(\pi_0 G_0(\tau x) + (1-\pi_0) G_1(\tau x)) \vee 1} \le 0;$$

and from (D.6),

$$\lim_{m\to\infty} \sup_{0\le x\le 1} \left\{ \frac{\pi_0 G_0(\tau) x}{(\pi_0 G_0(\tau x) + (1-\pi_0) G_1(\tau x)) \vee 1} - \frac{|S_\tau| x}{|S_{\tau x}| \vee 1} \right\} \le 0.$$

□

Our next goal is to generalize Proposition D.1 to cases in which the value of τ is selected on the basis of the available data. Thus, in Proposition D.2 below we consider a random variable $\hat{\tau}$ which describes the selection rule, meaning that $\tau = \hat{\tau}(\text{data})$.

Proposition D.2. *Assume that condition (A1) or (A2) is satisfied for each p-value. Moreover, assume that the convergence assumptions of equations* (D.1)–(D.3) *hold for the PC p-values* $\{p_j^{\gamma/s}, j = 1, \ldots, m\}$.

For any given value $\tau \in (0, 1]$*, define the (random) function* $\hat{F}_\tau : [0, 1] \to [0, 1]$ *by* $\hat{F}_\tau(x) = |S_{\tau x}|/|S_\tau|$. *Let* $\hat{\tau} \equiv \hat{\tau}_m$ *denote a* $(0, 1]$*-valued random variable which is measurable with respect to (the* σ*-field generated by) the available (random) data. Assume that the sequence* $\{\hat{\tau}_m\}_{m\geq 1}$ *possesses an almost sure limiting value* $\tau_\infty \in (0, 1]$*, and that* $||\hat{F}_{\hat{\tau}_m} - \hat{F}_{\tau_\infty}||_\infty \to 0$ *almost surely as* $m \to \infty$.

Then, the FDP of the BH procedure at level α *on* $\{p_j^{\gamma/s}/\hat{\tau}_m, j \in S_{\hat{\tau}_m}\}$ *is asymptotically almost surely at most* α.

Proof. Consider the following representation of the BH threshold $\hat{x} \equiv \hat{x}(\tau)$ for a given τ (and a given α):

$$\begin{aligned} \hat{x}(\tau) &= \max\left\{x \in (0, 1] : \hat{F}_\tau(x) \geq \frac{x}{\alpha}\right\}, \\ &= \max\left\{x \in (0, 1] : \frac{x}{\hat{F}_\tau(x)} \leq \alpha\right\}, \end{aligned} \tag{D.8}$$

if the maximum in (D.8) exists, and $\hat{x}(\tau) = 0$ otherwise (see, e. g., Lemma 5.7 in Dickhaus (2014)). The graph of the function $x \mapsto x/\alpha$ appearing in (D.8) is occasionally referred to as the "Simes line". Proposition D.1 then yields (under the stated assumptions), that choosing $\hat{x}(\tau)$ as the rejection threshold for the conditional PC p-values leads to an FDP which is asymptotically almost surely upper-bounded by α for any fixed τ. In particular, considering $\tau = \tau_\infty$ in (D.8) leads to an FDP which is asymptotically almost surely upper-bounded by α, because τ_∞ is a fixed constant in the interval $(0, 1]$, by assumption.

Analogously, the BH threshold for a random (selected) value $\hat{\tau}_m$ is given by

$$\hat{x}(\hat{\tau}_m) = \max\left\{x \in (0, 1] : \frac{x}{\hat{F}_{\hat{\tau}_m}(x)} \leq \alpha\right\}, \tag{D.9}$$

if the maximum in (D.9) exists, and $\hat{x}(\hat{\tau}_m) = 0$ otherwise. Clearly, this representation implies that $\hat{x}(\hat{\tau}_m)$ depends on the data only via $\hat{F}_{\hat{\tau}_m}$, as soon as $\hat{\tau}_m$ has been chosen. By our assumptions, $\hat{\tau}_m$ converges almost surely to τ_∞ and $|\hat{F}_{\hat{\tau}_m} - \hat{F}_{\tau_\infty}|$ converges uniformly and almost surely to zero. Furthermore, the mapping $\hat{F}_\tau \mapsto \hat{x}(\tau)$ is continuous. From these assertions, we conclude that $|\hat{x}(\hat{\tau}_m) - \hat{x}(\tau_\infty)|$ converges to zero almost surely as m tends to infinity. However, as argued before, choosing $\hat{x}(\tau_\infty)$ as the rejection threshold for the conditional PC p-values leads to an FDP which is asymptotically almost surely upper-bounded by α, which yields the assertion of the proposition. □

The proof of Proposition 3.3 follows from Proposition D.2 since $\hat{\tau}_m$ depends on the data only via $\{|S_{\tau_i}|/m, \hat{F}_{\tau_i}, i = 1, \ldots, K\}$. Since $|S_{\tau_i}|/m$ and $\hat{F}_{\tau_i}$ converge almost surely to well defined limiting functions for $i = 1, \ldots, K$, there exists a limiting value $\tau_\infty \in \{\tau_1, \ldots, \tau_K\}$ that $\hat{\tau}_m$ converges to almost surely, and $||\hat{F}_{\hat{\tau}_m} - \hat{F}_{\tau_\infty}||_\infty \to 0$ almost surely as $m \to \infty$.

