UNIVERSITY OF CALIFORNIA SAN DIEGO

Multiple Testing and False Discovery Rate Control: Theory, Methods and Algorithms

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Mathematics (with a Specialization in Statistics)

by

Shiyun Chen

Committee in charge:

Professor Ery Arias-Castro, Chair
Professor Ian Abramson
Professor Karen Messer
Professor Dimitris Politis
Professor Armin Schwartzman

2019

The dissertation of Shiyun Chen is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California San Diego

2019

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deep and sincere gratitude to my advisor Professor Ery Arias-Castro, for his continuous support of my Ph.D. study and research, for his motivation, patience and professional knowledge in mathematics and statistics. He provided me invaluable guidance throughout my research, and encouraged me to carry on research projects independently. It was a great privilege and honor to work and study under his supervision.

Besides, I would like to thank the rest of my doctoral committee: Professor Ian Abramson, Professor Karen Messer, Professor Dimitris Politis, and Professor Armin Schwartzman for offering me excellent courses, insightful comments and encouragement on my research. Outside UCSD, I am immensely grateful for the numerous helpful discussions with Professor Dean Foster, Dr. David Heckerman, Dr. Shiva Kasiviswanathan, Dr. Nina Mishra, and Professor Robert Stine during my internship at Amazon.

I would also like to express my gratitude to all my friends in the Department of Mathematics in UCSD, including but not limited to Yuchao, Andrew, Xiao, Kuangyi, Pengbo, and Rong, for sharing fruitful research ideas, discussions and relaxing time with me. Moreover, I would like to thank all faculty and staff members in the Department of Mathematics who provide me tremendous guidance and help during my Ph.D. life.

Last but not least, I would like to thank my whole family for their support and understanding of my PhD study and all my career. In particular, my husband Jiaqi Gu, from Department of Computer Science, University of California Los Angeles, provides me immense help with the large-sample size numerical experiments and technical solutions related to computer science problems.

Chapter 1, partially, is a version of the paper “Distribution-free Multiple Testing” , Electronic Journal of Statistics, Arias-Castro, Ery; Chen, Shiyun, Volume 11, Number 1 (2017). The dissertation author was the co-author of this paper.

Chapter 2, partially, is a version of the paper “A Scan Procedure for Multiple Testing”,

Chen, Shiyun; Ying Andrew; Arias-Castro, Ery. The manuscript has been submitted for publication in a major statistical journal. The dissertation author was the primary investigator and author of this material.

Chapter 3, partially, is a version of the paper "Sequential Multiple Testing", Chen Shiyun; Arias-Castro, Ery. The manuscript has been submitted for publication in a major statistical journal. The dissertation author was the primary investigator and author of this material.

Chapter 4, partially, is a version of paper "Contextual Online False Discovery Rate Control", Chen, Shiyun; Kasiviswanathan, Shiva. The manuscript has been submitted to a major machine learning conference. The dissertation author was the primary investigator and author of this material.

VITA

2013	Bachelor of Science in Mathematics, East China Normal University
2014-2019	Teaching Assistant, University of California San Diego
2014-2019	Research Assistant, University of California San Diego
2019	Doctor of Philosophy in Mathematics (with a Specialization in Statistics), University of California San Diego

PUBLICATIONS

Arias-Castro, Ery; Chen, Shiyun. "Distribution-free Multiple Testing", *Electronic Journal of Statistics*, 11.1 (2017): 1983-2001.

Chen, Shiyun; Arias-Castro, Ery. "Sequential Multiple Testing", *Submitted*, 2017.

Chen, Shiyun; Ying, Andrew; Arias-Castro, Ery. "A Scan Procedure for Multiple Testing", *Submitted*, 2018.

Chen, Shiyun; Kasiviswanathan, Shiva. "Contextual Online False Discovery Rate Control", *Submitted*, 2018.

ABSTRACT OF THE DISSERTATION

Multiple Testing and False Discovery Rate Control: Theory, Methods and Algorithms

by

Shiyun Chen

Doctor of Philosophy in Mathematics (with a Specialization in Statistics)

University of California San Diego, 2019

Professor Ery Arias-Castro, Chair

Multiple testing, a situation where multiple hypothesis tests are performed simultaneously, is a core research topic in statistics that arises in almost every scientific field. When more hypotheses are tested, more errors are bound to occur. Controlling the false discovery rate (FDR) [BH95], which is the expected proportion of falsely rejected null hypotheses among all rejections, is an important challenge for making meaningful inferences. Throughout the dissertation, we analyze the asymptotic performance of several FDR-controlling procedures under different multiple testing settings. In Chapter 1, we study the famous Benjamini-Hochberg (BH) method [BH95] which often serves as benchmark among FDR-controlling procedures, and show that it is asymptotic optimal in a stylized setting. We then prove that a distribution-free

FDR-controlling method of Barber and Candès [FBC15], which only requires the (unknown) null distribution to be symmetric, can achieve the same asymptotic performance as the BH method, thus is also optimal. Chapter 2 proposes an interval-type procedure which identifies the longest interval with the estimated FDR under a given level and rejects the corresponding hypotheses with P-values lying inside the interval. Unlike the threshold approaches, this procedure scans over all intervals with the left point not necessary being zero. We show that this scan procedure provides strong control of the asymptotic false discovery rate. In addition, we investigate its asymptotic false non-discovery rate (FNR), deriving conditions under which it outperforms the BH procedure. In Chapter 3, we consider an online multiple testing problem where the hypotheses arrive sequentially in a stream, and investigate two procedures proposed by Javanmard and Montanari [JM15] which control FDR in an online manner. We quantify their asymptotic performance in the same location models as in Chapter 1 and compare their power with the (static) BH method. In Chapter 4, we propose a new class of powerful online testing procedures which incorporates the available contextual information, and prove that any rule in this class controls the online FDR under some standard assumptions. We also derive a practical algorithm that can make more empirical discoveries in an online fashion, compared to the state-of-the-art procedures.

Introduction

Multiple hypotheses testing - controlling overall error rates when performing simultaneous hypothesis tests - is a well-established area in statistics with applications in a variety of scientific disciplines [DvdL07, Dic14, Roq11]. This is the so-called discovery science, where the scientist proceeds by formulating multiple hypotheses, testing each one of them on data simultaneously, and selecting the most promising ones. Multiplicity of tests should be taken into consideration even if the experiments are designed and carried out correctly, otherwise the probability that the true null hypotheses are rejected by chance alone may be overwhelmingly high. For instance, consider a setting where n hypotheses are tested, but only a few of them, say n_1, are non-nulls. If we use a fixed significance level α for each test, then we can expect $\alpha(n-n_1)$ true nulls that are falsely rejected, which can substantially exceed the number n_1 of true non-nulls. Such a problem has become even more important with modern data science, where standard data pipelines involve performing a large number of hypothesis tests on complex datasets.

Typically, each test being performed returns one P-value, which we use to decide whether to reject the corresponding null hypothesis, i.e., claim it is false. A hypothesis is rejected if the corresponding P-value is below some significance level. The rejected hypotheses are called *discoveries*, and the subset of these that are true but mistakenly rejected are called *false discoveries*. In the seminal work of [BH95], Benjamini and Hochberg defined the notion of *false discovery rate* (FDR), namely the expected fraction of discoveries that are falsely rejected among all discoveries (a formal definition will be provided later), which is a widely used criterion of type I error rate

for statistical inference in multiple hypothesis testing problems.

They also developed a standard method, known as BH procedure, which can control FDR at a pre-assigned level. Given the P-values $P_1, \ldots, P_n$ where each P_i is associated with the hypothesis $\mathbb{H}_i$, and a desired control level q, the BH procedure proceeds as follows:

1. Let $P_{(i)}$ be the i-th P-value in the (increasing) sorted order, and define $P_{(0)} = 0$. Let

$$i_{\mathrm{BH}} \equiv \max\{0 \le i \le n : P_{(i)} \le iq/n\}. \tag{0.1}$$

2. Reject H_j for every hypothesis with $P_j \le P_{i_{\mathrm{BH}}}$.

As mentioned above, this method can control FDR under the level q when the P-values are independent or under some form of positive dependence. The BH procedure (with some improvement) remains the benchmark in the context of multiple hypothesis testing, and has been studied across the related literature. Since then, FDR-controlling methods have been proposed and in turn adopted by practitioners faced with large-scale testing problems. See [Roq11] for a survey.

However, standard FDR-controlling techniques, such as BH procedure, have some restrictions which do not necessarily apply in practice.

First, it is often assumed, in a substantial proportion of the corresponding literature, that the P-values are available, which implicitly assumes that the null distribution of each test statistic is known perfectly. But this may not always be the case, especially in applications. Therefore it inspires the occurrence of distribution-free or nonparametric procedures in multiple testing, like the recent one proposed by [FBC15].

Second, although a number of variants have been proposed, most of these methods are threshold-type procedures which compute a threshold function based on the P-values and reject the null hypotheses with the corresponding P-values below that threshold [GW04, Sto02, STS04, FBC15]. However, we argue that this is not so obvious in the context of multiple testing, particularly in harder cases where the alternatives are not easily identified and in which most of the

smallest P-values come from true null hypotheses. Under these circumstances, for a pre-chosen level, a more powerful procedure that we consider should not only focus on the those smallest P-values, but also look at the moderate or even larger P-values for candidates of rejection, so that it can obtain more power under the same FDR control level. In applications, such procedure can help us identify the most promising hypotheses for further research. This was already understood by [Chi07] who studied how to modify the BH procedure in order to improve the power. He proposed a complex method which applies the BH procedure at multiple locations in the unit interval, with each location playing the role of the origin, to get multiple interval rejection regions, and then rejects all P-values lying in the union of these intervals. The resulting method was shown to be more powerful than the BH method by both analysis and simulation.

Third, the traditional multiple testing research has focused on the offline setting, which means we have the access to the entire batch of hypotheses and the corresponding P-values. For example, standard FDR-controlling methods like the BH procedure require aggregating P-values for all the tests and processing them jointly. This makes it impossible to utilize them in a number of applications which are best modeled as an online hypothesis testing problem. In this scenario, it is assumed that an infinite sequence of hypotheses arrives in a stream, and decisions are made only based on previous results before the next hypothesis arrives. For example, in marketing research a sequence of A/B tests can be carried out in an online fashion, or in a pharmaceutical drug test a sequence of clinical trials are conducted over time. [FS08] introduced this model and designed the first online alpha-investing rules which use and earn wealth in order to control a modified definition of FDR (referred to as mFDR), which was later extended to a class of generalized alpha-investing (GAI) rules by [AR14]. [JM15] and [JM18] showed that a monotone class of GAI rules can control online FDR in this setting. Of special note is a procedure within this class called LORD (*Levels based on Recent Discovery*) that performs consistently well in practice. [RYWJ17] modified the GAI class (referred as to GAI++) to improve the power of GAI algorithms (uniformly) while controlling FDR, and the improved LORD++ method arguably

represents the current state-of-the-art method in online multiple hypothesis testing. Very recently, [RZWJ18] demonstrated that using adaptiveness, a method called SAFFRON makes some further improvements in the power over LORD++ empirically. We survey these online testing procedures in more detail in Chapter 4.

Fourth, current online testing procedures make a decision by taking the P-value as an input at each time. However, these procedures ignore additional information that is often available in modern applications. In addition to the P-value, each hypothesis could also have some contextual information (also referred to as prior or side information) related to the tested hypothesis. The problem of using prior information has been studied in the offline setting [GRW06, IKZH16, LF18, LB16b, RYWJ17, XZZT17], but not in the online setting where the p-values and contextual features are not available at the onset. We will call such a problem a contextual online hypothesis testing problem and would like to investigate online testing rules which are not only capable of controlling the FDR in an online manner, but also improving the power by incorporating the contextual features into the decision process.

In this dissertation, we investigate and analyze the performance of some existing procedures which have been considered to address some of the above scenarios in traditional (offline) or online multiple testing settings. Besides, for the second scenario, we propose a new simple method which is based on interval rejection type and has better performance compared to the threshold-type BH procedure under certain conditions. For the fourth scenario, we propose a new class of powerful online testing procedures where the rejections thresholds (significance levels) are learned sequentially by incorporating contextual information so far and previous decisions, and compare its performance to the state-of-art online testing procedures.

