On change-point perspectives in multiple testing and weak signal inference

Anica Kostic

Department of Statistics

London School of Economics and Political Science

This dissertation is submitted for the degree of

Doctor of Philosophy

December 2022

To my parents

Declaration

I certify that the thesis I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it). The copyright of this thesis rests with the author. Quotation from it is permitted, provided that full acknowledgement is made. This thesis may not be reproduced without my prior written consent. I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party. I declare that my thesis consists of 33140 words.

Anica Kostic

December 2022

Acknowledgements

First and foremost, I would like to thank my supervisor, Professor Piotr Fryzlewicz, for accepting me as one of his students, and for his endless patience and continuous guidance during the past four years. I am extremely grateful for his constant support, invaluable feedback and for giving me the freedom and encouraging me to pursue my curiosity and ideas. It was a great privilege working under his supervision and learning from him.

I am also very grateful for the financial support from the London School of Economics Statistics PhD Scholarship. Thanks to all the staff in the Department of Statistics at the London School of Economics and especially to Penny Montague for her patience and for taking good care of the Statistics PhD students.

I would like to thank my parents and my sister for being good role models, for their unconditional love and support and for always encouraging me to pursue education. I also extend my appreciation to all of my friends from Belgrade, and especially Ana Arsenijević and Dejana Čolak who are also about the finish their PhD journeys. I would like to thank my professors, teachers and former colleagues at the Faculty of Mathematics at the University of Belgrade. I am deeply grateful for the quality education that I received and for the opportunity to work there as a graduate teaching assistant.

Finally I would like to thank my friends and office-mates on the seventh floor, JingHan Tee, José Manuel Pedraza-Ramírez, Gianluca Giudice, Sahoko Ishida, Davide

de Santis and KaiFang Zhou. My thanks extend to Shakeel Gavioli-Akilagun, Camilo Cárdenas-Hurtado and Eduardo Ferioli-Gomes who are always fun to be around. Although a large part of my PhD experience was marked by the pandemic, online meetings and a rather empty Columbia House, I am grateful for the time spent with my colleagues. My special thanks go to Sahoko Ishida for being a true friend and to Gabriel Wallin. I greatly appreciate their support during this last year.

Abstract

This thesis studies the problem of multiple testing from the change-point perspective, and the problem of inference in the Gaussian sequence model.

In the first part of the thesis, we propose a method for estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. The idea is to consider the sequence of sorted p-values as an approximately piecewise linear function with one change-point in slope. We propose an estimator for this change point, which can be further used in combination with Storey's family of estimators to get the estimator of the false null proportion. Our proposed estimator is conservative and we provide consistency results using the tools from the theory of quantile processes. We compare our estimator to various others proposed in the literature through simulations.

Secondly, building on the ideas from the first part, we consider possible applications of some recent multiple change-point methods in multiple testing. We propose to use algorithms for estimating multiple change-points in mean on the sequence of p-values spacings, thus approximating the local FDR with a piecewise constant function. This naturally divides p-values into groups based on their significance. Additionally, we highlight some lesser-known existing change-point interpretations of the global testing methods.

Lastly, we propose a thresholding method for inference in the Gaussian sequence model. We analyse it from both multiple testing and signal estimation perspective, and consider its asymptotic behaviour. Starting from the full sequence of values, the

method sequentially excludes the largest values one by one, until the remaining values resemble noise. The idea is to consider values in groups in order to retain more signals in the case when signal is weak but dense, shared among many coordinates. We consider a possible application in the change point literature.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

Multiple testing is the problem of testing many hypotheses simultaneously, and it is a part of both classical and modern statistical literature. In applications, multiple testing adjustments are critical in order to guard against drawing false conclusions, the probability of which increases as the number of tests increases. In this thesis we approach the problem of multiple testing from a change-point perspective, we propose new methods that use ideas from the change-point literature and we note some existing connections between the two topics. Through thresholding estimators, multiple testing methods are related to the problem of signal estimation in the Gaussian sequence model. We also propose a new thresholding method for inference in the Gaussian sequence model.

In Chapter 2 we provide a literature review on the topics of multiple testing and signal estimation. An overview of the multiple testing problem and the current state of the literature are given in Section 2.1, and this section is relevant to Chapters 3, 4 and 5. Section 2.2, on estimating the proportion of false null hypotheses, is relevant to Chapter 3, and Section 2.3 on the Gaussian sequence model to Chapter 5.

In Chapter 3, we propose the Difference of Slopes (DOS) method, a new method for estimating the proportion of false null hypotheses in multiple testing problem.

This two step procedure first fits a piecewise linear function with two segments to the sorted sequence of p-values by estimating the "change-point" in slope using the proposed DOS statistic. It is interpreted as a threshold separating small, mostly false-null p-values from larger, mostly true null p-values. In the second step, this threshold is used in combination with Storey's estimator (Storey, 2002) to get the estimate of the proportion. The theoretical results show that the proposed estimator is asymptotically conservative estimator of the false null proportion, and characterise the limiting behaviour of the estimated change-point location and the false null proportion. Simulations show that our approach works particularly well in sparse settings, when the proportion of false null hypotheses is small, and yields estimates with small mean squared error.

In Chapter 4, building on the change-point interpretation suggested in Chapter 3, we explore some lesser known connections between the problems of multiple testing and change-point detection. The bridge between the two topics is the problem of testing for a change in the rate function of a Poisson process. Furthermore, we investigate possible applications of multiple change-point methods in analysing the sequence of p-values. We propose to segment p-values into groups based on their significance using some suitably modified existing multiple change-point algorithms. The results of such analysis are illustrated and possible usefulness of this approach in the applied literature is discussed. Statistical methods for grouping p-values based on their significance have not been considered in the literature, and usually some fixed thresholds are used for this purpose, which makes our proposal new. We comment on the possible applications of this approach to solving different multiple testing problems, such as estimating the local false discovery rate and the false null proportion.

In Chapter 5, we propose the Tail-Summed Scores (TSS) method for inference on the signal in the Gaussian sequence model. Starting from the full set of values,

the pseudo-sequential TSS procedure excludes the largest absolute values one by one, which are then declared as signal, until the remaining set of values begins to resemble noise as a group. This is achieved by comparing the norm of the remaining signal with a certain threshold at each step, and stopping the procedure when this norm drops below the threshold. The idea is to consider values in groups in order to detect as many as possible signal components when the signal is weak. The conservativeness of the procedure can be adjusted by choosing different sequences of thresholds. As this is a thresholding procedure, we analyse it from both the signal estimation and the multiple testing perspective. We discuss its applications to the problem of estimating the proportion of coordinates with change in the panel data change-point model, and its potential to improve the accuracy of a change-point estimation method.

Finally, Chapter 6 gives a brief summary of the contributions of this thesis and proposes possible directions for the future research.

Chapter 2

Literature review

2.1 Multiple testing problem

In this section we introduce the topic of multiple testing which is relevant to all of the remaining chapters. We introduce some elementary concepts of multiple testing with basic approaches in Sections 2.1.1 and 2.1.2. State of the art multiple testing methods are described in Section 2.1.3 and applications in Section 2.1.4.

Multiple testing problem arises when many statistical hypotheses are tested simultaneously. Let $H_0^1, \ldots, H_0^n$ be the sequence of null hypotheses, $T_1, \ldots, T_n$ the sequence of test statistics, and $p_1, \ldots, p_n$ the corresponding sequence of p-values. If the null hypothesis H_0^i does not hold, we refer to the corresponding p-value p_i as false null p-value. Otherwise, as true null p-value. Of interest are the problems of global testing, that is of testing whether all null hypotheses are true, and of simultaneous inference, deciding which null hypotheses should be rejected. Global testing problem is also called multiple testing of a single hypothesis. It comes down to a single hypothesis testing problem, where not rejecting the global null hypothesis means not rejecting any of the null hypotheses H_0^i, but by rejecting the global null we do not make decisions for the individual tests. In simultaneous inference we make a decision for each hypothesis,

and for each we have the unknown ground truth (true H_0^i or false H_0^i) and the decision made by the test (to not reject H_0^i or to reject H_0^i). Therefore, testing can result in two types of errors: type I error, that we make when the null hypothesis holds but we reject it, or type II error, when the null hypothesis is false but we do not reject it. Commonly, when testing a single hypothesis, tests are evaluated based on their ability to not reject H_0 when H_0 holds, keeping the probability of type I error below the predetermined significance level α:

$$P_{H_0}(\text{rejected}) \leq \alpha.$$

When testing multiple hypotheses, each test can result in type I error, type II error or the correct decisions of rejecting the false null, or not-rejecting the true null hypothesis. Let $\mathbb{I}$ be the indicator function, such that $\mathbb{I}(H_0^i \text{ is rejected}) = 1$ if H_0^i is rejected and 0 otherwise. We define the random variables V and R as

$$\begin{aligned} V &= \sum_{H_i=0} \mathbb{I}(H_0^i \text{ is rejected}), \\ R &= \sum_{i=1}^{n} \mathbb{I}(H_0^i \text{ is rejected}). \end{aligned} \tag{2.1}$$

R is the total number of rejected hypotheses, while V is the number of falsely rejected null hypotheses. Table 2.1 below, shows the common notation for the number of errors and right decisions made when simultaneously testing n hypotheses. The random variables U, T and S are defined analogously to V and R in (2.1).

	non-rejected	rejected	total
true H_0	U	V	n_0
false H_0	T	S	$n - n_0$
total	$n - R$	R	n

Table 2.1 Random variables denoted in the cells of the table are obtained by summing the indicators of each of the four possible outcomes over H_i.

Most of the methods discussed here are based on p-values derived from continuous and known distributions of test statistics T_i under the null. Given a sample $\boldsymbol{X} = (X_1, \ldots, X_m)$, p-value is defined as a statistic for which it holds that $0 < p(\boldsymbol{X}) < 1$ and that for any $0 \leq \alpha \leq 1$,

$$P_{H_0}(p(\boldsymbol{X}) \leq \alpha) \leq \alpha.$$

A p-value can be constructed that has a uniform distribution on an interval $[0, 1]$. To illustrate this, let H_0 be a simple hypothesis, containing only one probability distribution for the sample, and let T be the test statistic with a continuous cumulative distribution function F_T, such that large values of T are critical and in favour of rejecting H_0. For $p = 1 - F_T(T)$ it holds that

$$\begin{aligned} P_{H_0}(p < x) &= P(1 - F_T(T) < x) \\ &= P(F_T(T) > 1 - x) \\ &= P(T > F_T^{-1}(1 - x)) \\ &= 1 - F_T(F_T^{-1}(1 - x)) \\ &= x, \end{aligned}$$

which proves that under H_0, $p \sim U[0, 1]$.

2.1.1 Global testing

Global testing is the problem of testing whether all null hypotheses are true. This is called the intersection, or the global null hypothesis, and it can be seen as a conjunction of the individual hypotheses:

$$H_0 : \bigwedge_{i=1}^{n} H_0^i,$$

although $\cap_{i=1}^{n} H_0^i$ is a more common notation. Consider the global null test where we reject the global null hypothesis if there is at least one p-value smaller than α. This test has the probability of type I error going to 1, that we now explain. Let $H_1, ..., H_n$ be the sequence of numbers corresponding to the sequence of hypotheses, where $H_i = 1$ if H_0^i is true, otherwise $H_i = 0$. The rejection is made if $V > 0$, where V can be written as

$$\begin{aligned} V &= \sum_{H_i=0} \mathbb{I}(p_i \leq \alpha) \\ &= \mathbb{I}\{\min_i p_i \leq \alpha\}. \end{aligned}$$

Using the fact that the true null p-values are independent and have $U[0,1]$ distribution under the global null, the probability of making a type I error is

$$P_{H_0}(V > 0) = 1 - (1 - \alpha)^n.$$

This means that using a constant significance level leads to falsely rejecting the global null hypothesis with probability going to 1 as the number of hypotheses increases. To control the type I error we need a threshold for p-values that goes to zero as n increases.

We introduce some of the most common global testing methods below. The Bonferroni test rejects the global null hypothesis if and only if

$$\min_{i=1,...,n} p_i \leq \alpha/n.$$

This test controls the probability of type I error at level α, without any assumptions on the dependence structure of the p-values:

$$\begin{aligned} P_{H_0}(\min_i p_i \leq \alpha/n) &\leq \sum_{i=1}^{n} P_{H_0}\left(p_i \leq \alpha/n\right) \\ &\leq n\alpha/n = \alpha. \end{aligned}$$

The name originates from the use of the union bound, which belongs to the group of Bonferroni inequalities, in the proof of the type I error control. Let $p_{(1)}, \ldots, p_{(n)}$ be an increasingly sorted sequence of p-values. Graphically, if points $(i/n, p_{(i)})$ are plotted on a coordinate plane, the Bonferroni method rejects H_0 if all p-values fall below the line $y = \alpha/n$.

The Simes test (Simes, 1986) rejects H_0 if and only if there exists $i \in \{1, \ldots, n\}$ such that

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

Graphically, this means that H_0 is rejected if any of the sorted p-values $(i/n, p_{(i)})$ fall below the line with slope α/n. The smallest p-value is compared to the same threshold as with the Bonferroni, but larger p-values are compared to larger thresholds. It follows that:

$$\begin{aligned} \{\min_i p_i \leq \alpha/n\} &= \bigcap_{i=1}^{n} \{p_{(i)} \leq \alpha/n\} \\ &\subseteq \bigcap_{i=1}^{n} \{p_{(i)} \leq \alpha i/n\}, \end{aligned}$$

showing that the rejection set for the Bonferroni is smaller than for the Simes procedure. Hence, the Simes procedure is less conservative than the Bonferroni.

For the Fisher's combined probability test (Fisher, 1946) the distribution of the p-values is assumed to be $U[0, 1]$ under the null and they are assumed to be independent.

From the probability transformation it follows that the distribution of $-\log(p_i)$ is standard exponential Exp(1). Combining p-values we get a test statistic for the global null test,

$$-\sum_{i=1}^{n} 2\log(p_i),$$

which has χ^2_{2n} distribution under the null. As implied by the Fisher's combined test, if the correct distribution of the test statistics under the null is known, then testing the global null hypothesis can be seen as a goodness-of-fit test for the uniform distribution. This means that the statistics such as the Kolmogorov-Smirnov (Kolmogorov, 1933), Cramér-von Mises (Cramér, 1928) or Anderson-Darling (Anderson and Darling, 1952) can be used to this end. However, as the false null p-values tend to take smaller values, we are only interested in testing one-sided hypothesis and with a focus on the left tail of the p-value distribution, which should be taken into account.

Recently, global tests for some specific models were proposed that are not necessarily based on a given sequence of p-values. In Sur et al. (2017) and Ma et al. (2021), global testing methods for the parameters of the high-dimensional logistic regression model are proposed and their asymptotic distribution under the null derived. Detection boundary for the problem of detecting sparse regression models is studied in Ingster et al. (2010) and Arias-Castro et al. (2011). Global test for the equality of the coefficients of two high-dimensional multivariate regression models is proposed in Xia et al. (2018). Detection boundary for sparse binary regression models was studied in Mukherjee and Johnstone (2015). Global testing of covariance structure of a multivariate sample is considered in Cai (2017). Global testing against sparse alternatives under Ising models is considered in Mukherjee et al. (2018).

2.1.2 Simultaneous inference

In this section we introduce some basic approaches to the problem of simultaneous inference in multiple testing, when a decision is to be made for each individual hypothesis H_0^i. If all hypotheses are independent and true null, and each hypothesis is tested at a constant level α, the number of type I errors made will be $n\alpha$ on average, which follows from the uniform distribution of the p-values under the null. For large n, this number can be unacceptably large, and therefore multiple testing procedures are defined to control it or other related quantities (error rates). We introduce some of these error rates and methods that control them in the sections below. Which error rate to use depends on the research objective, and how harmful false rejections are considered to be. The familywise error rate (FWER) is defined as

$$\text{FWER} = P(V > 0),$$

where V is defined in Table 2.1. Although this looks similar to the probability of type I error in global null testing, for a multiple testing procedure it is of interest to control this probability under any configuration of true and false null hypotheses, not simply under the global null. If a multiple testing procedure controls the FWER under the global null, then we say that it control the FWER weakly. If it controls the FWER under any configuration of true and false null hypotheses, then we say that it controls the FWER strongly.

The closure principle provides a way to adapt any global testing procedure to a multiple testing procedure that controls the FWER in a strong sense. First, a collection containing all possible intersection hypotheses is defined as

$$\mathcal{C} = \{H_I : I \subseteq \{1, \ldots, n\}\}, \tag{2.2}$$

where $H_I = \bigwedge_{j \in I} H_j$. Given a global testing procedure, the corresponding closure procedure rejects H_i if all intersection hypotheses H_I are rejected at level α, for any $I \subseteq \{1, \ldots, n\}$ such that $i \in I$. This implies a top-down order of testing of hypotheses on a tree. The global null hypothesis is the root node and the children are obtained by excluding the individual hypotheses from the intersection. For a node H_I, for some $I \subseteq \{1, \ldots, n\}$, its children are $\{H_J : J = I \setminus \{i\}, i \in I\}$, so the tree is of depth n, with individual hypotheses as leaves. If H_I is not rejected, then we do not proceed to test the children hypotheses, and the individual hypotheses H_i, for $i \in I$, are not rejected. The resulting procedure controls the FWER in a strong sense at level α. Let $\mathcal{H}_0 = \{i : H_i^0 \text{ holds}\}$. To make a false rejection it is necessary to reject $H_{\mathcal{H}_0} = \bigwedge_{i \in \mathcal{H}_0} H_i^0$. It follows that

$$P(V > 0) \leq P(H_{\mathcal{H}_0} \text{ is rejected}) \leq \alpha. \tag{2.3}$$

A disadvantage of using the closure principle is the large number of tests that needs to be performed, which can make the resulting procedure complicated and slow to compute. However, for some global testing methods, the closure procedure is simple. The multiple testing procedure obtained from the closure principle and the Bonferroni procedure is the Holm procedure (Holm, 1979). Closure principle with the Simes procedure leads to the Hommel procedure (Hommel, 1988).

The FWER may be seen as too restrictive a criterion. When testing a large number of hypotheses, it can happen that some true null p-values take very small values, so the false rejections happen early on. FWER controlling procedures would therefore restrict the number of true rejections. Furthermore, falsely rejecting some smaller proportion of the true null hypotheses may not be harmful if the goal is to identify interesting hypotheses for further analysis. In the seminal paper by Benjamini and Hochberg (1995) a new multiple testing error rate, called false discovery rate, is proposed together with a method for controlling it. To define it, let V and R be defined as in Table 2.1.

The false discovery proportion (FDP) is defined as

$$\text{FDP} = \begin{cases} V/R, & R > 0 \\ 0, & R = 0. \end{cases}$$

The false discovery rate (FDR) is defined as the expected value of the FDP:

$$\text{FDR} = \mathbb{E}(\text{FDP}).$$

Let $q \in (0, 1)$. The Benjamini-Hochberg (BH) FDR-controlling procedure rejects $p_{(1)}, ..., p_{(\hat{k}_{BH})}$, where

$$\hat{k}_{BH} = \max \left\{ i : p_{(i)} \leq \frac{qk}{n} \right\}. \tag{2.4}$$

We can see this method as thresholding at level $t_{BH} = p_{(\hat{k}_{BH})}$. In Benjamini and Hochberg (1995) it is proven that, if all true null p-values are independent, level-q BH procedure controls the FDR conservatively at level q, under any configuration of true and false null hypotheses. More precisely, for the BH procedure it holds that

$$\mathbb{E}(\text{FDP}|p_1 = p_1, \ldots, p_n = p_n) \leq \frac{n_0}{n} q, \tag{2.5}$$

where n_0 is the number of true null hypotheses. Hence, the BH procedure controls the FDR at level q, for any configuration of the true and the false null p-values, conditional on their values. Using the property of the conditional expectation, it holds that

$$\text{FDR} \leq \frac{n_0}{n} q. \tag{2.6}$$

In Finner and Roters (2001) and Genovese and Wasserman (2002) it is proven that equality holds in (2.6). In Section 2.2 we explain how the estimator of the false

null (equivalently true null) proportion can be used to increase the power of the BH procedure. The relationship between the FWER and the FDR is:

$$\text{FDR} = \mathbb{E}(\text{FDP}) \leq \mathbb{E}(I\{V > 0\}) = P(V > 0) = \text{FWER},$$

so any procedure controlling the FWER also controls the FDR.

In Storey (2002), the following mixture distribution model for the p-values was first considered:

$$p \sim \pi_1 F_1 + \pi_0 U[0, 1],$$

where $\pi_0 = 1 - \pi_1$, the distribution under the null is $U[0, 1]$ and under the alternative F_1. This facilitated further development of the theory for the FDR. In Genovese and Wasserman (2004), the FDR is considered as a function of a given threshold for rejecting p-values. They defined the stochastic process $\text{FDP}(t)$ and its expectation $\text{FDR}(t)$:

$$\text{FDP}(t) = \frac{\sum_{H_i=0} \mathbb{I}(p_i < t)}{\sum_{i=1}^{n} \mathbb{I}(p_i < t)},$$
$$\text{FDR}(t) = \mathbb{E}(\text{FDP}(t)).$$

As $\mathbb{E}(\text{FDP}(t))$ involves the mean of the ratio of random variables, a more convenient quantity is considered

$$Q(t) := \frac{\pi_0 t}{\pi_1 F_1(t) + \pi_0 t}. \tag{2.7}$$

$Q(t)$ is proved to be an asymptotic mean of the $\text{FDP}(t)$ process when $n \to \infty$, assuming the mixture distribution above, in the sense that $\mathbb{E}(\text{FDP}(t)) = Q(t) + o(1)$. This approach allowed them to develop the theory of the FDR control of the BH method, including the consistent estimation of $\text{FDR}(t)$ and the asymptotic validity of

the proposed plug-in method for the FDR control under the mixture model assumption for the p-values.

While the FDR is the most frequently used multiple testing error rate, we mention a few others defined in the literature. In Genovese and Wasserman (2002), the false non-discovery rate (FNR) is defined as

$$\mathrm{FNR} = \begin{cases} \mathbb{E}\left(T/A\right), & A > 0 \\ 0, & A = 0, \end{cases}$$

where $A = n - R$ is the total number of non-rejected hypotheses, and T is the number of falsely non-rejected hypotheses defined in Table 2.1. The authors consider the risk function that includes both the FDR and the FNR, and describe a procedure that can be used to minimise this risk function.

The positive FDR (pFDR), and a method that controls it are proposed in Storey (2003), where the pFDR is defined as

$$\mathrm{pFDR} = \mathbb{E}\left(\frac{V}{R}\middle| R > 0\right).$$

pFDR has a Bayesian interpretation, as the probability of falsely rejecting the null hypothesis given the rejection region. In Efron (2007), assuming the mixture distribution model for the test statistics (or p-values), the local FDR is defined for each observation as the posterior probability of it being from the null distribution:

$$\mathrm{lfdr}(t) = \frac{\pi_0 f_0(t)}{\pi_0 f_0(t) + \pi_1 f_1(t)},$$

where f_0 is the density under the null and f_1 under the alternative. To highlight the fact that the local FDR is a density based quantity, as opposed to the tail-area based

FDR, lowercase letters are used for the notation. Efron (2007) proposes a method for estimating the parameters of the mixture distribution, thus estimating the lfdr function. The significance of each hypothesis is then measured with the lfdr value, and thresholding the lfdr values defines a multiple testing procedure. If the density of the false null p-values is decreasing, then $\mathrm{FDR}(t)$ is smaller than $\mathrm{lfdr}(t)$. The cutoff threshold used for FDR is usually 0.05 or 0.1, while for lfdr a larger threshold is used, for example 0.2 in Efron (2007). Another related quantity is the marginal FDR (mFDR) defined as

$$\mathrm{mFDR} = \frac{\mathbb{E}(V)}{\mathbb{E}(R)},$$

and related to $Q(t)$ defined in (2.7). The relationships between the FDR, pFDR and mFDR are as follows. The mFDR and the pFDR are equal if the hypotheses come from a two component mixture distribution (see Corollary 1 in Storey (2003)). For the FDR and the mFDR, Genovese and Wasserman (2002) proved that

$$\mathrm{mFDR} = \mathrm{FDR} + O(n^{-1/2}).$$

Among the mentioned error rates, the most widely used are the FDR and the local FDR. The Benjamini-Hochberg method is used for its simplicity and theoretical guarantees, while the local FDR is used for its convenient Bayesian probability interpretation.

2.1.3 Modern multiple testing literature

Since the early days of multiple testing research, state of the art data sets arising from applications motivated the development of multiple testing procedures for increasingly complex models. Below we describe such settings in more detail, and review the relevant literature. Some of these settings include assumptions on the structure of p-values, group or hierarchical, or additional data such as weights or covariates containing

information on how likely a test is to be a true null, or how "important" it is, in the sense of the loss incurred by making a wrong decision.

If p-values are initially divided into groups, then the goal can be to control the group FDR, within-group FDR or overall FDR. Grouped p-values were considered in Cai and Sun (2009), Hu et al. (2010) and Liu et al. (2016b). Hierarchical structure of the groups of p-values is considered in Yekutieli (2008), Benjamini and Bogomolov (2014), Barber and Ramdas (2017), Ramdas et al. (2019) and Katsevich et al. (2021). In Bogomolov et al. (2021) a hierarchical structure for the p-values is considered, where an error rate is defined at each level of the hierarchical structure and in a bottom-up way, from children to parent. The proposed procedure is shown to control this error under certain dependency assumptions on the tree structure of p-values. In Basu et al. (2018), a weighted FDR-controlling procedure is proposed, where weights describe the severity of a false positive decision and the power gain of a true positive decision. Multiple testing procedure where prior information is given as p-value weights is first considered in Genovese et al. (2006) and Roquain and Van de Wiel (2009). In Ignatiadis et al. (2016) and Ignatiadis and Huber (2021), an FDR-controlling procedure is proposed where weights are calculated from the additional covariate containing information on the power of each test and the probability of it being true null. The covariate given for each p-value is independent of it if the hypothesis is true null. This setting was also considered in Zhang et al. (2019), where multidimensional covariates are allowed. In Lei and Fithian (2018), an iterative procedure for multiple testing with side information is proposed that controls the FDR in finite samples. In Lei et al. (2021) a multiple testing procedure is proposed with the possibility of generic structural constraint on the rejected set of hypotheses. Multiple testing for spatial data is considered in Cai et al. (2022). In Cao et al. (2022), an FDR controlling multiple testing procedure was proposed where auxiliary information given for each test induces an ordered sequence

of hypotheses. This order is not strictly followed, as the procedure does not necessarily reject the initial block of p-values. This flexibility can be useful when the auxiliary information is not very strong. p-values with heterogeneous distributions are considered in Habiger et al. (2017) and Chen et al. (2020).

Multiple testing concepts are also used for solving the problem of variable selection in linear models, where the FDR control is also of interest. Computing p-values for the coefficients of a linear model is not always possible and even if it is, the conditions on their distribution and the dependence structure needed for many multiple testing procedures are not met. The concept of knockoffs was introduced in Barber and Candés (2015) for variable selection in linear regression models, and a procedure that controls the FDR of the selected variables is proposed, that does not rely on computing the p-values.

A possible application of the method proposed in Chapter 3 to some of these modern settings is discussed in Section 3.6. A possible application of the method proposed in Chapter 4 for choosing the weights in Basu et al. (2018) is discussed in Section 4.3.

2.1.4 Applications

Multiple testing problem is popularised as it appears in many applications, such as medical research, genomics, social sciences and so forth. It is necessary to be addressed in order to adequately interpret results. Below we review some of the datasets used in applications of the modern multiple testing algorithms. The most common application of the large scale multiple testing methods is in DNA microarray experiments, and most often for the purpose of analysing gene expression data. First we introduce some terms from genetics that will be used to describe the datasets.

Gene expression is a process where the information from the genes is used in a synthesis of a final gene product, which for protein-coding genes is a protein. The

process is divided into two stages: transcription and translation. For protein-coding genes, in the transcription stage, mRNA is produced, while in the translation stage the information from the mRNA is decoded and protein molecules are synthesised. Proteins play an important role in a cell, as they are involved in most of the cell functions. Thus, alterations at the transcriptional level can lead to abnormal functions of proteins, causing the development of cancer. DNA microarray (chip) is a technology that can be used, among other things, to get the quantitative measurements for the expression of genes by measuring the amount of mRNA in a cell. To explore the genetic component of a disease, the gene expression levels between healthy tissues or cell lines and those of patients affected by a disease are compared. Identifying differentially expressed genes is of interest because they are related to the disease. Studying them is useful for understanding the pathological process, and for development of personalised treatments. There are usually thousands of genes, and for each a measure of difference in expressions is given, which might be significant or not. This forms a sequence of hypotheses tests and test statistics. For gene i H_0^i is the null hypothesis that gene i is not differentially expressed. Analysing gene expression data can then help in early diagnosis, or it can be used for individualised treatment of the disease.

Single-nucleotide polymorphism (SNP) is a variation at a single position in a DNA sequence among individuals. DNA is a polymer, made up of long chains of nucleotides, and a nucleotide contains a nucleobase. In the DNA there are four nucleobases: guanine (G), adenine (A), cytosine (C) and thymine (T). SNPs occur when at a specific base position in the genome there is a variation, for example, such that most people have G-nucleotide, while some minority has an A-nucleotide at that position. Often, for such a variation to be considered an SNP, it should be present in at least one percent of the population. A more general term is single-nucleotide variants SNVs that includes SNPs and also rare mutations (present in less than 1% of the population), however

this distinction is sometimes disregarded and any variation is considered an SNP. The process of determining the DNA sequence, that is the order of nucleotides in DNA is called DNA sequencing, and SNPs can be detected through this analysis. SNP array, a type of a DNA microarray, is another tool that can be used to detect SNPs within a population. An important use of SNPs is in genome-wide association studies (GWASs). GWASs aim to discover SNPs that are associated with a phenotype (an observable characteristic) or with a disease, such as heart disease, diabetes and cancer. Large-scale multiple testing problems arise naturally in GWASs. For example, all human beings have 99.9% identical DNA, so the frequency of nucleotide variations is once in every 1000 nucleotides. Considering different populations (geographical or ethnic groups) more than 600 million SNVs in total have been identified. In GWASs this large number of SNPs is tested for association with a disease or a phenotype and multiple testing corrections are necessary to be addressed.

Using the notions of SNPs and gene expression introduced above, we describe the datasets used in some of the papers described in Section 2.1.3.

In Basu et al. (2018) a weighted multiple testing procedure was applied to a data set from the Framingham Heart Study. This long-term cohort study started observing healthy individuals from the town of Framingham, Massachusetts with no symptoms of cardiovascular diseases (CVD). The primary goal of the study was to identify the risk factors for the CVD by monitoring the participants over the years. The study began in 1948 and is still ongoing, now studying third generation offsprings of the original cohort, and including additional cohorts reflecting the more diverse population. The study also expanded the research questions and the type of data collected. This includes a GWAS of two generations, with the aim to study how genetic variants contribute to phenotypes that are risk factors for the CVD (Cupples et al., 2007). In Basu et al. (2018) the data from the older generation was used to define weights to

be used for the multiple testing procedure on the younger generation, with the aim of identifying SNPs with significant association to BMI. The p-values of SNPs from the older generation data, measuring the significance of the relationship to BMI, are divided into three groups, describing the strength of the relationship. These groups are: $\mathcal{G}_1$ - less than 0.001, $\mathcal{G}_2$ - between 0.001 and 0.01, and $\mathcal{G}_3$ - greater than 0.01. The weights given to the younger generation participants are based on the group which the parent belongs to. The values associated to the three groups are arbitrary, but should be decreasing respectively, for example $(4, 2, 1)$ and $(5, 2, 1)$ were used in the paper. Larger values correspond to the gain when a hypothesis is rejected correctly.

In Cao et al. (2022), data from two GWASs on coronary artery disease (CAD) was used for the illustration of their proposed multiple testing method with auxiliary data. Each data set contains p-values for over 500000 common SNPs, measuring their association with CAD. The proposed procedure uses p-values from one study as an auxiliary information that induces ordering of the p-values from the other study, which is necessary for the method.

In Ignatiadis et al. (2016) and Ignatiadis and Huber (2021) the proposed hypothesis weighting methods are applied to the study from Grubert et al. (2015) where the associations between SNPs and chromatin activity is studied. In particular the associations between SNPs and the levels of H3K27ac are of interest. H3K27ac is a code for a specific modification to the basic proteins histones, around which DNA is wrapped around. This modification is associated with enhanced transcription of the genes. Under the null hypothesis, SNP values are independent of the H3K27ac levels at all locations, while under the alternative the values are associated. The hypotheses are formed by comparing SNP levels at each genomic location with H3K27ac at each location. The procedure identifies pairs of SNPs and genomic regions (H3K27ac peaks) where those are correlated.

In Lei and Fithian (2018) several real datasets of gene expression levels were used for illustration of their proposed multiple testing method with side information. The data is obtained through RNA sequencing (RNA-Seq), which is a sequencing technology that has some advantages to microarray-based methods, see Kukurba and Montgomery (2015). RNA-Seq data is used by some of the multiple testing methods with auxiliary information, since in addition to the gene expression levels (RPKS), the information about the preciseness of the measurements (raw count data) can be used as a measure of reliability.

In Liu et al. (2016a), GWAS data on breast cancer with more than 500000 SNPs is used to identify SNPs associated with breast cancer. As SNPs that are nearby tend to be highly correlated, methods for dealing with multiple testing under dependence are proposed. The dependency is modeled by Markov random field, making the proposed FDR controlling procedure a multiple testing procedure for graphical models.

An example of hierarchical structure of hypotheses can be found in Bogomolov et al. (2021), where the goal is to identify SNPs that influence the expression of genes in a multi-tissue analysis. Highest level hypotheses are tissues, as gene expression levels vary across tissues. In each tissue, hypotheses are tested to find which SNPs inside the gene affect its expression. The goal is to find shared and tissue specific SNPs affecting the genes. The proposed procedure controls error rates at multiple levels of resolution. Another dataset studied in this paper is on the association of gut microorganisms and colorectal cancer. The hierarchical data structure comes from the taxonomic classification of these microorganisms.

Gene expression level data from a GWAS on HIV, with the goal to find the genes that are differentially expressed in HIV positive was considered in Efron (2008) and Lynch et al. (2017). Data on HIV amino acid sequences comparing the mutation rates at different locations between the subtypes B and C of the HIV virus is studied in

order to help develop a vaccine that would be efficient against both subtypes. This data was used in Chen et al. (2018).

Gene/drug response data is used in Li and Barber (2017) and Li and Barber (2019). The data given is a gene expression data of the breast cancer cells in response to no dose, low dose and high dose setting of oestrogen treatment. The aim is to find which genes are affected by the low dosage, by using the information on the effect of the high dosage.

In Hu et al. (2010) a method for multiple testing of grouped hypotheses is applied on a breast cancer microarray gene expression dataset. The grouping of hypotheses is done according to the Gene Ontology (GO) categories, or using clustering techniques.

In Benjamini and Bogomolov (2014), data from a GWAS on Alzheimer's disease is considered to find which SNPs are associated to some regions in the brain given voxel data of the brain volume in patients with Alzheimer's disease. Hypotheses are formed for each voxel and each SNP, and groups (families) are formed by fixing the SNP in consideration. Their proposed method selects significant groups and tests the hypotheses within them.

In Ma et al. (2021), the proposed multiple testing method for the parameters in high-dimensional logistic regressions is used to test the association between different faecal metabolites and pediatric Crohn's disease.

Adequate yearly progress (AYP) data of California elementary schools is used for illustration of group hypothesis testing procedures proposed in Cai and Sun (2009), Liu et al. (2016b) and Sarkar and Zhao (2017). The data was collected with the aim to compare the academic performances of socioeconomically advantaged (SEA) to socioeconomically disadvantaged (SED) students. The hypotheses are initially divided in three groups, and correspond to small, medium and large schools. Multiple testing

procedure is applied to identify the schools where the difference in the performance is unusually large or unusually small.

Multiple testing method for detecting spatial signals in imaging data is proposed in Zhang et al. (2011). The method is applied to the fMRI dataset obtained in an emotional control study, where the goal is to detect regions of activation in brain when subjects are shown series of images. To this end, their proposed method takes into account both spatial and temporal correlation in the data.

Cai et al. (2022) propose a multiple testing method for hypotheses located on a lattice with spatial patterns, with applications in the analysis of two- and three-dimensional images. First, the sparse spatial structure is estimated and this information is used to construct the weights, inducing the ordering of p-values. The ordered p-values are then used in a BH-like procedure controlling the FDR asymptotically. For three-dimensional multiple testing, MRI data was used from the study of attention deficit hyperactivity disorder (ADHD). Through reduction in the resolution of the MRI images, the information on brain activity is aggregated and p-values from testing for the difference in brain activity between subjects with and without ADHD are calculated. The proposed spatial multiple testing procedure is used to reveal the regions where the activity is different between the two groups.

In Barber and Ramdas (2017), fMRI data was considered for the problem of multiple testing of whether the activity in a voxel v in the brain is related to the semantic features of the text presented to the subjects s seconds earlier. The method proposed in this paper simultaneously controls the FDR across multiple partitions of p-values. The p-values are grouped in three ways, fixing the value of s, fixing the value of v, and considering a groups of voxels belonging to each ROI for all s.

2.2 Estimating the proportion of false null hypotheses

Estimating the proportion of false null hypotheses can be of interest in its own right, as an overall measure of the extent of the observed changes. However, more often, proportion estimators are used in further analysis, to improve on the methods for identifying the subset of false null hypotheses while controlling a certain multiple testing error rate. In particular, proportion estimators can be used for improving the FDR-controlling Benjamini-Hochberg (BH) procedure, introduced in Section 2.1.2 . The idea of making the BH procedure adapt to the unknown proportion was first proposed in Benjamini and Hochberg (2000). Let $\hat{n}_0$ be a conservative estimator of n_0, the number of true null hypotheses, such that $P(\hat{n}_0 \geq n_0) = 1$. Note that estimating the proportion or the number of true or false null hypotheses is considered to be the same problem. Consider the procedure that thresholds the p-values at $t_{\text{adapt}} = p_{(\hat{k}_{\text{adapt}})}$, where

$$\hat{k}_{\text{adapt}} = \max \left\{ i : p_{(i)} \leq \frac{qk}{\hat{n}_0} \right\}. \tag{2.8}$$

A procedure incorporating a proportion estimator as above is called an adaptive BH procedure, it controls the FDR at the same level q as the classical BH procedure but has larger power as the threshold is larger. This follows by replacing q in (2.5) by $q' = qn/\hat{n}_0$:

$$\text{FDR}(t_{\text{adapt}}) = \frac{qn_0}{\hat{n}_0} \leq q \text{ a.s..}$$

If $\hat{n}_0$ is only asymptotically conservative, in the sense that

$$\frac{n_0}{\hat{n}_0} \to c, \text{ a.s. as } n \to \infty,$$

where $c \leq 1$, then the adaptive estimator will control the FDR asymptotically. As the threshold of the adaptive BH procedure is larger than that of the classical BH procedure, the adaptive BH has an increased power, in the sense that the probability of true rejections is larger. A comparison of different adaptive FDR-controlling procedures is investigated in Blanchard and Roquain (2009).

Aside from improving the power of the BH procedure, an interest in studying the proportion estimators also comes from the connection to the problem of estimation in a two-component mixture model. In the literature on proportion estimators, p-values are most often modeled as having a mixture distribution with two components, a uniform distribution and an unknown false null distribution:

$$F(t) = \pi_0 t + \pi_1 F_1(t), \quad t \in (0, 1), \tag{2.9}$$

where $\pi_1 = 1 - \pi_0$. This model is considered in the majority of papers reviewed below (Cai et al., 2007; Genovese and Wasserman, 2004; Langaas et al., 2005; Meinshausen and Rice, 2006). Estimating the false null proportion then means estimating the parameter π_1. A mixture model with one known component arises in some applications. A potential application of a proportion estimator to astronomy data was proposed in Meinshausen and Rice (2006), for estimating the number of objects in the Kuiper belt. Swanepoel (1999) considers application in astrophysics, where the proportion parameter models the strength of the pulsed signal.

We review some of the methods proposed in the literature for estimating the proportion of the false null hypotheses. We focus on the case of independent p-values. Estimating the proportion under dependence is less studied in the literature, see for example Friguet and Causeur (2011), Ostrovnaya and Nicolae (2012) and Neumann et al. (2021).

2.2.1 p-value plot and the CDF-based methods

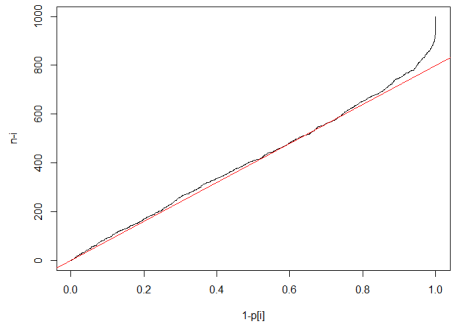
The earliest papers on this topic consider a fixed number of the true null hypotheses n_0 with no assumptions on the distribution of the alternative p-values. The only assumption is that the true null p-values have $U[0,1]$ distribution while the false null p-values tend to be smaller. This model is considered first in Schweder and Spjøtvoll (1982) and Benjamini and Hochberg (2000). Many of the later papers are based on their idea of analysing the p-value plot which we introduce now. If there is a fixed number n_0 of true null hypotheses and $\lambda \in (0,1)$ is large such that there are not many (or not any) alternative p-values larger than λ, then

$$\mathbb{E}\left(\sum_{i=1}^{n} \mathbb{I}\left\{p_i > \lambda\right\}\right) \approx \mathbb{E}\left(\sum_{i:H_i=0} \mathbb{I}\{p_i > \lambda\}\right) \approx n_0(1-\lambda). \tag{2.10}$$

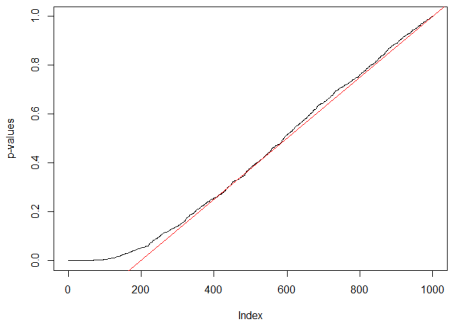
Let $W(\lambda) = \sum_{i=1}^{n} \mathbb{I}\{p_i > \lambda\}$. This means that empirically, we expect

$$W(\lambda) \approx n_0(1-\lambda).$$

The values of function $W(\lambda)$ only change at points $\lambda = p_{(i)}$, so we consider points $(1-p_{(i)}, W(p_{(i)})) = (1-p_{(i)}, n-i)$. For large i, these should indicate a line with slope n_0 - this is referred to as the p-value plot in Schweder and Spjøtvoll (1982). Alternatively, the empirical CDF plot or the plot of $(i, p_{(i)})$ can be used as they contain the same information, but the former is more convenient in a sense that the slope is "equal" to what we are trying to estimate, and there is no need for further transformations. An illustration of the p-value plot and the quantile plot $(i, p_{(i)})$ is given in Figure 2.1. It shows that, to estimate the non-null proportion, we should estimate the slope of the linear part. By decreasing the nonzero mean, the p-value plot approaches the linear function, and the curvature becomes smaller, while increasing the nonzero mean makes the line approximately piecewise linear. Similar holds for the quantile plot of p-values.



(a) Points $(1-p_{(i)}, n-i)$ and the line with slope n_0



(b) Points $(i, p_{(i)})$ and the line with slope $1/n_0$

Fig. 2.1 p-value plot and the quantile plot for the fixed-proportion model with $n = 1000, n_1 = 200$. 2-sided p-values where the null distribution of the test statistics is $N(0,1)$ and the alternative is $N(3,1)$.

