Thèse n° 8102

Multiple testing with test statistics following heavy-tailed distributions

Présentée le 19 mars 2021

Faculté des sciences de base
Chaire de statistique appliquée
Programme doctoral en mathématiques

pour l'obtention du grade de Docteur ès Sciences

par

Zhiwen JIANG

Acceptée sur proposition du jury

Prof. C. Hongler, président du jury
Prof. S. Morgenthaler, directeur de thèse
Prof. O. Renaud, rapporteur
Prof. W. Zhang, rapporteur
Prof. A. Davison, rapporteur

École
polytechnique
fédérale
de Lausanne

2021

To my parents. . .

Acknowledgements

When I look back on the journey, the PhD is not a natural follow-up to the undergraduate education, but a preparation for the real adventure in academia. I joined the group of Prof. Stephan Morgenthaler as a doctoral student in the summer of 2016. Stephan is always patient and kind. He knows everything about the students, the stress and the ambition, the determination and the doubt. Throughout the four years Stephan has been guiding me in both teaching and research, and the discussions with him were always inspiring and encouraging. I would like to thank him for his patience and confidence in my PhD, and for his encouragement and guidance as a supervisor, a thesis advisor, a teacher and a mentor.

I am also thankful to Prof. Sofia Olhede for organising the evaluation of my thesis. I appreciate faithfully the jury of my oral exam, including my thesis advisor Stephan, the jury president Prof. Clément Hongler, the internal expert Prof. Anthony Davison and the external experts, Prof. Olivier Renaud from the University of Geneva and Prof. Weiping Zhang from the University of Science and Technology of China. I sincerely appreciate their careful evaluation and important feedback. I am also grateful to Prof. Victor Panaretos and Prof. Thomas Mountford for participating in my first year's candidacy exam.

Groups in statistics at EPFL have grown to a big family, and there were a lot of moments of gratitude in the past four years. I would like to thank my lab mates Helen, Djalel, Rémy, Daria and Zhengwei, for sharing their experience and feelings, especially at the moments of worries and hesitation. I am particularly thankful to Helen for correcting my writing and giving valuable feedback. I am indebted to my friends and colleagues, Sonia, Laya, Soumaya, Neda, Anda, Jonathan, Kartik, Tomas R., Tomas M., Timmy, Mario, Soham and Jacques. And I am also thankful to the secretaries of STAP and SDS, Nadia and Gabriella, and to the secretaries of EDMA, Anna and Anne, for their warm help.

I would like to thank my friends Panpan, Menglin, Weina, Weizhe, Haoqing, Shengquan, Yuan, and the sports team Tingyong, Ping, Qi, Lulu, Chaoqun, Xiaotong and Mingxiang for the encouragement and company. I would like to thank Hao for his love, respect and understanding.

Finally, this thesis is dedicated to my mother and my father, without whose love and decisive support I would not have been able to pursue and finish my PhD.

Lausanne, September 25, 2020

Abstract

In multiple testing problems where the components come from a mixture model of noise and true effect, we seek to first test for the existence of the non-zero components, and then identify the true alternatives under a fixed significance level α. Two parameters, namely the fraction of the non-null components ε and the size of the effects μ, characterise the two-point mixture model under the global alternative. When the number of hypotheses m goes to infinity, we are interested in an asymptotic framework where the fraction of the non-null components is vanishing, and the true effects need to be sizable to be detected. Donoho and Jin give an explicit form of the asymptotic detectable boundary based on the Gaussian mixture model under the classic calibration of the parameters of the mixture model. We prove the analogous results for the Cauchy mixture distribution as an example heavy-tailed case. This requires a different formulation of the parameters, which reflects the added difficulties.

We also propose a multiple testing procedure based on a filtering approach that can discover the true alternatives. Benjamini and Hochberg (BH) compare the observed p-values to a linear threshold curve and reject the null hypotheses from the minimum up to the last up-crossing, and prove the false discovery rate (FDR) is controlled. However, there is an intrinsic difference in heavy-tailed settings. Were we to use the BH procedure we would get a highly variable positive false discovery rate (pFDR). In our study we analyse the distribution of the p-values and devise a new multiple testing procedure to combine the usual case and the heavy-tailed case based on the empirical properties of the p-values. The filtering approach is designed to eliminate most p-values that are more likely to be uniform, while preserving most of the true alternatives. Based on the filtered p-values, we estimate the mode ϑ and define the rejection region $\mathcal{R}(\vartheta,\delta) = [\vartheta-\delta/2, \vartheta+\delta/2]$ such that the most informative p-values are included. The length δ is chosen by controlling the data-dependent estimation of FDR at a desired level.

Keywords: False discovery rate (FDR), filtering, heavy-tailed distribution, local FDR, mode estimation, multiple testing, operating characteristics, positive FDR.

Résumé

Dans un problème de tests multiples, où les variables suivent un modèle de mélange de bruit et d'effets réels, nous cherchons d'abord à tester l'existence des composantes non nulles, puis à identifier au niveau α les vraies alternatives parmi les mélanges. Deux paramètres, en l'occurrence la fraction des composantes non nulles ε et la taille des effets μ, caractérisent le modèle de mélange lorsque l'hypothèse alternative globale est vraie. Lorsque le nombre d'hypothèses m passe à l'infini, nous nous intéressons à une structure asymptotique, dans laquelle, la fraction des composantes non nulles diminue et la taille des effets réels est suffisamment grande pour être détectée. Donoho et Jin donnent une forme explicite de la frontière de détection asymptotique, basée sur le modèle de mélange Gaussien respectant les hypothèses classiques du modèle de mélange. Nous prouvons par analogie ce résultat pour la distribution de mélange de Cauchy, la distribution de Cauchy étant un exemple de cas de lois de probabilité à queue lourde. Cela nécessite une formulation différente des paramètres, qui reflète les difficultés supplémentaires.

Nous proposons également une procédure de tests multiples basée sur une approche par filtration permettant de découvrir les vraies alternatives. Benjamini et Hochberg (BH) comparent les p-valeurs observées à une courbe de seuil linéaire et rejettent les hypothèses nulles du minimum jusqu'au dernier croisement ascendant. Ils prouvent également que le taux de fausses découvertes (FDR) est contrôlé. Cependant, il existe une différence intrinsèque dans les paramètres à queue lourde. Si nous utilisions la procédure BH, nous obtiendrions un taux de fausses découvertes positives (pFDR) très variable. Dans notre recherche, nous analysons la distribution des p-valeurs et concevons une nouvelle procédure de tests multiples qui se base sur les propriétés empiriques des p-valeurs. Cette procédure permet de combiner le cas Gaussien et le cas à queue lourde. L'approche par filtration est conçue pour éliminer la plupart des p-valeurs qui sont plus susceptibles d'être uniformes, tout en préservant la plupart des vraies alternatives. Sur la base des p-valeurs filtrées, nous estimons le mode ϑ et définissons la région de rejet $\mathcal{R}(\vartheta,\delta) = [\vartheta - \delta/2, \vartheta + \delta/2]$ telle que les p-valeurs les plus informatives soient incluses. La longueur δ est choisie en contrôlant le FDR, estimé à partir des données, à un niveau souhaité.

Contents

1 Introduction

This work is aimed at developing an adaptive method for multiple testing problems where the test statistics follow a heavy-tailed distribution. The multiple testing problem gains more and more interest as large-scale simultaneous inference becomes an essential technique in many applications. In this chapter we first give a brief introduction to the multiple testing problem, and then explain why this topic is worth investigating from different perspectives.

1.1 Background

The general statistical decision problem investigates a set of observations which are the realised values x of random variables X whose distribution P_θ is at least partly unknown. The parameter θ is supposed to label the distribution of X, and is often assumed to vary in a parameter space Θ. The purpose of statistical inference is to obtain information from the observable results and understand the underlying distribution P_θ mathematically. A decision rule $\delta(x)$ is desired to provide guidance about the distribution and the parameter space, which simply maps the possible values x to the labels of decisions that should be chosen. To eliminate redundancy, we omit this notation δ in the following chapters and focus on the corresponding partition in the sample space of X, namely the rejection region and the acceptance region. Due to the randomness of the variable and the lack of knowledge of the background, uncertainty of the statistical decision is unavoidable. Every decision made between a number of possible candidates is associated with a consequence that should be evaluated quantitatively. In order to compare the conclusions mathematically, different rules and criteria have been developed. The main issues in hypothesis testing and its difficulties when generalised to multiple testing are briefly introduced in this chapter.

1.1.1 Hypothesis testing

In the 20th century, Egon Pearson and Jerzy Neyman made extensive contributions to the formalisation and development of statistical methods in hypothesis testing. As an essential topic in statistical inference, the purpose of hypothesis testing is to make the statistical decision of whether the prior hypothesis has been correctly formulated. This decision procedure is based on the current observations x of certain random variables X, of which the distribution P_θ is assumed to be varying in a class $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$. The null hypothesis H_0 is often formulated to claim some property of P_θ that is supposed to be verified, and an alternative hypothesis H_1 is involved when we decide between a favoured statement and another. This divides the parameter space Θ into Θ_0 and Θ_1 under the null and the alternative respectively. A choice of rejecting or not rejecting H_0 is made given the data.

Since the decisions depend on the value x of X, the sample space is divided into two regions, the rejection region $R_{(\cdot)}$ and the acceptance region $R^c_{(\cdot)}$, associated with a specified decision rule. We can define a critical function $\phi(x)$ of x to represent the probability of rejecting the null at the value x, and $1 - \phi(x)$ stands for the probability of not rejecting. In a randomised test, the probability $0 \leq \phi(x) \leq 1$ for all x, while in a nonrandomised test the critical function is reduced to an indicator of the critical region with $\phi(x) = 1$ or 0.

There are two types of errors when the decision is made. A type I error occurs when the null hypothesis is rejected but it is indeed true, while a type II error occurs when the null hypothesis is not rejected but is indeed false.

		True state	
		H_0 true	H_0 false
Decision	Reject H_0	*Type I error*	
	Not reject H_0		*Type II error*

The two types of errors are both worth investigating and they are obviously not equivalent. Unfortunately, it is not possible to minimise the probabilities of both simultaneously. Usually a bound $\alpha \in (0,1)$ is assigned to control the probability of type I error, and this threshold is called the significance level of the test. This value can be customised to the tolerance of false rejections. The rejection of null hypothesis H_0 is often referred to as a *discovery*. In this case the procedure is to minimise the probability of type II error subject to a pre-determined control of the probability of type I error

$$\mathbb{P}_\theta\{X \in R_\phi\} = \mathbb{E}_\theta \phi(X) = \int \phi(x) \mathrm{d}P_\theta(x) \leq \alpha \quad \text{for all } \theta \in \Theta_0 .$$

This is because we want the null hypothesis to be rejected carefully, and once we claim a rejection, the probability of the conclusion being wrong is bounded by α.

On the other hand, the type II error is closely related to the ability to detect false nulls. Minimising the type II error is equivalent to maximising the statistical *power* against the alternative H_1, which is for each $\theta \in \Theta_1$,

$$\mathbb{P}_\theta\{X \in R^c_\phi\} = \mathbb{E}_\theta \phi(X) = \int \phi(x) \mathrm{d}P_\theta(x),$$

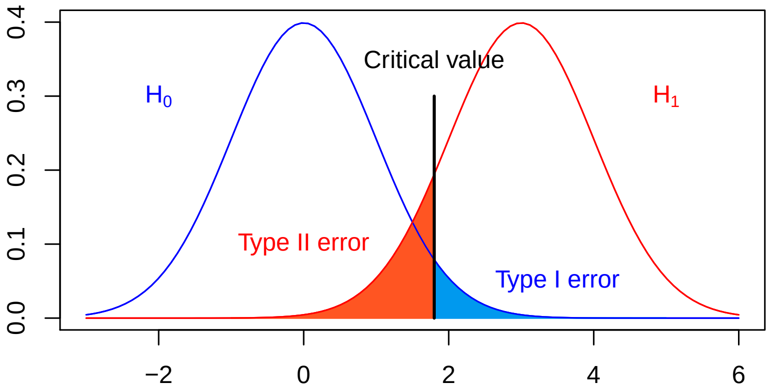


Figure 1.1 – Testing the shifted mean

i.e., the probability of rejection when the null hypothesis is false.

The main issues in hypothesis testing can be illustrated with the following example. When a new treatment is tested in a clinical trial, one must be very cautious to claim that it has a better effect than the existing methods. The decision is made between the null hypothesis ≪no improvement≫ and the alternative hypothesis ≪positive effect≫. Suppose the measurements follow a Gaussian $\mathcal{N}(\mu, 1)$ distribution, where μ is the increment in the effect of the new treatment. Mathematically the problem is to test

$$H_0 : \mu = 0 \quad \text{against} \quad H_1 : \mu > 0 \,.$$

Not rejecting the null hypothesis means no significant improvement is detected, while rejecting the null hypothesis leads to a positive discovery which may have a great impact in follow-up studies. A false rejection here means the researchers declare a significant improvement but it is indeed wrong. The belief in this new treatment may cause an abuse while it is not improving the result in reality, or even worse, is harmful to some patients. This situation is assumed to be limited by controlling the probability of type I error. Subject to the threshold of the probability of a false discovery, maximising the power results in the optimality of detecting the significant improvement of the new treatment. When the power is high, it implies that one will not miss a promising treatment when it is indeed a better solution.

The rejection region can also be based on the p-value of the test, which is defined as the probability that the current observation or more extreme observations occur when H_0 is true. In other words, it reflects the probability of H_0 being true by measuring how much the data contradict the null hypothesis. Suppose H_0 is true and the observations are the realisations of the random variable under H_0. If the p-value is below a pre-specified threshold, for example $p < 0.05$, it implies that the current observations or even more extreme ones are rare, while the fact that they do exist indicates the erroneousness of H_0.

1.1.2 Early development of simultaneous inference

Multiple testing is a subfield of simultaneous inference, which includes multiple comparison and estimation as well as testing. In the early-stage research of simultaneous testing, multiplicity correction is not involved and the principal ideas are not formulated mathematically. Fisher (1935) proposed the first method to do simultaneous tests, but it was not named in terms of a multiple testing methodology. The initial proposal for making multiple inferences was formulated by Tukey (1953). The theories and methodologies in this stage are developed mainly to provide multiple conclusions simultaneously, and it is necessary to associate with each conclusion a statistical measure of confidence. We refer to the book by Miller (1966) as a good summary of the early development of simultaneous inference.

Fisher's least significant difference test

Fisher (1935) proposed a test that locates the significant effects in the analysis of variance, which is considered as a predecessor of the multiple stage tests. The null hypothesis

$$H_0 : \mu_1 = \cdots = \mu_m, \tag{1.1}$$

can arise in any application of the normal linear model, where μ_i is the mean of the i-th population. Suppose $\{X_{ij}, j = 1, \dots, n_i\}$ is the sample from the i-th population, with $i = 1, \dots, m$ and $\sum_{i=1}^{m} n_i = N$. Denote the sample mean and the pooled variance

$$\bar{X}_{\cdot\cdot} = \frac{1}{N}\sum_{i,j} X_{ij}, \quad \bar{X}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}, \quad i = 1, \dots, m, \quad \text{and} \quad S^2 = \frac{1}{N-m}\sum_{i,j}(X_{ij} - \bar{X}_{i\cdot})^2.$$

In order to test the overall null hypothesis (1.1) in the analysis of variance, one will apply an F test which compares the test statistic

$$\frac{\sum_i (\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2/(m-1)}{\sum_{i,j}(X_{ij} - \bar{X}_{i\cdot})^2/(N-m)} \tag{1.2}$$

to the α quantile of the $F_{m-1,N-m}$ distribution. If the F value does reject the overall null hypothesis, one will apply a follow-up test to locate the significantly different pairs. For each pairwise mean comparison between μ_i and $\mu_{i'}$ considered in (1.1), a t-test at level α is utilised and the component hypothesis $\mu_i = \mu_{i'}$ is rejected if the t-value

$$\frac{\bar{X}_{i\cdot} - \bar{X}_{i'\cdot}}{S/\sqrt{\frac{n_i n_{i'}}{n_i + n_{i'}}}}, \quad i \neq i' \tag{1.3}$$

exceeds the $\alpha/2$ quantile of the t_{N-m} distribution. This is a two-stage test that locates the significant means with the overall significance guaranteed by the F test.

Tukey's studentized range test

Tukey (1953) provided a studentized range test to make comparisons of the means of the measurements in one-way classification with the overall erroneousness being controlled. Suppose $\{X_{ij},\ i = 1,\dots,m,\ j = 1,\dots,n\}$ are m independent balanced samples, each of which is a sample of n independent normal random variables with common mean and variance. Let μ_i be the mean of the i-th sample. The problem of testing the hypothesis (1.1) is equivalent to testing

$$H_0 : \mu_i = \mu_{i'} \quad \text{for all} \quad i \neq i'\,. \tag{1.4}$$

In order to test the difference between a single pair μ_i and $\mu_{i'}$, the test statistic is based on $\left|(\bar{X}_{i\cdot} - \bar{X}_{i'\cdot}) - (\mu_i - \mu_{i'})\right|$. Thus, for any paired difference $\mu_i - \mu_{i'}$, $i \neq i'$ to be bounded by a threshold, it is natural to consider that the following inequality

$$\frac{\max_{i,i'}\left\{\left|(\bar{X}_{i\cdot} - \bar{X}_{i'\cdot}) - (\mu_i - \mu_{i'})\right|\right\}}{S/\sqrt{n}} \leq c \tag{1.5}$$

holds under the null hypothesis. Notice that the numerator is the range of m independent $N(0,\sigma^2)$ random variables, and the sample variance S in the denominator is the square root of the pooled variance based on the whole sample. The test statistic is compared to the upper α quantile of the studentized range distribution with the parameters m, $m(n-1)$. This is called a Tukey's studentized range test, and sometimes also called a wholly significant difference (WSD) test, or Tukey's honestly significant difference (HSD) test.

Other related literature

The work of Dunnett (1955, 1964) and Duncan (1955) and other statisticians in the fifties and sixties also contributed to the field of simultaneous testing and multiple comparison. Scheffé developed a procedure based on the F-test in the analysis of variance to derive further conclusions for any possible contrasts $\sum_{i=1}^{m} c_i\mu_i$. Dunnett's method was aimed at comparing multiple treatment means with a control mean under the setting of equal sample sizes and equal variances. Duncan's work was more associated with the multiple stage tests first proposed by Newman (1939). Duncan used an α_p-level studentized range test in a follow-up study after the overall null hypothesis (1.1) is rejected, where $p = 2,3,\dots,m$ is the number of the means in a subset of interest. To compare the p means, the level of the studentized range test is chosen to be $\alpha_p = 1-(1-\alpha)^{p-1}$.

Remark. *Note that the concepts of multiple testing and multiple comparison are sometimes used interchangeably, while for statisticians, there are noticeable distinctions between the principals of multiple testing and multiple comparison. In multiple comparisons, the main goal is to compare the means of different groups, while in multiple testing, one is more interested* *in inference based on pre-specified null and alternative hypotheses.*

Some applications

In studies of clinical trials, genomics and neural networks, it is crucial to test whether each explanatory variable has an impact on the responses. The selected variables are often used in a follow-up study. For example, when comparing several treatments against a control case in a clinical study, each treatment generates a test of ≪no difference≫ against ≪improvement≫, namely

$$H_{0,i} : \mu_i = 0 \quad \text{against} \quad H_{1,i} : \mu_i > 0\,,$$

where μ_i denotes the mean response to the i-th treatment. The overall null hypothesis

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_m = 0\,,$$

which is the intersection of all the nulls, states that no significant improvement is detected among all the treatments. When the overall null hypothesis is accepted, the study is complete and no more information is required. If the overall null hypothesis is rejected, one would like to identify which treatments have a significant improvement.

On the other hand, one can also compare multiple treatments instead of testing their improvement over a control case. For example, with the null hypotheses being

$$H_{i,j} : \mu_i = \mu_j\,, \quad 1 \le i < j \le m\,,$$

a maximum of $\binom{m}{2}$ tests can be performed to compare each pair of treatments.

Another typical application of multiple testing and comparison is in genetics or genomics, where thousands of genes, or even more, are tested simultaneously. Nowadays the biotechnology such as the Next Generation Sequencing (NGS) allows us to tackle high-throughput data. Large-scale testing is often carried out at one time, and it may turn out that only a few genes among thousands of candidates are of interest.

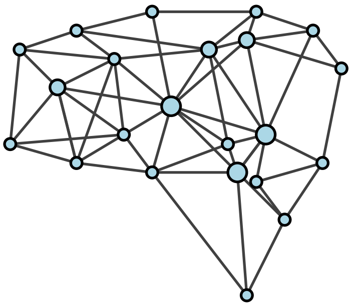
We take the study of differential gene expression as an introductory example, where one would like to test whether or not there is a difference in gene expression between the control subjects and the patients with a hereditary cancer. Suppose we receive the independent Gaussian measurements $\{X_{i,j}, Y_{i,j}, i = 1,\ldots,m, j = 1,\ldots,n\}$ for the levels of gene expressions of the two groups. We focus on testing the family of hypotheses

$$H_{0,i} : \mu_i^X = \mu_i^Y\,, \quad i = 1,\ldots,m\,,$$

where rejecting a null hypothesis $H_{0,i}$ indicates that for this gene i, the expression is significantly different between the control group and cancer group. We know that for a single gene i, the classical method provides a two-sample Student t-test between two samples of measurements $\{X_{i,j}, Y_{i,j}\}_{j=1}^n$ at significance level α. When considering multiple genes, for example $m = 5,000$, a choice of $\alpha = 0.05$ allows on average 250 false rejections if all the null hypotheses are true. This conclusion can be very misleading in the sense of causality and lead to a waste of time and resources.

Similar multiple testing studies also occur in neural networks, where a large number of tests are generated according to the architecture of the network. One can investigate the function or the structure of the different regions in the brain by analysing the measurements. For example, in network data analysis based on neuro-imaging, we seek to formalise the statistical properties based on the network data, and derive the principal features from the graphical model. These statistics will be used to compare different networks and detect the factors that most influence the networks.

Mathematically one can conclude the principal properties in a graph $G = (V, E)$, where V and E are the set of vertices and edges respectively. Using graphical methods one can test



for either the local properties reflected by the statistics related to the nodes and edges, such as connected components, triangles, and other higher-order structures. On the other hand, one can also test for the global structure such as the depth or the shortest-path length of the graph. The testing procedures are usually aimed at determining the critical features of a network, or comparing one network to another. In real applications, for example, multiple testing procedures are applied to investigate the connectivity between different regions of the brain or to compare the levels of gene expression measured for different subjects. Topological and geometric data analysis will also be involved in this field.

1.1.3 Multiplicity adjustment

We see from the previous sections that in multiple testing, the set of conclusions are often regarded as a whole instead of being evaluated based on each single hypothesis. Multiplicity adjustment is needed if the inferences are made simultaneously and the errors are considered jointly. The control of the erroneousness per test does not guarantee the equivalent control of wrong detections over the family of hypotheses. A standard choice of significance level $\alpha = 0.05$ or 0.01 may allow a number of false discoveries that is beyond the tolerance when considering the whole family of tests.

Consider the case when we are to test simultaneously a finite number of hypotheses $\{H_{0,i}\}_{i=1}^{m}$, of which m_0 are actually true. An intuitive idea is to test each hypothesis separately using a pre-determined significance level α. Suppose some of the null hypotheses are rejected, the mean value of the false rejections is

$$\begin{aligned}\mathbb{E}(\#\{\text{False rejections}\}) &= \sum_{i:\, H_{0,i} \text{ is true}} \mathbb{P}_{H_{0,i}}(H_{0,i} \text{ is rejected at level } \alpha) \\ &= \#\{i : H_{0,i} \text{ is true}\} \cdot \alpha \\ &= m_0 \alpha\,.\end{aligned}$$

Since the number of true null hypotheses is unknown, the number of falsely rejected nulls can be quite large regardless of the combination of true and false hypotheses. On the other hand, the probability of making at least one false rejection increases dramatically with m:

$$\begin{aligned}\mathbb{P}(\text{at least one false rejection}) &= 1 - \mathbb{P}(\text{no false rejection}) \\ &= 1 - \bigcap_{i=1}^{m} \{H_{0,i} \text{ is not falsely rejected at level } \alpha\} = 1 - (1-\alpha)^m\end{aligned}$$

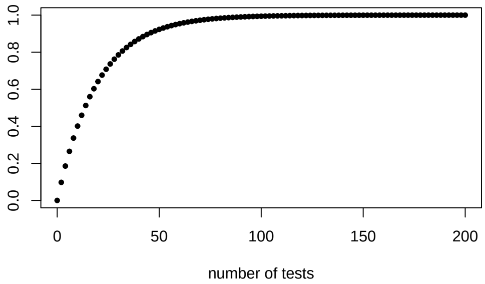


Figure 1.2 – The probability of making at least one false rejection

Figure 1.2 shows how the probability of making at least one false rejection among m tests grows with m, and we can see that, for example, this probability is already above 0.9 when we test 50 hypotheses simultaneously. A conclusion with a controlled type I error less than 0.05 per test is very misleading in terms of the erroneousness in testing the whole family. This phenomenon indicates that the level α needs to be adjusted.

In order to control the global probability of committing false rejections, the Bonferroni correction tests each null hypothesis at a reduced significance level α/m, and has been proved to control the family-wise error rate (FWER), that is, the probability of having at least one false rejection at the level α. However, the Bonferroni correction is known to be too conservative. For example, with 50 tests and the overall significance level $\alpha = 0.05$, the Bonferroni correction only rejects the null hypothesis if the p-value is less than 0.001. In reality m can go beyond thousands, and such a strict rule of rejection may lead to a high rate of false negatives, that is, a low power for detecting the true alternative hypotheses. But this method is still widely used in the confirmatory analyses, for example in clinical trials, when a single false positive can cause fatal consequences, so the researcher does not want any misleading discoveries. Holm (1979) developed the Bonferroni correction with slight modification. The Holm's procedure compares the p-values to an increasing threshold sequence

$$\frac{\alpha}{m}, \frac{\alpha}{m-1}, \cdots, \frac{\alpha}{1}$$

and is proved to control the FWER as well.

In some exploratory analyses where the discoveries reported from the tests will be examined in more detailed follow-up studies, the aim of controlling the type I error is not limited to avoiding any single false discovery, but rather controlling the proportion of false discoveries. Benjamini and Hochberg (1995) invented a step-wise procedure that controls a different criterion. The false discovery rate (FDR) is defined as

$$\mathrm{FDR} = \mathbb{E}\left(\frac{\#\{\text{False rejections}\}}{\#\{\text{Total rejections}\}}\right),$$

which is the expected proportion of false rejections among all rejections. The BH procedure tests the null hypotheses $H_{0,(1)}, H_{0,(2)}, \dots, H_{0,(m)}$ following the increasing order of the observed p-values

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}.$$

The ordered p-values are compared to the sequence of thresholds

$$\frac{\alpha}{m}, \frac{2\alpha}{m}, \dots, \alpha,$$

which is linear in $i = 1, 2, \dots, m$. It rejects the null hypothesis for the p-value $p_{(i)} \leq i\frac{\alpha}{m}$ and stops at the last crossing point. It is known that the Benjamini–Hochberg procedure controls the false discovery rate if all the test statistics are independent. This method is widely used since it provides control of error rate for a large-scale multiple testing study with promising power. Benjamini and Yekutieli (2001) improved the step-wise procedures under positive dependence. Much literature based on BH procedure and the control of FDR and its alternatives has appeared in the last decades.

Multiplicity correction is considered as the main issue in the field of multiple testing. Useful techniques to control error rates will be systematically explained in Chapter 2.

1.1.4 Weak and strong control of error rates

One of the main issues in multiple testing is to quantitatively control the erroneousness. Different criteria are defined and adopted depending on the setting. In this section we propose an intuitive illustration, of which a mathematical formalisation can be found in Chapter 2.

The control of the type I error rate only when all the null hypotheses in the family are true is referred to as *weak control*, for example the *experiment-wise error rate* control. It is in practice not interesting because it does not provide equivalent control of the rate of making an incorrect decision. For many statistical testing methods, the maximum probability of a false rejection occurs when only some of the null hypotheses are true, instead of all of the null hypotheses being true.

Most literature concentrates on the multiple testing procedures with *strong control*, that is, the control of type I error rate regardless the configuration of true and false null hypotheses. Apart from the two major choices, the FWER and the FDR mentioned above, there are alternative formulations of error rate considered in multiple testing. The *per-comparison error rate (PCER)* is also referred to as *error rate per hypothesis*, of which the average

$$\text{PCER} = \frac{\#\{\text{False rejections}\}}{\#\{\text{Total hypotheses}\}}$$

can be used as the expected value of the number of false discoveries over the total number of hypotheses. This criterion ignores the multiplicity correction in the sense that the PCER is controlled at α if each individual hypothesis is tested at level α. The *error rate per family (PFE)* is defined as the expected number of false discoveries in the family. The *marginal false discovery rate (mFDR)* is defined as

$$\text{mFDR} = \frac{\mathbb{E}(\#\{\text{False rejections}\})}{\mathbb{E}(\#\{\text{Total rejections}\})},$$

which is the ratio of the expected number of false rejections over the expected number of total rejections. With some assumptions this can be proved to be equivalent to FDR control. The *false discovery proportion (FDP)* is defined as the ratio of the number of false rejections to total rejections. Although it is ideal that the FDP is controlled at each realisation, this is impossible due to randomness.

1.2 Specific principles and topics in multiple testing

Problems of multiple testing and multiple comparison are widely discussed in many related disciplines, such as clinical design, genetics and genomics, where the decisions are made over multiple measurements. In recent decades, sub-fields in multiple testing driven by real data applications have become more and more influential in simultaneous inference. In this section we introduce some principles of high attention and some specific topics which have been developing rapidly.

1.2.1 Hierarchical problems: closed testing and partition testing

One of the most important branches of multiple testing problems is hierarchical testing, which originally arises in pharmaceutical studies where there are multiple doses and endpoints to be examined. A multiple-dose study is aimed at finding the exact dose such that the drug is both effective and safe. This problem is logically hierarchical when the responses to multiple doses, for example, the means of responses $\mu_1, \ldots, \mu_m$ corresponding to m increasingly ordered doses, are tested simultaneously.

Multiple-endpoint studies attempt to make significant conclusions on multiple endpoints of a clinical study. A primary endpoint is an endpoint in a clinical study such that the significant discovery on that particular endpoint alone is sufficient to make a conclusion for the whole study. A secondary endpoint is a clinical endpoint such that a significant finding on that endpoint alone is insufficient to make a conclusion for the whole study. The experiment is considered as a hierarchical problem, as the secondary endpoints are to be examined when there are no significant primary endpoints. The multiplicity correction has to be customised according to the experimental design.

In hierarchical testing problems as described above, the selection of dose and the decision path are of the highest importance. Because it is often desired to control the exact occurrence of false discoveries, the FWER is favoured instead of the FDR. The most powerful methods that control the FWER are the closed testing and the partition testing procedures.

Closed testing

Closed testing devises stepwise multiple testing procedures with fixed experiment-wise error. Fisher (1935), Tukey (1953) and Hartley (1955) had similar ideas before Marcus et al. (1976) first formulated the closed testing principle. He pointed out the essential feature of the closed testing procedures is that the sets of the hypotheses are closed under arbitrary intersection, that is, any subset intersection hypothesis involving members of the family of tests is also a member of the family.

Consider the problem of testing a group of null hypotheses $\{H_i, i \in I\}$. The primary hypotheses that do not imply the truth of any other hypothesis are called *minimal hypotheses.* All the possible intersections among the lower hypotheses are called *composite hypotheses.* Define the family of intersection hypotheses

$$\left\{H_J = \cap_{i \in J} H_i, J \subset I\right\} .$$

For example, given the minimal hypothesis $H_{ij} : \mu_i = \mu_j$, the intersection hypothesis $H_{ijk} = H_{ij} \cap H_{ik} : \mu_i = \mu_j = \mu_k$ is a composite hypothesis that is above H_{ij}, H_{ik} and H_{jk}. Each test is performed at the local significance level α, which equals the required level of the overall test.

A framework of closed testing is as follows.

Step 1. Test each minimal hypothesis at level α.

Step 2. Test each intersection hypothesis at level α.

Step 3. Conclusion.

Any hypothesis may be rejected when both the following conditions hold:

i) the test itself is significant;

ii) any intersection hypothesis that includes the test is rejected.

The closed testing procedure controls the family-wise error rate in a strong sense and is attractive because of its high power. However, some directional errors may occur in two-sided testing problems, which is difficult to solve in hierarchical testing. On the other hand, the design of the test procedure can be complex, because the number of intersection hypotheses, $2^m - 1$, increases rapidly as the problem grows.

Partition testing

For the family $\{H_i, i \in I\}$ and the corresponding parameter space $\Theta = \cup_{i\in I}\Theta_i$, the partition principle controls the FWER by partitioning the entire null space Θ into disjoint subspaces $\left\{\Theta_J^*, J \subset I\right\}$, such that

$$\{\Theta_J^* = \cap_{j\in J}\Theta_j \cap \left(\cup_{i\notin J}\Theta_i\right)^c, J \subset I\}$$

forms a partition of $\cup_{i\in I}\Theta_i$, given the fact that for any $J, K \subset I$,

$$\Theta_J^* \cap \Theta_K^* = \left(\cap_{i\in J}\Theta_i\right) \cap (\cap_{i\in K}\Theta_i) \cap \left(\cup_{i\in J}\Theta_i\right)^c \cap \left(\cup_{i\in J}\Theta_i\right)^c \cap \left(\cup_{i\notin J\cup K}\Theta_i\right)^c = \varnothing.$$

Because the all hypotheses based on $\left\{\Theta_J^*, J \subset I\right\}$ are disjoint, at most one of the nulls is true, which will naturally guarantee the probability of a false rejection without a multiplicity correction. The overall FWER is controlled at level α as long as each subspace is tested at level α. The partitioning principle was developed by Stefansson et al. (1988), and later on investigated by Finner and Strassburger (2002).

The advantage of testing the partitioned hypotheses is that it reduces the number of hypotheses to be tested compared to the intersection hypotheses generated in the general closed testing procedures. The partitioning also helps to decide a path of testing as well as increases the power, given the background knowledge that the null hypotheses of interest are hierarchically ordered. We recommend Shaffer (1995) for further detail.

1.2.2 Online testing

The problem of online multiple testing arises in some applications where it is permissible to have infinite number of hypotheses. Compared to classical off-line problems, online procedures make decisions immediately at the each stage instead of after receiving all the test statistics. This could happen in many applications where the studies and the corresponding tests are carried out sequentially. On the other hand, the scale of the tests could reach tens of thousands, which is large enough to be considered as infinite. Multiplicity corrections based on the total number of hypotheses are not applicable since this quantity is unknown before the termination of the whole procedure.

Suppose one is testing a sequence of hypotheses

$$\mathcal{H} = \{H_1, H_2, \ldots\}$$

of which the total number can be infinite. Up to the m-th test, define the set of the received hypotheses $\mathcal{H}(j) = \left\{H_1, H_2, \ldots, H_j\right\}$, and let $R(j) = \left\{R_1, R_2, \ldots, R_j\right\}$ denote the decisions for each test, of which the components are the indicators $R_j = \mathbb{1}\{\text{reject } H_j\}$. An α-spending procedure begins with an initial α-wealth $W(0)$, which is an allowance for type I error for the whole family $\mathcal{H}$, and performs each test H_j at α_j. At each step, the total α-wealth is reduced by α_j if H_j is rejected. No further test is conducted once the remaining α-wealth reaches 0.

Tukey (1991) had this idea of α-spending on sequential tests. Foster and Stine (2008) modified this online testing method and proposed an α-investing rule. Given the results $\left\{R_1, R_2, \ldots, R_{j-1}\right\}$, the level for testing H_j is decided by an investing rule $\mathcal{I}_{W(0)}$ such that

$$\alpha_j = \mathcal{I}_{W(0)}\left(\left\{R_1, R_2, \ldots, R_{j-1}\right\}\right).$$

The test for H_j is based on the p-value p_j. When H_j is accepted, it costs $\alpha_j / \left(1 - \alpha_j\right)$, which is called a *pay-off.* When H_j is rejected, the procedure earns a *pay-out* $\omega \in (0,1)$ that is carried on to the next tests. The change in the α-wealth is then concluded by

$$W(j) - W(j-1) = \begin{cases} \omega, & p_j \le \alpha_j, \\ -\alpha_j / \left(1 - \alpha_j\right), & p_j > \alpha_j. \end{cases}$$

One example for the α-investing rules is defined by

$$\mathcal{I}_{W(0)}\left(\left\{R_1, R_2, \ldots, R_{j-1}\right\}\right) = \frac{W(j-1)}{1 + j - k^*} \quad j > k^*,$$

where k^* denotes the index of the hypothesis last rejected before H_j. This rule spends half of the current α-wealth on the test right after a rejection, and the remaining wealth will soon decay to zero if significant alternatives occur consecutively.