Throughout the work, we evaluate the performance of each procedure from two perspectives. As false discovery rate is analogous to the type I error rate, the *false non-discovery rate* (FNR) is defined as the expected proportion of true alternatives that are not discovered among all true alternatives, which plays the role of type II error rate in multiple testing problem. We will

also regard the power of a testing procedure as 1 – FNR. [GW02] first introduced the dual notions of false non-discovery rate (slightly different from the definition we use here) and conducted the power analysis. They studied the deciding point (threshold of critical p-value) of BH procedure in the large sample limit. They also defined a risk function of procedures combining both FDR and FNR, and compare the risk of BH procedure with other methods. In this dissertation we consider the risk of a procedure as the sum of FDR and FNR. (See the formal definitions later.)

The dissertation is organized as follows. In Chapter 1, we analyze the asymptotic risk of BH method and the recent distribution-free method of [FBC15] (referred to as BC procedure) under asymptotically generalized Gaussian model (see Definition 1 of AGG model), which includes the normal model. In such a setting, the fraction of alternative hypotheses is tending to 0 with some parametrization. We first derive an asymptotic oracle risk lower bound for all threshold-type multiple testing procedures, and show the asymptotic optimality of BH method in the sense that it can achieve that lower bound. We then consider the distribution-free BC method which only assumes that the test statistics have a common null distribution symmetric about zero. This method has been proved to control FDR under a given level. We also show that in the same model BC procedure can achieve the same asymptotic performance as BH procedure to the first order, thus is also optimal.

In Chapter 2, we are looking beyond the traditional threshold-type procedures, and considering the general interval-type procedures which may yield more power. We propose a method that identifies the longest interval with estimated false discovery rate not exceeding the target level and rejects the corresponding null hypotheses. Unlike the BH method, which does the same but over intervals with an endpoint at the origin, the new procedure 'scans' all intervals. We show that this scan procedure provides a strong control of FDR asymptotically, and at the same time outperforms the BH method asymptotically in the power-law location models under some conditions.

Chapter 3 considers the online hypothesis testing scenario where a possibly infinite

sequence of hypotheses arrives sequentially in a stream, and decisions are made only based on previous decisions. In other words, the decisions have to be made without access to the number of hypotheses in the stream or the future P-values, but solely based on the previous decisions. We propose to use the recent sequential multiple testing procedures (LORD and LOND) of [JM15] which have been proved to control the FDR in an online manner, and study their asymptotic properties in an AGG model. We compare their performance with (static) BH method, and show that LORD can achieve the same power as BH method to the first asymptotic order. We also quantify the performance of LOND.

In Chapter 4, we consider the contextual online multiple testing problem where a vector of contextual features is associated with each hypothesis arriving over time, along with the P-value. Extending the state-of-art online testing rules in [AR14, JM18, RYWJ17], we propose a new broad class of online multiple testing rules, referred to as *contextual generalized alpha-investing* (CGAI) rules, which can incorporate contextual information in the testing process through a general way. We prove that any monotone rule in the class can control the online FDR under a given level with some standard assumptions of P-values. The mFDR control is also obtained under a weaker assumption on P-values. We then focus on a subclass of these rules for designing a practical online FDR control procedure and to compare the statistical power of various procedures. We show the superior performance of our method compared to the state-of-art methods theoretically under some sufficient conditions, and empirically by implementing experiments on both synthetic and real datasets.

Chapter 1

Distribution-free Multiple Testing

1.1 Abstract

We study a stylized multiple testing problem where the test statistics are independent and assumed to have the same distribution under their respective null hypotheses. We first show that, in the normal means model where the test statistics are normal Z-scores, the well-known method of [BH95] is optimal in some asymptotic sense. We then show that this is also the case of a recent distribution-free method proposed by [FBC15]. The method is distribution-free in the sense that it is agnostic to the null distribution — it only requires that the null distribution be symmetric. We extend these optimality results to other location models with a base distribution having fast-decaying tails.

1.2 Introduction

Multiple testing arises in a wide array of applied settings, ranging from anomaly detection in sensor arrays to the selection of genes that are differentially expressed [DvdL07, Dic14]. This is particularly true in so-called discovery science, where the scientist proceeds by formulating

hypotheses, testing each one of them on data, and following up on the most promising ones. Each step along the way is fraught with pitfalls, and even if the experiment was correctly designed and carried out, the scientist still needs to contend with the multitude of tests that were performed.

Multiple testing is now a well-established area in statistics. In a substantial proportion of the corresponding literature it is assumed that P-values are available. This, implicitly, assumes that the null distribution of each test statistic is known (perfectly). For example, the Benjamini-Hochberg (BH) procedure was proposed in this context [BH95]. See [Roq11] for a fairly recent and comprehensive review of the literature, as it pertains to mathematical results in the area.

Our contribution is two-fold. First, we prove that the BH method is asymptotically optimal to first order in the normal (location) model, which corresponds to an idealized setting where the tests being performed are Z-tests and the effect, when present, affects the mean. In fact, we show that this is the case in the much wider context of asymptotically generalized Gaussian models — see Definition 1. Second, we propose to use the recent distribution-free method of [FBC15] that only relies on the assumption that the test statistics have a common null distribution that is symmetric about 0 and show that, in the same normal model, it achieves the same asymptotic performance to first order. This method, proposed in the context of post-model selection inference, is also intimately related to our own work [ACW17] on distribution-free testing of the global null hypothesis.

1.2.1 The risk of a multiple testing procedure

Consider a setting where we want to test n null hypotheses, denoted $\mathbb{H}_1, \ldots, \mathbb{H}_n$. The test that we use for $\mathbb{H}_i$ rejects for large positive values of a statistic X_i. Throughout, we assume that $X_1, \ldots, X_n$ are independent. Denote the vector of test statistics by $\mathbf{X} = (X_1, \ldots, X_n)$. Let Ψ_i denote the survival function[1] of X_i and $\Psi = (\Psi_1, \ldots, \Psi_n)$.

Remark 1. In a large portion of the literature, it is assumed that P-values can be computed (or at

[1] In this chapter, the survival function of a random variable Y is defined as $y \mapsto \mathbb{P}(Y \geq y)$.

Table 1.1: This table summarizes the outcome of applying a multiple testing procedure $\mathcal{R}$ to a particular situation involving n null hypotheses.

	accept null	reject null	total
null true	$U_{\mathcal{R}}$	$V_{\mathcal{R}}$	$\lvert\mathcal{H}^0\rvert$
null false	$T_{\mathcal{R}}$	$S_{\mathcal{R}}$	$\lvert\mathcal{H}^1\rvert$
total	$W_{\mathcal{R}}$	$R_{\mathcal{R}}$	n

least approximated). The simplest such case is when $\mathbb{H}_i$ is a singleton, $\mathbb{H}_i = \{\Psi_i^{\text{null}}\}$, and the null distributions $\Psi_1^{\text{null}}, \dots, \Psi_n^{\text{null}}$ are known. In that case, the i-th P-value is defined as $P_i = \Psi_i^{\text{null}}(X_i)$, which is the probability of exceeding the observed value of the statistic under its null distribution. In this context, working with the statistics $X_1, \dots, X_n$ is equivalent to working with the P-values $P_1, \dots, P_n$.

We will let $\mathcal{H}^0 \subset [n] = \{1, \dots, n\}$ index the true null hypotheses, and $\mathcal{H}^1 \subset [n]$ index the false null hypotheses, meaning

$$\mathcal{H}^0 = \{i \in [n] : \Phi_i \in \mathbb{H}_i\}, \quad \mathcal{H}^1 = \{i \in [n] : \Psi_i \notin \mathbb{H}_i\}. \tag{1.1}$$

A multiple testing procedure $\mathcal{R}$ takes the test statistics $\mathbf{X}$ and return a subset of $\mathcal{R}(\mathbf{X}) \subset [n]$ representing the null hypotheses that the procedure rejects. Table 1.1 describes the outcome when applying some significance rule in such a setting and defines some necessary notations.

Given such a procedure $\mathcal{R}$, the false discovery rate is defined as the expected value of the false discovery proportion [BH95]

$$\text{FDR}_\Psi(\mathcal{R}) = \mathbb{E}_\Psi(\text{FDP}(\mathcal{R}(\mathbf{X}))), \quad \text{FDP}(\mathcal{R}(\mathbf{X})) := \frac{V_{\mathcal{R}}}{R_{\mathcal{R}} \vee 1}, \tag{1.2}$$

While the FDR of a multiple testing procedure is analogous to the type I error rate of a test procedure, the false non-discovery rate (FNR) plays the role of type II error rate and is here

defined as the expected value of the false non-discovery proportion.

$$\mathrm{FNR}_\Psi(\mathcal{R}) = \mathbb{E}_\Psi\big(\mathrm{FNP}(\mathcal{R}(\mathbf{X}))\big), \quad \mathrm{FNP}\big(\mathcal{R}(\mathbf{X})\big) := \frac{T_\mathcal{R}}{|\mathcal{H}^1| \vee 1}. \tag{1.3}$$

where we denoted the cardinality of a set $\mathcal{A} \subset [n]$ by $|\mathcal{A}|$. Note that this definition is different from that introduced in [GW02], where FNP is defined as $T_\mathcal{R}/(n - R_\mathcal{R})$. We find our definition more intuitive, especially under the sparse mixture model where the number of false nulls is very sparse compared to that of true nulls.

In analogy with the risk of a test — which is defined as the sum of the probabilities of type I and type II error — we define the risk of a multiple testing procedure $\mathcal{R}$ as the sum of the false discovery rate and the false non-discovery rate

$$\mathrm{risk}_\Psi(\mathcal{R}) = \mathrm{FDR}_\Psi(\mathcal{R}) + \mathrm{FNR}_\Psi(\mathcal{R}). \tag{1.4}$$

Remark 2. The procedure that never rejects and the one that always reject both achieve a risk of 1, so that any method that has a risk exceeding 1 is useless.

1.2.2 Threshold procedures

We say that a multiple testing procedure $\mathcal{R}$ is of threshold type if it is of the form

$$\mathcal{R}(X_1, \ldots, X_n) = \{i : X_i \geq \tau(X_1, \ldots, X_n)\}, \tag{1.5}$$

for some threshold function τ. For example, the BH method is a threshold procedure based on the P-values — see (1.12).

Because they are so natural in the present context, we will restrict the discussion to threshold procedures. A sizable proportion of the papers in the literature do the same [Roq11], for example, the case of [Sto07]. In particular, the lower bound that we develop (Theorem 1) is

only meant to apply to such procedures.

1.2.3 The normal model and the optimality of the BH method

This model corresponds to the setting above with $X_i \sim \Psi_i = \mathcal{N}(\mu_i, 1)$ and $\mathbb{H}_i : \mu_i = 0$, so that $\mathbb{H}_i$ is a singleton equal to $\Psi_i^{\text{null}} = \mathcal{N}(0,1)$. In this context it is compelling to ask how large the μ_i's need to be in order for the risk of the BH procedure to tend to zero. To the best of our knowledge, this question has not been directly answered in the literature.

Our inspiration for considering the normal (location) model comes from the seminal work of Ingster [Ing97, IS03] and [DJ04] on testing the global null $\bigcap_i \mathbb{H}_i$. In [Ing97] we find the following first-order asymptotic result. Assume a prior under which $m \le n$ randomly picked μ_i's are set to $\sqrt{2r\log n}$ and the others are set to 0. An interesting parameterization happens to be $m/n \sim n^{-\beta}$ with $\beta > 0$ fixed. Focusing on the so-called sparse regime, where $\beta > 1/2$, one finds that the detection boundary is at $r = \rho(\beta)$, where

$$\rho(\beta) = \begin{cases} \beta - 1/2, & 1/2 < \beta \le 3/4, \\ (1 - \sqrt{1-\beta})^2, & 3/4 < \beta < 1. \end{cases} \tag{1.6}$$

This means that, taking r to be fixed, when $r < \rho(\beta)$ all tests have risk at least 1 in the large sample limit (which is as bad as random guessing), while when $r > \rho(\beta)$ the likelihood ratio test has risk 0 in the large sample limit. [DJ04] propose an adaptive test procedure based on Tukey's higher criticism that achieves this optimal detection boundary. (The higher criticism also achieves the detection boundary over $\beta \le 1/2$ not displayed here.)

Returning to the question of identifying the false null hypotheses, which is our concern here, we know that $r > 1$ allows for the identification of the false nulls with a control of the family-wise error rate (FWER) at any fixed level. In fact, if we define the corresponding risk as the sum of FWER and the probability of at least one false non-discovery, then $r = 1$ is the

precise boundary for this to be controlled, and the Bonferroni procedure achieves the boundary over $\beta \in (0,1)$. In this chapter, we focus instead on controlling the risk (1.4) involving FDR. The following is a special case of a more general lower bound appearing later in the chapter.