D.4 Additional simulations

D.4.1 Results using Stouffer and Simes combination p-values

In the setting considered in Section 5, with $m = 20000$ and $\pi_1 = 0.002$, we applied the novel procedures using Stouffer and Simes combination p-values. Tables D.1 and D.2 show the average number of true discoveries (our measure of power) and FDP for the novel procedures, using the selection threshold prespecified as $\tau = 0.1$ or adaptively chosen as in Zhao et al. (2019) with $\beta = 0.1, 0.5$. Compared with the unconditional approach of applying BH on the PC p-values, we see that the power is greater with the novel approach, for every γ. The power advantage over the unconditional approach is very large, especially when the p-values are combined using Simes method. Combining p-values using Stouffer is more powerful than using Simes. However, a comparison with Table 1 in the main manuscript shows that the power is highest using the Fisher combining method. Although Fisher combining has better power properties than Stouffer and Simes in many data generation settings, this is not always true: Wang and Owen (2019) showed that any combining function increasing in its coordinates, that uses the $s - \gamma + 1$ largest p-values to provide a valid PC p-value, is admissible.

From Table D.2 it is clear that the conditional and unconditional procedures are below the nominal 0.05 level, that the unconditional approach is the most conservative (i.e., with lowest FDR level), and that using Simes (rather than Stouffer or Fisher) for combining is most conservative.

Table D.1: In the symmetric block dependent setting, the average number of true discoveries for $\gamma = 2, 3, 4, 5$ using the Stouffer combining method (rows 1–4) and the Simes combining method (rows 5–8), for the following procedures at level 0.05: BH on PC p-values, BH and adaptive BH (denoted aBH) on conditional PC p-values using selection threshold $\tau = 0.1$, adaptive threshold at $\beta = 0.1$ and adaptive threshold at $\beta = 0.5$. Based on 5000 repetitions.

					conditional			
			τ=0.1		$\hat{\tau}$ with $\beta = 0.1$		$\hat{\tau}$ with $\beta = 0.5$	
Combining method	γ	BH	BH	aBH	BH	aBH	BH	aBH
Stouffer	2	17.7	20.0	19.7	19.9	19.7	20.2	20.1
	3	9.0	12.7	12.4	12.6	12.4	12.8	12.9
	4	4.0	7.4	7.4	7.1	7.0	7.4	7.5
	5	1.6	4.1	4.2	3.8	3.7	4.0	4.0
Simes	2	6.3	17.9	17.3	18.1	18.3	18.5	19.2
	3	0.5	7.2	7.8	6.7	6.5	6.9	7.6
	4	0.0	1.5	1.7	1.3	1.1	1.4	1.4
	5	0.0	0.3	0.4	0.3	0.2	0.3	0.3

Table D.2: In the symmetric block dependent setting, the average FDP (estimated FDR) for $\gamma = 2, 3, 4, 5$, using the Stouffer combining method (rows 1–4) and the Simes combining method (rows 5–8), for the following procedures at level 0.05: BH on PC p-values, BH and adaptive BH (denoted aBH) on conditional PC p-values using selection threshold $\tau = 0.1$, adaptive threshold at $\beta = 0.1$ and adaptive threshold at $\beta = 0.5$. Based on 5000 repetitions.

					conditional			
			τ=0.1		$\hat{\tau}$ with $\beta = 0.1$		$\hat{\tau}$ with $\beta = 0.5$	
Combining method	γ	BH	BH	aBH	BH	aBH	BH	aBH
Stouffer	2	0.005	0.016	0.014	0.015	0.014	0.017	0.017
	3	0.003	0.017	0.016	0.017	0.016	0.019	0.020
	4	0.001	0.020	0.020	0.017	0.016	0.019	0.021
	5	0.000	0.016	0.017	0.013	0.012	0.015	0.016
Simes	2	0.001	0.011	0.010	0.012	0.012	0.012	0.014
	3	0.000	0.006	0.007	0.005	0.005	0.006	0.007
	4	0.000	0.001	0.002	0.001	0.001	0.001	0.001
	5	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table D.3: In the symmetric block dependent setting with block size 100, the average number of true discoveries for $\gamma = 2, 3, 4, 5$ for the following procedures at level 0.05: adaFilter by Wang et al. (2021), BH on PC p-values, BH and adaptive BH (denoted aBH) on conditional PC p-values using selection threshold $\tau = 0.1$, adaptive threshold at $\beta = 0.1$ and adaptive threshold at $\beta = 0.5$. Based on 5000 repetitions.