In general, online methods are believed to be more sensitive to the order of the hypotheses compared to off-line testing procedures, and the spending rate can be designed according to prior knowledge. The methods are particularly powerful when the false null hypotheses come to the process in the first places and appear in clusters. The results of testing the past hypotheses will definitely influence the α-spending and the chance of rejecting the upcoming ones. In this case, post-selection inference is also combined with online testing in order to reduce the false rejections due to the outliers.

1.3 Outline

The rest of the thesis proceeds as follows. In Chapter 2 we set up the multiple testing problem for mixture models. Most of the literature works on Gaussian mixture models on the level of the original observations, the p-values or the z-values. The test statistics that measure the difference between the mixture and the pure null distribution are established and analysed. As a second step of locating the alternatives, an individual rejection is made when the observed p-value is less than a threshold.

In Chapter 3 we unveil the problem caused by the heavy-tailedness in multiple testing, and discuss this phenomenon from several perspectives. Since the Gaussian assumption is not always satisfied, it may occur that the test statistic has a heavy-tailed distribution, which will lead to incorrect assumptions about the distribution of the p-values under the alternatives. We explain this added difficulty in Section 3.2 and 3.3, and define an asymptotic framework for testing the global null hypothesis, where the fraction of true alternatives vanishes with m increasing, and the size of the non-null effects must be moderately large to be detected. In Section 3.4 we provide the asymptotic detection boundary for Cauchy mixture models based on the convergency of Kullback–Leibler divergence, which is the expected log-likelihood ratio statistic.

In Chapter 4 we propose a filtering method that works on the ordered p-values and their local concentration. When we are to identify the true alternatives among the mixture of null and non-null components, the whole literature of rejecting the smallest p-values will not be adapted to heavy-tailed testings. We give an illustration by investigating the positive FDR. In order to locate the true alternatives, we propose a filtering approach that eliminates the p-values that are more likely to be uniform. In a follow-up step we estimate the mode of the p-values left and construct a rejection region centralised at the mode, with the length to be decided by data-dependent FDR control.

Chapter 5 includes a discussion on robust testing and a summary of the whole text. In general, an ideal multiple testing procedure is desired to be robust to the tail index, multimodality, and to dependence. In order to characterise the p-value clusters that may occur in a finite-dimensional mixture model, we propose another method based on the p-value gap statistics and their local discrepancy, which is more data-dependent and adaptive. We also briefly discuss the critical conditions that influence the testing procedures for heavy-tailed and light-tailed distributions, and introduce the existing results on testing for multimodality.

2 Multiple testing for mixture models

The problem of detecting significant components appears in many applications, where among a large number of observations only a small proportion are informative. The significant components may come from, for example, the true signals in a communication system, a disease-causing mutation in the genome, the increments in the responses to a new drug treatment and so on. This chapter gives the problem set-up based on mixture models and discusses the methodologies for detecting and locating the non-null components.

2.1 Preliminaries

Given an independent sample $X_1, X_2, \ldots, X_m$ from a mixture model, identifying the significant components individually is a multiple testing problem in which we test a family of hypotheses

$$H_{0,i} \, : \, X_i \sim F_0 \quad \text{against} \quad H_{1,i} \, : \, X_i \sim F_1, \quad i = 1, \ldots, m,$$

simultaneously at a given significance level, where F_0 and F_1 are the distributions of the observations under the null and alternative hypothesis respectively. We are interested in this multiple testing problem because it is difficult to control the increased type I error of testing a family of hypotheses.

In most cases the proportion of alternatives is unknown, and it is even a problem to tell whether it is possible to detect them. Thus it is reasonable to begin with the detection of the existence of non-null components.

2.1.1 The overall null hypothesis

First we focus on when to reject the global intersection null hypothesis, that is to say, we test the existence of a fraction of the sample from the alternative distribution F_1 against the joint null hypothesis $H_0^{(m)}$ that all the observations are i.i.d. from the null distribution F_0, which is equivalent to testing

$$H_0^{(m)} : X_i \overset{\text{i.i.d.}}{\sim} F_0 \quad \text{against} \quad H_1^{(m)} : X_i \overset{\text{i.i.d.}}{\sim} (1-\varepsilon)F_0 + \varepsilon F_1,$$

where ε is the proportion of the significant components.

The random effects model is convenient, although not necessary, to interpret the mixture model in multiple testing, with normality of the test statistics assumed. Consider the random effects model $X_i = \mu_i + z_i, \quad i = 1,\dots,m,$ where $z_i \overset{\text{i.i.d.}}{\sim} N(0,1)$ for $1 \le i \le m$ are white noise. The effects μ_i are supposed to have different distributions under the null and the alternative hypotheses, of which the convolution with a standard normal distribution will give the distributions of X_i under $H_{0,i}$ and $H_{1,i}$ respectively.

As a motivating example, we consider the mixture model where the null effects occur with the probability $\mathbb{P}(\mu_i = 0) = 1-\varepsilon$, and for the non-null effects $\mu_i \neq 0$, we assume $\mu_i \sim H$, which is a distribution concentrated at μ. This leads to testing the global null distribution against the two-point Gaussian mixture

$$H_0^{(m)} : X_i \overset{\text{i.i.d.}}{\sim} N(0,1) \quad \text{against} \quad H_1^{(m)} : X_i \overset{\text{i.i.d.}}{\sim} (1-\varepsilon)N(0,1) + \varepsilon N(\mu,1), \tag{2.1}$$

which is the core model that most literature in multiple testing works with.

2.1.2 From test statistics to p-values and z-values

Here we make a remark that our work will be mainly based on the level of the p-values.

For a significance level $\alpha_i \in (0,1)$, define the nested rejection region $\mathcal{R}_{\alpha_i}$ for the value of the test statistic X_i such that

$$\mathbb{P}_{H_{0,i}}(x_i \in \mathcal{R}_{\alpha_i}) = \alpha_i\,.$$

Since the p-value is defined as the smallest value of this significance level, that is,

$$p(x) = \inf_{\alpha_i \in (0,1)} \{x_i \in \mathcal{R}_{\alpha_i}\},$$

a threshold $u \in (0,1)$ for the p-value is equivalent to the rejection region $\mathcal{R}_u$ for x_i, namely

$$\mathbb{P}_{H_{0,i}}(p(x_i) \le u) = \mathbb{P}_{H_{0,i}}(x_i \in \mathcal{R}_u) = u\,.$$

It is straightforward to see the link between the test statistics and the p-values.

Now we introduce the definition of the z-values as follows. Suppose for the test statistics $X_1,\dots,X_m$, the p-values $P_1,\dots,P_m$ are computed from $P_i = \mathbb{P}_{H_{0,i}}(X_i > x_i)$.

Definition 2.1.1 (z-values)**.** *Define the z-values*

$$z_i = \Phi^{-1}(p_i)\,, \quad i = 1,\dots,m \tag{2.2}$$

where $p_1, \dots, p_m$ are the p-values, and Φ is the cumulative distribution function of the standard normal.

Under the null hypothesis $H_0^{(m)} : X_i \overset{\text{i.i.d.}}{\sim} N(0,1)$, the z-values will have a standard normal distribution

$$H_{0,i} : z_i \sim N(0,1)\,, \quad i = 1, \dots, m\,.$$

Efron and Storey made important contributions on multiple testing procedures based on the z-values, of which the relavent part will be reviewed in Section 2. Usually it is agreed that the z-values offer the chance to pursue an improvement in power which benefits from the normal theory.

Both the p-values and the z-values are convenient to work with, and it is up to the statisticians to choose a suitable one to investigate.

2.2 Testing the overall null hypothesis

The problem of testing the frequency $\varepsilon = 0$ against $\varepsilon > 0$ in the mixture model described above was first studied by Ingster (1998). He discussed the theory and methods using the likelihood ratio test when μ and ε are known.

In more recent literature, the statisticians investigated various statistics that measure the departure of the sample distribution from the theoretical model, namely the uniform distribution if the test is based on the p-values. The existence of the alternative components is summarised by an informative test statistic based on the whole sample.

On the other hand, it is also welcomed if the proportion ε can be directly inferred, or at least an approximate lower bound is derived from the observations. The test (2.1) is therefore based on the estimate $\hat{\varepsilon}$.

2.2.1 Comparing the p-values to the uniform

Tukey (1976) proposed the *second-level significance test*, based on the *Higher Criticism (HC)* test statistic

$$\mathrm{HC}_{\alpha,m} = \frac{\sqrt{m}[(\text{ Fraction Significant at } \alpha) - \alpha]}{\sqrt{\alpha(1-\alpha)}}\,. \tag{2.3}$$

The "fraction significant at α" is the proportion of the detected non-null components. This quantity is simply

$$\text{Fraction Significant at } \alpha = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{p_i \le \alpha\}\,, \tag{2.4}$$

given a decision rule based on the p-values and the significance level α. This value is comparable to α itself when all the nulls are true and the rejections are derived randomly. Therefore, the conclusions as the result of a first-order significant test could be equally bad as random rejections if a distinction between the standardised values of (2.4) and α is not evident. Tukey's argument considers the hypotheses jointly and admits the rejection of the overall null hypothesis (2.1) only when $\mathrm{HC}_{\alpha,m}$ exceeds a critical bound.

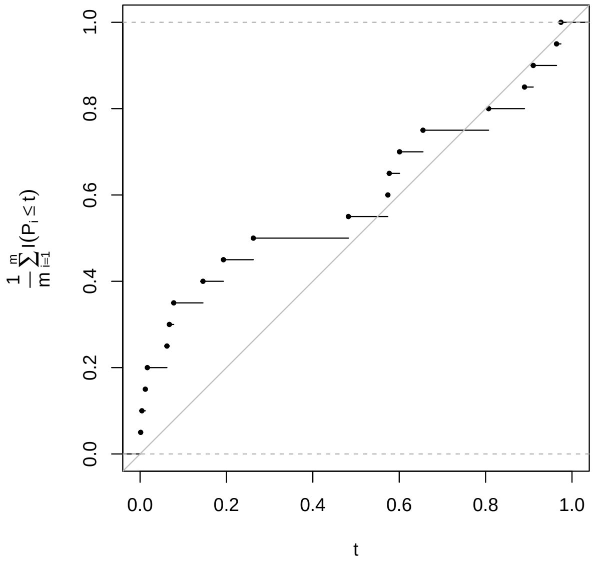


Figure 2.1 – Empirical c.d.f. of the p-values from the Gaussian mixture model

The higher criticism test is a standardised Kolmogorov-Smirnov test, of which the test statistic (2.3) is a measurement on how much the distribution of the observed p-values lies away from the uniform. Under the joint null hypotheses where all the observations are from the same null distribution, the p-values defined as $P(X_i) = \mathrm{P}(N(0,1) > X_i)$ are independent and identically distributed uniform random variables. When the overall null hypothesis is not true, the departure of the observed p-values from the uniform distribution becomes a quantity that measures the discrepancy between $H_0^{(m)}$ and $H_1^{(m)}$.

In a more detailed computation where we desire to compare the p-values to the uniformly distributed random variables $U_1, U_2, \ldots, U_m \overset{\text{i.i.d.}}{\sim} U(0,1)$, one can define the empirical cumulative distribution function

$$F_m(t) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{U_i \le t\},$$

and let $F_m(t)$ also stand for $\frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{P_i \le t\}$ under $H_0^{(m)}$. Figure 2.1 shows the empirical cumulative distribution function of a sample from the Gaussian mixture model with the frequency of non-zero effects $\varepsilon = 0.1$ and the positive effect $\mu = 2$. A noticeable distinction between the sample distribution and the uniform distribution is observed in this plot.

Since the normalized *uniform empirical process*

$$W_m(t) = \frac{\sqrt{m}\,[F_m(t) - t]}{\sqrt{t(1-t)}}$$

is asymptotically $N(0,1)$ distributed (see Shorack and Wellner (1986)), it guarantees the asymptotic properties of the test statistics based on the higher criticism. Tukey proposed a table of critical values to reject the overall null hypothesis (2.1).

Donoho and Jin (2004) proposed a modified version of the HC statistic

$$\mathrm{HC}_m^* = \max_{1\le i\le \alpha_0 m} \frac{\sqrt{m}\left[\frac{i}{m} - p_{(i)}\right]}{\sqrt{p_{(i)}\left(1-p_{(i)}\right)}}, \tag{2.5}$$

whose distribution is given by

$$\mathrm{HC}_m^* \overset{d}{=} \max_{1\le t\le \alpha_0} \frac{\sqrt{m}\left[\frac{\sum_{i=1}^m \mathbb{1}\{p_{(i)}\le t\}}{m} - t\right]}{\sqrt{t(1-t)}} = \max_{1\le t\le \alpha_0} W_m(t)$$

under the null $H_0^{(m)}$, and $\alpha_0 \in (0,1)$ is a fixed level. Since the distribution of $\max_{1\le t\le\alpha_0} W_m(t)$ has an iterated logarithm law, this method rejects $H_0^{(m)}$ when HC_m^* exceeds the critical value $h(n,\alpha_n) = \sqrt{2\log\log(n)}(1+o(1))$. Donoho and Jin proved that this test also has full power in the asymptotic case, as achieved by the likelihood ratio test.

Remark. *The power optimality is only achieved under certain restrictions. With the null and alternative distributions properly parametrised, the detection boundary is defined in the parameter space to separate the regions where it is possible or not to test the presence of non-null signals. We will propose our related work on the asymptotically detectable boundary in Chapter 3.*

2.2.2 Alternative methods for estimating ε and the null distribution

The mixture distribution of the p-values has been extensively analysed by Efron et al. (2001), Efron and Tibshirani (2002), Storey (2002), Genovese and Wasserman (2004), Meinshausen and Rice (2006), Jin and Cai (2007) and Cai et al. (2011), from the perspective of statistical inference. Their contributions are essential to the development of multiple testing based on the mixture model. Here we review the key ideas of their methods and explain why these methods cannot be directly applied to test heavy-tailed statistics.

Meinshausen and Rice's estimated proportion

Meinshausen and Rice (2006) established a lower $100(1-\alpha)\%$ confidence bound on the proportion of false null hypotheses based on the empirical distribution of the p-values. One can reject the intersection of null hypotheses when this bound is greater than zero. They proposed an estimate for ε,

$$\widehat{\varepsilon} = \sup_{t\in(0,1)} \frac{F_m(t) - t - \beta_{m,\alpha}\delta(t)}{1-t} \tag{2.6}$$

where $F_m(t)$, as defined before, is the empirical distribution function of the p-values. The bounding function $\delta(t)$ and the bounding sequence $\beta_{m,\alpha}$ are defined as follows.

A bounding function $\delta(t)$ is any real-valued function on $[0,1]$ that is strictly positive on $(0,1)$. For example, one can take a linear function $\delta(t) = t$, a constant $\delta(t) = 1$ or a standard deviation-proportional function $\delta(t) = \sqrt{t(1-t)}$. Let $U_m(t)$ be the empirical cumulative distribution function of m independent realisations of a random variable with uniform distribution on $[0,1]$. Define $V_{m,\delta}$ as the supremum of the weighted empirical distribution

$$V_{m,\delta} = \sup_{t\in(0,1)} \frac{U_m(t) - t}{\delta(t)}.$$

A series $\beta_{m,\alpha}$ is called a bounding sequence for a bounding function $\delta(t)$ if, for a constant level α the following two conditions are satisfied:

i) $m\beta_{m,\alpha}$ is monotonically increasing with m;

ii) $\mathbb{P}(V_{m,\delta} > \beta_{m,\alpha}) < \alpha$ for all m.

Here we explain how the bounding function and the bounding sequence help to estimate the proportion of true alternatives. The key idea behind this established estimator is again to compare the empirical distribution $F_m(t)$ to $U_m(t)$.

One will have the cumulative distribution function $U(t) = t$ for an exact uniform distribution, while the realised form $U_m(t)$ can frequently exceed t. Now we consider an enlarged bound $t + \beta_{m,\alpha}\delta(t)$, where the $\beta_{m,\alpha}$ and $\delta(t)$ are chosen such that the probability of $U_m(t)$ exceeding $t + \beta_{m,\alpha}\delta(t)$ can be upper bounded by α. This property is guaranteed by the condition ii), since

$$\begin{aligned}
\mathbb{P}(V_{m,\delta} > \beta_{m,\alpha}) < \alpha &\Longrightarrow \mathbb{P}\left(\sup_{t\in(0,1)} \frac{U_m(t)-t}{\delta(t)} > \beta_{m,\alpha}\right) < \alpha \\
&\Longrightarrow \mathbb{P}\left(\frac{U_m(t)-t}{\delta(t)} > \beta_{m,\alpha}\right) < \alpha \quad \text{simultaneously for all } t \\
&\Longrightarrow \mathbb{P}\big(U_m(t) > t + \beta_{m,\alpha}\delta(t)\big) < \alpha.
\end{aligned}$$

When comparing the behaviour of $F_m(t)$ to $U_m(t)$, notice that the function $F_m(t)$, that is, the frequency of the p-values less than or equal to t, exceeds the bound $t + \beta_{m,\alpha}\delta(t)$, is due to the non-null p-values. The difference $F_m(t) - t - \beta_{m,\alpha}\delta(t)$ is therefore considered as an estimate of ε.

With the estimator (2.10), it follows that

$$\mathbb{P}(\hat{\varepsilon} \le \varepsilon) \ge 1 - \alpha \tag{2.7}$$

under the overall null hypothesis. Therefore, one is able to conclude that for finite sample size m, the overall null hypothesis is rejected at level α when $\hat{\varepsilon} > 0$. In addition, the asymptotic control

$$\limsup_{m\to\infty} \mathbb{P}(\hat{\varepsilon} \le \varepsilon) \ge 1 - \alpha \tag{2.8}$$

is achieved with $\beta_{m,\alpha}$ properly chosen to be an asymptotic bounding sequence.

Cai and Jin's estimation of a two-point Gaussian mixture

Inspired by the possibility that the non-null proportion could be consistently estimated from the observations, Jiashun Jin and Tony Cai made a series of contributions to large-scale multiple testing and comparison based on Gaussian mixtures, including both the sparse case and the non-sparse case.

Cai et al. (2007) focused on the sparse two-point Gaussian mixture model, where the sparsity means the parameters μ and ε vary in a region such that the proportion of non-zero components is relatively small, and the values of the significant elements are large enough to be detected. They worked on the mixture model based on the observations rather than the p-values, and they developed a parametric approach to directly estimate the fraction ε.

Since there are two unknown parameters μ and ε in the mixture model

$$F(t) = (1-\varepsilon)\Phi(t) + \varepsilon\Phi(t-\mu),$$

one can determine the values of μ and ε precisely by solving the equations of $F(t)$, which are established on the realisations evaluated at $t = \tau$ and $t = \tau'$. Let

$$D(\mu;\tau,\tau') = \frac{\Phi(\tau) - F(\tau)}{\Phi(\tau') - F(\tau')} = \frac{\Phi(\tau) - \Phi(\tau-\mu)}{\Phi(\tau') - \Phi(\tau'-\mu)}, \tag{2.9}$$

and note that $D(\mu;\tau,\tau')$ is a monotone function of μ. Therefore, ε is determined by

$$\varepsilon = \frac{\Phi(\tau) - F(\tau)}{\Phi(\tau) - \Phi(\tau-\mu)}, \tag{2.10}$$

with μ solved from $D(\mu;\tau,\tau')$.

Intuitively, the estimates of μ and ε can be derived from the equations (2.9) and (2.10) with $F(\tau)$ and $F(\tau')$ replaced by estimates. Since the empirical estimates of $F(\tau)$ and $F(\tau')$ are influenced by the choices of τ and τ', Cai et al. (2007) proposed the slightly biased estimates

$$F^+(\tau) \ge F(\tau) \quad \text{and} \quad F^-(\tau') \le F(\tau'),$$

which leads to an estimator

$$\hat{\mu} \geq \mu ,$$

and therefore

$$\hat{\varepsilon} = \frac{\Phi(\tau) - F^{+}(\tau)}{\Phi(\tau) - \Phi(\tau - \hat{\mu})} \leq \frac{\Phi(\tau) - F(\tau)}{\Phi(\tau) - \Phi(\tau - \mu)} = \varepsilon .$$

The choices of $F^{+}(\tau)$ and $F^{-}(\tau')$ will be given below.

In practice, the knowledge of the distribution of

$$\sqrt{m} \frac{|F_m(t) - F(t)|}{\sqrt{F(t)(1 - F(t))}}$$

guarantees that, with a fixed ξ_m such that

$$\mathbb{P}\left(W_m^* = \sup_{t \in S_m} \left\{ \sqrt{m} \frac{|F_m(t) - F(t)|}{\sqrt{F(t)(1 - F(t))}} \right\} \leq \xi_m \right) = 1 - \alpha ,$$

of which the two roots $F_{\xi_m}^{-}$ and $F_{\xi_m}^{+}$ can be solved from the equality, and $S_m \subset (-\infty, \infty)$, a simultaneous confidence envelop

$$\mathbb{P}\left(F_{\xi_m}^{-}(t) \leq F(t) \leq F_{\xi_m}^{+}(t) \right) = 1 - \alpha$$

can be used to choose the $F^{+}(\tau)$ and $F^{-}(\tau')$ close to the true values. A one-sided confidence interval for ε was provided with the mean squared error bounded.

Jin and Cai (2007) worked on the estimation of the null distribution and the non-null effects especially for non-sparse cases. They estimated the null normal distribution $N(\mu_j, \sigma_j^2)$ and the frequency ε based on the z-scale and the empirical characteristic function of the normal test statistics.

Remark. *The methodologies in testing large-scale multiple hypotheses are almost all initiated based on the Gaussian mixtures. The normal tail and the Gumbel maxima play the key roles in developing the distribution of the test statistics and bounding the error rates. As a consequence, these methods are not applicable to multiple testing problems based on heavy-tailed distributions. We will give the examples in Chapter 3 and 4.*

2.3 Control of error rates

After detecting the existence of non-null components, we will naturally focus on a follow-up study, to identify the alternatives individually. Multiple testing procedures are used to test families of hypotheses simultaneously with the false rejections under control. Most of the procedures analyse the p-values of each individual hypothesis and reject those with the corresponding p-values below a certain threshold.

We will reformulate the error rates using the notation initiated by Benjamini and Hochberg. Defined on $\{1,2,\ldots,m\}$, R is the number of total rejections, V and S are the numbers of the false and true discoveries respectively, while U and T are the numbers of the true and false negatives respectively.

	True null	False null	Total
Declared significant	V	S	R
Declared non-significant	U	T	$m-R$
Total	m_0	$m-m_0$	m

Notice that m is a known constant, m_0 is unknown, and only R is an observable random variable, while V, S, U, T are unobservable random variables.

2.3.1 Cut-off threshold

There are many theories on the multiplicity correction. The Bonferroni correction uses a reduced level α/m as the significance level of each test, and it is straightforward that the desired significance level of the whole family is again bounded by α. Consider the problem of testing m hypotheses

$$\mathcal{H} = \{H_{0,i}\,,\, i = 1,\ldots,m\}$$

simultaneously at a given significance level α, and an unknown number m_0 of them are true nulls. Let $p_1,\ldots,p_m$ be the p-values corresponding to each hypothesis. The Bonferroni correction rejects the null hypothesis $H_{0,i}$ for the p-value less than the adjusted threshold α/m, and the probability of making at least one false rejection is bounded by α.

With the notations above, the family-wise error rate (FWER)

$$\text{FWER} = \mathbb{P}(V \geq 1)$$

is defined as the probability of erroneously rejecting at least one true null hypothesis.

Theorem 2.3.1 (Bonferroni Procedure)**.** *If, for $i = 1,\ldots,m$, hypothesis $H_{0,i}$ is rejected when the p-value $p_i \leq \frac{\alpha}{m}$, then the family-wise error rate for the simultaneous testing of $H_{0,1},\ldots,H_{0,m}$ satisfies FWER $\leq \alpha$.*

Proof. Suppose hypotheses $H_{0,i}$ with $i \in I$ are true and the remainder false, with the cardinality $|I|$ being m_0. We obtain the FWER

$$\begin{aligned}\text{FWER} = \mathbb{P}(\text{reject any } H_{0,i} \text{ with } i \in I) &= \mathbb{P}\left(\bigcup_{i\in I}\left\{p_i \leq \frac{\alpha}{m}\right\}\right)\\ &\leq \sum_{i\in I}\mathbb{P}\left(p_i \leq \frac{\alpha}{m}\right) \leq m_0\frac{\alpha}{m} \leq \alpha\,.\end{aligned}$$

□

The Bonferroni method is an example of a single-step procedure that assigns a common cut-off threshold and applies to all the hypotheses without further adjustment. With the conservative control of the probability of type I error per test at a horizontal level α/m, it results in low statistical power, that is, the ability to correctly detect the false null hypotheses is not satisfactory. From this perspective, one would consider slightly increasing the threshold of rejecting a single null hypothesis while maintaining the control of FWER.

Rather than controlling the FWER, one may consider the k-FWER, which is the probability of making at least k false rejections,

$$k\text{-FWER} = \mathbb{P}(V \geq k).$$

A simple procedure that controls the k-FWER for a given $k \leq m$ rejects any hypothesis for the p-value $p_i \leq \frac{k\alpha}{m}$.

2.3.2 Step-wise procedures

Other than the k-FWER control, various improvements can be made to increase the statistical power by lifting up the threshold for rejecting each single null hypothesis. A huge class of step-wise procedures has been developed based on comparing the sequence of the p-values to a proper threshold curve.

Holm's method

Holm (1979) developed a step-wise method that controls the FWER with improved power. Based on the Bonferroni correction, he reformulated the testing procedure as a sequentially rejective multiple testing method, where the total number of the remaining tests is reduced by one after each rejection. In this sense, one would perform the first test at level α/m, and then the second at level $\alpha/(m-1)$ if the first null hypothesis is rejected. The order of the hypotheses being tested is defined by the order of the p-values, such that the hypothesis with the highest probability of being rejected comes first. Let

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(i)} \leq \cdots \leq p_{(m)}$$

denote the non-decreasingly ordered p-values. The p-values are compared to the sequence of the thresholds

$$p_{(1)} \leq \frac{\alpha}{m}, \quad p_{(2)} \leq \frac{\alpha}{m-1}, \ldots, \quad p_{(i)} \leq \frac{\alpha}{m-i+1}, \cdots$$

instead of the constant significance level $\frac{\alpha}{m}$, and the procedure stops when the first $p_{(j)} > \frac{\alpha}{m-j+1}$ appears. It was proved that the overall significance level for the family of hypotheses is upper bounded by α.

Theorem 2.3.2 (Holm procedure)**.** *The Holm procedure satisfies FWER* $\leq \alpha$.

Proof. Suppose I is the set of the indices of the true null hypotheses. Let s be the smallest index satisfying

$$p_{(s)} = \min_{i \in I} p_i \,,$$

and $p_{(s)}$ denotes the smallest p-value of the true null hypotheses. Following the Holm procedure, $p_{(s)}$ will be tested only if $p_{(1)}, p_{(2)}, \ldots, p_{(s-1)}$ are tested and rejected, otherwise the procedure will stop at the first acceptance and commit no false rejection. Since all the hypotheses $H_{0,(1)}, H_{0,(2)}, \ldots, H_{0,(s-1)}$ are supposed to be correctly rejected before $H_{0,(s)}$ is tested, the first false rejection will occur when $p_{(s)} \le \alpha/(m-s+1)$. If not, the procedure stops at the acceptance of $H_{0,(s)}$ and no further hypothesis is tested. Therefore, the probability of at least one false rejection is

$$\begin{aligned}
\text{FWER} &= \mathbb{P}\left(H_{0,(1)}, H_{0,(2)}, \ldots, H_{0,(s)} \text{ are rejected}\right) \\
&= \mathbb{P}\left(\bigcap_{i=1}^{s} \left\{p_{(i)} \le \frac{\alpha}{m-i+1}\right\}\right) \le \mathbb{P}\left(p_{(s)} \le \frac{\alpha}{m-s+1}\right) = \mathbb{P}\left(\min_{i \in I} p_i \le \frac{\alpha}{m-s+1}\right) \\
&= \mathbb{P}\left(\bigcup_{i \in I} \left\{p_i \le \frac{\alpha}{m-s+1}\right\}\right) \le \sum_{i \in I} \mathbb{P}\left(p_i \le \frac{\alpha}{m-s+1}\right) \le \frac{|I|\,\alpha}{m-s+1} \le \alpha \,,
\end{aligned}$$

bounded by α, where the last inequality results from $s \le m - |I| + 1$. □

We recommend the book by Lehmann and Romano (2005) as a good summary with generalisations.

Step-up and step-down procedures

The Holm procedure is a step-down procedure. A *step-down* procedure starts from testing the null hypothesis that has the highest probability of being rejected, and stops at the first acceptance with all the previous hypotheses rejected. The order of the null hypotheses is usually obtained by ordering the p-values, since the p-value summarises each individual test and reveals its significance. In this sense, the step-down methods test the null hypotheses at the levels

$$\alpha_1, \alpha_2, \ldots, \alpha_m$$

following the order of

$$p_{(1)} \le p_{(2)} \le \cdots \le p_{(m)} \,,$$

and stops at the first crossing point.

On the other hand, a *step-up* procedure stops at the last crossing point of the p-values with the threshold curve. Suppose the null hypotheses are ordered as $H_{0,(1)}, H_{0,(2)}, \ldots, H_{0,(m)}$ according to the order of the p-values. One would keep on comparing $p_{(i)}$ to α_i for $i = 1, 2, \ldots, m$ and observe the $1 \le k \le m$ such that no more rejection occurs after obtaining $p_{(k)} \le \alpha_k$. One can equivalently start from the opposite direction, that is, testing if $p_{(m)} \le \alpha_m$. If $H_{0,(m)}$ is rejected, the procedure will stop on rejecting all the null hypotheses. If not, one will keep on testing $H_{0,(m-1)}, H_{0,(m-2)}, \dots$ until a first rejection $H_{0,(k)}$ occurs, and reject the null hypotheses $H_{0,(1)}, H_{0,(2)}, \dots, H_{0,(k)}$.

Benjamini–Hochberg procedure and the false discovery rate

Benjamini and Hochberg (1995) introduced the breakthrough idea of controlling the false discovery rate (FDR)

$$\mathrm{FDR} = \mathbb{E}\left(\frac{V}{R \vee 1}\right), \tag{2.11}$$

which is the expected proportion of erroneous rejections V among all rejections $R = V + S$. This proportion $Q = V/(V+S)$ is defined as false discovery proportion (FDP), and is set to be zero when $V + S = 0$.

The choice of controlling the FWER or the FDR may differ from case to case. In general, FDR creates more rejections but also makes more false discoveries. We prove the following proposition that shows the relation between the FDR and the FWER numerically.

Proposition 2.3.3. *Given a fixed significance level α, control of FWER implies control of FDR, in the sense that*

$$\mathit{FDR} \leq \mathit{FWER}. \tag{2.12}$$

Proof. We first note that

$$\frac{V}{R \vee 1} \leq \mathbb{1}\{V \geq 1\},$$

because for $V \geq 1$, we obtain that $V/R \leq 1 = \mathbb{1}\{V \geq 1\}$, since $V \leq R$. Otherwise, $\frac{V}{R \vee 1} = 0 = \mathbb{1}\{V \geq 1\}$ when $V = 0$. Notice that $V = R$ leads to the equality $V/R = \mathbb{1}\{V \geq 1\}$ regardless of the number of rejections. Then it follows that

$$\mathrm{FDR} = \mathbb{E}(\mathrm{FDP}) = \mathbb{E}\left(\frac{V}{R \vee 1}\right) \leq \mathbb{E}(\mathbb{1}\{V \geq 1\}) = \mathbb{P}(V \geq 1) = \mathrm{FWER}.$$

Therefore, control of the FWER at level α implies control of the FDR, and the equality holds when all hypotheses are true. □

Based on the ordered p-values, the Benjamini–Hochberg (BH) procedure rejects the null hypotheses $H_{0,(1)}, H_{0,(2)}, \dots, H_{0,(k)}$ with

$$k = \max_{1 \leq i \leq m} \left\{ i : \, p_{(i)} \leq \frac{i}{m}\alpha \right\}, \tag{2.13}$$

and if no such i exists, it rejects no hypothesis. The BH procedure is a step-up procedure.

Theorem 2.3.4 (BH procedure)**.** *For independent test statistics and for any configuration of false null hypotheses, the BH procedure controls the FDR at level α.*

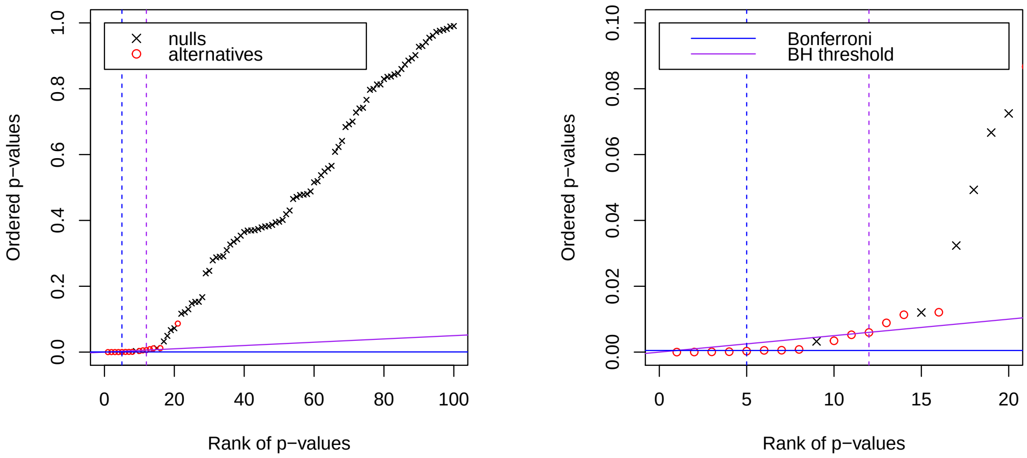


Figure 2.2 – Discoveries based on the Bonferroni correction and the BH procedure *(Left-side: rejections based on 100 p-values; right-side: a zoom-in of the left-side plot)*

Figure 2.2 shows the rejected p-values provided by the Bonferroni correction and the BH procedure. Using the Gaussian mixture model with $\varepsilon = 0.15$ and $\mu = 3$, we simulate a sample of size $m = 100$, of which 15 follow an $N(3,1)$ distribution (plotted in red circles), and the others are white noise (plotted in black crosses). In order to visualise the rejections based on the p-values, we order them and compare them to the threshold curves. We take the significance level $\alpha = 0.05$. The Bonferroni correction is a cut-off threshold $\alpha/m = 0.0005$, which is the horizontal blue line in both figures. The BH procedure compares the ordered p-values to a linear threshold plotted in purple, of which the slope equals $\alpha/m = 0.0005$. The vertical dashed lines show the last rejected p-values for each procedure.

The p-values from the alternatives tend to be smaller than those from the null, due to the positive shift in the normal distribution. The right-hand panel focuses on the smallest p-values near the threshold. The Bonferroni correction gives 5 rejections, corresponding to the p-values below 0.0005, and all of them are true positives. The BH procedure gives the first acceptance when the p-value exceeds the threshold, and rejects all the hypotheses with p-values below this value. In this experiment the BH procedure gives 12 rejections and commits a false rejection at $p_{(9)}$. The false discovery proportion in this case will be $1/12 \approx 0.083$. The BH procedure controls the expected value of the false discovery proportion instead of its exact value in each realisation. Compared to the Bonferroni correction, the BH procedure gives more false rejections but increases the power.

General control of the FDR

As a general formulation, step-up procedures can be summarised as follows. With the p-values ordered non-decreasingly, a step-up procedure rejects the null hypotheses with the set of indices

$$\left\{ i : p_{(i)} \leq \frac{\beta(k)}{m} \alpha \right\}, \tag{2.14}$$

and the critical index k is chosen to be

$$k = \max_{1 \leq i \leq m} \left\{ i : \; p_{(i)} \leq \frac{\beta(i)}{m} \alpha \right\} \tag{2.15}$$

where $\beta : \{1, \ldots, m\} \to \mathbb{R}^+$ is a nondecreasing function that can be customised according to the setting.

Benjamini and Hochberg gave the original proof that the BH procedure controls the FDR under certain assumptions. Based on the distribution of p-values, many improvements have been done by Benjamini and Yekutieli (2001), Benjamini et al. (2006), Sarkar (2002), Genovese and Wasserman (2002), Storey (2002), Blanchard et al. (2008) and Roquain et al. (2011).

Efron's Bayesian approach

Efron et al. (2001) and Efron (2004) developed multiple testing procedures for the two-groups model, and contributed to real data applications in genomics and DNA microarrays. They worked on the z-values instead of the p-values when testing the significant elements from the mixtures. The proportions of the two groups, namely π_0 for the nulls and π_1 for the alternatives, are regarded as prior probabilities. The corresponding density functions are $f_0(z)$ under the null and $f_1(z)$ under the alternative.