Corollary 1. *In the normal model, assume that* $\beta \in (0,1)$ *and* $r \geq 0$ *are both fixed. If* $r < \beta$*, then the risk of any threshold procedure has limit inferior at least 1 as* $n \to \infty$.

In our context, we know that Corollary 1 is tight because the BH method (which is a threshold procedure) achieves the stated selection boundary with FDR control level set at some $q \to 0$ slowly. The following is also a special case of a more general result appearing later in the chapter.

Corollary 2. *In the setting of Corollary 1, if instead* $r > \beta$*, then the risk of the BH procedure (properly calibrated) tends to 0 as* $n \to \infty$.

It is worth remembering that BH method is known to control the FDR at the prescribed level [BH95], so the result is really about its (asymptotic) control of the FNR.

Together, Corollary 1 and Corollary 2 establish the BH procedure as asymptotically optimal to first order in the normal model among threshold procedures. We will see that this remains true for a much wider class of models.

Remark 3. While the equation $r = \rho(\beta)$ defines the "detection boundary" in the (β, r) plane for testing the global null $\bigcap_i \mathbb{H}_i$ in the normal model, the equation $r = \beta$ is the "selection boundary" for the same multiple testing problem. Intuitively, detection is "easier" than selection, and this is confirmed in the fact that the detection boundary is entirely below the selection boundary — indeed, $\beta > \rho(\beta)$ for all $\beta \in (0,1)$.

1.2.4 Multiple testing under symmetry

The P-values are based on the assumed knowledge of the null distribution of each test statistic. In many practical settings, this is not strictly the case, resulting in P-values that are only

approximately uniformly distributed under their respective null hypothesis. This can jeopardize the control of the FDR. In the same way that it may be appealing in some situations to use a distribution-free test such as the signed-rank test instead of the t-test, it may also be desirable to use a distribution-free procedure for multiple testing.

Our working assumption is the following

$$X_1, \dots, X_n \text{ are independent with common null distribution that is symmetric about 0.} \tag{1.7}$$

This assumption might be reasonable in some crossover trials. Although testing the global null is more typical in such a setting (and one might apply the signed-rank test), a proper multiple testing analysis may be carried out when it is desired to identify which subjects truly benefited from treatment.

The assumption of symmetry is at the very core of the literature on nonparametric tests [Het84]. And it is also quite natural in the context of multiple testing. For example, under these assumptions, [DR06] consider testing the global null and propose a test based on sign flips, while [ACW17] propose a nonparametric analog to the higher criticism. Beyond testing the global null, [ABR10] propose a resampling procedure also based on sign flips with the purpose of controlling the FWER in a setting that also allows for dependence, while [FBC15] propose a nonparametric analog to the BH method.

We call the latter the Barber-Candès (BC) procedure — see Section 1.5 for a proper definition. We study this method and show that, under fairly general conditions, it achieves the selection boundary. In particular, it does as well as the BH procedure which requires the knowledge of the null distributions. The following is a special case of a more general results appearing later on.

Corollary 3. *The conclusions of Corollary 2 apply to the BC procedure.*

The BC method is shown in [FBC15] to control the FDR at the desired level, so the result

is really about its (asymptotic) control of the FNR.

1.2.5 More related work

Our contribution is thus twofold: we obtain an asymptotic oracle risk bound for multiple testing and then show that the BH achieves that bound; and we show that the distribution-free BC method also achieves that bound.

Various oracle bounds are available in the literature. In a context where the P-values are uniformly distributed under the null and have the same distribution under the alternative, [GW02] consider an oracle that knows the number of false null hypotheses $|\mathcal{H}^1|$ and the common alternative distribution. See also [SC07, Sto07, BCFG11, NR12]. [MMB11] also discusses oracle bounds but in a different setting where FWER control is the goal.

The notion of risk considered here (1.4), although natural to us, seems new. More common is the risk corresponding to Hamming loss, very popular in the classification literature. In our notation, for a procedure $\mathcal{R}$, this risk is defined as follows

$$\mathrm{risk}_{\Psi}^{\mathrm{Hamming}}(\mathcal{R}) = \mathbb{E}_{\Psi}(|\mathcal{R} \triangle \mathcal{H}^1|). \tag{1.8}$$

For example, this risk is considered in [GW02, SC07, JJ12, JK14, BST15, BCFG11, NR12]. All these papers provide some asymptotic analysis of the Hamming risk, whether from a minimax or oracle perspective. In this context, [GW02, BCFG11, NR12] compare the performance of BH method to that of an oracle, concluding that the BH method comes close to achieving the oracle bound under some conditions.

Other distribution-free procedures have been suggested in the literature. Most are based on resampling [WY93, GDS03, YB99, RW07, ABR10]. These methods are not applicable in the setting assumed here. They are typically applied to situations, as in microarray analysis, where each test statistic is based on comparing two (or more) samples. Another class of methods consist

in estimating the null distribution — assumed common to all test statistics — and the alternative distribution — also assumed to be common to all test statistics — with the goal of imitating the oracle thresholding method based on that knowledge. This is advocated in [Efr04, PvdL04], for example. [Sto07] and [SC07] discuss such procedures and derive performance bounds. Such methods rely on the ability to estimate the mixture consistently. There is work in that direction in [JC07, CJ10].

Although not as directly related, [ABDJ06] consider the problem of estimating the mean vector $(\mu_1, \ldots, \mu_n)$ in the normal model, and show that hard thresholding with the BH threshold is asymptotically minimax in some settings.

1.2.6 Content

In Section 1.3 we derive an oracle bound on the boundary for multiple testing in a location model where the base distribution is asymptotically generalized Gaussian. This comprises the normal model. In Section 1.4 we analyze the performance of the BH procedure based on the full knowledge of the null distribution, while in Section 1.5 we analyze the performance of the BC procedure. We present the result of some numerical experiments in Section 1.6. The proofs are gathered in Section 1.7.

1.3 The AGG model

We start by defining an oracle threshold procedure, which will serve as benchmark on a family of location models where the base distribution is asymptotically polynomial in log-scale — which in particular encompasses the normal model. The result is an oracle risk bound.

1.3.1 The oracle procedure

We consider an oracle that provides $\mathcal{H}^1$ and use that information to optimize the threshold in terms of minimizing the risk at a particular realization, namely,

$$\tau_o(\mathbf{X}) \in \operatorname*{arg\,min}_{t\in\mathbb{R}} \ \mathrm{FDP}(\mathcal{R}_t(\mathbf{X})) + \mathrm{FNP}(\mathcal{R}_t(\mathbf{X})), \qquad \mathcal{R}_t(\mathbf{X}) := \{i : X_i \geq t\}. \tag{1.9}$$

In words, with full knowledge of the set of false null distributions $\mathcal{H}^1$, the procedure chooses a threshold that partitions the test statistics in a way that minimizes the sum of the false discovery and non-discovery proportions. The expected risk of this procedure is what we call below the oracle risk.

Remark 4. Of course, if one knew $\mathcal{H}^1$, one would simply reject $\mathbb{H}_i$ for all $i \in \mathcal{H}^1$ and, in the end, there would not any multiple testing problem to deal with! The oracle procedure is, however, constrained to be of threshold type, with the goal of serving as a benchmark for threshold-type procedures.

Our oracle is the strongest possible, in the sense that it provides $\mathcal{H}^1$, and we use the oracle information to optimize the threshold. Most other publications that discuss oracle bounds, such as [GW02, SC07, Sto07, BCFG11, NR12], operate in a setting where the statistics have the same null distribution and the same alternative distribution, and consider an oracle that provides these two distributions together with the number of false null hypotheses $|\mathcal{H}^1|$; this oracle information is then used to optimize a *constant* threshold. [MMB11] consider an oracle that provides $\mathcal{H}^1$, and well as the joint distribution of the P-values indexed by $\mathcal{H}^0$, and use that oracle information to obtain an optimized single-step procedure for FWER control.

1.3.2 Asymptotically generalized Gaussian model

In a location model, we assume that we know the null survival function Ψ, assumed to be continuous for simplicity, and consider $\Psi(\cdot - \mu)$ as a location family of distributions. We then

assume that the test statistics are independent with respective distribution $X_i \sim \Psi_i = \Psi(\cdot - \mu_i)$, where $\mu_i = 0$ under the null $\mathbb{H}_i$ and $\mu_i > 0$ otherwise. Both minimax and Bayesian considerations lead to considering a prior on the μ_i's where $m \leq n$ randomly picked μ_i's are set equal to some $\mu > 0$ and the others are set to 0. The prior is therefore defined based on m and μ, which together control the signal strength.

Beyond the normal model, we consider other location models where the base distribution has a polynomial right tail in log scale.

Definition 1. A survival function Ψ is asymptotically generalized Gaussian (AGG) on the right with exponent $\gamma > 0$ if $\lim_{x\to\infty} x^{-\gamma} \log \Psi(x) = -1/\gamma$.

The AGG class of distributions is nonparametric and quite general. It includes the parametric class of generalized Gaussian (GG) distributions with densities $\{\psi_\gamma, \gamma > 0\}$ given by $\log \psi_\gamma(x) \propto -|x|^\gamma/\gamma$, which comprises the normal distribution ($\gamma = 2$) and the double exponential distribution ($\gamma = 1$).

Remark 5. We note that the scale (e.g., standard deviation) is fixed, but this is really without loss of generality as both the BH and BC methods are scale invariant. For the BH method, this is because the P-values are scale invariant. However, this is so because we provide the BH method with the null distribution, including the scale. The BC method, by contrast, can operate without knowledge of the scale.

[DJ04] consider the problem of testing the global null in a GG location model and derived the detection boundary. We use the same prior, where m nulls chosen uniformly at random are designated to be false and all positive μ_i's are set equal to μ, with

$$m = \lfloor n^{1-\beta} \rfloor, \quad \text{with } 0 < \beta < 1 \quad \text{(fixed)}, \tag{1.10}$$

and

$$\mu = \mu_\gamma(r) = (\gamma r \log n)^{1/\gamma}, \quad \text{with } r > 0 \quad \text{(fixed)}. \tag{1.11}$$

[NR12] obtain general bounds on the excess (Hamming) risks of Bayesian FDR and the BH method relative to an oracle, which they specialize to the GG model, showing that under similar conditions the BH method achieves an oracle bound.

Theorem 1. *Consider a location model where the base distribution is AGG with exponent* $\gamma > 0$*, with prior described above, and with the parameterization* (1.10)-(1.11)*. If* $r < \beta$*, then the oracle risk has limit inferior at least 1 as* $n \to \infty$*.*

1.4 The performance of the BH method

We order the X_i's in *decreasing* order, to obtain the following order statistics $X_{(1)} \geq \cdots \geq X_{(n)}$. Given a desired FDR control at q, the BH procedure of [BH95] is defined as the threshold procedure (1.5), with threshold

$$\tau_{\mathrm{BH}} = X_{(\iota_{\mathrm{BH}})}, \quad \iota_{\mathrm{BH}} := \max\left\{i : X_{(i)} \geq \Psi^{-1}(iq/n)\right\}. \tag{1.12}$$

This procedure is shown in [BH95] to control the FDR at q when the tests are independent — which we assume throughout.

Typically, q is set to a small number, like $q = 0.10$. In this chapter we allow $q \to 0$ as $n \to \infty$, but slowly. Specifically, we always assume that

$$q = q(n) > 0 \text{ such that } n^a q(n) \to \infty \text{ for all fixed } a > 0. \tag{1.13}$$

The following result establishes the BH procedure as optimal in the AGG model, in the sense that it achieves the selection boundary ($r = \beta$) stated in Theorem 1.

Theorem 2. *In the setting of Theorem 1, if instead* $r > \beta$*, then the BH procedure with* q *satisfying* (3.11) *has* FNR *tending to 0 as* $n \to \infty$*. In particular, if* $q \to 0$*, then it has risk tending to 0 since the procedure has* $\mathrm{FDR} \leq q$*.*

Remark 6. For any multiple testing procedure, $\text{FNR} \to 0$ if and only if $\text{FNP} \to 0$ in probability. Indeed, one direction is justified by Markov's inequality, and the other direction is justified by dominated convergence and the fact that $\text{FNP} \leq 1$.