				conditional					
				τ=0.1		$\hat{\tau}$ with $\beta = 0.1$		$\hat{\tau}$ with $\beta = 0.5$	
π_1	γ	adaFilter	BH	BH	aBH	BH	aBH	BH	aBH
0.002	2	25.4	25.8	28.6	28.4	26.8	26.3	27.5	27.1
	3	16.9	14.4	20.0	20.1	17.7	16.8	18.6	18.0
	4	10.5	6.2	12.0	12.3	10.0	9.2	10.8	10.2
	5	6.9	2.0	6.3	6.5	4.9	4.3	5.4	4.9
0.02	2	338.7	302.4	320.1	328.4	312.3	307.4	316.7	314.3
	3	281.5	202.4	242.8	269.7	238.7	231.0	245.5	244.1
	4	218.1	110.1	160.3	196.1	165.7	158.7	172.1	172.3
	5	168.7	46.7	91.8	123.7	104.6	100.0	109.7	112.5

Table D.4: In the symmetric block dependent setting with block size 100, the average FDP (estimated FDR) for $\gamma = 2, 3, 4, 5$ for the following procedures at level 0.05: adaFilter by Wang et al. (2021), BH on PC p-values, BH and adaptive BH (denoted aBH) on conditional PC p-values using selection threshold $\tau = 0.1$, adaptive threshold at $\beta = 0.1$ and adaptive threshold at $\beta = 0.5$. Based on 5000 repetitions.

				conditional					
				τ=0.1		$\hat{\tau}$ with $\beta = 0.1$		$\hat{\tau}$ with $\beta = 0.5$	
π_1	γ	adaFilter	BH	BH	aBH	BH	aBH	BH	aBH
0.002	2	0.046	0.003	0.012	0.013	0.006	0.004	0.008	0.007
	3	0.044	0.001	0.014	0.015	0.005	0.004	0.009	0.007
	4	0.043	0.001	0.013	0.015	0.005	0.004	0.008	0.006
	5	0.043	0.000	0.010	0.011	0.004	0.003	0.006	0.005
0.02	2	0.034	0.004	0.010	0.017	0.007	0.005	0.009	0.008
	3	0.031	0.002	0.007	0.019	0.007	0.005	0.008	0.008
	4	0.027	0.001	0.006	0.016	0.007	0.006	0.008	0.008
	5	0.022	0.000	0.003	0.009	0.005	0.004	0.006	0.007

D.4.2 Results with strong dependence among the p-values

In the main manuscript we showed results for weak dependence using a dependence block size of 10, and 2000 independent blocks, for a total of $m = 20000$ PC null hypotheses. In this section we alter the block size to 100, keeping all other configurations of the data generation the same. This is a setting with strong dependence, since the blocks are large and we only have 200 independent blocks, for a total of $m = 20000$ PC null hypotheses.

Tables D.3 and D.4 show the average number of true discoveries and FDP for the novel procedures, using the selection threshold pre-specified as $\tau = 0.1$ or adaptively chosen as in Zhao et al. (2019) with $\beta = 0.1, 0.5$, versus competitors. The average number of true discoveries and FDP are remarkably similar to the averages in Tables 1 and 2 in the main manuscript for adaFilter, the unconditional approach, and the conditional approach with $\tau = 0.1$. However, the power of the novel procedures with the selection threshold adaptively chosen have lower power compared with the weak dependence setting in the main manuscript. Table D.5 shows the reason for the power deterioration: when the dependence is stronger, the estimated selection threshold tends to be higher than for weak dependence.

D.4.3 Results using the adaptive method of Hoang and Dickhaus (2022, 2021b) for estimating π_0

Storey's estimator (Storey et al., 2004) for the fraction of null hypotheses tends to be conservative if the p-values from null hypotheses have a distribution that is stochastically larger than uniform. Therefore, Hoang and Dickhaus (2022, 2021b) suggested to randomize the p-values first, and then apply Storey's estimation method on the vector of randomized (rather than original) p-values.

The method of Hoang and Dickhaus (2022, 2021b) can provide a better estimate of the fraction of PC null hypotheses among the selected than Storey's estimator (Storey et al., 2004), in settings where the conditional PC p-values are still stochastically larger than uniform. We set out to examine this in

Table D.5: In the symmetric block dependent setting with block size 100,, the average adaptively selected τ using the method of Zhao et al. (2019) with $\beta = 0.1, 0.5$.

π_1	γ	$\hat{\tau}$ with $\beta = 0.1$	$\hat{\tau}$ with $\beta = 0.5$
0.002	2	0.63	0.40
	3	0.52	0.35
	4	0.54	0.41
	5	0.57	0.46
0.02	2	0.64	0.43
	3	0.55	0.40
	4	0.56	0.45
	5	0.59	0.50

our simulation settings. Both methods were implemented using the fixed parameter $\lambda = 0.5$. Specifically, Storey's estimate is $2 \times \left(\sum_{i \in S_\tau} \mathbb{I}\left(\frac{p_i^{\gamma/s}}{\tau} > 0.5 \right) + 1 \right)$; the estimator of Hoang and Dickhaus (2022, 2021b) is similar, except $\frac{p_i^{\gamma/s}}{\tau}$ is replaced by

$$U\mathbb{I}\left\{ \frac{P_i^{\gamma/s}}{\tau} > 0.5 \right\} + 2 \times \frac{P_i^{\gamma/s}}{\tau} \mathbb{I}\left\{ \frac{P_i^{\gamma/s}}{\tau} \leq 0.5 \right\},$$

where U is a Uni$[0, 1]$ random variable.