Following the definition

$$z_i = \Phi^{-1}(P_i), \quad i = 1, \ldots, m,$$

the theoretical null hypothesis is based on

$$z_i | H_{0,i} \text{ is true } \sim N(0,1).$$

The null density of the z-values is therefore

$$f_0(z) = \varphi(z) = e^{-z^2/2} \big/ \sqrt{2\pi}.$$

In reality, the empirical null distribution has a peak near zero, which is referred to as the central peak, where the distribution of the large majority of z-values of the true nulls is not influenced by the positive shift in the alternatives. The empirical null hypothesis

$$z_i | H_{0,i} \text{ is true } \sim N(\mu, \sigma^2)$$

is derived from the central peak, this density function provides an estimate $f_0(z)$.

Efron also estimated the density of the mixture $f(z) = \pi_0 f_0(z) + \pi_1 f_1(z)$ by Poisson regression fitting, using the fact that the number of z-values falling into each evenly spaced interval is proportional to its density, which can be reasonably estimated using a Poisson variable. The estimated densities $f(z)$ and $f_0(z)$ are used to construct a new criterion $f_0(z)/f(z)$, which is the ratio of the null density over the mixture density evaluated at the z-values. We will look into this method in Chapter 4. Further details and examples can be found in Efron (2010).

2.4 Motivation and innovation

The methodologies mentioned above are of the highest importance in multiple testing problems for mixture models. Among the frequently mentioned assumptions, the normality of the test statistics and the concavity of the distribution of the p-values from the alternatives guarantee the key principles behind the existing methods.

In the problem of detecting the alternatives among the mixtures, we find that the step-wise threshold methods are not able to solve multiple testing problems when the distribution of test statistics has heavy tails. Heavy-tailed data arises in many applications, and the heavy-tailedness leads to intrinsic problems in analysing the distribution of the test statistics and the p-values. To illustrate this, we discuss the behaviour of the p-values under the Cauchy distribution, which is a representative of this case. We propose an adaptive multiple testing procedure that works in the heavy-tailed situations, and can be generalised and robustified as well.

3 Testing globally for the existence of alternative

Our main contribution begins with answering the following question: What are the basic assumptions that should be guaranteed before step-wise threshold procedures are applied to multiple testing problems?

In this chapter we discuss one of the most important assumptions, normality, which is not inherently true in different applications. For example in financial models, the assets often have a large fluctuation and a narrow peak, and are not believed to have a normal distribution. Although the normality is not explicitly required in the methodologies mentioned in Chapter 2, the light-tailedness does contribute to the desired properties. We study the behavior of Cauchy test statistics as a motivating example, and propose a new method to detect the alternatives from the mixtures where the test statistics are heavy-tailed instead of Gaussian.

We are in particular interested in the asymptotic case where the hypothesis is asymptotically detectable with the parameters (ε, μ) under proper restrictions. We give the formulation and the results of the asymptotic detection based on the Cauchy mixture model.

3.1 Problem set-up

For $i = 1, \dots, m$, we test the hypothesis $H_{0,i}$ for the distribution of X_i, which is denoted by P_{θ_i} and is assumed to be varying in a class of distributions $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ modelled by the parameter θ. Suppose the null hypothesis $H_{0,i}$ is true if and only if

$$P_{\theta_i} \in \mathcal{P}_{0,i} = \left\{P_\theta : \theta \in \Theta_{0,i}\right\},$$

that is, the distribution of X_i belongs to a pre-specified class of distributions. Assume

$$m_0 = \sum_{i=1}^{m} \mathbb{1}\left\{H_{0,i} \text{ is true}\right\}$$

is the unknown number of the true null hypotheses, and $m_1 = m - m_0$ the number of true alternatives. Intuitively when m_1 is too small compared to m, which is referred to as the sparse case, it will be very difficult to detect the true alternatives. Therefore, in asymptotic studies m_0 needs to increase together with m in order to be detectable. We will give the formulation in Section 3.3.

Throughout this work we discuss the multiple testing problem based on the random effects model

$$X_i = \mu_i + \sigma Z_i\,, \quad i = 1, \ldots, m \tag{3.1}$$

where Z_i are independent random noises following a unimodal distribution with median zero, and $\sigma > 0$ is a scaling parameter. Note that σ need not be the square root of the variance since we do not require the existence of the moments. Let μ_i represent the size of the effect, of which the majority are assumed to be nulls. Suppose there exists a small fraction $\varepsilon \in (0,1)$ of non-null effects with $\mu_i \neq 0$, such that the probability of receiving a true null observation is

$$\mathbb{P}(\mu_i = 0) = 1 - \varepsilon\,.$$

For simplification we first assume that the non-zero effects concentrate at a positive value $\mu > 0$. This leads to the following statistical test for the null distribution against a two-point mixture model

$$\begin{aligned} &H_0^{(m)} : X_i \overset{\text{i.i.d.}}{\sim} F_0(x) \\ &\qquad \text{against} \\ &H_1^{(m)} : X_i \overset{\text{i.i.d.}}{\sim} F(x) = (1-\varepsilon)F_0(x) + \varepsilon F_0(x-\mu)\,. \end{aligned} \tag{3.2}$$

If the overall null hypothesis $H_0^{(m)}$ is accepted, none of the effects are declared significant and no further analysis is carried out. When the overall null hypothesis is rejected, the presence of an effect makes it desired to test

$$H_{0,i} : X_i \sim F_0(x) \quad \text{against} \quad H_{1,i} : X_i \sim F_1(x) = F_0(x-\mu), \quad i = 1, \ldots, m, \tag{3.3}$$

that is, to locate the true alternatives subject to a tolerated control of error rate.

Remark. *In the multiple testing problems, we often use m to represent the number of the hypotheses, while another index n that stands for the sample size of each test, such as the number of repeated measurements, is sometimes omitted. In fact there are two types of the interpretation of the random variable X_i as follows:*

i) X_i is taken as a single observation of which the distribution is P_{θ_i};

ii) X_i is a test statistic derived from a sample of size n_i.

In the first case, there exist situations when only one observation is collected per study. In the second case one would assume that in the i-th study, $X_{i1}, \ldots, X_{in_i}$ is an independent sample of *size n_i, which is usually true when repeated measurements are accessible. Therefore, the sample size n_i's are often omitted in the multiple testing procedures since they have already contributed to the computation of the test statistic X_i and the p-value P_i.*

Following this set-up we emphasise the heavy-tailedness, since both cases mentioned above do not inherently guarantee a Gaussian-type test statistic. In the second case when we test for $H_{0,i}$ given only a few repeated measurements, for example, X_{i1} and X_{i2} with unknown variance, then the test statistic X_i is apparently non-Gaussian. These are the most important situations where our work provides a new solution with critical thinking.

3.2 Heavy-tailed test statistics in hypothesis testing

The multiple testing procedures that reject the null hypotheses with the smallest p-values are based on the fact that, for the Gaussian distribution, the smallest p-values are most likely to be associated with the largest observations, and ideally indicate the alternatives. Since the generic p-values are defined as the probability of exceeding the current observation, they are clearly linked with the tail distribution of the true effects. In this section we first illustrate this phenomenon by analysing the extreme distribution of maxima, and then propose a reasonable parametrization under the asymptotic consideration.

3.2.1 Tail distribution and extreme distribution of maxima

Most literature works with the normality assumption of the test statistics, so we take the Gaussian distribution as an introductory example and a representative of light-tail distributions as well. The standard normal distribution belongs to the Gumbel maximum attraction domain, of which the extreme distribution of maxima is given as

$$\lim_{m\to\infty} \Phi(b_m + x/b_m)^m = \exp(-\exp(-x)), \quad \text{for all } x \in \mathbb{R},$$

where Φ is the cumulative distribution function of the standard normal, and $\{b_n, n \in \mathbb{N}\}$ is a normalising sequence. The extreme distribution of the normal random variables is also light-tailed, although heavier than Gaussian.

In addition, the exponential, lognormal, Gamma and Weibull distributions also lie in the maximum domain of attraction of the Gumbel distribution. However, for a heavy-tailed distribution such as Cauchy or Pareto, the distribution of maxima is also heavy-tailed.

Cauchy test statistics

We focus on the Cauchy distribution as an extreme case of heavy-tailed distributions. First we offer an explanation of our original interest in testing Cauchy components.

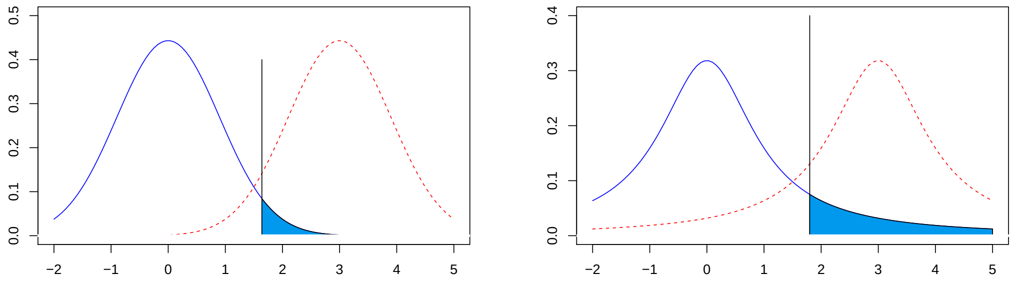


Figure 3.1 – Test the positive shift. (*Left-side: Gaussian; right-side: Cauchy*)

Consider a single test $H_0 : \mu = 0$ against $H_1 : \mu > 0$, Figure 3.1 compares the null and the shifted densities, with the test statistics following the Gaussian distribution (on the left) and the Cauchy distribution (on the right) respectively. The blue curves in both plots are the null densities with median zero, while the dashed red curves are the alternative densities with a positive shift. The shaded areas are the rejection regions defined as $(x_c, +\infty)$, where x_c is the critical value. We establish the one-side test for the normal mean by rejecting a large value of the normal variable that exceeds the critical value. However, for a large value of the Cauchy test statistic, the probability of it coming from the null is comparable to the probability of it being a true alternative.

Now we compare the tails of the test statistics. To distinguish two light-tailed distributions, the largest observations are supposed to be from the alternative distributions, because the ratio between the alternative tail and the null tail tends to zero exponentially when the alternative distribution has a positive shift. However, in mixtures of heavy-tailed distributions, the non-null components are not easily distinguished from the extremes of the null components. The difficulty is caused by the speed of the tail distribution decaying to zero.

Given an independent sample $X_1, X_2, \dots, X_m$, we want to test whether each observation comes from either a standard Cauchy$(0,1)$ or a shifted Cauchy$(\mu,1)$. The standard Cauchy distribution with the density function defined as

$$f(x) = \frac{1}{\pi(1+x^2)}$$

has an extremely heavy tail. We compute the extreme distribution of Cauchy maxima as follows.

Since

$$F^{-1}(t) = \tan[\pi t - \pi/2]\,,$$

the limiting result

$$\lim_{\delta\to 0}\frac{F^{-1}(1-\delta)-F^{-1}(1-2\delta)}{F^{-1}(1-2\delta)-F^{-1}(1-4\delta)}=\lim_{\delta\to 0}\frac{\tan(\pi/2-\delta\pi)-\tan(\pi/2-2\delta\pi)}{\tan(\pi/2-2\delta\pi)-\tan(\pi/2-4\delta\pi)}=2^1=2^{-\kappa} \tag{3.4}$$

indicates that the shape parameter of the extreme distribution is $\kappa=-1<0$. We will explain how it determines the extreme value distribution below.

To derive an explicit form of the Cauchy maxima, let $M_{m:m}=\max(X_1,X_2,\dots,X_m)$ be the largest observation among a standard Cauchy sample, we look for the distribution of

$$\mathbb{P}(M_{m:m}\le x)=(F(x))^m\,.$$

In order to have a non-degenerate limit distribution of the maximum, define the distribution of a linear transformation

$$\mathbb{P}\left(\frac{M_{m:m}-a_m}{b_m}\le\frac{x-a_m}{b_m}\right)=H\left(\frac{x-a_m}{b_m}\right),$$

where a_m and b_m are constant sequences depending only on m. The distributional properties are given by the following lemma.

Lemma 3.2.1 (Generalised extreme value distribution(GEVD$_M$))**.** *The only non-degenerate family of distributions satisfying*

$$\lim_{m\to\infty}H_m(a_m+b_mx)=\lim_{m\to\infty}[F(a_m+b_mx)]^m=H(x)\quad \textit{for any } x$$

is

$$H_\kappa(x;\lambda,\delta)=\exp\left\{-\left[1-\kappa\left(\frac{x-\lambda}{\delta}\right)\right]^{1/\kappa}\right\},\quad 1-\kappa\left(\frac{x-\lambda}{\delta}\right)\ge 0,\quad \kappa\ne 0, \tag{3.5}$$

and

$$H_0(x;\lambda,\delta)=\exp\left[-\exp\left(\frac{x-\lambda}{\delta}\right)\right],\quad \kappa=0. \tag{3.6}$$

Recall that the calculation (3.4) with $\kappa=-1$ leads to the Fréchet family. We compute the corresponding regularisation parameters

$$a_m=0,\quad b_m=F^{-1}\left(1-\frac{1}{m}\right)\doteq\frac{m}{\pi},$$

So, the distribution of Cauchy maxima is given by

$$f_{M_{m:m}}(x)=\frac{m\mathrm{e}^{-\frac{m}{\pi x}}}{\pi x^2}, \tag{3.7}$$

whose mode is proportional to m.

The distribution of Cauchy maximum is still very long-tailed, which is completely different from the light-tailed distributions in the Gumbel maximum domain of attraction. For the detection of heavy-tailed mixtures, the largest components are not simply drawn from a positively shifted distribution, but a mixture of both the null and the alternative. The indiscriminate use of the one-side rejection region and the p-values naturally implies the emphasis on the largest test statistics, and the error rate of false positives can be very large when the test statistics are in reality heavy-tailed.

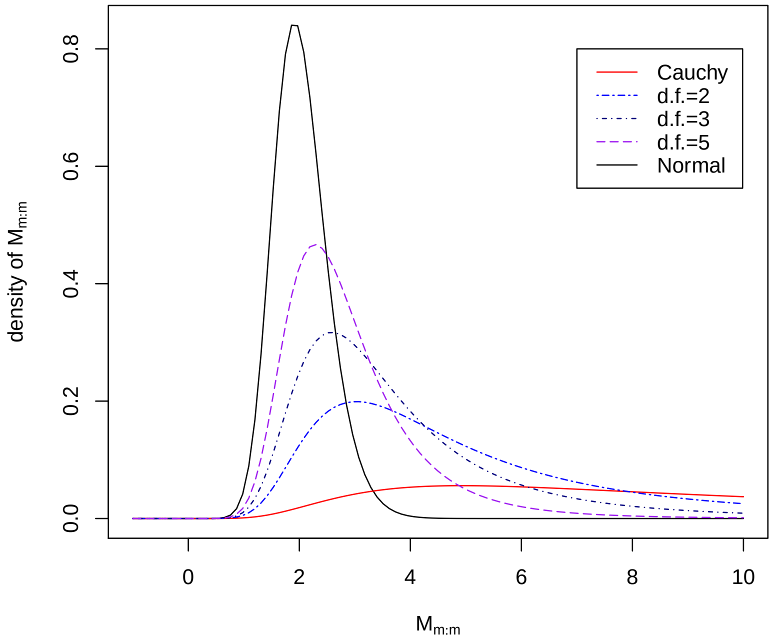


Figure 3.2 – Densities of the extreme distribution of maxima of t_ν distributions

Figure 3.2 compares the densities of the extreme values with the random variables having t_ν-distributions, with $\nu = 1, 2, 3, 5, \infty$ and sample size $m = 30$. Note that $\nu = 1$ is the Cauchy density and $\nu = \infty$ corresponds to the normal density. With the tail of f_0 and f_1 getting longer, the extreme value distribution has a longer tail and larger variance. As a consequence, the heavy-tailed test statistics do not show an immediate distinction from the extreme values from the nulls.

Figure 3.3 shows the densities of Cauchy maxima (left-side) and of t_2 maxima (right-side) compared to the other extreme densities mentioned above. When the heavy-tailedness appears with different tail distributions, any increment in the tail index makes a huge difference to the distributional property of the test statistics.

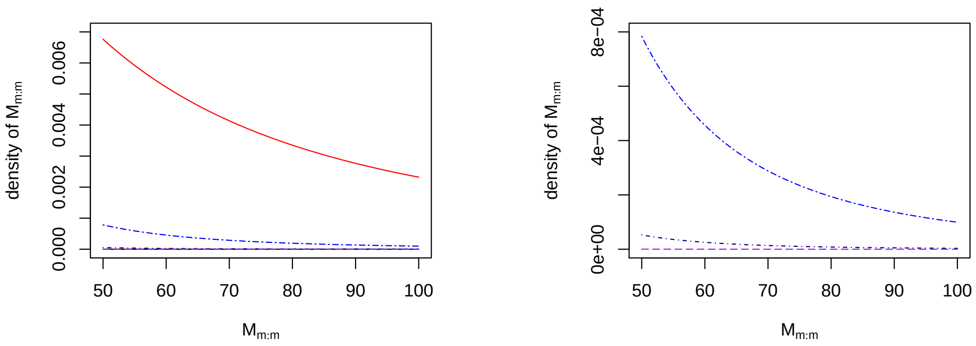


Figure 3.3 – Zoom-in of the densities of the extreme values

3.2.2 Formulation of the parameter space

Now we discuss the parametrisation such that the asymptotic hypothesis (3.2) can be investigated in the parameter space of (ε, μ) .

We are particularly interested in testing (3.2) with $m \to \infty$. The problem gets subtle when the number of the alternatives is not enough to be detected, or the size of the non-zero components cannot be distinguished from the block maxima of the nulls due to screening. As we consider the asymptotic case, the parameters (ε, μ) need to vary with m as follows.

i) For the parametrisation of ε_m we prefer the case such that the number of the true nulls and alternatives, namely m_0 and m_1, both tend to infinity as m grows.

ii) For the parametrisation of μ_m we conclude that the quality of testing (3.2) depends on how large the test statistics from the alternatives are compared to the maxima from the nulls. The increasing of μ_m with m needs to be comparable to the maximum, otherwise it is either too trivial or impossible to be detected.

We propose the formulation for (ε_m, μ_m) in the parameter space as follows.

Definition 3.2.2 (Asymptotic framework)**.** *Define the test of hypothesis (3.2) a desired asymptotic detection problem with ε_m and μ_m parametrised as*

$$\varepsilon_m = m^{-\gamma}, \quad \mu_m = M_a(m^r), \tag{3.8}$$

where the constant parameters $0 < \gamma < 1$, $r > 0$, and $M_a(m)$ denotes the order of the block maxima with respect to the size m, such that

$$\lim_{m\to\infty} \frac{M_{m:m}}{M_a(m)} \longrightarrow 1 \quad \textit{in probability.} \tag{3.9}$$

Note that for the Gaussian mixture with shifted mean, the desired parametrisation is $\mu_m = \sqrt{2\log(m^r)} = \sqrt{2r\log m}$, and for the Cauchy we seek to test $\mu_m = m^r$.

3.3 Likelihood ratio test for mixture models

In this section we discuss the likelihood ratio test for the global null $H_0^{(m)}$ against the mixture model $H_1^{(m)}$ as defined in (3.2). We introduce the ratio of the shifted density over the null density

$$g(x) = g_m(x;\ \mu,\varepsilon) = \frac{f_1(x)}{f_0(x)} = \frac{f_0(x-\mu)}{f_0(x)}. \tag{3.10}$$

Therefore, the logarithm of the likelihood ratio statistic for an independent sample $X_1, X_2, \dots, X_m$ is given by

$$\log(\mathrm{LR}_m) = \sum_{i=1}^{m} \log(\mathrm{LR}_m(X_i)) = \sum_{i=1}^{m} \log\big(1-\varepsilon_m + \varepsilon_m g_m(X_i)\big), \tag{3.11}$$

where $\mathrm{LR}_m(X_i)$ is the likelihood ratio based on one observation X_i.

3.3.1 Heavy-tailedness

We explain the decisive role played by the tail of the distribution of the test statistics.

The behaviour of the function $g(x)$ is essential to the effectiveness of likelihood ratio test. The likelihood ratio of the shifted Cauchy to the standard Cauchy is not monotone, and

$$g(x) = \frac{1+x^2}{1+(x-\mu)^2}$$

is a slowly varying function of x. Thus the largest values of the Cauchy test statistics are not necessarily attributed to the alternative distributions.

Figure 3.4 shows the function $g(x)$ for the Gaussian mixture and the Cauchy mixture distribution respectively, with $\varepsilon = 0.15$ and $\mu = 3$. The likelihood ratio with respect to the Gaussian mixture model is exponentially increasing with the test statistic, with $g(x) = e^{\mu x - \frac{1}{2}\mu^2}$. Therefore, the likelihood ratio test gives the generic one-sided rejection region such that any observed value of the test statistic that exceeds the critical value will imply a rejection of the null hypothesis. However, for the Cauchy mixture distribution, $g(x)$ is bounded and not monotone, and as a consequence, the most informative region is located around a central peak.

In general, we give the following conclusion for heavy-tailedness in the detection of a mixture model.

Condition 3.3.1 (Likelihood ratio). *Define the multiple testing problem (3.2) as a heavy-tailed*

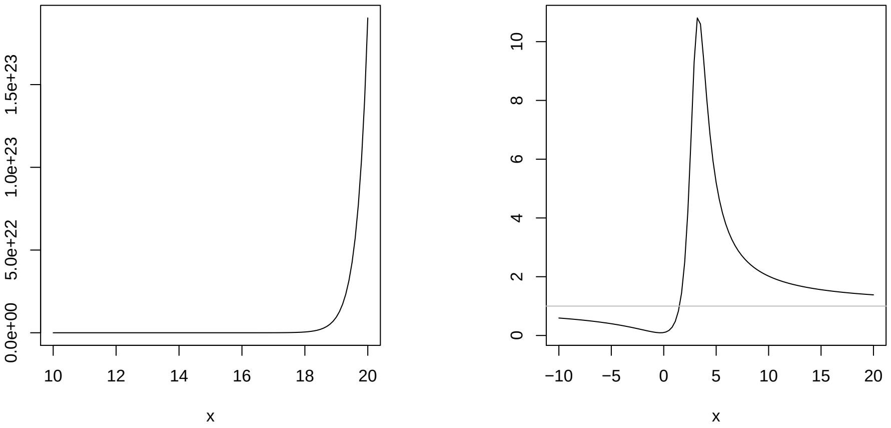


Figure 3.4 – The function $g(x)$ for Gaussian mixture *(left side)* and Cauchy mixture *(right side)*

testing problem if and only if the distribution of the test statistic satisfies

$$g(\sigma x)/g(x) \longrightarrow \textit{Constant}, \quad x \to \infty, \tag{3.12}$$

for any $\sigma > 0$, *where* $g(x) = f_1(x)/f_0(x) = f_0(x-\mu)/f_0(x)$.

3.3.2 Asymptotics

The likelihood ratio test for the problem (3.2) is the most powerful test subject to a fixed significant level. We refer to this optimality as detailed below.

Definition 3.3.2 (Optimal test)**.** *Let* α_0 *denote the tolerance of the significance level, and let*

$$\beta = \mathbb{P}_{H_1^{(m)}}\left(\textit{accept } H_0^{(m)}\right)$$

be the probability of type II error. An optimal testing procedure minimises β *subject to* $\alpha \le \alpha_0$.

The probabilities of the two types of errors are linked to the behaviour of the likelihood ratio statistic, and are inherently influenced by the tail distribution of the test statistic. We give the following results that conclude the asymptotic property of the expected log-likelihood ratio statistic.

Theorem 3.3.3. *Following the parametrisation (3.8)*

$$\varepsilon_m = m^{-\gamma} \quad \textit{and} \quad \mu_m = M_a(m^r)$$

of the asymptotic framework (3.2.2), there exists an asymptotically detectable region in the $\gamma - r$ *space such that with the parameters* (γ, r) *in this region, the following two conclusions hold:*

i) the log-likelihood ratio test statistic

$$\log(\mathrm{LR}_m) \longrightarrow \infty \quad \textit{in probability;} \tag{3.13}$$

ii) the sum of the probabilities of type I and type II errors tends to zero,

as $m \to \infty$.

In reality, one can derive the detection boundary utilising the maximum likelihood estimators of the parameters, namely ε and μ in terms of the mixture model. However, for most distributions the likelihood estimators can be very complicated, and can only be calculated numerically without the explicit form. Therefore, we propose to use the expected value of $\log \mathrm{LR}_m$ to test for the mixture distribution, of which the effectiveness can be guaranteed by the following lemma.

Lemma 3.3.4. *Suppose the random variable X has the density function f_0 under the global null $H_0^{(m)}$, and the mixture distribution $f(x) = (1-\varepsilon)f_0(x) + \varepsilon f_0(x-\mu)$ under the alternative $H_1^{(m)}$. Therefore, the test for $H_0^{(m)}$ against $H_1^{(m)}$ is fully characterised by the ratio $g(x) = f_1(x)/f_0(x)$, in the sense that the log-likelihood ratio of the test (3.2) satisfies*

$$\mathbb{E}_{f_0}(\log \mathrm{LR}_m) = O\left(m\varepsilon^2 \left(\mathbb{E}_{f_0}\left[(g(X))^2 - 1\right]\right)\right). \tag{3.14}$$

Proof. The proof of this lemma is based on a second order Taylor expansion.

$$\begin{aligned}
\mathbb{E}_{f_0}(\log \mathrm{LR}_m) &= \sum_{i=1}^{m} \int \log\left(\frac{(1-\varepsilon)f_0(x) + \varepsilon f_0(x-\mu)}{f_0(x)}\right) f_0(x)\,\mathrm{d}x \\
&= m \int \log\left(1 + \varepsilon(g(x)-1)\right) f_0(x)\,\mathrm{d}x \\
&\approx m \int \left(-\frac{\varepsilon^2}{2} g^2(x) + (\varepsilon^2 + \varepsilon) g(x) - \left(\frac{\varepsilon^2}{2} + \varepsilon\right)\right) f_0(x)\,\mathrm{d}x \\
&= -\frac{m\varepsilon^2}{2} \int \left(g^2(x) - 1\right) f_0(x)\,\mathrm{d}x \\
&= O\left(m\varepsilon^2 \left(\mathbb{E}_{f_0}\left[(g(X))^2 - 1\right]\right)\right),
\end{aligned}$$

which is true for any f_0 and f_1. □

In real data applications, there are various approaches to get a consistent estimator of the expected logarithm of the likelihood ratio. One is able to design asymptotically detectable tests based on the detection boundary proposed in Section 3.4.

3.4 Asymptotic detection boundary

In this section we provide the optimal detectable region of the test for the global null hypothesis against the two-point mixture with ε_m and μ_m formulated as (3.8).

Proposition 3.4.1 (Asymptotically detectable)**.** *Consider the test*

$$H_0^{(m)} \,:\, X_i \overset{i.i.d.}{\sim} F_0 \quad \textit{against} \quad H_1^{(m)} \,:\, X_i \overset{i.i.d.}{\sim} (1-\varepsilon)F_0 + \varepsilon F_1\,, \tag{3.15}$$

with the parametrisation

$$\varepsilon_m = m^{-\gamma} \quad \textit{and} \quad \mu_m = M_a(m^r)\,.$$

The detectable boundary is a curve in the (γ, r) space that partitions it into the detectable region and the non-detectable region under asymptotic consideration, such that

i) in the interior of the detectable region, the sum of the probabilities of the type I and type II error goes to zero as m tends to infinity;

ii) in the non-detectable region, the sum of the probabilities of the two types error goes to one as m tends to infinity.

The results on the asymptotic detection imply that when the number of hypotheses goes to infinity, we can only achieve an optimal test in the interior of the detectable region, which is a parametric restriction that limits the variation of the frequency ε_m and the size μ_m.

As for the detection problem based on the heavy-tailed test statistics, we are motivated to develop a new approach out of the following considerations:

- The effectiveness of the existing methodologies often relies on the assumptions that are not fulfilled with the test statistics following heavy-tailed distributions. Most approaches are adapted to the normal test statistics. Although there are weaker conditions other than the normality, such as the monotonicity of $f_0(x-\mu)/f_0(x)$ in x required in the estimation procedures mentioned before, these assumptions are still not true for f of power law.

- Taking the Cauchy distribution as a representative of the heavy-tailed distributions, the methods with moment estimation will fail, and the maximum likelihood estimators do not have explicit forms.

We propose a method based on the Kullback-Leibler divergence introduced in Kullback and Leibler (1951), which can be used to measure the distributional distinctions between the null and the alternative. More applications and interpretations can be found in information theory, and we recommend the works by Cover and Thomas (1991) and Kullback (1997) for further details.

3.4.1 On the Kullback–Leibler (KL) divergence

We let f_0 and f denote the densities of the test statistic X_i under the null hypothesis $H_0^{(m)}$ and the alternative hypothesis $H_1^{(m)}$ in (3.15) respectively, where

$$f(x) = (1-\varepsilon) f_0(x) + \varepsilon f_1(x) .$$

Definition 3.4.2. *Suppose the densities f_0 and f are absolutely continuous with respect to one another, and have the same support on $\mathbb{R}$.*

i) Define the Kullback-Leibler (KL) divergence from f_0 to f,

$$I(f_0 \| f) = \int \log\left(\frac{f_0(x)}{f(x)}\right) f_0(x)\, dx ,$$

which is the expected information for discrimination between f_0 and f per observation from f_0.

ii) Define the KL distance between f_0 and f,

$$\begin{aligned} KLD(f_0; f) &= I(f_0 \| f) + I(f \| f_0) \\ &= \int \log\left(\frac{f_0(x)}{f(x)}\right) f_0(x)\, dx + \int \log\left(\frac{f(x)}{f_0(x)}\right) f(x)\, dx , \end{aligned}$$

which is the symmetrised sum of KL divergences from f_0 to f and from f to f_0.

The KL divergence, which is also called discrimination information and a member of a large class of relative entropies, is designed to measure the distributional distinction in information theory. $I(f_0 \| f)$ is the information loss when f is used to estimate the true distribution f_0, and $I(f \| f_0)$ is the opposite. When we consider the KL divergence as a test statistic, either a large $I(f_0 \| f)$ or $I(f \| f_0)$ implies a significant difference between $H_0^{(m)}$ and $H_1^{(m)}$. This KL distance $I(f_0 \| f) + I(f \| f_0)$ measures the difference between f_0 and f from both directions. We investigate each of them and the sum as well.

The KL divergence is not symmetrical, that is, $I(f_0 \| f) \neq I(f \| f_0)$, and does not satisfy the triangle inequality, but there are the following properties such that the utility of the KL divergence as a measure of similarity between density functions is supported.

Proposition 3.4.3 (Non-negativity). *For any density functions f and g that satisfy the conditions in definition (3.4.2), the KL divergence $I(f \| g) \geq 0$ with equality if and only if $f \equiv g$.*

Proposition 3.4.4 (Additivity). *For independent observations $X_1, X_2, \ldots, X_m$,*

$$I_{X_1,\ldots,X_m}(f \| g) = \sum_{i=1}^{m} I_{X_i}(f \| g) ,$$

where $I_X(f \| g)$ is the realisation of $I(f \| g)$.

Proposition 3.4.5 (Convexity). *The KL divergence $I(f\|g)$ is convex in the pair (f,g), that is, for two pairs of probability density functions (f_1,g_1) and (f_2,g_2),*

$$\lambda I(f_1\|g_1)+(1-\lambda)I(f_2\|g_2)\geq I\big((\lambda f_1+(1-\lambda)f_2)\|(\lambda g_1+(1-\lambda)g_2)\big)$$

for any $\lambda\in[0,1]$.

Proposition 3.4.6 (Transformation invariance). *The KL divergence $I(f\|g)$ remains invariant under non-singular transformation.*

Propositions (3.4.3)-(3.4.6) are easy to prove and we recommend the paper by Kullback and Leibler (1951) for further detail.

The non-negativity (3.4.3) of KL divergence between any probability distributions follows directly from Jensen's inequality. Together with (3.4.4), it implies that when $\mu>0$ and $\varepsilon>0$ are fixed, the KL divergence evaluated at the i.i.d. sample $X_1,X_2,\ldots,X_m$ satisfies

$$\exists\,\delta>0,\quad I_{X_1,\ldots,X_m}(f_0\|f)=mI_{X_1}(f_0\|f)>m\delta, \tag{3.16}$$

which will tend to infinity as m grows. This property coincides with the classic testing approaches that separates $H_0^{(m)}$ and $H_1^{(m)}$ for fixed and known parametric distributions. However, it becomes a subtle problem when we model μ and ε as functions of m. We are interested in the asymptotic behaviour of KL divergence and will provide a new method of testing such hypotheses.

3.4.2 Asymptotic detection based on the KL divergence

Now we consider the test for the global null hypothesis $H_0^{(m)}$ against the mixture distribution under $H_1^{(m)}$, as defined in (3.15), utilising the asymptotic property of the KL divergence between the null and alternative distributions.

The best error exponent is given by Stein's lemma.

Lemma 3.4.7 (Stein's lemma). *Given a sample of size m, let α_m, β_m be the probabilities of type I and type II error associated with the test (3.15) of m test statistics. Define the optimal β_m^* such that*

$$\beta_m^*=\min_{\text{all tests s.t.}\alpha_m\leq\alpha_0}\beta_m\,.$$

It follows that

$$\beta_m^*\sim\mathrm{e}^{-mI(f_0\|f)},\quad m\to\infty. \tag{3.17}$$

We give a sketch of the proof.

Proof. For i.i.d. test statistics $X_1,\dots,X_m$ under the null $H_0^{(m)}$, we obtain

$$\lim_{m\to\infty} \frac{1}{m} \log \frac{\prod_{i=1}^m f_0(X_i)}{\prod_{i=1}^m f(X_i)} \xrightarrow{p} I(f_0\|f) \tag{3.18}$$

by the weak law of large numbers. Therefore, for any significance level $0 < \alpha_m \leq \alpha_0$, there exists $0 < \eta < \alpha_m$ such that

$$\lim_{m\to\infty} \mathbb{P}_{H_0^{(m)}} \left(e^{m(I(f_0\|f)-\eta)} < \frac{\prod_{i=1}^m f_0(X_i)}{\prod_{i=1}^m f(X_i)} < e^{m(I(f_0\|f)+\eta)} \right) = 1\,. \tag{3.19}$$

Construct a sequence of acceptance regions

$$A_m = \left\{ X \in \mathscr{X} \,:\, e^{m(I(f_0\|f)-\eta)} < \frac{\prod_{i=1}^m f_0(X_i)}{\prod_{i=1}^m f(X_i)} < e^{m(I(f_0\|f)+\eta)} \right\}. \tag{3.20}$$

Therefore,

$$\lim_{m\to\infty} \mathbb{P}_{H_0^{(m)}}(A_m) = 1\,. \tag{3.21}$$

On the other hand, by definition (3.20) of A_m,

$$e^{-m(I(f_0\|f)+\eta)}\mathbb{P}_{H_0^{(m)}}(A_m) \;\leq\; \mathbb{P}_{H_1^{(m)}}(A_m) \;\leq\; e^{-m(I(f_0\|f)-\eta)}\mathbb{P}_{H_0^{(m)}}(A_m)\,, \tag{3.22}$$

of which the limit leads to

$$\lim_{m\to\infty} (\beta_m^*)^{1/m} \longrightarrow \mathrm{e}^{-I(f_0\|f)}, \quad m\to\infty. \tag{3.23}$$

Note that the last equation follows from the fact that A_m is the optimal acceptance region with the minimal probability of type II error. Here the optimality can be proved by stating that no other acceptance/rejection region can achieve a smaller α_m and a smaller β_m simultaneously. Then the proof is complete. □

Similar proofs can be found in Chapter 12 of Cover and Thomas (1991), and Chapter 5, Section 3 of Kullback (1997) .

This lemma gives a direct relation between the KL divergence and the probabilities of type I and type II error, which guarantees that KL divergence could be used to perform hypothesis tests of large sample size. So we give the following criterion.

Definition 3.4.8 (Asymptotically detectable (KL method))**.** *Given an independent sample of size m, the test problem (3.15) is asymptotically detectable if mKLD tends to ∞ as m → ∞; if mKLD tends to 0, there is not enough information to detect the existence of non-null effects.*

3.4.3 Detect the mixture models

In the last part of this chapter, we give the results of the asymptotic detection of the mixture model by offering a detection boundary in the parameter space of (γ, r) that partitions the whole space into a detectable region and a non-detectable region. We first explore the KLD method for the Gaussian mixture, and compare our conclusion to the classic result given by Ingster, Donoho and Jin, and then derive the result for the Cauchy mixture, which is taken as a representative of the heavy-tailed mixture model.

The Gaussian mixture model

Suppose we test for the standard normal distribution against the mixture model, that is,

$$H_0^{(m)} \,:\, X_i \overset{\text{i.i.d.}}{\sim} N(0,1) \quad \text{against} \quad H_1^{(m)} \,:\, X_i \overset{\text{i.i.d.}}{\sim} (1-\varepsilon)N(0,1) + \varepsilon N(\mu,1)\,. \tag{3.24}$$

Following the behaviour of mKLD we conclude that, for fixed ε and μ, the test (3.24) is automatically detectable for sufficiently large m. So we focus on testing $\varepsilon_m = 0$ against $\varepsilon_m > 0$ when the fraction ε_m and shift μ_m are calibrated as

$$\varepsilon_m = m^{-\gamma}, \quad \mu_m = M_a(m^r) = \sqrt{2r\log m}, \tag{3.25}$$

and we give the statistical interpretation of the parameters.