1.5 The performance of the BC method

Under the assumption of symmetry, given the desired FDR control level q, the Barber-Candès (BC) procedure defines the data-dependent threshold τ_{BC} as:

$$\tau_{\text{BC}} = \inf\left\{t \in |\mathbf{X}| : \widehat{\text{FDP}}_{\text{BC}}(t) \leq q\right\}, \tag{1.14}$$

where, as usual, the infimum is infinite if the set is empty, $|\mathbf{X}| := \{|X_i| : i = 1, \ldots, n\}$ is the set of sample absolute values, and

$$\widehat{\text{FDP}}_{\text{BC}}(t) := \frac{1 + \#\{i : X_i \leq -t\}}{1 \vee \#\{i : X_i \geq t\}}, \tag{1.15}$$

is a measure of how asymmetric the set of observations $\{X_i : |X_i| \geq t\}$ is.

The notation is borrowed from [FBC15] and is justified by the fact that this quantity aims at estimating $\text{FDP}(\mathcal{R}_t)$, where $\mathcal{R}_t = \{i : X_i \geq t\}$ as in (1.9).

The BC procedure is shown in [FBC15] to control the FDR at level q.

The following result shows that, although agnostic to the null distribution, the BC procedure achieves the selection boundary in a AGG model as long as the underlying distribution is symmetric.

Theorem 3. *In the setting of Theorem 1, and assuming that the null distribution Ψ is symmetric about 0, if instead $r > \beta$, then the BC procedure with q satisfying* (3.11) *has* FNR *tending to 0 as $n \to \infty$. In particular, if $q \to 0$, then it has risk tending to 0 since the procedure has* $\text{FDR} \leq q$.

1.6 Numerical experiments

In this section, we perform simple simulations to compare the BH and BC procedures on finite data, with the goal of illustrating the theory we established. We consider the normal model and the double-exponential model. It is worth repeating that the BH procedure requires knowledge of null distribution as it is based on the P-values. In contrast, the BC method does not require knowledge of the null distribution.

1.6.1 Fixed sample size

In this first set of experiments, the sample size is chosen large at $n = 10^5$. The FDR control level is set at $q = 0.05$. We draw m observations from the alternative distribution $\Psi(\cdot - \mu)$, and the other $n - m$ from the null distribution Ψ. All the models are parameterized as described in Section 1.3.2, in particular, (1.10) and (1.11). We choose a few values for the parameter β so as to exhibit different sparsity levels, while the parameter r takes values in a grid of spanning $[0, 1]$. Each situation is repeated 500 times and we report the average FDP and FNP for each procedure.

Normal model

In this model Ψ is the standard normal distribution. The simulation results are reported in Figure 1.1 and Figure 1.2. In Figure 1.1 we report the FDP. Recall that the methods are set to control the FDR at the desired level ($q = 0.05$). We see that the BC method becomes more conservative than the BH method as β increases. In Figure 1.2 we report the FNP. We see that the BC method performs comparably to the BH method at $\beta = 0.3$ and $\beta = 0.5$, but is clearly less powerful in the sparsest regime $\beta = 0.7$. This is in line with the earlier observation that the BC method becomes more conservative with increasing values of β. It can also be explained by the fact, at $\beta = 0.7$, the number of false nulls ($m = 31$ out of $n = 10^5$) is too small to reveal the asymptotic power of the BC method. Finally, we remark that the transition from high FNP to low

FNP happens in the vicinity of the theoretical threshold ($r = \beta$).

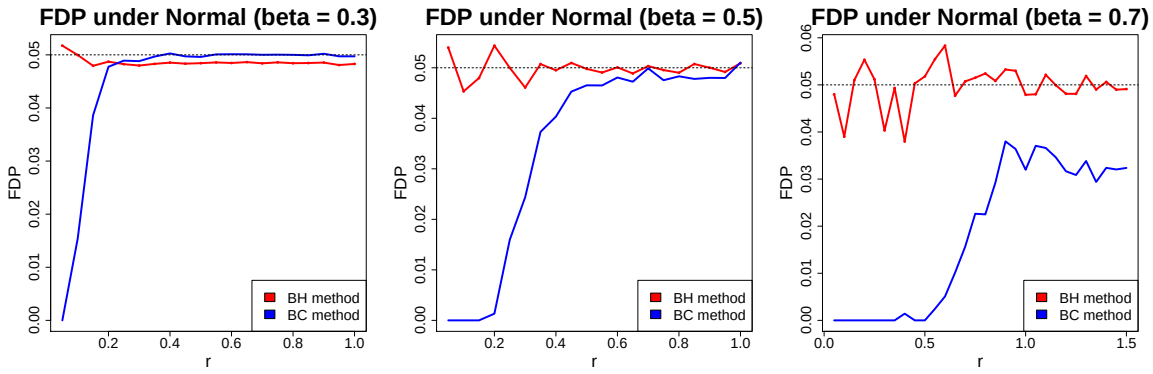

Figure 1.1: Simulation results showing the FDP for the BH and BC methods under the normal model in three distinct sparsity regimes. The black horizontal line delineates the desired FDR control level ($q = 0.05$).

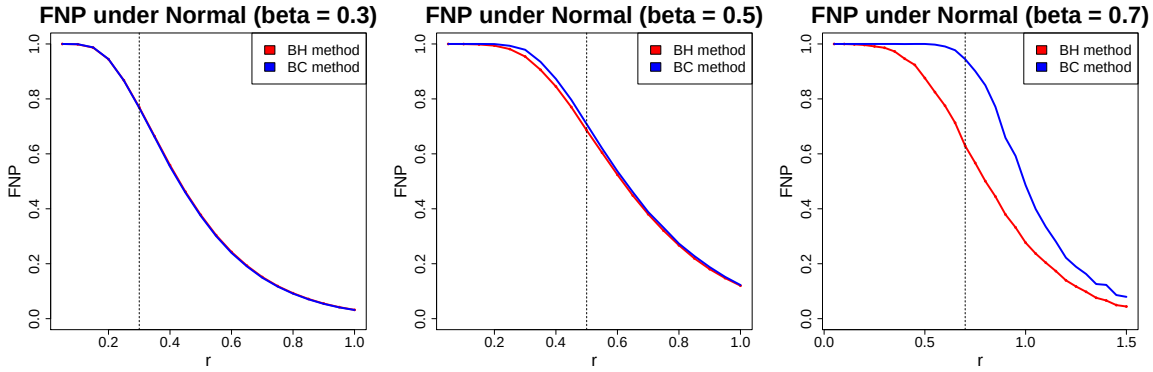

Figure 1.2: Simulation results showing the FNP for the BH and BC methods under the normal model in three distinct sparsity regimes. The black vertical line delineates the theoretical threshold ($r = \beta$).

Double-exponential model

In this model Ψ is double-exponential distribution with variance of 1. The simulation results are reported in Figure 1.3 (FDP) and Figure 1.4 (FNP). Here we observe that the BC method is rather conservative regardless of β. The two methods are again comparable in terms of FNP, in fact a bit more so than in the normal setting. The transition from FNP near 1 to FNP near 0 happens, again, in the vicinity of the theoretical threshold, but is much sharper here.

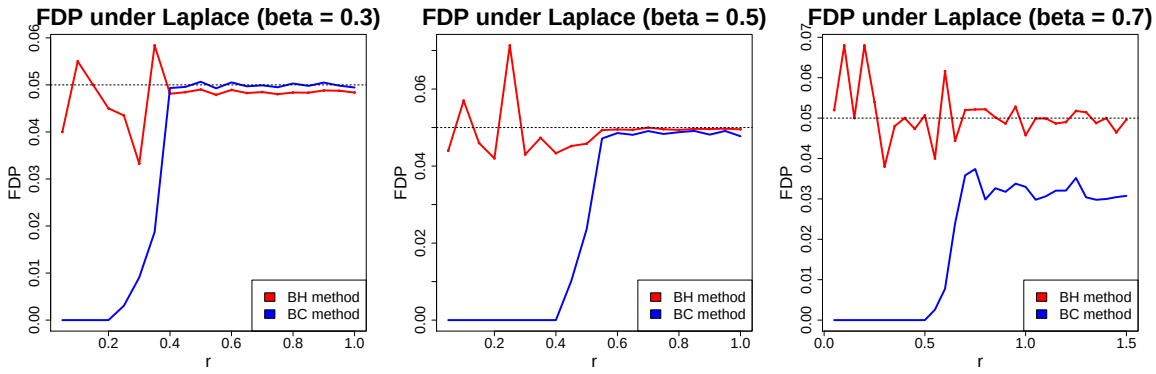

Figure 1.3: Simulation results showing the FDP for the BH and BC methods under the double-exponential model in three distinct sparsity regimes. The black horizontal line delineates the desired FDR control level ($q = 0.05$).

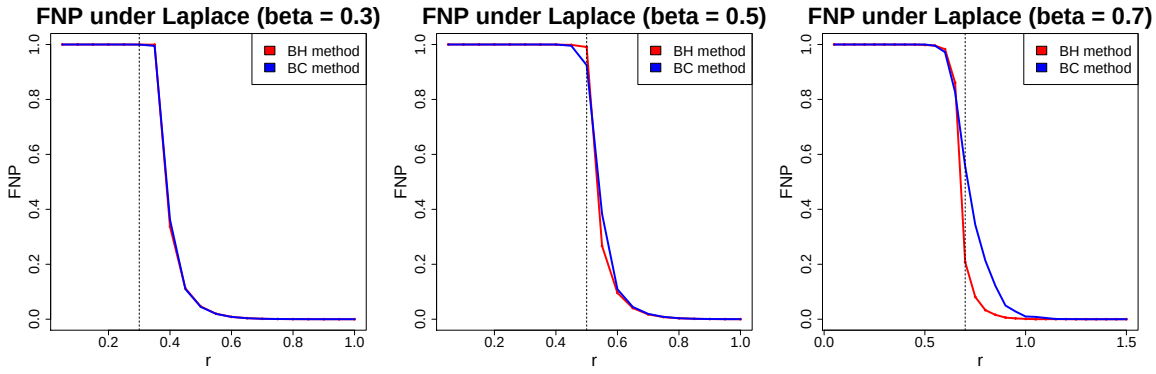

Figure 1.4: Simulation results showing the FNP for the BH and BC methods under the double-exponential model in three distinct sparsity regimes. The black vertical line delineates the theoretical threshold ($r = \beta$).

1.6.2 Varying sample size

In this second set of experiments, we examine the effect of various sample sizes on the risk of BH and BC procedures under the standard normal model and the double-exponential model (with variance 1). We simultaneously explore the effect of letting the desired FDR control level q tend to 0, in accordance with (3.11). Specifically, we set it as $q = q_n = 1/\log n$. We choose n on a log scale, specifically, $n \in \{10^2, 10^3, 10^4, 10^5, 10^6\}$. Each time, we fix a value of (β, r) such that $r > \beta$.

In the first setting, we set $(\beta, r) = (0.4, 0.9)$. The simulation results are reported in

Figure 1.5 and Figure 1.6. We see that, in both models, the risks of the two procedures decrease to zero rapidly as the sample size gets larger. The BH method clearly dominates (in terms of FNP) up until $n = 10^3$, and after that the two methods behave similarly.

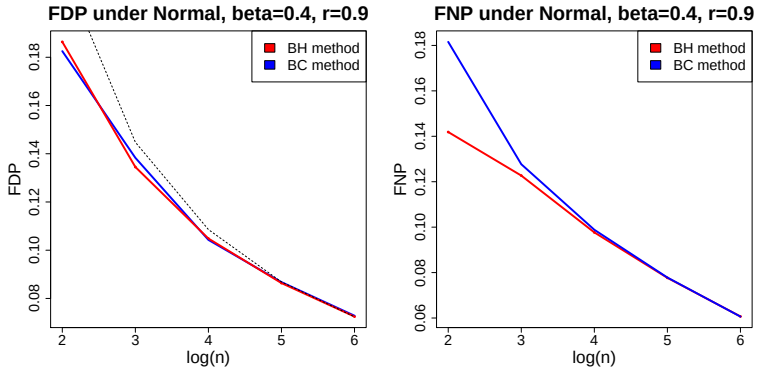

Figure 1.5: FDP and FNP for the BH and BC methods under the normal model with $(\beta, r) = (0.4, 0.9)$ and varying sample size n.