Schweder and Spjøtvoll (1982) propose using the p-value plot just as an "overview of the situation" when "no quantified estimate of n_0 is needed". They suggest that the method can be formalized using the least squares estimate of the slope, but that it would be difficult to assess the properties of the estimator since p-values are not independent, so they do not proceed in that direction. Instead, a family of plug-in estimators based on (2.10) is proposed. For $\lambda \in (0,1)$, the proposed estimator of the number of true null hypotheses is

$$\hat{n}_{0,SS}(\lambda) = \frac{W(\lambda)}{1-\lambda}. \tag{2.11}$$

Now, assume that p-values come from a mixture distribution (2.9). Denote as $\hat{F}_n$ the empirical CDF of p-values. As $W(\lambda) = n(1-\hat{F}_n(\lambda))$, this leads to the following family of plug-in estimators for the true null and the false null proportion:

$$\hat{\pi}_0(\lambda) = \frac{1-\hat{F}_n(\lambda)}{1-\lambda},$$

$$\hat{\pi}_1(\lambda) = \frac{\hat{F}_n(\lambda) - \lambda}{1 - \lambda}. \tag{2.12}$$

Note that $\hat{\pi}_0(\lambda)$ is the slope of the line connecting two points on the ECDF plot, $(\lambda, F_n(\lambda))$ and $(1, 1)$, and the reciprocal of the slope connecting the corresponding points on the quantile plot. The family of estimators (2.12) is usually referred to as Storey's estimator in the literature as it was also considered in Storey (2002) for the purpose of adaptive pFDR control, however the idea first suggested by Schweder and Spjøtvoll (1982). Schweder and Spjøtvoll (1982) do not propose a method for choosing λ, but they note that if λ is too small, we will include some false null p-values leading to a biased estimator underestimating π_1. However, if λ is close to 1, the variance of the estimator will be too large because of the factor $(1 - \lambda)$ in the denominator. This is clear from the following formulas for bias and variance:

$$\begin{aligned} \text{Bias}(\hat{\pi}_1(\lambda)) &= -\frac{1 - F_1(\lambda)}{1 - \lambda}, \\ \text{Var}(\hat{\pi}_1(\lambda)) &= \frac{\pi_0}{n}\left(\frac{1}{1 - \lambda} - \pi_0\right). \end{aligned}$$

Below we describe different approaches proposed in the literature for choosing λ to be used in (2.12).

We start with Benjamini and Hochberg (2000), where the proportion estimator is constructed and first used for the purpose of improving the power of the BH procedure. Taking $\lambda = p_{(j)}$ for some j in (2.11) we have

$$\hat{n}_{0,SS}(p_{(j)}) = \frac{n - j}{1 - p_{(j)}} \approx \left(\frac{1 - p_{(j)}}{n + 1 - j}\right)^{-1}. \tag{2.13}$$

We would have the equality in (2.13) if we defined $W(\lambda) := \sum_{i=1}^{n} I\{p_i \geq \lambda\}$ or if we had added $(n+1)$th p-value $p_{(n+1)} = 1$, however the difference is negligible. The value

for λ proposed by Benjamini and Hochberg (2000) is $\lambda = p_{(\hat{j}_{BH})}$ where

$$\hat{j}_{BH} = \min\left\{i : \frac{1 - p_{(i)}}{n + 1 - i} > \frac{1 - p_{(i+1)}}{n + 1 - (i + 1)}\right\},$$

so according to (2.13) the proposed estimator of the number of the true null hypotheses is

$$\hat{n}_{0,BH} = \lceil \hat{n}_{0,SS}(p_{(\hat{j}_{BH})}) \rceil.$$

$\hat{j}_{BH}$ is the first index for which the slopes stop increasing, which is a sign we are probably past all the smaller alternative p-values. The illustration can be seen in Figure 2.2. No theoretical guarantees are provided for this estimator, however it is

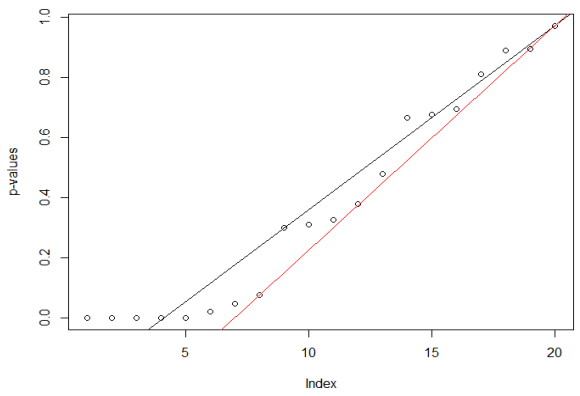


Fig. 2.2 Benjamini and Hochberg (2000) method illustration: 5 out of 20 p-values are false null. Slopes of lines connecting smallest p-values with the last one are increasing at first. The red line is connecting $(\hat{j}_{BH}, p_{(\hat{j}_{BH})})$ and $(20, p_{(20)})$ and the black line that passes through $(\hat{j}_{BH} + 1, p_{(\hat{j}_{BH}+1)})$ has a smaller slope. In this example $\hat{j}_{BH} = 8$ and the estimated number of false nulls is 5.

strongly negatively biased, and limiting the gain in power of the adapted BH procedure. For this reason it is usually not used in practice.

In Storey (2002), the problem of estimating the proportion is considered indirectly, as a means for getting a better estimate of the positive FDR (pFDR). λ is chosen using bootstrap method with the goal of minimising the MSE of the resulting pFDR

estimator. In Storey and Tibshirani (2003) a cubic spline is fitted to the function $\hat{n}_{0,SS}(\lambda)$ in (2.11) and their proposed estimator for the number of true nulls $\hat{n}_{0,ST}$ is the value of the fitted spline at $\lambda = 1$. In Storey et al. (2004), similarly as in Storey (2002), the authors choose λ using bootstrap, now with the objective of finding the value of λ that minimises the MSE of the resulting estimator.

Without any additional assumptions on the distribution of false null hypotheses, theoretical guarantees of proportion estimators cannot be studied. Let p -values come from a mixture distribution with CDF as in (2.9). If the density exists it is given by

$$f(t) = (1 - \pi_1) + \pi_1 f_1(t), \quad t \in (0, 1). \tag{2.14}$$

Under this model, the number of alternative hypotheses is a random variable with distribution $\text{Bin}(n, \pi_1)$ and the goal is to estimate π_1 (or π_0). The assumptions on the alternative distribution F_1 are usually mild. The *purity* assumption, introduced in Genovese and Wasserman (2004) is defined as:

$$\operatorname*{essinf}_t f_1(t) = 0.$$

This holds for example, if f_1 drops to zero, such that $f_1(t) = 0$ for $t \geq b$ where $b < 1$. The purity assumption makes the theoretical analysis easier, as it eliminates the *identifiability* problem. The identifiability problem arises when the unknown components (π_1, F) of the mixture model (2.9) are not uniquely identified, such that for all t,

$$\begin{aligned} F(t) &= \pi_0 t + \pi_1 F_1(t) \\ &= \pi_0^* t + \pi_1^* F_1^*(t), \end{aligned}$$

for some $\pi_0^* \neq \pi_0$ and $F_1^* \neq F_1$. In general, if a model is not identifiable then its parameters cannot be uniquely estimated. However, under the purity assumption, π_1 is uniquely determined as

$$\pi_1 = 1 - \operatorname*{essinf}_t f(t),$$

and

$$F_1(t) = \frac{F(t) - \pi_0 t}{\pi_1}.$$

Under the purity assumption, the consistency and the asymptotic normality of the estimators from Storey's family (2.12) as estimators for $\frac{F(\lambda)-\lambda}{1-\lambda}$ follows easily (see Proposition 3.2 in Genovese and Wasserman (2004)). If λ is such that $F_1(\lambda) = 1$, then $\hat{\pi}_1(\lambda)$ is a consistent estimator of the false null proportion π_1.

We finish this section by describing the two methods that can be seen as variants of Storey's estimator, involving some additional steps and calculations. In Jiang and Doerge (2008), the authors propose the average estimate approach. The idea is to improve Storey's estimator by aggregating information for different values of λ. For a given B, we define $\lambda_k = (k-1)/B$, $k = 1, \ldots, B$. To reduce the underestimation caused by small λ values, few of the smallest λ's will be taken out of consideration, so only $\lambda_{i-1}, \ldots, \lambda_B$ are considered. The choice of i and B is explained below. The resulting true null proportion estimator is the average of Storey's estimators:

$$\hat{\pi}_0^{JD} = \frac{1}{B-i+2} \sum_{j=i-1}^{B} \frac{1 - \hat{F}_n(\lambda_j)}{1-\lambda_j}, \tag{2.15}$$

To explain how i is chosen, notice that for small λ_i, we expect $\hat{F}_n(\lambda_{i+1}) - \hat{F}_n(\lambda_i) \geq \frac{1}{B-i+1}(1 - \hat{F}_n(\lambda_i))$, because of the effect of the false null distribution. Having the opposite hold is a sign of weakening alternative, thus i is defined as

$$i = \min \left\{ i : \hat{F}_n(\lambda_{i+1}) - \hat{F}_n(\lambda_i) \leq \frac{1}{B-i+1}(1 - \hat{F}_n(\lambda_i)) \right\}.$$

Regardless, the sum in (2.15) starts from $i-1$. The authors explain that λ_{i-1} is included in order to decrease the variance. B is chosen from set $I = \{5, 10, 20, 50, 100\}$ using bootstrap. The optimal B is the one leading to the estimator whose MSE is the smallest.

In Cheng et al. (2015), a version of Storey's method based on estimating the tail of the alternative distribution is proposed. Considering the mixture model (2.9), we have

$$\begin{aligned}\pi_0 &= \frac{F_1(x) - F(x)}{F_1(x) - x}, \quad x \in [0,1] \\ &= \frac{n\bar{F}(x) - n\bar{F}_1(x)}{n(1-x) - n\bar{F}_1(x)},\end{aligned}$$

where $\bar{F} = 1 - F$ and $\bar{F}_1 = 1 - F_1$. Let $Q(x) := \bar{F}_1(x)$ be the tail of the alternative distribution of p-values and let $\hat{Q}(x)$ be an estimator of $Q(x)$. The proposed family of estimators is

$$\hat{\pi}_0^{CGT}(\lambda) = \frac{n(1 - \hat{F}_n(\lambda)) - n\hat{Q}(\lambda)}{n(1-\lambda) - n\hat{Q}(\lambda)}.$$

The final estimator is obtained by averaging

$$\hat{\pi}_0^{CGT} = \frac{1}{J} \sum_{\lambda_j \in \Lambda} \min\{1, \max\{0, \hat{\pi}_0^{CGT}(\lambda)\}\}.$$

where $\Lambda = \{0.20, 0.25, \ldots, 0.50\}$ and consequently $J = 7$. This method is realised only for Gaussian mean testing. The nonzero means are allowed to be different for each test, such that the distribution of the test statistics under the ith false null hypothesis is $N(\mu_i, \sigma_i)$. The distribution of the p-values (one or two-sided) under the false null is a function of the effect size $\delta = \mu/\sigma$. It is assumed that for each test a sample of size m is available, making the nonzero effect estimation, and thus the tail estimation, more precise. However, the tail estimator requires an initial estimator of the proportion, in

order to select the small p-values used to estimate the tail of the alternative. As the initial estimator they use the one form Storey et al. (2004).

2.2.2 Density based methods

Density mixture model (2.14) with purity assumption is considered in Swanepoel (1999). Precisely, they assume that there exists $\theta \in (0, 1)$ such that $f_1(\theta) = 0$. Their method for estimating the proportion is based on estimating the minimum value of the density denoted as $\widehat{f(\theta)}$, since it yields a plug-in estimator for the alternative proportion:

$$\hat{\pi}_1^S = 1 - \widehat{f(\theta)}. \tag{2.16}$$

$\widehat{f(\theta)}$ can be estimated using kernel density estimator, but their proposed estimator for $\widehat{f(\theta)}$ is based on spacings between p-values. Given the increasingly sorted sample of random variables $p_{(1)}, \ldots, p_{(n)}$, spacings are defined as differences $p_{(i+1)} - p_{(i)}$. This definition can be generalised to k-spacings, defined as $p_{(i+k)} - p_{(i)}$. Their statistic is based on the maximal $2s_n$ spacing defined as

$$M_n = \max_{1 \leq i \leq n} \left(p_{(i+s_n)} - p_{(i-s_n)} \right),$$

where $s_n \to \infty$ as $n \to \infty$ is a nonrandom sequence of integers. From the earlier literature on spacings (Barbe, 1992; Deheuvels, 1984), it is known that M_n is related to the minimum of the density function. Intuitively, maximal spacing is likely to correspond to the interval where f has dropped to its minimum, such that it is constant on an interval $(\frac{i+s_n}{n}, \frac{i-s_n}{n})$. On this interval, the quantile function is linear with slope $1/f(\theta)$. It holds that

$$M_n = \frac{2s_n/n}{f(\theta)},$$

which motivates the following estimator:

$$\widehat{f(\theta)} = \frac{2s_n/n}{M_n}.$$

They prove the consistency and the asymptotic normality of their estimator (2.16) by proving that those properties hold for $\widehat{f(\theta)}$ as an estimator of $f(\theta)$.

In Langaas et al. (2005) they propose three estimators based on the idea of estimating the minimum value of the density, similarly as in Swanepoel (1999). However, they do not give any theoretical results for these estimators. The initial assumption is that f is decreasing and $f(1) = 0$. The first estimator is based on the minimum of the Grenander nonparametric maximum likelihood density estimator for decreasing densities. The second estimator uses the value of the Grenander density estimate on the longest constant interval - which is assumed to be the the interval where f drops to zero. For the third estimator they add another assumption, requiring that f is convex. In this case, the estimator of the minimum value is the nonparametric maximum likelihood estimator for convex densities evaluated at 1.

In Celisse and Robin (2010), similarly to the second method in Langaas et al. (2005), they estimate the proportion by estimating the value of the density on the longest interval where it is constant. For the density estimation they use histograms and cross-validation method to choose the best histogram. This means that they approximate the density with a piecewise constant functions. The proportion estimate is given by the height of the histogram on an interval where it is minimum. They prove the consistency of their estimator under the assumption that $f_1(t) = 0$ for $t \geq b$ and some $b \in [0, 1]$.

Related to the density estimation approach is the approach by Efron (2007). The focus of this paper is in estimating the local false discovery rate, and the proportion estimator is obtained as a by-product. The main difference from the other density

approaches is that instead of considering a mixture model for p-values, they transform p-values to z-values (for example, $z = \Phi^{-1}(1-p)$ for one-sided testing). The true null z-values are then expected to have $N(0,1)$ distribution, and be concentrated around zero. Consequently, the problem of estimating the proportion by estimating the minimum value of the p-values density translates to the problem of fitting a standard normal density curve around 0 to the density of the z-values. Additionally, Efron (2007) considers the true null density estimation of the z-values, that sometimes might not be standard normal but $N(\mu_0, \sigma_0^2)$. No theoretical guarantees for the proportion estimator are provided.

2.2.3 Empirical process-based methods

Another category of methods are those based on the empirical CDF of p-values, that use the results from the empirical process theory for the proof of theoretical guarantees. Let $E_n(t)$ be the empirical CDF of a uniform $U[0,1]$ sample of size n. The uniform empirical process is defined by

$$\left\{\sqrt{n}(E_n(t) - t), 0 \leq t \leq 1\right\}.$$

In the empirical process literature, inequalities bounding the supremum of the uniform empirical process are of the form

$$P\left(\sup_{t\in(0,1)} \frac{E_n(t) - t}{\delta(t)} \geq \beta_{n,\alpha}\right) \leq \alpha, \tag{2.17}$$

for different bounding functions $\delta(t)$ and bounding sequences $\beta_{n,\alpha}$. For proportion estimators based on the empirical CDF of p-values, these inequalities can be used to

prove the results of the form

$$P(\hat{\pi}_1 \leq \pi_1) \geq 1 - \alpha.$$

This implies underestimation of π_1 with high probability, and the interval $(\pi_1, 1)$ can be seen as a level $1 - \alpha$ confidence interval.

One of the inequalities of this type is the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality, where $\delta(t) = 1$. DKW inequality upper bounds the maximum distance between the empirical and the true CDF. The bound does not depend on the true distribution, and can be used to determine $\beta_{n,\alpha}$ in (2.17). For the uniform distribution, the DKW inequality is given by:

$$P\left(\sup_{t\in(0,1)} |E_n(t) - t| > \varepsilon\right) \leq Ce^{-2n\varepsilon^2}, \quad \text{for every } \varepsilon > 0. \tag{2.18}$$

In Genovese and Wasserman (2004) it is used to prove the confidence statements for their estimator (see Theorem 3.1 therein):

$$\hat{\pi}_1^{\mathrm{GW}} = \sup_{t\in(0,1)} \frac{F_n(t) - t - \beta'_{n,\alpha}}{1 - t},$$

where the bounding sequence $\beta'_{n,\alpha} = \sqrt{\log(2/\alpha)/2n}$ is obtained from the DKW inequality. Their idea was further developed in Meinshausen and Rice (2006) where they found the bounding function yielding the "best" estimator is $\delta(t) = \sqrt{t(1-t)}$ (see Theorem 3 therein). Their proposed estimator is

$$\hat{\pi}_1^{\mathrm{MR}} = \sup_{t\in(0,1)} \frac{F_n(t) - t - \beta''_{n,\alpha}\delta(t)}{1 - t}. \tag{2.19}$$

The bounding sequence $\beta''_{n,\alpha}$ is obtained using the result on the limiting Gumbel distribution of the supremum of weighted empirical process. Aside from the confidence

interval result which follows immediately, they also prove the consistency of their estimator, in the sense that

$$\frac{\hat{\pi}_1^{\mathrm{MR}}}{\pi_1} \xrightarrow{P} 1,$$

under some conditions on the strength of the alternative distribution and π_1 . We also note that, as in most of the other papers, the estimator in Meinshausen and Rice (2006) is also a slope-based estimator, similar to Storey's plug-in family of estimators given in (2.12).

With Storey's family, the slopes containing the information about the proportion are calculated from point $(\lambda, F_n(\lambda))$ for a given λ, to $(1, 1)$, leading to the family of estimators (2.12). For the estimators in Genovese and Wasserman (2004) and Meinshausen and Rice (2006), the slopes are calculated from point $(t, F_n(t) - \beta_{n,\alpha}\delta(t))$ to $(1, 1)$ and then a supremum is taken over t. For a fixed t, this would result in a larger slope than Storey's, as $F_n(t) - \beta_{n,\alpha}\delta(t) < F_n(t)$, leading to smaller estimates for π_1. However, taking the supremum as in (2.19) gives an estimator for which $\hat{\pi}_1 < \pi_1$ holds with large probability, with minimal amount of underestimation. $F_n(t) - \beta_{n,\alpha}\delta(t)$ acts as a uniform lower bound for the true CDF, $F(t)$, and the proof of underestimation relies on inequality (2.17). Following this idea, a confidence interval for π_1 is also proposed by Li and Siegmund (2015).

The most restrictive model, which also allows for detailed theoretical analysis, is Gaussian mixture model as a model for the distribution of test statistics rather than p-values:

$$T \sim \pi_0 N(0, 1) + \pi_1 N(\mu, 1).$$

For this model, the proportion estimator that is also a lower boundary of the confidence interval of the form $(\hat{\pi}_1, 1)$ is proposed in Cai et al. (2007). The consistency of their estimator is proved to hold over the whole detectable region of the rare-weak model (2.26) introduced in Section 2.1.1. It follows that the problem of estimating π_1 is not

harder than the problem of testing if $\pi_1 = 0$. They also provide the minimax theory for the problem and prove that their estimator has the MSE within the logarithmic factor of the minimax risk.

2.3 Signal estimation and the Gaussian sequence model

In this section we review the literature on the Gaussian sequence model that is considered in Chapter 5, where we propose a new thresholding estimator for this model. High level overview of the rich literature on this model is given in Section 2.3.1. The essential notions of sparsity and thresholding are described in Section 2.3.2. The Gaussian sequence model can also be studied from the multiple testing perspective, which is discussed in Section 2.3.3, along with the relevant literature.

2.3.1 The Gaussian sequence model

The Gaussian sequence model is defined as

$$X_i = \mu_i + \varepsilon_i, \quad i = 1, \dots, n, \tag{2.20}$$

where $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, $\mu = (\mu_1, \dots, \mu_n)$ is the unknown mean vector, and σ^2 is the noise variance. This model is commonly studied with the goal to make inference on the mean vector, under some assumptions on its structure, based on the sample $X_1, \dots, X_n$. As the number of unknown parameters is growing with sample size, this is a problem of nonparametric statistics. This simple model is encountered in a wide range of research areas and is a part of both classical and modern statistical literature. It gained popularity in the '90s where it was notably applied to nonparametric function

estimation and statistical signal processing (Donoho and Johnstone, 1994, 1995, 1998). One way to see the model given in (2.20) in this light is directly as a model for sampled values of the underlying function in the time domain. The other way is to consider (2.20) in the frequency domain, as a model for the Fourier coefficients of a signal at different frequencies, or as a model for the wavelet coefficients of a signal in the wavelet domain. Consequently, the methods for estimating the mean vector μ extend to methods for estimating signals, functions or images when those are transformed to a frequency domain.

An unpublished monograph by Johnstone (2017) covers the rich theory of minimax estimation of μ under various constraints on its structure. It also contains many of the author's work on the topic of nonparametric function estimation using wavelets, which started with the seminal paper on nonparametric function estimation using wavelets (Donoho and Johnstone, 1994).

Mean vector is usually assumed to belong to a subset $\Theta_n \subset \mathbb{R}^n$. Some examples of Θ_n include n-dimensional balls and ellipsoids in l_p norms. The performance of an estimator is measured by comparing the rate of its worst case risk to the rate of the minimax risk $R_n(\Theta_n)$, where

$$R_n(\Theta_n) = \min_{\hat{\mu}} \max_{\mu \in \Theta_n} \mathbb{E}_\mu L(\hat{\mu}, \mu),$$

over a given set Θ_n and loss function $L(\hat{\mu}, \mu)$. For a given Θ_n and loss function, the estimator is minimax if its worst case risk is equal to the minimax risk $R_n(\Theta_n)$. The estimator is asymptotically minimax if its worst case risk is asymptotically equal to the minimax risk. The estimator is adaptively asymptotically minimax if its worst case risk is of the same order as the minimax risk across different parameter spaces or loss functions.

The most common loss function is the squared error loss $L(\hat{\mu}, \mu) = \|\hat{\mu} - \mu\|_2^2$, but loss functions based on other norms l_r, for $1 \leq r < 2$, or quasi norms for $r < 1$, are also considered $L(\hat{\mu}, \mu) = \|\hat{\mu} - \mu\|_r^r$.

Broadly, estimators of μ can be categorised as linear or non-linear. Linear estimators are linear in the data, and have the form $\hat{\mu} = C\mathbf{X}$ where C is an $n \times n$ matrix and $\mathbf{X} = (X_1, \ldots, X_n)$. C cannot depend on the sample in a way that would make the estimator non-linear in the data. The most popular class of non-linear estimators are thresholding estimators. Particularly, given a threshold level λ, we can consider hard or soft thresholding function

$$t_\lambda^H(x) = x\mathbb{I}\{|x| > \lambda\},$$
$$t_\lambda^S(x) = (|x| - \lambda)\,\mathrm{sgn}(x)\mathbb{I}\{|x| - \lambda \geq 0\}.$$

Applying these functions coordinate-wise on the sample vector $\mathbf{X}$, we get hard and soft thresholding estimators of μ:

$$\hat{\mu}^H = t_\lambda^H(\mathbf{X}),$$
$$\hat{\mu}^S = t_\lambda^S(\mathbf{X}).$$

These estimators are appropriate when some form of sparsity of the mean vector μ is assumed, limiting the proportion of large coordinates. We review different definitions of sparsity and thresholding estimators in Section 2.3.2.

In the recent literature, there is an interest in considering convex constraints on Θ_n in the Gaussian sequence model. The reason for this is that some important methods such as LASSO, isotonic regression and other shape constrained problems, can naturally be translated to the problem of signal estimation in the Gaussian sequence model under convex constraints. Much work has been done on developing minimax theory for

these problems. We provide a brief review of some of the relevant papers. Han et al. (2022) study the asymptotic normality and the optimality of the likelihood ratio test in model (2.20), testing $H_0 : \mu = \mu_0 \in \Theta_n$ versus $H_0 : \mu \neq \mu_0$ under convex constraints on $\Theta_n \ni \mu$. Properties of the least squares estimator under convex constraints on Θ_n, such as the magnitude of the estimation error and the optimality in the minimax sense, are considered in Chatterjee (2014). Extensions to his work and some results on the penalised least squares estimators under convex constraints on Θ_n are considered in Chen et al. (2017).

In Cai and Wei (2022), distributed estimation for the Gaussian sequence model under the communication constraints is considered. A new signal recovery method when μ is a sum of a sparse and a dense signal is proposed in Chernozhukov et al. (2017). Minimax estimation of linear and quadratic functionals of the Gaussian sequence model is studied in Collier et al. (2017) and Collier et al. (2018). Adaptive minimax estimators of l_0-sparse signal its unknown variance parameter σ^2 and l_2 norm are proposed in Comminges et al. (2021). The model considered therein generalises the Gaussian sequence model by considering iid noise with an unknown distribution, not necessarily Gaussian.

Most of the mentioned papers do not consider the problem of estimating the nuisance parameter σ. When the number of nonzero coordinates is small, Johnstone (2017) and Comminges et al. (2021) propose to use the median M-estimator of scale for estimating σ as

$$\hat{\sigma} = \frac{\text{MAD}}{0.6745}.$$

Another possible method for estimating the standard deviation in the Gaussian sequence model is proposed in Lenth (1989), and it is based on removing the large absolute values (that possibly have a nonzero mean), and assuming that the trimmed sample is a sample from $N(0, 1)$. The standard deviation estimator is obtained from the

median of the absolute values of the trimmed sample, as the median of the half-normal distribution with parameter σ^2 is

$$\sigma\sqrt{2}\text{erf}^{-1}(1/2),$$

where erf^{-1} is the inverse of the Gaussian error function. The estimator proposed is

$$\hat{\sigma} = 1.5 \times \text{med}(|X_i|, |X_i| \leq 2.5s_0), \tag{2.21}$$

where $s_0 = 1.5 \times \text{med}(|X_1|, \ldots, |X_n|)$. The constant 1.5 is used as $1.5 \approx 1/(\sqrt{2}\text{erf}^{-1}(1/2))$. In the remainder of this section we assume $\sigma = 1$.

2.3.2 Sparsity and thresholding

We now focus on sparse μ vectors. There are different ways of defining sparsity, the most intuitive one limits the proportion of nonzero coordinates in μ to be at most η:

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{\mu_i \neq 0\} \leq \eta.$$

More generally, vectors with only a small proportion of significantly large coordinates are considered sparse. It can be achieved by imposing the power-law decay of the decreasingly sorted absolute values in the sequence, $|\mu|_{(n)}, \ldots, |\mu|_{(1)}$ as:

$$|\mu|_{(n-k+1)} \leq \eta n^{1/p} k^{-1/p}.$$

For given η and p, the sequences satisfying the inequality above define a *weak l_p ball* of radius η. Another way of imposing sparsity is by considering *strong l_p balls*. Strong l_p

ball of average radius η contains sequences $(\mu_1, \ldots, \mu_n)$ such that

$$\frac{1}{n}\sum_{i=1}^{n}|\mu_i|^p \leq \eta^p.$$

Specifically, to impose sparsity for strong l_p balls, $p < 2$ is considered. As an illustrative example for this, consider sequences $(1/\sqrt{n}, \ldots, 1/\sqrt{n})$ and $(1, 0, \ldots, 0)$ which have equal l_2 norm. However, the l_p norm of the first one is increasing as p decreases. Thus, out of the signals with equal l_2 norm, sparse are those that have small l_p norm for $p < 2$.

We introduce a few thresholding methods commonly found in the literature. The methods introduced below assume $\sigma = 1$ and do not discuss the case when σ is unknown and to be estimated.

A threshold value that proved optimal in the minimax sense for the Gaussian sequence model, and usually referred to as the universal threshold is $t_U = \sqrt{2\log n}$. For minimaxity results of the soft and hard thresholding estimators we refer the reader to Johnstone (2017). The factor $\sqrt{2\log n}$ was chosen as it is with large probability greater than the maximum in a sequence of standard Gaussian variables. Precisely, it holds that:

$$P(\max_{i=1,\ldots,n}|Z_i| \geq \sqrt{2\log n}) \leq \frac{1}{\sqrt{\pi \log n}},$$

which goes to zero for $n \to \infty$. Another thresholding method is based on Stein (1981), where an unbiased estimate of the l_2 risk is proposed for an arbitrary estimate of the multivariate Gaussian mean is considered. In the special case of soft-thresholding estimators, the risk is estimated as the following function of the threshold:

$$\hat{U}(t) = n + \sum_{k=1}^{n} X_k^2 \wedge t^2 - 2\sum_{k=1}^{n} I\left\{X_k^2 \leq t^2\right\}.$$

Based on this estimate, a thresholding method is proposed by Donoho and Johnstone (1995) which uses threshold that minimises the above risk estimate.

Bayesian approaches can be categorised as empirical Bayesian or fully Bayesian. Sparsity of μ is usually imposed through selecting the appropriate prior distribution, a mixture of Dirac delta and some other distribution. Such prior is also called spike and slab prior in the literature. This was notably used in Johnstone and Silverman (2004), where an empirical Bayes method is proposed for signal estimation in the Gaussian sequence model. The estimator is based on an observation that if a non-zero component of the prior distribution for μ is symmetric about zero and unimodal, then the median of the posterior distribution has a thresholding property. This posterior median vector defines the estimator of the mean vector in Johnstone and Silverman (2004). The threshold depends on the sparsity parameter, so the method adapts to the unknown sparsity.

More Empirical Bayes methods are proposed in Jiang and Zhang (2009), Martin and Walker (2014), Banerjee et al. (2020), Belitser and Nurushev (2020). Construction of credible sets for the estimated parameters following spike and slab prior distributions is considered in Castillo and Szabó (2020). Some fully Bayesian approaches can be found in Carvalho et al. (2010), Castillo and van der Vaart (2012) and Bhattacharya et al. (2015). Algorithms improving the numerical accuracy and computational speed of some Bayesian methods is proposed van Erven and Szabó (2020).

2.3.3 Multiple testing in the Gaussian sequence model

First we will describe a popular global testing method that considers the Gaussian sequence model. The Higher Criticism statistic (HC), is proposed by Donoho and Jin (2004) for the problem of detecting sparse mixtures, which can also be seen as a problem of testing the global null hypothesis. We describe the problem under the

Gaussian model which is the primary model considered in Donoho and Jin (2004). Consider the following multiple testing problem given by the sequence of hypotheses:

$$\begin{aligned} H_0^i &: X_i \sim N(0,1) \\ H_1^i &: X_i \sim N(\mu,1), \quad \mu > 0. \end{aligned} \tag{2.22}$$

If the global null holds, then $X_i \sim N(0,1)$, for all $i = 1, \ldots, n$, otherwise some X_i have $N(\mu, 1)$ distribution. The global null hypothesis test can also be stated as:

$$H_0 : X_i \sim N(0,1), \quad 1 \leq i \leq n \tag{2.23}$$

$$H_1 : X_i \sim (1-\varepsilon)N(0,1) + \varepsilon N(\mu,1), \quad 1 \leq i \leq n, \tag{2.24}$$

where it is of interest to consider small values of the parameters μ and ε, and have a test be able to reject the null hypothesis and detect a sparse mixture of the sample, even in these weak alternative cases. The HC statistic is proposed for this testing problem, and it can be used for mixtures other than Gaussians, which are also considered in Donoho and Jin (2004). The sample is first transformed to the sequence of p-values. For the one sided alternative as in (2.22), for example, the p-values are calculated as

$$p_i = 1 - \Phi(X_i),$$

where Φ is the CDF of the $N(0,1)$ distribution. Under the global null, the p-values have $U[0,1]$ distribution. Let $p_{(1)} \leq \cdots \leq p_{(n)}$ be the order statistics of the p-values sequence. The HC statistic is defined as

$$\max_{i=1,\ldots,n} \mathrm{HC}(i),$$

where

$$\mathrm{HC}(i) = \sqrt{n}\frac{i/n - p_{(i)}}{\sqrt{p_{(i)}(1-p_{(i)})}}, \quad i = 1, \ldots, n. \tag{2.25}$$

The use of HC statistic beyond the global testing problem is discussed in Chapter 4. Its optimality for the problem of global testing is discussed in terms of its power to detect a rare-weak Gaussian mixture. The distribution under the alternative is called rare-weak Gaussian mixture if the following conditions hold, keeping both the proportion of nonzero means and their value small:

$$\begin{aligned} \varepsilon = \varepsilon_n = n^{-\beta}, \quad \beta \in \left(\frac{1}{2}, 1\right) \\ \mu = \mu_n = \sqrt{2r\log(n)}, \quad r \in (0,1). \end{aligned} \tag{2.26}$$

The conditions on (r, β) under which the optimal Neyman-Pearson likelihood ratio test has the sum of type I and type II errors go to 0 are called distinguishability conditions. In other words, the alternative hypothesis is distinguishable if it is strong enough so that there exists a test that has asymptotically full power. These results can be found for example in Chapter 8 of Ingster and Suslina (2003). Using the notation as in Donoho and Jin (2004), the detection boundary is

$$\rho^*(\beta) = \begin{cases} \beta - 1/2, & 1/2 < \beta \leq 3/4 \\ (1-\sqrt{1-\beta})^2, & 3/4 < \beta < 1, \end{cases}$$

meaning that for $r > \rho^*(\beta)$ the sum of type I and type II errors goes to 1 for the likelihood ratio test, while for $r < \rho^*(\beta)$ it goes to 0. The two dimensional domain (r, β) was later termed as phase space, as $\rho^*(\beta)$ divides it into two regions or phases. Below the curve $r = \rho^*(\beta)$ is the undetectable region, and above is the detectable region. In Donoho and Jin (2004) it is proved that the HC test has full power in the

detectable region. Cai et al. (2007) show that ε_n can be consistently estimated, in the sense that $\hat{\varepsilon}_n/\varepsilon_n \xrightarrow{P} 1$ whenever (r, β) belongs to the detectable region. In Xie et al. (2011) an "almost exact recovery" boundary was identified,

$$\rho_{\text{recovery}}(\beta) = (1 + \sqrt{1-\beta})^{1/2},$$

such that for the parameters above it, the oracle procedure by Sun and Cai (2007) identifies all the signals correctly with probability going to 1. Let F_n be the empirical CDF of p-values and F_0 the $U[0,1]$ CDF. HC objective sequence can be written as:

$$\text{HC}(i; p_{(i)}) = \sqrt{n}\frac{F_n(p_{(i)}) - F_0(p_{(i)})}{\sqrt{F_n(p_{(i)})(1 - F_n(p_{(i)}))}}.$$

Finally, we note that, since it is comparing empirical to the theoretical distribution of p-values, the HC statistic is related to the Kolmogorov-Smirnov and the Anderson-Darling statistics.

In general, thresholding estimators of the mean in the Gaussian sequence model (2.20) allow for multiple testing interpretation. Related multiple testing problem can be described with the sequence of hypotheses

$$\begin{aligned} H'_{0,i} &: X_i \sim N(0,1) \\ H'_{1,i} &: X_i \sim N(\mu, 1), \quad \mu \neq 0, \end{aligned} \tag{2.27}$$

and multiple testing methods that choose which hypotheses are false null can be seen as hard thresholding methods of signal estimation. Let $S = \{i : \mu_i \neq 0\}$ be the subset of nonzero mean, signal variables. Both multiple testing methods and hard thresholding

estimators give an estimate for S

$$\hat{S} = \{i : \hat{\mu}_i \neq 0\}.$$

We now list some relevant papers, that consider multiple testing in the Gaussian sequence model. In Rabinovich et al. (2020), as an error rate they consider the sum FNR + FDR. They prove that both the classical BH procedure and the one in Barber and Candés (2015) are rate optimal in this sense in case of (generalised) GSM. This error rate was also investigated in Abraham et al. (2021).

Multiple testing and signal estimation procedures have different objectives. While in multiple testing the goal is to control a certain error rate, in signal estimation the goal is to minimize the distance between the true and the estimated signal. However, multiple testing procedures have been used for signal estimation in model (2.20), see for example Abramovich et al. (2006) and Abramovich et al. (2010) where the BH procedure is proven to be asymptotically minimax for signal estimation. Gaussian mixture model with unknown mean and variance of both components was considered in Cai and Jin (2010) from a multiple testing point of view. FDR control of the empirical Bayesian procedure using spike and slab prior, which was previously used for signal estimation (see Castillo and Szabó (2020)) is considered in Castillo and Roquain (2020) and Abraham et al. (2022). Jeng (2016) proposed a method for dividing set of p-values into three sets: signal, noise and indistinguishable set. In Du et al. (2021), a multiple testing method for the multivariate Gaussian model is considered.

Chapter 3

Difference of Slopes method for estimating the false null proportion

In this chapter we propose a new method for the problem of estimating the proportion of false null hypotheses in multiple testing. We propose the Difference of Slopes statistic, which yields a threshold separating the small p-values from the large ones using change-point ideas. This threshold is then used in combination with Storey's estimator (Storey, 2002) to get the proportion estimator.

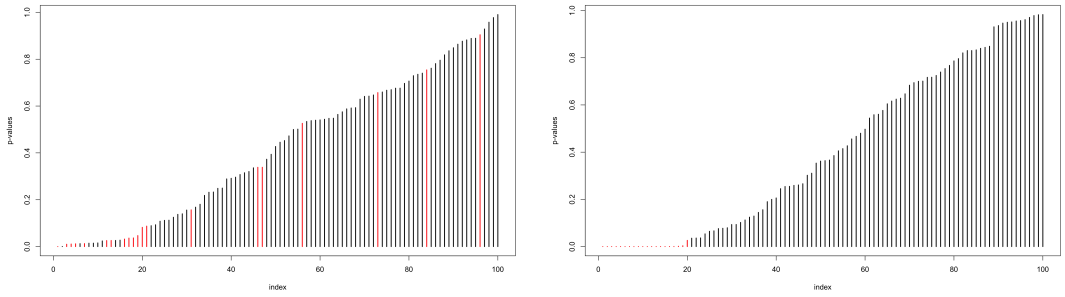
This chapter is organised as follows. In Section 3.1 we motivate our methodology by describing the relevant literature. In Section 3.2 we describe our proposed method. Theoretical results are provided in Section 3.3, and simulation results in Section 3.4. We look at possible extensions in Section 3.5 and a brief discussion is given in Section 3.6.

3.1 Motivation

The novelty of our approach to the problem of estimating the false null proportion is that we adopt a particular change-point perspective of the problem that has not

been explored before. The idea behind the proposed method is to approximate the quantile function of the p-values with a piecewise linear function with one change-point in slope. For this, it is enough to estimate the change-point in slope from the sample of p-values, as the quantile function takes value 0 at 0 and 1 at 1. This is achieved by our proposed Difference of Slopes (DOS) statistic. To illustrate our perspective, consider the sequence of sorted p-values $(i/n, p_{(i)})$ (the quantile plot), shown in Figure 3.1. This sequence approximates the quantile function of the p-values. If the alternative is very strong, then we might have all the false null p-values smaller than all the true null p-values, as seen in Figure 3.1 on the right hand plot. In that case a change-point exists, separating the false null from the true null p-values. This is an unrealistic assumption, and it is more common to observe a smooth change problem of transition between the false null and the true null p-values, with purely false null p-values at the beginning, the mixing interval in the middle, and purely true null p-values at the end . However, if the alternative is "rare and weak" we can expect the smallest true null p-value to be smaller than the smallest false null p-value (see Meinshausen and Rice (2006) for a discussion on this), in which case there is no interval at the beginning containing only the false null p-values . Similarly, for weak alternatives, false null p-values can take values close to 1. An illustration for the weak alternative case is given in the left-hand plot in Figure 3.1. In this case there is no change-point in the usual sense, however the estimated change-point would act as a classification threshold, separating smaller, mostly false null p-values, from the larger, mostly true null p-values.

After estimating the change-point, to get the false proportion estimate, we consider p-values smaller than the DOS threshold. The estimate is obtained by subtracting the expected number of true null hypotheses from the number of p-values smaller than the threshold. This comes down to applying Storey's estimator (Schweder and Spjøtvoll, 1982; Storey, 2002) with the DOS threshold as the tuning parameter value.



(a) Weak alternative (b) Strong alternative

Fig. 3.1 Sorted sequence of 100 p-values. Red bars correspond to p-values of false null hypotheses and black bars to true null hypotheses.

It implies ignoring the false null p-values larger than the threshold, and thus possibly underestimating the false null proportion. Denote the false null proportion by π_1. In Theorem 1 we prove that our estimator $\hat{\pi}_1^{DOS}$ is asymptotically conservative, in the sense that

$$\hat{\pi}_1^{DOS} \xrightarrow{a.s.} \tilde{\pi}_1, \tag{3.1}$$

where $\tilde{\pi}_1$ is a constant such that $\tilde{\pi}_1 \leq \pi_1$. Thus, our estimator asymptotically has non-positive bias and it will not overestimate the true proportion. Conservative (or asymptotically conservative) estimators of the false null proportion are preferable as they guarantee the (asymptotic) FDR control of the adaptive Benjamini-Hochberg procedure (Benjamini and Hochberg, 2000). This approach is discussed in Section 2.2. With the exception of a few existing proportion estimators, such as the empirical process-based ones proposed in Meinshausen and Rice (2006) and Genovese and Wasserman (2004), theoretical properties were not investigated for most of the proposed proportion estimators in the literature. A simulation study in Section 3.4 shows that, for moderate sample sizes ($n = 1000$), the underestimation of the DOS estimator is less severe than of the consistent estimator proposed in Meinshausen and Rice (2006). Furthermore, simulations show that the variance of our method is among the lowest in sparse cases,

when the proportion of false null hypotheses is small. In this setting our estimator also has the lowest mean squared error. The performance of the DOS method is also investigated for small samples ($n = 50, 100$) and the results show that it outperforms the competitors for various values of π_1 and the strength of the effect under the alternative.

Using change-point ideas for solving multiple testing problems has been suggested in several papers. In Benjamini and Hochberg (2000), the authors remark that a possible approach to estimating the false null proportion would be to use some change-point detection method to find the end of the linear part in the quantile plot of the p-values. To the best of our knowledge, this idea was only explored in Turkheimer et al. (2001), where they propose a pseudo-sequential procedure that tests for uniformity by iteratively excluding the smallest p-values. However, this method is very conservative and we do not include it in the simulation study. Another link to the change-point literature is the Higher Criticism (HC) statistic by Donoho and Jin (2004), proposed for testing the global null hypothesis, of whether there are any false null p-values. The HC statistic is introduced it in Section 2.3.3. It is very closely related to Pontogram statistic by Kendall and Kendall (1980) used for testing for a change-point in the intensity of a Poisson process. The relationship between the two statistics is explored in Chapter 4.

3.2 Difference Of Slopes method

Let p-values come from a mixture distribution with CDF given by

$$F(t) = \pi_1 F_1(t) + \pi_0 t, \quad t \in [0, 1], \tag{3.2}$$

where $F_1(t)$ is a (weakly) concave function and the distribution F_1 is stochastically smaller than $U[0,1]$ distribution, in the sense that $F_1(t) \geq t$ for all $0 \leq t \leq 1$. The mixture model is common in the multiple testing literature, and the concavity assumption of the false null distribution is reflecting the assumption that the concentration of the false null p-values is decreasing on $[0,1]$. Let $p_{(1)}, \ldots, p_{(n)}$ be the sequence of increasingly sorted p-values from the model (3.2). We define the Difference of Slopes statistic $\hat{k}_{DOS}$ by

$$\hat{k}_{DOS} = \underset{nc_n \leq i \leq n/2}{\operatorname{argmax}} \, d(i), \tag{3.3}$$

where

$$d(i) = \frac{p_{(2i)} - p_{(i)}}{i/n} - \frac{p_{(i)}}{i/n} \tag{3.4}$$

$$= \frac{p_{(2i)} - 2p_{(i)}}{i/n}, \tag{3.5}$$

and c_n is such that $c_n \to 0$ and $nc_n \to \infty$. The choice of c_n is discussed in Sections 3.3 and 3.4.