As explained in the previous sections, the maximum of standard normal observations from a sample of size m is the same order with $\sqrt{2\log m}$, which gives the range where a shift μ_m could be intuitively visible. On the other hand, at least one non-null component is expected in a sample of size m, so we set $0 < \gamma < 1$, $0 < r < 1$ to be the region of interest. It is then necessary to discuss different settings of r and γ to find a detection boundary based on the asymptotic behaviour of mKLD.

The following theorem is our result obtained by analysing the KL divergence $mI(f_0 \| f)$.

Theorem 3.4.9. *In the Gaussian mixture detection problem (3.24) with $\varepsilon_m = m^{-\gamma}$ and $\mu_m = M_a(m^r) = \sqrt{2r\log m}$,*

$$mI(f_0 \| f) \to \infty, \quad m \to \infty,$$

if and only if $r > \rho(\gamma)$, where

$$\rho(\gamma) = \begin{cases} 0, & 0 < \gamma < \frac{1}{2}, \\ \gamma - \frac{1}{2}, & \frac{1}{2} < \gamma < \frac{3}{4}, \\ (1-\sqrt{1-\gamma})^2. & \frac{3}{4} < \gamma < 1, \end{cases} \tag{3.26}$$

is the asymptotic detection boundary.

Proof. Under the calibration $\varepsilon_m = m^{-\gamma}$, $\mu_m = \sqrt{2r\log m}$, the KL divergence from f_0 to f of a single observation is given by

$$\begin{aligned}
I(f_0\|f) &= \int \log\left(\frac{f_0(x)}{f(x)}\right) f_0(x)\,\mathrm{d}x \\
&= -\int \log\left(1+\varepsilon_m\left(\mathrm{e}^{\mu_m x-\mu_m^2/2}-1\right)\right) f_0(x)\,\mathrm{d}x \\
&= -\left(\int_{-\infty}^{\sqrt{2q\log m}} + \int_{\sqrt{2q\log m}}^{\infty}\right)\log\left(1+\varepsilon_m\left(\mathrm{e}^{\mu_m x-\mu_m^2/2}-1\right)\right) f_0(x)\,\mathrm{d}x \\
&\sim -\int_{-\infty}^{\sqrt{2q\log m}} \varepsilon_m\left(\mathrm{e}^{\mu_m x-\mu_m^2/2}-1\right) f_0(x)\,\mathrm{d}x - \int_{\sqrt{2q\log m}}^{\infty} \varepsilon_m\left(\mathrm{e}^{\mu_m x-\mu_m^2/2}-1\right) f_0(x)\,\mathrm{d}x \\
&\quad + \frac{\varepsilon_m^2}{2}\int_{\sqrt{2q\log m}}^{\infty}\left(\mathrm{e}^{2\mu_m x-\mu_m^2}-2\mathrm{e}^{\mu_m x-\mu_m^2/2}+1\right) f_0(x)\,\mathrm{d}x \\
&= \frac{\varepsilon_m^2}{2}\left(\mathrm{e}^{\mu_m^2}\bar{\Phi}\left(\sqrt{2q\log m}-2\sqrt{2r\log m}\right)-2\bar{\Phi}\left(\sqrt{2q\log m}-\sqrt{2r\log m}\right)+\bar{\Phi}\left(\sqrt{2q\log m}\right)\right)
\end{aligned}$$

with $0<q<r<1$, of which the approximation is valid if and only if

$$\left|\varepsilon_m(\mathrm{e}^{\mu_m x-\mu_m^2/2}-1)\right| < 1. \tag{3.27}$$

Note that

$$\int_a^\infty \mathrm{e}^{-t^2/2}\mathrm{d}t \sim \frac{1}{a}\mathrm{e}^{-a^2/2},$$

we can substitute $\varepsilon_m = m^{-\gamma}$, $\mu_m = \sqrt{2r\log m}$ in and obtain

$$mI(f_0\|f) \sim m^{1-2\gamma+2r-(\sqrt{q}-2\sqrt{r})^2} - m^{1-2\gamma-(\sqrt{q}-\sqrt{r})^2} + m^{1-2\gamma-q}. \tag{3.28}$$

Then we discuss the following three cases.

i) When $0<\gamma<\frac{1}{2}$, it is obvious that $mI(f_0\|f)\to\infty$.

ii) When $\frac{1}{2}<\gamma<\frac{3}{4}$ and $0<r<\frac{1}{4}$, condition (3.27) holds and $\sqrt{q}=2\sqrt{r}<1$ gives the leading order of m in (3.28),

$$\begin{aligned}
mI(f_0\|f)\to\infty &\Longleftrightarrow 1-2\gamma+2r-(\sqrt{q}-2\sqrt{r})^2>0 \text{ when } \sqrt{q}=2\sqrt{r} \\
&\Longleftrightarrow r>\gamma-\frac{1}{2}.
\end{aligned} \tag{3.29}$$

iii) When $\frac{1}{2}<\gamma<\frac{3}{4}$ and $\frac{1}{4}<r<1$, condition (3.27) holds and $q=1<2\sqrt{r}$ gives the leading order of m in (3.28),

$$\begin{aligned} mI(f_0\|f)\to\infty &\Longleftrightarrow 1-2\gamma+2r-(\sqrt{q}-2\sqrt{r})^2>0 \text{ when } q=1 \\ &\Longleftrightarrow r>(1-\sqrt{1-\gamma})^2. \end{aligned} \tag{3.30}$$

Conclusion of the three cases gives the theorem. □

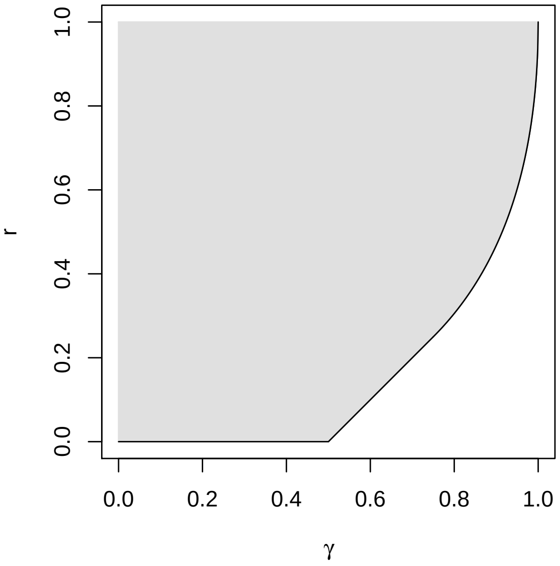


Figure 3.5 – The asymptotic detectable region (the grey part) of testing Gaussian mixtures

As shown in Figure 3.5, the detection boundary is a curve in the $\gamma - r$ plane that separates the whole parameter space into two parts, the detectable region (shaded) and the non-detectable region. Our KLD method achieves the optimal detectable boundary, which is the same as the one given by the likelihood ratio test.

The Cauchy mixture model

As the KLD method is validated in detecting the non-null components in Gaussian mixture model, we use it to tackle the heavy-tailed detection problem. In most of these detection problems, the tail index of the distribution has a considerable influence on the critical point of rejection. We start with the two-point Cauchy mixture model and test

$$\begin{aligned} H_0^{(m)} &: X_i \overset{\text{i.i.d.}}{\sim} f_0(x)=\frac{1}{\pi\left(1+x^2\right)}, \\ H_1^{(m)} &: X_i \overset{\text{i.i.d.}}{\sim} f(x)=\frac{1-\varepsilon_m}{\pi\left(1+x^2\right)}+\frac{\varepsilon_m}{\pi\left(1+(x-\mu_m)^2\right)}, \end{aligned} \tag{3.31}$$

since the case with fixed μ and ε is trivial in asymptotics.

As we proved in Section 3.2 that the Cauchy block maximum $M_a(m) \sim m$, which leads to the asymptotic parametrisation

$$\varepsilon_m = m^{-\gamma}, \quad \mu_m = M_a(m^r) = m^r. \tag{3.32}$$

We propose the asymptotic detectable boundary for the parameters (γ, r) in the following theorem.

Theorem 3.4.10. *The detection problem (3.31) with ε_m and μ_m given as (3.32) is asymptotically detectable if the parameters (γ, r) are in the region of*

$$\left\{(\gamma, r) : \left\{0 < \gamma < \frac{1}{2},\, 0 < r < 1\right\} \bigcup \left\{\frac{1}{2} \le \gamma < 1,\, \gamma - \frac{1}{2} < r < 1\right\}\right\}. \tag{3.33}$$

Proof. The integral of the log function embedded with the polynomials is a tough problem. Thus, we seek to approximate the ratio according to its asymptotic behaviour. Recall that the KL distance between f_0 and f is

$$\mathrm{KLD}(f_0; f) = I(f_0 \| f) + I(f \| f_0) = \int \log\left(\frac{f_0(x)}{f(x)}\right) f_0(x)\mathrm{d}x + \int \log\left(\frac{f(x)}{f_0(x)}\right) f(x)\mathrm{d}x,$$

where the log-likelihood ratio $\mathrm{LLR}^*(x) = \log(f/f_0)$ is

$$\mathrm{LLR}(x) = \log\left(\frac{(1-\varepsilon)f_0(x) + \varepsilon f_1(x)}{f_0(x)}\right) = \log\left(1 + \varepsilon\left(\frac{f_1(x)}{f_0(x)} - 1\right)\right).$$

For the Cauchy mixture model, the ratio $f_1(x)/f_0(x)$ is bounded, slowly varying in x, and tends to one when $x \to \pm\infty$. An intuitive approximation of the $\mathrm{LLR}(x)$ is the expansion using

$$\log(1+a) \simeq a$$

when $|a|$ is small. Otherwise when $|a|$ is large and $\frac{1}{|a|}$ is relatively small, we have

$$\log(1+a) = \log\left[a\left(1 + \frac{1}{a}\right)\right] = \log a + \log\left(1 + \frac{1}{a}\right) \simeq \log a + \frac{1}{a}.$$

Then we propose a piecewise approximation

$$\mathrm{LLR}^*(x) = \begin{cases} \dfrac{2\varepsilon\mu\left(x - \frac{\mu}{2}\right)}{1 + (x-\mu)^2}, & x \le \frac{\mu}{2}, \\ \min\left\{\dfrac{2\varepsilon\mu\left(x - \frac{\mu}{2}\right)}{1 + (x-\mu)^2}, \log\left(\dfrac{2\varepsilon\mu\left(x - \frac{\mu}{2}\right)}{1 + (x-\mu)^2}\right) + \dfrac{1 + (x-\mu)^2}{2\varepsilon\mu\left(x - \frac{\mu}{2}\right)}\right\}, & x > \frac{\mu}{2}. \end{cases} \tag{3.34}$$

Note that the approximation is applied when $\varepsilon^2\mu^2 + \varepsilon\mu^2 - 1 \geq 0$ and there exist the two roots $x_{1,2} = \mu(1+\varepsilon) \pm \sqrt{\varepsilon^2\mu^2 + \varepsilon\mu^2 - 1}$ at equality.

Therefore,

$$\mathrm{LLR}^*(x) = \begin{cases} \dfrac{2\varepsilon\mu\left(x - \frac{\mu}{2}\right)}{1+(x-\mu)^2}, & x \leq x_1 \text{ or } x \geq x_2\,, \\ \log\left(\dfrac{2\varepsilon\mu\left(x-\frac{\mu}{2}\right)}{1+(x-\mu)^2}\right) + \dfrac{1+(x-\mu)^2}{2\varepsilon\mu\left(x-\frac{\mu}{2}\right)}, & x_1 < x < x_2\,. \end{cases} \tag{3.35}$$

We desire to give the explicit form of the KL distance and provide a detectable boundary by analysing the asymptotic behaviour with the parameters γ and r varying.

$$\begin{aligned} \mathrm{KLD}(f_0;f) &= I(f_0\|f) + I(f\|f_0) \\ &= \int \log\left(\frac{f_0(x)}{f(x)}\right) f_0(x)\mathrm{d}x + \int \log\left(\frac{f(x)}{f_0(x)}\right) f(x)\mathrm{d}x \\ &= \int -\log\left(\frac{f(x)}{f_0(x)}\right) f_0(x)\mathrm{d}x + \int \log\left(\frac{f(x)}{f_0(x)}\right) f(x)\mathrm{d}x \\ &= \int -\log\left(1 + \frac{2\varepsilon\mu\left(x-\frac{\mu}{2}\right)}{1+(x-\mu)^2}\right)\frac{1}{\pi\left(1+x^2\right)}\mathrm{d}x \\ &\quad + \int \log\left(1 + \frac{2\varepsilon\mu\left(x-\frac{\mu}{2}\right)}{1+(x-\mu)^2}\right)\left[(1-\varepsilon)\frac{1}{\pi\left(1+x^2\right)} + \varepsilon\,\frac{1}{\pi\left(1+(x-\mu)^2\right)}\right]\mathrm{d}x \\ &= \int \log\left(1 + \frac{2\varepsilon\mu(x-\frac{\mu}{2})}{1+(x-\mu)^2}\right)\left[\frac{\varepsilon}{\pi(1+(x-\mu)^2)} - \frac{\varepsilon}{\pi(1+x^2)}\right]\mathrm{d}x \\ &= \int \log\left(1 + \frac{2\varepsilon\mu\left(x-\frac{\mu}{2}\right)}{1+(x-\mu)^2}\right)\left[\frac{2\varepsilon\mu(x-\frac{\mu}{2})}{\pi(1+x^2)(1+(x-\mu)^2)}\right]\mathrm{d}x. \end{aligned}$$

We discuss the two cases of the approximation.

i) $0 < r < \frac{\gamma}{2}$.

$$\begin{aligned} \mathrm{KLD}(f_0;f) =& I(f_0\|f) + I(f\|f_0) \\ \simeq& \int_{-\infty}^{+\infty} \left(\frac{2\varepsilon\mu\left(x-\frac{\mu}{2}\right)}{1+(x-\mu)^2}\right)\left[\frac{2\varepsilon\mu(x-\frac{\mu}{2})}{\pi(1+x^2)(1+(x-\mu)^2)}\right]\mathrm{d}x \\ =& \frac{\varepsilon^2}{2\pi}\left(\frac{\mu\left(\mu x - \mu^2 - 2\right)}{1+(x-\mu)^2} + \left(\mu^2-2\right)\arctan(x-\mu) + 2\arctan(x)\right)\Bigg|_{-\infty}^{+\infty}. \end{aligned}$$

It is easy to see that

$$\lim_{x\to\pm\infty} \frac{\varepsilon^2}{2\pi}\left(\frac{\mu\left(\mu x - \mu^2 - 2\right)}{1+(x-\mu)^2} + \left(\mu^2-2\right)\arctan(x-\mu) + 2\arctan(x)\right) = \pm\frac{\varepsilon^2\mu^2}{4},$$

which leads to

$$m\mathrm{KLD}(f_0;f) = O\left(m^{1-2\gamma+2r}\right).$$

ii) $r \geq \frac{\gamma}{2}$.

$$\begin{aligned}
\mathrm{KLD}(f_0;f) &= I(f_0 \| f) + I(f \| f_0) \\
&= \left(\int_{-\infty}^{x_1} + \int_{x_2}^{\infty} + \int_{x_1}^{x_2}\right) \log\left(1 + \frac{2\varepsilon\mu\left(x-\frac{\mu}{2}\right)}{1+(x-\mu)^2}\right) \left[\frac{2\varepsilon\mu(x-\frac{\mu}{2})}{\pi(1+x^2)(1+(x-\mu)^2)}\right] \mathrm{d}x \\
&= I_A + I_B + I_C .
\end{aligned}$$

For $x \leq x_1$ or $x \geq x_2$,

$$\begin{aligned}
I_A + I_B &\simeq \left(\int_{-\infty}^{x_1} + \int_{x_2}^{\infty}\right) \left(\frac{2\varepsilon\mu\left(x-\frac{\mu}{2}\right)}{1+(x-\mu)^2}\right) \left[\frac{2\varepsilon\mu(x-\frac{\mu}{2})}{\pi(1+x^2)(1+(x-\mu)^2)}\right] \mathrm{d}x \\
&= \frac{\varepsilon^2}{2\pi} \left(\frac{\mu\left(\mu x - \mu^2 - 2\right)}{1+(x-\mu)^2} + \left(\mu^2 - 2\right) \arctan(x-\mu) + 2\arctan(x)\right) \left(\Big|_{-\infty}^{x_1} + \Big|_{x_2}^{\infty}\right)
\end{aligned}$$

For $x_1 < x < x_2$,

$$\begin{aligned}
I_C &\simeq \int_{x_1}^{x_2} \left[\log\left(\frac{2\varepsilon\mu\left(x-\frac{\mu}{2}\right)}{1+(x-\mu)^2}\right) + \frac{1+(x-\mu)^2}{2\varepsilon\mu\left(x-\frac{\mu}{2}\right)}\right] \left[\frac{2\varepsilon\mu(x-\frac{\mu}{2})}{\pi(1+x^2)(1+(x-\mu)^2)}\right] \mathrm{d}x \\
&= \underbrace{\int_{x_1}^{x_2} \log\left(\frac{2\varepsilon\mu\left(x-\frac{\mu}{2}\right)}{1+(x-\mu)^2}\right) \left[\frac{2\varepsilon\mu(x-\frac{\mu}{2})}{\pi(1+x^2)(1+(x-\mu)^2)}\right] \mathrm{d}x}_{I_{C_1}} + \underbrace{\int_{x_1}^{x_2} \frac{1+(x-\mu)^2}{2\varepsilon\mu\left(x-\frac{\mu}{2}\right)} \left[\frac{2\varepsilon\mu(x-\frac{\mu}{2})}{\pi(1+x^2)(1+(x-\mu)^2)}\right] \mathrm{d}x}_{I_{C_2}},
\end{aligned}$$

where the second integral

$$I_{C_2} = \int_{x_1}^{x_2} \frac{1+(x-\mu)^2}{2\varepsilon\mu\left(x-\frac{\mu}{2}\right)} \left[\frac{2\varepsilon\mu(x-\frac{\mu}{2})}{\pi(1+x^2)(1+(x-\mu)^2)}\right] \mathrm{d}x = \int_{x_1}^{x_2} \frac{1}{\pi(1+x^2)} \mathrm{d}x = \frac{1}{\pi}[\arctan(x_2) - \arctan(x_1)].$$

Thus,

$$\begin{aligned}
mI_{C_2} &= \frac{1}{\pi} m \left(\frac{1}{\mu(1+\varepsilon) - \sqrt{\varepsilon^2\mu^2 + \varepsilon\mu^2 - 1}} - \frac{1}{\mu(1+\varepsilon) + \sqrt{\varepsilon^2\mu^2 + \varepsilon\mu^2 - 1}}\right) \\
&= O(m^{1-r-\frac{\gamma}{2}}).
\end{aligned}$$

The first part I_{C_1} can be approximated by the integral of a Cauchy density with a central peak at μ,

$$\begin{aligned} I_{C_1} &= \int_{x_1}^{x_2} \log\left(\frac{2\varepsilon\mu\left(x-\frac{\mu}{2}\right)}{1+(x-\mu)^2}\right)\left[\frac{2\varepsilon\mu(x-\frac{\mu}{2})}{\pi(1+x^2)(1+(x-\mu)^2)}\right]\mathrm{d}x \\ &\simeq \int_{x_1}^{x_2}\left(\log(\varepsilon\mu^2)\frac{\varepsilon\mu^2}{\pi(1+\mu^2)}\right)\frac{1}{1+(x-\mu)^2}\mathrm{d}x \\ &= \left(\log(\varepsilon\mu^2)\frac{\varepsilon\mu^2}{\pi(1+\mu^2)}\right)\arctan(x-\mu)\Big|_{x_1}^{x_2} \\ &= \left(\log(\varepsilon\mu^2)\frac{\varepsilon\mu^2}{\pi(1+\mu^2)}\right)(\arctan(x_2-\mu)-\arctan(x_1-\mu)). \end{aligned}$$

$$\begin{aligned} mI_{C_1} &= m\left(\log(\varepsilon\mu^2)\frac{\varepsilon\mu^2}{\pi(1+\mu^2)}\right)\left(\arctan\left(\mu\varepsilon+\sqrt{\varepsilon^2\mu^2+\varepsilon\mu^2-1}\right)-\arctan\left(\mu\varepsilon-\sqrt{\varepsilon^2\mu^2+\varepsilon\mu^2-1}\right)\right) \\ &= O\left(m^{1-\gamma}\log m\left(1+(m^{\frac{\gamma}{2}-r})\right)\right). \end{aligned}$$

A summary of the results above leads to the conclusion in the theorem. □

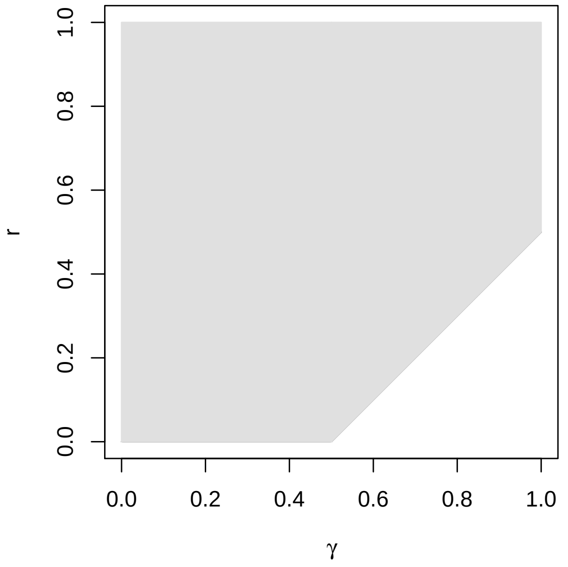


Figure 3.6 – The asymptotic detectable region (the grey part) of testing Cauchy mixtures

Figure 3.6 shows the asymptotically detectable region for the Cauchy mixture model, with ε_m and μ_m formulated as functions of m and the parameters (γ, r). For example, with the total number of hypotheses m=10000, a multiple testing problem for sparse Cauchy components has $\gamma > \frac{1}{2}$, which leads to the fraction $\varepsilon < 0.01$, that is, the number of the true non-zero components $m\varepsilon_m = 100$. In order that the alternatives are asymptotically detectable, the non-zero Cauchy effects must be at least $\mu_m > 10$. On the other hand, for the Cauchy effects of size $\mu_m = 10$ to be detectable, there must exist at least 100 true alternatives among the sample of 10000 observations.

Remark (Sparsity). *The region of $\gamma > \frac{1}{2}$ is also referred to as the sparse case, with the proportion of the alternatives $\varepsilon_m < \frac{1}{\sqrt{m}}$. The asymptotic detection theorems indicate that with the frequency of the alternatives being relatively small, the size of the non-zero effects must be moderately large to be detected. Apparently, the increasing of μ_m of the Cauchy elements needs to be faster than the normal means.*

4 Detecting individual alternatives

In this chapter we propose an adaptive multiple testing procedure with the test statistics following heavy-tailed distributions.

4.1 Distribution of p-values and control of error rates

Throughout this work we focus on the p-values instead of the test statistics or the z-values. Multiple hypothesis testing can equivalently be based upon the test statistics and the generic p-values, and the p-value plays an important role in terms of statistical inference and interpretation. There is a rich literature discussing the advantages and disadvantages of p-values, and we will emphasise that it is the correct definition and adaptation that influence the effectiveness of p-value based methods. In this section we first give the definition of p-values, and then discuss how p-values behave in multiple testing problems with heavy-tailed test statistics.

4.1.1 p-values

Depending on the random variable X_i and the null hypothesis $H_{0,i}$, the p-value denoted by $P_i = P_i(X_i)$ is also a random variable that reflects how strongly the value of X_i contradicts $H_{0,i}$. Recall that the p-value is computed as the probability of having the current observation or an even more extreme value under the null hypothesis. Therefore, it is equivalently regarded as the smallest significance level that would be taken to reject the null hypothesis for the observation of X_i. A small p-value indicates that the rejection is a correct decision with high probability.

Formally, let $\alpha \in (0,1)$ be a potential significance level, and let $\mathcal{R}_\alpha$ be the corresponding rejection region in the sample space of X. We call the region $\mathcal{R}_\alpha$ *nested* in the sense that

$$\mathcal{R}_{\alpha'} \subseteq \mathcal{R}_\alpha \quad \text{for any } 0 < \alpha' < \alpha < 1\,.$$

The p-value is formally defined as the smallest level α such that the null hypothesis is rejected by a nested rejection region $\mathcal{R}_\alpha$, that is,

$$\mathbb{P}(X) = \inf_\alpha \{X \in \mathcal{R}_\alpha\}.$$

By definition, there are the following distributional properties of the p-values.

Distribution of the p-values under the null

Proposition 4.1.1 (p-value)**.** *Suppose random variable X has distribution P_θ for some $\theta \in \Theta$, and the null hypothesis H specifies $\theta \in \Theta_H$. Assume the rejection regions $\mathcal{R}_\alpha$ are nested.*

i) If

$$\sup_{\theta \in \Theta_H} \mathbb{P}_{X \sim P_\theta}(X \in \mathcal{R}_\alpha) \leq \alpha, \quad 0 < \alpha < 1,$$

then it follows that the distribution of the p-value P under $\theta \in \Theta_H$ satisfies

$$\mathbb{P}_{X \sim P_\theta}(P \leq u) \leq u, \quad 0 \leq u \leq 1. \tag{4.1}$$

ii) If the rejection region $\mathcal{R}_\alpha$ is constructed such that, for $\theta \in \Theta_H$,

$$\mathbb{P}_{X \sim P_\theta}(X \in \mathcal{R}_\alpha) = \alpha, \quad 0 < \alpha < 1,$$

then the p-values are uniformly distributed, that is,

$$\mathbb{P}_{X \sim P_\theta}(P \leq u) = u, \quad 0 \leq u \leq 1. \tag{4.2}$$

The second case of Proposition 4.1.1 is known as the *uniformity* of the p-values under the null, while the first case is referred to as the *super-uniformity.* There are a series of methodologies aiming at weighting and adjusting the p-values instead of using the raw ones, and we recommend Benjamini et al. (2006), Genovese et al. (2006) and Roquain et al. (2009) for further interest.

Throughout the present thesis, we consider the definition

$$P_i = \mathbb{P}\left(X > X_i | H_{0,i} \text{ is true}\right) = 1 - F_0(X_i) \tag{4.3}$$

that is universally adopted.

The distribution of p-values under the null hypotheses given by Proposition 4.1.1 is not influenced by the number of hypotheses or the distribution family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. When the alternative hypotheses are true, the distribution of p-values will depend on the tail distribution and the number of hypotheses as well.

4.1.2 Multiple testing based on p-values

Now we investigate the behaviour of the p-values under the alternative, and explore how it influences the statistical testing.

Consider the multiple testing problem for the mixture model

$$H_{0,i}: X_i \sim F_0 \quad \text{against} \quad H_{1,i}: X_i \sim F_1, \quad i = 1, \dots, m, \tag{4.4}$$

let $H^m = (H_1, \dots, H_m)$ be the indicator variables where $H_i = 0$ if and only if the i-th null hypothesis is true. Given the independence of the tests, we assume that $H_i \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\varepsilon)$. Let $P^m = (P_1, \dots, P_m)$ denote the p-values, of which the realised values are denoted by $(p_1, \dots, p_m)$. The generic p-value is given by $p_i = \mathbb{P}_{H_{0,i}}(X > X_i) = 1 - F_0(X_i)$. We assume that P^m are marginally drawn from the probability distribution

$$P_i \sim (1-\varepsilon)U + \varepsilon F_p\,,$$

where U is the probability distribution of a Uniform(0,1), and

$$P_i | H_i = 1 \sim F_p\,,$$

where $H_i = 1$ corresponds to a true alternative.

Distribution of the p-values under the alternative

Recall that we assume the test statistics follow the mixture model

$$X_i \sim (1-\varepsilon)F_0 + \varepsilon\, F_1\,,$$

so the marginal distribution of the p-values under the alternatives $\{H_{1,i},\, i = 1, \dots, m\}$ is

$$F_p(t) = \mathrm{P}_{H_{1,i}}(P_i \le t) = \mathrm{P}_{H_{1,i}}(1 - F_0(X_i) \le t) = \mathrm{P}_{H_{1,i}}(X_i \ge F_0^{-1}(1-t)) = 1 - F_1(F_0^{-1}(1-t))\,,$$

and the density function is

$$f_p(t) = \frac{f_1(F_0^{-1}(1-t))}{f_0(F_0^{-1}(1-t))}\,.$$

Therefore, the distribution of a p-value under the mixture model is formulated as

$$\tilde{F}_p(t) = 1 - \underbrace{[(1-\varepsilon)F_0 + \varepsilon\, F_1]}_{\text{mixture model}}(F_0^{-1}(1-t)) = (1-\varepsilon)t + \varepsilon(1 - F_1(F_0^{-1}(1-t))),$$

$$\tilde{f}_{p_i}(t) = (1-\varepsilon) + \varepsilon\, f_1(F_0^{-1}(1-t))(F_0^{-1}(1-t))' = (1-\varepsilon) + \varepsilon\, \frac{f_1(F_0^{-1}(1-t))}{f_0(F_0^{-1}(1-t))},$$

for $0 < t < 1$.

Given $F_1(x) = F_0(x-\mu)$, the joint distribution of the p-values under the global alternative $H_1^{(m)}$ is influenced by both ε, μ and m. We will restrict the multiple testing problems to the asymptotically detectable region given in Chapter 3.

4.1.3 A heavy-tailed framework in multiple testing

Let $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$ be the ordered p-values, and let α be a pre-specified significance level bounding the probability of type I error. Multiple testing procedures based on rejecting the smallest p-values were originally developed and verified based on Gaussian test statistics, of which the effectiveness is guaranteed by the fact that the true alternatives are easily distinguished from the block maxima of the nulls. That is why the classic theory and methodologies we described in the previous chapters work well under the normality assumption. For heavy-tailed test statistics, however, we will point out that even though the control of the false discovery rate is still maintained, classical step-wise procedures will fail to find rejections among the smallest p-values.

In Chapter 3 we discussed the heavy-tailedness with respect to the likelihood ratio and provided the Condition 3.3.1

$$f_1(x)/f_0(x) \longrightarrow \text{Constant}, \quad x \to \infty,$$

to characterise the heavy-tailed distributions. In addition, we now assume the distribution of the p-values of a heavy-tailed test statistic satisfies the following conditions.

Condition 4.1.2 (C1)**.** *Assume that the distribution of the test statistic under the alternative hypothesis, namely F_1, satisfies*

i) *F_1 is symmetric and unimodal.*

ii) (Asymptotic scale invariance.) *For any $\sigma > 0$, $\bar{F}_1(\sigma x)\big/\bar{F}_1(x) \to C_\sigma$ as $x \to \infty$, where $\bar{F}_1(x) = \mathbb{P}(X > x)$, $C_\sigma > 0$ is a constant related to σ.*

iii) (Long − tailedness.) *For any $t > 0$, $\bar{F}_1(x+t)\big/\bar{F}_1(x) \to 1$ as $x \to \infty$.*

Condition 4.1.3 (C2)**.** *Assume that the cumulative distribution and the density of the p-values under the alternative, namely F_p and f_p, satisfy*

i) *F_p is not concave,*

ii) $\lim_{t\to 0^+} \frac{dF_p(t)}{dt} = 1,$

iii) $\lim_{t\to 0^+} \frac{df_p(t)}{dt} = 1,$

iv) *f_p is unimodal,*

v) $f_p(t)$ *is uniformly continuous in* t.

We will later on explain how these conditions change the detection of alternatives with test statistics having heavy-tailed distributions. As an illustration, we consistently take Cauchy distribution as an example that follows Conditions 4.1.2 and 4.1.3.

p-values of Cauchy test statistics

In the Cauchy distribution where

$$F_0(x) = \frac{\arctan(x)}{\pi} + \frac{1}{2} \quad \text{and} \quad F_0^{-1}(t) = \tan\left(\pi t - \frac{\pi}{2}\right),$$

we obtain the distribution of a single p-value $P_i(X_i)$ under the alternative Cauchy distribution with a shift μ as follows,

$$F_p(t) = \frac{1}{2} - \frac{1}{\pi}\arctan\left(\tan\left(\frac{\pi}{2} - \pi t\right) - \mu\right),$$

$$f_p(t) = \frac{1 + (F_0^{-1}(1-t))^2}{1 + (F_0^{-1}(1-t) - \mu)^2} = \frac{1 + \tan^2(\frac{\pi}{2} - \pi t)}{1 + \left(\tan(\frac{\pi}{2} - \pi t) - \mu\right)^2}.$$

Therefore, we derive the distribution of the p-values under the mixture model

$$\tilde{f}_p(t) = (1-\varepsilon) + \varepsilon \frac{1 + \frac{1}{\tan^2(\pi t)}}{1 + \left(\frac{1}{\tan(\pi t)} - \mu\right)^2}.$$

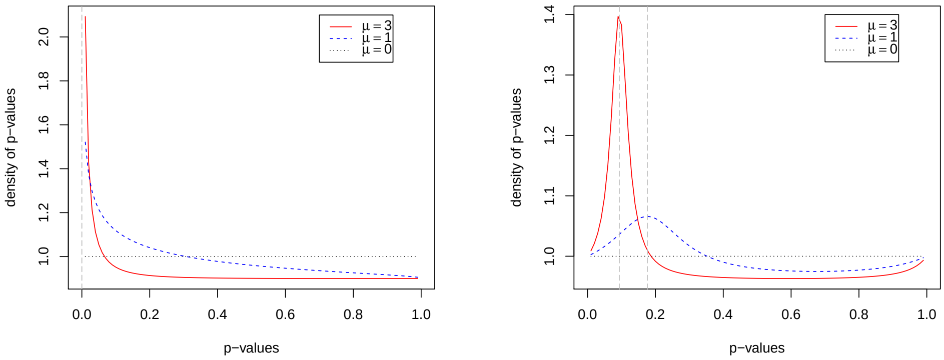


Figure 4.1 – The densities of the p-values for Gaussian mixtures *(left-side)* and Cauchy mixtures *(right-side)*

Figure 4.1 compares $(1-\varepsilon) + \varepsilon f_p$, that is, the density of the p-value under the alternative $H_1^{(m)}$, with the test statistics following the Gaussian distribution (on the left) and the Cauchy distribution (on the right) respectively. The frequency of the true alternatives is $\varepsilon = 0.1$, and the positive shift is $\mu = 3$ (solid red lines) and $\mu = 1$ (dashed blue lines). The horizontal lines correspond to $f(t) = 1$, which is the density of the uniform distribution under the global null $H_0^{(m)}$.

With normal test statistics, the marginal distribution of the p-values under the alternative goes to infinity fast when the p-value goes to zero. This explains why the non-null components are relatively easily detected from the Gaussian mixtures. When we consider the Cauchy test statistics, the distribution f_p is bounded, such that with a small frequency ε, the density εf_p is not easily distinguished from $f(t) = 1$. In addition, compared to the uniform distribution on $(0,1)$, it is obvious that the most informative region of the p-values (indicating the true alternatives) is not centralised at zero, but rather located around a critical value which is always larger than zero.

In fact, for the Cauchy mixture model, we obtain the explicit form of the local concentration of the alternative p-values

$$p_c = \frac{1}{\pi}\left(\arctan\left(-\sqrt{1+\frac{\mu^2}{4}}-\frac{\mu}{2}\right)\right)+\frac{1}{2}, \tag{4.5}$$

where the density of the p-values is maximised. Note that $p_c > 0$, which indicates that the p-values from the alternatives are not the smallest ones.

Therefore, we are interested in formulating a multiple testing procedure based on the *mode* of the p-values under the alternative, and we aim at locating the alternatives by detecting the mode.

4.1.4 Quality of the test: operating characteristics

Before addressing the proposed testing procedure, we provide another way to evaluate a test, which inspired our distributional formulation of the test statistics and the p-values.

In practice, the quality of a classifier or a testing procedure can be quantified utilising the ROC, which stands for the Receiver Operating Characteristic, or equivalently, the Relative Operating Characteristic. This concept is brought from electrical engineering to statistical testing and comparison. Two of the most widely discussed operating characteristics are the true positive rate (TPR) and the false positive rate (FPR), and the curve of the TPR as a function of the FPR is referred to as the ROC curve.

When evaluating a test, we seek to maximise the power subject to control of the false positive rate. The ROC curve provides a graphical perspective on the evaluation of testing procedures. We propose the following formulation based on the distribution of the p-values under mixture models.

Definition 4.1.4. *Suppose the p-values follow the marginal distribution*

$$\tilde{F}_p(t) = (1-\varepsilon)t + \varepsilon F_p(t),$$

which is a random mixture with $\varepsilon = \mathbb{P}(H_i = 1)$. Define the probabilities of having a true positive and a false positive as

$$G_1(t,\Delta) = \int_{\mathcal{R}_{t,\Delta}} f_p(u)\,du \tag{4.6}$$

$$G_0(t,\Delta) = \int_{\mathcal{R}_{t,\Delta}} 1\,du \tag{4.7}$$

respectively, where $\mathcal{R}_{t,\Delta}$ is a rejection region for the p-values, centered at a threshold $t \in [0,1]$ and of length $\Delta \in [0,1]$.