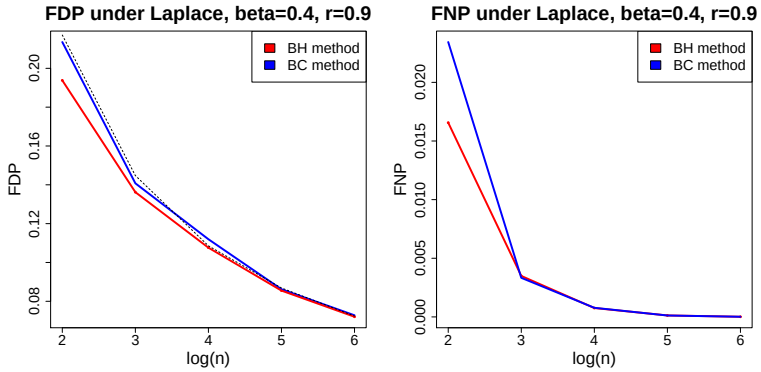

Figure 1.6: FDP and FNP for the BH and BC methods under the double-exponential model with $(\beta, r) = (0.4, 0.9)$ and varying sample size n.

In the second setting, we set $(\beta, r) = (0.7, 1.5)$ for normal model and $(\beta, r) = (0.7, 1.2)$ for double-exponential model. The simulation results are reported in Figure 1.7 and Figure 1.8. In this sparser regime, we can see that the BC method is much more conservative than BH method when n is relatively small. But as n gets larger, this is less pronounced. The BH method clearly dominates (in terms of FNP) up until $n = 10^3$ and past $n = 10^4$ the two methods behave similarly. The difference is much more dramatic here, in line with our findings in Section 1.6.1.

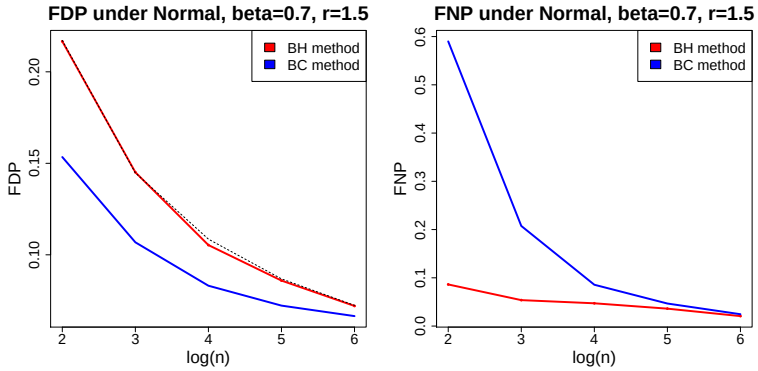

Figure 1.7: FDP and FNP for the BH and BC methods under the normal model with $(\beta, r) = (0.7, 1.5)$ and varying sample size n.

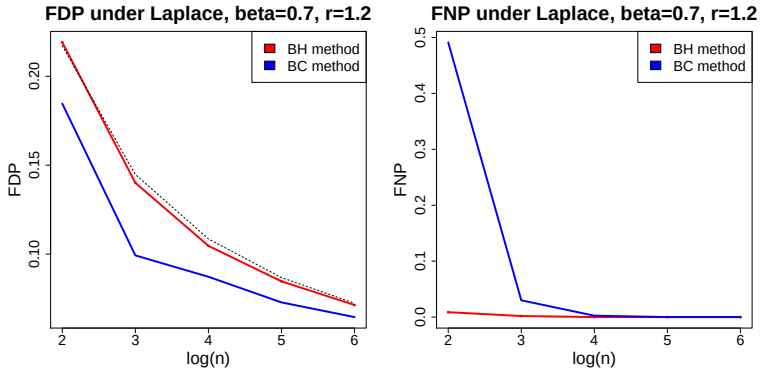

Figure 1.8: FDP and FNP for the BH and BC methods under the double-exponential model with $(\beta, r) = (0.7, 1.2)$ and varying sample size n.

1.7 Proofs

We prove our results in this section.

1.7.1 Proof of Theorem 1

For $t \in \mathbb{R}$, recall that $\mathcal{R}_t = \{i : X_i \geq t\}$ and define

$$\text{the number of type I errors:} \quad \mathrm{I}(t) = \left|\mathcal{R}_t \setminus \mathcal{H}^1\right| ; \tag{1.16}$$

$$\text{the number of type II errors:} \quad \mathrm{II}(t) = \left|\mathcal{H}^1 \setminus \mathcal{R}_t\right| . \tag{1.17}$$

Set $\delta = \log\log n$. We distinguish between two cases.

- When $t \le \mu + \delta$, using the fact that $t \mapsto \mathrm{I}(t)$ is non-increasing, we have

$$\mathrm{FDP}(\mathcal{R}_t) = \frac{\mathrm{I}(t)}{|\mathcal{R}_t|} \ge \frac{\mathrm{I}(t)}{\mathrm{I}(t) + |\mathcal{H}^1|} \ge \frac{\mathrm{I}(\mu+\delta)}{\mathrm{I}(\mu+\delta) + m}. \tag{1.18}$$

- When $t > \mu + \delta$, using the fact that $t \mapsto \mathrm{II}(t)$ is non-decreasing, we have

$$\mathrm{FNP}(\mathcal{R}_t) = \frac{\mathrm{II}(t)}{|\mathcal{H}^1|} \ge \frac{\mathrm{II}(\mu+\delta)}{m}. \tag{1.19}$$

(Recall that $m = |\mathcal{H}^1|$ in our model.) Hence, we conclude that for any $t \in \mathbb{R}$,

$$\mathrm{FDP}(\mathcal{R}_t) + \mathrm{FNP}(\mathcal{R}_t) \ge \frac{\mathrm{I}(\mu+\delta)}{\mathrm{I}(\mu+\delta) + m} \wedge \frac{\mathrm{II}(\mu+\delta)}{m}. \tag{1.20}$$

Consequently, to show that the oracle threshold risk has limit inferior at least 1 as n tends to infinity, by dominated convergence it suffices to show that the RHS tends to 1 in probability, or put differently, that

$$\frac{\mathrm{I}(\mu+\delta)}{m} \to \infty \quad \text{and} \quad \frac{\mathrm{II}(\mu+\delta)}{m} \to 1, \quad \text{in probability as } n \to \infty. \tag{1.21}$$

On the one hand, we have $\mathrm{I}(\mu+\delta) \sim \mathrm{Bin}(n-m, \Psi(\mu+\delta))$, so that for $\mathrm{I}(\mu+\delta)/m$ to diverge to ∞ in probability it suffices that $(n-m)\Psi(\mu+\delta)/m \to \infty$. And indeed, this is the case since

$$\log\Big[(n-m)\Psi(\mu+\delta)/m\Big] = \log(n/m) + o(1) + \log\Psi(\mu+\delta) \tag{1.22}$$

$$= \log(n/n^{1-\beta}) + o(1) - \tfrac{1}{\gamma}(\mu+\delta)^{\gamma}(1+o(1)) \tag{1.23}$$

$$= \beta\log n + o(1) - (r+o(1))\log n \tag{1.24}$$

$$= (\beta - r + o(1))\log n \to \infty, \tag{1.25}$$

using the fact that $m \sim n^{1-\beta}$, that Ψ is AGG with exponent γ, that $\mu + \delta \sim \mu$ with μ defined in (1.11), and that $r < \beta$.

On the other hand, we have $\mathrm{II}(\mu+\delta) \sim \mathrm{Bin}(m, 1-\Psi(\delta))$, so that for $\mathrm{II}(\mu+\delta)/m$ to converge to 1 in probability it suffices that $\Psi(\delta) \to 0$, which is the case since $\delta \to \infty$.

1.7.2 Proof of Theorem 2

Let Ψ denote the null survival function, assumed to be AGG with parameter $\gamma > 0$. Let $\hat{G}$ denote the empirical survival function

$$\hat{G}(t) = \frac{1}{n} \sum_{i \in [n]} \mathbb{I}\{X_i \geq t\}. \tag{1.26}$$

Let $Y_i = X_i - \mu_i$ and note that $(Y_i : i \in [n])$ are IID with distribution Ψ. Define the empirical survival functions

$$\hat{W}_{\text{true}}(y) = \frac{1}{n-m} \sum_{i \notin \mathcal{H}^1} \mathbb{I}\{Y_i \geq y\}, \quad \hat{W}_{\text{false}}(y) = \frac{1}{m} \sum_{i \in \mathcal{H}^1} \mathbb{I}\{Y_i \geq y\}, \tag{1.27}$$

so that

$$\hat{G}(t) = (1-\varepsilon)\hat{W}_{\text{true}}(t) + \varepsilon \hat{W}_{\text{false}}(t-\mu). \tag{1.28}$$

where $\varepsilon := m/n \sim n^{-\beta}$ under (1.10).

We need the following result to control the deviations of the empirical distributions.

Lemma 1 ([Eic79])**.** *Let $Z_1, \ldots, Z_k$ be IID with continuous survival function Q. Let $\hat{Q}_k$ denote their empirical survival function and define $\zeta_k = \sqrt{2\log\log(k)/k}$ for $k \geq 3$. Then*

$$\frac{1}{\zeta_k} \max_z \frac{\hat{Q}_k(z) - Q(z)}{\sqrt{Q(z)(1-Q(z))}} \to 1, \text{ in probability as } k \to \infty. \tag{1.29}$$

In particular,

$$\hat{Q}_k(z) = Q(z) + O_{\mathbb{P}}(\zeta_k)\sqrt{Q(z)(1-Q(z))}, \quad \text{uniformly in } z. \tag{1.30}$$

Applying Lemma 1, we get

$$\hat{G}(t) = (1-\varepsilon)\big[\Psi(t) + O_{\mathbb{P}}(\zeta_n)\sqrt{\Psi(t)(1-\Psi(t))}\big] \tag{1.31}$$

$$+\varepsilon\big[\Psi(t-\mu) + O_{\mathbb{P}}(\zeta_m)\sqrt{\Psi(t-\mu)(1-\Psi(t-\mu))}\big]. \tag{1.32}$$

From this we get

$$\hat{G}(t) = G(t) + \hat{R}(t), \tag{1.33}$$

where

$$G(t) := \mathbb{E}[\hat{G}(t)] = (1-\varepsilon)\Psi(t) + \varepsilon\Psi(t-\mu), \tag{1.34}$$

and

$$\hat{R}(t) = O_{\mathbb{P}}\big(\zeta_n\sqrt{\Psi(t)(1-\Psi(t))} + \zeta_m\varepsilon\sqrt{\Psi(t-\mu)(1-\Psi(t-\mu))}\big), \quad \text{uniformly in } t \in \mathbb{R}. \tag{1.35}$$

Let $\iota = \iota_{\mathrm{BH}}$ be defined as in (1.12). We have $\hat{G}(X_{(i)}) = i/n$, so that $X_{(i)} \geq \Psi^{-1}(q\hat{G}(X_{(i)}))$ for $i \leq \iota$ and $X_{(i)} < \Psi^{-1}(q\hat{G}(X_{(i)}))$ for $i > \iota$. Based on that, and the fact that $\hat{G}$ is constant between two consecutive X_i's, we have that there is $\tau \in (X_{(\iota+1)}, X_{(\iota)}]$ such that

$$\tau = \min\big\{t : t \geq \Psi^{-1}(q\hat{G}(t))\big\} = \min\big\{t : t = \Psi^{-1}(q\hat{G}(t))\big\}. \tag{1.36}$$

Note that the BH procedure coincides with $\mathcal{R}_\tau$, the threshold method with threshold τ. In particular,

$$\mathrm{FNP}(\mathcal{R}_\tau) = 1 - \hat{F}(\tau), \quad \hat{F}(t) := \frac{1}{m}\sum_{i\in\mathcal{H}^1} \mathbb{I}\{X_i \geq t\}, \tag{1.37}$$

so that it suffices to show that $\hat{F}(\tau) \to 1$ in probability. As above, by Lemma 1,

$$\hat{F}(t) = \hat{W}_{\mathrm{false}}(t-\mu) = \Psi(t-\mu) + O_{\mathbb{P}}(\zeta_m)\sqrt{\Psi(t-\mu)(1-\Psi(t-\mu))}, \tag{1.38}$$

and in particular $\hat{F}(\tau) = \Psi(\tau - \mu) + o_{\mathbb{P}}(1)$, so it suffices to show that $\tau - \mu \to -\infty$ in probability.