Figure D.3 shows that both estimation methods are conservative, and tend to overestimate the fraction of PC null hypotheses among the selected. However, for $\gamma = 2, 3$ the method of Hoang and Dickhaus (2022, 2021b) is far less conservative. For $\gamma = 4, 5$, there is little difference between the methods. Table D.6 shows the gain in power from using each method.

Table D.6: In the symmetric block dependent setting with $\pi_1 = 0.02$, for $\gamma = 2, 3, 4, 5$, the average number of true rejections using our conditional approach with selection threshold $\tau = 0.25$, when the multiple testing procedure on the conditional PC p-values is: BH (column 4); adaptive BH using Storey's plug-in estimate (column 5); and adaptive BH using the method of Hoang and Dickhaus (2022, 2021b) (column 6). Columns 2 and 3 provide the average value for the estimated fraction of PC nulls among the selected, using Storey's plug-in estimate and the method of Hoang and Dickhaus (2022, 2021b), respectively. Based on 5000 simulations.

γ	Storey's $\hat{\pi}_0$	Hoang and Dickhaus's $\hat{\pi}_0$	BH	adaptive BH Storey's $\hat{\pi}_0$	adaptive BH Hoang and Dickhaus's $\hat{\pi}_0$
2	1.02	0.88	320.30	319.88	322.75
3	0.85	0.73	250.38	255.69	260.89
4	0.68	0.65	176.03	192.79	195.01
5	0.60	0.64	109.62	133.62	130.54

D.5 A different approach of choosing τ adaptively

As mentioned in Section 3.1, one can consider any function $G(\tau) = H(|S_\tau|, \tau)$, where H is increasing in $|S_\tau|$ and decreasing in τ, and minimize G. The idea is always to minimize the number $|S_\tau|$ of non-discarded p-values while trying to keep the penalty $1/\tau$ small. For example, minimize $G(\tau) = |S_\tau| - km\tau$, where m is the number of p-values, and $k \in (0, \infty)$ is a pre-specified constant.

Since $|S_\tau| = m\hat{F}_m(\tau)$, we approximate $G'(\tau)$ with

$$m \left[\frac{\hat{F}_m(\tau + \omega) - \hat{F}_m(\tau)}{\omega} - k \right],$$

where $\hat{F}_m$ is the ecdf of the p-values, and $\omega > 0$ is a small pre-specified constant.

Given a sequence of τ's, $\tau_1 < \ldots < \tau_K$, we go from τ_i to τ_{i-1}, starting with τ_K, if there is sufficient evidence that $G'(\tau_i) > 0$. We can compare $\hat{F}_m(\tau_i + \omega) - \hat{F}_m(\tau_i)$ and $k\omega$ directly or we can try to reject $q > k$, where $m(\hat{F}_m(\tau_i + \omega) - \hat{F}_m(\tau_i)) \sim$ Binomial$(m, q\omega)$ at a pre-specified significance level β (cf. Zhao et al. (2019)).