Typically, if the rejection region of the p-values is $[0,\Delta]$ where Δ is a fixed threshold, we obtain the following formulations under the mixture model,

$$G_1(\Delta) = G_1(0,\Delta) = F_p(\Delta), \tag{4.8}$$

$$G_0(\Delta) = G_0(0,\Delta) = \Delta. \tag{4.9}$$

In this case, the probability of having a correctly declared positive, denoted by $G_1(\Delta)$, and the probability of having a falsely declared positive, denoted by $G_0(\Delta)$, are the cumulative distribution functions of the p-values under the null $H_{0,i}$ and alternative $H_{1,i}$ respectively. The ratio is therefore

$$\frac{\mathbb{P}(\mathrm{TP})}{\mathbb{P}(\mathrm{FP})} = \frac{G_1(\Delta)}{G_0(\Delta)} = \frac{F_p(\Delta)}{\Delta}. \tag{4.10}$$

For any distribution F_p differentiable on $[0,1]$, it follows that

$$\lim_{\Delta\to 0} \frac{G_1(\Delta)}{G_0(\Delta)} = \lim_{\Delta\to 0} f_p(\Delta). \tag{4.11}$$

As we explained in Section 4.1.2, $\lim_{\Delta\to 0} f_p(\Delta)$ is indeed influenced by the tail distribution of the test statistics.

Example 4.1.5. *The left-side plot in Figure 4.2 compares the ROC curves based on the Gaussian (the dashed blue curve) and Cauchy (the solid red curve) test statistics respectively. The right-side plot shows the corresponding functions $G_1(\Delta)/G_0(\Delta)$, namely $G_1(\Delta)/\Delta$ given the rejection region $\mathcal{R}_{t,\Delta} = [0,\Delta]$. The Gaussian and the Cauchy distributions stand for two regimes in the detection of mixture models. For Gaussian test statistics, $G_1(\Delta)/\Delta$ is monotonically decreasing and the ROC curve is concave, such that with a rejection region $\mathcal{R}_{t,\Delta} = [0,\Delta]$, the ratio of the increment in power over the increment in type I error is maximised at $t = 0$. However, the Cauchy test statistics which satisfy the conditions 4.1.2 and 4.1.3 cannot be detected using the rejection region $\mathcal{R}_{t,\Delta} = [0,\Delta]$, given the fact that the false positives increase as fast as the true positives at $t = 0$. The maximum slope is obtained at a positive value which is greater than zero.*

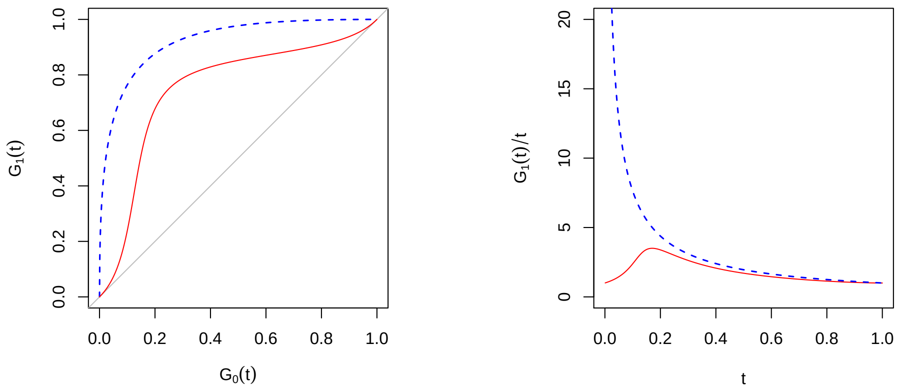


Figure 4.2 – The ROC curve (left-side) and the ratio TPR/FPR (right-side) for Gaussian mixture model (dashed blue curves) and Cauchy mixture model (solid red curves)

Now we set up the criterion of our testing procedure and define the rejection region based on the variation of the true and false discoveries.

Definition 4.1.6 (Significance center)**.** *In the multiple testing problem (4.4), we define the rejection region of the p-values as*

$$\mathcal{R} = \mathcal{R}_{t,\Delta} \subseteq [0,1]\,. \tag{4.12}$$

Define $G_1'(t)$ and $G_0'(t)$ as

$$G_1'(t) = \frac{\partial G_1(t,\Delta)}{\partial \Delta}\bigg|_{\Delta=0}, \tag{4.13}$$

$$G_0'(t) = \frac{\partial G_0(t,\Delta)}{\partial \Delta}\bigg|_{\Delta=0}, \tag{4.14}$$

which can be interpreted as the true positive rate and false positive rate evaluated near threshold t respectively. The optimal threshold t^* is defined as the significance center that maximises the ratio

$$t^* = \arg\max_t \frac{G_1'(t)}{G_0'(t)}\,. \tag{4.15}$$

We use (4.15) as a criterion because it maps out the slope of the ROC curve with the significant center t varying from zero to one, and helps to determine the desired control of error rates. We seek to obtain a significance region such that the number of true positives increases rapidly with only few false positives included. Maximising (4.15) as a function of $t \in [0,1]$ will give a feasible solution to the rejection region.

Recall that step-wise procedures reject the p-values of Gaussian test statistics at $p_i \leq \alpha_i$, which is equivalent to defining the rejection region $\mathcal{R}_{t,\Delta} = [0,\Delta]$. We give the following condition that must be guaranteed to have $t^* = 0$.

Condition 4.1.7 (C3). *$G_1(t)/t$ is decreasing in t on $[0,1]$, where $G_0(t)$ and $G_1(t)$ are defined by (4.8) and (4.9).*

Generally, in multiple testing problems for mixture models, a feasible testing procedure depends on the tail distributions of the test statistic, namely the light-tailed distribution and the heavy-tailed distribution respectively. We conclude the following two regimes formally, where the step-wise multiple testing procedures should be established according to the distributional property of the test statistics.

Lemma 4.1.8. *Consider the rejection region $\mathcal{R}_{t,\Delta}$ of the p-values associated with the multiple testing problem (4.4), we give the following two cases:*

i) Suppose the distribution of the p-values satisfies Condition 4.1.7. Then the rejection region $\mathcal{R}_{t,\Delta} = [0,\Delta]$ is optimal, in the sense that $t^* = 0$.

ii) Suppose the distribution of the test statistics and the p-values, namely F_1 and F_p, satisfy Conditions 4.1.2 and 4.1.3, with Condition 4.1.7 violated. Then any decision rule based on $\mathcal{R}_{t,\Delta} = [0,\Delta]$ is infeasible, in the sense that $t^* > 0$.

Proof. i) For the first case that Condition 4.1.7 is satisfied,

$$\left(\frac{G_1(t)}{t}\right)' \leq 0 \implies G_1'(t) \leq \frac{G_1(t)}{t}.$$

As $G_1(\Delta)/\Delta$ is decreasing, we obtain

$$\arg\max_t \frac{G_1(t)}{t} = 0,$$

and

$$\lim_{t\to 0} \frac{G_1(t)}{t} = \lim_{t\to 0} f_p(t) = \lim_{t\to 0} G_1'(t),$$

and it is easy to conclude that the significance center

$$t^* = \arg\max_t G_1'(t) = 0. \tag{4.16}$$

ii) If Conditions 4.1.2 and 4.1.3 are satisfied, then Condition 4.1.7 is violated. It is easy to show that the significance center $t^* = 0$ contradicts the condition that $G_1(\Delta)/\Delta$ is decreasing. In addition,

$$\lim_{t\to 0} \frac{G_1(t)}{t} \longrightarrow 1 \tag{4.17}$$

implies that the quality of the test based on $\mathcal{R}_{t,\Delta} = [0,\Delta]$ for very small Δ is equally bad as rejecting the hypotheses randomly.

□

Example 4.1.9. *As an example, in Cauchy mixture model we verify that the ratio $G_1(t)/G_0(t)$ is not monotone, so Condition 4.1.7 is violated. We obtain the significant center p-value at*

$$\arg\max_t \frac{\mathrm{d}\mathbb{P}(P_i \le t|H_i = 1)}{\mathrm{d}\mathbb{P}(P_i \le t|H_i = 0)} = \arctan\left(-\sqrt{1+\frac{\mu^2}{4}} - \frac{\mu}{2}\right)\Big/\pi + \frac{1}{2} > 0,$$

which is greater than zero, while at $t = 0$ we obtain $G_1'(0)/G_0'(0) = 1$. This implies that the region containing the majority of true alternatives is always away from zero, and the false positives are increasing as fast as the true positives at a threshold of p-values near zero.

Remark. *The framework of Lemma 4.1.8 coincides with our previous argument that the rejection region based on the p-values should be defined as an interval centered at the mode of $f_p(t)$.*

Up to now we have explained from different perspectives why the classic step-wise threshold methods based on $\mathcal{R}_t = [0, t]$ are not useful for the heavy-tailed multiple testing problems, given the non-adjusted p-values. From now on we start to present our results of the proposed testing procedure. If a classic multiple testing procedure, such as BH, is applied to this heavy-tailed situation, the power is extremely low even though the control of FDR$\le \alpha$ is maintained. We will first discuss the criteria considered.

4.2 Control of FDR and positive FDR

Given a family of m independent hypotheses $\{H_{0,i},\ i = 1,\dots,m\}$, we discuss the limitation of the control of error rate with FDR used alone. We consider the procedures that reject $H_{0,i}$ for $p_i \le \alpha_i$ based on the generic p-values $p_i = \mathbb{P}_{H_{0,i}}(X > x_i)$. Let R denote the cardinality of total rejections, and let V denote the number of false rejections, that is,

$$R = \sum_{i=1}^m R_i = \sum_{i=1}^m \mathbf{1}\{P_i \le \alpha_i\}, \quad V = \sum_{i=1}^m V_i = \sum_{i=1}^m \mathbf{1}\{P_i \le \alpha_i \cap H_i = 0\};$$

the false discovery rate is thus

$$\mathrm{FDR} = \mathbb{E}\left[\frac{V}{R \vee 1}\right].$$

Recall that the BH procedure introduced by Benjamini and Hochberg (1995) controls the FDR by rejecting the nulls $H_{0,(1)}, H_{0,(2)},\dots,H_{0,(k)}$ with the critical k

$$k = \max_{1\le i\le m}\left\{i:\ p_{(i)} \le \frac{i}{m}\alpha\right\},$$

and if no such i exists, one rejects no hypothesis and the sample FDP equals zero.

We make a remark on the effectiveness of the BH procedure by considering the following question:

- Does the quality of the BH procedure rely on the distribution of the test statistics?

In terms of FDR control, it is known that under independence, the BH procedure is a distribution-free approach that controls the FDR regardless of the distribution of the test statistics. However, the quality of the BH procedure does rely on the correct distributional assumptions of the test statistics and the p-values.

Recall that we introduced the operating characteristics to evaluate the quality of the test, and we provide Lemma 4.1.8 such that the effectiveness of the BH procedure is guaranteed when testing Gaussian or other light-tailed components. When the normality of the test statistics is not evident, and in the meantime the BH procedure declares no significance, it is not necessarily true that the alternative components do not exist. We may reasonably suspect that the condition of applying a rejection region $\mathcal{R}_t = [0, t]$ to the p-values, such as Condition 4.1.7, is violated. Thus, we are motivated to discuss the distribution of the test statistics that fails/guarantees the effectiveness of the classic step-down procedures. One example that causes this failure would be the heavy-tailed framework we defined.

Positive FDR.

Given the limitation of the FDR, we study another control of error rate, called the *positive false discovery rate (pFDR)*, which was first defined by Storey (2002), Storey (2003) as:

$$\mathrm{pFDR} = \mathbb{E}\left[\frac{V}{R}\middle| R > 0\right] = \frac{\mathrm{FDR}}{\mathrm{P}(R > 0)}. \tag{4.18}$$

This criterion provides another consideration beyond the FDR control, namely, we should also specify how often we are able to detect the alternatives, before we identify the true positives among the findings. When the methods are not likely to report discoveries, the control of the FDR is not convincing since the positive FDR can be quite high, which means the false discoveries among the rejected hypotheses can be dominant. In other words, pFDR can be quite large although FDR is bounded.

Figure 4.3 shows the simulation result of the rejections declared by the BH procedure given a sample from a Cauchy mixture distribution. We generate $m = 200$ p-values with the fraction $\varepsilon = 0.15$ having a positive shift $\mu = 5$. The p-values from the alternatives are not the ones near zero, and cannot be detected by the BH or the other step-wise procedures we introduced before. The dashed line is the classic threshold with $\alpha = 0.05$. In the current sample, only one smallest p-value $p_{(1)}$ is rejected, though it is in reality from the null. As a consequence, the false discovery proportion with respect to this sample is $\mathrm{FDP}_{\mathrm{BH}} = 1$. In addition, raising the tolerance of false discoveries will not help to detect the true alternatives. We also show the threshold lines with $\alpha = 0.3$ (dotted line) and $\alpha = 0.4$ (dot-dashed line). In order to have a critical k chosen by the step-down procedure, we need a relatively large slope of the threshold curve, which leads to a value of α which is not preferred. The FDR is thus controlled at a high level with a large number of nulls unavoidably rejected.

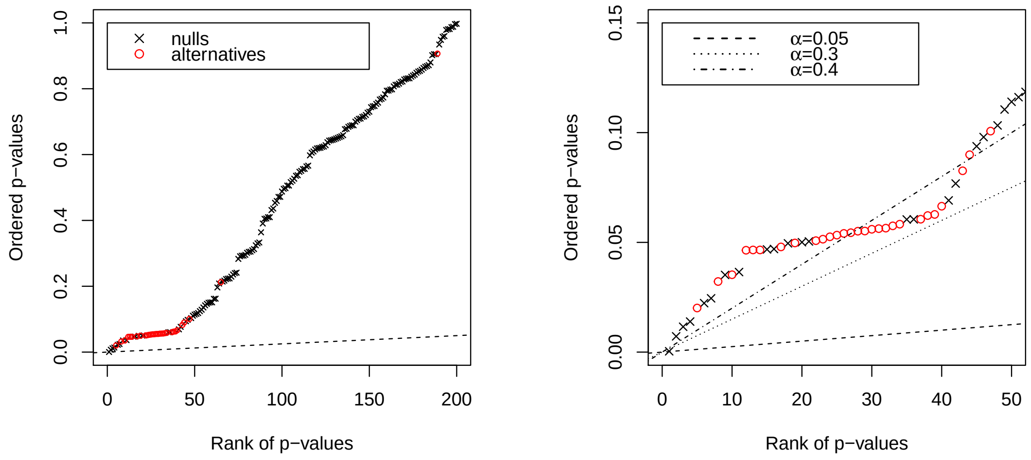


Figure 4.3 – Test of the Cauchy alternatives by the BH procedure *(The right panel is a zoom-in of the left panel with the first quarter of the ordered p-values plotted)*

We take the Cauchy distribution as an example of the heavy-tailed framework, and the BH procedure as an example of the step-down procedures based on the p-values. We provide the following theorem that captures quantitatively the ability to detect alternatives.

Theorem 4.2.1 (BH procedure for Cauchy mixtures)**.** *In the multiple testing problem based on Cauchy test statistics, with the significance level $\alpha_m = O((\log\log(m))^{-1})$ slowly decaying, the probability of declaring at least one rejection by the BH procedure is*

$$\mathbb{P}(R > 0) = 1 - \mathrm{e}^{-\alpha_m} + o\left(\frac{2\mathrm{e}^{-\alpha_m}\alpha_m}{1-\alpha_m}\right) \tag{4.19}$$

which is approximately α_m, regardless of the parametrisation of ε_m and μ_m.

Proof. Based on the mixture model $X_i \overset{\text{i.i.d.}}{\sim} (1-\varepsilon)F_0 + \varepsilon F_1$, the BH procedure rejects the hypotheses up to the last crossing point of the ordered p-values $p_{(1)},\dots,p_{(m)}$ with the threshold sequence $\{\alpha i/m,\ i = 1,\dots,m\}$. The probability of finding no rejection is thus

$$\begin{aligned}\mathbb{P}(R=0) &= \prod_{k=1}^{m} \mathbb{P}\left(p_{(k)} > \frac{k\alpha}{m}\right) \\ &= \prod_{k=1}^{m}\left(1 - \sum_{j=k}^{m}\binom{m}{j}\left[\tilde{F}_p\left(\frac{k\alpha}{m}\right)\right]^j\left[1-\tilde{F}_p\left(\frac{k\alpha}{m}\right)\right]^{m-j}\right),\end{aligned} \tag{4.20}$$

where the mixture distribution

$$\tilde{F}_p(t) = (1-\varepsilon)t + \varepsilon\left[\frac{1}{2} - \frac{1}{\pi}\arctan\left(\tan\left(\frac{\pi}{2} - t\right) - \mu\right)\right] = t + O(t^2)$$

for t close to zero. Therefore, for a fixed α, the probability that at least one hypothesis is rejected can be approximated by

$$\mathbb{P}(R > 0) \approx 1 - \left(1 - \frac{\alpha}{m}\right)^m \longrightarrow 1 - \mathrm{e}^{-\alpha}.$$

Through a more precise calculation we show that, for any $k \leq m$,

$$\begin{aligned}
\prod_{j=1}^{k} \mathbb{P}\left(p_{(j)} > \frac{\alpha j}{m}\right) &= \prod_{j=1}^{k}\left(1 - \sum_{i=0}^{j-1}\binom{m}{i}\left(\frac{\alpha j}{m}\right)^i\left(1 - \frac{\alpha j}{m}\right)^{m-i}\right) \\
&= \prod_{j=1}^{k}\left(\mathrm{e}^{-\alpha j}\sum_{i=0}^{j-1}\frac{\alpha^i j^i}{i!}\right) \\
&= \mathrm{e}^{-\alpha}\prod_{j=2}^{k}\left(\mathrm{e}^{-\alpha j}\left(\mathrm{e}^{\alpha j} - \frac{\alpha^j j^j}{j!} + O\left(\frac{(\alpha j)^{j+1}}{(j+1)!}\right)\right)\right) \qquad (4.21) \\
&= \mathrm{e}^{-\alpha}\prod_{j=2}^{k}\left(1 - \mathrm{e}^{-\alpha j}\frac{\alpha^j j^j}{j!} + O\left(\frac{\mathrm{e}^{-\alpha j}(\alpha j)^{j+1}}{(j+1)!}\right)\right) \\
&= \mathrm{e}^{-\alpha}\left(\left(1 - \frac{2\mathrm{e}^{-2\alpha}\alpha^2}{1-\alpha}\right) + o(\alpha^2\mathrm{e}^{-2\alpha})\right)
\end{aligned}$$

Thus, the probability of having at least one rejection in BH procedure is approximated as

$$\mathbb{P}(R > 0) = 1 - \mathrm{e}^{-\alpha}\left(1 - \frac{2\mathrm{e}^{-2\alpha}\alpha^2}{1-\alpha}\right) + o(\alpha^2\mathrm{e}^{-3\alpha}). \tag{4.22}$$

For reasonably small α, the probability of rejecting at least one hypothesis is

$$\mathbb{P}(R > 0) = 1 - \mathrm{e}^{-\alpha} + o\left(\frac{2\mathrm{e}^{-\alpha}\alpha}{1-\alpha}\right), \tag{4.23}$$

and the value of (4.23) is almost α regardless of the fraction ε and the size μ of the non-null effects.

□

Remark. *The pFDR is controlled at the level*

$$\text{pFDR} = \frac{\text{FDR}}{\mathbb{P}(R > 0)} \leq \frac{\alpha}{1 - \mathrm{e}^{-\alpha} + o\left(\frac{2\mathrm{e}^{-\alpha}\alpha}{1-\alpha}\right)} \approx 1$$

with small α. As a consequence, the BH procedure either finds no rejections, or has a large false discovery proportion among the rejected null hypotheses. In other words, the FDR control is maintained due to a small probability $\mathbb{P}(R > 0)$.

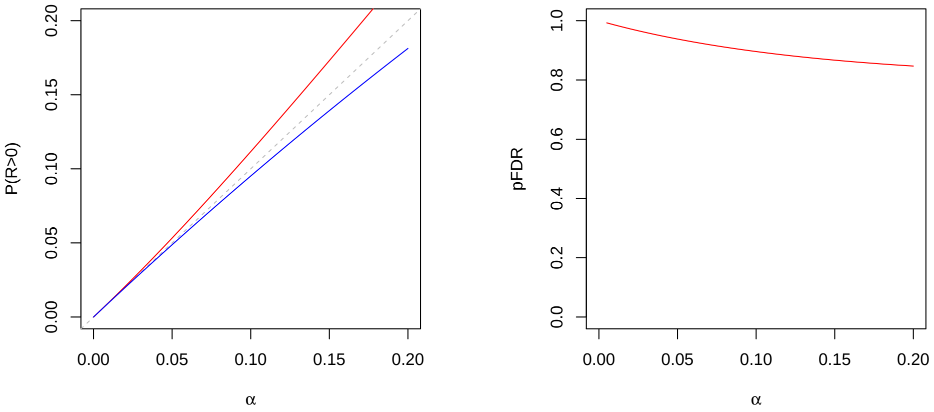


Figure 4.4 – The actual bound of pFDR with respect to $\mathbb{P}(R > 0)$

The left-side plot in Figure 4.4 shows the approximate value of $\mathbb{P}(R > 0)$ utilising the BH procedure in Cauchy mixture model. The blue curve is a lower bound $1 - \mathrm{e}^{-\alpha}$, and the red curve is the approximation given by the equation (4.23). It can be seen that for small α, the probability $\mathbb{P}(R > 0)$ is approximately α. The right-side plot shows the actual control of pFDR, which is given by $\mathrm{FDR}/\mathbb{P}(R > 0)$. Together with the simulation in Figure 4.3 we can conclude that the BH procedure finds discoveries with a low probability, and commits a great proportion of false discoveries among the declared significant ones.

Remark. *The phenomenon illustrated in Theorem 4.2.1 cannot be improved by changing the slope or the threshold curve in the step-up or step-down procedures where the rejections start from the smallest p-values. The approximation in the proof is guaranteed by Condition 4.1.3, and Cauchy is a heavy-tailed example. In the following section we propose a decision rule defined by the rejection region $\mathcal{R}_\vartheta$ centered at $\vartheta > 0$, that commits fewer false rejections and detects more true alternatives.*

4.3 Filtering approaches for multiple testing

We have explained from different perspectives the difficulties that heavy-tailed test statistics bring into the definition of the rejection region in multiple testing problems. Now we introduce our testing procedure, which is aimed at formulating the local concentration and the gaps of the ordered p-values.

In general, we aim at solving the following two problems in multiple testing:

i) *Where are the alternatives most likely to occur?*

ii) *How many hypotheses should be rejected?*

In order to answer the two questions, we propose a filtering method to select a subset of the p-values that are presumably alternatives. In a follow-up step, we determine the p-value that maximises the increasing of true discoveries over the increasing of false discoveries as defined by Definition 4.1.6. This is equivalent to maximising the density of the observed p-values over $[0,1]$.

4.3.1 Outline of the method

We first give an outline of our method. Suppose $\tilde{f}_p(t) = (1-\varepsilon) + \varepsilon f_p(t)$ is the marginal density of the p-values drawn from the mixtures, and there exists a unique mode ϑ such that

$$\tilde{f}_p(\vartheta) = \max_t \tilde{f}_p(t)\,.$$

According to our criterion described by the Definition 4.1.6, we seek to find the "most significant" p-values from the sample, that is, to estimate the mode and provide a good interpretation as well.

One can consider the kernel density estimate of $\tilde{f}_p$:

$$\widehat{\tilde{f}}_{p_{m,h}}(t) = \widehat{\tilde{f}}_{p_{m,h}}(t;\, p_1, \ldots, p_m) = \frac{1}{mh}\sum_{i=1}^{m} K\left(\frac{t-p_i}{h}\right),$$

where the bandwidth h is a function of m and $p_1, \ldots, p_m$, and K is a well-chosen kernel. Efron et al. (2001), Efron (2004), Genovese and Wasserman (2004) and Jin and Cai (2007) have contributed to the estimation of the sample distribution of the z-values and the p-values drawn from Gaussian mixture models. However, their methods are not ideal to solve heavy-tailed multiple testing problems, due to the large variation of the true alternative test statistics.

In order to have an accurate estimator of the mode ϑ of f_p from the mixtures, we propose a method that reduces the randomness caused by the majority of the null p-values. Suppose $\xi_m \in (0,1)$ is a tuning parameter such that we filter the observed p-values at the level ξ_m before we estimate the sample distribution. In other words, we delete $100(1-\xi_m)\%$ of the original p-values according to a filtering rule such that the true nulls are more likely to be deleted. In terms of the filtered p-values, we can *(i)* derive the asymptotic proportion of true and false discoveries, and *(ii)* use the filtered p-values to obtain a precise estimate of ϑ, and build up a finite-sample rejection region. We will provide results for both cases.

4.3.2 Filtering the p-values

Our filtering approach is designed to partition the sample p-values into a preferably alternative subset and a presumably uniform subset.

As a toy example, we consider a random elimination procedure that deletes the p-values randomly from the mixture P^m. In this case, for any pre-specified filtering parameter $\xi \in (0,1)$, the remaining p-values are still a mixture of U and F_p with fractions $1-\varepsilon$ and ε respectively. The proportion of the alternatives among the mixture is neither reduced nor enlarged. We use this filter as a worst case that provides no improvement in locating the alternatives or estimating the mode.

In general, let $p^m = \{p_1, \dots, p_m\}$ denote the realised p-values given the test statistics for the hypotheses H_i, $i = 1, \dots, m$. Define a filtering operator $\mathscr{T}$ such that

$$\mathscr{T}\left(p^m\right) = \mathscr{T}\left(\{p_1, \dots, p_m\}\right) = \left\{p_{(i)} : i \in I_s^{\mathscr{T}}\right\} = \mathcal{S}^{\mathscr{T}}, \tag{4.24}$$

where $\mathcal{S}^{\mathscr{T}}$ denotes the set of the remaining p-values, of which the ranks among the whole sample is denoted by $I_s^{\mathscr{T}}$, with $|I_s^{\mathscr{T}}| = m_1^{\mathscr{T}}$. Let $\mathcal{U}^{\mathscr{T}}$ denote the set of the excluded p-values

$$\mathcal{U}^{\mathscr{T}} = \left\{p_{(i)}, i \in I_u^{\mathscr{T}}\right\}, \tag{4.25}$$

where $I_u^{\mathscr{T}}$ is the set of ranks with $|I_u^{\mathscr{T}}| = m_0^{\mathscr{T}}$, such that

$$I_s^{\mathscr{T}} \cap I_u^{\mathscr{T}} = \varnothing, \quad I_s^{\mathscr{T}} \cup I_u^{\mathscr{T}} = \{1, \dots, m\}.$$

We use $\mathcal{U}$ for the presumably uniform p-values that we desire to delete, and use $\mathcal{S}$ for the remaining ones that are more likely to be true alternatives. Our notation is consistent with the true positives "S" and true negatives "U" used by Benjamini and Hochberg in classic multiple testing. In addition, denote the order statistics of $\mathcal{S}^{\mathscr{T}}$ by

$$\mathcal{S}^{\mathscr{T}} = p^m \setminus \mathcal{U}^{\mathscr{T}} = \left\{p_1^*, \dots, p_{m_1^{\mathscr{T}}}^*\right\} \tag{4.26}$$

for simplicity of notation, such that $p_1^* \le p_2^* \le \cdots \le p_{m_1^{\mathscr{T}}}^*$.

Now we investigate the distributions of $\mathcal{U}^{\mathscr{T}}$ and $\mathcal{S}^{\mathscr{T}}$ and compare them to the uniform distribution U and the alternative F_p.

Suppose there are some falsely selected p-values in the filter $\mathscr{T}$ of which the two types are the remaining null p-values and the deleted alternatives. Let $\mathrm{FE}^{\mathscr{T}}$ be the set of the *falsely excluded* p-values, which is defined as

$$\mathrm{FE}^{\mathscr{T}} = \left\{P_i \in P^m : P_i \in (\mathcal{U}^{\mathscr{T}} \cap P^{I_1})\right\},$$

while $\mathrm{FI}^{\mathscr{T}}$ denotes the set of the *falsely included* p-values in the remaining set

$$\mathrm{FI}^{\mathscr{T}} = \left\{P_i \in P^m : P_i \in (\mathcal{S}^{\mathscr{T}} \cap P^{I_0})\right\},$$

with P^{I_0} and P^{I_1} being the sets of true nulls and alternatives respectively. Both types of false selections appear in the filtration and contaminate the distribution of $\mathcal{U}^{\mathscr{T}}$ and $\mathcal{S}^{\mathscr{T}}$, which leads to the fact that the distribution of $\mathcal{S}^{\mathscr{T}}$ is still a mixture of the nulls and alternatives. The goal of a filtering method is to enlarge the proportion of the alternatives, such that the estimate based on $\mathcal{S}^{\mathcal{T}}$ is consistent with F_p. In addition, the asymptotic properties of the filter can be derived as m tends to infinity.

Remark. *For any $P_i^* \in \mathcal{S}^{\mathcal{T}}$, we define the distribution*

$$P_i^* \sim (1-w)U^* + wF_p^*,$$

of which w is an oracle proportion. $\mathcal{S}_\xi = \left\{P^*_{(i)}, i \in I_s^{\mathcal{T}}\right\}$ is not an independent sample, which makes it a subtle problem to discuss the finite-sample control of the filtering properties. We consider the asymptotic behaviour of the filter.

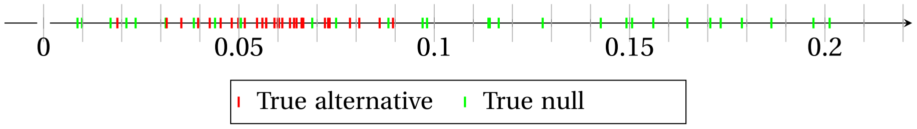


Figure 4.5 – Ordered p-values of Cauchy mixture

Figure 4.5 is a plot of the null (in green) and alternative (in red) p-values drawn from a Cauchy mixture distribution, with the number of hypotheses $m = 200$, the frequency of the alternatives $\varepsilon = 0.15$ and the positive shift $\mu = 3$. The plot is a zoomed-in version in the region $[0, 0.2]$ such that the most informative ones can be seen. The p-values from the nulls are roughly uniform, while those from the alternatives have a local concentration far from zero. A filtering approach is to delete the p-values that are presumably uniforms, and preserve most of the true alternatives.

4.3.3 Randomised filtering

We first propose a filter that deletes the p-values randomly from fixed grids, and we prove that it enlarges the proportion of the true alternatives among the remaining p-values.

Define a filter $\mathcal{T}^R$ that works as follows.

Given a pre-specified filtering parameter $\xi \in (0,1)$, we seek to delete $m_\xi^{\mathcal{T}} = \lceil (1-\xi)m \rceil$ p-values with this procedure. Define the bins $D_j = \left[(j-1)/m_\xi^{\mathcal{T}}, j/m_\xi^{\mathcal{T}}\right)$ for $j = 1, \ldots, m_\xi^{\mathcal{T}} - 1$ and $D_{m_\xi^{\mathcal{T}}} = \left[(m_\xi^{\mathcal{T}} - 1)/m_\xi^{\mathcal{T}}, 1\right]$ with the length $d = 1/m_\xi^{\mathcal{T}}$ and the mid-points $\left\{c_1, \ldots, c_{m_\xi^{\mathcal{T}}}\right\}$.

For $j = 1, \ldots, m_\xi^{\mathcal{T}}$, the filter $\mathcal{T}^R$ looks for the observed p-values located in D_j.

- If there exists at least one p-value in D_j, $\mathscr{T}^R$ deletes one of them randomly.

- If no such p-value exists, $\mathscr{T}^R$ deletes nothing in D_j and moves to D_{j+1}.

Compared to the random elimination over all the p-values, $\mathscr{T}^R$ conducts random exclusion in each evenly spaced interval, which guarantees that the alternative p-values with a local concentration around the mode are not much influenced. Note that this requires the parametrisation of ε_m and μ_m in the asymptotic detectable region. In addition, a subspace where the filter classifies the nulls and the alternatives is defined.

Theorem 4.3.1 (Asymptotic filtering)**.** *Consider the Cauchy mixture model*

$$(1-\varepsilon_m)F_0(x)+\varepsilon_m F_0(x-\mu_m),$$

where $\varepsilon_m = m^{-\gamma}$ and $\mu_m = m^r$. The expected ratio of the false exclusions committed by $\mathscr{T}^R$ over the total number of the true alternatives converges to zero, that is,

$$\frac{\mathbb{E}\left|FE^{\mathscr{T}}\right|}{m_1} \longrightarrow 0, \quad m \to \infty, \tag{4.27}$$

if the parameters (γ, r) satisfy

$$r > 1-\frac{\gamma}{2}. \tag{4.28}$$

Proof. The expected value of the false exclusions is

$$\mathbb{E}\left|\mathrm{FE}^{\mathscr{T}}\right| = \mathbb{E}\sum_{i:H_i=1}\mathbb{1}\left\{P_i \in \mathfrak{U}^{\mathscr{T}}\right\}. \tag{4.29}$$

Considering the intervals $\left\{D_j,\ j=1,\dots,m_\xi^{\mathscr{T}}\right\}$ on $[0,1]$, we obtain

$$\begin{aligned}
\mathbb{E}\left|\mathrm{FE}^{\mathscr{T}}\right| &= \mathbb{E}\sum_{i:H_i=1}\mathbb{1}\left\{P_i \in \mathfrak{U}^{\mathscr{T}}\right\}\\
&= \sum_{i:H_i=1}\sum_{j=1}^{m_\xi^{\mathscr{T}}}\mathbb{P}\left(P_i \in (\mathfrak{U}^{\mathscr{T}}\cap D_j)\right)\\
&\approx m_1\sum_{j=1}^{m_\xi^{\mathscr{T}}}\frac{\varepsilon f_p(c_j)}{(1-\varepsilon)+\varepsilon f_p(c_j)}\\
&\approx m_1 m_\xi^{\mathscr{T}}\int_0^1\frac{\varepsilon f_p(t)}{(1-\varepsilon)+\varepsilon f_p(t)}\,\mathrm{d}t\\
&= \varepsilon m_1 m_\xi^{\mathscr{T}}\int_0^1\frac{1}{(1-\varepsilon)\frac{1}{f_p(t)}+\varepsilon}\,\mathrm{d}t.
\end{aligned}$$

Since the function $1/f_p$ is bounded on $[0,1]$ and has a single peak at

$$t = \frac{1}{\pi}\left(\arctan\left(-\sqrt{1+\frac{\mu^2}{4}}-\frac{\mu}{2}\right)\right)+1 \geq \frac{1}{2},$$

we have the integral over $[0,1]$ upper bounded by twice the integral over the right half. Thus,

$$\begin{aligned}
\mathbb{E}\left|\mathrm{FE}^{\mathcal{T}}\right| &\leq 2\varepsilon m_{\xi}^{\mathcal{T}} m_1 \int_{\frac{1}{2}}^{1} \frac{1}{(1-\varepsilon)\frac{1+(\tan(\pi/2-\pi t)-\mu)^2}{1+\tan^2(\pi/2-\pi t)}+\varepsilon}\,\mathrm{d}t \\
&= \frac{2\varepsilon m_{\xi}^{\mathcal{T}} m_1}{\pi} \int_{0}^{\frac{\pi}{2}} \frac{1}{(1-\varepsilon)\frac{1+(\tan x+\mu)^2}{1+\tan^2 x}+\varepsilon}\,\mathrm{d}x \\
&\leq \frac{2\varepsilon m_{\xi}^{\mathcal{T}} m_1}{\pi} \int_{0}^{\frac{\pi}{2}} \frac{1}{(1-\varepsilon)(1+\mu^2)\cos^2 x+\varepsilon}\,\mathrm{d}x \\
&= \frac{2\varepsilon m_{\xi}^{\mathcal{T}} m_1}{\pi} \int_{0}^{\frac{\pi}{2}} \frac{\sec^2 x}{(1-\varepsilon)(1+\mu^2)+\varepsilon(1+\tan^2 x)}\,\mathrm{d}x \\
&= \frac{2\varepsilon m_{\xi}^{\mathcal{T}} m_1}{\pi} \int_{0}^{\infty} \frac{1}{(1-\varepsilon)(1+\mu^2)+\varepsilon+\varepsilon y^2}\,\mathrm{d}y \\
&= \frac{2\varepsilon m_{\xi}^{\mathcal{T}} m_1}{\pi} \frac{1}{(1-\varepsilon)(1+\mu^2)+\varepsilon} \int_{0}^{\infty} \frac{1}{1+\left(\frac{\sqrt{\varepsilon}y}{\sqrt{(1-\varepsilon)(1+\mu^2)+\varepsilon}}\right)^2}\,\mathrm{d}y \\
&= \varepsilon m_{\xi}^{\mathcal{T}} m_1 \frac{1}{\sqrt{\varepsilon}\sqrt{(1-\varepsilon)(1+\mu^2)+\varepsilon}}.
\end{aligned}$$

Recall that we consider the parametrisation

$$\varepsilon_m = m^{-\gamma}, \quad \mu_m = m^r$$

with (γ, r) in the asymptotically detectable region. Therefore, the expected proportion

$$\frac{\mathbb{E}\left|\mathrm{FE}^{\mathcal{T}}\right|}{m_1} \leq \frac{\sqrt{\varepsilon} m_{\xi}^{\mathcal{T}}}{\sqrt{(1-\varepsilon)(1+\mu^2)+\varepsilon}} = \frac{1-\xi}{m^{r+\gamma/2-1}(1+o(1))} \longrightarrow 0 \tag{4.30}$$

as $m \to \infty$, if

$$r > 1 - \frac{\gamma}{2}. \tag{4.31}$$

□

This result provides a subspace in the asymptotically detectable region in Chapter 3. Using this randomised filtering procedure, the true alternatives can be well kept in the sequence $\mathcal{S}^{\mathcal{T}}$ as described in (4.27), while the total number of deleted p-values is designed to be $100(1-\xi_m)\%m$.