Since $r > \beta$ and $\beta < 1$, we may take a real number $r_* \in (\beta, r \wedge 1)$. Define $t_* = (\gamma r_* \log n)^{1/\gamma}$. Since $t_* - \mu \to -\infty$, it suffices to show that $\tau \le t_*$ with probability tending to 1. We have

$$G(t_*) = (1-\varepsilon)\Psi(t_*) + \varepsilon\Psi(t_* - \mu). \tag{1.39}$$

The first term is $\sim \Psi(t_*)$, with

$$\Psi(t_*) = n^{-r_* + o(1)}, \tag{1.40}$$

by Definition 1, which says that $\log \Psi(t) \sim -t^\gamma/\gamma$ as $t \to \infty$. The second term is $\sim n^{-\beta}$ by (1.10) and the fact that $\Psi(t_* - \mu) \to 1$ since, again, $t_* - \mu \to -\infty$. Together, we obtain $G(t_*) \sim n^{-\beta}$, using also the fact that $r_* > \beta$. In addition, by (1.35) we have

$$\hat{R}(t_*) = O_{\mathbb{P}}\big(\zeta_n \sqrt{\Psi(t_*)}\big) + o_{\mathbb{P}}(\varepsilon) = o_{\mathbb{P}}(n^{-\beta}), \tag{1.41}$$

since $\zeta_n \sqrt{\Psi(t_*)} = n^{-\frac{1}{2}(r_*+1)+o(1)}$ (any poly-logarithmic factor was absorbed in $n^{o(1)}$), again by (1.40), and $\beta < r_* < 1$. Hence, applying (1.33), we obtain

$$\hat{G}(t_*) = G(t_*) + \hat{R}(t_*) \sim_{\mathbb{P}} G(t_*) \sim n^{-\beta}. \tag{1.42}$$

Together with (1.40), and using by (3.11), we have

$$\hat{G}(t_*)/\Psi(t_*) = n^{(r_* - \beta) + o_{\mathbb{P}}(1)} \gg 1/q. \tag{1.43}$$

This, together with (1.36), implies that $\tau \le t_*$ with probability tending to 1, hence $\mathrm{FNP}(\mathcal{R}_\tau) \to 0$ in probability. By Remark 6, we have $\mathrm{FNR}(\mathcal{R}_\tau) \to 0$ as $n \to \infty$.

1.7.3 Proof of Theorem 3

The proof borrows a number of arguments from Section 1.7.2. We use the same notation and assume as before that the X_i's are distinct. We order the absolute values of statistic $|X|$ in decreasing order, meaning that $|X|_{(1)} \geq \cdots \geq |X|_{(n)}$. Recall that Ψ is now symmetric about 0.

Define the threshold

$$\tau = \inf\left\{t : \widehat{\mathrm{FDP}}_{\mathrm{BC}}(t) \leq q\right\}. \tag{1.44}$$

The difference with τ_{BC} in (1.14) is that the range is not limited to $|\mathbf{X}|$. It can be seen that $\tau = |X|_{(\iota_{\mathrm{BC}}+1)}$ if $\iota_{\mathrm{BC}} < n$ and $\tau = 0$ if $\iota_{\mathrm{BC}} = n$. This, in particular, implies

$$\mathrm{FNP}(\mathcal{R}_\tau) \leq \mathrm{FNP}(\mathcal{R}_{\tau_{\mathrm{BC}}}) \leq \mathrm{FNP}(\mathcal{R}_\tau) + \tfrac{1}{m}. \tag{1.45}$$

Since in our model $m \to \infty$, it suffices to show that $\mathrm{FNP}(\mathcal{R}_\tau) \to 0$ in probability. As before, (1.37) holds true, so it suffices to show that $\hat{F}(\tau) \to 1$ in probability. For that, we saw earlier that it suffices to show that $\tau \leq t_*$ with probability tending to 1.

We have

$$\widehat{\mathrm{FDP}}_{\mathrm{BC}}(t_*) = \frac{1 + n(1 - \hat{G}(-t_*))}{1 \vee n\hat{G}(t_*)}. \tag{1.46}$$

We already saw that $\hat{G}(t_*) \sim n^{-\beta}$, so the denominator above is $\sim n^{1-\beta}$ as $n \to \infty$. For the numerator, by (1.33), we have

$$1 - \hat{G}(-t_*) = 1 - G(-t_*) - \hat{R}(-t_*). \tag{1.47}$$

By (3.15),

$$1 - G(-t_*) = (1-\varepsilon)(1 - \Psi(-t_*)) + \varepsilon(1 - \Psi(-t_* - \mu)) \tag{1.48}$$

$$= (1-\varepsilon)\Psi(t_*) + \varepsilon\Psi(t_* + \mu) \quad \text{[by symmetry of } \Psi\text{]} \tag{1.49}$$

$$\sim \Psi(t_*) = n^{-r_* + o(1)}. \quad \text{[by (1.40)]} \tag{1.50}$$

By (1.35),

$$\hat{R}(-t_*) = O_{\mathbb{P}}\Big(\zeta_n\sqrt{1-\Psi(-t_*)} + \zeta_m\varepsilon\sqrt{1-\Psi(-t_*-\mu)}\Big) \tag{1.51}$$

$$= O_{\mathbb{P}}(\zeta_n\sqrt{\Psi(t_*)} + \zeta_m\varepsilon\sqrt{\Psi(t_*+\mu)}) \quad \text{[by symmetry of } \Psi\text{]} \tag{1.52}$$

$$= O_{\mathbb{P}}(n^{-\frac{1}{2}(r_*+1)+o(1)} + o(n^{-\frac{1}{2}(r_*+\beta+1)+o(1)})) \quad \text{[by (1.40)]} \tag{1.53}$$

$$= O_{\mathbb{P}}(n^{-\frac{1}{2}(r_*+1)+o(1)}). \tag{1.54}$$

(Again, any poly-logarithmic factor was absorbed in $n^{o(1)}$. Combined with the fact that $r_* < 1$, we get $1-\hat{G}(-t_*) \sim n^{-r_*+o(1)}$, and therefore

$$\widehat{\mathrm{FDP}}_{\mathrm{BC}}(t_*) = \frac{n^{1-r_*+o(1)}}{n^{1-\beta}} = n^{\beta-r_*+o(1)} \ll q. \quad \text{[by (3.11) and } \beta < r_\star\text{]} \tag{1.55}$$

Hence, $\widehat{\mathrm{FDP}}_{\mathrm{BC}}(t_*) \le q$ with probability tending to 1, and when this is the case, $\tau \le t_*$, by definition of τ above. This also implies $\mathrm{FNP}(\mathcal{R}_\tau) \to 0$ in probability. By Remark 6, we have $\mathrm{FNR}(\mathcal{R}_\tau) \to 0$ as $n \to \infty$.

1.8 Acknowledgement

Chapter 1, partially, is a version of the paper "Distribution-free Multiple Testing" , Electronic Journal of Statistics, Arias-Castro, Ery; Chen, Shiyun, Volume 11, Number 1 (2017). The dissertation author was the co-author of this paper.

Chapter 2

A Scan Procedure for Multiple Testing

2.1 Abstract

In this chapter, we propose a method that identifies the longest interval with estimated false discovery rate not exceeding the target level and rejects the corresponding null hypotheses. Unlike the Benjamini-Hochberg method, which does the same but over intervals with an endpoint at the origin, the new procedure 'scans' all intervals. In parallel with [STS04], we show that this scan procedure provides strong control of asymptotic false discovery rate. In addition, we investigate its asymptotic false non-discovery rate, deriving conditions under which it outperforms the Benjamini-Hochberg procedure. For example, the scan procedure is superior in power-law location models.

2.2 Introduction

In previous chapter, we defined the false discovery rate (FDR), which serves as a much less conservative type I error criterion than the family-wise error rate (FWER), and the false non-discovery rate (FNR) which is analogous to the type II error rate in multiple testing. And

we introduced the two methods (BH and BC methods) which have been proved to control FDR under some assumptions on the P-values, for example, the assumption that the P-values are independent. Besides to these two procedures, many other FDR-controlling methods have also been proposed and adopted by practitioners faced with large-scale testing problems. However, most of these methods compute a threshold based on the P-values and reject the null hypotheses with the corresponding P-values below that threshold [GW04, Sto02, STS04]. See [Roq11] for a survey.

A threshold approach to multiple testing is natural stemming from the fact that the smaller a P-value is, the more evidence it provides against the null hypothesis being tested. However, we argue that this is not so obvious in the context of multiple testing, particularly in harder cases where the alternatives are not easily identified and in which most of the smallest P-values come from true null hypotheses. This was already understood by [Chi07], who studied how to modify the BH procedure in order to improve the power. He proposed a sophisticated method which applies the BH procedure at multiple locations in the unit interval, with each location playing the role of the origin, resulting in a rejection region made of possibly multiple intervals. The method has more power than the BH method.

In current chapter we propose a simpler approach based on the longest interval whose estimated FDR is below the prescribed level. Compared to [Chi07], the method is simpler and is already shown to outperform the BH method in some settings of potential interest, such as in power-law location models. The method can be seen as a direct extension of the approach of [Sto02]. It thus presents a sort of minimal working example where looking beyond threshold methods can be beneficial.

Scanning over intervals is a common procedure for detecting areas of interest in a point process at least since the work of [Nau65]. In this context, and its extension to discrete signals, the main task has been to test for homogeneity, and some articles have tackled such situations from a multiple testing angle [SZY11, PRBR17, BH07, CdCS06, PGVW07, PPGVW04]. While

these papers aim at controlling the FDR when scanning spatiotemporal data, here we consider a standard multiple testing situation with a priori no spatiotemporal structure, and offer scanning as a way to generalize and potentially improve upon threshold procedures.

The method we propose and discuss here may seem counterintuitive as it is not of threshold-type, however it can happen indeed. For example, when the alternative distribution has a much smaller variance than the null distribution, then the P-values from the true alternatives cannot be too small. If we set a small desired FDR level and apply a threshold procedure like the BH method under this circumstance, we will probably reject the smallest P-values but miss those from the true alternatives, and thus will lead to a loss in the power. Besides, we also find our approach valuable, as it turns out to offer some advantages as compared to threshold-type methods. In fact, we show that the method can improve on the BH method in the context of a power-law location model.

2.2.1 Framework

We consider a setting where we test n null hypotheses, denoted by $\mathbb{H}_1, \ldots, \mathbb{H}_n$. The test for $\mathbb{H}_i$ yields a P-value, denoted as P_i. In this context, a multiple testing procedure $\mathcal{R}$ takes the P-values, $\mathbf{P} = (P_1, \ldots, P_n)$, and returns a subset of indices representing the null hypotheses that the procedure rejects. We still let $\mathcal{H}^0 \subset [n] = \{1, \ldots, n\}$ index the true null hypotheses, and $\mathcal{H}^1 \subset [n] = \{1, \ldots, n\}$ index the false null hypotheses. Table 1.1 describes the outcome when applying some significance rule in such a setting and defines some necessary notations. Let $n_0 = |\mathcal{H}^0|$ and $n_1 = |\mathcal{H}^1|$. Given such a procedure $\mathcal{R}$, the false discovery rate (FDR) is defined the same as (1.2), and the false non-discovery rate (FNR) is defined as (1.3) in Chapter 1.

2.2.2 Threshold procedures

Remember the definition of (1.5) in Chapter 1, threshold procedures are of the form

$$\mathcal{R}(\mathbf{P}) = \big\{ i : P_i \le \tau(\mathbf{P}) \big\}, \tag{2.1}$$

where τ is some (measurable) function with values in $[0,1]$. As we stated earlier, most multiple testing procedures are of this form, including the BH method. Specifically, following [Sto02], we may describe the BH method as follows. For $0 \le t \le 1$, define the following quantities (see Table 1.1),

$$V(t) = \#\{ i \in \mathcal{H}^0 : P_i \le t \}, \quad S(t) = \#\{ i \notin \mathcal{H}^0 : P_i \le t \},$$

and

$$R(t) = V(t) + S(t) = \#\{ i : P_i \le t \},$$

as well as

$$\mathrm{FDR}(t) = \mathbb{E}\left(\frac{V(t)}{R(t) \vee 1} \right).$$

This is the FDR of the procedure with rejection region $[0,t]$. It is estimated by replacing $V(t)$ with nt, justified by the fact that $\mathbb{E}(V(t)) \le n_0 t \le nt$. (The first inequality is an equality when all the null P-values are uniformly distributed in $[0,1]$.) This yields

$$\widehat{\mathrm{FDR}}(t) = \frac{nt}{R(t) \vee 1},$$

and the BH method may be defined via the threshold,

$$\hat{\tau}_\diamond = \max \big\{ t : \widehat{\mathrm{FDR}}(t) \le q \big\},$$

if it is desired to control the FDR at $q \in (0,1)$.