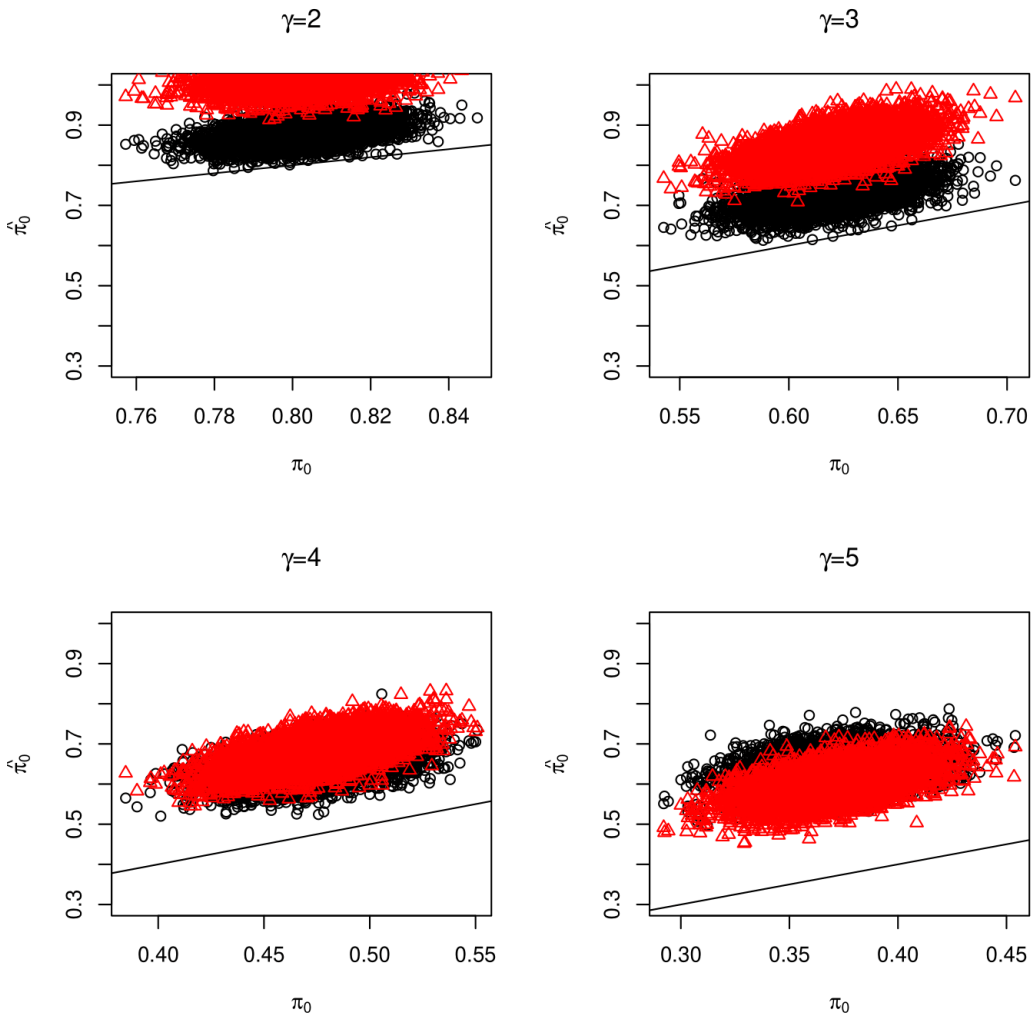


Figure D.3: The estimated fraction of PC null hypotheses, using the plug-in method of Storey et al. (2004) (red triangles) and using the method of Hoang and Dickhaus (2022, 2021b) (black circles), versus the true fraction of PC null hypotheses among the selected. Plotted are the results from 5000 data generations from the symmetric block dependence described in Section 5, with $\pi_1 = 0.02$. The threshold for selection was $\tau = 0.25$. The black line is the 45 degree diagonal line. Each panel is a different γ.

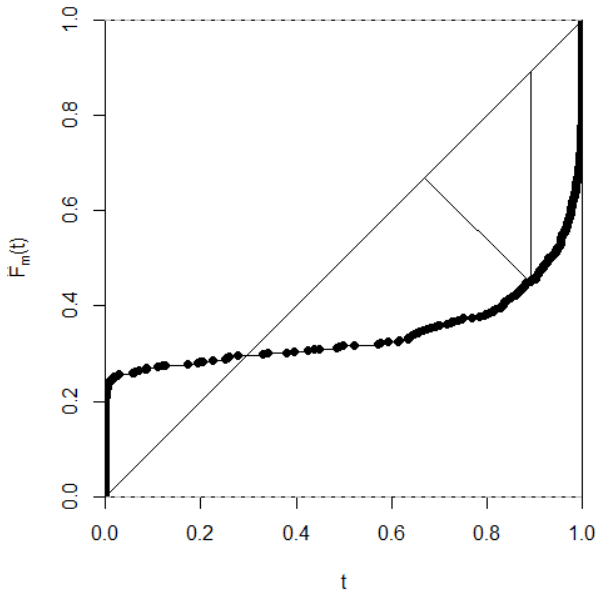


Figure D.4: Maximizing the distance (below the identity line) between the identity line and the ecdf $\hat{F}_m$ of the p-values. The length of the vertical line is $-G$.

Graphically, if $k = 1$, minimizing G is maximizing the vertical distance between the identity line and the ecdf $\hat{F}_m$ (below the identity line), which is equivalent to maximizing the distance between the identity line and the ecdf $\hat{F}_m$, see Figure $D.4$.

D.6 Selection before combination

As described in Algorithm 2.1, we select in step (ii) from the p-values $P_j^{\gamma/s}$, $j = 1, \ldots, m$, i.e. after combining the base p-values $P_{1,j}, \ldots, P_{s,j}$. According to Theorem 1, if the latter fulfil either $(A1)$ or $(A2)$, the partial conjunction p-value $P_j^{\gamma/s}$ is uniformly valid. However, both $(A1)$ and $(A2)$ mean that $P_{1,j}, \ldots, P_{s,j}$ are uniformly valid themselves. Thus, the conditional p-values $(P_{i,j}/\tau_{i,j}, \mid P_{i,j} \leq \tau_{i,j})_{i=1,\ldots,s}$, are valid for $H_{j,1}, \ldots, H_{j,s}$, respectively, for any $\tau_{i,j} \in [0, 1]$, $i = 1, \ldots, s$. For the sake of simplicity we set $\tau_{i,j} = \tau^{(1)}$ for all i, j.