The sequence $d(i)$, $1 \leq i \leq n/2$ is called the DOS sequence. The first term on the RHS of (3.4) is the slope of the line connecting points $(i/n, p_{(i)})$ and $(2i/n, p_{(2i)})$, and the second term is the slope of the line connecting $(0,0)$ and $(i/n, p_{(i)})$. Therefore, the sequence $d(i)$ can be seen as the sequence of slopes differences in the quantile plot, and $\hat{k}_{DOS}$ as the location where the difference in slopes is maximal. The procedure is illustrated in Figure 3.2, showing the quantile plot and the matching DOS sequence. We note that the DOS statistic can be seen as a method for finding an *elbow* or a *knee* in the quantile plot. Our proposed separation threshold t_{DOS} is the $\hat{k}_{DOS}$-th smallest p-value, $p_{(\hat{k})}$, and it is data-dependent:

$$t_{DOS} = p_{(\hat{k}_{DOS})}.$$

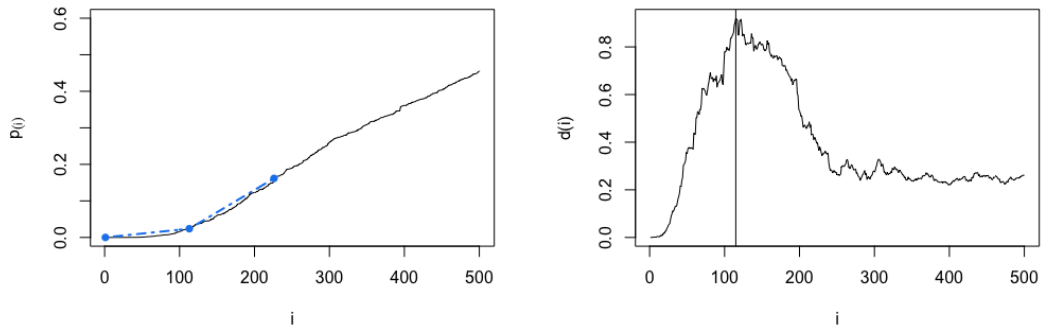


Fig. 3.2 The illustration of the DOS procedure on $n = 1000$ 1-sided p-values from the Gaussian model for the test statistics, where $H_0 : X \sim N(0, 1)$, $H_1 : X \sim N(3, 1)$ and the number of false null hypotheses is fixed $n_1 = 100$. Left: The sequence of the first 500 smallest p-values. The blue dash-dotted broken line reveals the detected change-point location and the corresponding symmetric interval with the largest slopes difference. Right: the DOS sequence $d_{(i)}$ with vertical line at the location of the maximum $\hat{k}_{DOS}$.

To obtain the proportion estimate using the DOS statistic, we plug t_{DOS} into Storey's family of estimators given in (3.6):

$$\hat{\pi}_1(\lambda) = \frac{\hat{F}_n(\lambda) - \lambda}{1 - \lambda}, \tag{3.6}$$

where $\hat{F}_n$ is the empirical CDF of the p-values, and get

$$\hat{\pi}_1^{DOS} = \frac{\hat{k}_{DOS}/n - t_{DOS}}{1 - t_{DOS}}. \tag{3.7}$$

In contrast to other Storey-based estimators in the literature that focus on values $\lambda \approx 1$ to reduce the bias, we aim to find the smallest possible λ leading to meaningful estimates. As described in Section 2.1.2, large values of λ increase the variance of the estimator, but taking λ too small causes underestimation. We remark that although the estimated change-point in Figure 3.2 is after the location $n_1 = 100$, this does not mean that our proportion estimate overestimates the proportion, as this change-point location is used with Storey's estimator to get the proportion estimate, and it is not

a proportion estimator itself. Our change-point approach provides a threshold t_{DOS} acting as an estimate of the upper bound of the support of the alternative distribution, and in that way reducing the bias and keeping the variance small in the estimator (3.7). This effect is best seen when the proportion of false null hypotheses is small, as shown in the simulation study in Section 3.4.

Excluding the first nc_n values from the search for maximum in (3.3) is a sufficient condition to guarantee the consistency results of Theorem 1, where we assume $c_n = n^{-\theta}$ for $\theta \in (0, 1)$. In Remark 2 we note that this is not the only possible option for c_n, and we discuss in detail the effects of different rates for c_n on the asymptotic results. Furthermore, the simulations in Section 3.4 show that excluding the values from the beginning of the sequence $d(i)$ does not affect the estimates considerably, so in practice it is not necessary to exclude any values.

We use symmetric intervals for slopes calculation, that is $[0, i/n]$ and $[i/n, 2i/n]$. In this way, for each candidate change-point, there is an (almost) equal number of p-values left and right from it. In this way, we focus more on the local patterns in the quantile plot, which is particularly useful for sparse cases, when π_1 is small. Furthermore, calculating slopes differences using the full set of p-values, that is $[0, i/n]$ and $[i/n, 1]$, results in severe underestimation of the proportion, as the values of the second slope are large and less variable in the beginning, than those of the first slope. In Section 3.2.1 we further motivate the use of symmetric intervals.

As ordered p-values are sample quantiles, we note that the DOS sequence is a proxy for the "ideal function" involving quantile function of the mixture distribution F, denoted by F^{-1}:

$$\frac{p_{(2i)} - 2p_{(i)}}{i/n} \approx \frac{F^{-1}(2i/n) - 2F^{-1}(i/n)}{i/n}. \tag{3.8}$$

Denote the ideal function by $h_F(t)$,

$$h_F(t) = \frac{F^{-1}(2t) - 2F^{-1}(t)}{t}. \tag{3.9}$$

It follows that the point of the maximum slopes difference is close to the maximum of the ideal function, and this is the change-point that the DOS is estimating:

$$\frac{\hat{k}_{DOS}}{n} \approx \tilde{\tilde{t}} := \operatorname*{argmax}_{t\in(0,1]} h_F(t). \tag{3.10}$$

Our main theoretical results are stated in Theorem 1 below. We discuss the assumptions and provide the proof in Section 3.3. They include the consistency results regarding (3.8) and (3.10), and the asymptotic conservativeness of the DOS proportion estimator, as mentioned in (3.1).

Theorem 1. *Let*

$$F(x) = \pi_1 F_1(x) + \pi_0 x, \quad x \in [0,1], \tag{3.11}$$

be the p-value distribution, where F_1 is a (weakly) concave function. Let $h_F(t)$ be defined as in (3.9), and assume that condition (A1) holds. Let $p_{(1)}, \ldots, p_{(n)}$ be the order statistics of the p-values sequence. Let $\hat{k}_{DOS}$ and $\hat{\pi}_1^{DOS}$ be as defined in (3.3) and (3.7), respectively, with $c_n = n^{-\theta}$, for some $\theta \in (0,1)$. It holds that

$$\hat{k}_{DOS}/n \overset{a.s.}{\to} \tilde{\tilde{t}} := \operatorname*{argmax}_{0\le t\le 1/2} \frac{F^{-1}(2t) - 2F^{-1}(t)}{t}, \tag{3.12}$$

$$p_{(\hat{k}_{DOS})} \overset{a.s.}{\to} F^{-1}(\tilde{\tilde{t}}), \tag{3.13}$$

$$\hat{\pi}_1^{DOS} \overset{a.s.}{\to} \frac{\tilde{\tilde{t}} - F^{-1}(\tilde{\tilde{t}})}{1 - F^{-1}(\tilde{\tilde{t}})} \le \pi_1. \tag{3.14}$$

3.2.1 Ideal behaviour and curvature interpretation

In this section we discuss the interpretation of the ideal quantities approximated by the DOS method. Firstly, we consider possible interpretations of the ideal change-point location $\operatorname{argmax}_t h_F(t)$, where $h_F(t)$ is the ideal function defined in (3.9). We also consider the relationship between the DOS statistic and the second derivative and curvature of the quantile function. Finally, we examine the amount of underestimation the DOS is asymptotically making in the Gaussian mixture case.

Let F^{-1}, the quantile function of the distribution F defined in (3.2), be continuous on $t \in [0,1]$, and its first two derivatives continuous on $(0,1)$. Define

$$\begin{aligned} H(t,a) &:= \frac{F^{-1}(t+a) - F^{-1}(t)}{a} - \frac{F^{-1}(t) - F^{-1}(t-a)}{a} \\ &= \frac{F^{-1}(t+a) - 2F^{-1}(t) + F^{-1}(t-a)}{a}, \end{aligned}$$

where $a \in (0,t]$ and $t \in [0,1]$. For $a = t$, $H(t,t) = \frac{F^{-1}(2t) - 2F^{-1}(t)}{t} = h_F(t)$, which is the ideal function the DOS sequence is approximating. Since $(F^{-1})'$ is an increasing function, by taking partial derivative in a, it follows that $H(t,a)$ is increasing in a for any fixed t. It follows that

$$\operatorname*{argmax}_{(t,a)} H(t,a) = \operatorname*{argmax}_{(t,t)} H(t,t) = (\tilde{\tilde{t}}, \tilde{\tilde{t}}), \tag{3.15}$$

meaning that the DOS "true change-point location" $\tilde{\tilde{t}}$ is actually the point of the largest slopes difference on symmetric intervals of arbitrary length in the quantile function. Because of the concavity assumption, it is enough to consider increasing symmetric intervals to estimate it. Because of the sample variability, scanning through all possible

window sizes by considering the statistic

$$\underset{i,j}{\operatorname{argmax}} \frac{p_{(i+j)} - 2p_{(i)} + p_{(i-j)}}{j} \tag{3.16}$$

would introduce noise in the estimator, as under the null $p_i \sim U[0,1]$, and it holds that

$$\operatorname{Var}\left(\frac{p_{(i+j)} - 2p_{(i)} + p_{(i-j)}}{j}\right) \approx \frac{2}{n^2 j}, \tag{3.17}$$

so the variance is larger for smaller window sizes j, and the statistic in (3.16) is inadequate. Note that if in (3.16) instead of j we divide by $\sqrt{j}$, the variance does not depend on j. Different scalings in the DOS sequence are considered in Section 3.5.

We now move on to discuss the interpretation of the point of maximum in (3.15), with the goal of understanding how the estimated DOS change-point is related to the underlying quantile function. For small a, function $H(t,a)$ has a second derivative interpretation. Second order Taylor expansion provides the following approximation:

$$H(t,a) = a(F^{-1})''(t) + o(a).$$

Thus, for small a_0 and any $t \in (0,1)$, $H(t,a_0) \approx a_0(F^{-1})''(t)$ so it holds that

$$\underset{t}{\operatorname{argmax}}\, H(t,a_0) \approx \underset{t}{\operatorname{argmax}} (F^{-1})''(t).$$

However, for larger a this approximation does not hold anymore. We use Mathematica for further analysis of the function $H(t,a)$ in the Gaussian case, that is when F is the CDF of the 1-sided p-values coming from the Gaussian mean testing problem $\pi_1 N(\mu,1) + \pi_0 N(0,1)$. An illustration of function $H(t,a)$ is given in Figure 3.3. As discussed above, the maximum of this function is achieved on the line $t = a$, and the point of maximum is the change-point that the DOS method approximately estimates.

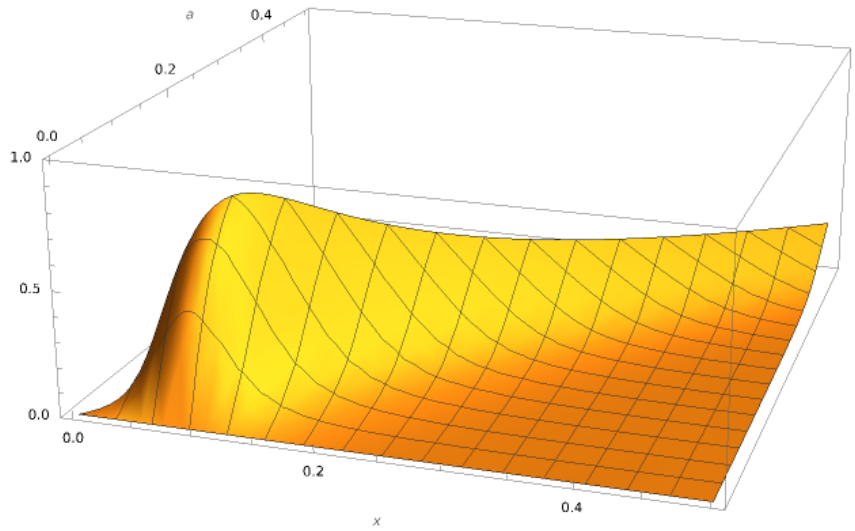


Fig. 3.3 Bivariate function $H(t, a)$ for the distribution of the 1-sided p-values from the Gaussian mean testing with $\pi_1 = 0.1$, $\mu = 3$.

To investigate how this point of maximum is related to the behaviour of F^{-1} directly, we recall that the DOS statistic can be seen as a method for finding the elbow in a (quantile) function. An elbow can be defined as a point of the maximum curvature of a function, $\operatorname{argmax}_t \kappa(t)$. We claim that the DOS change-point location is after the maximum curvature point of function F^{-1}. For the Gaussian mixture model, we investigate this relationship by comparing $\max_t h_F(t)$ to $\operatorname{argmax}_{t\in(0,1)} t(F^{-1})''(t)$ instead of the maximum curvature point, as it holds that $\operatorname{argmax}_{t\in(0,1)} t(F^{-1})''(t) > \operatorname{argmax}_{t\in(0,1)} \kappa_{h_F}(t)$. To prove this, consider

$$\begin{aligned}\operatorname*{argmax}_t t(F^{-1})''(t) &= \operatorname*{argmax}_{t\in[0,1]} t\frac{-f'(F^{-1}(t))}{f(F^{-1}(t))^3} \\ &= F\left(\operatorname*{argmax}_{y\in[0,1]} F(y)\frac{|f'(y)|}{f^3(y)}\right).\end{aligned}$$

Since $(\frac{f^2(y)}{1+f^2(y)})^{3/2}\frac{1}{F(y)}$ is a decreasing function of y we have

$$\operatorname*{argmax}_{y\in[0,1]} F(y)\frac{|f'(y)|}{f^3(y)} \geq \operatorname*{argmax}_{y\in[0,1]} \frac{|f'(y)|}{(1+f^2(y))^{3/2}},$$

which implies

$$\operatorname*{argmax}_t t(F^{-1})''(t) \geq F\left(\operatorname*{argmax}_{y\in[0,1]} \frac{|f'(y)|}{(1+f^2(y))^{3/2}}\right). \tag{3.18}$$

The expression on the RHS of (3.18) is the point of the maximum curvature of function F^{-1}, proving that $\operatorname{argmax}_{t\in(0,1)} t(F^{-1})''(t) > \operatorname{argmax}_{t\in(0,1)} \kappa_{h_F}(t)$. In Figure 3.4, the distance between $\tilde{\tilde{t}}$, the change-point the DOS is estimating, and $\operatorname{argmax}_{t\in(0,1)} t(F^{-1})''(t)$ is shown for the Gaussian mixture model for different values of μ (colour coded) and π_1 (on the x-axis). Numerical experiments show that, for the Gaussian mixture model, $\tilde{\tilde{t}}$ is always larger than $\operatorname{argmax}_{t\in(0,1)} t(F^{-1})''(t)$. This implies that the DOS change-point happens after the maximum curvature point of the quantile function.

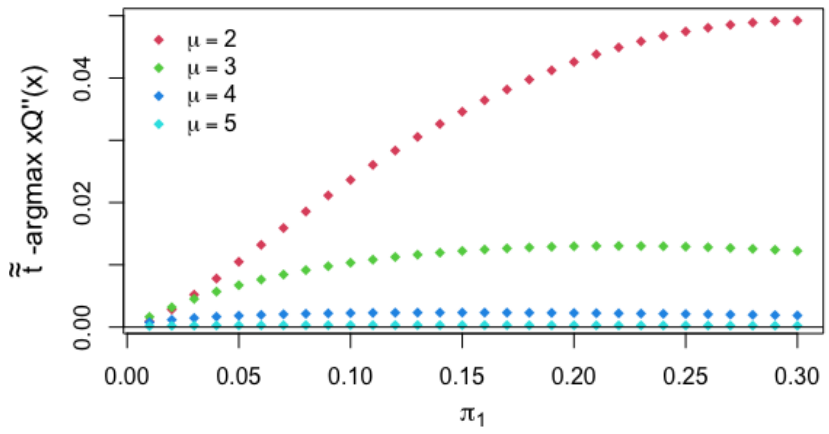


Fig. 3.4 The distance of the DOS true change-point and $\operatorname{argmax}_{x\in(0,1)} x(F^{-1})''(x)$.

We now turn to the issue of underestimation of the DOS estimator. As guaranteed by Theorem 1, the DOS estimator will underestimate the proportion, in the sense that $\hat{\pi}_1^{DOS}$ converges almost surely to $\tilde{\pi}_1 \leq \pi_1$. To investigate what part of the proportion our estimator is asymptotically able to estimate, in Figure 3.5 we plot the standardised difference $(\tilde{\pi}_1 - \pi_1)/\pi_1$ for different values of μ and π_1.

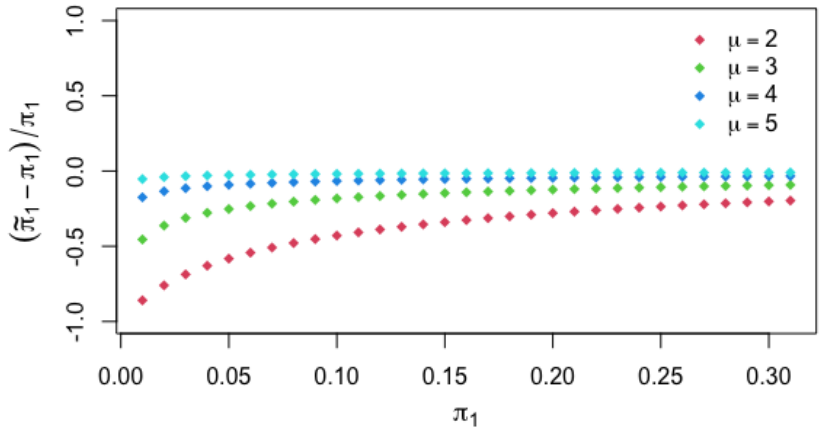


Fig. 3.5 The degree of proportion underestimating shown as the scaled distance between the DOS ideal proportion $\tilde{\pi}_1$ and the true proportion π_1 for the Gaussian mixture model for different values of μ.

Finally, we observe that the ideal change-point location $\tilde{\tilde{t}}$ gets close to the true false null proportion π_1 already for $\mu = 3$ in the Gaussian mixture case. This suggests that the change-point location itself can possibly be used for signal estimation and thresholding in the same way as the HC threshold of the HC statistic. This point of view is further discussed in Chapter 4. In Figures 3.6 and 3.7 we show the distance between $\tilde{\tilde{t}}$ and π_1.

3.3 Theoretical results

The consistency results presented in this section rely on the theory of quantile processes. Most of the existing theoretical results on this topic only cover the case of the uniform quantile process. Before presenting the main theorems we state two lemmas that connect the quantile process of the p-value distribution to the uniform quantile process. This will allow us to later use some existing results on the almost sure behaviour of the weighted uniform quantile process in the proof of Theorem 1.

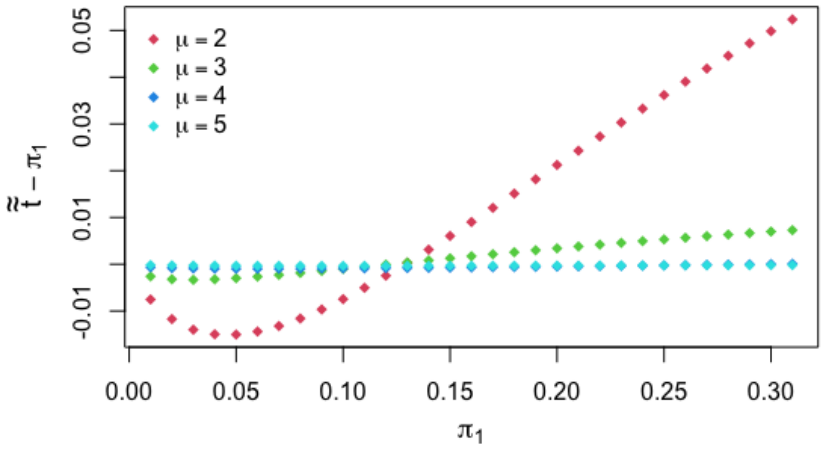


Fig. 3.6 The relationship between the DOS true change-point location and the true proportion for the Gaussian mixture model. For $\mu = 3$ already, the ideal change-point is close to the true proportion. This implies that if the signal is strong enough, the change-point location can be used as the proportion estimate and as a threshold for signal estimation.

3.3.1 Some useful lemmas

Lemma 1. *Let $X_1, \ldots, X_n$ be the sample from distribution*

$$F(x) = \pi_1 F_1(x) + \pi_0 x,$$

where F_1 is a weakly concave function. Let $\hat{F}_n$ be the empirical CDF of a sample $X_1, \ldots, X_n$, and $\hat{E}_n$ the empirical CDF of a sample of size n from $U[0,1]$ distribution. Let $\{q_n(y), y \in (0,1)\}$ be the quantile process of $X_1, \ldots, X_n$ and $\{u_n(y), y \in (0,1)\}$ the uniform quantile process defined as

$$q_n(y) = \sqrt{n}(\hat{F}_n^{-1}(y) - F^{-1}(y)),$$
$$u_n(y) = \sqrt{n}(\hat{E}_n^{-1}(y) - y),$$

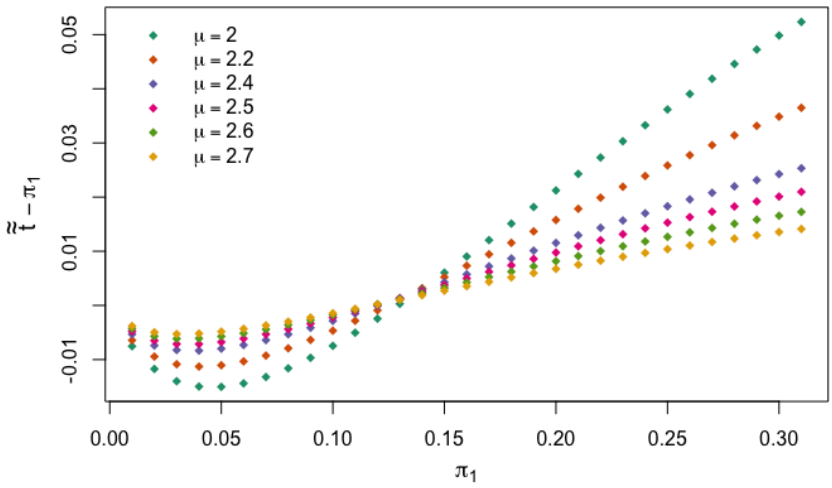


Fig. 3.7 The relationship between the DOS true change-point location and the true proportion for the Gaussian mixture model with small μ.

where $\hat{F}_n^{-1}$ *and* $\hat{E}_n^{-1}$ *are left continuous inverses of* $\hat{F}_n$ *and* $\hat{E}_n$*. It holds that*

$$q_n(y) \leq C u_n(y), \quad y \in (0,1), \tag{3.19}$$

where

$$C := \sup_{x,y} \frac{F^{-1}(y) - F^{-1}(x)}{y - x} = \frac{1}{\pi_0}. \tag{3.20}$$

Proof. As F_1 is a concave function on $[0,1]$, and F is a linear combination of F_1 and a linear function, F is also a concave function on $[0,1]$. It holds that the inverse of a continuous, concave and increasing function on an interval is convex on the same interval, so it follows that F^{-1} is a convex function. Using the upper bound for F

$$F(x) \leq \pi_1 + \pi_0 x, \quad x \in [0,1],$$

it holds that for any $0 \leq x < y \leq 1$

$$\begin{aligned}\frac{F^{-1}(y) - F^{-1}(x)}{y - x} &\leq \frac{1 - F^{-1}(x)}{1 - x} \\ &\leq \frac{1}{\pi_0},\end{aligned}$$

which proves (3.20). Let $X_{k;n}$ be the kth order statistic of the sample and let $(k-1)/n < y \leq k/n$. The following sequence of inequalities concludes the proof by proving (3.19).

$$\begin{aligned}q_n(y) &= \sqrt{n}(\hat{F}_n^{-1}(y) - F^{-1}(y)) \\ &= \sqrt{n}(X_{k;n} - F^{-1}(y)) \\ &= \sqrt{n}(F^{-1}(F(X_{k;n})) - F^{-1}(y)) \\ &= \sqrt{n}(F^{-1}(U_{k;n}) - F^{-1}(y)) \\ &= \sqrt{n}\frac{F^{-1}(U_{k;n}) - F^{-1}(y)}{U_{k;n} - y}(U_{k;n} - y) \\ &\leq \frac{1}{\pi_0}\sqrt{n}(\hat{E}_n^{-1}(y) - y) \\ &= \frac{u_n(y)}{\pi_0}.\end{aligned}$$

□

Lemma 2. *Under the assumptions of Lemma 1, it holds that*

$$P\left(\sup_{0<y<1} |q_n(y)| \geq x\right) \leq 2e^{-2x^2/C^2}.$$

Proof. To prove this we use the result from Lemma 1 and the relationship between the uniform empirical and the uniform quantile process. Let $\alpha(n) = \sqrt{n}(\hat{E}_n(u) - u)$ be the uniform empirical process. By the change of variable argument we have $\sup_{0<y<1} |u_n(y)| = \sup_{0<y<1} |\alpha_n(y)|$ (see Remark 1.4.1 in Csörgő (1983)). The result now follows from the Dvoretzky-Kiefer-Wolfowitz inequality, using the tight bound

from Massart (1990):

$$\begin{aligned} P\left(\sup_{0<y<1}|q_n(y)| \geq x\right) &\leq P\left(\sup_{0<y<1}|u_n(y)|C \geq x\right) \\ &= P\left(\sup_{0<y<1}|\alpha_n(y)| \geq x/C\right) \\ &\leq 2e^{-2x^2/C^2}. \end{aligned}$$

□

3.3.2 Consistency results

In this section, we consider the asymptotic behaviour of the statistics proposed in (3.3) and (3.7), with the proof of Theorem 1 provided below. Aside from the concavity assumption on the false null distribution F_1, we also need the following assumption regarding the function $h_F(t)$ defined in (3.9):

(A1) $h_F(t)$, for $t \in (0,1)$, has a unique point of local maximum, which we denote as $\tilde{\tilde{t}}$, where $\tilde{\tilde{t}} \leq 1/2$, signifying that $\tilde{\tilde{t}} = \text{argmax}_{t\in(0,1)}\, h_F(t)$.

With the assumption above we exclude the situations when $h_F(t)$ is flat around $\tilde{\tilde{t}}$ which is a very special case, that cannot be characterised easily in terms of conditions on F. For example, this assumption does not hold in some cases where F_1 is a mixture of uniform distributions with multiple components – that is, when the quantile function is piecewise linear. For instance, for a uniform mixture distribution whose quantile function is piecewise linear with change-points in slope at $0.1, 0.2, 0.3, 0.4$ and with increasing slopes on the first four segments equal to $0.1, 0.2, 0.4, 0.9$, the corresponding function $h_F(t)$ is constant on the interval $[0.3, 0.4]$ where its value is maximal.

The convergence rates of the statistics in Theorem 1 are also considered, and they depend on the differentiability of h_F at $\tilde{\tilde{t}}$, and also on how flat h_F is at $\tilde{\tilde{t}}$, which is

measured using higher order derivatives. This is discussed in the proof of the theorem and also in Remark 1.

Theorem 1. *Let*

$$F(x) = \pi_1 F_1(x) + \pi_0 x, \quad x \in [0,1], \tag{3.11}$$

be the p-value distribution, where F_1 is a (weakly) concave function. Let $h_F(t)$ be defined as in (3.9), and assume that condition (A1) holds. Let $p_{(1)}, \ldots, p_{(n)}$ be the order statistics of the p-values sequence. Let $\hat{k}_{DOS}$ and $\hat{\pi}_1^{DOS}$ be as defined in (3.3) and (3.7), respectively, with $c_n = n^{-\theta}$, for some $\theta \in (0,1)$. It holds that

$$\hat{k}_{DOS}/n \xrightarrow{a.s.} \tilde{\tilde{t}} := \operatorname*{argmax}_{0 \le t \le 1/2} \frac{F^{-1}(2t) - 2F^{-1}(t)}{t}, \tag{3.12}$$

$$p_{(\hat{k}_{DOS})} \xrightarrow{a.s.} F^{-1}(\tilde{\tilde{t}}), \tag{3.13}$$

$$\hat{\pi}_1^{DOS} \xrightarrow{a.s.} \frac{\tilde{\tilde{t}} - F^{-1}(\tilde{\tilde{t}})}{1 - F^{-1}(\tilde{\tilde{t}})} \le \pi_1. \tag{3.14}$$

Proof. Let

$$h_n(t) := \frac{\hat{F}_n^{-1}(2t) - 2\hat{F}_n^{-1}(t)}{t}.$$

The empirical function $h_n(t)$ is approximating the ideal function

$$h_F(t) := \frac{F^{-1}(2t) - 2F^{-1}(t)}{t}.$$

Function h_F is positive on $(0,1)$. This holds as F^{-1} is a convex function, which follows from the concavity assumption on F_1. The convexity of F^{-1} is demonstrated in the proof of Lemma 1. The idea of the proof is to show that the two functions are uniformly close

$$h_n(t) \approx h_F(t), \quad \forall t \in (0,1). \tag{3.21}$$

From this, the consistency of the change-point estimator follows:

$$\hat{k}_{DOS}/n = \arg\max_t h_n(t) \approx \arg\max_t h_F(t), \tag{3.22}$$

which also implies the consistency of the estimators (3.3) and (3.7). We start with the following sequence of inequalities, aiming to upper bound the rate of difference $|h_n(t) - h_F(t)|$, uniformly for $t \in (c_n, 1)$, using strong limit theorems for weighted uniform quantile processes.

$$\begin{aligned}
|h_n(t) - h_F(t)| &= \left| \frac{\hat{F}_n^{-1}(2t) - 2\hat{F}_n^{-1}(t)}{t} - \frac{F^{-1}(2t) - 2F^{-1}(t)}{t} \right| \\
&\leq 2 \left| \frac{\hat{F}_n^{-1}(2t) - F^{-1}(2t)}{2t} \right| + 2 \left| \frac{\hat{F}_n^{-1}(t) - F^{-1}(t)}{t} \right| \\
&\leq \frac{2}{\sqrt{n}} \frac{|q_n(2t)|}{2t} + \frac{2}{\sqrt{n}} \frac{|q_n(t)|}{t} \\
&\leq \frac{4}{\sqrt{n}} \sup_{t \in (c_n, 1)} \frac{|q_n(t)|}{t} \\
&\leq \frac{C}{\sqrt{n}} \sup_{t \in (c_n, 1)} \frac{|u_n(t)|}{t}.
\end{aligned} \tag{3.23}$$

In the last inequality we used the result from Lemma 1. To bound the weighted uniform quantile process $u_n(t)/t$, we use Theorem 2 case (III) from Einmahl and Mason (1988), setting $\nu = 0, a_n = \log(n)/n$ using the notation therein, to get

$$\limsup_{n \to \infty} \sup_{c_n \leq t \leq 1/2} \frac{c_n^{1/2} |u_n(t)|}{t \sqrt{\log \log n}} \overset{a.s.}{=} 2. \tag{3.24}$$

Note that this result only considers $t \leq 1/2$, but since for $t > 1/2$, the weight function $1/t$ is bounded we have

$$\sup_{1/2 \leq t \leq 1} \frac{|u_n(t)|}{t} \leq \sup_{1/2 \leq t \leq 1} 2|u_n(t)|.$$

Now we apply the Chung-Smirnov law of iterated logarithm for the uniform empirical process (Chung (1949)) in (3.25) to get

$$\begin{aligned}\limsup_{n\to\infty}\sup_{1/2\le t\le 1}\frac{|u_n(t)|}{t\sqrt{\log\log n}} &\le 2\limsup_{n\to\infty}\sup_{0\le t\le 1}\frac{|u_n(t)|}{\sqrt{\log\log n}},\\ &\overset{a.s.}{=}\sqrt{2},\end{aligned} \tag{3.25}$$

which implies

$$\limsup_{n\to\infty}\sup_{1/2\le t\le 1}\frac{c_n^{1/2}|u_n(t)|}{t\sqrt{\log\log n}}\overset{a.s.}{=}0.$$

Thus, we have

$$\limsup_{n\to\infty}\sup_{c_n\le t\le 1}\frac{c_n^{1/2}|u_n(t)|}{t\sqrt{\log\log n}}\overset{a.s.}{=}2.$$

It means that, for any $\varepsilon>0$ and large enough n, on a set of probability 1, it holds that

$$\frac{1}{\sqrt{n}}\sup_{c_n\le t\le 1}\frac{|u_n(t)|}{t}\le\frac{\sqrt{\log\log n}}{n^{\frac{1-\theta}{2}}}(2+\varepsilon). \tag{3.26}$$

Finally, (3.26) and (3.23) give a uniform upper bound for $|h_n(t)-h_F(t)|$ on $t\in(c_n,1]$

$$\sup_{t\in(c_n,1]}|h_n(t)-h_F(t)|\le C\frac{\sqrt{\log\log n}}{n^{\frac{1-\theta}{2}}}, \tag{3.27}$$

where C is a constant that for large n approaches 2. Denote

$$\hat{t}_n:=\arg\max h_n(t).$$

From (3.27) and the reverse triangle inequality, it follows that for n large enough

$$||h_n(t)|-|h_F(t)||<C\frac{\sqrt{\log\log n}}{n^{\frac{1-\theta}{2}}},$$

uniformly, for all $t \in (c_n, 1)$ with probability 1. Since for n large enough, $\tilde{\tilde{t}} > c_n$, the following sequence of inequalities holds with probability 1:

$$h_F(\tilde{\tilde{t}}) \leq \left|h_n(\tilde{\tilde{t}})\right| + C\frac{\sqrt{\log\log n}}{n^{\frac{1-\theta}{2}}} \leq \left|h_n(\hat{t}_n)\right| + C\frac{\sqrt{\log\log n}}{n^{\frac{1-\theta}{2}}} \leq h_F(\hat{t}_n) + 2C\frac{\sqrt{\log\log n}}{n^{\frac{1-\theta}{2}}}.$$

It implies

$$\left|h_F(\hat{t}_n) - h_F(\tilde{\tilde{t}})\right| \leq 2C\frac{\sqrt{\log\log n}}{n^{\frac{1-\theta}{2}}}. \tag{3.28}$$

We prove the consistency of $\hat{t}_n$ by contradiction. However, the rate of convergence depends on the differentiability of h, and we separate three different cases. For some additional discussion on this see Remark 1.

Case 1: h has a second derivative at $\tilde{\tilde{t}}$, and $h''(\tilde{\tilde{t}}) \neq 0$.

Case 2: h has a second derivative at $\tilde{\tilde{t}}$, and $h''(\tilde{\tilde{t}}) = 0$.

Case 3: h is not differentiable at $\tilde{\tilde{t}}$.

We start with Case 1, and note that a sufficient condition for h to be twice differentiable is that F is twice differentiable on $(0, 1)$. Let $|\hat{t}_n - \tilde{\tilde{t}}| > \sqrt{\frac{\log\log n}{n^{\frac{1-\theta}{2}}}}$. It holds that

$$\begin{aligned}\left|h_F(\tilde{\tilde{t}}) - h_F(\hat{t}_n)\right| &= (\hat{t}_n - \tilde{\tilde{t}})^2|h''(\tilde{\tilde{t}})| + o((\hat{t}_n - \tilde{\tilde{t}})^2) \\ &\geq C_1\frac{\log\log n}{n^{\frac{1-\theta}{2}}}.\end{aligned}$$

For large n, the last inequality is in contradiction with (3.28), so it must hold that

$$\left|\hat{t}_n - \tilde{\tilde{t}}\right| \leq C\sqrt{\frac{\log\log n}{n^{\frac{1-\theta}{2}}}}, \tag{3.29}$$

which proves the consistency in (3.12). For Case 2, if $h''(\tilde{\tilde{t}}) = 0$, the consistency still holds, since not all derivatives can be zero, but the rate of convergence is slower

accordingly. In Case 3, when h is not differentiable at $\tilde{\tilde{t}}$, such that left and right derivatives at $\tilde{\tilde{t}}$ are not equal, but lower bounded by a constant larger than zero in an interval around $\tilde{\tilde{t}}$, we can get a better convergence rate:

$$\left|\hat{t}_n - \tilde{t}\right| \leq C \frac{\log\log n}{n^{\frac{1-\theta}{2}}}.$$

This is the case for example when F is a mixture of uniform distributions (see Corollary 1). If a one-sided derivative approaches zero at $\tilde{\tilde{t}}$, then we similarly have (3.29) to hold. We proceed under Case 1, assuming that $h''(\tilde{\tilde{t}}) \neq 0$ holds, while the results for other cases can be obtained similarly. The following sequence of inequalities holds almost surely and proves the consistency in (3.13):

$$\begin{aligned}
\left|p_{(\hat{k}_{DOS})} - F^{-1}(\tilde{\tilde{t}})\right| &= \left|\hat{F}_n^{-1}(\hat{t}_n) - F^{-1}(\tilde{\tilde{t}})\right| \\
&\leq \left|\hat{F}_n^{-1}(\hat{t}_n) - F^{-1}(\hat{t}_n)\right| + \left|F^{-1}(\hat{t}_n) - F^{-1}(\tilde{\tilde{t}})\right| \qquad (3.30) \\
&\leq C_1 \sqrt{\frac{\log\log n}{n}} + C_2 \sqrt{\frac{\log\log n}{n^{\frac{1-\theta}{2}}}} \\
&\leq C_3 \sqrt{\frac{\log\log n}{n^{\frac{1-\theta}{2}}}}.
\end{aligned}$$

For the first term in (3.30) we use Lemma 1 and then the Chung-Smirnov law of iterated logarithm, to get the inequality which holds almost surely. For the second term we use the fact that F^{-1} is Lipschitz continuous, and the obtained rate of convergence in (3.29). The consistency of (3.14) follows similarly:

$$\begin{aligned}
\left|\hat{\pi}_1^{DOS} - \frac{\tilde{\tilde{t}} - F^{-1}(\tilde{\tilde{t}})}{1 - F^{-1}(\tilde{\tilde{t}})}\right| &= \left|\frac{\hat{t}_n - F^{-1}(\hat{t}_n)}{1 - F^{-1}(\hat{t}_n)} - \frac{\tilde{\tilde{t}} - F^{-1}(\tilde{\tilde{t}})}{1 - F^{-1}(\tilde{\tilde{t}})}\right| \\
&\leq \left|\frac{(1 - F^{-1}(\hat{t}_n))(\hat{t}_n - \tilde{\tilde{t}}) + \hat{t}_n(F^{-1}(\hat{t}_n) - F^{-1}(\tilde{\tilde{t}}))}{(1 - F^{-1}(\tilde{\tilde{t}}))(1 - F^{-1}(\hat{t}_n))}\right| \\
&\leq C \sqrt{\frac{\log\log n}{n^{\frac{1-\theta}{2}}}}.
\end{aligned}$$

Furthermore, using the inequality $F(t) \leq \pi_1 + (1 - \pi_1)t$, that holds for any t, we get that

$$\frac{\tilde{\tilde{t}} - F^{-1}(\tilde{\tilde{t}})}{1 - F^{-1}(\tilde{\tilde{t}})} \leq \pi_1.$$

□

Remark 1. Condition $h''(\tilde{\tilde{t}}) \neq 0$, where $\tilde{\tilde{t}} = \operatorname{argmax} h_F(t)$, is implicitly a condition on the strength of the alternative, as it requires h not to be flat around its point of maximum. Precisely, $h''(\tilde{\tilde{t}}) = 0$ is equivalent to

$$(F^{-1})''(\tilde{\tilde{t}}) - \frac{1}{2}(F^{-1})''(2\tilde{\tilde{t}}) = 0,$$

and

$$(F^{-1})''(2\tilde{\tilde{t}}) - (F^{-1})''(\tilde{\tilde{t}}) = (F^{-1})''(\tilde{\tilde{t}}).$$

It is enough to have that $\operatorname{argmax}(F^{-1})''(t) < \tilde{\tilde{t}}$ for $h''(\tilde{\tilde{t}}) \neq 0$ to hold, but it is not possible to further simplify this condition. In general, for $H(t, a)$ function, the condition

$$\frac{\partial^2}{\partial t^2} H(\tilde{\tilde{t}}, a) = 0,$$

where $\frac{\partial}{\partial t} H(\tilde{\tilde{t}}, a) = 0$ is equivalent to

$$(F^{-1})''(\tilde{\tilde{t}} + a) - (F^{-1})''(\tilde{\tilde{t}}) = (F^{-1})''(\tilde{\tilde{t}}) - (F^{-1})''(\tilde{\tilde{t}} - a).$$

The following Corollary 1 considers a special case when the p-values come from a mixture of two uniform distributions. In that case, F and F^{-1} are piecewise linear functions with one change-point where the slope changes. From Theorem 1 it follows that $\hat{k}_{DOS}/n$ consistently estimates this change-point and that $\hat{\pi}_1^{DOS}$ is an unbiased estimator of π_1.

Corollary 1. *Let $p_{(1)}, \ldots, p_{(n)}$ be the order statistics of the sequence of p-values coming from a distribution with CDF G, where G is a mixture of two uniform distributions*

$$\pi_1 U[0, b] + \pi_0 U[0, 1]. \tag{3.31}$$

Let $\hat{k}_{DOS}$ and $\hat{\pi}_1^{DOS}$, be the corresponding statistics proposed in (3.3) and (3.7), computed using the sequence of p-values with distribution (3.31) and with $c_n = n^{-\theta}$. It holds that

$$\begin{aligned} \hat{k}_{DOS}/n &\xrightarrow{a.s.} \pi_1 + b\pi_0, \\ p_{(\hat{k}_{DOS})} &\xrightarrow{a.s.} b, \\ \hat{\pi}_1^{DOS} &\xrightarrow{a.s.} \pi_1. \end{aligned}$$

Thus, $p_{(\hat{k}_{DOS})}$ and $\hat{\pi}_1^{DOS}$ are strongly consistent estimators of the uniform mixture parameters b and π_1, respectively.

Proof. In the case of uniform mixture, $h_G(t) = (G^{-1}(2t) - 2G^{-1}(t))/t$ is easy to calculate.

$$h_G(t) = \begin{cases} 0, & t \leq \frac{\pi_1 + b\pi_0}{2} \\ \frac{1}{t}\left(\frac{2t-\pi_1}{\pi_0} - \frac{2tb}{\pi_1 + b\pi_0}\right), & \frac{\pi_1 + b\pi_0}{2} < t \leq \pi_1 + b\pi_0 \\ \frac{\pi_1}{\pi_0 t}, & \pi_1 + b\pi_0 < t \leq 1/2 \\ \frac{1}{t}\left(1 - 2\frac{t-\pi_1}{\pi_0}\right), & 1/2 < t \leq 1 \\ 0, & t > 1 \end{cases}$$

Let $\tilde{t} = \operatorname{argmax} h_G(t) = \pi_1 + b\pi_0$ and $\hat{t}_n = \operatorname{argmax}_t h_n(t)$. The proof follows the same steps as the proof of Theorem 1. We note that since $h'_G(\tilde{t})$ does not exist, we can get a better rate of convergence for the statistics $\hat{k}_{DOS}/n, p_{(\hat{k}_{DOS})}$ and $\hat{\pi}_1^{DOS}$. Assume that

$|\hat{t}_n - \tilde{t}| > \frac{\log\log n}{n^{\frac{1-\theta}{2}}}$. For $\hat{t}_n > \tilde{t}$ we have

$$\begin{aligned}\left|h_G(\tilde{t}) - h_G(\hat{t}_n)\right| &= \left|\frac{\pi_1}{\pi_0\tilde{t}} - \frac{\pi_1}{\pi_0\hat{t}_n}\right| \\ &= \frac{\pi_1}{\pi_0}\left|\frac{1}{\tilde{t}} - \frac{1}{\hat{t}_n}\right| \\ &= \frac{\pi_1}{\pi_0}\frac{|\tilde{t} - \hat{t}_n|}{\hat{t}_n\tilde{t}} \\ &\geq C'|\tilde{t} - \hat{t}_n| \\ &\geq C''\frac{\log\log n}{n^{\frac{1-\theta}{2}}}.\end{aligned}$$

For $\hat{t}_n < \tilde{t}$ we get the same lower bound in a similar way. The last inequality is, for large n, in contradiction with the uniform bound in (3.28), so it must hold that

$$\left|\hat{t}_n - \tilde{t}\right| \leq C\frac{\log\log n}{n^{\frac{1-\theta}{2}}}.$$

Thus $\hat{t}_n$ is an a.s. consistent estimator for the change-point in the quantile function G^{-1}. The consistency of the derived estimators for b and π_1 follows as in the proof of Theorem 1, and it holds that

$$\begin{aligned}\left|p_{(\hat{k}_{DOS})} - b\right| &= \left|\hat{F}_n^{-1}(\hat{t}_n) - F^{-1}(\tilde{t})\right| \\ &\leq \left|\hat{F}_n^{-1}(\hat{t}_n) - F^{-1}(\hat{t}_n)\right| + \left|F^{-1}(\hat{t}_n) - F^{-1}(\tilde{t})\right| \qquad (3.32) \\ &\leq C_1\sqrt{\frac{\log n}{n}} + C_2\frac{\log\log n}{n^{\frac{1-\theta}{2}}} \\ &\leq C_3\frac{\log\log n}{n^{\frac{1-\theta}{2}}}.\end{aligned}$$

$$\begin{aligned}\left|\hat{\pi}_1^{DOS} - \pi_1\right| &= \left|\frac{\hat{t}_n - p_{(\hat{k}_{DOS})}}{1 - p_{(\hat{k}_{DOS})}} - \pi_1\right| \\ &= \left|\frac{\hat{t}_n - \tilde{t} + \pi_0(b - p_{(\hat{k}_{DOS})})}{1 - p_{(\hat{k}_{DOS})}}\right| \\ &\leq C_1\left|\hat{t}_n - \tilde{t}\right| + C_2|p_{(\hat{i})} - b| \\ &\leq C_3\frac{\log\log n}{n^{\frac{1-\theta}{2}}}.\end{aligned}$$

□

Corollary 2 shows that in the case when the alternative is a non-uniform distribution with support $[0, b]$, statistics $p_{(\hat{k}_{DOS})}$ and $\hat{\pi}_1^{DOS}$ will a.s. not overestimate both the support of the distribution and the alternative proportion.