2.2.3 Scan procedures

Effectively, threshold procedures examine intervals of the form $[0,t]$, where $t \in [0,1]$. We extend this family of procedures by considering all possible intervals, thus defining scan procedures as those of the form

$$\mathcal{R}(\mathbf{P}) = \{i : \sigma(\mathbf{P}) \le P_i \le \tau(\mathbf{P})\}, \tag{2.2}$$

where σ and τ are some (measurable) functions with values in $[0,1]$ and such that $\sigma \le \tau$ pointwise. Within this family of procedures, we define a specific procedure in analogy with the definition of the BH method given above.

For $0 \le s \le t \le 1$, define the following quantities (see Table 1.1),

$$V(s,t) = \#\{i \in \mathcal{H}^0 : s \le P_i \le t\}, \quad S(s,t) = \#\{i \notin \mathcal{H}^0 : s \le P_i \le t\},$$

and

$$R(s,t) = V(s,t) + S(s,t) = \#\{i : s \le P_i \le t\},$$

as well as

$$\mathrm{FDR}(s,t) = \mathbb{E}(\mathrm{FDP}(s,t)), \quad \text{where } \mathrm{FDP}(s,t) := \frac{V(s,t)}{R(s,t) \vee 1}.$$

This is the FDR of the procedure with rejection region $[s,t]$, which we estimate by replacing $V(s,t)$ with $n(t-s)$, which bounds its expectation, obtaining

$$\widehat{\mathrm{FDR}}(s,t) = \frac{n(t-s)}{R(s,t) \vee 1},$$

and our scan procedure is defined via the interval

$$(\hat{\sigma}, \hat{\tau}) = \arg\max \left\{ t - s : \widehat{\mathrm{FDR}}(s,t) \le q \right\}, \tag{2.3}$$

assuming, again, that we desire to control the FDR at $q \in (0,1)$. If there are several maximizing intervals, we choose the left-most interval.

Remark 7. By construction, relying on basic properties of the function $\widehat{\mathrm{FDR}}$, we have that $\hat{\sigma}$ and $\hat{\tau}$ correspond to P-values, and

$$\widehat{\mathrm{FDR}}(\hat{\sigma}, \hat{\tau}) \leq q. \tag{2.4}$$

2.2.4 Contribution and contents

In this chapter, following the work of [GW02] and of [Sto02, STS04], we consider an asymptotic setting where the scan procedure just defined is indeed able to control the FDR as desired. In the same framework, we also compare, in terms of FNR, the scan procedure with BH procedure, showing that the former is superior to the latter under some circumstances, including in power-law location models.

The rest of the chapter is organized as follows. In Section 2.3 we consider our scan procedure's ability to control the FDR. This is established in an asymptotic setting. In Section 2.4 we analyze the asymptotic FNR of our scan procedure and compare it with that of the BH procedure. In particular, we derive sufficient conditions under which the scan procedure outperforms the BH procedure. We present the results of numerical experiments in Section 2.5. We briefly discuss our results and possible extensions in Section 2.6. All proofs are gathered in Section 2.7.

2.3 False discovery rate

Large scale multiple testing appears in many areas of applications, where n is typically of the order of tens or hundreds of thousand. This has led to the consideration of an asymptotic setting where n tends to infinity [Sto02, STS04, GW02]. In detail, the asymptotic framework we consider requires the almost sure pointwise convergence of the empirical distribution of the null

P-values and of the empirical distribution of the non-null P-values, or in formula,

$$\lim_{n\to\infty} \frac{V(s,t)}{n_0} = t - s \quad \text{and} \quad \lim_{n\to\infty} \frac{S(s,t)}{n_1} = G(t) - G(s), \tag{2.5}$$

almost surely for any fixed $0 \le s < t \le 1$, where G is a continuous distribution function on the real line. We assume in addition that the following limit exists,

$$\pi_0 := \lim_{n\to\infty} \frac{n_0}{n} \in (0,1), \quad \pi_1 := 1 - \pi_0. \tag{2.6}$$

For the remaining results, we assume that Conditions (2.5)-(2.6) hold. Note that these conditions were also assumed in [STS04].

Remark 8. This asymptotic framework generalizes the Bayesian model where the null hypotheses are true with probability π_0 and not true with probability π_1, and the null P-values are uniform in $[0,1]$ and the non-null P-values are G-distributed, corresponding to a mixture model where the P-values are iid with distribution function $\pi_0 t + \pi_1 G(t)$.

Define

$$\overline{\mathrm{FDR}}^{\infty}(s,t) = \frac{t-s}{\pi_0(t-s) + \pi_1(G(t) - G(s))},$$

which is the pointwise (almost sure) limit of $\widehat{\mathrm{FDR}}(s,t)$ under the above assumptions. Our next result shows that the scan procedure controls the FDR asymptotically. Here, and everywhere else in the chapter, q will denote the level at which the FDR is to be controlled. We make the dependency of $(\hat{\sigma}, \hat{\tau})$ on n explicit, but note that other quantities, such as $\widehat{\mathrm{FDR}}, \mathrm{FDP}, \mathrm{FDR}$, also depend on n.

Theorem 4. *We have*

$$\limsup_{n\to\infty} \mathrm{FDP}(\hat{\sigma}_n, \hat{\tau}_n) \le q, \quad \textit{almost surely}, \quad \textit{and} \quad \limsup_{n\to\infty} \mathbb{E}\left[\mathrm{FDP}(\hat{\sigma}_n, \hat{\tau}_n)\right] \le q.$$

Remark 9. In our notation, $\text{FDP}(\hat{\sigma}_n, \hat{\tau}_n)$ is random and different from $\mathbb{E}\big[\text{FDP}(\hat{\sigma}_n, \hat{\tau}_n)\big]$. The latter is the FDR of the scan procedure with rejection region $[\hat{\sigma}_n, \hat{\tau}_n]$.

We consider the maximization in (2.3), but based on $\overline{\text{FDR}}^{\infty}$. Indeed, let $\mathcal{A}$ be the set of maximizers and δ be the value of the following optimization problem

$$\max\big\{t - s : \overline{\text{FDR}}^{\infty}(s,t) \le q\big\},$$

or, equivalently,

$$\max\big\{t - s : G(t) - G(s) = \beta(t-s)\big\},$$

where $\beta := \frac{1}{\pi_1}(\frac{1}{q} - \pi_0)$.

$$\begin{aligned}&\text{We assume that } \delta > 0 \text{ and that there is } (s,t) \in \mathcal{A} \text{ such that } u \mapsto \overline{\text{FDR}}^{\infty}(u,t) \text{ is}\\ &\text{strictly decreasing at } u = s \text{ or that } u \mapsto \overline{\text{FDR}}^{\infty}(s,u) \text{ is strictly increasing at } u = t.\end{aligned} \tag{2.7}$$

The strict monotonicity condition is true, for example, if G is concave on $[0,1]$, or more generally if G is differentiable as satisfies $G'(s) \vee G'(t) < (G(t) - G(s))/(t-s)$ at some $(s,t) \in \mathcal{A}$.

Theorem 5. *If* (2.7) *holds, then, almost surely, any accumulation point of* $(\hat{\sigma}_n, \hat{\tau}_n)$ *belongs to* $\mathcal{A}$.

Remark 10. This result is analogous to Theorem 1 in [GW02], which establishes a similar limit for the BH method under similar conditions. Specifically, they show that, almost surely, $\hat{\tau}_{\diamond,n}$ converges to

$$\delta_{\diamond} := \max\big\{t : \overline{\text{FDR}}^{\infty}(t) \le q\big\}, \tag{2.8}$$

with $\overline{\text{FDR}}^{\infty}(0,t)$. Alternatively, $\delta_{\diamond}$ may also be defined as the right-most solution to the equation $G(t) = \beta t$.

2.4 False non-discovery rate

Having established that the scan procedure asymptotically controls the FDR at the desired level, we now turn to examining its false non-discovery rate (FNR). We do so under the same asymptotic framework.

Theorem 6. *If* (2.7) *holds, then*

$$\lim_{n\to\infty} \mathbb{E}\left[\mathrm{FNP}(\hat{\sigma}_n, \hat{\tau}_n)\right] = 1 - \beta\delta.$$

As could be anticipated from Theorem 5, the limiting value is the asymptotic FNR of any deterministic rule given by an interval $[s,t]$ with $(s,t) \in \mathcal{A}$.

Remark 11. This result is analogous to Theorem 3 in [GW02], which establishes a similar limit for the BH method, specifically,

$$\lim_{n\to\infty} \mathbb{E}\left[\mathrm{FNP}(\hat{\tau}_{\diamond,n})\right] = 1 - \beta\delta_\diamond.$$

We now turn our attention to comparing the scan method and the BH method. The following theorem provides some sufficient conditions under which the scan procedure outperforms the BH procedure. It happens exactly when $\hat{\sigma}_n$ is bounded away from zero.

Theorem 7. *Assume that G is differentiable. If* (2.7) *holds, and in addition*

$$G'(0) < G'(\delta_\diamond), \tag{2.9}$$

then the scan procedure has strictly smaller asymptotic FNR than the BH procedure.

The BH procedure is known to be optimal in various ways under generalized Gaussian location models [ACC17, RRJW17]. It turns out that under such models, the distribution G is strictly concave, implying that the scan method will coincide with the BH method asymptotically.

(This is illustrated in the numerical experiment of Figure 2.1.) We therefore consider power-law location models, where the statistics have heavy tail. More specifically, we consider a mixture model where

$$X_1, \dots, X_n \overset{\text{iid}}{\sim} \pi_0 \Psi(x) + \pi_1 \Psi(x - \mu), \tag{2.10}$$

where Ψ is a continuous distribution on the real line and $\mu > 0$. These are meant to represent the test statistics, whose large values weigh against their respective null hypotheses. In particular, Ψ is the null distribution and μ is the effect size. The P-values are then computed as usual, meaning $P_i = \bar{\Psi}(X_i)$ where $\bar{\Psi} := 1 - \Psi$, and are seen to follow a mixture model

$$P_1, \dots, P_n \overset{\text{iid}}{\sim} \pi_0 t + \pi_1 G(t), \quad \text{where } G(t) := \bar{\Psi}(\bar{\Psi}^{-1}(t) - \mu). \tag{2.11}$$

Theorem 8. *Consider a mixture model* (2.10) *in the asymptotic defined by* (2.6)*. Then the condition* (2.5) *holds. Assume in addition that* Ψ *has a density* ψ *which can be taken to be strictly positive everywhere and such that* $\psi(x) \to 0$ *as* $x \to \infty$ *and* $\psi(x) \sim x^{-\gamma-1}(\log x)^c$ *as* $x \to \infty$ *for some* $\gamma > 0$ *and some* $c \in \mathbb{R}$*. Then there is* $\mu_0 > 0$ *(depending on* Ψ *and* β*) such that* (2.9) *holds for all* $\mu > \mu_0$*.*

Remark 12. The result does not say anything about (2.7), which is also required in Theorem 6, but this condition is fulfilled in most cases of interest.

2.5 Numerical experiments

In this section, we perform simple simulations to see the performance of the BH and scan procedures on finite data. We consider the normal and Cauchy mixture models, as in (2.10).

In the set of experiments, the sample size $n \in \{2, \dots, 8\} \times 10^3$. We draw $m = n(1 - \varepsilon)$ observations from the alternative distribution $\Psi(\cdot - \mu)$, and the other $n - m$ from the null distribution Ψ. Each situation is repeated 100 times and we report the average FDP and FNP for each

procedure together with error bars. The FDR control level was set at $q = 0.10$.

2.5.1 Normal model

In this model Ψ is the standard normal distribution. We set $\varepsilon = 0.05$ and $\mu = 4$. The choice of the μ is to put the signals in a detectable (but not easy) region. This is because under the global null hypothesis where $Z_i \sim \mathcal{N}(0,1)$ for all $i = 1,\dots,n$, we have $\max_{i\in[n]} Z_i \sim \sqrt{2\log n}$ with high probability. Therefore we choose the alternative mean of this order, that is, for $n \in \{2,\dots,8\}\times 10^3$, $\sqrt{2\log n} \approx 4$. See Figure 2.1, where we have plotted G, the P-value distribution under the alternative defined in (2.11). This is a situation where G is concave, so we expect the two methods to behave similarly. This is confirmed numerically. In fact, the scan procedure was observed, in these experiments, to coincide with the BH method. (This does not happen at smaller signal-to-noise ratios, e.g., when μ is smaller.) See Figure 2.2, where we have plotted the FDP and FNP of both procedures.