Furthermore, it is easy to see that $(P_{i,j}/\tau_{i,j}, \mid P_{i,j} \leq \tau_{i,j})_{i=1,\ldots,s}$ satisfy $(A1)$ or $(A2)$ if $P_{1,j}, \ldots, P_{s,j}$ do, respectively. Therefore, one can condition again, after combining $(P_{i,j}/\tau_{i,j}, \mid P_{i,j} \leq \tau_{i,j})_{i=1,\ldots,s}$ for the partial conjunction null hypothesis $H_j^{\gamma/S_{j,\tau^{(1)}}}$, where $S_{j,\tau^{(1)}} = \#\{P_{i,j}/\tau^{(1)}, i = 1, \ldots, s \mid P_{i,j} \leq \tau^{(1)}\}$. In case of $\gamma > S_{j,\tau_{(1)}}$, we retain $H_j^{\gamma/s}$.

There are two sources of conservativity when dealing with partial conjunction p-values. The base, null p-values $P_{1,j}, \ldots, P_{s,j}$ can be stochastically larger than Uni$[0, 1]$, or the partial conjunction p-value $P_j^{\gamma/s}$ can be conservative due to the nature of partial conjunction null hypotheses. In the paper we deal with the second kind directly and with the first kind only indirectly.

For this section, we call $\tau_1 = \cdots = \tau_m = \tau^{(2)}$ the selection parameter applied after combining (this was called τ in the paper). For fixed $\tau^{(1)}, \tau^{(2)}$, let $P_j^{0,0} = P_j^{\gamma/s}$, $P_j^{1,0}$, $P_j^{0,1}$, $P_j^{1,1}$, be the unconditional p-values, the selected p-values after conditioning with $\tau^{(1)}$ only before combining, the selected p-values after conditioning with $\tau^{(2)}$ only after combining, and the selected p-values after conditioning with $\tau^{(1)}$ and $\tau^{(2)}$ before and after combining, respectively. The index sets $S_{0,0}, S_{1,0}, S_{0,1}, S_{1,1}$ comprise of the selected indices $j \in \{1, \ldots, m\}$. In the paper, we only consider $P_j^{0,0}$ and $P_j^{0,1}$, where $S_{0,0} = \{1, \ldots, m\}$ and $S_{0,1} = S_{\tau^{(2)}}$.

For our simulations, we set $\tau^{(1)} = 0.7$ and $\tau^{(2)} = 0.5$. For our analysis, we consider Model 1 from Section 4, where $\theta^\star_{i,j} = 0$ and $n_{i,j} = 50$ for all i, j. We consider $s = 6, \gamma = 3$ for a fixed endpoint j, parameter values $\boldsymbol{\theta}_j = (\theta_{1,j}, \ldots, \theta_{s,j}) \in H_j^{\gamma/s}$, and $P_j^{\gamma/s}$ the Fisher combination. Via Monte-Carlo simulations with 100,000 repetitions, we approximate the probability for j to be selected, $j \in S_{0,0}, S_{1,0}, S_{0,1}, S_{1,1}$,

Table D.7: The considered parameter values $\boldsymbol{\theta}_j$ in the simulations.

Setting	$\theta_{1,j}\sqrt{n_{1,j}}$	$\theta_{2,j}\sqrt{n_{1,j}}$	$\theta_{3,j}\sqrt{n_{1,j}}$	$\theta_{4,j}\sqrt{n_{1,j}}$	$\theta_{5,j}\sqrt{n_{1,j}}$	$\theta_{6,j}\sqrt{n_{1,j}}$
(0,0)	0	0	0	0	-5	-5
(1,0)	1	1	0	0	-5	-5
(0,1)	0	0	0	0	0	-5
(1,1)	1	1	1	0	0	-5

and, if selected, approximate the cdfs of $P_j^{0,0}$, $P_j^{1,0}$, $P_j^{0,1}$, $P_j^{1,1}$. We consider the parameter values $\boldsymbol{\theta}_j$ as described in Table $D.7$. Under the (LFC-) parameter value $\boldsymbol{\theta}_j = (0,0,0,0,-\infty,-\infty)$, $P_j^{3/6}$ is uniformly distributed, so $P_j^{3/6}$ is close to Uni$[0,1]$ under Setting (0,0). Under Setting (1,0), the conservativity of $P_j^{3/6}$ comes from the conservative base p-values $P_{1,j}$ and $P_{2,j}$, whereas under Setting (0,1), $P_j^{3/6}$ is conservative due to the number of base, null p-values. In the last setting both sources of conservativity are present. The results are in Figure $D.5$.

Between $P_j^{1,0}$ and $P_j^{0,1}$, the first works better in Setting (1,0) and the second in Setting (0,1). The unconditional p-value $P_j^{0,0}$ produces the most conservative p-values in all settings, and the p-value $P_j^{1,1}$ is left with the least conservative p-values, after selection, in almost each setting. The probability of selecting j is lowest for $P_j^{0,1}$ and $P_j^{1,1}$. Note that the discarding event $j \notin S_{1,0}$ happens if and only if $\gamma > S_{j,\tau^{(1)}}$.