Corollary 2. *Let $[0, b]$, $b \leq 1$ be the support of the alternative distribution F_1, where F_1 is stochastically smaller than $U[0, b]$ distribution, in the sense that $F_1(t) \geq t/b$ for all $0 \leq t \leq b$. Statistics $p_{(\hat{k}_{DOS})}$ and $\hat{\pi}_1^{DOS}$ almost surely underestimate the parameters b and π_1 respectively.*

Proof. Let G be the CDF of a uniform mixture $\pi_1 U[0, b] + \pi_0 U[0, 1]$. Since F_1 is stochastically smaller than $U[0, b]$ it holds that

$$F_1(x) \geq \frac{x}{b}, \quad 0 \leq x \leq b.$$

This implies

$$\begin{aligned}F(t) &\geq G(t), \quad t \in [0, b], \\ F(t) &= G(t), \quad t \in [b, 1],\end{aligned}$$

and also

$$F^{-1}(t) \leq G^{-1}(t), \quad t \in [0, \pi_1 + b\pi_0],$$
$$F^{-1}(t) = G^{-1}(t), \quad t \in [\pi_1 + b\pi_0, 1].$$

Let $\tilde{t} = \pi_1 + b\pi_0$. The following sequence of inequalities shows that the maximum of the ideal sequence will not be achieved after $\tilde{t}$:

$$\begin{aligned}
\sup_{t\in(0,\tilde{t}]} \frac{F^{-1}(2t) - 2F^{-1}(t)}{t} &\geq \frac{F^{-1}(2\tilde{t}) - 2F^{-1}(\tilde{t})}{\tilde{t}} \\
&= \frac{G^{-1}(2\tilde{t}) - 2G^{-1}(\tilde{t})}{\tilde{t}} \\
&= \sup_{t\in(0,\tilde{t}]} \frac{G^{-1}(2t) - 2G^{-1}(t)}{t} \\
&\geq \sup_{t\in[\tilde{t},1/2]} \frac{G^{-1}(2t) - 2G^{-1}(t)}{t} \\
&= \sup_{t\in(\tilde{t},1/2]} \frac{F^{-1}(2t) - 2F^{-1}(t)}{t}.
\end{aligned}$$

We got that

$$\tilde{\tilde{t}} := \sup_{t\in(0,1/2]} \frac{F^{-1}(2t) - 2F^{-1}(t)}{t} \leq \tilde{t}.$$

From the consistency in Theorem 1 it follows that

$$\frac{\hat{k}_{DOS}}{n} \overset{a.s.}{\to} \tilde{\tilde{t}} \leq \tilde{t},$$

which implies

$$\begin{aligned}
p_{(\hat{k}_{DOS})} &\overset{a.s.}{\to} F^{-1}(\tilde{\tilde{t}}) \leq F^{-1}(\tilde{t}) = b, \\
\hat{\pi}_1 &\overset{a.s.}{\to} \frac{\tilde{\tilde{t}} - F^{-1}(\tilde{\tilde{t}})}{1 - F^{-1}(\tilde{\tilde{t}})} \leq \frac{\tilde{\tilde{t}} - F^{-1}(\tilde{\tilde{t}})}{F_1(F^{-1}(\tilde{\tilde{t}})) - F^{-1}(\tilde{\tilde{t}})} = \pi_1.
\end{aligned}$$

□

Remark 2. It is possible to use different rates for c_n in Theorem 1. c_n plays a role when applying the strong limit theorem for the weighted uniform quantile process $|u_n(t)|/t$ from Einmahl and Mason (1988). This is needed for bounding $|h_n(t) - h_F(t)|$ in (3.23). The theorem from Einmahl and Mason (1988) states that the sufficient conditions on c_n, where $c_n \to 0$, in order for (3.24) to hold are:

$$\frac{nc_n}{\log\log n} \to \infty,$$
$$\frac{\log\log(1/c_n)}{\log\log n} \to C < \infty.$$

We note that this also holds for example for $c_n = \log n/n$, or $(\log\log n)^2/n$. Additionally, we can trivially take $c_n = \varepsilon \in (0, 1)$, in which case the denominator in (3.23) is not a problem as the interval for the supremum is bounded away from zero, and we can just use the Chung-Smirnov law, stated in the proof of Theorem 1. The choice of c_n affects the rate of convergence for the statistics from the statements of Theorem 1. For c_n in general it holds that

$$\sup_{t\in(c_n,1]} |h_n(t) - h_F(t)| \leq \frac{C}{\sqrt{n}} \frac{\sqrt{\log\log n}}{\sqrt{c_n}},$$

and the term on the RHS reveals the rate of convergence of the considered statistics. Obviously, the slower c_n goes to zero, the better rate of convergence our estimators will have. From the work of Einmahl and Mason (1988) we see that the choice of c_n is very important in the theory of uniform quantile processes. As the false-null distribution F_1 is unknown, using the result from Lemma 2 we bound the quantile process of distribution F by a uniform quantile process. This approximation is convenient as most of the results in the theory of quantile processes are given only for the uniform quantile process. However, the behaviour of the weighted uniform quantile process

around 0 will be more variable than that of weighted quantile process of a distribution F. F is more concentrated around zero and the sample quantiles will be closer to the true quantiles than in the case of uniform distribution, reducing the boundary problem.

3.4 Simulations

We compare the performance of our method to different proportion estimators in the literature:

1. STS – Storey-Taylor-Siegmund by Storey et al. (2004), implemented in R package 'qvalue' by Storey et al. (2020)
2. MGF – Moment Generating Function method by Broberg (2005), implemented in R package 'SAGx' by Broberg (2020)
3. LLF – Langaas-Lindqvist-Ferkingstad by Langaas et al. (2005)
4. MR – Meinshausen-Rice by Meinshausen and Rice (2006)
5. JD – Jiang-Doerge by Jiang and Doerge (2008)
6. FIX – Storey's estimator (2.12) with $\lambda = p_{(n/2)}$

The STS is a bootstrap method by Storey et al. (2004) that finds the optimal λ to use in (2.12) by minimising the resulting estimator's MSE. The LLF estimator by Langaas et al. (2005) estimates the true null proportion as the longest constant interval of the density estimator by Grenander (1956). The MR estimator by Meinshausen and Rice (2006) gives a lower bound for the confidence interval for π_1, of form $(\hat{\pi}_1^{MR}, 1]$. This lower bound $\hat{\pi}_1^{MR}$ is guaranteed to consistently estimate the false null proportion, given that the signal is not too weak. The JD estimator by Jiang and Doerge (2008) is a bootstrap Storey-based estimator using the average estimator approach. It estimates

the false null proportion as the average of values in (2.12), for different λ's, where the number of summands and the corresponding λ values are calculated using bootstrap.

All of the methods above are introduced in detail in Section 2.2. We now introduce the moment generating function-based method by Broberg (2005) (MGF). We found that this estimator performs well, although it is rarely included in the simulation studies in the literature. The MGF estimates the proportion by estimating the moment generating function (mgf) of the mixture distribution of p-values, which is a weighted sum of the $U[0,1]$ and the false-null distribution mgf:

$$M_F(t) = \pi_1 M_{F_1}(t) + \pi_0 M_{U[0,1]}(t).$$

The proportion can be written as a ratio

$$\pi_1 = \frac{M_F(t) - M_{U[0,1]}(t)}{M_{F_1}(t) - M_{U[0,1]}(t)}. \tag{3.33}$$

In contrast to the other methods from the introduction that only exploit the behaviour of the p-values under the null, the MGF also estimates the behaviour under the alternative, by estimating $M_{F_1}(t)$. Additionally, they use the average of the essentially constant ratios involving the estimated mgf plugged in the equation (3.33) above. This makes the MGF estimator very precise in some cases, which is reflected in the simulation results.

Additionally, for control, we include FIX estimator, a simple Storey's estimator with $\lambda = p_{(n/2)}$. We choose $\lambda = p_{(n/2)}$ as from the definition of the DOS statistic, $1/2$ is the largest possible location for the change-point, which leads to using $\lambda = p_{(n/2)}$ in Storey's estimator. In this way we estimate the proportion using the larger half of the p-values.

Simulation results for the Gaussian mean testing model for sample size $n = 1000$ are given in Table 3.1, and for small sample sizes $n = 50$ and $n = 100$ in Table 3.2.

Under the true null hypothesis, the test statistics have $N(0, 1)$ distribution and under the alternative $N(\mu, 1)$, $\mu > 0$. In Table 3.1, for the DOS method, the number of excluded values from the beginning of the sequence $d(i)$ is set to $nc_n = 5$, and the estimates are not sensitive to this value. The (lack of) effect of different values of nc_n on the resulting estimates is investigated later in this section. In fact, the problem only appears when nc_n is close to the true number of false null p-values.

As guaranteed by Theorem 1, the DOS method underestimates the false null proportion. However, for moderate sample sizes it underestimates less so than the MR method. The DOS works particularly well in sparse cases, when its MSE is among the lowest ones as shown in Table 3.1. In the denser cases, when π_1 is larger, and the signal strength is weaker, the DOS method exhibits larger variance, which can be seen in the bottom two settings in 3.1, when $\mu = 2$ and $\pi_1 = 0.2$ or $\pi_1 = 0.3$. This variability comes from the smoother shape of the quantile function. When the signal is weak, the corresponding ideal function h_F is flatter around its maximum, which causes greater variability of the maximum point of the sequence $d(i)$.

Simulations show that for larger sample sizes $n > 5000$ the consistency for the MR method sets in, and the MR yields less biased estimates than the DOS.

	STS	MGF	LLF	MR	JD	DOS	FIX
$n_1 = 10$, $\mu = 3.5$							
BIAS	27.1	4.5	34.0	-4.2	8.6	**0.4**	8.5
SD	37.6	14.7	52.9	7.1	21.6	**4.7**	21.6
MSE	2148	236	3954	68	540	**22**	538
$n_1 = 30$, $\mu = 3.5$							
BIAS	24.6	**0.9**	13.7	-9.1	2.6	-2.8	3.1
SD	33.9	17.8	33.2	**4.0**	27.2	5.4	25.6
MSE	1754	317	1289	98	746	**37**	665
$n_1 = 50$, $\mu = 3$							
BIAS	22.6	-3.1	1.4	-17.3	-3.3	-8.5	**-0.8**
SD	35.3	18.8	28.3	**7.5**	30.4	8.8	28.2
MSE	1757	363	803	355	935	**150**	796
$n_1 = 100$, $\mu = 2$							
BIAS	19.1	-13.8	-9.8	-44.8	-7.2	-38.3	**-4.7**
SD	39.5	17.6	37.6	**13.8**	30.2	19.9	26.0
MSE	1925	**500**	1510	2197	964	1863	698
$n_1 = 100$, $\mu = 3$							
BIAS	24.5	-3.3	-4.5	-24.3	-4.0	-14.2	**-0.9**
SD	41.2	17.8	29.4	**8.4**	31.4	10.9	27.6
MSE	2298	328	885	661	1002	**320**	763
$n_1 = 200$, $\mu = 2$							
BIAS	16.4	-27.0	-19.3	-60.6	**-9.3**	-45.5	-13.6
SD	41.5	**15.9**	45.2	16.5	29.8	25.3	22.0
MSE	1991	982	2416	3945	975	2710	**669**
$n_1 = 300$, $\mu = 2$							
BIAS	11.1	-43.1	-23.8	-74.3	**-14.5**	-50.9	-27.2
SD	41.7	**16.6**	50	19.3	33.2	25.1	17.2
MSE	1862	2133	3066	5893	1312	3221	**1036**

Table 3.1 Bias, standard deviation and the MSE of the estimators, given the number of false null hypotheses n_1 and the non-zero mean μ, for a sample of size $n = 1000$, based on 1000 repetitions. Bold and underlined values correspond to the smallest values in each row.

We also consider the performance of the DOS procedure for small sample sizes. In this setting, we were unable to use the estimators STS and MGF as they require larger number of p-values around 1 in order to compute the estimates. For small sample sizes this is often not satisfied and the implemented functions return errors. In Table 3.2 we give simulation results for $n = 50$ and $n = 100$, for different numbers of false null hypotheses n_1 and nonzero mean parameter μ. The results show that our estimator has the smallest mean squared error among the considered estimators.

Finally, we consider the impact of c_n on the DOS estimates. Again, we consider the Gaussian setting, for different values of sample size $n \in \{10000, 1000, 100\}$, the nonzero mean μ and the false null proportion π_1. The results are shown in Figures 3.8, 3.9 and 3.10. We notice that the estimates are stable in many of the considered settings. The estimates become sensitive in two cases: 1) when we exclude "too many" values which in general happens as c_n approaches π_1, and 2) when the signal is very weak, in which case we also cannot count on the DOS estimator to perform well, see the lower right plots (d) in Figures 3.8, 3.9 and 3.10. The problem described in the first case can be seen in all three figures. The problem of the second case also shows the weakness of our estimator which is that even if the sample is very large such as $n = 10000$ in Figure 3.8, the estimator will be variable and dependent on c_n if the signal is not strong enough. However, for moderately large values of signal, we see that in practice it is not necessary to truncate the sequence $d(i)$.

3.5 Extensions

This section discusses some possible extensions of the DOS method, and ideas for correcting some of the weaknesses of our method.

3.5.1 A family of estimators

A possible generalisation of our method comes from introducing a parameter in the DOS sequence. For $\alpha \in [1/2, 1]$ we consider a family of statistics

$$\hat{k}_\alpha^{DOS} = \underset{nc_n \le i \le n/2}{\operatorname{argmax}}\, , d_\alpha(i),$$

where

$$\begin{aligned} d_\alpha(i) &= \frac{p_{(2i)} - p_{(i)}}{(i/n)^\alpha} - \frac{p_{(i)}}{(i/n)^\alpha} \\ &= \frac{p_{(2i)} - 2p_{(i)}}{(i/n)^\alpha}. \end{aligned}$$

For $\hat{k}_\alpha^{DOS}$ and the induced proportion estimator $\pi_{1,\alpha}^{DOS}$, statements analogous to those in Theorem 1 hold.

Corollary 3. *Under the same conditions as in Theorem 1 it holds that*

$$\begin{aligned} \frac{\hat{k}_\alpha^{DOS}}{n} &\xrightarrow{a.s.} \tilde{\tilde{t}}_\alpha := \underset{0 \le t \le 1/2}{\operatorname{argmax}} \frac{F^{-1}(2t) - 2F^{-1}(t)}{t^\alpha}, \\ p_{(\hat{k}_\alpha^{DOS})} &\xrightarrow{a.s.} F^{-1}(\tilde{\tilde{t}}_\alpha), \\ \pi_{1,\alpha}^{DOS} := \frac{\hat{k}_{DOS}/n - p_{(\hat{k}_\alpha^{DOS})}}{1 - p_{(\hat{k}_\alpha^{DOS})}} &\xrightarrow{a.s.} \frac{\tilde{\tilde{t}}_\alpha - F^{-1}(\tilde{\tilde{t}}_\alpha)}{1 - F^{-1}(\tilde{\tilde{t}}_\alpha)} \le \pi_1. \end{aligned}$$

Denote

$$h_F^\alpha(t) := \frac{F^{-1}(2t) - 2F^{-1}(t)}{t^\alpha}.$$

Note that for $\alpha_1 < \alpha_2$, since $1/t^{\alpha_2 - \alpha_1}$ is a decreasing function, it holds that

$$\operatorname*{argmax}_t h_F^{\alpha_1}(t) > \operatorname*{argmax}_t h_F^{\alpha_2}(t).$$

This implies that for larger α's the "change-point" happens later. However, in the case of bounded support of the alternative $[0, b]$ the ideal change-point location will never be after $\pi_1 + b\pi_0$ for any α, that is

$$\tilde{\tilde{t}}_\alpha \leq \pi_1 + b\pi_0.$$

Additionally, for the family of ideal functions $h_F^\alpha(t)$ we can draw some similar conclusions as for $h_F(t) = h_F^1(t)$ in Section 3.2.1, assuming that $(F^{-1})'(t)$ exists for $t \in (0, 1)$. Let

$$H^\alpha(x, a) := \frac{F^{-1}(x + a) - F^{-1}(x)}{a^\alpha} - \frac{F^{-1}(x) - F^{-1}(x - a)}{a^\alpha} \tag{3.34}$$

for $a \in (0, x]$ and $x \in [0, 1]$. For $a = x$, $H^\alpha(x, x) = h_F^\alpha(x)$. Since $(F^{-1})'$ is an increasing function, by taking the partial derivative in a, it follows that $H^\alpha(x, a)$ is increasing in a for any fixed x. This implies

$$\underset{(x,a)}{\operatorname{argmax}}\, H^\alpha(x, a) = \underset{(x,x)}{\operatorname{argmax}}\, H^\alpha(x, x),$$

suggesting that the generalised α-DOS statistic also estimates the point of the largest scaled difference on symmetric intervals of arbitrary length in the quantile function.

Consider $\alpha = 1/2$ and the sequence $d_{1/2}(i)$, that approximates the function $h_F^{1/2}(t)$. For each i, $d_{1/2}(i)$ can be written as a scaled difference of means in the sequence of p-value spacings $p_{(i+1)} - p_{(i)}$

$$d_{1/2}(i) = \frac{1}{\sqrt{i/n}} \left[\frac{1}{2} \sum_{j=1}^{2i} (p_{(j)} - p_{(j-1)}) - \sum_{j=1}^{i} (p_{(j)} - p_{(j-1)}) \right]. \tag{3.35}$$

Let $s_i = p_{(i+1)} - p_{(i)}$. The sequence in (3.35) in particular is familiar from the change-point literature. At each i we are taking the scaled difference in means, the mean of the first i and the mean of the first $2i$ values in the sequence s_i. Arguing as in Section

3.2.1, it is enough to consider the expanding intervals with as function $H^{1/2}(t, a)$ is increasing in a. This means that we can view $\text{argmax}_i\, d_{1/2}(i)$ as a *scan statistic* with variable window size. The maximum in the sequence $d_{1/2}$ marks the location of the largest difference in means in the sequence of spacings s_i on symmetric intervals.

In Table 3.3, simulation results for different values of α show that as the signal decreases and the number of false null hypotheses increases, it is better to use smaller α's, to reduce the amount of underestimation. However, in dense cases, for example $n_1 = 200$, this generalisation does not help, as the signal is too weak for our method, and we cannot beat FIX estimator. A possible approach, requiring further analysis, would be to choose the optimal value of α before using the generalised DOS estimator. This could be done by using bootstrap for example, to find α such that the resulting estimator has the smallest MSE.

3.5.2 Reducing the underestimation

While our method performs the best in the sparse cases, the MGF method performs remarkably well in the dense cases. There are two competitive advantages of the MGF method over ours. The first one is that the final estimator is averaged across multiple values, increasing its precision, while DOS only uses a single value (estimated change-point location) to then calculate the proportion. The second one is that they estimate the behaviour under the alternative, while we just try to "skip" the alternative. Both of these differences result in the underestimation for our method. Consequently there are two approaches to reducing the underestimation - averaging or estimating the behaviour under the alternative. Here we consider the aggregated approach to amend the first problem. The idea of the aggregated DOS estimator is to use multiple values after the detected change-point to estimate the proportion. Similarly to the JD method, the estimator would be an average of Storey's estimators for different

λ's. Instead of only using the detected change-point $\lambda_{DOS} = \hat{k}^{DOS}/n$ in (2.12), we can use multiple values $\lambda_{DOS} \leq \lambda_1 < \cdots < \lambda_N < 1$ and Storey's estimator to get the aggregated estimator

$$\hat{\pi}_1^{aggr} = \frac{1}{N} \sum_{i=1}^{N} \hat{\pi}_1(\lambda_i).$$

We consider $[\lambda_1, \lambda_N] = [\lambda_{DOS}, 2\lambda_{DOS}]$ and $[\lambda_1, \lambda_N] = [\lambda_{DOS}, 1]$ with $N = 100$. Simulations showed that by considering $[\lambda_{DOS}, 1]$ interval, the resulting proportion estimator gives very variable estimates as the effect of the starting point gets lost in the larger variance brought by larger values in the interval. This is also the case for $[\lambda_{DOS}, 2\lambda_{DOS}]$ when the signal is dense, suggesting that possibly a smaller set of values around the change-point should be used. In sparse cases however, the averaging approach can introduce an additional variance compared to using just λ_{DOS}. Boxplots showing simulation results for the aggregated estimator on $[\lambda_{DOS}, 2\lambda_{DOS}]$ can be seen in Figures 3.11 and 3.12. The results show that, although there is some decrease in bias for the aggregated estimators, particularly for $\alpha = 1$, it is not too significant. This suggests that a more careful approach to choosing the interval for the aggregation might be needed.

3.6 Discussion

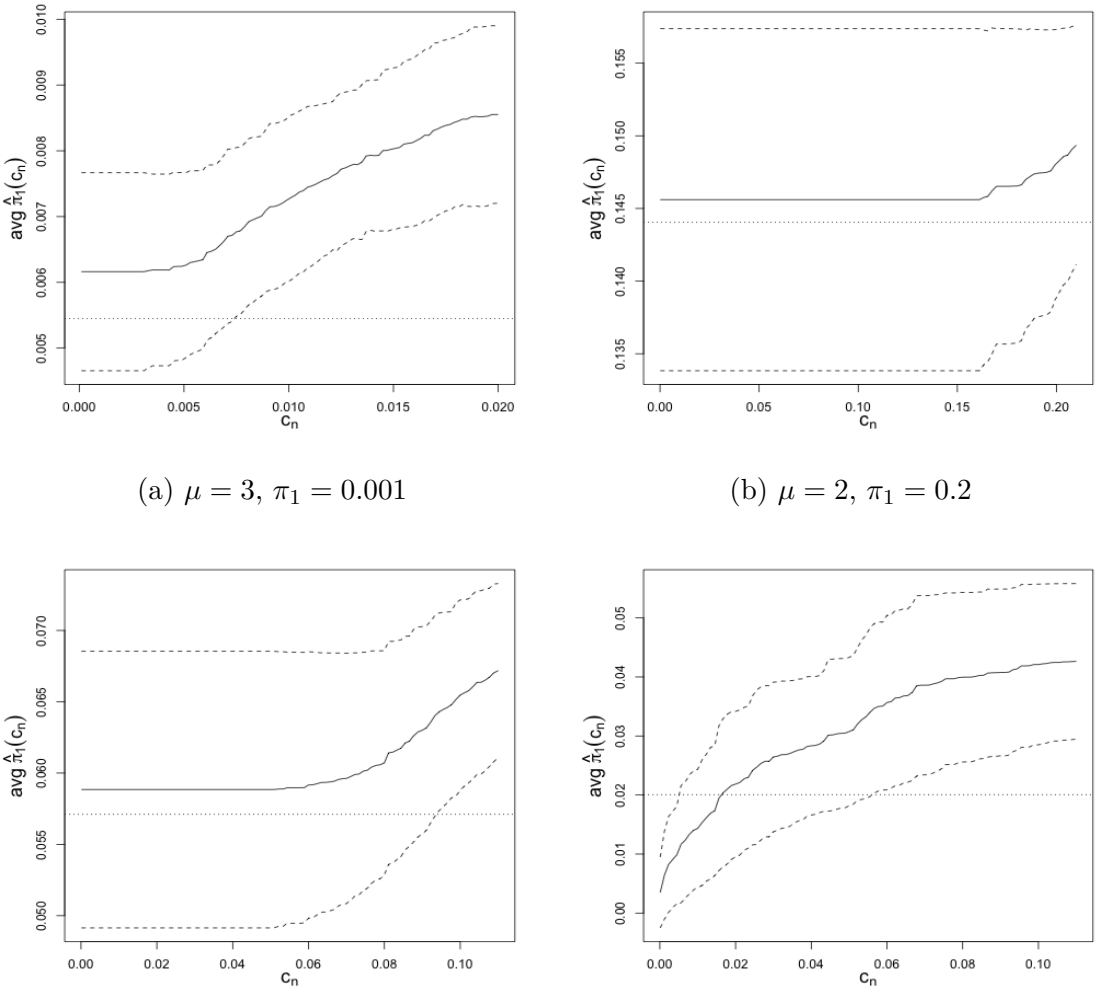
Another possible extension of the DOS method is to make it an iterative procedure. In this way we can estimate multiple "change-points" in the quantile function of the p-values which would give us a piecewise linear approximation to the quantile function. After estimating the first change-point in the quantile plot, we look for the next one in the group of p-values left of it, in order to filter out the remaining false-null p-values. This procedure would continue until the uniformity hypothesis of the remaining p-values can no longer be rejected. In general, there are no change-points in the usual sense,

and the estimated values would depend on the properties of the quantile function similarly as in Theorem 1. Estimating multiple change-points means approximating the distribution of the p-values with a mixture of multiple uniform distributions, and additionally, approximating the local FDR function, as defined in Efron (2007), with a piecewise constant function. This can be used for grouping p-values based on the estimated change-points in the following way: for the p-values between the two change-points we estimate a constant probability of them being from the true null hypothesis. This could be used in further analysis for deciding which hypotheses to reject.

We now comment on the possible use of the DOS proportion estimator in multiple testing procedures. The role of the proportion estimators in the multiple testing literature is mainly in making the Benjamini-Hochberg procedure "adapt to the unknown proportion", as a way to increase its power. In Blanchard and Roquain (2009), the performance of the adaptive Benjamini-Hochberg procedure using different proportion estimators is compared. The authors report that, based on the simulations, the best overall procedure for the FDR control at level q is the one that uses Storey's proportion estimator with $\lambda = q$. This is pointed out to be a "surprising result", as it is usually larger values λ that are used for estimating the proportion. Adaptive Benjamini-Hochberg with Storey's $\lambda = q$ estimator is also reported to be a "much more robust procedure in the case of dependent p-values, at the price of being slightly more conservative". Similarly to Storey's $\lambda = q$, our estimator also uses smaller values for λ, however the value is not constant but data-dependent, $\lambda = p_{(\hat{k}^{DOS})}$. The results of Blanchard and Roquain (2009) are encouraging the use of slightly conservative but low-variance proportion estimators such as DOS.

Many of the modern multiple testing procedures, that we introduced in Chapter 2, include prior knowledge on the p-values or additional assumptions on their structure. At the core of these procedures usually lies the Benjamini-Hochberg procedure, and

therefore there is a possibility in making these procedures more powerful by making them adaptive. Estimating the false null proportion in these modern settings has not been considered much in the literature, however it was suggested for example by Hu et al. (2010), Barber and Ramdas (2017), Basu et al. (2018) and Katsevich et al. (2021), as a way of increasing the power of the procedures proposed therein.



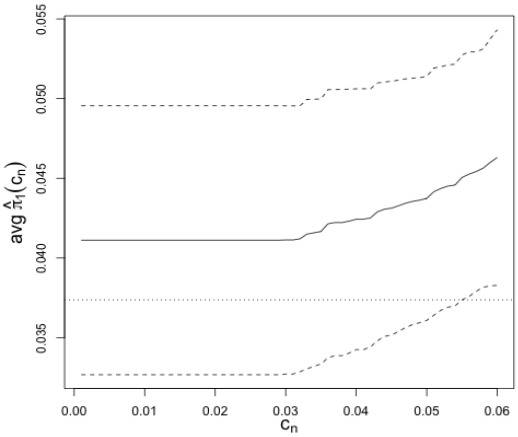
(a) $\mu = 3$, $\pi_1 = 0.001$

(b) $\mu = 2$, $\pi_1 = 0.2$

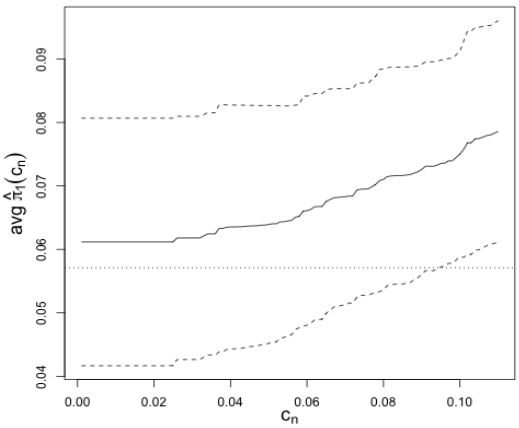
(c) $\mu = 2$, $\pi_1 = 0.1$

(d) $\mu = 1$, $\pi_1 = 0.1$

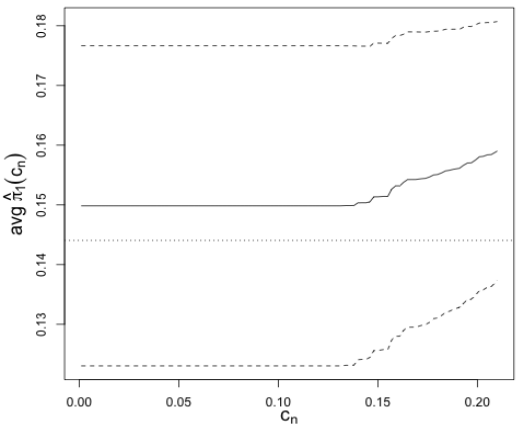
Fig. 3.8 The dependence of the proportion estimator on c_n shown for varying μ, π_1 and $n = 10000$. Solid line - Average estimated proportion over $N = 10000$ repetitions, shown as a function of c_n; Dashed lines - average proportion $\pm$ average standard deviation as a function of c_n; Dotted line - the limit of $\hat{\pi}_1^{DOS}$ from Theorem 1.



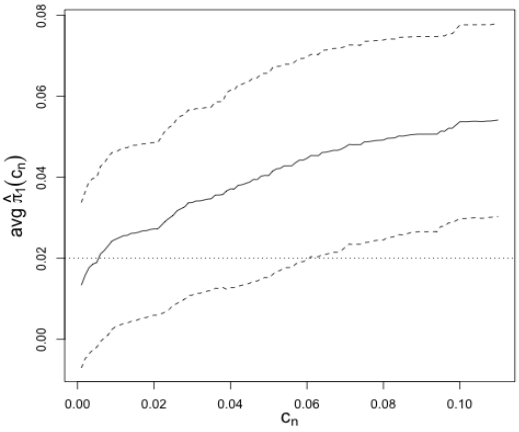
(a) $\mu = 3$, $\pi_1 = 0.05$



(b) $\mu = 2$, $\pi_1 = 0.2$

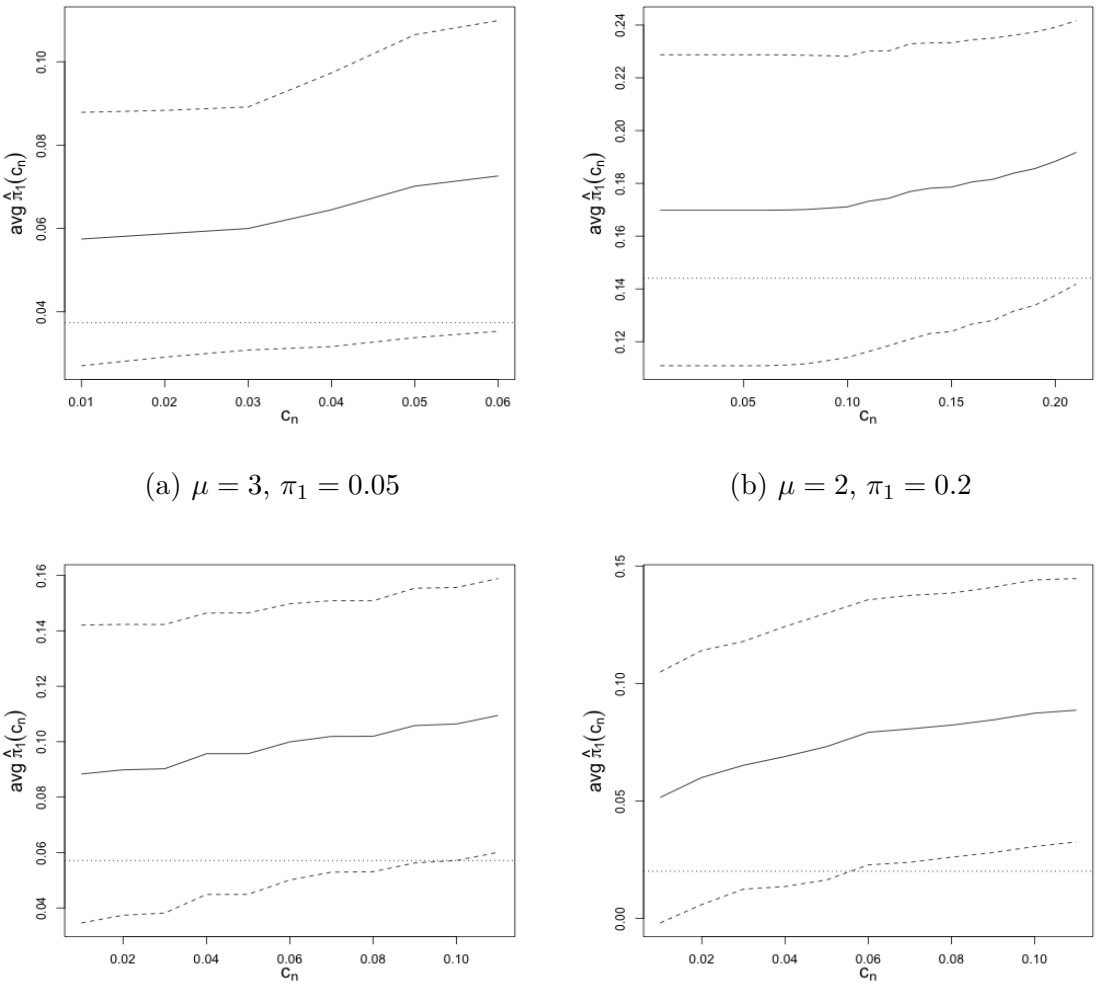


(c) $\mu = 2$, $\pi_1 = 0.1$



(d) $\mu = 1$, $\pi_1 = 0.1$

Fig. 3.9 The dependence of the proportion estimator on c_n shown for varying μ, π_1 and $n = 1000$. Solid line - Average estimated proportion over $N = 10000$ repetitions, shown as a function of c_n; Dashed lines - average proportion $\pm$ average standard deviation as a function of c_n; Dotted line - the limit of $\hat{\pi}_1^{DOS}$ from Theorem 1.



(a) $\mu = 3$, $\pi_1 = 0.05$

(b) $\mu = 2$, $\pi_1 = 0.2$

(c) $\mu = 2$, $\pi_1 = 0.1$

(d) $\mu = 1$, $\pi_1 = 0.1$

Fig. 3.10 The dependence of the proportion estimator on c_n shown for varying μ, π_1 and $n = 100$. Solid line - Average estimated proportion over $N = 10000$ repetitions, shown as a function of c_n; Dashed lines - average proportion $\pm$ average standard deviation as a function of c_n; Dotted line - the limit of $\hat{\pi}_1^{DOS}$ from Theorem 1.

	LLF	MR	JD	DOS
$n = 50, n_1 = 5, \mu = 3$				
BIAS	3.6	-3.3	0.7	**0.2**
SD	7	**1.4**	5	2.6
MSE	62	13	25	**7**
$n = 50, n_1 = 10, \mu = 2$				
BIAS	1.1	-6.5	**-0.6**	-1.7
SD	6.9	**2.2**	5.9	3.6
MSE	49	47	35	**15**
$n = 50, n_1 = 20, \mu = 2$				
BIAS	-2	-9.7	**-1.5**	-4.2
SD	7	**2.7**	5.6	2.9
MSE	53	101	34	**25**
$n = 100, n_1 = 5, \mu = 3$				
BIAS	6.8	-3.4	1.5	**0.2**
SD	11	**1.8**	7	3.5
MSE	167	15	52	**12**
$n = 100, n_1 = 10, \mu = 2.5$				
BIAS	3.4	-5.8	**0.7**	-0.8
SD	9.7	**2.3**	7.8	4.4
MSE	106	39	61	**20**
$n = 100, n_1 = 20, \mu = 2$				
BIAS	**-0.6**	-11.1	-1.3	-3.6
SD	9.6	**3.2**	8.7	5.8
MSE	93	133	77	**47**
$n = 100, n_1 = 40, \mu = 2$				
BIAS	-4	-16.1	**-2.7**	-7
SD	10.8	**4.1**	8.3	4.2
MSE	132	278	75	**66**

Table 3.2 Bias, standard deviation and the MSE of the estimators, given the total number of hypotheses n, the number of false null hypotheses n_1 and the non-zero mean μ, based on 1000 repetitions. Bold and underlined values correspond to the smallest values in each row. The DOS method consistently achieves the smallest MSE.

	$\alpha = 1/2$	$\alpha = 3/4$	$\alpha = 1$	FIX
		$n_1 = 10, \mu = 3.5$		
BIAS	10.6	3.2	**1.0**	8.5
SD	16.0	7.3	4.3	21.1
MSE	368	63	**20**	518
		$n_1 = 30, \mu = 3.5$		
BIAS	6.4	**-0.1**	-2.9	2.8
SD	14.1	7.2	5.5	25.5
MSE	240	53	**38**	661
		$n_1 = 50, \mu = 3$		
BIAS	4.5	-4.0	-8.8	**0.3**
SD	16.2	10.3	8.2	27.9
MSE	283	**123**	144	777
		$n_1 = 100, \mu = 2$		
BIAS	**-4.9**	-20.9	-36.8	-5.6
SD	23.9	23.1	19.4	26.9
MSE	**596**	970	1729	754
		$n_1 = 100, \mu = 3$		
BIAS	**0.9**	-7.9	-13.8	-0.9
SD	16.4	12.2	10.4	27.0
MSE	271	**211**	298	728
		$n_1 = 200, \mu = 2$		
BIAS	-16.5	-29.5	-47.6	**-14.3**
SD	22.8	26.1	25.8	22.2
MSE	793	1553	2930	**700**

Table 3.3 Comparing the performance of α-DOS methods for different values of α, and the model parameters. FIX corresponds to Storey's estimator as in (2.12) with $\lambda = p_{(n/2)}$. Bold and underlined values correspond to the smallest values in each row.

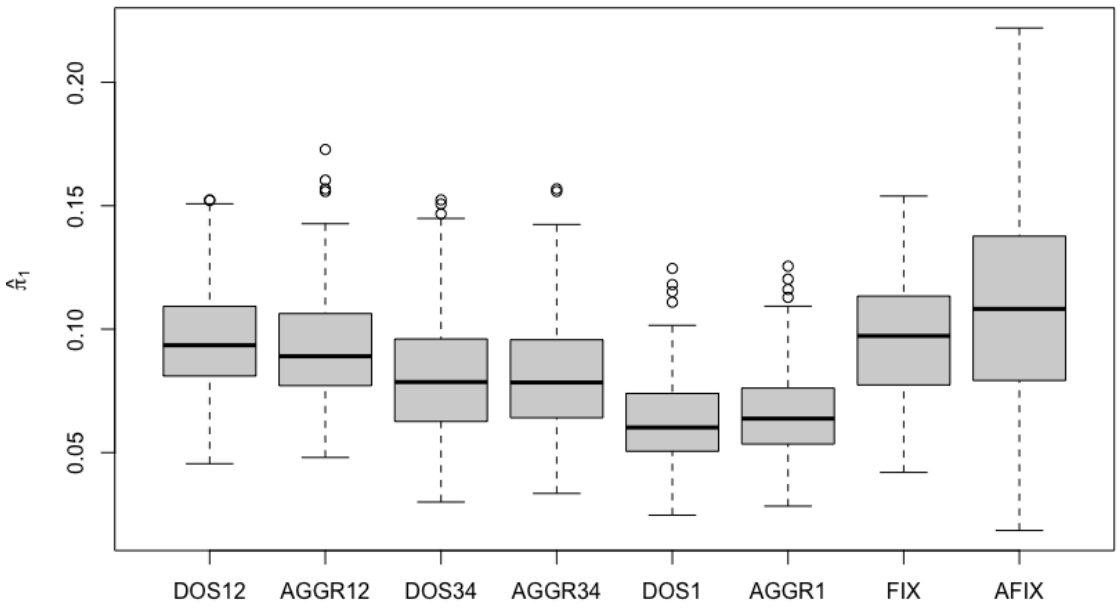


Fig. 3.11 Boxplots of the false null proportion estimates for the general α-DOS and the aggregated α-DOS procedure, for different powers $\alpha \in \{1/2, 3/4, 1\}$. The model is Gaussian mixture with $\pi_1 = 0.1$ and $\mu = 2$. FIX corresponds to Storey's estimator with $\lambda = p_{(n/2)}$.The number or repetitions is $N = 1000$.

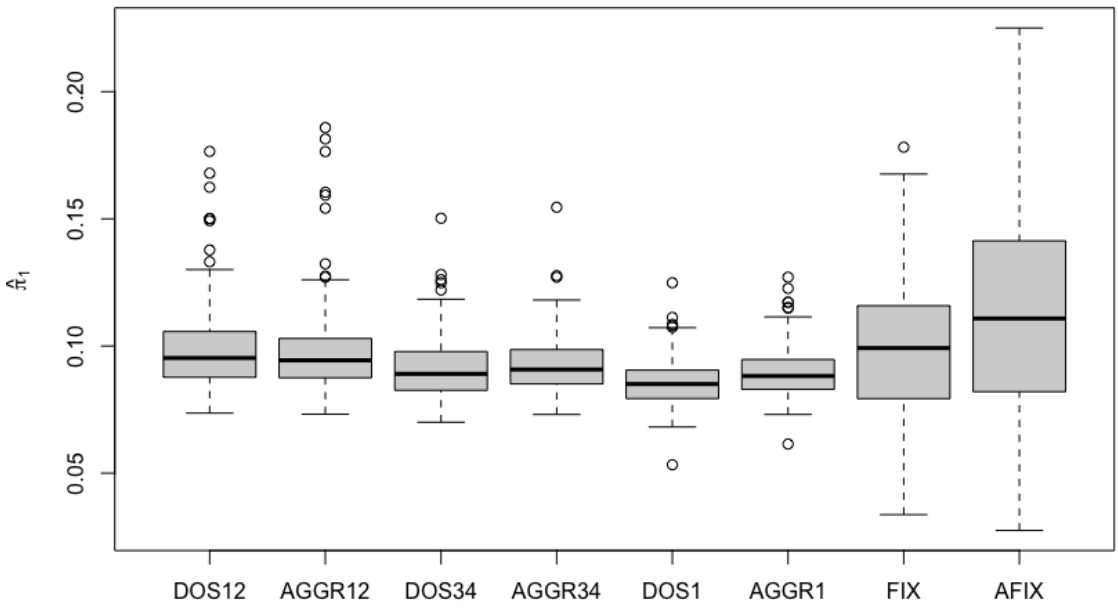


Fig. 3.12 Boxplots of the false null proportion estimates for the general α-DOS and the aggregated α-DOS procedure, for different powers $\alpha \in \{1/2, 3/4, 1\}$. The model is Gaussian mixture with $\pi_1 = 0.1$ and $\mu = 3$. FIX corresponds to Storey's estimator with $\lambda = p_{(n/2)}$. The number or repetitions is $N = 1000$.

Chapter 4

Interpretations and applications of change-point methods in multiple testing

In this chapter we highlight some existing methods that are used for both the change-point detection and the global testing problem, and explore the ways in which change-point inference (detection and estimation) can be used for solving multiple testing problems. We propose to use multiple change-point methods to divide the p-values into groups based on their significance. A potential usefulness of the p-value grouping is supported by a few examples in the applied literature, where p-values are divided into groups based on their significance but using seemingly arbitrary values.

This chapter is organised as follows. In Section 4.1 we describe a connection between the Higher Criticism and the CUSUM statistic used for change-point detection, and between the Berk-Jones statistic and the LR test for a change in the Poisson process. In Section 4.2 we discuss the properties of the spacings between p-values. In Section 4.3 we propose grouping p-values based on their significance by applying multiple change-point methods on the sequence of, possibly transformed, p-value spacings. We use some

existing, but suitably modified where necessary, algorithms for multiple change-point estimation. In Section 4.4 we briefly discuss possible applications for estimating the local false discovery rate, or estimating the proportion of false null hypotheses.