We can equivalently prove that the expected proportion of false inclusions tends to zero, such that the true alternative p-values dominate the filtered sequence $\mathcal{S}^{\mathcal{T}}$. The mode estimator based on $\mathcal{S}^{\mathcal{T}}$ is thus consistent for the true mode of f_p.

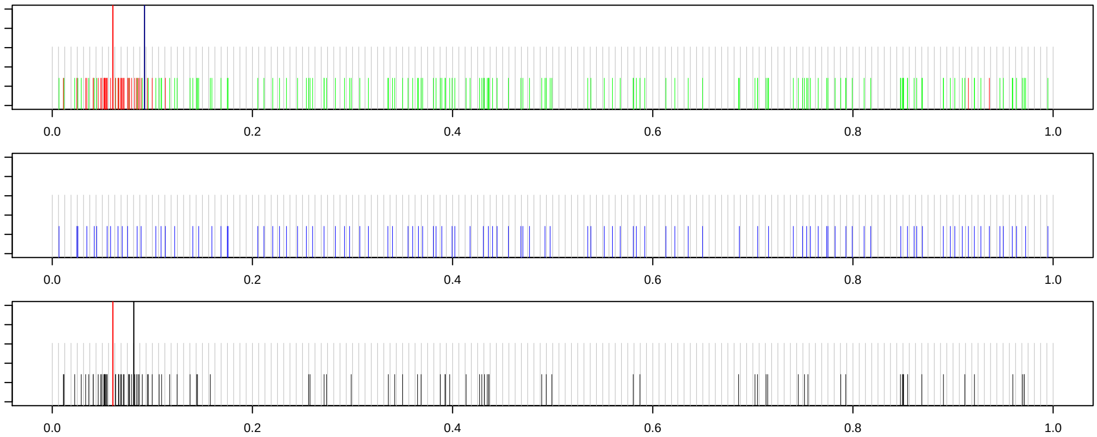


Figure 4.6 – The randomised filtering approach for the p-values corresponding to Cauchy mixtures *(with $m = 200$, $\varepsilon = 0.15$, $\mu = 3$ and filtering parameter $\xi = 0.1$)*

Figure 4.6 shows the allocation of the p-values before and after the randomised filtering $\mathcal{T}^R$. We take a sample of the mixture of null and alternative p-values with $m = 200$, frequency $\varepsilon = 0.15$ and positive shift $\mu = 3$. The plot on the top is the oracle allocation of the mixture, of which the green bars are the nulls and the red bars stand for the true alternatives. The theoretical mode given by equation (4.5) is computed and plotted as the vertical red line. With $\xi = 0.1$, the grids in gray provide the intervals $\left\{D_j,\ j = 1, \ldots, m_\xi^{\mathcal{T}}\right\}$, of width $d = 1/m_\xi^{\mathcal{T}} = 1/180$. The plot in the middle shows the deleted p-values, namely $\mathcal{U}^{\mathcal{T}}$, selected by the randomised filter $\mathcal{T}^R$. All the exclusions are based on the mixture with no knowledge on the distribution of the nulls and the alternatives. Both distributions may contribute to the selection, since the difference is invisible through the observations. The plot at the bottom gives $\mathcal{S}^{\mathcal{T}}$, the remaining p-values left out by the filter, of which the estimated mode is shown by the vertical line in black.

Although the estimated mode from $\mathcal{S}^{\mathcal{T}}$ is visibly better than the estimate based on the whole sample P^m, it may still differ from the theoretical mode (in red) due to randomness in finite-sample studies. This is because the width of the intervals $1/m_\xi$ can be very small when the parameter ξ gets close to zero. Given the randomness of the observations, the $\mathcal{T}^R$ filter deletes less than we desired. The number of eliminations

$$m_0^{\mathcal{T}} = \left|\mathcal{U}^{\mathcal{T}}\right| = \sum_{j=1}^{m_\xi^{\mathcal{T}}} \mathbb{1}\left\{ \sum_{i \in I_u^{\mathcal{T}}} \mathbb{1}\left\{P_{(i)} \in D_j\right\} \geq 1 \right\} \tag{4.32}$$

is upper bounded by an ideal case in which each interval D_j has at least one p-value located.

Therefore,

$$\left|\mathcal{U}^{\mathcal{T}}\right| \le m_\xi^{\mathcal{T}}.$$

It follows that the value of (4.29) for finite m is upper bounded by the ideal case $\left|\mathcal{U}^{\mathcal{T}}\right| = m_\xi^{\mathcal{T}}$; that is, for any $j = 1, \dots, m_\xi^{\mathcal{T}}$, $\exists i \in I_u^{\mathcal{T}}$ such that $P_{(i)} \in D_j$. Additional calculation shows that

$$\mathbb{P}\left(\left|\mathcal{U}^{\mathcal{T}}\right| = m_\xi^{\mathcal{T}}\right) \longrightarrow 0, \quad m \to \infty.$$

Thus, the filter $\mathcal{T}^R$ might delete much less than desired.

A better solution we propose is fixed-length filtering that excludes an exact number of p-values by examining neighbouring intervals when one interval turns out to be empty.

4.3.4 fixed-length filtering

In order to have the number of exclusions fixed at a desired value, we propose a rule to tackle the empty intervals occurring when the null p-values are not quite evenly spaced. We define a selection rule using the minimal distance from the p-values to the center of each interval, which takes the advantage of the uniform distribution of the p-values under the null.

With a filtering parameter $\xi \in (0,1)$, the number of the excluded p-values is determined to be $|\mathcal{U}^{\mathcal{T}}| = m_0^{\mathcal{T}} = \lceil (1-\xi)m \rceil$, and the number of the p-values left is thus $|\mathcal{S}^{\mathcal{T}}| = m_1^{\mathcal{T}} = \lceil \xi m \rceil$. We define a filter $\mathcal{T}_\xi^F$ that works as follows.

- For $j = 1$, denote the excluded p-value by p_1^ξ, of which the distance to the center of the first grid D_1 is minimised, that is,

$$p_1^\xi = \underset{p_i}{\arg\min} \left| p_i - \frac{1/2}{m_0^{\mathcal{T}}} \right|.$$

- For $j = 2, \dots, m_0^{\mathcal{T}}$, let

$$\mathcal{U}_{j-1}^{\mathcal{T}} = \bigcup_{i=1}^{j-1} \left\{ p_i^\xi \right\}$$

  denote the set of deleted p-values in the first $j-1$ steps. Then the j-th element added to $\mathcal{U}_j^{\mathcal{T}}$ is defined by

$$p_j^\xi = \underset{p_i \notin \mathcal{U}_{j-1}^{\mathcal{T}}}{\arg\min} \left| p_i - \frac{j - \frac{1}{2}}{m_0^{\mathcal{T}}} \right|. \tag{4.33}$$

Let $\mathcal{U}_{m_0^{\mathcal{T}}}^{\mathcal{T}} = \mathcal{U}^{\mathcal{T}}$ denote the set of excluded p-values resulting from the filter $\mathcal{T}_\xi^F$, i.e.,

$$\mathcal{U}^{\mathcal{T}} = \left\{ p_{(i)} : i \in I_u^{\mathcal{T}} \right\} = \left\{ p_1^\xi, p_2^\xi, \dots, p_c^\xi \right\},$$

which consists of the closest p-values to the uniform quantiles.

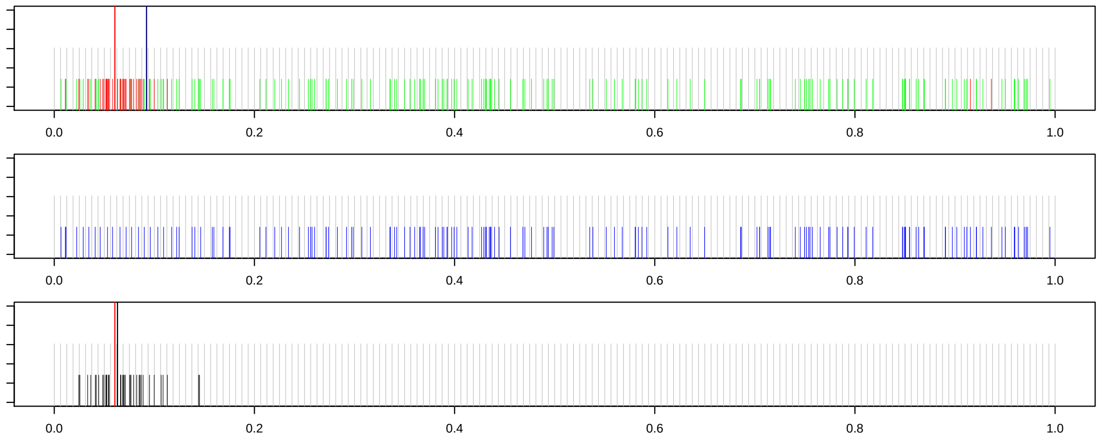


Figure 4.7 – The fixed-length filtering approach for the p-values corresponding to Cauchy mixtures *(with* $m = 200$*,* $\varepsilon = 0.15$*,* $\mu = 3$ *and filtering parameter* $\xi = 0.1$*)*

Figure 4.7 shows how the fixed-length filtering method $\mathcal{T}_\xi^F$ deletes the p-values that are highly likely to be uniformly distributed. With $m = 200$, $\varepsilon = 0.15$ and $\mu = 3$, the first plot is the mixture of the p-values from the nulls and the alternatives, in green and red respectively. The vertical line in red shows the true mode of the alternative p-values, which is unknown and to be estimated from the observations. Given the filtering parameter $\xi = 0.10$, the intervals are evenly spaced with the width $d = 1/m_0^{\mathcal{T}} = 1/180$, shown by the vertical lines in gray. The second plot shows the deleted p-values selected by the fixed-length filter $\mathcal{T}_\xi^F$. Based on the same simulated realisation, the fixed-length filter apparently deletes more p-values than the randomised filter. The last plot shows the remaining p-values $\mathcal{S}^{\mathcal{T}}$, of which the estimated mode is shown by the vertical line in black. The estimated mode from $\mathcal{S}^{\mathcal{T}}$ is consistent of the theoretical mode plotted in red. Later on we will show the consistency of the mode estimate.

Remark. *We propose the fixed-length filter* $\mathcal{T}_\xi^F$ *because we desire to guarantee how many* p*-values are to be excluded according to a spacing rule which is not far from the uniform. Note that* ξ *is not necessary, although preferred, to be a good estimator of the true proportion* ε*. We propose two ways to choose* ξ *in practice.*

i) Since the filter $\mathcal{T}_\xi^F$ *has the false exclusion* $FE^{\mathcal{T}}$ *bounded, which leads to the tendency of maintaining the majority of the true alternatives, it is reasonable to use a relatively small* ξ *to delete as many* p*-values as we expect to be from the nulls. We suggest to choose a moderately small* ξ *less than or equal to an estimator of the true proportion of the alternatives, that is,* $\xi \le \hat{\varepsilon}$*, while taking into account that the left-out* p*-values are sufficient for estimating the mode.*

ii) The parameter ξ *can be chosen by a recurrence procedure that picks the value of* ξ *that stabilises the estimate of the mode. With an initial* $\xi = 0.5$ *for example, the mode of* $\mathcal{S}^{\mathcal{T}}$

is estimated repeatedly with ξ decreasing. Take the optimal ξ such that the sequence of estimate $\hat{\vartheta}$ converges.

Now we compare the deleted p-values $\mathcal{U}^{\mathcal{T}}$ to the uniform distribution.

We first investigate a subset of $\mathcal{U}^{\mathcal{T}}$. Recall that

$$c_j = \frac{j - 1/2}{m_0^{\mathcal{T}}}, \quad j = 1, \dots, m_0^{\mathcal{T}},$$

is the mid point of D_j, and define

$$Q_j = \operatorname*{argmin}_{P_i \in P^m} |P_i - c_j| \tag{4.34}$$

to be the nearest p-value to the j-th mid point, for $j = 1, \dots, m_0^{\mathcal{T}}$, where duplicates are allowed. Let

$$\mathcal{U}_Q^{\mathcal{T}} = \left\{Q_1, Q_2, \dots, Q_{m_0^{\mathcal{T}}}\right\}$$

denote the sequence of Q_j's defined by (4.34). We first compare $\mathcal{U}_Q^{\mathcal{T}}$ to $\mathcal{U}^{\mathcal{T}}$ and investigate the difference between the two sequences.

Theorem 4.3.2. *Suppose $P^m = \{P_1, \dots, P_m\}$ are i.i.d. p-values following the marginal distribution*

$$P_i \sim (1 - \varepsilon)U + \varepsilon F_p, \quad i = 1, \dots, m,$$

with F_p satisfying Condition 4.1.3. Then the se of excluded p-values $\mathcal{U}^{\mathcal{T}}$ given by the filter $\mathcal{T}_\xi^F$ has a subsequence $\mathcal{U}_Q^{\mathcal{T}}$ defined by (4.34), which is asymptotically uniformly distributed, in the sense that

$$\lim_{m \to \infty} \sup_{t \in (0,1)} \left|F_{Q, m_0^{\mathcal{T}}}(t) - t\right| = 0, \tag{4.35}$$

where

$$F_{Q, m_0^{\mathcal{T}}}(t) = \frac{1}{m_0^{\mathcal{T}}} \sum_{i=1}^{m_0^{\mathcal{T}}} \mathbb{1}(Q_i \le t)$$

is the empirical distribution of the excluded p-values.

In order to prove this theorem, we utilise the exchangeability of the set of permuted p-values $\left\{P_{\pi(1)}^\xi, P_{\pi(2)}^\xi, \dots, P_{\pi(m_0^{\mathcal{T}})}^\xi\right\}$, where $\left(\pi(1), \pi(2), \dots, \pi(m_0^{\mathcal{T}})\right)$ is a random permutation of $(1, 2, \dots, m_0^{\mathcal{T}})$. For simplicity of notation, we refer to $\left\{P_{\pi(1)}^\xi, P_{\pi(2)}^\xi, \dots, P_{\pi(m_0^{\mathcal{T}})}^\xi\right\}$ as $\pi \circ \mathcal{U}^{\mathcal{T}} = \pi \circ \left\{P_1^\xi, P_2^\xi, \dots, P_{m_0^{\mathcal{T}}}^\xi\right\}$. The filtering approach gives the same output, namely $\mathcal{U}^{\mathcal{T}}$ and $\pi \circ \mathcal{U}^{\mathcal{T}}$, for a fixed sequence P^m. In this case, the exchangeability of $\mathcal{U}^{\mathcal{T}}$ and $\mathcal{S}^{\mathcal{T}}$ is respected.

Before giving the proof of the theorem, we present the following properties of the excluded p-values $\mathcal{U}^{\mathcal{T}}$ and $\mathcal{U}_Q^{\mathcal{T}}$.

Lemma 4.3.3. *Given a fixed sequence P^m from the mixture model, the filtering procedures defined by (4.33) and (4.34) delete the sequences $\mathcal{U}^{\mathcal{T}}$ and $\mathcal{U}^{\mathcal{T}}_Q$ respectively. It follows that*

$$\mathcal{U}^{\mathcal{T}}_Q \subseteq \mathcal{U}^{\mathcal{T}}, \tag{4.36}$$

with equality if and only if all the Q_j's are distinct.

Proof. For any $Q_j \in \mathcal{U}^{\mathcal{T}}_Q$, it follows that $Q_j \in P^m = \mathcal{U}^{\mathcal{T}}_{j-1} \cup \left(P^m \backslash \mathcal{U}^{\mathcal{T}}_{j-1}\right)$,

$$\begin{aligned}\mathbb{1}\left\{Q_j \in \mathcal{U}^{\mathcal{T}}\right\} &= \mathbb{1}\left\{Q_j \in \mathcal{U}^{\mathcal{T}} | Q_j \in \mathcal{U}^{\mathcal{T}}_{j-1}\right\} \mathbb{1}\left\{Q_j \in \mathcal{U}^{\mathcal{T}}_{j-1}\right\} + \mathbb{1}\left\{Q_j \in \mathcal{U}^{\mathcal{T}} | Q_j \notin \mathcal{U}^{\mathcal{T}}_{j-1}\right\} \mathbb{1}\left\{Q_j \notin \mathcal{U}^{\mathcal{T}}_{j-1}\right\} \\ &= \mathbb{1}\left\{Q_j \in \mathcal{U}^{\mathcal{T}}_{j-1}\right\} + \mathbb{1}\left\{Q_j \in \mathcal{U}^{\mathcal{T}} | Q_j \notin \mathcal{U}^{\mathcal{T}}_{j-1}\right\} \mathbb{1}\left\{Q_j \notin \mathcal{U}^{\mathcal{T}}_{j-1}\right\}.\end{aligned}$$

Since

$$\left|Q_j - c_j\right| \leq \min_{P_i \in P^m \backslash \mathcal{U}^{\mathcal{T}}_{j-1}} \left|P_i - c_j\right|,$$

we obtain

$$Q_j \in \mathcal{U}^{\mathcal{T}}_{j-1} \cup \left\{P^{\xi}_j\right\} = \mathcal{U}^{\mathcal{T}}_j, \quad j = 1, \ldots, m.$$

Therefore $\mathcal{U}^{\mathcal{T}}_Q \subseteq \mathcal{U}^{\mathcal{T}}$.

□

Lemma 4.3.4 (Monotonicity). *The rule defined by (4.34) selects a non-decreasing sub-sequence of $\mathcal{U}^{\mathcal{T}}$ such that*

$$Q_1 \leq Q_2 \leq \cdots \leq Q_{m_0^{\mathcal{T}}}. \tag{4.37}$$

Proof. Let $d = 1/m_0^{\mathcal{T}}$ be the spacing of the grid. If there exists j such that $Q_{j+1} < Q_j$, then by definition

$$\left|Q_j - c_j\right| < \left|Q_{j+1} - c_j\right|, \tag{4.38}$$

$$\left|Q_{j+1} - c_{j+1}\right| < \left|Q_j - c_{j+1}\right|, \tag{4.39}$$

the first equation of which leads to

$$\left|Q_j - c_j\right| < \left|Q_{j+1} - c_{j+1} + c_{j+1} - c_j\right| \leq \left|Q_{j+1} - c_{j+1}\right| + d < \left|Q_j - c_{j+1}\right| + d.$$

If $Q_j \geq c_{j+1}$, then $Q_j > c_j$, so

$$Q_j - c_j < Q_j - c_{j+1} + d$$

gives a contradiction by stating that $d < d$. Therefore, $Q_j < c_{j+1}$. On the other hand, the second equation leads to

$$\left|Q_{j+1} - c_{j+1}\right| < \left|Q_j - c_j + c_j - c_{j+1}\right| \leq \left|Q_j - c_j\right| + d < \left|Q_{j+1} - c_j\right| + d,$$

which is true only when $Q_{j+1} > c_j$. Combining the two conclusions we obtain that the only case that satisfies the two arguments is

$$c_j < Q_{j+1} < Q_j < c_{j+1}\,.$$

However, this conflicts with the definition of Q_j and Q_{j+1} guaranteed by (4.34). Therefore, $Q_{j+1} \geq Q_j$ for $j = 1, \ldots, m_0^{\mathcal{T}} - 1$. □

Now we investigate the sequence $\mathcal{U}_\xi \backslash \left\{Q_1, Q_2, \ldots, Q_{m_\xi}\right\}$, which consists of the different elements between the two excluded sequences $\mathcal{U}^{\mathcal{T}}$ and $\mathcal{U}_Q^{\mathcal{T}}$. The questions we are interested in are:

i) How many elements are eliminated by $\mathcal{U}^{\mathcal{T}}$ in addition to $\mathcal{U}_Q^{\mathcal{T}}$?

ii) Are they part of the nulls or the alternatives?

For the sequence $\mathcal{U}_Q^{\mathcal{T}}$ selected by (4.34), suppose there exist some duplicated elements, of which the first index of the first tie is denoted by i_1. We assume there exist integers $i_1, i_2, \ldots, i_\tau$ and $k_1, k_2, \ldots, k_\tau$ such that the sequence $\mathcal{U}_Q^{\mathcal{T}}$ is regarded as a combination of the monotone pieces and identical parts:

$$\begin{aligned} Q_1 < \cdots < Q_{i_1} = \cdots = Q_{i_1+k_1} < Q_{i_1+k_1+1} < \cdots \\ < Q_{i_2} = \cdots = Q_{i_2+k_2} < Q_{i_1+k_2+1} < \cdots \\ < Q_{i_\tau} = \cdots = Q_{i_\tau+k_\tau} < Q_{i_1+k_\tau+1} < \cdots < Q_{m_0^{\mathcal{T}}}\,. \end{aligned}$$

Note that $\mathcal{U}_Q^{\mathcal{T}} = \mathcal{U}^{\mathcal{T}}$ if and only if $k_1 = k_2 = \cdots = k_\tau = 0$.

Define

$$K_\xi = \#\left\{\mathcal{U}^{\mathcal{T}} \backslash \mathcal{U}_Q^{\mathcal{T}}\right\} = \sum_{j=1}^{m_0^{\mathcal{T}}} \mathbb{1}\left\{P_j^\xi \notin \mathcal{U}_Q^{\mathcal{T}}\right\}, \tag{4.40}$$

i.e., the number of non-identical elements between the excluded p-values $\mathcal{U}^{\mathcal{T}}$ defined by the filter (4.33) and the sequence $\mathcal{U}_Q^{\mathcal{T}}$ defined by (4.34). The following lemma characterises the size of K_ξ.

Lemma 4.3.5. *Consider the filter $\mathcal{T}_\xi^F$ and the sequences $\mathcal{U}^{\mathcal{T}}$ and $\mathcal{U}_Q^{\mathcal{T}}$, it follows that*

i)

$$K_\xi = \sum_{i=1}^{\tau} k_i = \#\left\{j = 2, \ldots, m_0^{\mathcal{T}} \;:\; Q_{j-1} = Q_j\right\}. \tag{4.41}$$

ii) Asymptotically, $\mathcal{U}^{\mathcal{T}}$ increases a fixed proportion of the deleted p-values compared to $\mathcal{U}_Q^{\mathcal{T}}$, in the sense that

$$K_\xi = O(m)\,. \tag{4.42}$$

Proof. Recall that we defined the mid-points $c_j = \frac{j-1/2}{m_0^{\mathcal{T}}}$ of the intervals $D_j = \left[(j-1)/m_0^{\mathcal{T}}, j/m_0^{\mathcal{T}}\right)$ for $j = 1, \ldots, m_0^{\mathcal{T}} - 1$ and $D_{m_0^{\mathcal{T}}} = \left[(m_0^{\mathcal{T}} - 1)/m_0^{\mathcal{T}}, 1\right]$. The length of the intervals is $d = 1/m_0^{\mathcal{T}}$. For any $j \in \{2, \ldots, m_0^{\mathcal{T}}\}$ such that $Q_j = Q_{j-1}$, the fact that

$$\left|Q_j - c_{j-1}\right| + \left|Q_j - c_j\right| \geq c_j - c_{j-1} = d$$

leads to

$$\left|Q_j - c_{j-1}\right| \vee \left|Q_j - c_j\right| \geq \frac{d}{2}.$$

By definition of Q_{j-1} and Q_j, the event $Q_j = Q_{j-1}$ implies

$$\left(\min_{P_i \in P^m} \left|P_i - c_{j-1}\right|\right) \vee \left(\min_{P_i \in P^m} \left|P_i - c_j\right|\right) \geq \frac{d}{2},$$

which indicates that at least one of the two neighbouring intervals D_{j-1} and D_j has no p-values located. Therefore,

$$\begin{aligned}
\mathbb{P}(Q_j = Q_{j-1}) \leq & \mathbb{P}\left(D_{j-1} \cap P^m = \varnothing\right) + \mathbb{P}\left(D_j \cap P^m = \varnothing\right) - \mathbb{P}\left((D_j \cup D_{j-1}) \cap P^m = \varnothing\right) \\
= & \left(1 - \int_{D_{j-1}} (1-\varepsilon) + \varepsilon f_p(t)\,\mathrm{d}t\right)^m + \left(1 - \int_{D_j} (1-\varepsilon) + \varepsilon f_p(t)\,\mathrm{d}t\right)^m \\
& - \left(1 - \int_{D_j \cup D_{j-1}} (1-\varepsilon) + \varepsilon f_p(t)\,\mathrm{d}t\right)^m \\
\leq & 2\left(1 - \frac{1-\varepsilon}{m_0^{\mathcal{T}}}\right)^m - \left(1 - \frac{2(1-\varepsilon)}{m_0^{\mathcal{T}}}\right)^m \\
\longrightarrow & 2\exp\left(-\frac{1-\varepsilon}{1-\xi}\right) - \exp\left(-\frac{2(1-\varepsilon)}{1-\xi}\right), \quad m \longrightarrow \infty.
\end{aligned}$$

On the other hand, define

$$D_j^* = \left[c_j - \min_i |P_i - c_j|, \; c_j + \min_i |P_i - c_j|, \right]$$

i.e., the shortest interval that contains the nearest p-value to c_j. The event $Q_j = Q_{j-1}$ is thus a result from $|D_{j-1}^* \cup D_j^*| = 2\left(\min_{P_i \in P^m} \left|P_i - c_{j-1}\right| + \min_{P_i \in P^m} \left|P_i - c_j\right|\right) \geq 2d$. The probability

$$\begin{aligned}
\mathbb{P}(Q_j = Q_{j-1}) \geq \mathbb{P}&\left((D_{j-1}^* \cup D_j^*) \cap P^m = 1, \; \min_{P_i \in P^m} \left|P_i - c_{j-1}\right| + \min_{P_i \in P^m} \left|P_i - c_j\right| = 2d\right) \\
= m & \left(1 - \int_{D_{j-1}^* \cup D_j^*} (1-\varepsilon) + \varepsilon f_p(t)\,\mathrm{d}t\right)^{m-1} \int_{D_{j-1}^* \cup D_j^*} (1-\varepsilon) + \varepsilon f_p(t)\,\mathrm{d}t \\
\longrightarrow & \frac{2(1-\varepsilon)}{1-\xi} \exp\left(-\frac{2(1-\varepsilon)}{1-\xi}\right), \quad m \longrightarrow \infty.
\end{aligned}$$

Therefore,

$$\frac{\mathbb{E}K_\xi}{m} \longrightarrow \text{Constant},$$

as m tends to infinity.

□

Let $\mathrm{FE}_\xi^{\mathcal{T}}$ and $\mathrm{FI}_\xi^{\mathcal{T}}$ denote the false eliminations and false inclusions due to the fixed-length filtering $\mathcal{T}^F$. The following theorem shows that the proportion of alternative p-values deleted by fixed-length filtering tends to zero.

Theorem 4.3.6. *Consider the two-point Cauchy mixture model where $\varepsilon_m = m^{-\gamma}$ and $\mu_m = m^r$. The expected ratio of the false exclusions committed by $\mathcal{T}^F$ over the total number of the true alternatives m_1 converges to zero, that is,*

$$\frac{\mathbb{E}\left|\mathrm{FE}_\xi^{\mathcal{T}}\right|}{m_1} \longrightarrow 0, \quad m \to \infty, \tag{4.43}$$

if the parameters (γ, r) satisfy

$$r > 1 - \frac{\gamma}{2}. \tag{4.44}$$

Proof. We provide an analogous proof of the randomised filtering approach. The sequence $\mathcal{U}^{\mathcal{T}} = \left\{P_j^\xi,\ j = 1, \ldots, m_0^{\mathcal{T}}\right\}$ forms a partition on $(0, 1]$, which is denoted by

$$D_j^* = \left(P_{(j-1)}^\xi,\ P_{(j)}^\xi\right], \quad j = 1, \ldots, m_0^{\mathcal{T}} + 1,$$

with $P_{(0)}^\xi = 0$ and $P_{(m_0^{\mathcal{T}}+1)}^\xi = 1$. Let $\left\{c_j^*,\ j = 1, \ldots, m_0^{\mathcal{T}} + 1\right\}$ be the mid-points of $\left\{D_j^*,\ j = 1, \ldots, m_0^{\mathcal{T}} + 1\right\}$. It follows that

$$\begin{aligned}
\frac{\mathbb{E}\left|\mathrm{FE}_\xi^{\mathcal{T}}\right|}{m_1} &= \frac{1}{m_1}\mathbb{E}\sum_{i:H_i=1} \mathbb{1}\left\{P_i \in \mathcal{U}^{\mathcal{T}}\right\} \\
&= \frac{1}{m_1}\sum_{i:H_i=1}\sum_{j=1}^{m_0^{\mathcal{T}}+1} \mathbb{P}\left(P_i \in (\mathcal{U}^{\mathcal{T}} \cap D_j^*)\right) \\
&\approx \sum_{j=1}^{m_0^{\mathcal{T}}+1} \frac{\varepsilon f_p(c_j^*)}{(1-\varepsilon) + \varepsilon f_p(c_j^*)} \\
&\approx (m_0^{\mathcal{T}} + 1)\int_0^1 \frac{\varepsilon f_p(t)}{(1-\varepsilon) + \varepsilon f_p(t)}\,\mathrm{d}t \\
&= \varepsilon(m_0^{\mathcal{T}} + 1)\int_0^1 \frac{1}{(1-\varepsilon)\frac{1}{f_p(t)} + \varepsilon}\,\mathrm{d}t \\
&= \frac{1-\xi}{m^{r+\gamma/2-1}(1+o(1))} \longrightarrow 0, \quad m \longrightarrow \infty,
\end{aligned}$$

if $r > 1 - \frac{\gamma}{2}$, which is the same asymptotic boundary as (4.27). □

Actually, one can equivalently prove

$$\frac{\mathbb{E}\sum_{i:H_i=0}\mathbb{1}\{P_i \in \mathcal{U}^{\mathcal{T}}\}}{m_0} \longrightarrow 1, \quad m \longrightarrow \infty, \tag{4.45}$$

in the same region of the (γ, r) space. The asymptotic property of the filtering approaches is given by the ability to delete the null p-values, as well as to preserve the true alternatives.

Proof of Theorem 4.3.2.

Proof. Consider the selected sequence $\mathcal{U}_Q^{\mathcal{T}}$, we first prove the convergence of Q_1 to the uniform quantile. In order to prove

$$\left|Q_1 - c_1\right| = \left|\operatorname*{argmin}_{P_i \in P^m} \left|P_i - c_1\right| - c_1\right| \to 0 \quad \text{almost surely,} \tag{4.46}$$

we define $P_0^{\xi} = \operatorname*{argmin}_{P_i \in P^m} |P_i - 0|$, and to avoid ambiguity, we refer to $P_{(1)}$ among $\{P_1, \dots, P_m\}$ as $P_{1:m}$, and prove that

$$P_0^{\xi} \to 0 \quad \text{almost surely} \quad m \longrightarrow \infty. \tag{4.47}$$

Because the P_i's are marginally drawn from the mixture model $P_i \sim \tilde{F}_p = (1-\varepsilon)U + \varepsilon F_p$, $i = 1, \dots, m$, of which the density is denoted by $\tilde{f}_p$, the density of the minimum $P_{1:m}$ is

$$f_{p_{1:m}}(t) = m\tilde{f}_p(t)[1 - \tilde{F}_p(t)]^{m-1}, \quad 0 < t < 1.$$

We obtain the second moment

$$\begin{aligned}
\mathbb{E}(|P_0^{\xi}|^2) =& m\mathbb{E}\left(P_1^2 \mathbb{1}_{\{P_1 \le P_2, \dots, P_1 \le P_m\}}\right) \\
=& m\int_0^1 t^2 \left[(1-\varepsilon)1 + \varepsilon f_p(t)\right]\left[1 - [(1-\varepsilon)t + \varepsilon F_p(t)]\right]^{m-1} \mathrm{d}t \\
\le& m\left[(1-\varepsilon) + \varepsilon C_p\right]\int_0^1 t^2 \left[(1-\varepsilon)(1-t) + \varepsilon(1 - F_p(t))\right]^{m-1} \mathrm{d}t \\
\le& m\left[(1-\varepsilon) + \varepsilon C_p\right]\int_0^1 t^2 (1-t)^{m-1} \mathrm{d}t \\
=& m\left[(1-\varepsilon) + \varepsilon C_p\right]\frac{2(m-1)!}{(m+2)!} \\
=& \frac{2\left[(1-\varepsilon) + \varepsilon C_p\right]}{(m+2)(m+1)}.
\end{aligned}$$

For any $\eta > 0$, the probability of $P_{1:m}$ exceeding η is bounded by Chebyshev's inequality, which gives

$$\mathbb{P}(|P_0^{\xi}| > \eta) \le \frac{\mathbb{E}(|P_0^{\xi}|^2)}{\eta^2} \le \frac{2\left[(1-\varepsilon) + \varepsilon C_p\right]}{\eta^2(m+2)(m+1)},$$

of which the sum over m is finite,

$$\sum_{m=1}^{\infty} \mathbb{P}(|P_0^{\xi}| > \eta) \leq \sum_{m=1}^{\infty} \frac{2\left[(1-\varepsilon)+\varepsilon C_p\right]}{\eta^2(m+2)(m+1)} < +\infty.$$

From the Borel–Cantelli lemma we obtain that

$$\mathbb{P}\left(\limsup_{m\to\infty} |P_0^{\xi}| > \eta\right) = 0$$

for any positive value $\eta > 0$, which leads to

$$P_0^{\xi} \to 0 \quad \text{almost surely.}$$

Using the same strategy we can prove that, for any $j = 1,\ldots,m_{\xi}$,

$$\left|\operatorname*{argmin}_{P_i\in P^m} \left|P_i - c_j\right| - c_j\right| = \left|Q_j - c_j\right| \to 0 \quad \text{almost surely,} \quad m \longrightarrow \infty. \tag{4.48}$$

Therefore,

$$\lim_{m\to\infty} \sup_{t\in(0,1)} \left| \frac{1}{m_0^{\mathcal{T}}} \sum_{i=1}^{m_0^{\mathcal{T}}} \mathbb{1}(Q_i \leq t) - t \right| = 0.$$

□

Remark. *The filtering method is non-parametric in a sense that it can be applied to any mixture model before knowing the distribution of the test statistics.*

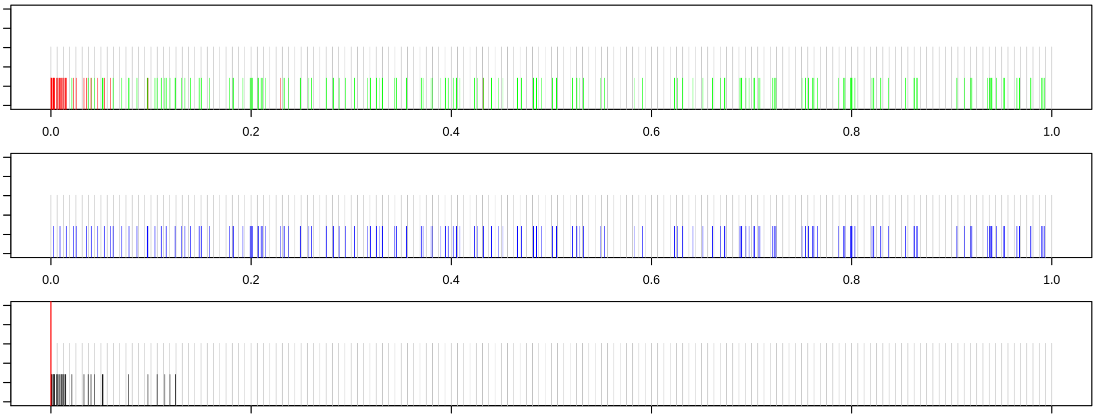


Figure 4.8 – fixed-length filtering applied to Gaussian mixtures

Figure 4.8 shows fixed-length filtering applied to a Gaussian mixture model with total number of hypotheses $m = 200$, the proportion of true alternatives $\varepsilon = 0.15$, size of the positive effect $\mu = 2$, and filtering parameter $\xi = 0.1$. The estimated mode of the alternative p-values is almost zero, which is consistent with the true mode based on normal test statistics obtained by $f_p(0) = \infty$.