D.7 Regarding the combination p-value $P^{\gamma/s}$

We assumed that the combination function $P^{\gamma/s}(P_1,\dots,P_s)$ is increasing in each p-value $P_1,\dots,P_s$ and valid for $H^{\gamma/s}$. Since it holds $H_s^1 \subseteq H_s^2 \subseteq \cdots \subseteq H_s^s$, valid p-values for H_s^γ need not be valid for $H_s^{\gamma'}$, $\gamma < \gamma'$. However, valid p-values for H_s^γ can be easily derived from (increasing) combination functions for $H_{s-\gamma+1}^1$ by essentially combining the $s-\gamma+1$ largest p-values, as shown by Benjamini and Heller (2008). For instance, the Fisher p-value

$$P^{1/(s-\gamma+1)}(P_1,\dots,P_{s-\gamma+1}) = 1 - \Phi\left(\frac{1}{\sqrt{s-\gamma+1}}\sum_{i=1}^{s-\gamma+1}\Phi^{-1}\left(1-P_i\right)\right)$$

which is valid for $H^{1/(s-\gamma+1)}$ implies that

$$P^{\gamma/s}(P_1,\dots,P_s) = 1 - \Phi\left(\frac{1}{\sqrt{s-\gamma+1}}\sum_{i=\gamma}^{s}\Phi^{-1}\left(1-P_{(i)}\right)\right)$$

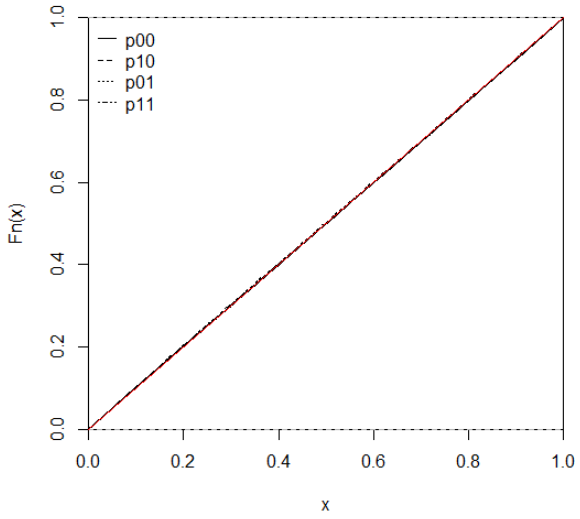
is also valid for $H^{\gamma/s}$.

Furthermore, we assumed that $P^{\gamma/s}$ is uniform under the LFC $\pi(0,\dots,0,U_1,\dots,U_{s-\gamma+1})$ in $H^{\gamma/s}$, where π is any permutation vector of s elements. This assumption is sufficient (together with $(A1)$ or $(A2)$) to imply that $P^{\gamma/s}(P_1,\dots,P_s)$ is a uniformly valid p-value for $H^{\gamma/s}$, as claimed in Theorem 1. This is, for example, fulfilled if $P^{\gamma/s}$ is derived from a combination p-value $P^{1/(s-\gamma+1)}$, as described above, that is uniform under the LFC $(U_1,\dots,U_{s-\gamma+1})$. As mentioned in Remark 2, the assumption is not necessary, and

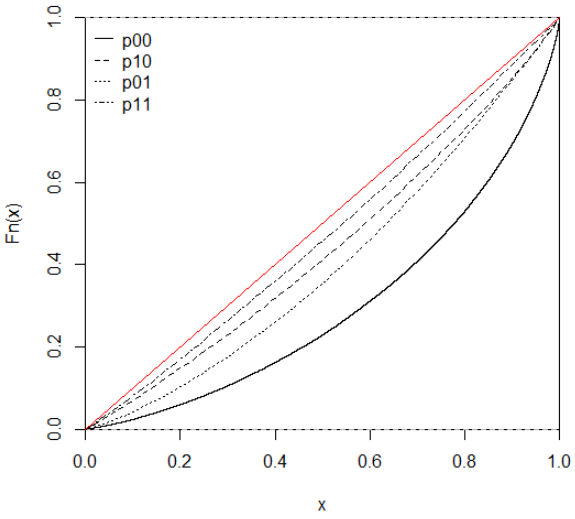
$$\text{Uni}[0,1] \leq_{\text{rh}} P^{\gamma/s}(\pi(0,\dots,0,U_1,\dots,U_{s-\gamma+1}))$$

is also sufficient.