4.1 Change-point detection statistics for testing the global null

4.1.1 The Higher Criticism and the CUSUM statistic

In this subsection we describe the connection between the Higher Criticism (HC) and the Cumulative Sum (CUSUM) statistics which, to the best of our knowledge, has not been addressed in the literature. The CUSUM statistic comes from the change-point literature and is defined below, and the HC statistic is introduced in Section 2.3.3 as a global testing method. Indirectly, the relationship between the two statistics has been noted to hold as both are closely related to the Pontogram statistic by Kendall and Kendall (1980), which we also introduce below.

The HC statistic given by (2.25) is initially proposed as a statistic for global testing, however it has applications beyond this problem. In Donoho and Jin (2008) and Donoho and Jin (2009) the HC threshold is used for selecting useful features when training linear classifiers. In Klaus and Strimmer (2013), the properties of the HC as a thresholding method for signal identification are analysed. In this setting, when the global null is known to be false, and the goal is to identify all the false null p-values, the following modified version of the HC statistic is used:

$$\mathrm{HC}' = \max_{1 \leq i \leq n} \mathrm{HC}'(i),$$

where

$$\mathrm{HC}'(i) = \sqrt{n}\frac{i/n - p_{(i)}}{\sqrt{i/n(1-i/n)}}, \quad 1 \leq i \leq n.$$

The $\hat{k}_{\mathrm{HC}}$ smallest p-values are then declared as signal, where

$$\hat{k}_{\mathrm{HC}} = \underset{1\leq i\leq n}{\operatorname{argmax}}\, \mathrm{HC}'(i). \tag{4.1}$$

In Klaus and Strimmer (2013), the HC threshold is found to approximate a natural class boundary of a two-class linear discriminant analysis problem. The class boundary is a point where the Bayesian probabilities of a value coming from the null and the alternative component are equal – which is by definition the point where the local false discovery rate, defined in Section 2.1.2, is equal to 1/2.

To introduce the CUSUM statistic, we consider the problem of testing for a change in mean in a Gaussian sequence $X_1, \ldots X_n$, where $X_i \sim N(\mu_i, 1)$ and the hypotheses are given by:

$$H_0 : \mu_1 = \cdots = \mu_n \tag{4.2}$$

$$H_1 : \mu_1 = \cdots = \mu_k \neq \mu_{k+1} = \cdots = \mu_n. \tag{4.3}$$

The CUSUM statistic, which is the generalised likelihood ratio statistic for the testing problem above, is the maximum of the absolute CUSUM sequence:

$$\max_{i=1,\ldots,n} |\mathrm{CUSUM}(i)|, \tag{4.4}$$

where

$$\mathrm{CUSUM}(i) = \sqrt{\frac{i(n-i)}{n}}\left(\frac{1}{i}\sum_{k=1}^{i} X_k - \frac{1}{n-i}\sum_{k=i+1}^{n} X_k\right), \quad 1 \leq i \leq n. \tag{4.5}$$

The interpretation is that, at each candidate change-point i, the scaled difference of the means left and right from it is calculated, and if there is a change in mean, it will likely be where this difference is the largest. The CUSUM statistic is widely used in the change-point literature, also as a nonparametric method of testing for a change. Large values of $\max_{i=1,\dots,n} |\mathrm{CUSUM}(i)|$ indicate a presence of a change-point and the position of the maximum in the CUSUM sequence, $\operatorname{argmax}_{i=1,\dots,n} |\mathrm{CUSUM}(i)|$ is the estimated location of the change-point.

We now describe the relationship between the HC and the CUSUM statistic. Calculating the CUSUM statistic for the sequence of scaled p-values spacings $s_i := n(p_{(i)} - p_{(i-1)})$, $i = 1, \dots, n$, where $p_0 = 0$, gives the following approximation:

$$\begin{aligned}
\mathrm{CUSUM}(i) &= \sqrt{\frac{n}{i(n-i)}} \left[\sum_{j=1}^{i} s_j - \frac{i}{n} \sum_{j=1}^{n} s_j \right] \\
&= \sqrt{\frac{n}{i(n-i)}} \left[\sum_{j=1}^{i} n(p_{(j)} - p_{(j-1)}) - i \sum_{j=1}^{n} (p_{(j)} - p_{(j-1)}) \right] \\
&= \sqrt{\frac{n}{i(n-i)}} \left[np_{(i)} - ip_{(n)} \right] \\
&\approx \sqrt{n} \frac{np_{(i)} - i}{\sqrt{i(n-i)}} \\
&= \sqrt{n} \frac{p_{(i)} - i/n}{\sqrt{(i/n)(1-i/n)}} \\
&= -\mathrm{HC}'(i).
\end{aligned}$$

It follows that

$$|\mathrm{CUSUM}(i)| \approx |\mathrm{HC}'(i)|, \quad i = 1, \dots, n$$

In the above sequence of equalities, we used $p_{(n)} \approx 1$ approximation, as for large n their difference is negligible. Thus, we get that the Higher Criticism statistic comes very close to applying the CUSUM statistic for change in mean on a sequence of

scaled spacings of p-values s_i. However, the CUSUM statistic here is to be considered as a nonparametric method of change-point detection, because the model under the alternative is not specified in general and it is not Gaussian. The position of the maximum in the CUSUM (HC) sequence is then the location of the estimated change-point in mean in the spacings sequence. The means before and after the change-point are estimated by sample means of the values left and right from the estimated change-point. Note that the problem of estimating the change in mean in the sequence of p-values spacings is equivalent to the problem of estimating the change-point in slope in the sequence of sorted p-values. The illustration of the relationship between the HC and the CUSUM statistic is given in Figures 4.2 and 4.1. In Figure 4.1 a sequence of spacings is shown with the estimated change-point location that determines the piecewise constant approximation for the sequence of scaled spacings. In Figure 4.2 the HC sequence is shown, and the sequence of sorted p-values alongside the piecewise linear approximation induced by the HC threshold.

The relationship between the HC and the CUSUM statistic poses a question about a possible application of other change-point methods in the multiple testing literature, using statistics and procedures related to the CUSUM statistic. This reveals more existing relationships. Scan statistic, that has been used in Arias-Castro and Ying (2019) for global testing, can be seen as a scan statistic on the sequence of p-values spacings used in Olshen et al. (2004) for multiple change-point inference. For each subinterval of $\{1, \ldots, n\}$ the scan statistic compares the mean of scaled spacings in that subinterval with the total mean of the spacings, which is equal to 1. Another related method from the change-point literature is the Moving Sum (MOSUM) statistic by Eichinger and Kirch (2018). The MOSUM considers differences on symmetric intervals and is used for testing for the existence of multiple change-points and their estimation.

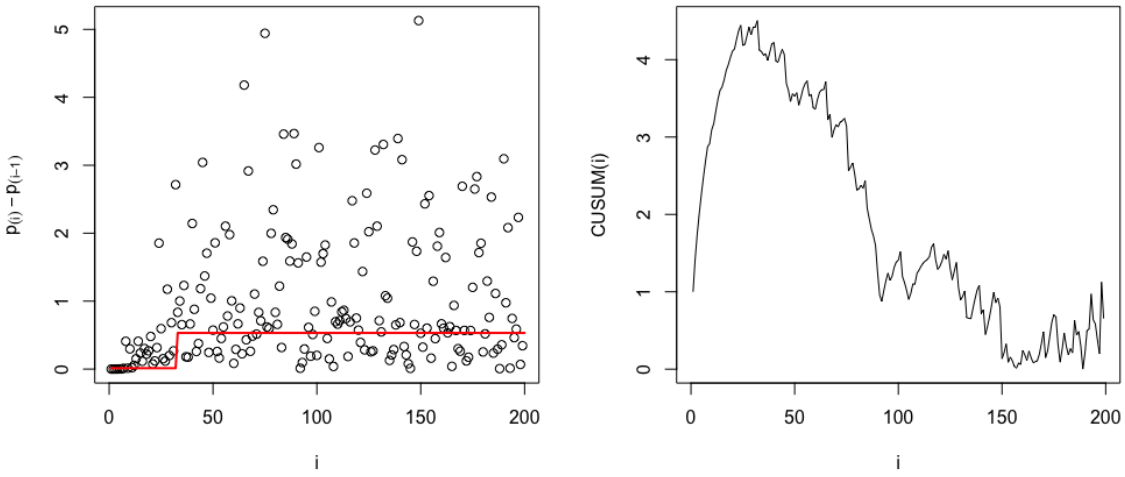


Fig. 4.1 The sample of 200 1-sided p-values, 20 of them non-null, from the Gaussian model with nonzero mean $\mu = 3$. Left: the sequence of spacings of the p-values and the piece-wise constant approximation. The change-point location is the location of the maximum of the CUSUM sequence. Right: The corresponding CUSUM sequence.

Given the window size G, the differences are first calculated on symmetric intervals

$$T_{k,n}(G) = \frac{\sum_{i=k+1}^{k+G} s_i - \sum_{i=k-G+1}^{k} s_i}{\sqrt{2G}} = n\frac{p_{k+G} - 2p_{k-G}}{\sqrt{2G}}, \quad G \le k \le n - G. \tag{4.6}$$

The MOSUM statistic is defined by taking the maximum of $T_{k,n}(G)$ over k:

$$\mathrm{MOSUM}(G) = \max_{G \le k \le n-G} \sqrt{n}\frac{p_{(k+G)} - 2p_{(k-G)}}{\sqrt{2G/n}}. \tag{4.7}$$

This procedure resembles the Difference of Slopes procedure introduced in Chapter 3. Particularly, its generalised form for $\alpha = 1/2$ described in Section 3.5. The difference is that the MOSUM considers sliding intervals of fixed width, while the DOS considers expanding symmetric intervals. However, the observations made in Section 3.2.1 demonstrate that under the usual assumptions on the p-value distribution, it is enough

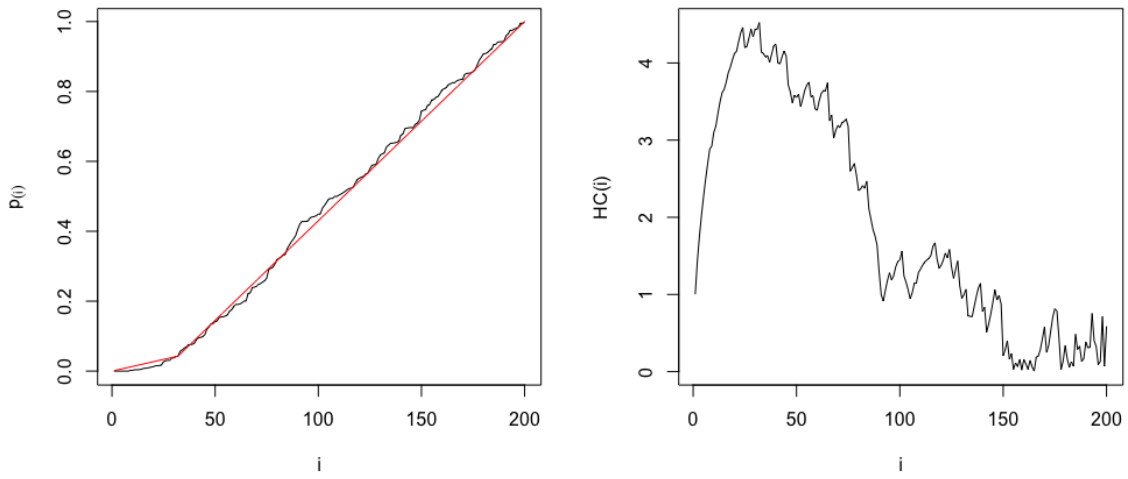


Fig. 4.2 The sample of 200 1-sided p-values, 20 of them non-null, from the Gaussian model with nonzero mean $\mu = 3$. Left: the sequence of sorted p-values and the piece-wise linear approximation. The location of the change in slope is the location of the maximum of the HC sequence. Right: The HC sequence.

to consider the expanding intervals, as false null p-values will be concentrated in the beginning.

Note that the CUSUM, MOSUM and the scan statistic from Olshen et al. (2004) all operate by comparing the scaled means between intervals, and all of them choose the objective intervals in a different way. Denote the discrete interval $\{s, s+1, \ldots, e\}$ as $[s, e]$. The CUSUM considers intervals $[1, i], [i, n]$, for all $i = 1, \ldots, n-1$, the MOSUM $[i - G + 1, i]$ and $[i + 1, i + G]$, for $G \leq i \leq n - G$, and the scan statistic in Olshen et al. (2004) compares means between intervals $[i, j]$ and $[1, n]$ for all $1 \leq i < j \leq n$. Furthermore, as Figure 4.2 suggests, we could consider some existing methods from the change-point literature for detecting changes in slope in a piecewise linear model, for example Baranowski et al. (2016). The piecewise linear model can be fitted to the sorted sequence of p-values, instead of the piecewise constant model for the spacings. However, the advantage of considering spacings instead of order statistics for p-values

is that, under the null, the p-values are uniformly distributed, and uniform spacings are asymptotically independent. Distributional properties of p-values spacings are further discussed in Section 4.2.

Below we describe how the relationship between the HC and the CUSUM statistic originates from the literature on testing for a change in the rate function of a Poisson process.

4.1.2 The Berk-Jones statistic and the Poisson process

In this subsection, we first define the Berk-Jones statistic and its applications. The relationship between the Berk-Jones and the Higher Criticism statistic is established in the literature through the Pontogram statistic by Kendall and Kendall (1980). The Berk-Jones statistic will be used in Section 4.2, as an alternative to the HC/CUSUM statistic.

The Berk-Jones statistic

Berk and Jones (1979) suggested the following goodness-of-fit statistic for testing if the given sample of $X_1, \ldots, X_n$ comes from the uniform $U[0,1]$ distribution:

$$\mathrm{BJ}^* = \max_{x \in (0,1)} \mathrm{BJ}(x), \tag{4.8}$$

where

$$\mathrm{BJ}(x) = F_n(x) \log \frac{F_n(x)}{x} + [1 - F_n(x)] \log \frac{1 - F_n(x)}{1 - x}. \tag{4.9}$$

and $F_n(x)$ is the empirical CDF of the sample. However, to the best of our knowledge, Chernoff and Rubin (1956) were the first to propose this statistic for a related but different problem. This statistic is derived as maximum likelihood estimator for the location of the discontinuity in a piecewise constant density on $[0,1]$. Equivalently,

we can see it as the MLE for the location of the change in slope in the corresponding CDF (or quantile function). The close relationship between the inference on piecewise constant density and Poisson process with piecewise constant rate function has been noted in Rubin (1961). To clarify, let $\lambda(t)$ be the intensity function of a Poisson process $\{N(t), t \in [0,1]\}$, and consider testing for a jump in the rate function $\lambda(t)$:

$$\begin{aligned} H_0 &: \lambda(t) = \lambda \\ H_1 &: \lambda(t) = \begin{cases} \lambda_1, & t \leq t_0 \\ \lambda_2, & t > t_0. \end{cases} \end{aligned} \tag{4.10}$$

Given the sequence of arrival times, and conditional on the number of events in a given interval, the GLR statistic for the above testing problem is equal to the GLR statistic for the analogous problem of testing for a discontinuity in a piecewise constant density. In the Poisson process context $F_n(x) = \frac{1}{n}N(x)$, where $N(1) = n$ and we condition on this event. The Berk-Jones statistic becomes

$$\mathrm{BJ}^* = \sup_{t \in (0,1)} \left\{ N(t)/n \log \frac{N(t)/n}{t} + (1 - N(t)/n) \log \frac{1 - N(t)/n}{1-t} \right\}.$$

Furthermore, the MLE for the location t_0 of the change in $\lambda(t)$ is equal to the MLE of the discontinuity point of a piecewise constant density with one jump. The consistency of the Berk-Jones statistic as an estimator of discontinuity point of a density is proven in Chernoff and Rubin (1956).

Pontogram

Pontogram is a method introduced by Kendall and Kendall (1980) for testing if there exist planned alignments in a set of points on a plane. Planned alignments are characterised by a large number of near-collinear triads of points. The authors identify

this problem with the problem of change-point detection in the rate of a Poisson process. The testing considered is as in (4.10). Under the null, the generating process is homogeneous and under the alternative the rate function is piecewise constant with one change-point such that the rate is larger before the change-point. The test statistic they propose is defined as:

$$\max_t Z(t), \tag{4.11}$$

where

$$Z(t) = \left(\frac{N(t)}{t} - \frac{n - N(t)}{1 - t} \right) \sqrt{\frac{t(1-t)}{n}}, \quad t \in (0, 1), \tag{4.12}$$

and $N(t)$ is the number of events in the interval $[0, t]$ and $n = N(1)$. Pontogram is then defined as a plot of $Z(t)$ against $t \in (0, 1)$. An illustration of Pontogram for a Poisson process with change in intensity can be seen in Figure 4.3.

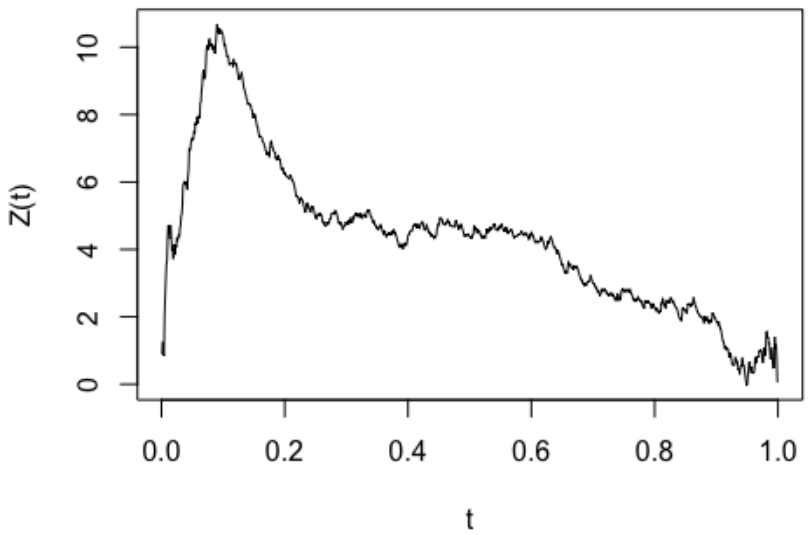


Fig. 4.3 The Pontogram for a realisation of a Poisson process with doubled intensity in the first tenth portion of the time-span.

In Donoho and Jin (2004), the authors briefly mention a close relationship between the Higher Criticism statistic and Pontogram. Specifically, it holds that the values in the Higher Criticism sequence are equal to the values of Pontogram function at arrival times $p_{(i)}$. We get this by replacing $N(t)$ with $nF_n(t)$, where $F_n(t)$ is the empirical CDF of the p-values (arrival times). Letting $t = p_{(i)}$, we have

$$\begin{aligned} Z(t) &= \left(\frac{N(t)}{t} - \frac{n - N(t)}{1-t}\right)\sqrt{\frac{t(1-t)}{n}} \\ &\left(\frac{nF_n(t)}{t} - \frac{n - nF_n(t)}{1-t}\right)\sqrt{\frac{t(1-t)}{n}} \\ &\left(\frac{i}{p_{(i)}} - \frac{n-i}{1-p_{(i)}}\right)\sqrt{\frac{p_{(i)}(1-p_{(i)})}{n}} \\ &= \sqrt{n}\frac{i/n - p_{(i)}}{\sqrt{p_{(i)}(1-p_{(i)})}} = \mathrm{HC}(t). \end{aligned} \tag{4.13}$$

We note that as a possible change-point location estimator we could use $\max_t Z(t)$. This statistic has been considered in Akman and Raftery (1986). They provide asymptotic theory and rates of convergence for $\max_t Z(t)$ as an estimator of the change-point location. Kendall and Kendall (1980) mention that, under the null hypothesis, $\max_t Z(t)$ is asymptotically close to the Berk-Jones statistic. For the proof of closeness between the HC and the Berk-Jones statistic see for example Lemma A.4. in Donoho and Jin (2004).

4.2 Spacings of p-values

In this section we discuss the properties of the p-values spacings and how best to fit them into the change-point methodology. Under the global null, p-values have $U[0,1]$ distribution, and the distributions of the ordered p-values are $p_{(i)} \sim \mathrm{Beta}(i, n)$. The distribution of the spacings $p_{(i+1)} - p_{(i)}$ is $\mathrm{Beta}(1, n)$ for any i. Additionally, the

scaled spacings $s_i = n(p_{(i+1)} - p_{(i)})$ are asymptotically independent and exponentially distributed with Exp(1) distribution.

In the previous section we have not made any assumptions on the p-value distribution under the alternative, however this assumption is implicitly made by considering the testing problem (4.10). To describe the behaviour of spacings in the presence of false null p-values, we first consider the uniform mixture model for p-values

$$p \sim \pi_1 U[0, b] + \pi_0 U[0, 1]. \tag{4.14}$$

Its density is piecewise constant with one change-point. This distribution is convenient as the generalised likelihood ratio test for the following global null testing problem against the mixture alternative:

$$\begin{aligned} H_0 &: p \sim U[0, 1] \\ H_1 &: p \sim \pi_1 U[0, b] + \pi_0 U[0, 1], \end{aligned} \tag{4.15}$$

is the Berk-Jones statistic. Furthermore, $\operatorname{argmax}_{t\in(0,1)} \mathrm{BJ}(t)$ is the MLE for b, the point where the uniform mixture density in (4.15) jumps, which is explained in the previous section. The parameter b also acts as a change-point in the distribution of the p-value spacings. The spacings between the p-values smaller than b all have the same distribution, different to the distribution of the spacings where p-values larger than b figure. The conditional distribution is given by:

$$p_{(i)} - p_{(i-1)} \sim \begin{cases} b\mathrm{Beta}(1, n(\pi_1 + (1-\pi_1)b)), & \text{if } p_{(i)} \leq b \\ (1-b)\mathrm{Beta}(1, n(1-\pi_1)(1-b))), & \text{if } p_{(i-1)} > b. \end{cases} \tag{4.16}$$

It follows that asymptotically, the distribution of s_i before the change converges to $\text{Exp}(\lambda_1)$, and after the change to $\text{Exp}(\lambda_2)$, where the rate parameters are given by $\lambda_1 = \frac{\pi_1 + b(1-\pi_1)}{b}$ and $\lambda_2 = 1 - \pi_1$. The location of the change is random, $\tau \sim \text{Bin}(n, \pi_1 + b(1-\pi_1))$.

The uniform mixture model for p-values is unrealistic, and in general there will be no change-point in the distribution in the sequence of p-values spacings. It is common to consider p-values that come from the Gaussian mean testing. The distribution of the p-values coming from the one-sided Gaussian mean testing, where under the alternative the distribution is $N(\mu, 1)$, $\mu > 0$, and under the null it is $N(0, 1)$, is a mixture with CDF

$$F_p(x) = \pi_1(1 - \Phi(\Phi^{-1}(x) - \mu)) + \pi_0 x, \quad x \in [0, 1]. \tag{4.17}$$

In this case, the distribution of the spacings under the alternative is intractable. p-values spacings, where the p-values come from the uniform mixture and from the mixture in (4.17) are shown in Figure 4.4. In the Gaussian testing setting, as the support of the null and the alternative component are both $[0, 1]$, we might not have an interval towards the end of the sequence that contains exclusively uniform spacings.

Assuming that the distribution of the p-values under the alternative has a support $[0, b]$, where $b < 1$ is of intermediate difficulty. In that case there is still a change-point separating the uniform spacings from the spacings in the beginning, where false null p-values figure, and it is justified to use the change-point method to estimate b . However, the false null spacings will be dependent and their distribution unknown in general. If the alternative distribution drops to zero at some $b \in (0, 1)$, then this can be considered a smooth change problem in the sequence of spacings. Looking at the sequence of scaled spacings s_i from right to left ($i = n$ to $i = 1$), the appearance of the first false null p-values changes the distribution of spacings and marks the onset of change. However, we do not consider the smooth change problem here.

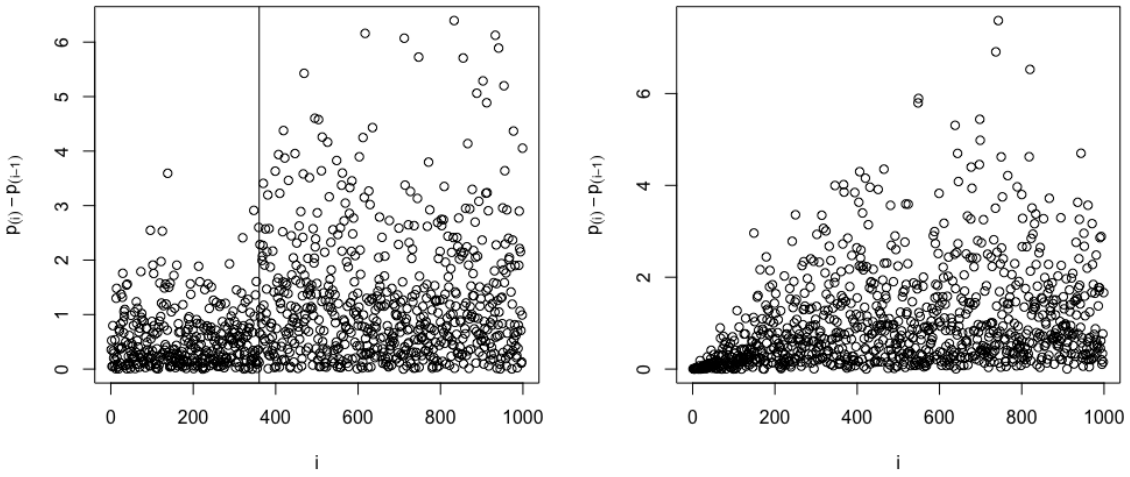


Fig. 4.4 Left: the sequence of p-values spacings from the uniform mixture model (4.14) with $\pi_1 = 0.2$, $b = 0.2$ and $n = 1000$, and the approximate change-point location at $n(\pi_1 + b(1 - \pi_1))$. Right: the sequence of p-values spacings from the model in (4.17) with $\pi_1 = 0.2$ and $\mu = 2$.

The most general statement on spacings is given in Pyke (1965), where the asymptotic exponentiality and independence of spacings is proved to hold for any CDF F of spacings with density f. Let $i/n \to t$, as $n \to \infty$. It holds that

$$s_i = n(p_{(i)} - p_{(i-1)}) \xrightarrow{D} \text{Exp}(f(F^{-1}(t)). \tag{4.18}$$

One of the common assumptions is that p-values come from a distribution with a decreasing density. This is considered by many methods for estimating the proportion of false null hypotheses introduced in Section 2.2. If f is decreasing, then the rate parameter $f(F^{-1})(t)$ is decreasing in t. This implies that the spacings of p-values have an increasing mean and variance and essentially approximate the *density quantile function* $f(F^{-1}(t)$. This is the case for example for the Gaussian p-values as seen in Figure 4.4.

To summarise, in the uniform mixture case, applying a change-point estimation method for a single change-point would effectively estimate the p-value distribution. Similarly, if the p-value distribution is a mixture of multiple uniform components, multiple change-point methods can be used for estimating the distribution. For p-value distributions with continuous densities there are no change-points. However, approximating the spacings with a piecewise constant function can be seen as a nonparametric function estimate giving some insights into the distribution of the p-values.

In the next section we explore possible applications of the change-point methods for modelling the sequence of p-values. Before applying a change-point algorithm to the sequence of spacings, we might want to deal with the unequal variance of the p-value spacings. We now consider some transformations of the spacings aiming to equalise the variance.

4.2.1 Transformed spacings

As the variance of spacings is increasing with the mean, we propose to use log or power transformation for the spacings before applying a change-point algorithm. First we consider the effects of such transformations on the spacings when p-values come from a uniform mixture (4.14). Let $X \sim \text{Exp}(1)$ and $Y \sim \text{Gumbel}(0, 1)$. It holds that

$$-\log X \stackrel{D}{=} Y.$$

This implies that logged scaled spacings, precisely $-\log(s_i)$, asymptotically have shifted Gumbel distribution. Before the change, that is for p-values smaller than b, the distribution of spacings is asymptotically $\log \lambda_1 + \text{Gumbel}(0, 1)$ and for p-values larger than b it is asymptotically $\log \lambda_2 + \text{Gumbel}(0, 1)$. Since $\lambda_1 > \lambda_2$, the jump is to a lower

level:

$$-\log\Big(n(p_{(i)}-p_{(i-1)})\Big) \sim \begin{cases} \log\lambda_1 + \text{Gumbel}(0,1), & p_{(i)} \le b \\ \log\lambda_2 + \text{Gumbel}(0,1), & p_{(i-1)} > b. \end{cases} \tag{4.19}$$

With log transformation we have transformed the problem of estimating the parameters of (4.14) into the change in mean problem, but instead of the usual standard Gaussian noise we have standard Gumbel noise, which has mean γ and variance $\pi^2/6$ where $\gamma \approx 0.5772$ is the Euler-Mascheroni constant. Thus, the log transformation solves the problem of increasing variance. The result on the asymptotic distribution of spacings from Pyke (1965) suggests that this holds in the general case as well.

Another possible transformation is the power transformation of the spacings

$$(n(p_{(i)}-p_{(i-1)}))^{1/4}. \tag{4.20}$$

The fourth-root transformation of the exponential random variable is considered for example in Kittlitz (1999). This power is chosen as it leads to a distribution with skewness that is very close to zero. As the distribution of the spacings is right skewed, this can be used to make it symmetric. Other exponents, close to 1/4 have also been used for this purpose. If $Y \sim \text{Exp}(\lambda)$ then the density of the transformed variable $Y^{1/4}$ is

$$f_{Y^{1/4}}(y) = 4y^3\lambda e^{-\lambda y^4}, \quad y \ge 0. \tag{4.21}$$

Note that this transformation maintains the quadratic relationship between the mean and the variance. Thus, this transformation only makes the noise symmetric.

The effect of log and power transformation on the p-values from the Gaussian model are shown in Figures 4.5 and 4.6.

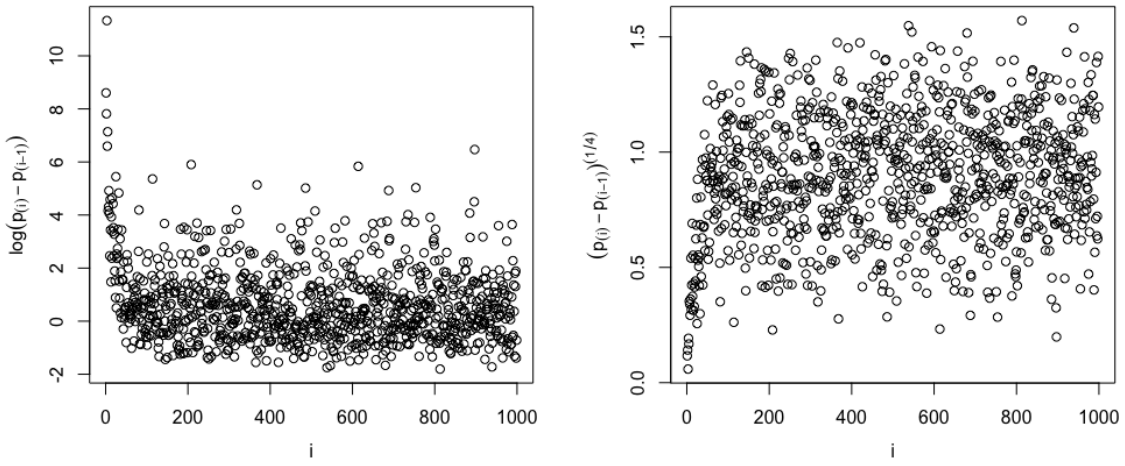


Fig. 4.5 Transformed spacings from the Gaussian model with $\pi_1 = 0.05$ and $\mu = 3$, $n = 1000$. Left: log-transformed spacings $-\log(s_i)$. Right: power-transformed spacings $s_i^{1/4}$.

4.3 Multiple change-point algorithms for p-values

The observations made in Section 4.1 motivated the following question: Can some recent advances in the change-point literature be used for solving more complex multiple testing problems, other than the global testing? In this section we propose to divide the p-values into multiple groups based on their significance. This approach to analysing the p-values of multiple testing has not been considered before, but it might be of interest in applications as described below. To this end we use the methods for multiple change-point estimation. Although, strictly speaking, there might not be change-points in the sequence of spacings, these algorithms are used as they result in nonparametric estimates for the underlying function of the (transformed) spacings. The change-points in the piecewise constant approximation for s_i induce the piecewise linear approximation of the CDF of p-values.

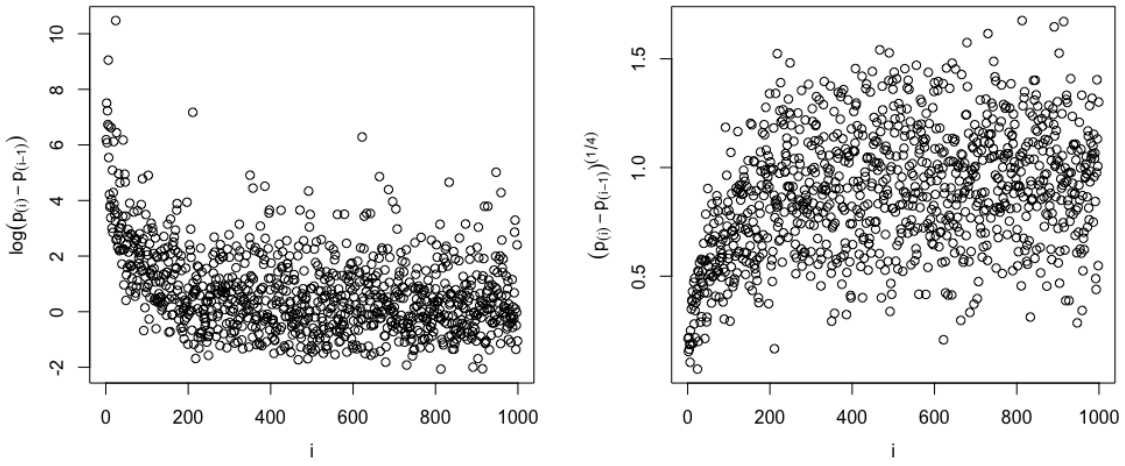


Fig. 4.6 Transformed spacings from the Gaussian model with $\pi_1 = 0.2$ and $\mu = 2$, $n = 1000$. Left: log-transformed spacings $-\log(s_i)$. Right: power-transformed spacings $s_i^{1/4}$.

4.3.1 Segmenting p-values into groups

Grouping p-values based on their significance is seen in some applied papers dealing with multiple testing problem. For that purpose some seemingly arbitrary constants are chosen, such as 0.1^{-k} or 5×0.1^{-k} which are commonly seen in the literature. A data-driven approach to grouping p-values might be of interest in applications, as in this way the groups of p-values are formed based on the behaviour of the whole sequence, rather than on some arbitrarily chosen constants. We list a few examples. In Raffaello et al. (2006) where gene expression data is considered, genes are divided into three classes, extremely, moderately and low significant genes. The classes are formed using their FDR values, such that highly significant p-values have FDR less than 0.01, moderately significant either 0.05 or 0.1, and low significant either 0.15 or 0.2. In Davidson and Shanks (2017) significant genes are divided into two groups, of highly and moderately significant genes. In Duwadi et al. (2018) significantly expressed genes

are divided into groups based on their value as: extremely significant ($p < 0.1^{-55}$), highly significant ($p < 0.1^{-15}$) and significant ($p < 0.001$). Aside from the applied literature, grouping p-values can be used for assigning the weights to p-values, when analysing p-values with additional information, precisely as in Basu et al. (2018), a method we described in Section 2.1.4. Similarly, the thresholds for the p-value groups formed in Basu et al. (2018) are constants of the form 0.1^{-k} or 5×0.1^{-k}.

4.3.2 Tools from the change-point literature

Here we review three different change-point algorithms that we use for the purpose of segmenting the sequence of p-values spacings. They are largely based on the existing methods, with some modifications introduced to make it more appropriate for the structure of the spacings sequence.

IDetect with Berk-Jones statistic

In Anastasiou and Fryzlewicz (2022), a method called Isolate-Detect (ID) is proposed for estimating multiple change-points in the piecewise constant mean. The CUSUM statistic is used for comparing the means on different intervals, and intervals are chosen as follows. Given a sample X_i, $i = 1, \ldots, n$, the ID procedure starts from the endpoints of the interval, and considers left- and right-expanding intervals until the interval on which the CUSUM statistic is greater than a pre-specified threshold is found. Given the step parameter h for the increasing length of the intervals the procedure alternately considers right and left intervals in the following order until the first change-point is detected:

$$[1, h], [n, n - h + 1], [1, 2h], [n, n - 2h + 1], \ldots .$$

The first change-point is detected when the first interval on which the CUSUM statistic is greater than a given threshold is found. If the change is detected at location b^* in one of the left expanding intervals $[n, n - ih + 1]$ for example, the intervals considered in the next step are

$$[1, h], [b^* - h + 1, b^*], [1, 2h], [b^* - 2h + 1, b^*], \ldots .$$

That is, the procedure is restarted considering the smaller interval $[1, b^*]$. Similarly, if the first change-point is detected in one of the right-expanding intervals, we consider $[b^*, n]$ in the next step. Note that the smallest possible value for h is $h = 3$.

With a few modifications we use this procedure for the problem of multiple change-point estimation in the spacings sequence. First, instead of the CUSUM statistic, we use the scaled Berk-Jones statistic for each subinterval $[s, e]$, defined by

$$\max_{s \le i \le e} \left\{ (i - s) \log\left(\frac{i - s}{p_{(i)} - p_{(s)}} \frac{p_{(e)} - p_{(s)}}{e - s} \right) + (e - i) \log\left(\frac{e - i}{p_{(e)} - p_{(i)}} \frac{p_{(e)} - p_{(s)}}{e - s} \right) \right\}. \quad (4.22)$$

The difference from the regular Berk-Jones statistic introduced above is the scaling factor $(p_{(e)} - p_{(s)})/(e - s)$. We use this scaling to make the values of the statistic between the intervals of different sizes comparable. With it, we scale both the domain and the codomain of the subinterval of p-values $p_{(s)}, \ldots, p_{(e)}$. The factor $1/(e - s)$ scales the domain to $[0, 1]$ while the factor $1/(p_{(e)} - p_{(s)})$ ensures that the spacings in the interval sum to 1. The second difference to the original ID procedure is that in search for the change-points we consider only left expanding intervals. As the spacings between the large p-values are likely uniform spacings, the idea is to proceed including the smaller p-values until there is enough evidence to reject the null. This first change-point defines the group of true null p-values. In the next steps the subset of p-values containing some false null p-values is segmented leading to increasingly

significant groups. For implementing the ID procedure with Berk-Jones statistic, the threshold is chosen empirically, such that the probability of the type I error 0.05. The empirical threshold guarantees that with large probability the procedure will not detect a change-point if there is not any. The illustration of the application of this procedure is shown in Figure 4.7. Instead of s_i, the sequence of power transformed spacings $s_i^{1/4}$ is shown in order to make the change-points in the beginning more visible, and also to make the results comparable to those of another method introduced below. We remark that, as the Berk-Jones statistic takes large values for both positive and negative jumps, the increasing signal is not guaranteed, so some post-processing might be needed.

The jumps here are to be interpreted as defining the linear approximation of the distribution while keeping the number of change-points small. In Figure 4.7, the p-values considered are the 1-sided p-values of the Gaussian mean testing, and their distribution is given in (4.17). Although there are only two regimes here, the estimated number of change-points will be larger than two in general, with more change points estimated for stronger alternatives. Only if the distribution of the p-values is a mixture of uniform distributions we can expect the number of groups estimated to be equal to the number of components in the uniform mixture.

Unbalanced Haar-Fisz technique

In Fryzlewicz (2007), a wavelet-based method for nonparametric function estimation is proposed for the model

$$X_i = f(i/n) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{4.23}$$

where $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. We use a slightly modified version of this procedure for approximating the sequence of p-values spacings with a piecewise constant function. The modification will account for the non-constant variance of the spacings, which is seen

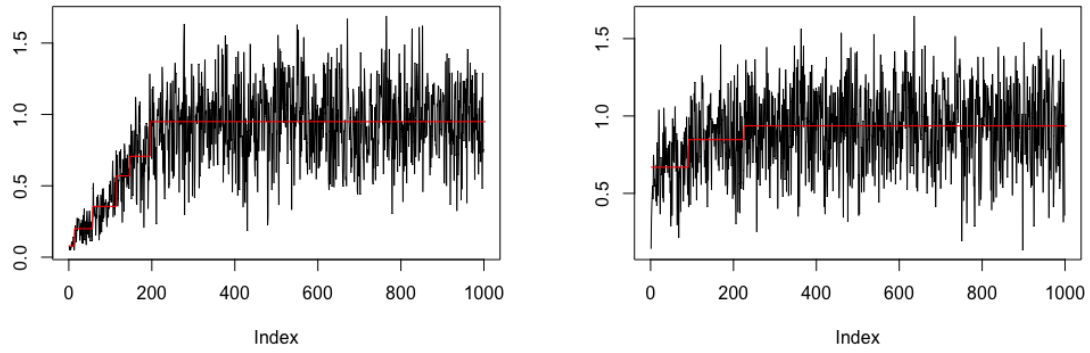


Fig. 4.7 The IDetect with Berk-Jones statistic for p-values from the Gaussian model. Left: The sequence $s_i^{1/4}$, and the fitted piecewise constant function where the model parameters are $\mu = 3, \pi_1 = 0.2$ and the sample size is $n = 1000$. Right: The sequence $s_i^{1/4}$, and the fitted piecewise constant function where $\mu = 2, \pi_1 = 0.1$.

.

in Figure 4.4 and also in the sequence of power transformed spacings $s_i^{1/4}$ in Figures 4.5 and 4.6. We explain the procedure in detail below.