A comparison of the proposed filtering approaches can be found in the simulation results.

4.4 Convergence of the filtered mode

To test the heavy-tailed components from a two-point mixture model, we assume that there exists a unique central peak of the alternative p-values, defined as

$$\vartheta = \arg\max_t \tilde{f}_P(t) = \arg\max_t \big((1-\varepsilon) + \varepsilon f_P(t)\big) = \arg\max_t f_P(t)\,. \tag{4.49}$$

For simplicity of notation, here we denote $|\mathcal{S}^{\mathcal{T}}| = s = m_1^{\mathcal{T}}$. As we saw before, the filtered p-values $\mathcal{S}^{\mathcal{T}} = P^m \backslash \mathcal{U}^{\mathcal{T}} = \{P_1^*, \ldots, P_s^*\}$ have distribution function F_p^* and density function f_p^*.

We assume that the true density $f_p^*(t)$ satisfies Condition (4.1.3), and ϑ^{ξ} is the unique mode defined by

$$f_p^*(\vartheta^{\xi}) = \max_{0\le t\le 1} f_p^*(t)\,. \tag{4.50}$$

With a bandwidth $h = h_s$ and a properly chosen kernel function K, the filtered kernel density estimate based on $\mathcal{S}^{\mathcal{T}}$ then becomes

$$f_{P(h)}(t) = f_{P(h)}(t; p_1^*, \ldots, p_s^*) = \frac{1}{hs}\sum_{j=1}^{s} K\left(\frac{t-p_j^*}{h}\right), \tag{4.51}$$

of which the sample mode is proved to converge to the true mode ϑ^{ξ}.

Definition 4.4.1. *The random variable ϑ_s^{ξ} such that*

$$f_{P(h)}(\vartheta_s^{\xi}) = \max_{0\le t\le 1} f_{P(h)}(t) \tag{4.52}$$

is called the sample mode.

In kernel density estimation, it is important to choose a proper bandwidth such that the bias and the variance of the estimator are balanced. We assume $h = h_s$ satisfies

$$\lim_{s\to\infty} h_s = 0, \quad \lim_{s\to\infty} s h_s^2 = \infty. \tag{4.53}$$

Theorem 4.4.2. *Suppose ϑ_s^{ξ} is the sample mode of the filtered p-values $\mathcal{S}^{\mathcal{T}}$ defined by (4.52), and h is a bandwidth satisfying (4.53). Then $\vartheta_s^{\xi} \to \vartheta$ in probability, that is, for any $\epsilon > 0$*

$$\mathbb{P}(|\vartheta_s^{\xi} - \vartheta| > \epsilon) \to 0, \quad s\to\infty. \tag{4.54}$$

where ϑ is the mode of the distribution F_p satisfying Condition (4.1.3).

We use the following propositions to prove the theorem.

Proposition 4.4.3 (Unimodal distribution)**.** *For any unimodal distribution with the mode ϑ, suppose the probability density function $f(t)$ is uniformly continuous. Then it follows that, for any $\epsilon > 0$, there exists an $\epsilon' > 0$ such that, for $0 < t < 1$,*

$$|\vartheta - t| \geq \epsilon \Longrightarrow \left| f(\vartheta) - f(t) \right| \geq \epsilon' .$$

Proposition 4.4.4. *Let ϑ^{ξ} be the unique mode based on the theoretical distribution of the filtered p-values $\mathcal{S}^{\mathcal{T}}$, and ϑ the mode with respect to the distribution of p-values under the alternatives. Then $\vartheta^{\xi} \to \vartheta$ in probability, as $s \to \infty$.*

Proof. Proposition 4.4.3 is easy to obtain. The proof of Proposition 4.4.4 follows from Hoeffding's inequality applied to

$$\frac{1}{m_1} \sum_{i:\, H_i = 1} \mathbb{1}\left\{ P_i \in \mathcal{S}^{\mathcal{T}} \right\},$$

where $\mathbb{1}\{P_i \in \mathcal{S}^{\mathcal{T}}\}$ are bounded random variables. □

Proof of Theorem 4.4.2

Proof. Consider the density function $f_p^*(t)$ and the true mode ϑ^{ξ} based on the distribution of $\mathcal{S}^{\mathcal{T}}$. Since ϑ_s^{ξ} is the sample mode derived from the kernel density estimator with filtration, we prove that

$$\vartheta_s^{\xi} \to \vartheta^{\xi}, \quad s \to \infty. \tag{4.55}$$

As we assumed that $f_p^*(t)$ is uniformly continuous and has a unique mode ϑ^{ξ}, it follows that the Proposition 4.4.3 holds. Therefore, it is enough to prove the convergence of $f_p^*(\vartheta_s^{\xi})$ in probability, that is,

$$f_p^*(\vartheta_s^{\xi}) \xrightarrow{p} f_p^*(\vartheta^{\xi}), \quad s \to \infty. \tag{4.56}$$

In order to derive the convergence of the estimated density function, we investigate the characteristic function of the sample. Let $\{\varphi_s\}_{s=1}^{\infty}$ be the sequence of sample characteristic functions,

$$\varphi_s(u) = \mathbb{E}\, e^{iuP_j^*} = \int_{-\infty}^{\infty} e^{iux}\, \mathrm{d}F_{p,s}(x).$$

Correspondingly, we construct the Fourier transform of the kernel function K. Suppose we choose a proper kernel $K(u)$ such that the Fourier transform

$$k(t) = \int_{-\infty}^{\infty} e^{itu} K(u)\, \mathrm{d}u$$

is absolutely integrable. Then we derive the kernel density estimator in the form of the sample characteristic function. It follows that the kernel density estimator can be written as

$$\begin{aligned} f_{p_{(h)}}(t) &= \frac{1}{hs}\sum_{j=1}^{s} K\left(\frac{t-P_j^*}{h}\right) \\ &= \frac{1}{h}\int K\left(\frac{t-x}{h}\right) \mathrm{d}F_{p,s}(x) = \frac{1}{h}\int K_h\,(t-x)\,\mathrm{d}F_{p,s}(x), \end{aligned}$$

where $K_h(t) = K(t/h)$. Let $\mathcal{F}\circ g$ denote the Fourier transform of a function g. The Fourier transform of $f_{p_{(h)}}(t)$ is then

$$\begin{aligned} \mathcal{F}\circ f_{p_{(h)}}(t) &= \frac{1}{h}\mathcal{F}\circ K_h(t)\cdot\mathcal{F}\circ F_{p,s}(t) \\ &= \frac{1}{h}\int K_h(ut)e^{iut}\,\mathrm{d}u\cdot\int e^{iut}\,\mathrm{d}F_{p,s}(u) \\ &= \frac{1}{h}\int K\left(\frac{ut}{h}\right)e^{iut}\,\mathrm{d}u\cdot\int e^{iut}\,\mathrm{d}F_{p,s}(u) \\ &= \int K(ut)e^{ihut}\,\mathrm{d}u\cdot\int e^{iut}\,\mathrm{d}F_{p,s}(u) \\ &= k(ht)\varphi_s(t), \end{aligned}$$

and it follows that the estimator $f_{p_{(h)}}(t)$ can be written as

$$\begin{aligned} f_{p_{(h)}}(t) = \mathcal{F}^{-1}\circ\mathcal{F}\circ f_{p_{(h)}}(t) &= \frac{1}{2\pi}\int e^{-iut}\mathcal{F}\circ f_{p_{(h)}}(u)\,\mathrm{d}u \\ &= \frac{1}{2\pi}\int e^{-iut}k(hu)\varphi_s(u)\,\mathrm{d}u. \end{aligned} \tag{4.57}$$

In order to prove the convergence of $f_{p_h}(t)$, we consider

$$\left|f_{p_{(h)}}(t) - f_p^*(t)\right|^2 = \left|f_{p_{(h)}}(t) - \mathbb{E}[f_{p_{(h)}}(t)] + \mathbb{E}[f_{p_{(h)}}(t)] - f_p^*(t)\right|^2,$$

and we utilise the fact that the kernel density estimate is asymptotically unbiased such that

$$\sup_t\left|\mathbb{E}[f_{p_{(h)}}(t)] - f_p^*(t)\right| \longrightarrow 0, \quad s\to\infty,$$

so it is enough to prove

$$\sup_t\left|f_{p_{(h)}}(t) - \mathbb{E}[f_{p_{(h)}}(t)]\right| \longrightarrow 0. \tag{4.58}$$

Following the form of (4.57), we have

$$\begin{aligned} \left|f_{p_{(h)}}(t) - \mathbb{E}[f_{p_{(h)}}(t)]\right| &\le \frac{1}{2\pi}\int\left|e^{iut}k(hu)(\varphi_s(u) - \mathbb{E}[\varphi_s(u)])\right|\mathrm{d}u \\ &\le \frac{1}{2\pi}\int|k(hu)|\left|\varphi_s(u) - \mathbb{E}[\varphi_s(u)]\right|\mathrm{d}u. \end{aligned}$$

Consider the $L^2(\mathcal{P})$ norm of (4.58), it suffices to prove that

$$\mathbb{E}^{\frac{1}{2}}\left[\sup_t \left|f_{p_{(h)}}(t) - \mathbb{E}[f_{p_{(h)}}(t)]\right|^2\right] \longrightarrow 0\,.$$

Notice that

$$\begin{aligned}\mathbb{E}^{\frac{1}{2}}\left[\sup_t \left|f_{p_{(h)}}(t) - \mathbb{E}[f_{p_{(h)}}(t)]\right|^2\right] &\le \mathbb{E}^{\frac{1}{2}}\left[\left|\frac{1}{2\pi}\int |k(hu)|\left|\varphi_s(u) - \mathbb{E}[\varphi_s(u)]\right| \mathrm{d}u\right|^2\right]\\ &\le \frac{1}{2\pi}\int |k(hu)|\mathbb{E}^{\frac{1}{2}}\left[\left|\varphi_s(u) - \mathbb{E}[\varphi_s(u)]\right|^2\right] \mathrm{d}u\,,\end{aligned}$$

which is a straightforward result from the Minkowski's integral inequality, and $\mathbb{E}^{\frac{1}{2}}\left[\left|\varphi_s(u) - \mathbb{E}[\varphi_s(u)]\right|^2\right]$ is the square root of the variance of $\varphi_s(u)$, which is bounded by definition

$$\mathrm{Var}(\varphi_s(u)) = \mathrm{Var}\left(\frac{1}{s}\sum_{j=1}^{s} e^{iuP_j^*}\right) \le \frac{1}{s}\mathbb{E}\left|e^{iuP_j^*}\right|^2 \le \frac{1}{s}\,.$$

Therefore,

$$\mathbb{E}^{\frac{1}{2}}\left[\sup_t \left|f_{p_{(h)}}(t) - \mathbb{E}[f_{p_{(h)}}(t)]\right|^2\right] \le \frac{1}{\sqrt{s}h}\int |k(u)|\,\mathrm{d}u \longrightarrow 0$$

if the bandwidth is chosen to satisfy

$$sh^2 \to 0\,. \tag{4.59}$$

Thus, (4.58) is proved and it follows that

$$f_{p_{(s,h)}}(t) \xrightarrow{p} f_p^*(t), \quad s \to \infty, \tag{4.60}$$

and equivalently, as $m \to \infty$. By Lemma 4.4.3, we conclude that the estimator of the mode with filtration converges to the theoretical mode ϑ^ξ of the filtered p-values, and therefore, converges to the mode ϑ. □

Remark. *In order to target the small p-values near the border of the interval $[0,1]$, we propose a transform ϱ from $(0,1)$ to $(0,+\infty)$*

$$\varrho(p_j^*) = -\log(p_j^*)\,, \quad j = 1,\ldots,s\,.$$

The kernel density estimator based on

$$-\log(p_1^*), -\log(p_2^*), \ldots, -\log(p_s^*)$$

is more reliable in practice, particularly when the mode of the alternative p-values is near zero, such as those of the normal test statistics.

Up to now we have solved the first part of the question, namely, to determine the location of the most informative p-values.

Given the property that the p-values of the Cauchy alternatives have a symmetric central peak around the mode, we propose an interval-type rejection region with the half-length $\delta \in [0, \widehat{\vartheta}]$. The inference for FDR and pFDR can be used to choose the optimal length of the rejection region.

4.5 Finite-sample control of FDR

In order to detect the p-values of the heavy-tailed alternatives for the full sample, we define the rejection region $\mathcal{R}_{\vartheta,\delta} = [\vartheta - \delta/2, \vartheta + \delta/2]$ that is described by two parameters, i.e. the center ϑ and the length δ. The total number of rejections is thus

$$R = R(\vartheta,\delta) = \#\{i : p_i \in \mathcal{R}_{\vartheta,\delta}\} = \sum_{i=1}^{m} \mathbb{1}\{|p_i - \vartheta| \le \delta/2\}.$$

4.5.1 Inference for FDR and pFDR

We now estimate the FDR and the pFDR. Given the definition introduced by Storey (2002), Storey (2003) and Storey et al. (2004), the pFDR can be formulated as below.

Proposition 4.5.1. *Suppose the p-values follow the random mixture model*

$$P_i | H_i \sim (1 - H_i) \cdot U + H_i \cdot F_p,$$

where the indicator $H_i \sim \mathrm{Bernoulli}(\varepsilon)$ is also a random variable. Then the pFDR *based on the rejection region $\mathcal{R}_{\vartheta,\delta}$ of the p-values can be written as*

$$\mathrm{pFDR}(\mathcal{R}_{\vartheta,\delta}) = \mathbb{P}(H_i = 0 | P_i \in \mathcal{R}_{\vartheta,\delta}). \tag{4.61}$$

The pFDR has a Bayesian interpretation, as the probability (4.61) is a posterior probability as $H_i \sim \mathrm{Bernoulli}(\varepsilon)$ is regarded as the prior.

In order to compute pFDR, we find that

$$\begin{aligned} \mathrm{pFDR}(\mathcal{R}_{\vartheta,\delta}) &= \frac{\mathbb{P}(P_i \in \mathcal{R}_{\vartheta,\delta} \cap H_i = 0)}{\mathbb{P}(P_i \in \mathcal{R}_{\vartheta,\delta})} = \frac{(1-\varepsilon)\mathbb{P}(P_i \in \mathcal{R}_{\vartheta,\delta} | H_i = 0)}{\mathbb{P}(P_i \in \mathcal{R}_{\vartheta,\delta})} = \frac{(1-\varepsilon)\delta}{\mathbb{P}(P_i \in \mathcal{R}_{\vartheta,\delta})} \qquad (4.62) \\ &= \frac{m(1-\varepsilon)\delta}{m\mathbb{P}(P_i \in \mathcal{R}_{\vartheta,\delta})} = \frac{\mathbb{E}(V(\vartheta,\delta))}{\mathbb{E}(R(\vartheta,\delta))} = \mathrm{mFDR}, \qquad (4.63) \end{aligned}$$

where mFDR stands for the *marginal false discovery rate.* Thus, we can estimate the pFDR by estimating the numerator and denominator of (4.63) respectively.

Remark. *The control of pFDR and mFDR is not equivalent, since the control of pFDR requires the inference for $\mathbb{P}(R > 0)$ and the conditional behaviour of $V|R$. In addition, there is no strong control of pFDR and mFDR, because both values depend on the configuration of the nulls and* *alternatives.*

The denominator of (4.62) can be written as

$$\mathbb{P}(P_i \in \mathcal{R}_{\vartheta,\delta}) = \mathbb{P}(P_i \in \mathcal{R}_{\vartheta,\delta} \,|\, R > 0)\mathbb{P}(R > 0)\,,$$

where $\mathbb{P}(P_i \in \mathcal{R}_{\vartheta,\delta} \,|\, R > 0)$ can be estimated using the observed value of R, that is,

$$\widehat{\mathbb{P}}(P_i \in \mathcal{R}_{\vartheta,\delta} \,|\, R > 0) = \frac{R \vee 1}{m}\,.$$

According to Section 4.2, the probability $\mathbb{P}(R > 0)$ depends on the distribution of the test statistics and the procedures as well. The optimal rejection region $\mathcal{R}_{\vartheta,\delta}$ captures the allocation of the alternative p-values correctly, such that the p-values in $\mathcal{R}^c_{\vartheta,\delta}$ are presumably uniforms. This leads to an estimate of $\mathbb{P}(R > 0) = 1 - \mathbb{P}(R = 0) \le 1 - (1-\delta)^m$.

The pFDR is therefore estimated as

$$\widehat{\text{pFDR}}(\vartheta,\delta) = \frac{(1-\hat{\varepsilon})m\delta}{\{R(\vartheta,\delta) \vee 1\}(1-(1-\delta)^m)}\,, \tag{4.64}$$

and the FDR, without conditioning on $R > 0$, is estimated by

$$\widehat{\text{FDR}}(\vartheta,\delta) = \frac{(1-\hat{\varepsilon})m\delta}{\{R(\vartheta,\delta) \vee 1\}}\,. \tag{4.65}$$

We now propose the estimate of ε.

i) The simplest idea is to avoid estimating ε by using $1-\varepsilon \le 1$. Therefore, the numerator of (4.69), which is the number of accepted components, is estimated by $m\delta$. Although this estimator

$$\widehat{\text{FDR}}(\vartheta,\delta) = \frac{m\delta}{\{R(\vartheta,\delta) \vee 1\}}$$

is widely used, and does provide a bound of the FDR control, we seek to have a precise and interpretable estimate of ε, or an estimate of V as (4.63) is considered.

ii) Based on our filtering procedure, we can select an optimal filtering parameter ξ as an estimator of ε. As we described before, the oracle filtering parameter is $\xi = \varepsilon$ such that we eliminate as many of the null p-values as possible. Suppose the filtering procedure starts with an initial $\xi_0 = 1/2$. In the iterations with $\{\xi_n,\ n = 1,2,\ldots,\}$ we estimate the mode from the filtered p-values. As the procedure is not stopped, we repeat the filtering estimation with a reduced parameter ξ_{n+1}. Thus, the ξ_n that stabilises the estimate $\hat{\vartheta}_n$ can be taken as $\hat{\varepsilon}$, as long as a measurement $\left|\hat{\vartheta}^{\xi_n} - \hat{\vartheta}^{\xi_{n-1}}\right| \le C_n$ is fulfilled. Note that C_n is chosen to measure the convergence of $\hat{\vartheta}_n$, and ξ here is regarded as a data-dependent tuning parameter. Thus, the estimated FDR is

$$\widehat{\mathrm{FDR}}(\vartheta,\delta) = \frac{(1-\xi)m\delta}{\{R(\vartheta,\delta) \vee 1\}}.$$

iii) Storey (2002) introduced the estimator based on a fixed rejection region pre-specified by a tuning parameter $\lambda \in (0,1)$, which is given by

$$1-\hat{\varepsilon} = \frac{W(\lambda)}{(1-\lambda)m},$$

where $W(\lambda) = \#\{p_i > \lambda\}$ is the number of accepted nulls with the length of acceptance region $1-\lambda$. The parameter λ is selected by minimising the mean square error of the FDR estimator. This estimator was later discussed in Storey (2003), Storey et al. (2004), Genovese and Wasserman (2004), Benjamini et al. (2006) and other related works. Benjamini used this estimator and replaced m by $\hat{m}_0 = m(1-\hat{\varepsilon})$ in the denominator of the linear step-up threshold, that is, to reject the nulls for the p-values $p_{(i)} \le i\alpha/\hat{m}_0$. Others used this estimator in the inference and control of FDR.

There are also the methods given by Benjamini and Hochberg (2000), Meinshausen and Rice (2006), Cai et al. (2007), Jin and Cai (2007) that contribute to the estimation of ε, but the estimators are too complicated or not applicable to our study.

We use an estimator of ε defined by

$$1-\hat{\varepsilon} = \frac{W_\vartheta(\xi)}{(1-\xi)m}, \tag{4.66}$$

which is an analogous estimator of Storey's method, and $W_\vartheta(\xi) = \#\left\{|p_i-\vartheta| > \xi/2\right\}$. Note that

$$\mathbb{E}(\hat{\varepsilon}) = 1 - \frac{\mathbb{E}(\sum_{i=1}^m \mathbb{1}\{|p_i-\vartheta|>\xi/2\})}{(1-\xi)m} \le 1 - \frac{\mathbb{E}(\sum_{H_i=0} \mathbb{1}\{|p_i-\vartheta|>\xi/2\})}{(1-\xi)m} = 1 - \frac{(1-\xi)m_0}{(1-\xi)m} = \varepsilon. \tag{4.67}$$

Since the simplest structure we assume is the two-point mixture model, we can expect that the p-values outside the estimated region $\mathcal{R}_{\vartheta,\xi}$ are dominated by the true nulls with frequency $(1-\varepsilon)(1-\xi)$. In other words, $W_\vartheta(\xi)$ is roughly $(1-\xi)(1-\varepsilon)m$, since the p-values from the nulls are assumed to be uniformly distributed over the region $\mathcal{R}^c_{\vartheta,\delta}$, of which the length $1-\delta$ is replaced by $1-\xi$. Thus, $W_\vartheta(\xi)/(1-\xi)$ is analogous to $m-R$, and is therefore utilised to estimate $m-m_1$.

Remark. *Although we desire to find an accurate estimate of the fraction of the effects, a lower bound of ε suffices, and the parameter ξ needs not be an estimator of ε. The estimator (4.66) is slightly biased, and can be used as a lower bound of ε. In reality it is acceptable to claim that the proportion of true alternatives is no less than the declared frequency $\hat{\varepsilon}$ in expectation.*

The pFDR is therefore estimated as

$$\widehat{\mathrm{pFDR}}(\vartheta,\delta) = \frac{W_\vartheta(\xi)\delta}{(1-\xi)\{R(\vartheta,\delta)\vee 1\}(1-(1-\delta)^m)}, \tag{4.68}$$

and the FDR is estimated by

$$\widehat{\mathrm{FDR}}(\vartheta,\delta) = \frac{W_\vartheta(\xi)\delta}{(1-\xi)\{R(\vartheta,\delta)\vee 1\}}. \tag{4.69}$$

Based on the estimator of FDR and pFDR, we are able to give the following data-dependent algorithm to detect and locate the alternatives.

Data-dependent rejection region.	
Step 1.	Compute the p-values for each test $p^m = \{p_1,\dots,p_m\}$, and order them non-decreasingly, i.e. $p_{(1)} \le p_{(2)} \le \cdots \le p_{(m)}$.
Step 2.	Apply the filtering approach $\mathscr{T}$ with a parameter $\xi \in (0,1)$, and get the filtered sequence $\mathcal{S}^{\mathscr{T}}$.
Step 3.	Estimate the density of $\mathcal{S}^{\mathscr{T}}$ and the mode $\hat{\vartheta}$.
Step 4.	Given $i = 1,\dots,m$ and a significance level α, reject $H_{0,(i)}$ if $\lvert p_i - \hat{\vartheta}\rvert \le \lvert p_{(\tau)} - \hat{\vartheta}\rvert$, where $\tau = \max\{\tau : \widehat{\mathrm{FDR}}(\hat{\vartheta}, 2(p_{(\tau)} - \hat{\vartheta})) \le \alpha\}$.

The control of FDR is based on the estimator (4.69). An appropriate $\hat{\delta}$ turns out to be the largest length of the rejection region $\mathcal{R}_{\vartheta,\delta}$ subject to the control of $\widehat{\mathrm{FDR}}$.

4.5.2 Data-dependent control of FDR for finite sample

We propose a control of FDR for finite-sample case in the following theorem. Note that our method is data-dependent since the estimate FDR is a plug-in estimator.

Theorem 4.5.2. *Based on the rejection region $\mathcal{R}_{\vartheta,\delta}$ and the estimator of FDR given by (4.69),*

$$\mathbf{E}(\widehat{\mathrm{FDR}}(\vartheta,\delta)) \ge \mathrm{FDR}(\vartheta,\delta) \tag{4.70}$$

for any valid (ϑ,δ).

Proof. We take the difference

$$\mathbb{E}(\widehat{\mathrm{FDR}}(\vartheta,\delta)) - \mathrm{FDR}(\vartheta,\delta) = \mathbb{E}\left[\frac{\delta W_\vartheta(\xi)/(1-\xi)}{\{R(\vartheta,\delta)\vee 1\}}\right] - \mathbb{E}\left[\frac{V(\vartheta,\delta)}{\{R(\vartheta,\delta)\vee 1\}}\right] = \mathbb{E}\left[\frac{\delta W_\vartheta(\xi)/(1-\xi) - V(\vartheta,\delta)}{\{R(\vartheta,\delta)\vee 1\}}\right]$$

$$\ge \mathbb{E}\left[\frac{\delta W_\vartheta(\xi)/(1-\xi) - V(\vartheta,\delta)}{R(\vartheta,\delta)}\mathbb{1}\{R(\vartheta,\delta)>0\}\right].$$

Recalling that

$$R(\vartheta,\delta) = S(\vartheta,\delta) + V(\vartheta,\delta),$$

we condition on $S(\vartheta,\delta)$ and tackle the $V(\vartheta,\delta)$ in both the numerator and the denominator. We obtain that the last equation above equals

$$\begin{aligned}
&\mathbb{E}\left[\frac{\delta W_\vartheta(\xi)/(1-\xi) - V(\vartheta,\delta)}{S(\vartheta,\delta)+V(\vartheta,\delta)}\mathbb{1}\{R(\vartheta,\delta)>0\}\right] \\
=&\mathbb{E}\left[\mathbb{E}\left[\frac{\delta W_\vartheta(\xi)/(1-\xi) - V(\vartheta,\delta)}{S(\vartheta,\delta)+V(\vartheta,\delta)}\mathbb{1}\{R(\vartheta,\delta)>0\}\Big|S(\vartheta,\delta)\right]\right] \qquad (4.71)\\
\geq&\mathbb{E}\left[\frac{\mathbb{E}\left[(\delta W_\vartheta(\xi)/(1-\xi) - V(\vartheta,\delta))\,\mathbb{1}\{R(\vartheta,\delta)>0\}\,\big|S(\vartheta,\delta)\right]}{\mathbb{E}\left[(S(\vartheta,\delta)+V(\vartheta,\delta))\mathbb{1}\{R(\vartheta,\delta)>0\}\,\big|S(\vartheta,\delta)\right]}\right],
\end{aligned}$$

with the last inequality obtained by Jensen's inequality on $V(\vartheta,\delta)$, given the fact that

$$\frac{W-V}{S+V} = \frac{W+S}{V+S} - 1$$

is a convex function of V with $W+S>0$. Since

$$\mathbb{E}\,[\delta W_\vartheta(\xi)/(1-\xi) - V(\vartheta,\delta)] \geq \delta m(1-\varepsilon)(1-\xi)/(1-\xi) - m(1-\varepsilon)\delta = 0, \qquad (4.72)$$

we conclude that

$$\mathbb{E}(\widehat{\mathrm{FDR}}(\vartheta,\delta)) \geq \mathrm{FDR}(\vartheta,\delta).$$

□

Similarly, we obtain that

$$\mathbb{E}(\widehat{\mathrm{pFDR}}(\vartheta,\delta)) \geq \mathrm{pFDR}(\vartheta,\delta), \qquad (4.73)$$

with our estimator having $1-(1-\delta)^m \geq \mathbb{P}(R>0)$ in the denominator of (4.68).

Following this theorem we can get control of the true FDR by limiting the estimated $\widehat{\mathrm{FDR}}(\vartheta,\delta)$ below a desired level. Our rejection region $\mathcal{R}(\vartheta,\delta)$ is nested, and the monotonicity of power is guaranteed.

Proposition 4.5.3 (Monotonicity of power)**.** *For fixed center ϑ, the decision rule defined by the rejection region $\mathcal{R}(\vartheta,\delta)$ has monotone power in a sense that*

$$\beta_{\mathcal{R}(\vartheta,\delta)} \geq \beta_{\mathcal{R}(\vartheta,\delta')} \quad \textit{for any} \quad 0<\delta\leq\delta'\leq 2\vartheta.$$

Remark. *Storey (2003) also considered the asymptotic control of the FDR and pFDR with $\varepsilon_m = \varepsilon$ fixed. We are not interested in this parametrisation since the number of significant components can be moderately large if it is proportional to m.*

4.6 Discussion

In this section we compare our procedures to the most related methods by Efron et al. (2001), Efron and Tibshirani (2002), Storey (2002), Storey (2003), and Cai and Sun (2017).

4.6.1 Positive FDR, local FDR and the empirical Bayesian interpretation

We discuss Storey's positive FDR and Efron's local FDR together because they both have a Bayesian interpretation, and are inherently linked to one another.

pFDR

As we mentioned near before, Storey's pFDR is also referred to as the posterior FDR, since they apply the Bayes formula to FDR with respect to the prior probability of H_i. The control of the pFDR

$$\mathrm{pFDR}(\mathcal{R}) = \mathbb{P}(H_i = 0|P_i \in \mathcal{R}) = \frac{(1-\varepsilon)\mathbb{P}(P_i \in \mathcal{R}|H_i = 0)}{(1-\varepsilon)\mathbb{P}(P_i \in \mathcal{R}|H_i = 0) + \varepsilon\mathbb{P}(P_i \in \mathcal{R}|H_i = 1)} \tag{4.74}$$

is also the same as our control of operating characteristics TPR/FPR, as $\mathbb{P}(P_i \in \mathcal{R}|H_i = 0)$ is the type I error and $\mathbb{P}(P_i \in \mathcal{R}|H_i = 1)$ is the power.

However, Storey and other authors of the related work only discussed the case when $\mathbb{P}(P_i \in \mathcal{R}|H_i = 1)/\mathbb{P}(P_i \in \mathcal{R}|H_i = 0)$ is decreasing, which is the same condition as we defined by (4.1.7) for detecting the light-tailed alternatives, without mentioning the solution of the rejection rules to the heavy-tailed cases. In addition, they discussed the asymptotic cases based on the assumption that

$$\sum_{i=1}^{m}(1-H_i)/m \longrightarrow \pi_0, \quad m \to \infty,$$

while we consider the asymptotic framework with

$$\sum_{i=1}^{m}(1-H_i)/m = 1-\varepsilon_m = 1-m^{-\gamma} \longrightarrow 1, \quad m \to \infty.$$

Although the numbers of the true nulls and alternatives both tend to infinity, the ratio will be difficult to detect, which also encouraged us to discuss the asymptotically detectable region.

Local FDR

Efron defined the *local false discovery rate* for the z-values instead of the test statistics or the p-values. His local FDR method was established on the theory of empirical Bayes inference. With the fixed prior probabilities π_0 and π_1, the z-values follow the distribution $f_0(z)$ under the nulls and $f_1(z)$ under the alternatives. $f(z) = \pi_0 f_0(z) + \pi_1 f_1(z)$ is thus the distribution of the z-values, of which $f_0(z) = \varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$. Efron defined the *local Bayes false discovery rate* as

$$\mathrm{Lfdr}(z) = \mathbb{P}(\text{null} \mid \text{test statistic } z) = \frac{\pi_0 f_0(z)}{\pi_0 f_0(z) + \pi_1 f_1(z)}, \tag{4.75}$$

where the densities in the numerator and the denominator are to be estimated respectively.

In our work, we consider the distribution of the p-values, and we maximise the ratio $G_1'(t)/G_0'(t)$ to get the significance center such that a large number of true positives are discovered subject to a small increment of the false positives. We can equivalently define the local FDR for the p-values as

$$\mathrm{Lfdr}_p(t) = \frac{1-\varepsilon}{1-\varepsilon+\varepsilon f_p(t)}.$$

Efron's density estimation benefits from the normal distribution of the z-values, and we utilise the uniform distribution of the majority p-values from the nulls.

It is necessary to conclude that maximising $G_1'(t)/G_0'(t)$ is equivalent to minimising the local FDR, taking the Lfdr(t) as a point-wise threshold sequence defined for the p-values, and in addition, equivalent to minimising the pFDR as well. We are particularly interested in looking for the most informative region of the p-values without a pre-determined rejection rule. Our method is adaptive and data-dependent, and is easily interpretable as well.

4.6.2 Screening for high-throughput data

A similar idea to our filtering method appeared in Cai and Sun (2017) for a different purpose. In high-dimensional multiple testing, one of the main issues is to reduce the dimension according to the capacity of the experiments. They discussed a screening approach applied to high-throughput applications, which was considered a multi-stage procedure.

In their screening approach, the selection rule is defined by

$$\delta_i = \mathbb{1}\{\hat{T}(Z_i) \le t_i\},$$

where $\hat{T}(Z_i)$ is an estimator of the local FDR defined by Efron, and t_i is a critical value. They keep the values with $\delta_i = 1$ and provided the conditions for a valid screening procedure. However, they used classic kernel density estimation to get the densities and in addition, they estimate $\mathrm{Lfdr}(z)$, whose effectiveness relies on the distribution of the test statistics. We are more interested in properly estimating the marginal distribution with a filtering method that reduces the influence caused by the majority of the nulls.

4.7 Numerical study

In this section we provide simulation results for the proposed algorithms.

4.7.1 BH procedure for testing Cauchy mixtures

We perform $N = 10,000$ replications of the simulation and each replication tests a batch of $m = 1000$ hypotheses. In the r-th replication we apply the Bonferroni correction and the BH procedure, and record the total number of rejections R_m^r, the number of false discoveries V_m^r and the number of true discoveries S_m^r out of the m simultaneously tested hypotheses. In each replication we compute the FDR and use the average as an estimate. The parameters ε and μ_i are chosen in the region where the presence of true signals is clearly detectable.

ε	μ	$\mathbb{P}(R>0)$		FDR		pFDR	
		Bonferroni	BH	Bonferroni	BH	Bonferroni	BH
$\varepsilon = 0.10$	$\mu = 5$	0.0487	0.0501	0.0447	0.0462	0.9179	0.9218
	$\mu = 10$	0.0476	0.0485	0.0424	0.0434	0.8918	0.8938
$\varepsilon = 0.15$	$\mu = 5$	0.0525	0.0539	0.0446	0.0460	0.8495	0.8534
	$\mu = 10$	0.0484	0.0498	0.0425	0.0437	0.8781	0.8765
$\varepsilon = 0.20$	$\mu = 5$	0.0466	0.0479	0.0378	0.0389	0.8112	0.8114
	$\mu = 10$	0.0493	0.0514	0.0396	0.0411	0.8022	0.8003

Table 4.1 – Bonferroni correction and BH procedure in multiple testing for Cauchy mixtures

As Table 4.1 shows, even though control of the false discovery rate is maintained below a desired level $\alpha = 0.05$, the chance of finding true discoveries is very limited by rejecting the hypotheses with the smallest observed p-values. The probability $\mathbb{P}(R > 0)$ is estimated by the observed mean $\frac{1}{N}\sum_{r=1}^{N} \mathbb{1}\{R_m^r > 0\}$, and the pFDR is estimated based on the replications where there are the rejections. Notice that only hundreds of replications report a few discoveries over the total $N = 10,000$, which means in each replication among the $m = 1000$ tests, almost no null hypothesis is rejected. The majority of the tests give zero discoveries, which leads to the overall control of the FDR. The power is surprisingly low in this heavy-tailed problem. When twenty percent of the observations are true effects, the chance of declaring at least one positive is still around 0.05 on average, which could be misleading as an experimental conclusion.

4.7.2 Performance of the proposed methods

Simulation 2

In this section we show the results of the filtering algorithm for small sample Cauchy mixtures.