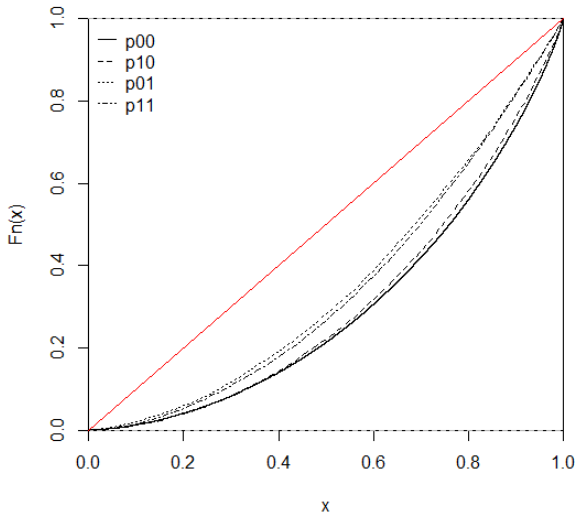
We give an example of a valid combination p-values $P^{\gamma/s}$ that fulfils neither condition, and one that only fulfils the less strict one. We choose any s and γ, so that $s-\gamma+1 = 10$ and consider the harmonic mean and the arithmetic mean, multiplied with constants such, that they are valid for $H^{\gamma/s}$. We approximated their cdfs under an LFC in $H^{\gamma/s}$ with a Monte-Carlo Simulation with 100,000 repetitions, see Figure $D.6$. Graphically, for a given parameter in the null, one can determine if a p-value is valid, if its cdf under that parameter is always below the identity line. It is uniformly valid, if the line that connects $(0,0)$ with the point $(\tau, F(\tau))$ is always above the cdf F, for all τ. If a p-value is valid under an LFC in the null, then it is valid everywhere in the null, thus the two p-values in Figure $D.6$ are indeed valid for $H^{\gamma/s}$. However, we can see that the harmonic mean is not uniformly valid under the LFC whereas the arithmetic mean is. This means that the first, used with base p-values that fulfil $(A1)$ or $(A2)$, is not necessarily uniformly valid, but the second is.



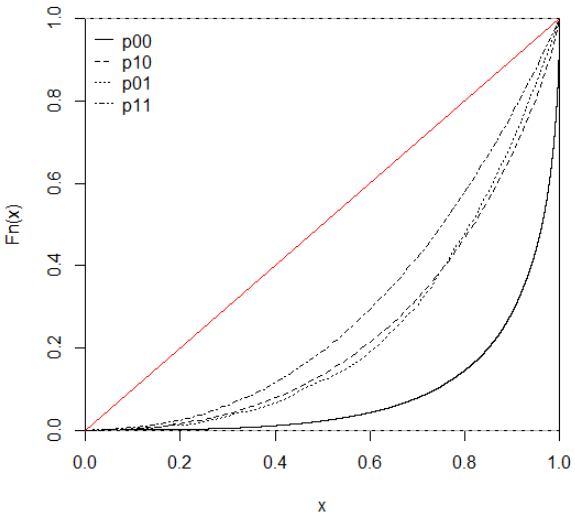
(a) Setting (0,0), no source of conservativity



(b) Setting (1,0), base p-values are conservative

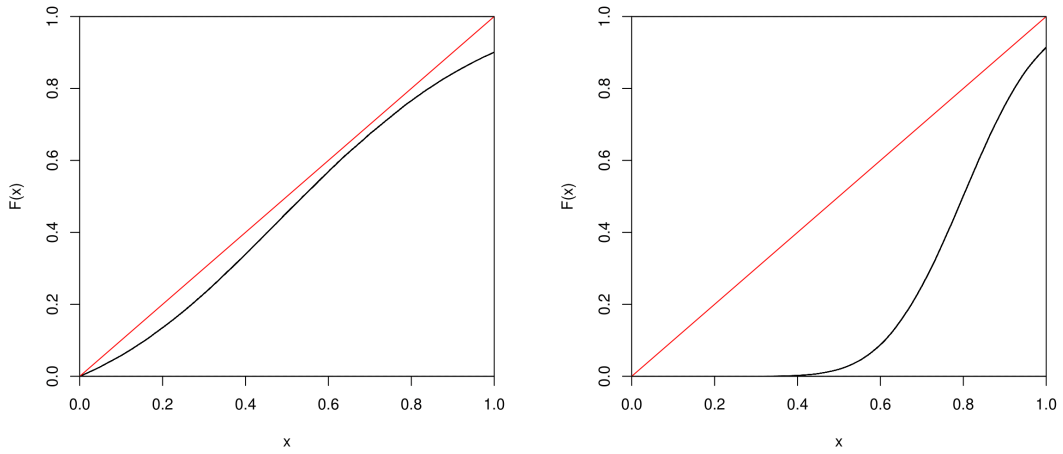


(c) Setting (0,1), more than $s - \gamma + 1$ null p-values



(d) Setting (1,1), both sources of conservativity

Figure D.5: Approximations of the cdfs of the (null) p-values $P^{0,0}$, $P^{1,0}$, $P^{0,1}$, $P^{1,1}$, given that the p-value is selected under different settings with different sources of conservativity.



(a) The cdf of the (adjusted) harmonic mean under an LFC in $H^{\gamma/s}$.

(b) The cdf of the (adjusted) arithmetic mean under an LFC in $H^{\gamma/s}$.

Figure D.6: The cdfs of the (adjusted) harmonic mean and the (adjusted) arithmetic mean under the LFC $\pi(0,\ldots,0,U_1,\ldots,U_{10})$ in $H^{\gamma/(\gamma+9)}$, approximated by a Monte-Carlo simulation with 100,000 repetitions.