A general approach to the nonparametric function estimation and denoising using wavelet thresholding consists of the following steps:

1. Choose the wavelet basis and transform the data to wavelet coefficients.

2. Threshold the wavelet coefficients.

3. Take the inverse wavelet transform of the thresholded coefficients to get the denoised function estimate.

Haar wavelet basis is the simplest wavelet basis consisting of rescaled step functions. Therefore, the procedure stated above using Haar basis would result in a piecewise constant estimate of the function. The unbalanced Haar (UH) procedure of Fryzlewicz (2007) follows these steps, and the basis used is Haar-like, but chosen adaptively from

the data. The basis vectors are of the form

$$\psi_{s,b,e}(l) = \left\{\frac{1}{e-s+1}\right\}^{1/2} \mathbb{I}(s \le l \le b) - \left\{\frac{1}{e-b} - \frac{1}{e-s+1}\right\}^{1/2} \mathbb{I}(b+1 \le l \le e), \tag{4.24}$$

where $\mathbb{I}$ is the indicator function. The number of vectors in the basis and their starting, jumping and ending points will depend on the data. Note that the classic Haar basis vectors can be seen as (4.24), where s and e are such that all dyadic subintervals of $[1, n]$ are considered, and the jump in the vector is set to the middle of the interval, $b = (e - s + 1)/2$. Unlike the classic Haar basis vectors, here the jump b in the basis vector is not always in the middle of the interval, and s, b and e are chosen from the data, as explained below. First, the interval $[s, e] = [1, n]$ is considered and its jump point $b_{0,1}$ is chosen as $b_{0,1} = \text{argmax}_{s \le b \le e} |\langle \mathbf{X}, \psi_{s,b,e} \rangle|$. The corresponding wavelet coefficient is $d_{0,1} = \langle \mathbf{X}, \psi_{s,b_{1,1},e} \rangle$. The jump point $b_{0,1}$ in the basis vector at the coarsest scale then becomes the endpoint of the basis vectors at a next, finer scale, and the procedure is repeated on vectors $[1, b_{0,1}]$ and $[b_{0,1}, n]$ to find the jump points $b_{1,1}$ and $b_{1,2}$ (and the coefficients) of the finer scale vector basis. This procedure is continued in a recursive way until the length of the basis vectors becomes too small, $e - s \ge 2$. The obtained sequence of wavelet coefficients is then thresholded at level $\hat{\sigma}\sqrt{2 \log n}$, where $\hat{\sigma}$ is the estimated standard deviation of the sample, and then transformed back to the original domain, yielding a nonparametric function estimate. In Fryzlewicz (2007), the relationship between this method and the binary segmentation procedure is noted. Binary segmentation (BS) by Vostrikova (1981) is a widely use method for detecting multiple change-points in mean by recursively applying the CUSUM transform on the subintervals. In the first step, the CUSUM statistic for the interval $[1, n]$ is calculated, and if it is large enough, the first change-point $b_{0,1}$ is detected at the location where the CUSUM sequence is maximised. Analogously to the UH procedure described above, the detected change-point is used to divide the interval into two, $[1, b_{0,1}]$ and

$[b_{0,1}+1, n]$, and then the CUSUM procedure is applied on both subintervals. If the CUSUM statistic is below the chosen threshold on a given interval, the segmenting stops on that interval. We want to apply the UH/BS procedure to the sequence of spacings, however straightforward application of this procedure is not appropriate as the method assumes Gaussian noise with constant variance. In our case, for scaled spacings s_i or power-transformed spacings $(s_i)^{1/4}$, the variance increases with the mean as shown in Figures 4.5 and 4.6. The relationship between the mean and the variance is quadratic, which follows from the exponentiality in Section 4.2. This corresponds to the multiplicative model for the data $X_i = s_i$ (or some transformation of s_i):

$$X_i = \sigma^2(i/n)\varepsilon_i^2, \quad i = 1, \ldots, n, \tag{4.25}$$

where $\varepsilon_t \overset{iid}{\sim} N(0,1)$. Applying the UH (BS) procedure, we can eliminate the effect of the changing variance from the CUSUM statistic by scaling it with the mean of the data on a given subinterval. That is

$$\frac{\text{CUSUM}(X[s:e])}{\frac{1}{e-s+1}\sum_{i=s}^{e} X_i}, \tag{4.26}$$

where $\text{CUSUM}(X[s:e])$ is the CUSUM statistic on the subsequence $X_s, \ldots, X_e$. This type of transformation for variance stabilisation has been considered in Fryzlewicz et al. (2007), where a wavelet-based data transformation method is proposed for stabilising the variance assumed to depend on the mean. It is based on the Haar-Fisz transform, proposed in Fryzlewicz and Nason (2004) and incorporated in a wavelet-based method for estimating the intensity of a Poisson process. The idea is to scale the Haar coefficients such that they have asymptotically Gaussian distribution and a constant variance. As the CUSUM statistic can be seen as a coefficient in the unbalanced Haar basis, (4.26) is a scaled unbalanced Haar coefficient. The threshold used is calculated

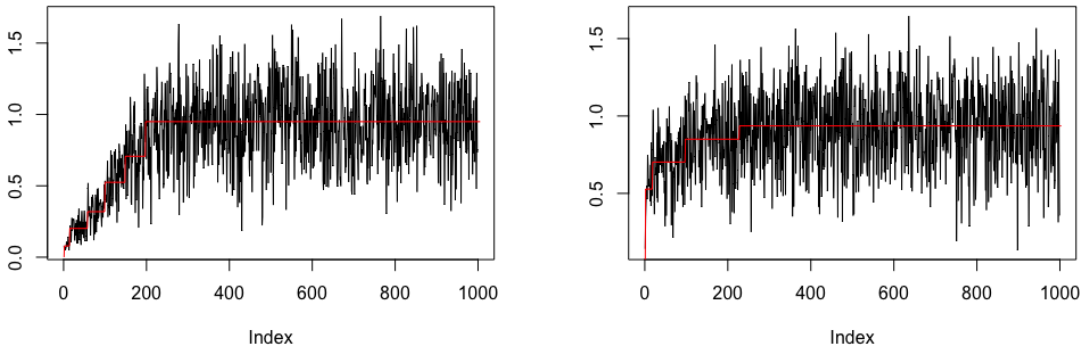


Fig. 4.8 The Unbalanced Haar-Fisz procedure applied on the sequence of scaled spacings s_i, of p-values from the Gaussian model. Left: The sequence $s_i^{1/4}$, and the fitted piecewise constant function where the model parameters are $\mu = 3, \pi_1 = 0.2$. Right: The sequence $s_i^{1/4}$, and the fitted piecewise constant function where $\mu = 2, \pi_1 = 0.1$.

empirically, such that the probability of detecting a change in the sequence of uniform spacings is 0.05. The illustration of the procedure applied on the sequence $X_i = s_i$ is shown in Figure 4.8, on the same data as in Figure 4.7. Again, instead of s_i we show the power transformed spacings in order to make the change-points in the beginning more visible, and to make the results comparable to those of the method introduced below. We note that in general the UH and the ID procedures estimate the change-points at similar locations.

NOT with piecewise constant mean and variance

In Baranowski et al. (2019), the Narrowest-Over-Threshold (NOT) method for estimating multiple change-points in the piecewise linear or piecewise constant mean function is proposed. The method also allows for the piecewise constant model for the variance. We use this algorithm for piecewise constant approximation of the sequence of transformed spacings assuming the piecewise constant model for the mean and variance. In contrast to the unbalanced Haar-Fisz method, here the variance is not assumed to be changing with mean and the statistic used is the GLR statistic for this problem. As

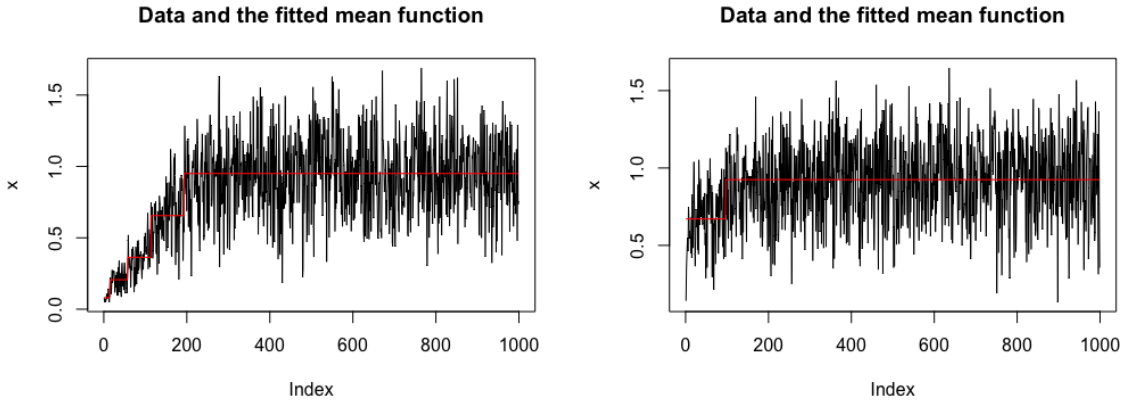


Fig. 4.9 The NOT procedure for the piecewise constant mean and variance applied on the sequence of power transformed spacings $s_i^{1/4}$ of p-values from the Gaussian model. Left: The sequence $s_i^{1/4}$, and the fitted piecewise constant function where the model parameters are $\mu = 3, \pi_1 = 0.2$. Right: The sequence $s_i^{1/4}$, and the fitted piecewise constant function where $\mu = 2, \pi_1 = 0.1$.

in the other multiple change-point procedures, this statistic is used on the subintervals of the data. The NOT procedure aims to isolate the change-points using short intervals that are randomly sampled from the data. At each step, a change-point is detected from the shortest interval on which the GLR statistic exceeds a pre-specified threshold. The method then proceeds recursively by looking for change-points on the two induced subintervals. We apply this procedure on the transformed spacings $(s_i)^{1/4}$ instead of s_i, which yields better results as their distribution is approximately symmetric. The illustration of the procedure is shown in Figure 4.9. In weak cases, as seen on the right-hand plot, this procedure results in fewer intervals in the segmentation of the spacings sequence.

4.3.3 Applications

Aside from grouping p-values, piecewise constant approximation of the p-value density can be applied, for example, for estimating the proportion of false null hypotheses or for constructing procedures controlling some multiple testing error rates. Piecewise

constant approximation of the p-value density induces piecewise constant approximation of the local FDR function, which is introduced in Section 2.1.2 and defined by

$$\text{lfdr}(t) = \frac{\pi_0}{\pi_1 f_1(t) + \pi_0}.$$

The local FDR is a Bayesian quantity, equal to the probability of a p-value being null given that it takes value t, where π_0 is the proportion of true nulls and f_1 is the density under the alternative.

In Figures 4.10 and 4.11, the NOT approximation of the lfdr function is shown alongside true lfdr function and the lfdr estimate from Strimmer (2008) implemented in R package 'fdrtool' by Klaus and Strimmer (2021). The method proposed in Strimmer (2008) first estimates the null proportion and the parameters of the null distribution (that is allowed to be non-uniform) using the truncated maximum likelihood approach. Using this information, a modified Grenander density estimator is used to compute the overall density, from which the lfdr estimate follows. When using this method we specify the uniform model for the true null p-values. We observe that the two lfdr estimates are close, however our change-point lfdr estimator is less precise, as the number of groups obtained from the change-point algorithm is always kept relatively small.

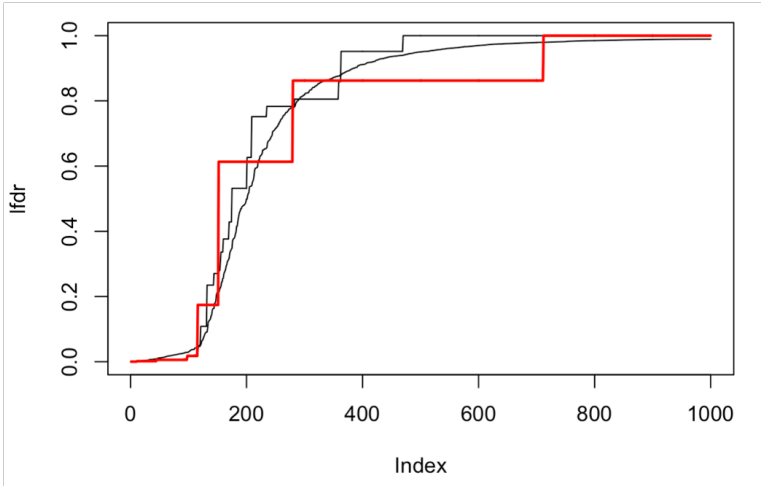


Fig. 4.10 Black piecewise constant function: The estimator of the local FDR using 'fdrtool' package. Black curve: The true local FDR function. Red piecewise constant function: The local FDR estimate using the change-point locations obtained by the ID procedure with Berk-Jones statistic. p-values come from the Gaussian model (4.17), where $\pi_1 = 0.3$ and $\mu = 2$.

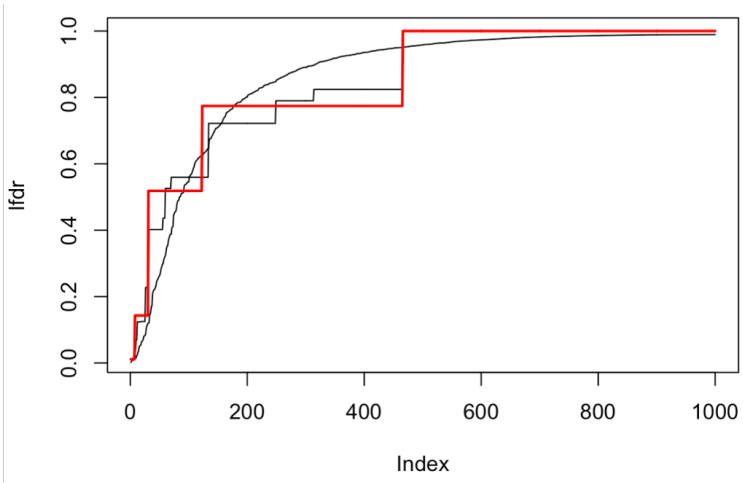


Fig. 4.11 Black piecewise constant function: The estimator of the local FDR using 'fdrtool' package. Black curve: The true local FDR function. Red piecewise constant function: The local FDR estimate using the change-point locations obtained by the ID procedure with Berk-Jones statistic. p-values come from the Gaussian model (4.17), where $\pi_1 = 0.1$ and $\mu = 2$.

4.4 Discussion

In this section we discuss some closely related ideas for future research. First, there are many options for the multiple change-point techniques that can be used. The Berk-Jones statistic was used with the ID algorithm, but we can incorporate the Berk-Jones statistic into any of the usual CUSUM-based algorithms. We can also consider logged spacings with CUSUM-based algorithms. It is of interest to study the induced grouping of p-values as a way of incorporating additional information obtained from the previous studies as in Basu et al. (2018) and in that way improving the power of the multiple testing procedure.

The resulting change-point lfdr estimate can be used for estimating the proportion of false null hypotheses, by finding the point where the lfdr estimate jumps to 1 and combining it with the Storey's estimator (Storey (2002)). Another possible application of the change-point lfdr estimate is in constructing a multiple testing procedure, similar to the one in Sun and Cai (2007). In Sun and Cai (2007), a multiple testing procedure based on the estimated local FDR values (l-values), is proved to be optimal in the sense that it minimises the FNR while controlling the FDR at a prescribed level. Let $\widehat{\text{lfdr}}(p_{(i)})$ be the estimated local FDR of the ith smallest p-value. The procedure rejects the hypotheses corresponding to the smallest k l-values, where

$$k = \max\{i : \frac{1}{i}\sum_{j=1}^{i} \widehat{\text{lfdr}}(p_{(i)}) < \alpha\}. \tag{4.27}$$

In Sun and Cai (2007) the lfdr is estimated by estimating the distribution of the z-values (test statistics, assumed to come from a mixture of Gaussians). They use a method from Jin and Cai (2007) for consistently estimating the null proportion and the distribution under the null assuming Gaussianity. A future direction would be to

use the change-point lfdr estimate in combination with this method to construct a new FDR-controlling method.

Chapter 5

Tail-summed Scores method for multiple testing and signal estimation

In this chapter we consider the problem of selecting the nonzero mean components in the Gaussian sequence model defined by

$$X_i = \mu_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{5.1}$$

where $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ is the unknown mean vector. The mean vector $\boldsymbol{\mu}$ is assumed to be sparse in a sense that there is a large proportion of terms that are equal to zero – the non-signal values. We denote by k the number of nonzero mean values, also referred to as the signal values. The exact assumptions regarding k will be introduced as we go along. We now assume that $\sigma^2 = 1$, and we work under this model. Possible methods for estimating σ^2 are introduced in Section 2.3.2.

This model and its importance are discussed in detail in Section 2.3.1. We propose a new thresholding method that we call the Tail-Summed Scores (TSS). There are two

approaches to analysing the thresholding estimators for the Gaussian sequence model, the signal estimation and the multiple testing approach. The optimality criterion depends on the approach taken. In the first case, the goal is to control the multiple testing error rate, while in the second one the goal is to minimise the mean squared error of the estimator. We discuss the performance of the TSS method in both cases. This chapter is organised as follows. In Section 5.1 we introduce the TSS method. Its properties and first theoretical results under some special cases are given in Section 5.2. Some theoretical results in the general case are stated in Section 5.3. Simulations and possible applications to change-point analysis are outlined in Section 5.4. A brief discussion is given in Section 5.5 and the proofs of the theoretical results can be found in Section 5.6.

5.1 TSS method

Let $X_1, \ldots, X_n$ be the sample from (5.1) with $\sigma^2 = 1$, and define $Y_i = X_i^2$, $i = 1, \ldots, n$. The non-signal values Y_i have χ_1^2 distribution, while the signal values have the noncentral χ_1^2 distribution with noncentrality parameter μ^2, denoted by $\chi_1^2(\mu^2)$. Let $S = \{i : \mu_i \neq 0\}$ be the subset of signal values and $k = |S|$ the number of signals. Let $Y_{(i)}$ be the i-th order statistic of the sequence $Y_1, \ldots, Y_n$, the order being increasing, i.e. $Y_{(1)} = \min_i Y_i, Y_{(n)} = \max_i Y_i$. Let $\rho_Y(j) = i$ if $Y_{(j)} = Y_i$. We wish to base our testing and recovery on the tail-sums of $Y_{(i)}$. Define the sequence of variables

$$T_i = \sum_{j=1}^{n-i+1} Y_{(j)}, \quad i = 1, \ldots, n. \tag{5.2}$$

The idea is to sequentially test each T_i, $i = 1, \ldots, n$, for the exceedance of a certain threshold $\lambda_i > 0$. Namely, we stop the sequential testing procedure as soon as we come

across an index i_0 for which $T_{i_0} < \lambda_{i_0}$. We then estimate the set S as

$$\hat{S} = \{i : \rho_Y(j) = i, \text{ for } j = n - i_0 + 2, \dots n\},$$

with $\hat{S} = \emptyset$ if $i_0 = 1$. We denote the estimated number of signal values as $\hat{k} = |\hat{S}|$. At each step, the TSS procedure excludes the largest of the remaining Y-values one by one, until the presence of the signal in the remaining set becomes insignificant, and the remaining set starts resembling the sample of χ^2_1 values. This requirement on the remaining set should be reflected in the sequence of thresholds, the choice of which we discuss below. The illustration of the procedure can be seen in Figure 5.1. Using the TSS procedure we aim to include more signal terms in the set $\hat{S}$ when there are many weak signals in the sequence. The weak signals would be aggregated so that the TSS procedure would proceed even if the remaining signals are indistinguishable on their own, as long as the total remaining signal is strong enough. The pseudo-code for the TSS procedure is given in Algorithm 1.

5.1.1 Choosing the thresholding sequence

The perfect separation case

We motivate the choice of the thresholding sequence by considering the *perfect separation* case, when the smallest signal variable is larger than the largest non-signal variable with large probability. Let $|S| = k$ and let

$$Y^S_{(k)} \geq Y^S_{(k-1)} \geq \dots \geq Y^S_{(1)},$$

be the decreasingly sorted signal values of the sequence $Y_1, \dots, Y_n$. Let

$$Y^{NS}_{(n-k)} \geq Y^{NS}_{(n-k-1)} \geq \dots \geq Y^{NS}_{(1)},$$

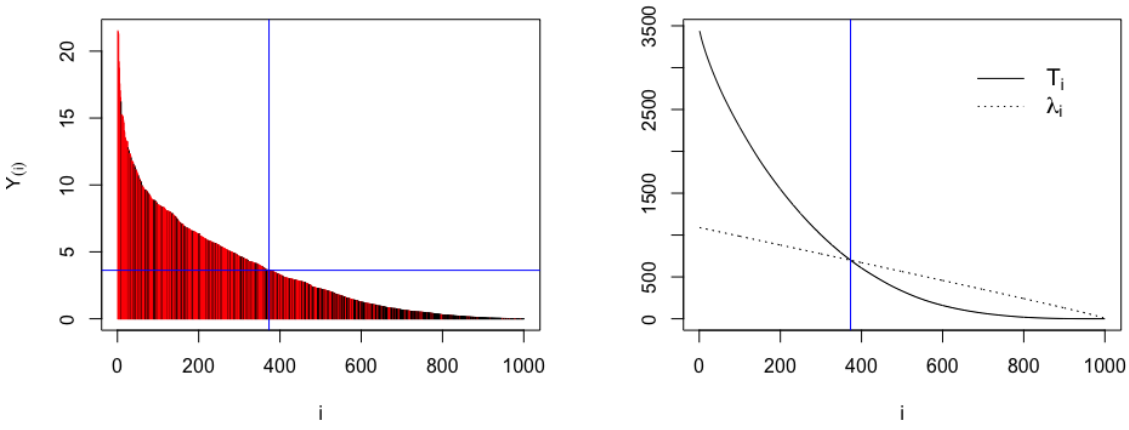


Fig. 5.1 The illustration of the TSS method, for the sample of size $n = 1000$ from model (2.20) where $\sigma^2 = 1$, $\mu_i = 2$, $i = 1, \ldots, n$ and $|S| = 600$. Left: $Y_{(i)}$ sequence, black bars correspond to the zero mean, red bars to the nonzero mean terms. Right: The TSS sequence T_i and the sequence of thresholds λ_i. Vertical blue line is at $i = 373$, which is the estimated number of signals. Horizontal and vertical blue line on the left plot mark the threshold value of the method.

be the decreasingly sorted non-signal values. Let Ω_n be the perfect separation event:

$$\Omega_n = \{Y^S_{(1)} \geq Y^{NS}_{(n-k)}\}. \tag{5.3}$$

If $P(\Omega_n) \to 1$, $n \to \infty$ then we say that the *perfect separation assumption* holds. To have this, a strong enough signal $\mu = \mu(n)$ is required, and we discuss a sufficient condition for this in Section 5.2. On the event Ω_n, $Y_{(1)} \leq \cdots \leq Y_{(n-k)}$ are the non-signal values, and $Y_{(n-k+1)} \leq \cdots \leq Y_{(n)}$ are the signal values. In general, conditional on Ω_n, the distribution of the sample changes, but the assumption $P(\Omega_n) \to 1$ allows us to use the unconditional chi-squared distribution for obtaining asymptotic statements. Note that if there is no signal, $k = 0$, we can also say that the perfect separation holds.

Algorithm 1: The TSS procedure

Data: the sample $X_1, \ldots, X_n$ and the thresholds $\lambda_1, \ldots \lambda_n$

Result: the estimated subset of signals $\hat{S}$, and the number of signals $\hat{k}$

$(Y_1, \ldots, Y_n) \leftarrow (X_1^2, \ldots, X_n^2)$;

$T \leftarrow \text{sum}(Y_1, \ldots, Y_n)$;

$(\rho_1, \ldots, \rho_n) \leftarrow \text{arg\_sort}(Y_1, \ldots, Y_n)$; `// the sorting permutation of indices`

$(Y_{(1)}, \ldots, Y_{(n)}) \leftarrow \text{sort}(Y_1, \ldots, Y_n)$;

$i \leftarrow 1$;

if $T < \lambda_1$ **then**

  return $(\emptyset, 0)$

end

while $T \geq \lambda_i$ **and** $i \leq n$ **do**

  $T \leftarrow T - Y_{(n-i+1)}$;

  $i \leftarrow i + 1$;

end

$\hat{S} := \{\rho_n, \ldots, \rho_{n-i+2}\}$;

$\hat{k} := i - 1$;

return $(\hat{S}, \hat{k})$

In the perfect separation case, the problem of multiple testing in the Gaussian sequence model is equivalent to testing the following sequence of nested hypotheses:

$$\begin{aligned} H_{0,i} &: \text{there are fewer than } i \text{ signal variables,} \\ H_{1,i} &: \text{there are at least } i \text{ signal variables.} \end{aligned} \tag{5.4}$$

In this case we can view the TSS procedure as a method for testing (5.4) which comes naturally as the statistics T_i are aggregating the sample values. T_i is a test statistic for the null hypothesis $H_{0,i}$, and λ_i is a critical value. If $H_{0,i}$ holds, on Ω_n, T_i has χ^2_{n-i+1}

distribution. We first consider the following sequence of thresholds:

$$\lambda_i = q_{\chi^2_{n-i+1}}(1-\alpha), \quad i = 1, \ldots, n, \tag{5.5}$$

where $q_{\chi^2_{n-i+1}}(\cdot)$ is the quantile function of the χ^2_{n-i+1} distribution. The thresholds in (5.5) guarantee asymptotic control of the FWER at level α under the perfect separation assumption, in the sense that

$$P(V > 0) \to \alpha, \quad n \to \infty, \tag{5.6}$$

where V is the number of false rejections made by the TSS procedure. Let k be the true number of signal values. It holds that:

$$\begin{aligned} P(V > 0) &= P(V > 0, \Omega_n) + P(\Omega_n^{\complement}) \\ &\leq P(\chi^2_{n-k+1} > \lambda_k) + P(\Omega_n^{\complement}) \\ &= P(\chi^2_{n-k+1} > q_{\chi^2_{n-k+1}}(1-\alpha)) + P(\Omega_n^{\complement}) \\ &\to \alpha. \end{aligned} \tag{5.7}$$

We note here that by aggregating the individual hypotheses of the multiple testing problem and considering instead the sequence of tests in (5.4) we avoid the multiplicity problem. The TSS procedure controls the FWER at level α without the need for adjusting for multiplicity. As it holds that FDR $\leq$ FWER under any configuration of true and false null hypotheses, shown in Section 2.1.2, with this choice of thresholds the TSS also controls the FDR asymptotically at level α in the perfect separation case.

The thresholds in (5.5) are motivated and justified under the assumption that $P(\Omega_n) \to 1$, which is restrictive as it requires large signal values μ. Below, we suggest some approximations for the proposed quantile sequence of thresholds, that make the

theory more convenient in the case when the perfect separation assumption does not hold.

Related threshold sequences

The concentration inequality for the upper tail of the chi-square distribution (Boucheron et al., 2013) gives rise to the following sequence of thresholds

$$\lambda_i^{H_i} = n - i + 1 + H_i\sqrt{2(n-i+1)}, \tag{5.8}$$

where

$$H_i = \sqrt{2\log\frac{1}{\alpha}} + \sqrt{\frac{2}{n-i+1}}\log\frac{1}{\alpha}. \tag{5.9}$$

Similarly, these thresholds control the FWER at the same level α as the exact quantile thresholds. In general, thresholds of the form (5.8) can be used for different sequences of values $H_i \geq 0$, $i = 1, \ldots, n$. H_i values can all be the same, or different for each i, and they can also depend on n as in (5.9), which we omit from the notation. For instance, Gaussian approximation of the chi-square quantiles

$$q_{\chi^2_k}(1-\alpha) \approx k + q_{N(0,1)}(1-\alpha)\sqrt{2k}. \tag{5.10}$$

also yields the thresholds of the form (5.8), where $H_i = q_{N(0,1)}(1-\alpha)$. For $\alpha \approx 0.05$, we get the following simplified sequence of thresholds

$$\lambda_i = n - i + 1 + 2\sqrt{2(n-i+1)}.$$

By tuning H_i, we can manipulate the conservativeness of our procedure. Asymptotic behaviour of the thresholds (5.8) is more obvious than of those in (5.5), hence the thresholds given by (5.8) will be used in most of the theoretical results below. Note

that for any $H_i = H_i(n) \to \infty$ we will have that the FWER of the TSS procedure converges to zero as $n \to \infty$ under the perfect separation assumption.

Thresholds $\lambda_i^{H_i}$ satisfy the following property:

$$T_i \leq \lambda_i^{H_i} \implies T_{i+1} \leq \lambda_{i+1}^{H_{i+1}},$$

where H_i is any non-decreasing sequence of values. This is stated and proved in Proposition 1 and it shows that $\lambda_i^{H_i}$ thresholds enable us to reject only the contiguous block of hypotheses at the beginning, under no additional assumptions on the values in the sample. A related statement is given in Proposition 2, where we bound the probability of overestimating the number of signals.

5.2 Asymptotic results in some special cases

In this section, we analyse the TSS method from three different angles. We start by considering its behaviour under the perfect separation assumption. In this case, the theoretical results we derive for the TSS procedure rely on the results obtained in Duval et al. (2007). In Duval et al. (2007), a method similar to the TSS is proposed for the problem of multiple testing and the theoretical results are obtained under the perfect separation assumption. In general, the perfect separation will not hold, and furthermore, the initial motivation for the TSS procedure was to use it when the signal is weak, in order to estimate a larger group of values as signal and increase the number of correctly identified signals. For this reason, we do not make assumptions on the signal strength in the remainder of the section. First, we introduce the *oracle TSS* procedure, the unattainable procedure with similar "stopping time" as the regular TSS, that provides the lower bound on the norm of the remaining, undetected signal. Finally, we analyse the expected behaviour of the TSS procedure by adopting the Gaussian

mixture model. This enables us to describe the stopping time of the TSS procedure asymptotically, under no assumptions on the signal strength.

5.2.1 The existing literature and the perfect separation case

In an unpublished manuscript by Duval et al. (2007) a method similar to the TSS is proposed for the problem of multiple testing in a more general model than the squared Gaussian sequence model that we are considering. Their model includes the degrees of freedom parameter so that the distribution under the null is χ^2_η, while the distribution under the alternative is the sum of η scaled noncentral χ^2_1 distributions with possibly different noncentrality parameters. Similarly to the TSS, their procedure, which we refer to as the DDLR procedure, thresholds the cumulative sums of the smallest order statistics. Using our notation, their proposed stopping criterion is

$$\hat{k}^{\mathrm{DDLR}} = \max_i \left\{ \frac{1}{n-i+1} T_i > 1 \right\}.$$

This corresponds to our procedure with thresholds (5.8) with $H_i = 0$ for all i, that is

$$\lambda_i^{\mathrm{DDLR}} = \lambda_i^0 = n - i + 1.$$

We note that for large n it holds that $\lambda_i^{DDLR} \sim \lambda_i^{H_i}$ for any non-decreasing sequence H_i used. The theoretical results in Duval et al. (2007) only cover the case when the perfect separation assumption holds. A sufficient condition for the perfect separation assumption to hold is given therein (see Lemma 1 in Duval et al. (2007)), and under that condition it holds that the FDR and FNR of the DDLR procedure converge to zero. In Lemma 3 below, we provide an alternative statement guaranteeing the perfect separation between the signals and the non-signals. Instead of the extreme value distribution method used in Duval et al. (2007), we use the tail bounds for the

maximum of Gaussian variables. Therefore, we provide an alternative technique for proving, but attain a minimum sufficient rate for the signal strength that is larger than that in Duval et al. (2007), making our technique sub-optimal.

Lemma 3. *Let* $U_1, \ldots, U_{n-k} \overset{iid}{\sim} \chi^2_1$ *and* $V_1, \ldots, V_k \overset{iid}{\sim} \chi^2_1(\mu^2)$ *where* $\mu \geq \sqrt{2\log(n-k)} + \sqrt{2\log k}$. *If* $k(n) \to \infty$ *and* $n - k \to \infty$, *the probability of perfect separation between the signal and the non-signal values converges to 1:*

$$P\left(\min_{j=1,\ldots k} V_j \geq \max_{j=1,\ldots,n-k} U_j\right) \xrightarrow{n\to\infty} 1. \tag{5.11}$$

In Theorem 2 below, we generalise the results from Duval et al. (2007), and prove the FDR control for the TSS method allowing different thresholding sequences $\lambda_i^{H_i}$. The proof is outlined below, however as it closely follows the one in Duval et al. (2007) some of the steps where the existing results are directly used are skipped. One of the assumptions of the theorem is the perfect separation, for which either our result given in Lemma 3 or Lemma 1 from Duval et al. (2007) can be used. The proof of the FDR and FNR control follows from proving that the proportion is asymptotically correctly estimated under the perfect separation assumption.

Theorem 2. *If the perfect separation assumption holds,* $k = \pi_1 n$ *for some* $\pi_1 \in (0,1)$, u_n *is such that* $u_n \to 0$ *and* $\sqrt{n}u_n \to \infty$, *and* $\lambda_i^{H_i}$ *is defined by (5.8), where* $H_i = H_i(n)$ *are such that* $H_i(n)/\sqrt{nu_n} \to 0$, *it holds that*

$$P\left(\left|\frac{\hat{k}}{n} - \pi_1\right| \geq u_n\right) \to 0, \quad n \to \infty, \tag{5.12}$$

where $\hat{k}$ *is the number of signals estimated by the TSS procedure. Furthermore, the FDR and the FNR of the TSS procedure go to zero.*

5.2.2 Oracle TSS

To get more insight into where the TSS procedure stops when the signal is weak we consider an alternative tail-summed sequence. Define the signal-first permutation of Y's, where the sorted signal values come before the sorted non-signal values:

$$Y^S_{(k)}, Y^S_{(k-1)}, \ldots, Y^S_{(1)}, Y^{NS}_{(n-k)}, Y^{NS}_{(n-k-1)}, \ldots, Y^{NS}_{(1)}. \tag{5.13}$$

If the perfect separation does not hold, the sequence in (5.13) is not decreasing, and this permutation is unknown to us. By tail-summing the terms of this sequence and finding the first point when it drops below the threshold we get a new thresholding rule – *the oracle TSS*. We introduce the notation below. Define the oracle TSS sequence:

$$T^O_i = \begin{cases} \sum_{j=1}^{n-i+1} Y^{NS}_{(j)}, & i \leq n-k \\ \sum_{j=1}^{n-k} Y^{NS}_{(j)} + \sum_{j=1}^{k-i+1} Y^S_{(j)}, & i > n-k, \end{cases}$$

The oracle TSS procedure selects the top $\hat{k}^O$ values, where

$$\hat{k}^O = \min\{i : T^O_i < \lambda_i\} - 1.$$

Let $\hat{k}$ be the number of rejections made with the regular TSS procedure and $\hat{k}^O$ the number of rejections made by using the oracle TSS. It holds that

$$\hat{k} \leq \hat{k}^O,$$

since at each step of the TSS procedure we exclude the largest of the remaining values. This means that $T^O_i \geq T_i$, for all i, and that $\max\{i : T^O_i - T_i\} = k+1$. In Figure 5.2 we see the regular TSS sequence, the oracle TSS sequence, and a sequence of

thresholds given by (5.8) with $H = 2$. Empirically, we observe that the threshold sequence "intersects" the regular and the oracle tail-summed sequences at points that are close by, thus yielding very similar thresholding rules. That is, in addition to giving the upper bound for the number of rejections made by the regular TSS procedure, the oracle TSS seems to approximate this number of rejections. This is shown in Figure 5.2 as the stopping location is approximately the same whether we use the regular or the oracle TSS. Additionally, in Figure 5.3, we provide simulations showing the average (over $N = 100$ repetitions) scaled distance between the oracle and the regular TSS stopping times as a function of sample size n, and for different values of parameters ε and μ.

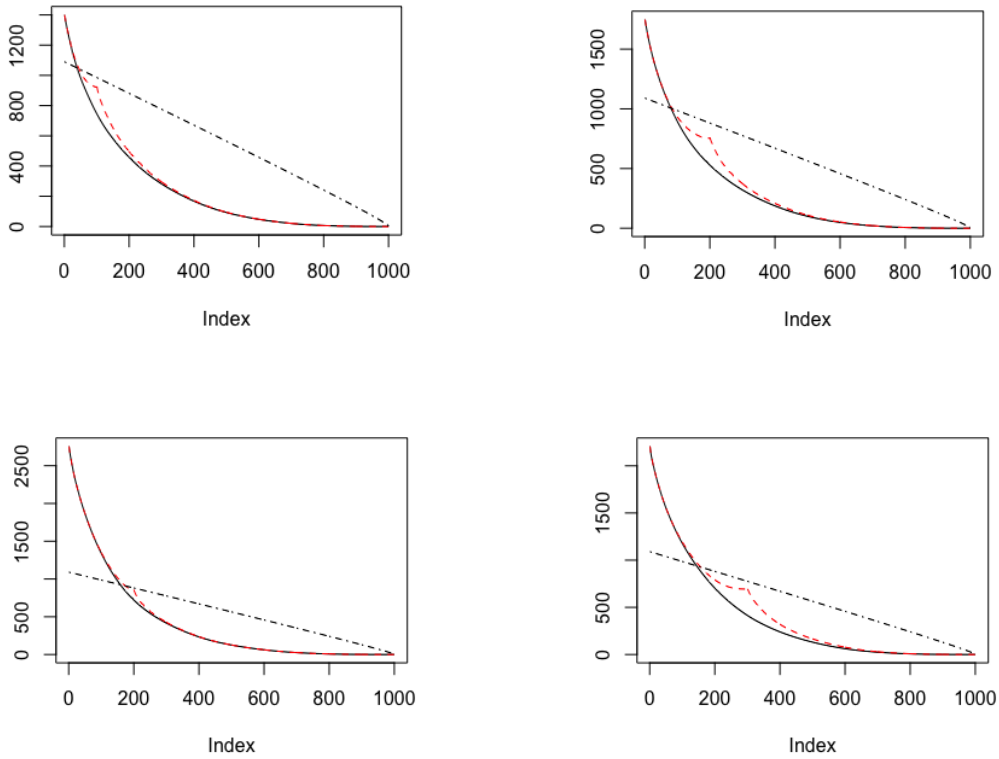
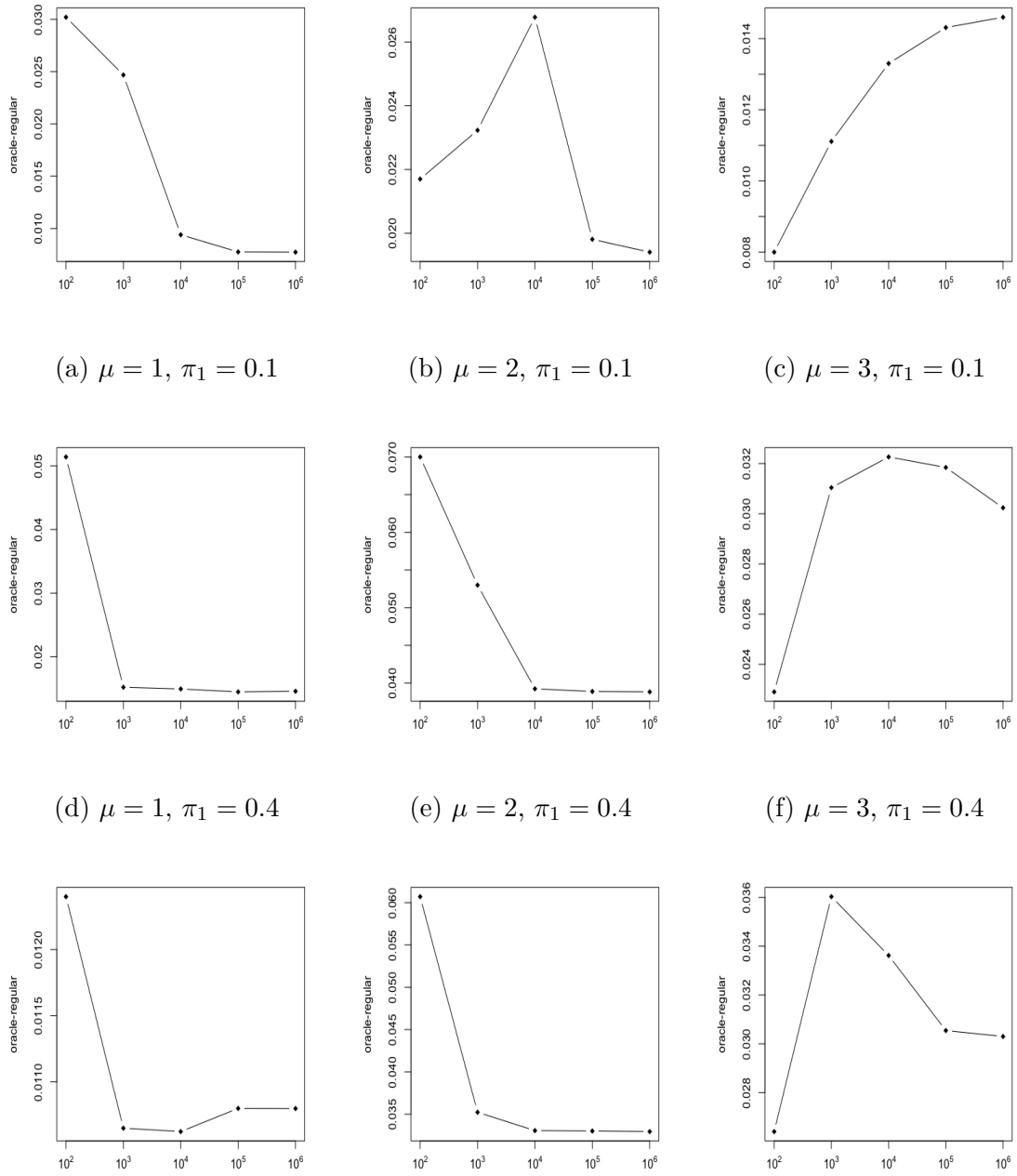


Fig. 5.2 The illustration of the regular and the oracle Tail-summed scores method, for different values of μ and k. Dash-dotted line is the sequence of thresholds $\lambda_i^{H_i}$ with $H_i = 2$. Black line - the regular TSS sequence. Red dashed line - the oracle TSS sequence. The values of the parameters and the estimated number of signal values by the TSS and by the oracle TSS are given as follows. Top left: $k = 100, \mu = 2$, TSS: 39, OTSS: 45, Top right $k = 200, \mu = 2$, TSS: 80, OTSS: 83, Bottom left: $k = 200, \mu = 3$, TSS: 143, OTSS: 156, Bottom right: $k = 300, \mu = 2$ TSS: 158, OTSS: 173



(a) $\mu = 1$, $\pi_1 = 0.1$ (b) $\mu = 2$, $\pi_1 = 0.1$ (c) $\mu = 3$, $\pi_1 = 0.1$

(d) $\mu = 1$, $\pi_1 = 0.4$ (e) $\mu = 2$, $\pi_1 = 0.4$ (f) $\mu = 3$, $\pi_1 = 0.4$

(g) $\mu = 1$, $\pi_1 = 0.7$ (h) $\mu = 2$, $\pi_1 = 0.7$ (i) $\mu = 3$, $\pi_1 = 0.7$

Fig. 5.3 Scaled difference of the oracle and the regular TSS procedure stopping times $(\hat{k}^O - \hat{k})/n$ are given for different values of μ and π_1 (the exact proportion of the signal values), and sample sizes $n \in \{10^2, 10^3, 10^4, 10^5, 10^6\}$ on the x-axis. The thresholds used are the asymptotic thresholds $\lambda_i^{H_i}$ with $H_i = 0$.

The signal-first sorting is infeasible in practice but we introduce it here to gain more insights on the behaviour of the regular TSS procedure, as the stopping times of the regular TSS and the oracle TSS are close, regardless of the number of signals k and the signal strength μ. The theoretical analysis is easier for the oracle TSS as at each step we know the distribution of the terms in the remaining sequence. The following theorem exploits this and proves that the oracle TSS procedure asymptotically stops no earlier than at the location for which it holds that the mean of the remaining signal is arbitrarily close to 1. As the TSS procedure stops earlier than the oracle procedure, and misses more signal values, it follows that the mean of the remaining signal, missed by the TSS procedure is larger than 1.

In the following theorem we consider the "asymptotic thresholds", that is (5.8) with $H_i = 0$, but the results can be generalised to other sequences H_i similarly as in Theorem 2.

Theorem 3. *For any $\varepsilon > 0$ define*

$$\tilde{y}_O(\varepsilon) = \min\left\{y : \int_0^y Q_{\chi_1^2(\mu^2)}(x)dx = \frac{1}{\pi_1}y(1+\varepsilon)\right\},$$

The quantity $\tilde{y}_O(\varepsilon)$ is asymptotically the proportion of the smallest signal values necessary to include so that their mean is at least $1+\varepsilon$. The oracle TSS procedure with thresholds given by (5.8) with $H_i = 0$ stops at the location $\hat{k}^O$ for which it holds that

$$P\left(\frac{\hat{k}^O}{n} - \frac{k}{n} < -\tilde{y}_O(\varepsilon)\right) \to 0, \quad n \to \infty.$$

meaning that asymptotically the mean of the undetected signal is no significantly larger than 1.

5.2.3 In-mean behaviour of the TSS

In this section, we adopt an alternative point of view to the problem, and instead of the Gaussian sequence model in (5.1), we consider the 2-point Gaussian mixture model for the sample:

$$X_i \sim \pi_1 N(\mu, 1) + \pi_0 N(0, 1), \quad i = 1, \ldots, n. \tag{5.14}$$

The difference between the two models is that in the Gaussian sequence model the number of signals is fixed and the nonzero means can be different, while in the Gaussian mixture model the number of signal variables is random, with binomial distribution, and the nonzero means are all equal. The Gaussian mixture model arises from the Gaussian sequence model when a sparse prior is assumed for the mean value of the vector $\boldsymbol{\mu} = (\mu, \ldots, \mu)$ with all equal components:

$$\mu \sim \pi_1 \delta_{\mu_0} + \pi_0 \delta_0,$$

where δ_{μ_0} is a Dirac measure centred at μ_0. The marginal distribution of X_i is then the mixture in (5.14) with $\mu = \mu_0$.

In the Gaussian mixture model all values in the sample come from the same mixture distribution, so by analysing the TSS procedure in this model, we avoid the problem of the unknown distributions in the TSS sequence. As earlier, let $Y_i = X_i^2$, $i = 1, \ldots, n$, where X_i is given in (5.14). The distribution of Y_i is

$$Y_i \sim \pi_0 \chi_1^2 + \pi_1 \chi_1^2(\mu^2), \quad i = 1, \ldots, n, \tag{5.15}$$

and for a given i, T_i is the sum of the smallest $n - i + 1$ order statistics from the mixture distribution (5.15). The in-mean point of view discussed below, analyses the asymptotic behaviour of the TSS procedure when $n \to \infty$. We start by motivating

the asymptotic behaviour, then describe the Lorenz curve interpretation of the TSS sequence, and finally using the convergence theorem for empirical Lorenz curves by Goldie (1977), we find the limiting value for the proportion of rejections made by the TSS procedure.