Figures 4.9-4.11 show the results of the proposed filtering procedures for $m = 50$ hypotheses, with $\varepsilon = 0.2$ and $\mu = 7$. In each panel, the plot on the top is the oracle allocation of the null and alternative p-values, of which the red bars stand for the true alternatives, and the blue bars

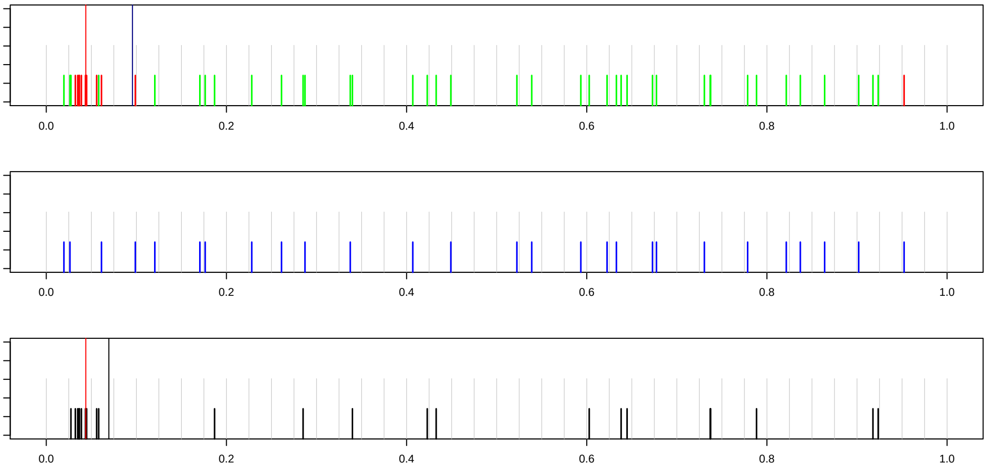


Figure 4.9 – The $\mathcal{U}^{\mathcal{T}}$ and $P^m \backslash \mathcal{U}^{\mathcal{T}}$ sequences of the randomised filter

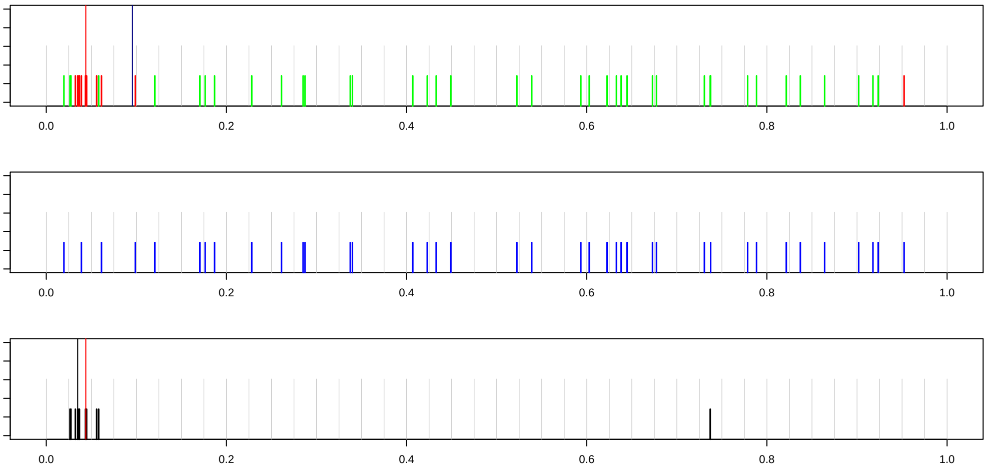


Figure 4.10 – The $\mathcal{U}^{\mathcal{T}}$ and $P^m \backslash \mathcal{U}^{\mathcal{T}}$ sequences of the fixed-length filter

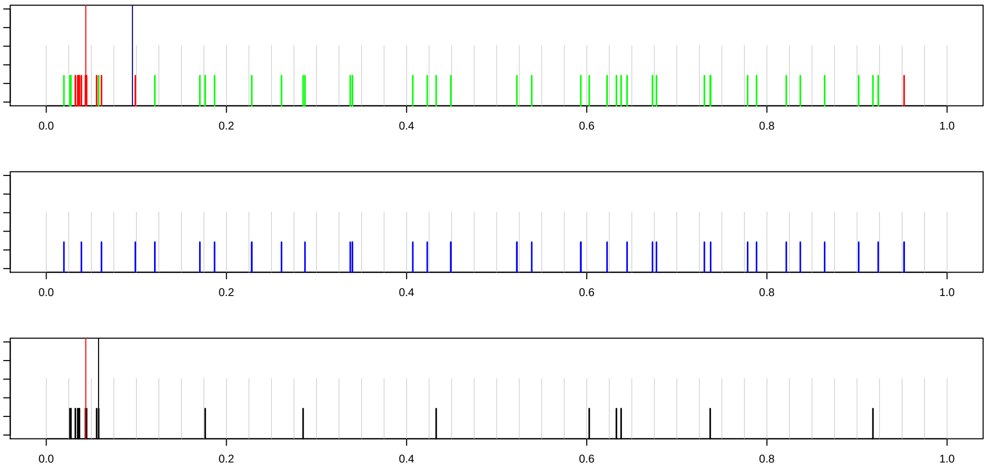


Figure 4.11 – The $\mathcal{U}_Q$ and $P^m \backslash \mathcal{U}_Q$ sequences of the fixed-length filter

stand for the true nulls. We apply the filtering approaches to a same sample, and with a same filtering parameter $\xi = 0.2$.

i) Figure 4.9 shows the results of the randomised filter $\mathscr{T}^R$. With the bins of length $d = 1/m_\xi = 1/40 = 0.025$, $\mathscr{T}^R$ deletes one p-value randomly from each interval, and the deleted ones are shown in the middle plot with the blue bars. For the empty bins, the filter $\mathscr{T}^R$ deletes nothing, so we proposed the fixed-length filtering to improve the elimination.

ii) Figure 4.10 shows the results of the fixed-length filter $\mathscr{T}^F$ with $\xi = 0.2$. Following the definition of $\mathscr{T}^F$ given by (4.33), the filter looks for the nearest p-value that is not eliminated in the past deletions. Therefore, although there are a number of empty bins, the filter $\mathscr{T}^F$ deletes enough p-values as required without removing the peak of the alternatives.

iii) Figure 4.11 shows the $\mathcal{U}_Q$ and $P^m \backslash \mathcal{U}_Q$ sequences of the fixed-length filter $\mathscr{T}^F$ with $\xi = 0.2$. The blue bars represent the sequence $\mathcal{U}_Q$ defined by (4.34), which is a subsequence of $\mathcal{U}^{\mathscr{T}}$ deleted by $\mathscr{T}^F$. Eliminating $\mathcal{U}_Q$ from P^m deletes more null p-values than the randomised filter plotted in Figure 4.11, but fewer than the fixed-length filter, due to the duplicated elements among the Q_j.

The deleted sequences are close to uniform, and the majority of the alternative p-values are preserved. Compared to the other figures, $\mathscr{T}^F$ deletes more p-values, of which most are nulls.

Simulation 3.

With $m = 1000$ hypotheses, we apply the proposed filtration algorithm to Cauchy mixtures with $\varepsilon = 0.05, 0.10, 0.15, 0.20, 0.25$, and $\mu = 6, 8, 10, 12, 14, 16, 18, 20$. For each configuration we run $N = 200$ replications and get the sample value of the parameters and the true and false discoveries.

$\varepsilon = 0.15$	$\mu = 6$	$\mu = 8$	$\mu = 10$	$\mu = 12$	$\mu = 14$	$\mu = 16$	$\mu = 18$	$\mu = 20$
ϑ	0.05121	0.03899	0.03142	0.02628	0.02258	0.01979	0.01761	0.01586
$\hat{\vartheta}$	0.05193	0.03935	0.03152	0.02636	0.02264	0.01984	0.01763	0.01587
$\hat{\delta}$	0.01573	0.01521	0.01412	0.0139	0.01317	0.01256	0.01241	0.01288
$\widehat{\mathrm{FDR}}$	0.08577	0.09389	0.08816	0.08660	0.08699	0.08493	0.08364	0.08377
FDR	0.08547	0.09238	0.08920	0.08258	0.08364	0.07820	0.07717	0.07451
TP/m_1	0.4924	0.4967	0.5994	0.7506	0.8175	0.8702	0.9006	0.9219

Table 4.2 – Simulation results of detecting the Cauchy alternatives

We present an example of the simulation results of detecting the Cauchy alternatives in

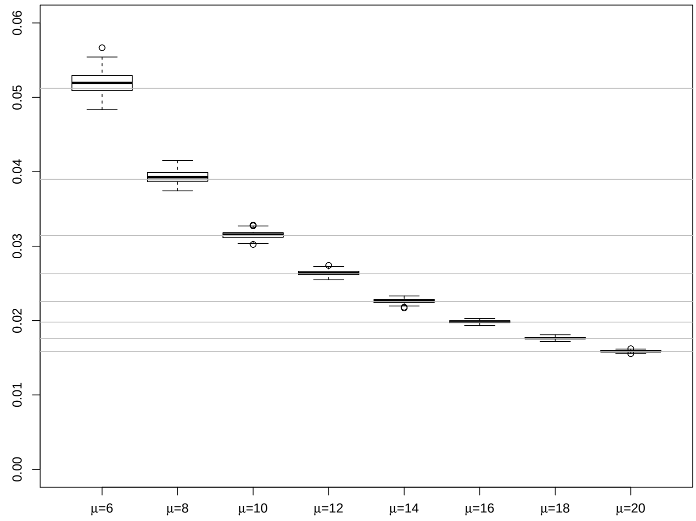


Figure 4.12 – The filtering estimate of the mode of the alternative p-values

Table 4.2. In each replication, we test $m = 1000$ hypotheses, of which the frequency of true alternatives is fixed at $\varepsilon = 0.15$. Let the size of the positive effect be $\mu = 6, 8, 10, 12, 14, 16, 18, 20$.

The estimates of the alternative mode $\hat{\vartheta}$ are shown by Figure 4.12, which is the result from the kernel density estimator based on the filtered p-values. We can see that the most informative p-value is decreasing with the positive shift μ increasing, which is true according to the tail area under the Cauchy density function. A pre-specified level $\alpha = 0.1$ is utilised to control the data-dependent estimator $\widehat{\mathrm{FDR}}(\vartheta, \delta) \leq \alpha$ given by (4.69). We choose the rejection region $\mathcal{R}_{\hat{\vartheta},\hat{\delta}}$ with the maximal length $\hat{\delta}$ subject to the control of $\widehat{\mathrm{FDR}}(\vartheta, \delta) \leq \alpha$. The average FDR is shown in the table by $\widehat{\mathrm{FDR}}$. The true value of FDR computed from the sample is different from the estimator $\widehat{\mathrm{FDR}}(\vartheta, \delta)$, which is influenced by the tuning parameter ξ as we propose in the filtering procedure. With the peak of the p-value getting narrow, the rejection region contains more true alternatives.

5 Further discussion and generalisation

In general, we are interested in the following three types of problems and methodologies in multiple testing:

- *Global tests.* Given a family of hypotheses, it is ideal to utilise one informative test statistic that measures the departure of the realisations from the theoretical distribution under the global null hypothesis. For example, the Higher Criticism test is a Kolmogorov-Smirnov test based on the level of p-values.

- *Individual tests (one-step/step-wise).* Other than the global tests that consider the null hypotheses jointly, there are many situations where individual tests are used to locate the significant components by assigning a significant level to each hypothesis. We introduced the Bonferroni correction, which is a cut-off threshold, and the step-up and step-down procedures in Chapter 2, as well as the hierarchical testing and sequential testing. Closed testing and partition testing identify the true discoveries with the structure of hypotheses properly designed. Online testing adjusts the α-wealth and the individual significance level according to the past rejections.

- *Inference-based tests.* Another huge class of multiple testing methods take the inference for the parameters as a first step. Since the parameters that capture the models are usually unknown, one of the main issues is to make inference for the parameters based on the observations before applying a testing procedure. We are also interested in the distributional properties of the variables that would help to develop a data-dependent approach. This field is highly linked to post-selection inference.

In the previous chapters we explained how the extreme values from the nulls confound the true effects from the alternatives due to the heavy-tailedness of the distribution, and proposed solutions in Chapter 3 and 4. In this chapter we will discuss several topics of generalisations that are still in progress, and give a summary of our main contributions at the end.

5.1 Generalisations

5.1.1 Heavy-tailed, long-tailed, and fat-tailed distributions

We explained in previous chapters how the distribution of the p-values is influenced by the tail distribution of the test statistics, and we proposed testing procedures based on the Cauchy distribution. Condition 3.3.1 on the likelihood ratio was discussed in Chapter 3 in terms of the existence of the most powerful test under asymptotic considerations. Conditions 4.1.2 and 4.1.3 captured the core features of the distribution of the heavy-tailed test statistics and the corresponding p-values. Those two conditions were required in our multiple testing study, and were referred to as a heavy-tailed framework when the usual methods require an opposite Condition 4.1.7. Furthermore, we are interested in different characterisations of the heavy-tailedness and seek to find a universal solution.

We first discuss the following concepts that are all considered as having a tail heavier than the exponential tail.

Definition 5.1.1. *Suppose $F(x)$ is the cumulative distribution function of the random variable X, and $\bar{F}(x) = \mathbb{P}(X > x)$ is the tail distribution.*

i) (Heavy-tailed.) A distribution F is referred to as heavy-tailed, if

$$\int_{-\infty}^{\infty} e^{tx} dF(x) = \infty, \quad t > 0, \tag{5.1}$$

or equivalently,

$$\lim_{x\to\infty} e^{tx}\bar{F}(x) = \infty, \quad t > 0. \tag{5.2}$$

ii) (Long-tailed.) A distribution F is referred to as long-tailed, if

$$\lim_{x\to\infty} \mathbb{P}(X > x + t \mid X > x) = 1, \tag{5.3}$$

or equivalently

$$\bar{F}(x+t) \sim \bar{F}(x), \quad x \to \infty. \tag{5.4}$$

iii) (Fat-tailed.) A distribution F is referred to as fat-tailed, if

$$\bar{F}(x) \sim x^{-\alpha}, x \to \infty, \quad \alpha > 0 \tag{5.5}$$

iv) (Regularly varying.) A distribution F is called regularly varying if

$$\bar{F}(x) = x^{-\rho} L(x), \tag{5.6}$$

where L is a slowly varying function.

All the definitions mentioned above overlap with our heavy-tailed framework. Therefore, in the text we did not specify the difference between heavy-tailed, long-tailed and fat-tailed distributions, but rather focused on the exact conditions that influence the quality of the test with normality violated. As long as Condition 4.1.3 is satisfied,

$$\lim_{t\to 0} f_p(t) = \lim_{t\to 0} \frac{f_1(F_0^{-1}(1-t))}{f_0(F_0^{-1}(1-t))} = 1, \tag{5.7}$$

the p-values are considered in the heavy-tailed framework.

Example 5.1.2 (Student's t distribution)**.** *In a two-point mixture model, suppose the random noise follows the distribution t_ν with ν being the degrees of freedom. The density function is*

$$f_{t_\nu}(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}. \tag{5.8}$$

The likelihood ratio is therefore $((1-\varepsilon)f_0(x)+\varepsilon f_1(x))/f_0(x) = (1-\varepsilon)+\varepsilon g(x)$, where

$$g(x) = \frac{f_{t_\nu}(x-\mu)}{f_{t_\nu}(x)} = \left(\frac{\nu+(x-\mu)^2}{\nu+x^2}\right)^{-(\nu+1)/2}$$

is not monotone.

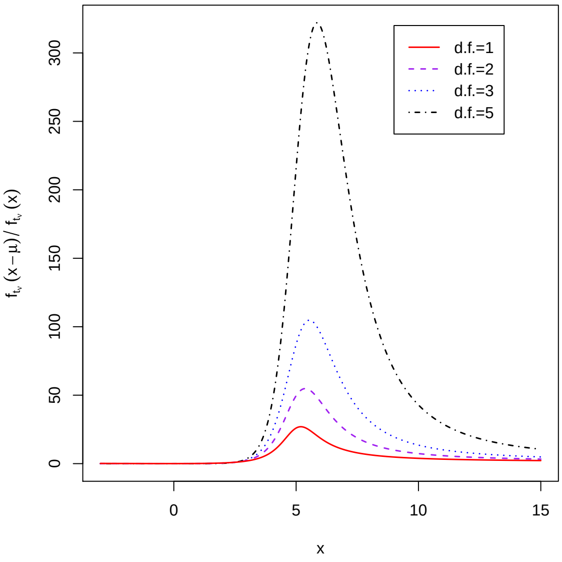


Figure 5.1 – The likelihood ratio for t_ν with $\nu = 1,2,3,5$

Figure 5.1 shows the function $g(x)$ for $\nu = 1,2,3,5$ and mean shift $\mu = 5$. As the degrees of freedom increase, the peak of the likelihood ratio increases rapidly, which makes the alternatives *easier to detect. For any* $\nu > 0$, *Condition 4.1.3 follows from*

$$\lim_{t \to 0} f_p(t) = \lim_{x \to \infty} \frac{f_{t_\nu}(x-\mu)}{f_{t_\nu}(x)} = 1\,,$$

which makes the generic p*-values near zero uninformative. When testing multiple hypotheses based on mixtures of* t *distributions, we would suggest using the rejection region method to identify the alternatives.*

Asymptotic detection based on Kullback-Leibler divergence, as proposed in Chapter 3, quantifies the detectable fraction ε and size μ of the non-null effects in the mixture model, of which the distribution family can be either light-tailed or heavy-tailed. The filtering approach we proposed in Chapter 4 is also a non-parametric method that identifies the location of the non-zero effects utilising the ordered sequence of p-values. Once the sample mode is estimated from the filtered p-values, the rejection region comes with no assumptions on the underlying distribution family.

5.1.2 Test statistics based on the gaps

In multiple testing problems based on two-point mixture models, as we explained in Chapter 4, the p-values from the alternatives have a local concentration near the mode. This phenomenon also explains why the testing procedures based on Gaussian noise declare rejections near zero. We may think of this problem from another perspective.

From a non-parametric point of view, a method that describes the local concentrations, or equivalently, the clusters, of the alternative p-values is desired. When we perform tests based on the p-values, the null distribution is taken to be uniform, while the p-values from the alternatives have the tendency to concentrate in very narrow regions, in which most of the p-values are from the true alternatives and only few of them are from the nulls. This property is true for both light-tailed and heavy-tailed alternative distributions.

In general, let p_c stand for the center, and $\delta > 0$ stand for the width of any interval-type rejection region based on the non-decreasingly ordered p-values. To have a better understanding of p_c and δ as the threshold parameters, we analyse the order statistics of the p-values. We point out that the alternative p-values have smaller gaps than the uniformly distributed ones.

Define the gap statistics

$$G_i = p_{(i+1)} - p_{(i)} \tag{5.9}$$

for $i = 1, \ldots, m-1$, which follow a Beta distribution $G_i \sim \text{Beta}(1, m)$ under the overall null hypothesis. We can intuitively compare the gaps to the expected value $\mathbb{E}(G_i) = 1/(1+m)$, though the raw p-values are noisy and the gaps have a large variation. We propose a modified version, called the *cumulative* p-values, and define the *weighted* gaps as below.

Figure 5.2 shows the gaps of the p-values based on Cauchy mixture model in the gray line. The

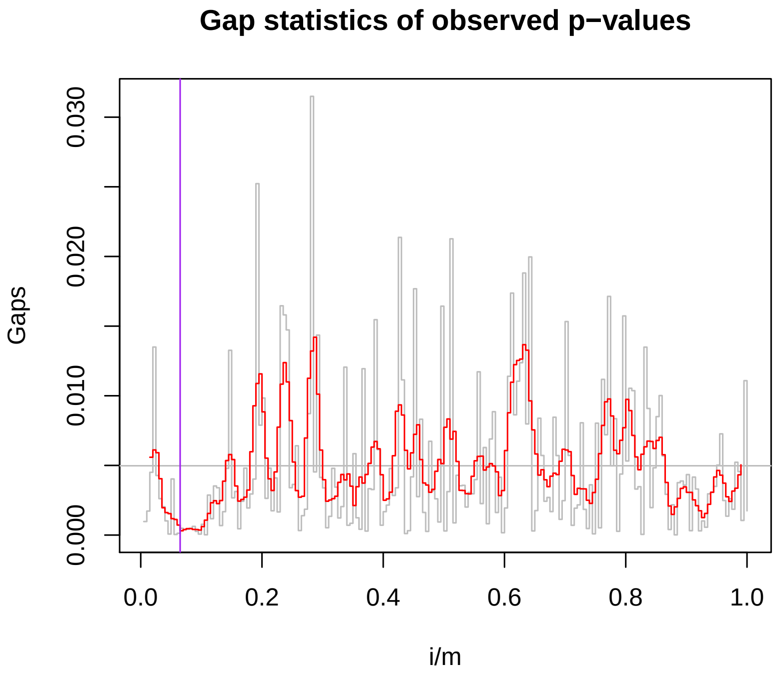


Figure 5.2 – The smoothed gaps of the p-values from the Cauchy mixture model

variation of the gaps is considerable, except for the interval containing most of the alternatives. The red line is the smoothed gaps with a linear smoother. The alternative p-values form a region that is visibly seen from the smoothed variation, with the theoretical location of the true alternatives (purple line) included. Although the gaps drawn from the raw p-values reveal the distributional property of the p-values under the alternatives, we need to find an appropriate way to stabilise the estimation.

Define a *smoothed* version of observed p-values

$$p_j^\dagger = \frac{1}{j}\sum_{i=1}^{j} p_{(i)}, \tag{5.10}$$

and for $j = 2, \ldots, m$, we define the weighted gap statistic

$$\begin{aligned} G_j^\dagger &= p_j^\dagger - p_{j-1}^\dagger \\ &= \frac{(j-1)(p_{(j)} - p_{(j-1)}) + (j-2)(p_{(j-1)} - p_{(j-2)}) + \cdots + (p_{(2)} - p_{(1)})}{j(j-1)} \\ &= \frac{(j-1)G_{j-1} + (j-2)G_{j-2} + \cdots + G_1}{j(j-1)}, \end{aligned} \tag{5.11}$$

which is in reality a weighted sum of the original gap statistics G_j's. We give a larger weight to G_i as it is closer to $G_j^\dagger$, which means that we pay more attention to the local properties of the gap statistics.

When the observations are i.i.d. from the null distribution, the p-values are uniformly distributed, and it follows that $G_i = p_{(i)} - p_{(i-1)} \sim \text{Beta}(1, m)$, with expectation $\mathbb{E}(p_{(j)} - p_{(j-1)}) = 1/(1+m)$. The weighted gaps have a Beta distribution with $\mathbb{E}(G_j^\dagger) = \mathbb{E}(p_j^\dagger - p_{j-1}^\dagger) = 1/(2+2m)$. Therefore, it is reasonable to compare the weighted gaps to $1/(2+2m)$ and find the region where the cluster of alternative p-values occurs, if any.

One can intuitively take the p-value that minimises the weighted gap, denoted by $\hat{p}_c$, to be the center of the cluster from the alternatives. The statistic $G_j^\dagger$ is a weighted sum of the gaps and gives a plausible estimation of the significance center. On the other hand, analysing the local change of the gaps is also a good way to discover the p-value cluster. Formally, define the local discrepancies

$$L_j = \left| \frac{1}{k} \sum_{i=j-k}^{j-1} G_j^\dagger - \frac{1}{2(m+1)} \right|, \quad U_j = \left| \frac{1}{k} \sum_{i=j}^{j+k-1} G_j^\dagger - \frac{1}{2(m+1)} \right|, \tag{5.12}$$

where L stands for "lower" and U stands for "upper". We want the L_j and U_j to recognise the segment when there is a distributional change in the gaps. In summary, we propose the following approach to locate the cluster of alternative p-values.

Local discrepancy of p-value gaps.	
Step i)	Compute the smoothed p-values $p_j^\dagger$ and the weighted gaps $G_j^\dagger$ for $j = 2, \ldots, m$.
Step ii)	Find $\hat{p}_c$ that minimises $G_j^\dagger$, $j = 2, \ldots, m$.
Step iii)	For a well-chosen bandwidth k, obtain the local discrepancy L_j's and U_j's.
Step iv)	Obtain the maximiser $\hat{p}_L^\dagger$ of L_j's and $\hat{p}_U^\dagger$ of U_j's, take the midpoint $\hat{p}_{c'} = (\hat{p}_L^\dagger + \hat{p}_U^\dagger)/2$.

Note that

$$p_j^\dagger = \frac{\sum_{i=1}^j p_{(i)}}{j} \le \frac{j p_{(j)}}{j} = p_{(j)}$$

is biased and increases more slowly than $p_{(j)}$. The region where the majority of the alternative p-values appear is maintained, and the center could be right-bounded by the minimiser of $G_j^\dagger$. To overcome its overshooting of p_c, we could compare the above results with the two-sided smoothed version of observed p-values

$$\tilde{p}_j = \begin{cases} \frac{1}{2j-1} \sum_{i=1}^{2j-1} p_{(i)}, & 2j-1 < m, \\ \frac{1}{2m-2j+1} \sum_{i=2j-m}^{m} p_{(i)}, & 2j-1 \ge m. \end{cases}$$

The left panel of the Figure 5.3 shows the scatter plot of the weighted gaps $G_j^\dagger$'s of the p-values, based on a sample from a Cauchy mixture distribution with $m = 200$, $\varepsilon = 0.1$ and $\mu = 8$. Notice

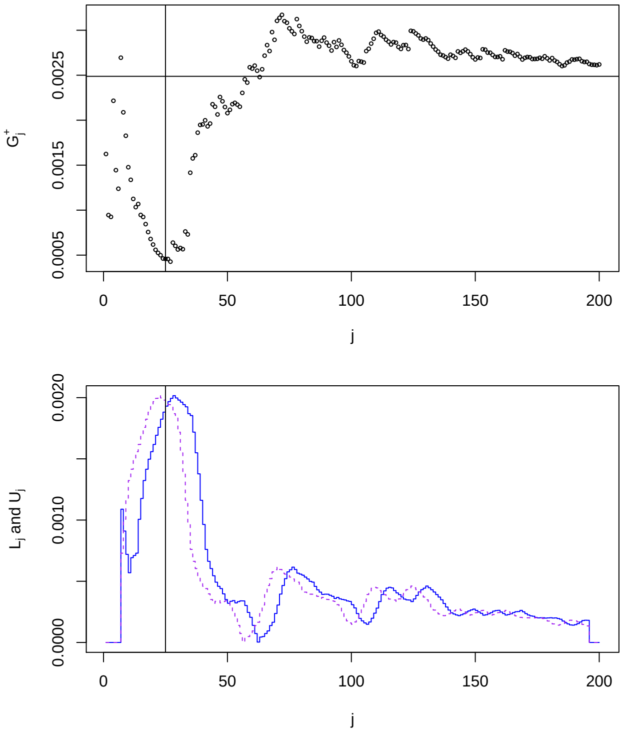


Figure 5.3 – The weighted gaps and the lower and upper local discrepancies

that the weighted gaps are near the expectation $1/(2+2m) = 0.00249$ of the Beta distribution, except for the region where the gaps are far below average. The right panel shows the local discrepancy L_j's and U_j's, in solid blue and dashed purple lines respectively. The midpoint of the two maximisers $\hat{p}_L^\dagger$ and $\hat{p}_U^\dagger$ estimates the center p_c well.

5.1.3 Multi-mode estimation and rejection sets

In this section we discuss another generalisation, the additive model with multi-modes.

Consider the mixture model

$$f(x) = \sum_{j=1}^{k} \pi_j f_0(x - \mu_j), \tag{5.13}$$

of which the proportions π_i's and the shifts μ_i's are unknown and not identical. Following the idea of the two-point mixture model, we propose the rejection sets

$$\mathcal{R} = \left\{ \bigcup_{i=1}^{k} \mathcal{R}^{(i)} \right\} \tag{5.14}$$

where $\mathcal{R}^{(i)} = \mathcal{R}_{\vartheta_i,\delta_i}$ is the i-th rejection interval.

This problem of detecting clustered alternative components is also mentioned in changepoint detection, as is analysed by Siegmund et al. (2011), Zhang et al. (2010), Cao and Biao Wu (2015) et al.

Under the consideration of robustness, we are also interested in testing for unimodality against multimodality. Hartigan and Hartigan (1985) developed the dip test of unimodality, where the dip test statistic is consistent for testing any unimodal against any multimodal distribution. In addition, it is also known that kernel density estimation can capture the multi modes by adjusting the bandwidth h. We recommend Hartigan (1977), Hartigan (1981) and Silverman (1981) for further discussion.

5.2 Conclusions

In general, we are working on multiple testing problems for mixture models, where we consider the test statistics following heavy-tailed distributions. We give a summary of our main contributions in the last part of the thesis.

We present results for testing the two-point mixture models based on the level of p-values, where the nulls and alternatives represent noise and true effects respectively. We seek to *(i)* detect the existence of the non-zero effects, and *(ii)* identify the alternatives from the mixtures. Our contribution makes a difference when the usual assumptions are violated, such as normality, which is a condition that most of the existing methods require. We emphasise the importance of analysing the tail distribution of the test statistics when the generic p-values are used in the testing procedure.

The existence of non-zero effects is investigated by testing the intersection null hypothesis $H_0^{(m)}$, which stands for the pure noise, against the alternative $H_1^{(m)}$, which indicates a mixture of the noise and non-zero effects. We described the difference between testing a light-tailed and a heavy-tailed random variable, of which we are particularly interested in the latter. We proposed the asymptotic detectable region utilising the Kullback–Leibler divergence based on Cauchy mixtures, and as a verification, we compared our methods and results to the classic methodologies developed mainly for Gaussian mixture models. We proved our KL method is equivalent to the likelihood ratio test, in terms of the convergency of the probabilities of type I and type II errors.

In order to locate the true alternatives, we proposed a filtering approach that filters out the p-values that are more likely to be uniformly distributed. With an appropriately defined filter $\mathscr{T}$, the sample of p-values are partitioned into two subsets, the remaining ones $\mathcal{S}^{\mathscr{T}}$ and the eliminated ones $\mathcal{U}^{\mathscr{T}}$. Basically, we achieved the following two goals with the filtering method:

i) Asymptotically, the excluded p-values are from the nulls, and the remaining ones are from the alternatives.

ii) For finite samples, we defined the rejection region based on the sample mode estimated from the filtered p-values. The length of the rejection region is maximised subject to a finite-sample control of FDR.

The proposed filtering approach increases the proportion of the alternatives among the mixture by eliminating the most likely uniformly distributed p-values without disturbing the majority of the alternatives. We proved that the expected value of the false eliminations tends to zero as m goes to zero. In addition, we proved that if the mode of the alternative distribution is unique, then the sample mode estimated using the filtered kernel density estimation is consistent. The mode estimator is thus utilised as the mid-point of the central peak of the p-values from the alternatives, which in our definition, serves as the significance center of the rejection region $\mathcal{R}$. Unlike for the Gaussian test procedures, we define the rejection region $\mathcal{R}_{\vartheta,\delta}$ centralised at the mode ϑ and of length δ. The center ϑ is estimated by the filtering kernel density estimation, and the length δ is chosen by data-dependent control of the FDR. We proved that the expected value of the estimator of FDR provides a good upper bound of the true value of the estimated FDR, such that this data-dependent control functions well. In this procedure we do not propose an estimate of δ. An optimal $\hat{\delta}$ is chosen to achieve the maximal power with the estimated $\widehat{\vartheta}$ bounded by α.

Furthermore, we proposed another non-parametric approach to estimate the local concentration of the p-values under the alternative. We use weighted gap statistics to characterise the sample of the p-values, based on the assumption that the p-values appear in clusters under the alternative. In terms of the multiple testing procedure for mixture models, most methodologies are developed based on the two-point mixture, of which the results are desired to be extended to the general case. We are also exploring the generalisation of our methods to the multimodal mixture models, where the filtering approach and the test for local discrepancy of the gaps will still function well. Ultimately, we are interested in non-parametric methods that are universally applicable to the problems of detecting the alternatives with a location parameter.

Bibliography

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 (Statistical Methodology)*, 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.

Benjamini, Y., Krieger, A. M., and Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3):491–507.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188.

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

Cai, T. T., Jeng, X. J., and Jin, J. (2011). Optimal detection of heterogeneous and heteroscedastic mixtures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):629–662.

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. (2017). Optimal screening and discovery of sparse signals with applications to multistage high-throughput studies. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1):197.

Cao, H. and Biao Wu, W. (2015). Changepoint estimation: another look at multiple testing problems. *Biometrika*, 102(4):974–980.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. John Wiley & Sons.

Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics*, 32(3):962–994.

Duncan, D. B. (1955). Multiple range and multiple f tests. *Biometrics*, 11(1):1–42.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*, 99(465):96–104.

Efron, B. (2010). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*. Cambridge University Press.

Efron, B. and Tibshirani, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology*, 23(1):70–86.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96(456):1151–1160.

Finner, H. and Strassburger, K. (2002). The partitioning principle: a powerful tool in multiple decision theory. *Annals of Statistics*, pages 1194–1213.

Fisher, R. A. (1935). *Design of Experiments*. Oliver & Boyd.

Foster, D. P. and Stine, R. A. (2008). α-investing: a procedure for sequential control of expected false discoveries. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(2):429–444.

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. *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.

Hartigan, J. A. (1977). Distribution problems in clustering. In van Ryzin, G. J., editor, *Classification and Clustering*, pages 45–71. Elsevier.

Hartigan, J. A. (1981). Consistency of single linkage for high-density clusters. *Journal of the American Statistical Association*, 76(374):388–394.

Hartigan, J. A. and Hartigan, P. M. (1985). The dip test of unimodality. *Annals of Statistics*, 13(1):70–84.

Hartley, H. (1955). Some recent developments in analysis of variance. *Communications on Pure and Applied Mathematics*, 8(1):47–72.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Ingster, Y. I. (1998). Minimax detection of a signal for l^n-balls. *Math. Methods Statist.*, 7(4):401–428.

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.

Kullback, S. (1997). *Information Theory and Statistics*. Dover Publications.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Lehmann, E. L. and Romano, J. P. (2005). Generalizations of the familywise error rate. *The Annals of Statistics*, 33(3):1138–1154.

Marcus, R., Eric, P., and Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660.

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.

Miller, R. G. (1966). *Simultaneous Statistical Inference*. McGraw-Hill Book Co.

Newman, D. (1939). The distribution of range in samples from a normal population, expressed in terms of an independent estimate of standard deviation. *Biometrika*, 31(12):20–30.

Roquain, E., Van De Wiel, M. A., et al. (2009). Optimal weighting for false discovery rate control. *Electronic Journal of Statistics*, 3:678–711.

Roquain, E., Villers, F., et al. (2011). Exact calculations for false discovery proportion with application to least favorable configurations. *Annals of Statistics*, 39(1):584–612.

Sarkar, S. K. (2002). Some results on false discovery rate in stepwise multiple testing procedures. *Annals of statistics*, 30(1):239–257.

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, 46(1):561–584.

Shorack, G. R. and Wellner, J. A. (1986). *Empirical Processes with Applications to Statistics*. John Wiley & Sons.

Siegmund, D., Zhang, N., and Yakir, B. (2011). False discovery rate for scanning statistics. *Biometrika*, 98(4):979–985.

Silverman, B. W. (1981). Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 43(1):97–99.

Stefansson, G., Kim, W.-C., and Hsu, J. C. (1988). On confidence sets in multiple comparisons. *Statistical Decision Theory and Related Topics IV*, 2:89–104.

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., 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.

Tukey, J. W. (1953). The problem of multiple comparisons. *Unpublished notes in private circulation*.

Tukey, J. W. (1976). T13 n: The higher criticism. *Course Notes, Statistics 411*.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical science*, 6(1):100–116.

Zhang, N. R., Siegmund, D. O., Ji, H., and Li, J. Z. (2010). Detecting simultaneous changepoints in multiple sequences. *Biometrika*, 97(3):631–645.

Zhiwen JIANG | Curriculum Vitae

PERSONAL PROFILE

Date of birth: 20 February, 1993

Nationality: Chinese

Address: Student Village No. C211, Chemin des Triaudes 4, Ecublens 1024, Switzerland

E-mail: zhiwenjzw@hotmail.com

Telephone: +41 78 648 10 52

EDUCATION

École Polytechnique Fédérale de Lausanne (EPFL) **Lausanne, Switzerland**
PhD in Statistics *2016–2020*
Thesis: Multiple Testing with Test Statistics Following Heavy-tailed Distributions

University of Science and Technology of China (USTC) **Hefei, China**
Master in Statistics *2014–2016*
Thesis: Regularized Covariance Analysis for Multiple Response Longitudinal Data

University of Science and Technology of China (USTC) **Hefei, China**
Bachelor in Statistics *2010–2014*
Thesis: Coordinate Descent on Sparse Estimation of Banding Covariance Matrices

RESEARCH

Problems in multiple testing with test statistics following heavy-tailed distributions
(2016-2020), Thesis advisor: Prof. Stephan Morgenthaler

In multiple testing problem where the observations come from a mixture model of noise and true effects, we want to first test for the existence of the non-null components, and then identify them individually subject to a fixed significance level α. We prove the results of asymptotic detectable boundary for the heavy-tailed distributions, and provide a multiple testing procedure that can discover the true alternatives.

Regularized Covariance Analysis for Multiple Response Longitudinal Data
(2014-2016), Thesis advisor: Prof. Weiping Zhang

For balanced multivariate longitudinal response with blocked covariance, we decompose the covariance into blocked coefficient matrices and covariance matrices of prediction error under generalised auto-regression, and propose the estimator based on the penalised loss function. We propose a modified graphical lasso algorithm and combine it with block-wise coordinate descent methods under the consideration of sparsity.

TEACHING

Autumn 2020: *Statistics for data science*, Master course, EPFL

Springs 2018, 2019: Chief instructor for *Probabilités et statistique*, Bachelor course, EPFL

Autumn 2018: Chief instructor for *Biostatistics*, Master course, EPFL

Autumn 2017: Chief instructor for *Probabilités*, Bachelor course, EPFL

Spring 2017: *Multivariate statistics*, Master course, EPFL

Autumn 2016: *Algèbre linéaire avancée*, Bachelor course, EPFL

Spring 2015: *Probability theory and mathematical statistics*, Bachelor course, USTC

TOPICS OF INTEREST

Multiple testing and multiple comparison
Longitudinal data analysis
Biostatistics and bioinformatics
Statistical learning and modeling

SKILLS

Computer skills: C/C++, R, Matlab, Mathematica, SAS, LaTeX

Languages: Mandarin *(native)*, English *(C1)*, French *(B1)*