Let F, Q, F_n, Q_n be, in this order, the CDF, quantile function, empirical CDF, and empirical quantile function of the distribution in (5.15). We recall the definition of the empirical quantile function, defined as the left continuous inverse of the empirical distribution function, that is $Q_n(0) = 0$ and $Q_n(q) = Y_{(i)}$, for $q \in (\frac{i-1}{n}, \frac{i}{n}]$, $i = 1, \ldots, n$. For each i, T_i can be written as

$$\begin{aligned} T_i &= \sum_{k=1}^{n} Y_k \mathbb{I}\{Y_k \leq Y_{(n-i+1)}\} \\ &= \sum_{k=1}^{n} Y_k \mathbb{I}\{Y_k \leq Q_n(1-y)\}, \end{aligned} \tag{5.16}$$

for any y for which $1 - y \in (\frac{n-i}{n}, \frac{n-i+1}{n}]$, or equivalently $y \in [\frac{i-1}{n}, \frac{i}{n})$. This motivates the definition of a stochastic process version of the TSS sequence where $y \in [0, 1]$ is the continuous argument. To clarify, we first adjust the notation by scaling the arguments $i = 1, \ldots, n$ of the sequence T to interval $[0, 1]$ and define

$$T_n(i/n) := T_i, \quad i = 1, \ldots, n. \tag{5.17}$$

We extend the definition of T_n to all points $y \in [0, 1]$ and define the stochastic process version of the TSS sequence as:

$$T_n(y) = \sum_{k=1}^{n} Y_k \mathbb{I}\{Y_k \leq Q_n(1-y)\}, \quad y \in [0, 1].$$

Thus, using the empirical quantile function instead of the order statistics in (5.16) allows us to consider the sequence of statistics T_i as a stochastic process approximating

the ideal function which we get by substituting the sample mean with the expectation and Q_n with Q as follows:

$$\begin{aligned} T_n(y) &= \sum_{k=1}^{n} Y_k \mathbb{I}\{Y_k \leq Q_n(1-y)\} \\ &\approx \sum_{k=1}^{n} Y_k \mathbb{I}\{Y_k \leq Q(1-y)\} \\ &\approx n\mathbb{E}(Y\mathbb{I}\{Y \leq Q(1-y)\}). \end{aligned} \tag{5.18}$$

This approximation can be formalised using the consistency results for empirical Lorenz curves found in Goldie (1977).The expectation on the RHS is related to the Lorenz curve of a distribution with quantile function Q. By substitution $z = F(x)$, $dx = \frac{dz}{f(F^{-1}(z))}$ in (5.18) we get

$$\begin{aligned} \mathbb{E}(Y\mathbb{I}\{Y \leq Q(1-y)\}) &= \int_0^{Q(1-y)} x f(x) dx \\ &= \int_0^{1-y} Q(z) dz. \end{aligned}$$

Lorenz curve of a distribution with density f, quantile function Q and mean m is defined as

$$L(x) = \frac{\int_{-\infty}^{Q(x)} t f(t) dt}{m} = \frac{1}{m} \int_0^x Q(t) dt, \quad x \in (0,1).$$

It follows that

$$\mathbb{E}(Y\mathbb{I}\{Y \leq Q(1-y)\}) = mL(1-y). \tag{5.19}$$

Similarly, the TSS process $T_n(y)$ is related to the empirical Lorenz curve $\hat{L}_n$, that can be expressed in terms of the TSS process as

$$T_n(y) = T_n(0)\hat{L}_n(1-y). \tag{5.20}$$

Using the consistency result for the convergence of empirical Lorenz curves from Goldie (1977) (see Theorem 1 therein), and the law of large numbers gives

$$\sup_{y\in[0,1]}\left|\frac{1}{n}T_n(y)-(1+\pi_1\mu^2)L(1-y)\right|\overset{a.s.}{\to}0 \tag{5.21}$$

The proof of this statement is included in the proof of Theorem 4 stated below. To describe the stopping time of the TSS procedure, we also consider the continuous argument for the thresholds. Starting from the thresholding sequence $\lambda_i^{H_i}$ we define the thresholding curve $\lambda_n(y)$:

$$\begin{aligned}\lambda_i^{H_i} &= n-i+1+H_i\sqrt{2(n-i+1)}, \quad i=1,\dots,n\\ &= n(1-i/n)+1/n+H_i\sqrt{2(n(1-i/n)+1)}\\ &\approx n(1-y)+1+H_i\sqrt{2(n(1-y)+1)}, \quad y\in[0,1]\\ &=: \lambda_n(y).\end{aligned} \tag{5.22}$$

Remark 3. Note that as $n\to\infty$, $\lambda_n(y)/n\to 1-y$, so the choice of constants H_i used for thresholds $\lambda_i^{H_i}$ in (5.8) do not affect the behaviour of the procedure asymptotically. The same holds if H_i's depend on n but are of order smaller than $\sqrt{n}$.

Thresholding the TSS process $T_n(y)$ using the thresholding curve $\lambda_n(y)$, yields a stopping time $\hat{y}_n$:

$$\hat{y}_n=\begin{cases}0, & \text{if } T_n(y)<\lambda_n(y) \text{ for } y\in(0,1)\\ \max\{y>0: T_n(y)\geq\lambda_n(y)\}, & \text{otherwise.}\end{cases} \tag{5.23}$$

Note that

$$\left|\frac{\hat{k}}{n} - \hat{y}_n\right| \leq 1/n. \tag{5.24}$$

(5.24) holds as $T_n(i/n) = T_i$, $\lambda_n(i/n) = \lambda_i$, $T_n(y)$ is a piecewise decreasing function and $\lambda_n(y)$ is a continuous decreasing functions. The approximations in (5.18) and (5.22) suggest that the corresponding stopping times should also be close, that is

$$\hat{y}_n \approx \tilde{y},$$

where

$$\begin{aligned}\tilde{y} &:= \max\{y : \mathbb{E}(Y\mathbb{I}\{Y \leq Q(1-y)\}) \geq 1-y\} \\ &= \max\left\{y \in [0,1] : L(1-y) \geq \frac{1-y}{1+\pi_1\mu^2}\right\}.\end{aligned} \tag{5.25}$$

As $L(1) = 1 > 1/(1+\pi_1\mu^2)$, the set above is non-empty and moreover it contains an interval around 0, so $\tilde{y}$ is well defined. In fact there is a unique point of "intersection", as L is a convex function, $L(0) = 0$, and it holds that:

$$L(1-\tilde{y}) = \frac{1-\tilde{y}}{1+\pi_1\mu^2}.$$

(5.25) implies that asymptotically the TSS procedure rejects the null hypothesis of no signal for the top $\tilde{y}100\%$ percent of the sample, and does not reject the null for the rest. It makes $Q(1-\tilde{y})$ our in-mean threshold.

The closeness between the sample and the asymptotic stopping time is formalised in the following Theorem 4.

Theorem 4. *Let $\hat{y}_n$ and $\tilde{y}$ be defined as in (5.23) and (5.25). It holds that*

$$\hat{y}_n \overset{a.s.}{\to} \tilde{y}, \quad n \to \infty.$$

5.3 Theoretical results in the general case

In this section we state some additional theoretical results in the general case, that is when no assumption is made on the signal strength. Note that Theorem 4 from the previous section is one such result.

The following lemma proves the contiguity of the TSS procedure, in the sense that there is only one intersection between the sequence of tail-summed statistics and the sequence of thresholds $\lambda_i^{H_i}$.

Proposition 1. *Let $\lambda_i^{H_i}$ be defined as in (5.8) where H_i is a non-decreasing sequence of real numbers. For all $i = 1, \ldots, n-1$ it holds that*

$$T_i < \lambda_i \implies T_{i+1} < \lambda_{i+1}.$$

The following proposition proves that the probability of the TSS procedure overestimating the number of signals is small. This statement is related to the FDR and FWER control only when the perfect separation assumption holds, which was discussed in Section 5.1.

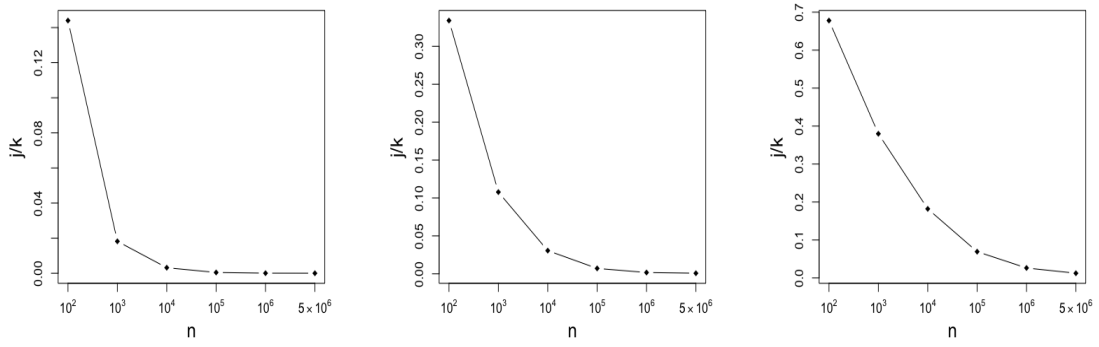
Proposition 2. *The probability of the TSS method overestimating the number of signals k is upper bounded by $1 - F_{\chi^2_{n-k}}(\lambda_{k+1})$, where λ_i, $i = 1, \ldots, n$ is the sequence of thresholds.*

The following theorem proves that the probability of the TSS procedure stopping before the mixing starts, and making zero false discoveries goes to zero as $n \to \infty$. In the proof we use some existing results for the asymptotic behaviour of trimmed sums. The review of the topic of trimmed sums can be found in Hahn et al. (1991). We again consider only the "asymptotic thresholds", that is (5.8) with $H_i = 0$, as the results can be generalised to other sequences H_i similarly as in Theorem 2.

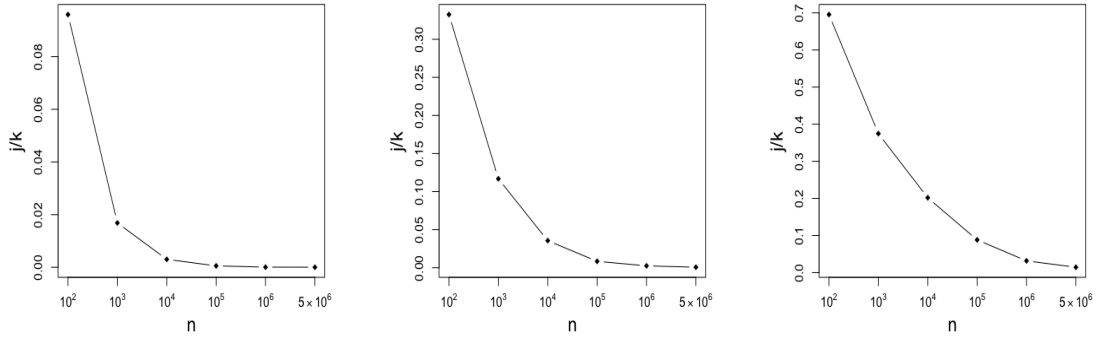
Theorem 5. *Let $\hat{k}_n$ be the estimated number of signals using the TSS procedure with thresholds given by (5.8) with $H_i = 0$. Let $k = \pi_1 n$ be the true number of signals, j the position of the largest non-signal variable in the decreasingly sorted sample, and assume that $j/k \to 0$, as $n \to \infty$. It holds that*

$$P\left(\hat{k}_n \leq j\right) \to 0, \quad n \to \infty.$$

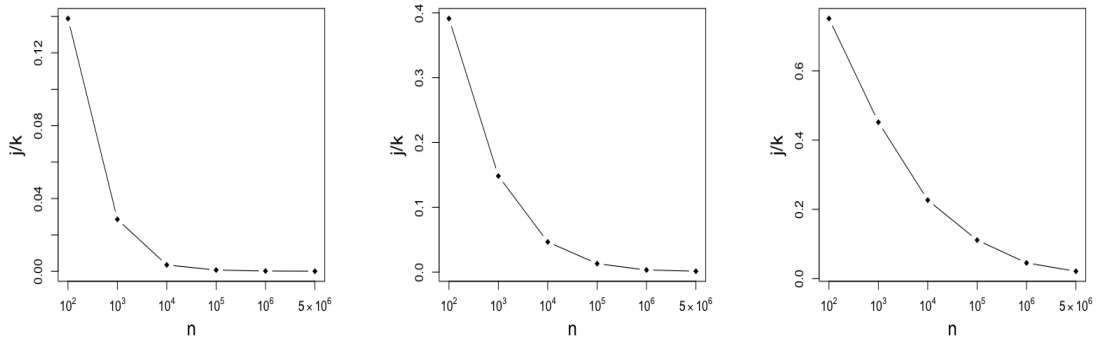
Remark 4. One of the assumptions of Theorem 5 is that $j/k \to 0$. This assumption does not hold if there is a perfect separation and $k = \pi_1 n$. However, empirically, we find that if the signal is weaker and does not increase with sample, this assumption is justified for a range of parameter values. This is illustrated in Figure 5.4 where it can be seen that the values j/k seem to approach zero when $n \to \infty$ for various values of $k = \pi_1 n$ and μ.



(a) $\mu = 1$, $\pi_1 = 0.1$ (b) $\mu = 2$, $\pi_1 = 0.1$ (c) $\mu = 3$, $\pi_1 = 0.1$



(d) $\mu = 1$, $\pi_1 = 0.4$ (e) $\mu = 2$, $\pi_1 = 0.4$ (f) $\mu = 3$, $\pi_1 = 0.4$



(g) $\mu = 1$, $\pi_1 = 0.7$ (h) $\mu = 2$, $\pi_1 = 0.7$ (i) $\mu = 3$, $\pi_1 = 0.7$

Fig. 5.4 For different values of μ and π_1 (the exact proportion of the signal values), the average (over $N = 100$ repetitions) scaled position j/k of the first non-signal variable in the decreasingly sorted sample is shown for different values of the sample size n.

5.4 Simulations and applications

5.4.1 Simulations

The behaviour of the TSS procedure can be controlled by adjusting H_i in the sequence of thresholds $\lambda_i^{H_i}$. Smaller H_i values will lead to more rejections. By setting $H_i = 2$ we achieve that when there is no signal, under the global null, we keep the probability of making a false discovery at approximately 0.05. In the following, unless specified, we use $H_i = 2$.

We first investigate the behaviour of the TSS procedure from the multiple testing point of view. For all the simulation results below a sample from the Gaussian sequence model with a given (nonrandom) k and a single nonzero mean parameter μ is considered. In Figures 5.5 and 5.6, the estimated FDR for the TSS procedure for different configurations of μ and k and for thresholds $\lambda_i^{H_i}$ with $H_i = 2$ are shown.

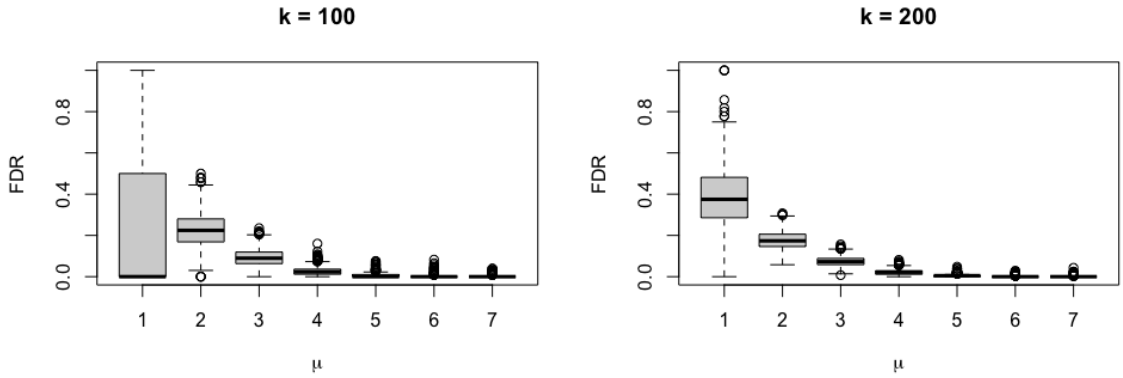


Fig. 5.5 Boxplots of the FDR values for the TSS method for different values of μ and k, where $H_i = 2$, for all i, based on the sample of size $n = 1000$ and $N = 1000$ repetitions.

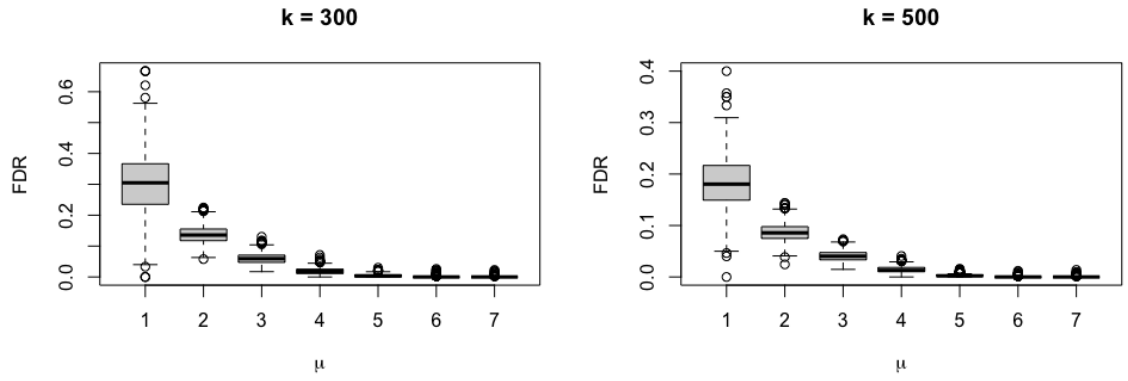


Fig. 5.6 Boxplots of the FDR values for the TSS method for different values of μ and k, where $H_i = 2$, for all i, based on the sample of size $n = 1000$ and $N = 1000$ repetitions.

For large μ, when the separation between the true-null and false-null values is better, the FDR of the TSS method is almost zero, which is in line with the statement of Theorem 2. On the other hand, for smaller μ values, the TSS does not control the FDR. When the signal is weak, mixing is greater but the TSS proceeds with detecting signals until the set of remaining values resemble the χ^2_1 sample. Note that a multiple testing procedure such as the FDR-controlling Benjamini-Hochberg, might stop later than the TSS in the strong signal case, but earlier than the TSS in the weak signal case. This holds because in the strong signal case, the TSS procedure with large enough H_i will still underestimate the number of signals, while the Benjamini-Hochberg procedure will make some controlled number of false rejections. However, if the signal is dense but weak, the TSS procedure will include more false rejections as the stopping time depends only on the strength of the remaining signal.

We can also investigate whether the TSS controls some other multiple testing error rates. The local FDR (lfdr) by Efron (2007), is defined in Section 2.1.2 as a posterior probability of a given value from the mixture distribution being from the

null distribution. In Figures 5.7 and 5.8, we see the estimated local FDR for the TSS threshold for different configurations of μ and k.

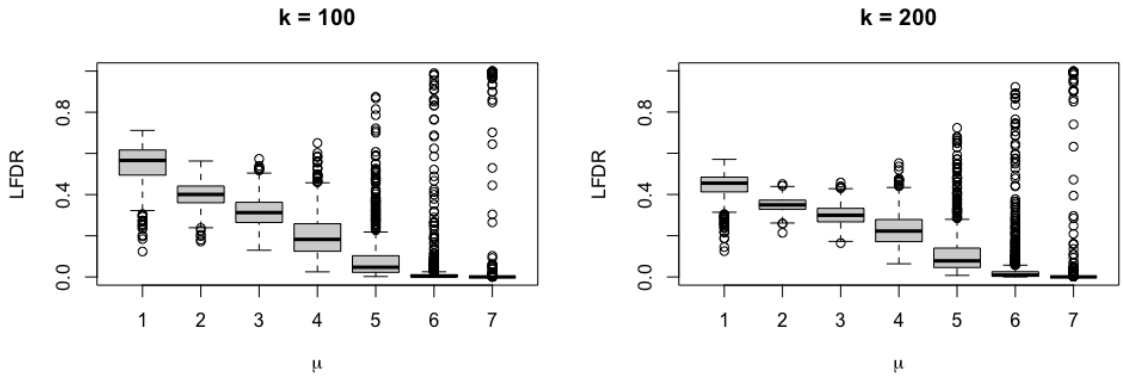


Fig. 5.7 Boxplots of the lfdr values for the TSS method for different values of μ and k, where $H_i = 2$, for all i, based on the sample of size $n = 1000$ and $N = 1000$ repetitions.

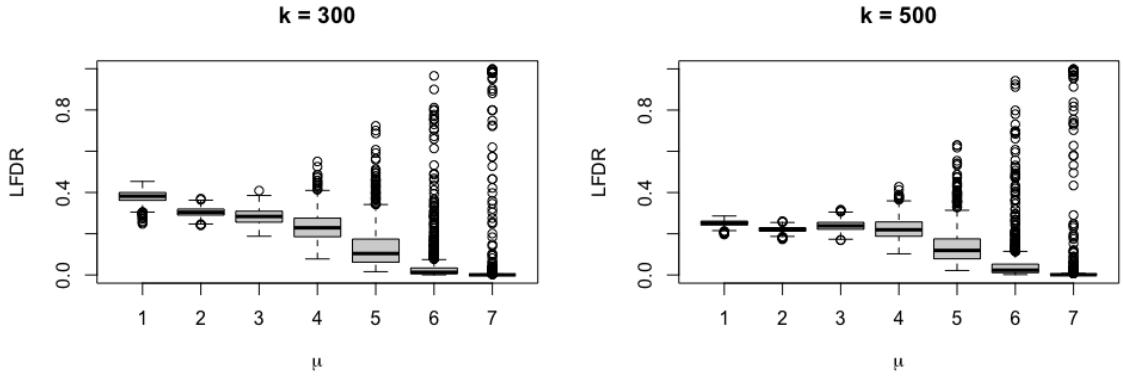


Fig. 5.8 Boxplots of the lfdr values for the TSS method for different values of μ and k, where $H_i = 2$, for all i, based on the sample of size $n = 1000$ and $N = 1000$ repetitions.

As k increases, the lfdr decreases, and as μ increases the lfdr also decreases on average. For large μ, the separation between null and alternative distribution is better, so although there are some large lfdr values, the FDR will still be small in those cases.

The simulations above confirm that the TSS method does not control the FDR or the lfdr when the signal is weak and the perfect separation condition does not hold. For that reason it cannot be compared to other multiple testing methods in that respect. We can also choose to view the TSS method as a signal estimation procedure and compare it to other thresholding signal estimation methods in the literature such as Donoho and Johnstone (1995), Johnstone and Silverman (2004) and Abramovich et al. (2006). The following methods used for comparison are introduced in Section 2.3:

1. UNI – universal threshold $t = \sqrt{2 \log n}$ by Donoho and Johnstone (1995)

2. FDR – the Benjamini-Hochberg FDR-controlling method at level q by Benjamini and Hochberg (1995) and Abramovich et al. (2006)

3. EBT – Empirical Bayes thresholding method by Johnstone and Silverman (2004)

4. SURE – thresholding estimator based on Stein's unbiased risk estimate, as considered in Donoho and Johnstone (1995)

The simulation results are shown in Table 5.1. The MSE of the TSS thresholding estimator is among the smallest when the signal is weak. Particularly, looser thresholds where $H_i = 0$ give better results for this purpose. In the case of strong signal, the performance of the TSS is not competitive, however, an alternative sequence of thresholds could lower down the MSE further.

	$\mu = 2$					$\mu = 3$				
Number nonzero	50	100	200	300	500	50	100	200	300	500
TSS $H = 2$	**206**	**377**	**662**	904	1300	275	434	689	888	1190
TSS $H = 0$	209	**367**	**634**	**859**	**1226**	**227**	374	609	798	1078
UNI	**203**	404	805	1206	2000	366	725	1433	2122	3391
FDR $q = 0.05$	314	383	785	1150	1795	313	519	836	1085	1460
FDR $q = 0.2$	227	421	673	900	1230	**226**	**350**	**527**	**659**	**850**
EBT	**203**	386	681	**874**	**944**	242	**361**	**494**	**585**	**749**
SURE	**203**	400	803	1206	2012	312	700	1443	2180	3640

Table 5.1 The estimated l_2 risk of different thresholding estimators based on $N = 1000$ repetitions for the sample of size $n = 1000$ from the Gaussian sequence model, with varying number of signals and for signal strength $\mu = 2$ and $\mu = 3$. The bold and underlined values correspond to the two smallest values in each column.

The simulations also show that, compared to the other thresholding procedures, the TSS behaves well if the goal is to estimate the quadratic functional of the Gaussian sequence, that is the l_2 norm of the signal

$$\|\mu\| = \sqrt{\sum_{i=1}^{n} \mu_i^2}.$$

The results are shown in Table 5.2, where among the other thresholding signal estimation procedures, the TSS has the smallest estimated risk, the risk being defined as

$$\mathbb{E}(\|\hat{\mu}\|^2 - \|\mu\|^2)^2.$$

The minimax estimator for this problem in the dense case, when $k > \sqrt{n}$, studied in Collier et al. (2017) for example, and defined as

$$\hat{\mu}_{\text{minimax}} = \sum_{i=1}^{n} X_i^2 - n,$$

outperforms the TSS. We note that the minimax estimator is a special case of the TSS sequence where a constant threshold is used:

$$\hat{k}_{\text{minimax}} = \max\{i : T_i \geq n\}.$$

This thresholding procedure is always more conservative than the TSS, as the threshold is larger than the sequence of thresholds that we propose. However, when the signal is weak, and the stopping time happens early on, it will hold that $\lambda_i^{H_i} \approx n$, and the behaviour of the minimax procedure and of the TSS is similar. This explains the good behaviour of the TSS for estimating the l_2 norm of the signal, which is true especially when the norm is smaller, as can be seen in Table 5.2.

Signal l_2 norm			30					50		
Number nonzero	50	100	200	300	500	50	100	200	300	500
TSS $H = 2$	9.34	7.33	7.41	8.10	10.3	15.3	12.4	19.7	36.0	75.3
UNI	15.4	198	474	570	620	11.9	14.0	524	1590	2900
FDR $q = 0.05$	6.56	31.2	243	460	674	13.8	28.2	23.6	32.3	2780
EBT	18.3	19.4	14.7	23.2	49.5	27.1	85.6	258	456	792
SURE	12.8	180	464	578	660	12.0	13.8	497	1640	3320
minimax	5.33	5.88	5.84	5.68	5.92	12.3	11.8	11.9	11.7	12.3

Table 5.2 The estimated risk (divided by 10^3) of different thresholding signal estimation procedures for estimating $\|\mu\|^2$ for given values of $\|\mu\|^2$ and k, where the sample size is $n = 1000$. The estimator in the last row is the minimax estimator of $\|\mu\|^2$, studied in Collier et al. (2017).

5.4.2 Applications in change-point inference

The initial motivation behind the TSS procedure was to use it for estimating the subset of coordinates with change in mean-plus-error panel data model. We now discuss this as a possible application. First we introduce the change-point model. Let

$$X_{ij} = \mu_{ij} + \varepsilon_{ij}, \quad i = 1, \ldots, p \quad j = 1, \ldots, n, \tag{5.26}$$

where $(\mu_{ij})_{j=1}^n$ is a piecewise constant signal for each coordinate $i = 1, ..., p$, and $(\varepsilon_{ij})_{j=1}^n$ is a mean-zero noise series. We say that a change-point is at location τ if $\mu_{i,\tau+1} \neq \mu_{i,\tau}$ for at least one coordinate i. Although it is enough for only one coordinate to change, we assume that a larger number of coordinates are affected by a change at the change-point location τ. Suppose that there is only one change point in the dataset X_{ij} at location τ and define the subset of coordinates with change as:

$$\mathcal{S} = \{j : \mu_{i,\tau+1} \neq \mu_{i,\tau}\} \subset \{1, ..., p\}.$$

Most of the papers studying model (5.26) consider the problem of testing for and estimating the unknown change-point location τ. A popular approach is via the CUSUM transformation. The CUSUM transformation of the data matrix $X \in \mathbb{R}^{p\times n} \mapsto Z \in \mathbb{R}^{p\times(n-1)}$ is

$$Z_{ij} = \sqrt{\frac{j(n-j)}{n}} \left(\frac{1}{n-j} \sum_{l=j+1}^{n} X_{il} - \frac{1}{j} \sum_{l=1}^{j} X_{il} \right), \quad i = 1, \ldots, p, \quad j = 1, \ldots, n-1.$$

For each row i, and each candidate change-point location j, this transformation compares the difference in sample means before and after it. The distribution of a term in the CUSUM matrix is

$$Z_{ij} \sim N(\theta_{ij}, 1),$$

where

$$\theta_{ij} = \sqrt{\frac{j(n-j)}{n}} \left(\frac{1}{n-j} \sum_{l=j+1}^{n} \mu_{il} - \frac{1}{j} \sum_{l=1}^{j} \mu_{il} \right), \quad i = 1, \ldots, p, \quad j = 1, \ldots, n-1.$$

Assume that the change-point location τ is known. The vector that is a column of the CUSUM matrix at location τ will have the largest absolute mean vector. Its

distribution is

$$Z_{\cdot,\tau} \sim N(\theta_{\cdot,\tau}, I_d), \text{ where } \theta_{\cdot,\tau} = \sqrt{\frac{\tau(n-\tau)}{n}}(\mu_{\cdot,\tau+1} - \mu_{\cdot,\tau}).$$

This separates the coordinates into two groups based on the distribution of $Z_{i,\tau}$:

$$Z_{i,\tau} \sim \begin{cases} N(0,1), & i \notin \mathcal{S} \\ N(\theta_{i,\tau}, 1), & i \in \mathcal{S}. \end{cases} \tag{5.27}$$

Thus, to estimate the subset of coordinates with change $\mathcal{S}$, based on vector $Z_{\cdot,\tau}$, we need a method for estimating the subset of non-zero coordinates in the Gaussian sequence model. This problem is largely unexplored in the change-point literature, but the concept of selecting an "influential" subset of coordinates has been investigated for example in Cho and Fryzlewicz (2015) and Cho (2016) as a way to reduce the effect of noisy coordinates on the change-point estimation. Assuming that the change-point location is shared between the coordinates, by selecting those coordinates believed to be affected by a change-point, we may improve the accuracy of the estimated change-point location. We mention two papers that address the problem of estimating $\mathcal{S}$. In Jirak (2015) a method for estimating the subset of coordinates with change is proposed, but the model considered does not assume shared change-point location, so the problem is considered coordinate-wise. The double CUSUM method by Cho (2016) estimates the shared change-point location by applying the CUSUM transformation twice, and as a byproduct also yields an estimate of $\mathcal{S}$. First, the CUSUM transformation is applied row-wise to reveal the likely jump locations in each row. Then, the values in each column of the absolute CUSUM matrix are sorted decreasingly before applying the CUSUM transformation for the second time, on the columns of that matrix. The idea behind applying the CUSUM transformation on the sorted columns is to find a

"jump" between the nonzero mean values of the coordinates with change and the zero mean values of the noisy coordinates. This second step provides an estimate of the set of coordinates that "contribute" to change, which is not claimed to be a consistent estimator of $\mathcal{S}$. Thus, the output of the double CUSUM method is both the estimated change-point location and the subset of influential coordinates. We now compare the estimated subsets of coordinates with change obtained by the double CUSUM and by the TSS method by calculating the estimated number of affected coordinates and the false discovery rate of the estimation. The TSS algorithm is applied using the change-point location estimated by the double CUSUM algorithm. That is, the TSS method is applied on the column of the CUSUM-transformed matrix where the double CUSUM algorithm had estimated the change-point location.

The results in Table 5.3 show that the TSS method selects a smaller number of coordinates compared to the double CUSUM method. The FDR of the estimated subset is significantly smaller than when using the double CUSUM method, so the TSS selects a smaller number of noisy coordinates. This motivates the question: Can we improve the accuracy of the change-point estimation by implementing the TSS procedure into a change-point estimation algorithm? This idea also extends to using other multiple testing methods to minimise the number of noisy coordinates selected and increase the precision of the estimated change-point location.

	$\|\mu\|_2 = 1.5$	$\|\mu\|_2 = 2$	$\|\mu\|_2 = 3$	$\|\mu\|_2 = 4$
		$k = 30$		
TSS	18 (0.083)	25 (0.030)	28 (0.005)	29 (0.005)
DC	194 (0.845)	136 (0.763)	31 (0.030)	30 (0.001)
		$k = 50$		
TSS	25 (0.133)	36 (0.067)	45 (0.008)	48 (0.002)
DC	205 (0.762)	160 (0.684)	58 (0.134)	50 (0.010)
		$k = 100$		
TSS	31 (0.164)	53 (0.096)	82 (0.033)	93 (0.009)
DC	224 (0.605)	202 (0.533)	135 (0.274)	106 (0.074)
		$k = 150$		
TSS	34 (0.148)	62 (0.101)	110 (0.050)	133 (0.020)
DC	234 (0.485)	225 (0.417)	191 (0.255)	161 (0.107)
		$k = 200$		
TSS	36 (0.132)	68 (0.095)	131 (0.028)	170 (0.086)
DC	242 (0.389)	240 (0.321)	229 (0.207)	214 (0.110)
		$k = 300$		
TSS	37 (0.086)	76 (0.063)	162 (0.040)	230 (0.026)
DC	247 (0.230)	255 (0.180)	276 (0.115)	288 (0.079)

Table 5.3 The average number of coordinates in the estimated signal set, an the false discovery rate (in parentheses), of the TSS and the double CUSUM (DC) procedure based on $N = 200$ repetitions. The parameter values are: $n = p = 500$, $\tau = 200$, and varying values of k - the number of true signals, and $\|\mu\|_2$ - the l_2 norm of the mean vector.

5.5 Discussion

The main idea of the TSS procedure is considering values in groups in order to detect more signals when the signal is weak. Naturally, the theory can be extended by assuming different distributions under the alternative – heavy tailed distributions under the alternative, for example, would make the problem easier. Another possibility is to consider different transformations of the sample before aggregating the values. We have used the squared values of the sample to exploit the additive property of the chi-square distribution, however simiar results can be obtained using other transformations and the central limit theorem. The sequence of thresholds used can also be manipulated to suit the goal of the analysis. The effect of more conservative thresholds than the one proposed here can be investigated if the aim is to control the FDR. In contrast, weaker thresholds might be of interest for signal estimation purposes. Additionally, the performance of the TSS can be examined for different problems, such as the problem of estimating the l_2 norm of the signal, which is implied in Section 5.4.

5.6 Proofs

5.6.1 Notation

Let $\mathcal{M}(n,k)$ be the collection of all subsets of $\{1,\ldots,n\}$ of cardinality k and $\mathcal{M}$ is the collection of all possible subsets of $\{1,\ldots,n\}$. $\prod_m v$ is the projection of a vector $v \in \mathbb{R}$ onto a subspace indexed by $m \in \mathcal{M}$. $\|\cdot\|$ is the l_2 norm. Let Z_i, $i = 1,...,n$ be the sequence of independent random variables where $k \leq n$ of them have $N(\mu,1)$ and $n-k$ have $N(0,1)$ distribution. Let $\mathcal{S} = \{i : Z_i \sim N(\mu,1)\}$ and $S^{\complement} = \{i : Z_i \sim N(0,1)\}$ and $U = (U_1,...,U_{n-k})$ be Z_i^2, $i \in \mathcal{S}^{\complement}$ and $V = (V_1,...,V_k)$ are Z_i^2, where $i \in \mathcal{S}$ (order not important).

5.6.2 Main results

Proposition 1. *Let $\lambda_i^{H_i}$ be defined as in (5.8) where H_i is a non-decreasing sequence of real numbers. For all $i = 1, \ldots, n-1$ it holds that*

$$T_i < \lambda_i \implies T_{i+1} < \lambda_{i+1}.$$

Proof. It holds that

$$\begin{aligned} T_{k+1} \geq \lambda_{k+1} &\implies X_{(1)} \geq \frac{\lambda_{k+1}}{n-k} \vee \cdots \vee X_{(n-k)} \geq \frac{\lambda_{k+1}}{n-k} \\ &\implies X_{(n-k)} \geq \frac{\lambda_{k+1}}{n-k} \\ &\implies X_{(n-k+1)} \geq \frac{\lambda_{k+1}}{n-k}. \end{aligned}$$

Since $T_k = T_{k+1} + X_{(n-k+1)}$ we have

$$\begin{aligned} T_{k+1} \geq \lambda_{k+1} &\iff T_{k+1} \geq \lambda_{k+1} \wedge X_{(n-k+1)} \geq \frac{\lambda_{k+1}}{n-k} \\ &\implies T_k \geq \lambda_{k+1}\left(1 + \frac{1}{n-k}\right). \end{aligned}$$

Now, it is enough to show that

$$\lambda_{k+1}\left(1 + \frac{1}{n-k}\right) \geq \lambda_k,$$

which is equivalent to

$$\begin{aligned} & H_k\sqrt{2(n-k)} + H_{k+1}\sqrt{\frac{2}{n-k}} \geq H_k\sqrt{2(n-k+1)} \\ \iff & H_{k+1}\left(\sqrt{n-k} + \frac{1}{\sqrt{n-k}}\right) \geq H_k\sqrt{n-k+1}. \end{aligned} \tag{5.28}$$

The inequality (5.28) holds since $\frac{\sqrt{n-k+1}}{\sqrt{n-k}+\frac{1}{\sqrt{n-k}}} = \sqrt{\frac{n-k}{n-k+1}} < 1$ and $H_{k+1} \geq H_k$ since H_k is non-decreasing. □

Remark 5. Note that (5.28) is a necessary and a sufficient condition for the contiguity to hold, and H_k being non-decreasing is a sufficient condition. As the contiguity depends on the ratios between consecutive H_k's, and they should be non-decreasing, there is no largest or smallest sequence of thresholds for which it holds. However, if we fix $H_1 \geq 0$, using the minimum condition from (5.28), we get the minimum allowed sequence of H_k's

$$\begin{aligned} H_k &= H_1 \prod_{i=1}^{k-1} \sqrt{\frac{n-k}{n-k+1}} \\ &= H_1 \sqrt{\frac{n}{n-k+1}}, \end{aligned}$$

giving the minimum sequence of thresholds to be

$$\lambda_k = n - k + 1 + H_1\sqrt{2n}.$$

We could also fix $H_1 < 0$ and get the maximum allowed thresholds. Note that simply having the contiguity property, does not mean reasonable thresholds, as we also want to be able to reject the hypothesis at an appropriate step, rather than just proceed accepting.

Lemma 3. *Let* $U_1, \ldots, U_{n-k} \overset{iid}{\sim} \chi^2_1$ *and* $V_1, \ldots, V_k \overset{iid}{\sim} \chi^2_1(\mu^2)$ *where* $\mu \geq \sqrt{2\log(n-k)} + \sqrt{2\log k}$. *If* $k(n) \to \infty$ *and* $n - k \to \infty$, *the probability of perfect separation between the signal and the non-signal values converges to 1:*

$$P\left(\min_{j=1,\ldots k} V_j \geq \max_{j=1,\ldots,n-k} U_j\right) \xrightarrow{n\to\infty} 1. \tag{5.11}$$

Proof. The proof follows from the proof of Theorem 4 in Jeng (2016), where a sufficient condition for the separation is given for Gaussian variables $N(0,1)$ and $N(\mu,1)$, $\mu > 0$. As expected, the sufficient rates for the signal strength stay the same as therein. We prove (5.11) by finding $a(n)$ such that the following holds

$$\begin{aligned} & P\left(\min_{j=1,\dots k} V_j < \max_{j=1,\dots,n-k} U_j\right) \\ & \leq P\left(\min_{j=1,\dots k} V_j \leq a(n)\right) + P\left(\max_{j=1,\dots,n-k} U_j > a(n)\right) \\ & \xrightarrow{n\to\infty} 0. \end{aligned} \tag{5.29}$$

For the second term in (5.29), for $a(n) = 2\log(n-k)$ it holds that

$$P(\max_{j=1,\dots,n-k} U_j > a(n)) \to 0, \quad n \to \infty. \tag{5.30}$$

This holds since

$$\begin{aligned} P(\max_{j=1,\dots,n-k} U_j > a(n)) &= P(\max_{j=1,\dots,n-k} |Z_j| > \sqrt{a(n)}) \\ &= 2P(\max_{j=1,\dots,n-k} Z_j > \sqrt{a(n)}) \\ &\leq 2(n-k)\frac{1}{\sqrt{a(n)}\sqrt{2\pi}}\exp\{-a(n)/2\} \quad = \frac{C}{\sqrt{2\log(n-k)}}. \end{aligned}$$

The last inequality follows from the union bound and the upper bound for the tail of Gaussian distribution $1 - \Phi(t) \leq \frac{e^{-t^2/2}}{t\sqrt{2\pi}}$. For the first term in (5.29) we have

$$\begin{aligned} P\left(\min_{j=1,\dots,k} V_j \leq a(n)\right) &= P\left(\min_{j=1,\dots,k} |Z_j + \mu| \leq \sqrt{a(n)}\right) \\ &\leq P\left(\min_{j=1,\dots,k} Z_j + \mu \leq \sqrt{a(n)}, \min_{j=1,\dots,k} Z_j \geq -\mu\right) + P(\min_{j=1,\dots,k} Z_j < -\mu) \\ &\leq P\left(\min_{j=1,\dots,k} Z_j + \mu < \sqrt{a(n)}\right) + P(\min_{j=1,\dots,k} Z_j < -\mu). \end{aligned} \tag{5.31}$$

For the second term in (5.31), using symmetry property, the arguments for the right tail of the maximum of Gaussians, and $\mu > \sqrt{2\log(k)}$ it holds that

$$P(\min_{j=1,\dots,k} Z_j < -\mu) \to 0, \quad n \to \infty.$$

The first term in (5.31) goes to zero as $\mu \geq \sqrt{2\log(n-k)} + \sqrt{2\log k}$. As both terms go to zero we have

$$P\left(\min_{j=1,\dots,k} V_j \leq a(n)\right) \to 0 \tag{5.32}$$

Combining (5.32) and (5.30) completes the proof. □

Proposition 2. *The probability of the TSS method overestimating the number of signals k is upper bounded by $1 - F_{\chi^2_{n-k}}(\lambda_{k+1})$, where λ_i, $i = 1, \dots, n$ is the sequence of thresholds.*

Proof.

$$\begin{aligned}
P\left(\hat{k} > k\right) &= P\left(T_{k+1} > \lambda_{k+1}\right) \\
&= P\left(\sum_{j=1}^{n-k} Y_{(j)} > \lambda_{k+1}\right) \\
&= P\left(\min_{m \in \mathcal{M}(n,n-k)} \left\|\prod_m X\right\|^2 > \lambda_{k+1}\right) \\
&\leq P\left(\sum_{i \in S^{\complement}} X_i > \lambda_{k+1}\right) \\
&= 1 - F_{\chi^2_{n-k}}(\lambda_{k+1})
\end{aligned}$$

□

Theorem 2. *If the perfect separation assumption holds, $k = \pi_1 n$ for some $\pi_1 \in (0,1)$, u_n is such that $u_n \to 0$ and $\sqrt{n}u_n \to \infty$, and $\lambda_i^{H_i}$ is defined by (5.8), where $H_i = H_i(n)$*

are such that $H_i(n)/\sqrt{nu_n} \to 0$*, it holds that*

$$P\left(\left|\frac{\hat{k}}{n} - \pi_1\right| \geq u_n\right) \to 0, \quad n \to \infty, \tag{5.12}$$

where $\hat{k}$ *is the number of signals estimated by the TSS procedure. Furthermore, the FDR and the FNR of the TSS procedure go to zero.*

Proof. We separate (5.12) into two terms:

$$P\left(\left|\frac{\hat{k}}{n} - \pi_1\right| \geq u_n\right) \leq P\left(\frac{\hat{k}}{n} - \pi_1 \leq -u_n\right) + P\left(\frac{\hat{k}}{n} - \pi_1 \geq u_n\right).$$

For the second term, the convergence to zero follows directly from Duval et al. (2007), as our thresholds are larger than the ones considered therein, making the probability considered even smaller. To get the convergence of the first term to zero it is enough to prove that

$$P1 = P\left(\left\{\frac{\hat{k}}{n} - \pi_1 \leq -u_n\right\} \cap \Omega_n\right) \to 0, \quad n \to \infty.$$

Note that it is more precise to write $\lceil nu_n \rceil$ From the contiguity property it follows that

$$\begin{aligned} P1 &\leq P\left(\{T_{k-nu_n+1} < \lambda_{k-nu_n+1}\} \cap \Omega_n\right) \\ &\leq P\left(\sum_{i=1}^{n-k} Y_{(i)} + \sum_{i=n-k+1}^{n-k+nu_n} Y_{(i)} \leq n - k + nu_n + H_{k-nu_n}\sqrt{2(n - k_n + nu_n)}\right) \\ &= P\left(\frac{1}{n-k+nu_n}\sum_{i=1}^{n-k} Y_{(i)}^{NS} + \frac{1}{n-k+nu_n}\sum_{i=1}^{nu_n} Y_{(i)}^{S} \leq 1 + \frac{H_{k-nu_n}\sqrt{2}}{\sqrt{n-k+nu_n}}\right). \end{aligned}$$

Now we separate the chi-square sum from the non-central chi-square sum:

$$P1 \leq P\left(\frac{1}{n-k+nu_n}\sum_{i=1}^{n-k} Y_{(i)}^{NS} \leq 1 - \frac{(1+\varepsilon)nu_n}{n-k+nu_n} + \frac{H_{k-nu_n}\sqrt{2}}{\sqrt{n-k+nu_n}}\right) + P\left(\frac{1}{n-k+nu_n}\sum_{i=1}^{nu_n} Y_{(i)}^{S} \leq \frac{(1+\varepsilon)nu_n}{n-k+nu_n}\right).$$

Denote the first term on the RHS as $P1'$ and the second one as $P1''$. $P1''$ goes to zero, which follows from the perfect separation condition, as in Duval et al. (2007). For $P1'$, the central limit theorem can be used to prove that it converges to zero.

$$\begin{aligned} P1' &= P\left(\frac{\sum_{i=1}^{n-k} Y_{(i)}^{NS} - (n-k)}{\sqrt{2(n-k)}} \leq -\frac{\varepsilon nu_n}{\sqrt{2(n-k)}} + H_{k-nu_n}\sqrt{1+\frac{nu_n}{n-k}}\right) \\ &= P\left(\frac{\sum_{i=1}^{n-k} Y_{(i)}^{NS} - (n-k)}{\sqrt{2(1-\pi_1)n}} \leq -\frac{\varepsilon\sqrt{n}u_n}{\sqrt{2(1-\pi_1)}} + H_{k-nu_n}\sqrt{1+\frac{u_n}{(1-\pi_1)}}\right). \end{aligned}$$

In order for $P1' \to 0$, from the central limit theorem, it is sufficient to have the expression on the RHS of the inequality go to $-\infty$. As $u_n \to 0$ and $\sqrt{n}u_n \to \infty$, it holds that

$$\sqrt{1+\frac{u_n}{(1-\pi_1)}} \to 1,$$

and

$$-\frac{\varepsilon\sqrt{n}u_n}{\sqrt{2(1-\pi_1)}} \to -\infty.$$

For H_n such that $\frac{\sqrt{n}u_n}{H_n} \to \infty$, the first term dominates the second and we have $P1' \to 0$. Finally, having the consistency of the proportion estimator, the proof of vanishing FDR and FNR follows as in Duval et al. (2007). □

Theorem 3. *For any $\varepsilon > 0$ define*

$$\tilde{y}_O(\varepsilon) = \min\left\{y : \int_0^y Q_{\chi_1^2(\mu^2)}(x)dx = \frac{1}{\pi_1}y(1+\varepsilon)\right\},$$

The quantity $\tilde{y}_O(\varepsilon)$ is asymptotically the proportion of the smallest signal values necessary to include so that their mean is at least $1+\varepsilon$. The oracle TSS procedure with thresholds given by (5.8) with $H_i = 0$ stops at the location $\hat{k}^O$ for which it holds that

$$P\left(\frac{\hat{k}^O}{n} - \frac{k}{n} < -\tilde{y}_O(\varepsilon)\right) \to 0, \quad n \to \infty.$$

meaning that asymptotically the mean of the undetected signal is no significantly larger than 1.

Proof. Note that

$$\begin{aligned} P\left(\frac{\hat{k}^O}{n} - \frac{k}{n} \le -\tilde{y}_O(\varepsilon)\right) &= P(\hat{k}^O \le k - n\tilde{y}_O(\varepsilon)) \\ &= P(T_{k-n\tilde{y}_O(\varepsilon)+1} < n - k + n\tilde{y}_O(\varepsilon)) \\ &= P(\chi^2_{n-k} + \tilde{S}_k^{k-n\tilde{y}_O(\varepsilon)} < n - k + n\tilde{y}_O(\varepsilon)), \end{aligned} \tag{5.33}$$

where $\tilde{S}_k^i$ is the sum of $n-i$ smallest order statistics from the sample of size k, from the noncentral chi-square distribution $\chi^2_1(\mu^2)$. Denote the signal values as $Y^S_{(1)}, \ldots, Y^S_{(k)}$ Let Q^S_k be the empirical quantile function of the signal variables. First, note that

$$\begin{aligned} \frac{1}{k}\tilde{S}_k^{k-n\tilde{y}_O(\varepsilon)} &:= \frac{1}{k}\sum_{i=1}^{n\tilde{y}_O(\varepsilon)} Y^S_{(i)} \\ &= \frac{1}{k}\sum_{i=1}^{k} Y^S_{(i)}\mathbb{I}\{Y^S_{(i)} \le Q^S_k(\tilde{y}_O(\varepsilon))\} \\ &\overset{a.s.}{\to} \int_0^{Q_{\chi^2_1(\mu^2)}(\tilde{y}_O(\varepsilon))} x dF(x), \quad n \to \infty \\ &= \int_0^{\tilde{y}_O(\varepsilon)} Q(y)dy = \frac{1}{\pi_1}\tilde{y}_O(\varepsilon)(1+\varepsilon). \end{aligned} \tag{5.34}$$

The almost sure convergence obtained above follows for example from the convergence of the empirical Lorenz curve studied in Goldie (1977) (see Theorem 1). Note that this result is also used in the proof of Theorem 4 below.

Let $0 < \varepsilon_0 < \varepsilon$. We split the probability in (5.33) into two terms, as in the proof of Theorem 2 and get

$$P\left(\frac{\hat{k}^O}{n} - \frac{k}{n} \leq -\tilde{y}_O(\varepsilon)\right) \leq P(\chi^2_{n-k} < n - k - \varepsilon_0 n\tilde{y}_O) + P\left(\tilde{S}_k^{k-n\tilde{y}_O} \leq n\tilde{y}_O(1+\varepsilon_0)\right) \tag{5.35}$$

The first term goes to zero, which follows from the central limit theorem. For the second term, it is enough to prove that

$$P\left(\frac{1}{n\tilde{y}_O}\tilde{S}_k^{k-n\tilde{y}_O} \leq 1 + \varepsilon_0\right) \to 0, \quad n \to \infty.$$

This holds since

$$P\left(\frac{k}{n\tilde{y}_O}\frac{1}{k}\tilde{S}_k^{k-n\tilde{y}_O} \leq 1 + \varepsilon_0\right) = P\left(\frac{\pi_1}{\tilde{y}_O}\frac{1}{k}\tilde{S}_k^{k-n\tilde{y}_O} - (1+\varepsilon) \leq \varepsilon_0 - \varepsilon\right) \to 0, \quad n \to \infty$$

where the last equality follows from the convergence result in (5.34) □

Theorem 4. *Let $\hat{y}_n$ and $\tilde{y}$ be defined as in (5.23) and (5.25). It holds that*

$$\hat{y}_n \xrightarrow{a.s.} \tilde{y}, \quad n \to \infty.$$

Proof. First, we note that $T_n(y)$ is closely related to the Lorenz curve estimator denoted as $\hat{L}_n(y)$ in Goldie (1977). Precisely, it holds that

$$\hat{L}_n(y) = \frac{T_n(1-y)}{T_n(0)}.$$

In Theorem 1 in Goldie (1977) it is proved that

$$\sup_{y\in[0,1]} |\hat{L}_n(y) - L(y)| \xrightarrow{a.s.} 0, \quad n \to \infty.$$

From the law of large numbers, we have that

$$\frac{1}{n}T_n(0) = \frac{1}{n}\sum_{i=1}^{n} Y_i \xrightarrow{a.s.} 1 + \pi_1\mu^2, \quad n \to \infty.$$

The uniform convergence of the scaled TSS process to the corresponding (scaled and reversed) Lorenz curve now follows directly from these last two asymptotic results and the following inequalities:

$$\begin{aligned}
&\sup_{y\in[0,1]} \left|\frac{1}{n}T_n(y) - (1+\pi_1\mu^2)L(1-y)\right| \\
&= \sup_{y\in[0,1]} \left|\frac{1}{n}T_n(1-y) - (1+\pi_1\mu^2)L(y)\right| \\
&= \frac{1}{n}T_n(0) \sup_{y\in[0,1]} \left|\hat{L}_n - \frac{1+\pi_1\mu^2}{\frac{1}{n}T_n(0)}L(y)\right| \\
&\leq \frac{1}{n}T_n(0) \left(\sup_{y\in[0,1]} |\hat{L}_n(y) - L(y)| + \left|\frac{1+\pi_1\mu^2}{\frac{1}{n}T_n(0)} - 1\right| \sup_{y\in[0,1]} |L(y)| \right) \\
&\xrightarrow{a.s.} 0, \quad n \to \infty.
\end{aligned}$$

As the sequence of thresholds is non-random it holds that $\lambda_n(y)/n \xrightarrow{a.s.} 1 - y$. It follows that the intersection point between $T_n(y)$ and $\lambda_n(y)$ ($\hat{y}_n$ by definition) has to converge almost surely to the unique intersection point between $(1+\pi_1\mu^2)L(1-y)$ and $1-y$ ($\tilde{y}$ by definition), that is:

$$\hat{y}_n = \max\{y : T_n(y) \geq \lambda_n(y)\} \xrightarrow{a.s.} \max\left\{y : (1+\pi_1\mu^2)L(1-y) \geq 1-y\right\} = \tilde{y},$$

which concludes the proof. □

Theorem 5. *Let $\hat{k}_n$ be the estimated number of signals using the TSS procedure with thresholds given by (5.8) with $H_i = 0$. Let $k = \pi_1 n$ be the true number of signals, j the position of the largest non-signal variable in the decreasingly sorted sample, and assume that $j/k \to 0$, as $n \to \infty$. It holds that*

$$P\left(\hat{k}_n \leq j\right) \to 0, \quad n \to \infty.$$

Proof. The probability that the TSS procedure stops in j or fewer steps, is the probability that the statistic at $(j+1)$st step is below the threshold.

$$P\left(\hat{k}_n \leq j\right) = P\left(\frac{1}{n-j}\sum_{i=1}^{n-j} Y_i \leq 1\right).$$

As j is the location of the largest non-signal value, the terms in the sum above are the whole χ^2_1 sample of size $n-k$ and $k-j$ signal values. By separating the sum above into the sum of signal and non-signal terms, we can separate the probability in two terms for some $\varepsilon > 0$:

$$\begin{aligned} P\left(\frac{1}{n-j}\sum_{i=1}^{n-j} Y_i \leq 1\right) \leq\ & P\left(\frac{1}{n-j}\sum_{i=1}^{n-k} Y^{NS}_{(i)} \leq 1 - \frac{(1+\varepsilon)(k-j)}{n-j}\right) \\ & + P\left(\frac{1}{n-j}\tilde{S}^j_{k_n} \leq \frac{(1+\varepsilon)(k-j)}{n-j}\right), \end{aligned} \tag{5.36}$$

where $\tilde{S}^j_{k_n} = \sum_{i=1}^{k-j} Y^S_{(i)}$ is the sum of $k-j$ smallest order statistics from the noncentral chi-square distribution. For the first term on the RHS we use the central limit theorem to get the convergence to zero. This probability comes down to

$$P\left(\frac{\sum_{i=1}^{n-k} Y^{NS}_{(i)} - (n-k)}{\sqrt{2(n-k)}} \leq -\frac{\varepsilon(k-j)}{\sqrt{2(n-k)}}\right).$$

As $k = \pi_1 n$ and $j/n \to 0$, as $n \to \infty$, it holds that

$$-\frac{\varepsilon(k-j)}{\sqrt{2(n-k)}} \to -\infty, \quad n \to \infty.$$

From the central limit theorem it follows that the above probability goes to zero. It is left to prove that the second term on the RHS in (5.36) goes to zero:

$$P\left(\frac{1}{k-j}\tilde{S}_k^j \leq 1+\varepsilon\right) \to 0, \quad n \to \infty. \tag{5.37}$$

If j behaves like a constant, that is $j/k \sim 1/k$, then $\tilde{S}_k^j$ is a lightly trimmed sum, and the central limit theorem holds as in the full-sample case (see for example Maller (1982), or Kesten (1993) for a more general discussion). The central limit theorem for the lightly trimmed sum from the noncentral chi-square distribution $\chi_1^2(\mu^2)$, implies

$$P\left(\frac{\tilde{S}_k^j - k(1+\mu^2)}{\sqrt{k(1+2\mu^2)}} \leq \frac{(1+\varepsilon)(k-j) - k(1+\mu^2)}{\sqrt{k(1+2\mu^2)}}\right) \to 0, \quad n \to \infty.$$

This holds since $\varepsilon > 0$ can be arbitrarily small, so we can chose it such that the value on the RHS goes to $-\infty$, and the probability above goes to zero. If j is such that $j/k \to 0$, but $j \to \infty$ then we can apply the weak convergence result for the moderately trimmed sums from Csörgő et al. (1986) (see Theorem 1). The theorem states that

$$\frac{S_k^j - k\int_{j/k}^{1-j/k} Q(u)du}{\sqrt{k\sigma^2(j/k)}} \xrightarrow{D} N(0,1), \quad n \to \infty. \tag{5.38}$$

where S_k^j is the symmetrically trimmed sum, excluding the smallest and the largest signal values $S_k^j = \sum_{i=j}^{k-j} Y_{(i)}^S$, and $\sigma^2(j/n)$ is a bounded function, such that $\sigma^2(j/n) \leq (1+\mu^2)^2$.

Since $S_k^j < \tilde{S}_k^j$, to have the vanishing probability in (5.37), it is enough to show that

$$P\left(\frac{S_k^j - k\int_{j/k}^{1-j/k} Q(u)du}{\sqrt{k(1+2\mu^2)}} \leq \frac{(1+\varepsilon)(k-j) - k\int_{j/k}^{1-j/k} Q(u)du}{\sqrt{k(1+2\mu^2)}}\right) \to 0,$$

as $n \to \infty$. The quantity on the RHS goes to $-\infty$ since $\varepsilon > 0$ is arbitrarily small and $\int_{j/k}^{1-j/k} Q(u)du \to \int_0^1 Q(u)du = 1 + \mu^2$. This observation and (5.38) conclude the proof. □

Chapter 6

Conclusions

In this thesis we approach the problem of multiple testing using ideas from the change-point literature, and we consider the problem of inference on the signal in the Gaussian sequence model. In this chapter, we provide a brief summary of our main contributions in Chapters 3, 4 and 5 and discuss possible directions for future research.

In Chapter 3, we propose the Difference of Slopes (DOS) method, a two-step method for estimating the proportion of false null hypotheses. In the first step, we approximate the sequence of sorted p-values with a piecewise linear function with one change-point in slope. In the second step, to get the proportion estimate, we apply the Storey's estimator by Storey (2002) using that change-point as the parameter. This essentially means approximating the quantile function of the p-value distribution with a piecewise linear function. The theoretical results use the theory of quantile processes and show that our estimator is asymptotically conservative in the sense that asymptotically, the false null proportion estimator has a non-positive bias. The simulation results show that for small ($n = 50$) and moderate ($n = 1000$) sample sizes, the DOS procedure works well when compared to the other proportion estimators in the literature. It works particularly well in sparse cases, when the proportion of false null hypotheses is small. If the proportion is large, but the signal is very weak, the single

change-point approximation might be inadequate causing the variance of the estimates to be larger. Also, for very large samples ($n = 10000$), and when the alternative is not strong enough, the DOS estimator can significantly underestimate the proportion. The reason for this is that the quantile function is smooth and the piecewise linear approximation is inadequate. Therefore, a possibility for future research is to make the DOS a sequential procedure, and estimate multiple change-points. This could improve the estimation in weak cases by using a more appropriate piecewise linear function for the approximation of the quantile function. Additionally, as mentioned in Section 3.5, a generalized version of the DOS estimator, α-DOS, should be further investigated, along with the methods for choosing α. As discussed in Section 3.6, a possible application of proportion estimators is in making the Benjamini-Hochberg procedure adaptive. For this purpose the almost sure conservativeness of a proportion estimator is a desirable property as it guarantees the asymptotic FDR control of the resulting adaptive procedure. Furthermore, as the classic Benjamini-Hochberg procedure lies at the core of many modern multiple testing procedures, the main direction for further research is to make these procedures adaptive by using the DOS proportion estimator.

In Chapter 4, we discuss some existing connections between the global testing problem and the change-point problem that have not been exploited in the literature. We discuss the relationship between the Higher Criticism (HC) statistic by Donoho and Jin (2004) and the Cumulative-Sum (CUSUM) statistic that is widely used in the change-point literature. Many change-point detection procedures use the CUSUM statistic on the subintervals of the data in order to improve testing and estimation when there are multiple change-points. We discuss some of these methods in Section 4.1 and 4.3. It might be of interest for future research to consider these alternative choices of intervals, and to adapt for example, the MOSUM by Eichinger and Kirch

(2018), the DOS or some other multiple change-point procedure for the global testing problem, by applying them on the sequence of spacings.

Extending our proposal in Section 4.3, we now mention some additional tools from the change-point literature that can be used on the sequence of spacings. These include smooth/gradual change detection or S-shaped function estimation for the sequence of spacings that would identify the smooth transition between the false null and the true null p-value spacings. Furthermore, shape constraints can be incorporated in the suggested methods as the direction of the change is known.

Gradual change - If we consider the sequence of p-values spacings, finding the point of the end of the linear part is naturally a gradual change problem. If the false null distribution is such that its support is $[0, b]$, then the spacings corresponding to the p-values larger than $[0, b]$ are uniform spacings, identically distributed and approximately independent. For spacings smaller than b, the distribution changes. The literature on abrupt change inference is substantially larger than on the gradual change. In Vogt and Dette (2015) a nonparametric method for testing and estimating the point of the onset of change is proposed. This method requires that the features that are changing in the locally stationary process are known . This includes mean, variance or covariance structure of the process. In Nie and Nicolae (2021), they propose a nonparametric kernel-based method for gradual change detection and localisation.

S-shaped functions - As the sequence of p-values approximates the quantile function of the p-value distribution, we can consider that the sequence of spacings approximates the derivative of the quantile function which is $Q'(t) = 1/f(Q(t))$, a reciprocal of the density quantile function. In the case of decreasing density, function $Q'(t)$ is S-shaped, as the values around $t = 0$ are close to zero, and for t large enough, $Q'(t) \approx 1/\pi_0$ is constant. In Feng et al. (2021) a least squares estimator as a nonparametric method for estimating S-shaped function is analysed and an efficient algorithm for computing

the estimator is proposed. The fitted function is piecewise linear, and in this sense also appropriate for the problem of grouping p-values.

In Chapter 5 we propose the Tail-Summed Scores (TSS), a method for inference in the Gaussian sequence model. TSS is a pseudo-sequential procedure for signal recovery that considers values in groups with the aim to increase the number of true signal discoveries when the signal is weak. Starting from the full sample, the procedure removes the largest absolute values one by one, which are then declared as signal, until the remaining set of values begins to resemble noise as a group. At each step, the norm of the remaining values is compared to a (possibly different) threshold, and the procedure is stopped as soon as the remaining norm is below the threshold. As the Gaussian values are squared and summed when calculating the norm, the proposed thresholds are quantiles of chi-square distributions with decreasing degrees of freedom. The theoretical analysis of the TSS method is given first in the case of a strong signal, that is when signal and non-signal values are well separated. These results rely on the results from Duval et al. (2007), where a similar procedure is proposed for the purpose of multiple testing. In this case the FDR of the TSS procedure is proved to go to zero. The theoretical analysis when no assumption on the signal strength is made is considerably more difficult. In this case we first consider the oracle TSS procedure. The oracle TSS uses the oracle sorting, separating the signal from the non-signal values, even when the signal is weak and signal values are smaller than the non-signal values. We explore the oracle TSS, as we find that empirically, the oracle TSS stops at a similar location as the regular TSS. We note the connection between the TSS sequence and the Lorenz curve of the squared sample, which yields the results on the asymptotic behaviour of the procedure. In the weak signal case, we prove that the TSS procedure will stop after the mixing between the signal and non-signal values start, so in general it will include some false positives. As mentioned

above, the conservativeness of the procedure can be adjusted by choosing different sequence of thresholds. In Section 5.4, a connection between the minimax estimator for the l_2 norm of the signal and the TSS method is noted. This suggests that different thresholds can be used for different purposes, and the topic of future research can be to find the best thresholds for each problem, while keeping the idea of considering values in groups. We can consider the problems of FWER or FDR control, estimating the signal or estimating the l_2 norm of the signal. Another possible extension of the TSS method comes from considering different distributions – not just the Gaussian sequence model. Finally, different transformation of the data can be considered, as an alternative to the squared values used by the proposed method.

References

Abraham, K., Castillo, I., and Roquain, E. (2021). Sharp multiple testing boundary for sparse sequences. *arXiv preprint arXiv:2109.13601*.

Abraham, K., Castillo, I., and Roquain, É. (2022). Empirical Bayes cumulative l-value multiple testing procedure for sparse sequences. *Electronic Journal of Statistics*, 16(1):2033–2081.

Abramovich, F., Benjamini, Y., Donoho, D. L., and Johnstone, I. M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *The Annals of Statistics*, 34(2):584–653.

Abramovich, F., Grinshtein, V., Petsa, A., and Sapatinas, T. (2010). On Bayesian testimation and its application to wavelet thresholding. *Biometrika*, 97(1):181–198.

Akman, V. E. and Raftery, A. E. (1986). Asymptotic Inference for a Change-Point Poisson Process. *The Annals of Statistics*, 14(4):1583–1590.

Anastasiou, A. and Fryzlewicz, P. (2022). Detecting multiple generalized change-points by isolating single ones. *Metrika*, 85(2):141–174.

Anderson, T. W. and Darling, D. A. (1952). Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes. *The Annals of Mathematical Statistics*, 23(2):193–212.

Arias-Castro, E., Candès, E. J., and Plan, Y. (2011). Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. *Annals of Statistics*, 39(5):2533–2556.

Arias-Castro, E. and Ying, A. (2019). Detection of sparse mixtures: higher criticism and scan statistic. *Electronic Journal of Statistics*, 13(1):208–230.

Banerjee, T., Fu, L. J., James, G. M., and Sun, W. (2020). Nonparametric Empirical Bayes Estimation on Heterogeneous Data. *arXiv preprint arXiv:2002.12586*.

Baranowski, R., Chen, Y., and Fryzlewicz, P. (2016). Narrowest-over-threshold detection of multiple change-points and change-point-like features. *arXiv preprint arXiv:1609.00293*.

Baranowski, R., Chen, Y., and Fryzlewicz, P. (2019). Narrowest-over-threshold detection of multiple change points and change-point-like features. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 81(3):649–672.

Barbe, P. (1992). Limiting distribution of the maximal spacing when the density function admits a positive minimum. *Statistics & probability letters*, 14(1):53–60.

Barber, R. F. and Candés, E. J. (2015). Controlling the false discovery rate via knockoffs. *Annals of Statistics*, 43(5):2055–2085.

Barber, R. F. and Ramdas, A. (2017). The p -filter: multilayer false discovery rate control for grouped hypotheses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1247–1268.

Basu, P., Cai, T. T., Das, K., and Sun, W. (2018). Weighted False Discovery Rate Control in Large-Scale Multiple Testing. *Journal of the American Statistical Association*, 113(523):1172–1183.

Belitser, E. and Nurushev, N. (2020). Needles and straw in a haystack: Robust confidence for possibly sparse sequences. *Bernoulli*, 26(1):191–225.

Benjamini, Y. and Bogomolov, M. (2014). Selective inference on multiple families of hypotheses. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 76(1):297–318.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Benjamini, Y. and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of educational and Behavioral Statistics*, 25(1):60–83.

Berk, R. H. and Jones, D. H. (1979). Goodness-of-fit test statistics that dominate the Kolmogorov statistics. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 47(1):47–59.

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015). Dirichlet–Laplace Priors for Optimal Shrinkage. *Journal of the American Statistical Association*, 110(512):1479–1490.

Blanchard, G. and Roquain, É. (2009). Adaptive false discovery rate control under independence and dependence. *Journal of Machine Learning Research*, 10:2837–2871.

Bogomolov, M., Peterson, C. B., Benjamini, Y., and Sabatti, C. (2021). Hypotheses on a tree: new error rates and testing strategies. *Biometrika*, 108(3):575–590.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

Broberg, P. (2005). A comparative review of estimates of the proportion unchanged genes and the false discovery rate. *BMC bioinformatics*, 6(1):199.

Broberg, P. (2020). *SAGx: Statistical Analysis of the GeneChip*. R package version 1.64.0.

Cai, T. and Jin, J. (2010). Optimal rates of convergence for estimating the null density and proportion of nonnull effects in large-scale multiple testing. *Annals of Statistics*, 38(1):100–145.

Cai, T. T. (2017). Global Testing and Large-Scale Multiple Testing for High-Dimensional Covariance Structures. *Annual Review of Statistics and Its Application*, 4(1):423–446.

Cai, T. T., Jin, J., and Low, M. G. (2007). Estimation and confidence sets for sparse normal mixtures. *Annals of Statistics*, 35(6):2421–2449.

Cai, T. T. and Sun, W. (2009). Simultaneous Testing of Grouped Hypotheses: Finding Needles in Multiple Haystacks. *Journal of the American Statistical Association*, 104(488):1467–1481.

Cai, T. T., Sun, W., and Xia, Y. (2022). LAWS: A Locally Adaptive Weighting and Screening Approach to Spatial Multiple Testing. *Journal of the American Statistical Association*, 117(539):1370–1383.

Cai, T. T. and Wei, H. (2022). Distributed nonparametric function estimation: Optimal rate of convergence and cost of adaptation. *The Annals of Statistics*, 50(2):1–35.

Cao, H., Chen, J., and Zhang, X. (2022). Optimal false discovery rate control for large scale multiple testing with auxiliary information. *The Annals of Statistics*, 50(2):807–857.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480.

Castillo, I. and Roquain, É. (2020). On spike and slab empirical Bayes multiple testing. *The Annals of Statistics*, 48(5).

Castillo, I. and Szabó, B. (2020). Spike and slab empirical Bayes sparse credible sets. *Bernoulli*, 26(1):127–158.

Castillo, I. and van der Vaart, A. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics*, 40(4):2069–2101.

Celisse, A. and Robin, S. (2010). A cross-validation based estimation of the proportion of true null hypotheses. *Journal of Statistical Planning and Inference*, 140(11):3132–3147.

Chatterjee, S. (2014). A new perspective on least squares under convex constraint. *The Annals of Statistics*, 42(6):2340–2381.

Chen, X., Doerge, R. W., and Heyse, J. F. (2018). Multiple testing with discrete data: Proportion of true null hypotheses and two adaptive FDR procedures. *Biometrical Journal*, 60(4):761–779.

Chen, X., Doerge, R. W., and Sarkar, S. K. (2020). A weighted FDR procedure under discrete and heterogeneous null distributions. *Biometrical Journal*, 62(6):1544–1563.

Chen, X., Guntuboyina, A., and Zhang, Y. (2017). A note on the approximate admissibility of regularized estimators in the Gaussian sequence model. *Electronic Journal of Statistics*, 11(2):4746–4768.

Cheng, D., He, Z., and Schwartzman, A. (2015). Multiple Testing of Local Extrema for Detection of Change Points. *Electronic Journal of Statistics*, 14(2):3705–3729.

Chernoff, H. and Rubin, H. (1956). The estimation of the location of a discontinuity in density. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 19–37. University of California Press.

Chernozhukov, V., Hansen, C., and Liao, Y. (2017). A lava attack on the recovery of sums of dense and sparse signals. *The Annals of Statistics*, 45(1):39–76.

Cho, H. (2016). Change-point detection in panel data via double cusum statistic. *Electron. J. Statist.*, 10:2000–2038.

Cho, H. and Fryzlewicz, P. (2015). Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(2):475–507.

Chung, K.-L. (1949). An Estimate Concerning the Kolmogroff Limit Distribution. *Transactions of the American Mathematical Society*, 67(1):36.

Collier, O., Comminges, L., and Tsybakov, A. B. (2017). Minimax estimation of linear and quadratic functionals on sparsity classes. *The Annals of Statistics*, 45(3):923–958.

Collier, O., Comminges, L., Tsybakov, A. B., and Verzelen, N. (2018). Optimal adaptive estimation of linear functionals under sparsity. *The Annals of Statistics*, 46(6A):3130–3150.

Comminges, L., Collier, O., Ndaoud, M., and Tsybakov, A. B. (2021). Adaptive robust estimation in sparse vector model. *The Annals of Statistics*, 49(3):1347–1377.

Cramér, H. (1928). On the composition of elementary errors. *Scandinavian Actuarial Journal*, 1928(1):13–74.

Csörgő, M. (1983). *Quantile Processes with Statistical Applications*. Society for Industrial and Applied Mathematics.

Csörgő, S., Horváth, L., and Mason, D. M. (1986). What portion of the sample makes a partial sum asymptotically stable or normal? *Probability Theory and Related Fields*, 72(1):1–16.

Cupples, L. A., Arruda, H. T., Benjamin, E. J., D'Agostino, R. B., Demissie, S., DeStefano, A. L., Dupuis, J., Falls, K. M., Fox, C. S., Gottlieb, D. J., Govindaraju, D. R., Guo, C.-Y., Heard-Costa, N. L., Hwang, S.-J., Kathiresan, S., Kiel, D. P., Laramie, J. M., Larson, M. G., Levy, D., Liu, C.-Y., Lunetta, K. L., Mailman, M. D., Manning, A. K., Meigs, J. B., Murabito, J. M., Newton-Cheh, C., O'Connor, G. T., O'Donnell, C. J., Pandey, M., Seshadri, S., Vasan, R. S., Wang, Z. Y., Wilk, J. B., Wolf, P. A., Yang, Q., and Atwood, L. D. (2007). The Framingham Heart

Study 100K SNP genome-wide association study resource: overview of 17 phenotype working group reports. *BMC Medical Genetics*, 8(S1):S1.

Davidson, M. A. and Shanks, E. J. (2017). 3q26-29 Amplification in head and neck squamous cell carcinoma: a review of established and prospective oncogenes. *The FEBS Journal*, 284(17):2705–2731.

Deheuvels, P. (1984). Strong limit theorems for maximal spacings from a general univariate distribution. *The Annals of Probability*, pages 1181–1193.

Donoho, D. and Jin, J. (2008). Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. *Proceedings of the National Academy of Sciences*, 105(39):14790–14795.

Donoho, D. and Jin, J. (2009). Feature selection by higher criticism thresholding achieves the optimal phase diagram. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4449–4470.

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

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455.

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to Unknown Smoothness via Wavelet Shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224.

Donoho, D. L. and Johnstone, I. M. (1998). Minimax estimation via wavelet shrinkage. *The Annals of Statistics*, 26(3):879–921.

Du, L., Guo, X., Sun, W., and Zou, C. (2021). False Discovery Rate Control Under General Dependence By Symmetrized Data Aggregation. *Journal of the American Statistical Association*, pages 1–15.

Duval, M., Delmas, C., Laurent, B., and Robert-Granié, C. C. (2007). A procedure based on partial sums of order statistics to detect differentially expressed genes. Working paper or preprint, available at https://hal.archives-ouvertes.fr/hal-00302355/file/Duval_etal_arxiv.pdf.

Duwadi, K., Austin, R. S., Mainali, H. R., Bett, K., Marsolais, F., and Dhaubhadel, S. (2018). Slow darkening of pinto bean seed coat is associated with significant metabolite and transcript differences related to proanthocyanidin biosynthesis. *BMC Genomics*, 19(1):260.

Efron, B. (2007). Size, power and false discovery rates. *Annals of Statistics*, 35(4):1351–1377.

Efron, B. (2008). Microarrays, empirical bayes and the two-groups model. *Statistical Science*, 23(1):1–22.

Eichinger, B. and Kirch, C. (2018). A MOSUM procedure for the estimation of multiple random change points. *Bernoulli*, 24(1):526–564.

Einmahl, J. H. J. and Mason, D. M. (1988). Strong Limit Theorems for Weighted Quantile Processes. *The Annals of Probability*, 16(4):1623–1643.

Feng, O. Y., Chen, Y., Han, Q., Carroll, R. J., and Samworth, R. J. (2021). Nonparametric, tuning-free estimation of S-shaped functions. *arXiv preprint arXiv:2107.07257*.

Finner, H. and Roters, M. (2001). On the False Discovery Rate and Expected Type I Errors. *Biometrical Journal*, 43(8):985.

Fisher, R. A. (1946). *Statistical methods for research workers*. Oliver and Boyd.

Friguet, C. and Causeur, D. (2011). Estimation of the proportion of true null hypotheses in high-dimensional data under dependence. *Computational Statistics and Data Analysis*, 55(9):2665–2676.

Fryzlewicz, P. (2007). Unbalanced Haar technique for nonparametric function estimation. *Journal of the American Statistical Association*, 102(480):1318–1327.

Fryzlewicz, P., Delouille, V., and Nason, G. P. (2007). GOES-8 X-ray sensor variance stabilization using the multiscale data-driven Haar–Fisz transform. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 56(1):99–116.

Fryzlewicz, P. and Nason, G. P. (2004). A Haar-Fisz Algorithm for Poisson Intensity Estimation. *Journal of Computational and Graphical Statistics*, 13(3):621–638.

Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 64(3):499–517.

Genovese, C. and Wasserman, L. (2004). A stochastic process approach to false discovery control. *The Annals of Statistics*, 32(3):1035–1061.

Genovese, C. R., Roeder, K., and Wasserman, L. (2006). False discovery control with p-value weighting. *Biometrika*, 93(3):509–524.

Goldie, C. M. (1977). Convergence theorems for empirical Lorenz curves and their inverses. *Advances in Applied Probability*, 9(4):765–791.

Grenander, U. (1956). On the theory of mortality measurement. *Scandinavian Actuarial Journal*, 1956(2):125–153.

Grubert, F., Zaugg, J., Kasowski, M., Ursu, O., Spacek, D., Martin, A., Greenside, P., Srivas, R., Phanstiel, D., Pekowska, A., Heidari, N., Euskirchen, G., Huber, W., Pritchard, J., Bustamante, C., Steinmetz, L., Kundaje, A., and Snyder, M. (2015). Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. *Cell*, 162(5):1051–1065.

Habiger, J., Watts, D., and Anderson, M. (2017). Multiple testing with heterogeneous multinomial distributions. *Biometrics*, 73(2):562–570.

Hahn, M. G., Mason, D. M., and Weiner, D. C., editors (1991). *Sums, Trimmed Sums and Extremes*. Progress in Probability. Birkhäuser Boston, Boston, MA.

Han, Q., Sen, B., and Shen, Y. (2022). High-dimensional asymptotics of likelihood ratio tests in the Gaussian sequence model under convex constraints. *The Annals of Statistics*, 50(1):376–406.

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

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

Hu, J. X., Zhao, H., and Zhou, H. H. (2010). False discovery rate control with groups. *Journal of the American Statistical Association*, 105(491):1215–1227.

Ignatiadis, N. and Huber, W. (2021). Covariate powered cross-weighted multiple testing. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 83(4):720–751.

Ignatiadis, N., Klaus, B., Zaugg, J. B., and Huber, W. (2016). Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature Methods*, 13(7):577–580.

Ingster, Y. I. and Suslina, I. A. (2003). *Nonparametric Goodness-of-Fit Testing Under Gaussian Models*, volume 169 of *Lecture Notes in Statistics*. Springer New York, New York, NY.

Ingster, Y. I., Tsybakov, A. B., and Verzelen, N. (2010). Detection boundary in sparse regression. *Electronic Journal of Statistics*, 4(none):1476–1526.

Jeng, X. (2016). Detecting weak signals in high dimensions. *Journal of Multivariate Analysis*, 147:234–246.

Jiang, H. and Doerge, R. (2008). Estimating the proportion of true null hypotheses for multiple comparisons. *Cancer Informatics*, 6:25–32.

Jiang, W. and Zhang, C. H. (2009). General maximum likelihood empirical bayes estimation of normal means. *Annals of Statistics*, 37(4):1647–1684.

Jin, J. and Cai, T. T. (2007). Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons. *Journal of the American Statistical Association*, 102(478):495–506.

Jirak, M. (2015). Uniform change point tests in high dimension. *The Annals of Statistics*, 43(6):2451–2483.

Johnstone, I. M. (2017). *Gaussian estimation: Sequence and wavelet models*. Book draft.

Johnstone, I. M. and Silverman, B. W. (2004). Needles and straw in haystacks: Empirical bayes estimates of possibly sparse sequences. *The Annals of Statistics*, 32(4):1594–1649.

Katsevich, E., Sabatti, C., and Bogomolov, M. (2021). Filtering the Rejection Set While Preserving False Discovery Rate Control. *Journal of the American Statistical Association*, pages 1–12.

Kendall, D. G. and Kendall, W. S. (1980). Alignments in two-dimensional random sets of points. *Advances in Applied Probability*, 12(2):380–424.

Kesten, H. (1993). Convergence in distribution of lightly trimmed and untrimmed sums are equivalent. *Mathematical Proceedings of the Cambridge Philosophical Society*, 113(3):615–638.

Kittlitz, R. G. (1999). Transforming the Exponential for SPC Applications. *Journal of Quality Technology*, 31(3):301–308.

Klaus, B. and Strimmer, K. (2013). Signal identification for rare and weak features: higher criticism or false discovery rates? *Biostatistics*, 14(1):129–143.

Klaus, B. and Strimmer, K. (2021). *fdrtool: Estimation of (Local) False Discovery Rates and Higher Criticism*. R package version 1.2.16.

Kolmogorov, A. (1933). Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari, Giorn.*, 4:83–91.

Kukurba, K. R. and Montgomery, S. B. (2015). Rna sequencing and analysis. *Cold Spring Harbor Protocols*, 2015(11):pdb–top084970.

Langaas, M., Lindqvist, B. H., and Ferkingstad, E. (2005). Estimating the proportion of true null hypotheses, with application to dna microarray data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(4):555–572.

Lei, L. and Fithian, W. (2018). AdaPT: an interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 80(4):649–679.

Lei, L., Ramdas, A., and Fithian, W. (2021). A general interactive framework for false discovery rate control under structural constraints. *Biometrika*, 108(2):253–267.

Lenth, R. V. (1989). Quick and Easy Analysis of Unreplicated Factorials. *Technometrics*, 31(4):469–473.

Li, A. and Barber, R. F. (2017). Accumulation Tests for FDR Control in Ordered Hypothesis Testing. *Journal of the American Statistical Association*, 112(518):837–849.

Li, A. and Barber, R. F. (2019). Multiple testing with the structure-adaptive Benjamini–Hochberg algorithm. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 81(1):45–74.

Li, J. and Siegmund, D. (2015). Higher criticism: p-values and criticism. *The Annals of Statistics*, 43(3):1323–1350.

Liu, J., Zhang, C., and Page, D. (2016a). Multiple testing under dependence via graphical models. *Annals of Applied Statistics*, 10(3):1699–1724.

Liu, Y., Sarkar, S. K., and Zhao, Z. (2016b). A new approach to multiple testing of grouped hypotheses. *Journal of Statistical Planning and Inference*, 179:1–14.

Lynch, G., Guo, W., Sarkar, S. K., and Finner, H. (2017). The control of the false discovery rate in fixed sequence multiple testing. *Electronic Journal of Statistics*, 11(2):4649–4673.

Ma, R., Tony Cai, T., and Li, H. (2021). Global and Simultaneous Hypothesis Testing for High-Dimensional Logistic Regression Models. *Journal of the American Statistical Association*, 116(534):984–998.

Maller, R. A. (1982). Asymptotic normality of lightly trimmed means – a converse. *Mathematical Proceedings of the Cambridge Philosophical Society*, 92(3):535–545.

Martin, R. and Walker, S. G. (2014). Asymptotically minimax empirical Bayes estimation of a sparse normal mean vector. *Electronic Journal of Statistics*, 8(2):2188–2206.

Massart, P. (1990). The Tight Constant in the Dvoretzky-Kiefer-Wolfowitz Inequality. *The Annals of Probability*, 18(3):1269–1283.

Meinshausen, N. and Rice, J. (2006). Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Annals of Statistics*, 34(1):373–393.

Mukherjee, G. and Johnstone, I. M. (2015). Exact minimax estimation of the predictive density in sparse Gaussian models. *The Annals of Statistics*, 43(3):937–961.

Mukherjee, R., Mukherjee, S., and Yuan, M. (2018). Global testing against sparse alternatives under Ising models. *The Annals of Statistics*, 46(5):2062–2093.

Neumann, A., Bodnar, T., and Dickhaus, T. (2021). Estimating the proportion of true null hypotheses under dependency: A marginal bootstrap approach. *Journal of Statistical Planning and Inference*, 210:76–86.

Nie, L. and Nicolae, D. (2021). A nonparametric method for gradual change problems with statistical guarantees. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. S., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 15721–15733. Curran Associates, Inc.

Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–572.

Ostrovnaya, I. and Nicolae, D. L. (2012). Estimating the proportion of true null hypotheses under dependence. *Statistica Sinica*, 22(4):1689–1716.

Pyke, R. (1965). Spacings. *Journal of the Royal Statistical Society: Series B (Methodological)*, 27(3):395–436.

Rabinovich, M., Ramdas, A., Jordan, M. I., and Wainwright, M. J. (2020). Optimal Rates and Tradeoffs in Multiple Testing. *Statistica Sinica*, 30(2):741–762.

Raffaello, A., Laveder, P., Romualdi, C., Bean, C., Toniolo, L., Germinario, E., Megighian, A., Danieli-Betto, D., Reggiani, C., and Lanfranchi, G. (2006). Denervation in murine fast-twitch muscle: short-term physiological changes and temporal expression profiling. *Physiological Genomics*, 25(1):60–74.

Ramdas, A. K., Barber, R. F., Wainwright, M. J., and Jordan, M. I. (2019). A unified treatment of multiple testing with prior knowledge using the P-filter. *Annals of Statistics*, 47(5):2790–2821.

Roquain, E. and Van de Wiel, M. A. (2009). Optimal weighting for false discovery rate control. *Electronic Journal of Statistics*, 3:678–711.

Rubin, H. (1961). The estimation of discontinuities in multivariate densities, and related problems in stochastic processes. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 563–574. University of California Press.

Sarkar, S. K. and Zhao, Z. (2017). Local False Discovery Rate Based Methods for Multiple Testing of One-Way Classified Hypotheses. *arXiv preprint arXiv:1712.05014*.

Schweder, T. and Spjøtvoll, E. (1982). Plots of P -values to evaluate many tests simultaneously. *Biometrika*, 69(3):493–502.

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

Stein, C. M. (1981). Estimation of the Mean of a Multivariate Normal Distribution. *The Annals of Statistics*, 9(6):1135–1151.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498.

Storey, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *Annals of Statistics*, 31(6):2013–2035.

Storey, J. D., Bass, A. J., Dabney, A., and Robinson, D. (2020). *qvalue: Q-value estimation for false discovery rate control*. R package version 2.22.0.

Storey, J. D., Taylor, J. E., and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):187–205.

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445.

Strimmer, K. (2008). A unified approach to false discovery rate estimation. *BMC Bioinformatics*, 9(1):303.

Sun, W. and Cai, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association*, 102(479):901–912.

Sur, P., Chen, Y., and Candès, E. J. (2017). The Likelihood Ratio Test in High-Dimensional Logistic Regression Is Asymptotically a Rescaled Chi-Square. *Probability Theory and Related Fields*, 175(1-2):487–558.

Swanepoel, J. W. (1999). The limiting behavior of a modified maximal symmetric $2s$-spacing with applications. *The Annals of Statistics*, 27(1):24–35.

Turkheimer, F. E., Smith, C. B., and Schmidt, K. (2001). Estimation of the number of "true" null hypotheses in multivariate analysis of neuroimaging data. *NeuroImage*, 13(5):920–930.

van Erven, T. and Szabó, B. (2020). Fast Exact Bayesian Inference for Sparse Signals in the Normal Sequence Model. *Bayesian Analysis*, -1(-1):1–28.

Vogt, M. and Dette, H. (2015). Detecting gradual changes in locally stationary processes. *Annals of Statistics*, 43(2):713–740.

Vostrikova, L. J. (1981). Detecting 'disorder' in multidimensional random processes. *Sovy. Math. Dokl.*, 24:55–59.

Xia, Y., Cai, T., and Cai, T. (2018). Two-Sample Tests for High-Dimensional Linear Regression with an Application to Detecting Interactions. *Statistica Sinica*, 28.

Xie, J., Cai, T. T., and Li, H. (2011). Sample size and power analysis for sparse signal recovery in genome-wide association studies. *Biometrika*, 98(2):273–290.

Yekutieli, D. (2008). Hierarchical false discovery rate-controlling methodology. *Journal of the American Statistical Association*, 103(481):309–316.

Zhang, C., Fan, J., and Yu, T. (2011). Multiple testing via FDRL for large-scale imaging data. *The Annals of Statistics*, 39(1):613–642.

Zhang, M. J., Xia, F., and Zou, J. (2019). Fast and covariate-adaptive method amplifies detection power in large-scale multiple hypothesis testing. *Nature Communications*, 10(1):3433.