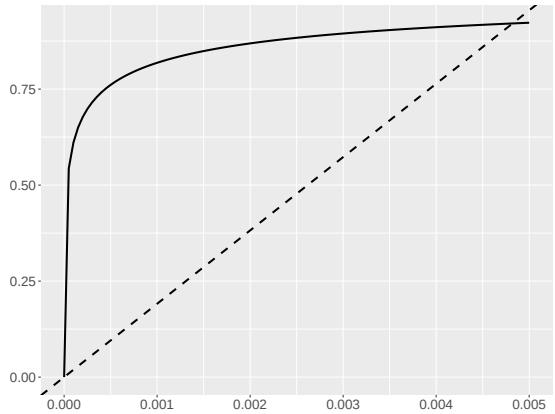

Figure 2.1: The alternative P-value distribution G in the normal mixture model with $\varepsilon = 0.05$ and $\mu = 4$ (solid black) and the line $y = \beta x$ (dashed black).

2.5.2 Cauchy model

In this model Ψ is the Cauchy distribution, which satisfies the assumption in Theorem 8 with $\gamma = 1, c = 0$. We set $\varepsilon = 0.10$ and $\mu = 37$. Again, we chose this value for μ in order to make the

Figure 2.2: FDP and FNP for the BH (red) and scan (blue) methods under normal mixture model. The methods are essentially identical. The FDR control was set at $q = 0.10$.

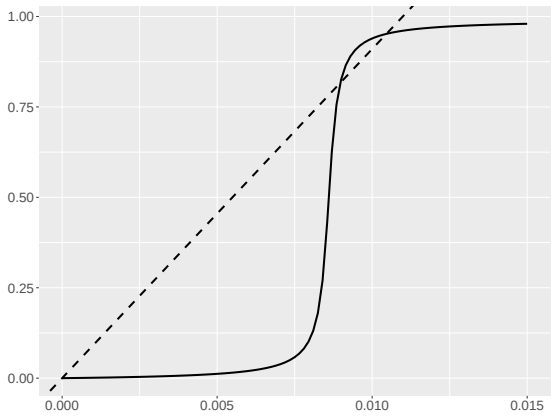

Figure 2.3: The alternative P-value distribution G in the Cauchy mixture model with $\varepsilon = 0.10$ and $\mu = 37$ (solid black) and the line $y = \beta x$ (dashed black).

problem neither too easy nor too hard, and in addition, this choice of parameters leads to a model that satisfies the condition (2.9). See Figure 2.3 for an illustration. Based on Theorem 7 we thus expect the scan procedure to outperform the BH procedure. This is confirmed in the numerical experiments. See Figure 2.4, where we have plotted the FDP and FNP of both procedures.

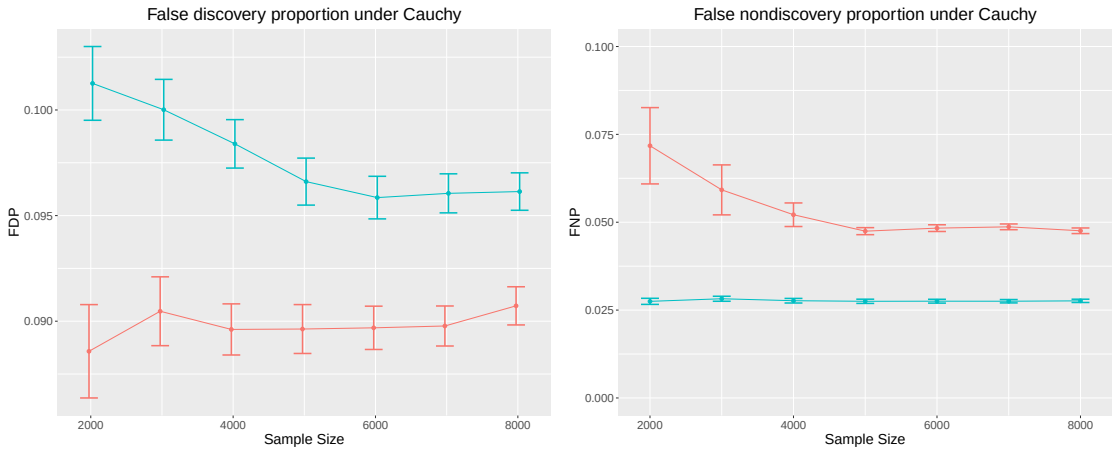

Figure 2.4: FDP and FNP for the BH (red) and scan (blue) methods under Cauchy mixture model. The FDR control was set at $q = 0.10$.

2.6 Discussion

Optimality. [GW02] argue that the BH method is not optimal among threshold procedures due to its being conservative in terms of FDR control. We expect the same to be true of our scan procedure. We could have pursued an improvement analogous to how the BH method was ameliorated in [Sto02, BH00] based on estimating the number of true null hypotheses (n_0 in our notation), but we chose not to do so for the sake of simplicity and focus.

Testing the global null hypothesis. In parallel work, we uncover a similar phenomenon in the context of testing the global null hypothesis. Indeed, continuing with the line of work coming out of [Ing97, DJ04], in [ACY19] we consider the problem of testing $\pi_1 = 0$ (same as testing $\mu = 0$) in mixture models like (2.10) and show that threshold tests can be strictly inferior to scan tests in power-law models. We did this in a different asymptotic regime where π_1 is small under the alternative.

2.7 Proofs

Henceforth, we assume that (2.5) and (2.6). Before proving our main results, we establish a few auxiliary lemmas.

Lemma 2. *For any fixed $d > 0$, almost surely,*

$$\lim_{n\to\infty} \sup_{t-s\geq d} \left| \widehat{\mathrm{FDR}}(s,t) - \overline{\mathrm{FDR}}^{\infty}(s,t) \right| = 0. \tag{2.12}$$

Proof. As is well-known, the pointwise convergences that we assume, namely (2.5), imply uniform convergences, so that, together with (2.6), we have

$$\lim_{n\to\infty} \sup_{s\leq t} \left| \frac{V(s,t)}{n} - \pi_0(t-s) \right| = 0, \tag{2.13}$$

$$\lim_{n\to\infty} \sup_{s\leq t} \left| \frac{S(s,t)}{n} - \pi_1(G(t)-G(s)) \right| = 0, \tag{2.14}$$

almost surely. Combining these also yields

$$\lim_{n\to\infty} \sup_{s\leq t} \left| \frac{R(s,t)}{n} - \left\{ \pi_0(t-s) + \pi_1(G(t)-G(s)) \right\} \right| = 0, \tag{2.15}$$

When $t - s \geq d$, we have

$$\pi_0(t-s) + \pi_1(G(t)-G(s)) \geq \pi_0 d > 0,$$

and it is thus straightforward to show that

$$\lim_{n\to\infty} \sup_{t-s\geq d} \left| \frac{n(t-s)}{R(s,t)\vee 1} - \frac{t-s}{\pi_0(t-s)+\pi_1(G(t)-G(s)} \right| = 0,$$

almost surely, which establishes our claim. □

Lemma 3. *Almost surely,*

$$\liminf_{n\to\infty} \inf_{s\le t} \left\{ \widehat{\mathrm{FDR}}(s,t) - \mathrm{FDP}(s,t) \right\} \ge 0.$$

Proof. For the first part, we have

$$\widehat{\mathrm{FDR}}(s,t) - \mathrm{FDP}(s,t) = \frac{n(t-s) - V(s,t)}{R(s,t) \vee 1} \ge 0 \;\Leftrightarrow\; t - s - \frac{V(s,t)}{n} \ge 0,$$

so that we only need to prove that

$$\liminf_{n\to\infty} \inf_{s\le t} \left\{ t - s - V(s,t)/n \right\} \ge 0.$$

But this simply comes from (2.13) and (2.6). □

Lemma 4. *If* (2.7) *holds, then, almost surely,*

$$\liminf_{n\to\infty} \left\{ \hat{\tau}_n - \hat{\sigma}_n \right\} \ge \delta. \tag{2.16}$$

Proof. Let (s,t) be as in (2.7), with (for example) $u \mapsto \overline{\mathrm{FDR}}^{\infty}(u,t)$ strictly decreasing at $u = s$. Then there is $\varepsilon > 0$ such that $\overline{\mathrm{FDR}}^{\infty}(u,t) < \overline{\mathrm{FDR}}^{\infty}(s,t)$ when $s < u \le s + \varepsilon$. With probability one, $\widehat{\mathrm{FDR}}(u,t)$ converges to $\overline{\mathrm{FDR}}^{\infty}(u,t)$, and when this is the case, $\widehat{\mathrm{FDR}}(u,t) \le q$ for n sufficiently large, then implying that $t - u \le \hat{\tau}_n - \hat{\sigma}_n$ by definition in (2.3). Hence, we have shown that for any such u, $\liminf_{n\to\infty} \left\{ \hat{\tau}_n - \hat{\sigma}_n \right\} \ge t - u$ almost surely, and we conclude by letting $u \searrow s$. (Recall that $\delta = t - s$ for any $(s,t) \in \mathcal{A}$.) □

2.7.1 Proof of Theorem 4

For the first part, using Lemma 3, we have

$$\liminf_{n\to\infty}\left[\widehat{\mathrm{FDR}}(\hat{\sigma}_n,\hat{\tau}_n)-\mathrm{FDP}(\hat{\sigma}_n,\hat{\tau}_n)\right]\geq 0,$$

almost surely, and we conclude with (2.4).

The second part just follows from the first part and Fatou's lemma.

2.7.2 Proof of Theorem 5

With probability one, a realization satisfies (2.12) with $d=\delta/2$, and (2.16). Consider such a realization and let (s^*,t^*) be an accumulation point of $(\hat{\sigma}_n,\hat{\tau}_n)$.

Because (2.16) holds, we have $t^*-s^*\geq\delta$.

We also have $\hat{\tau}_n-\hat{\sigma}_n\geq d$, eventually, and because (2.12) holds, this implies that

$$\lim_{n\in\mathcal{N}}\widehat{\mathrm{FDR}}(\hat{\sigma}_n,\hat{\tau}_n)-\overline{\mathrm{FDR}}^{\infty}(\hat{\sigma}_n,\hat{\tau}_n)=0.$$

Together with (2.4), we thus have

$$\limsup_{n\to\infty}\overline{\mathrm{FDR}}^{\infty}(\hat{\sigma}_n,\hat{\tau}_n)\leq q.$$

By continuity of $\overline{\mathrm{FDR}}^{\infty}$, this implies that $\overline{\mathrm{FDR}}^{\infty}(s^*,t^*)\leq q$, in turn implying that $t^*-s^*\leq\delta$.

We have thus established that (s^*,t^*) satisfies $t^*-s^*=\delta$ and $\overline{\mathrm{FDR}}^{\infty}(s^*,t^*)\leq q$, and therefore (s^*,t^*) belongs to $\mathcal{A}$ by definition.

2.7.3 Proof of Theorem 6

By definition of S in Table 1.1, we have

$$\text{FNP}(\hat{\sigma}_n, \hat{\tau}_n) = 1 - S(\hat{\sigma}_n, \hat{\tau}_n)/n_1.$$

By (2.14) and Theorem 5 together with the fact that $G(t) - G(s) = \beta\delta$ for any $(s,t) \in \mathcal{A}$, almost surely,

$$S(\hat{\sigma}_n, \hat{\tau}_n)/n_1 \to \beta\delta.$$

We thus have, almost surely,

$$\text{FNP}(\hat{\sigma}_n, \hat{\tau}_n) \to 1 - \beta\delta,$$

and we conclude using the Dominated Convergence theorem.

2.7.4 Proof of Theorem 7

Adapting the proof of Theorem 6, we can establish an analogous result for the BH method, specifically,

$$\mathbb{E}[\text{FNP}(\hat{\tau}_{\diamond,n})] \to 1 - \beta\delta_\diamond, \tag{2.17}$$

almost surely, where $\delta_\diamond$ was defined in (2.8). Therefore, to compare the asymptotic FNR of the scan and the BH procedures, we need to compare δ and $\delta_\diamond$.

Define $A_0 = \arctan(G'(0))$, $A_\diamond = \arctan(G'(\delta_\diamond))$, and $B = \arctan(\beta)$. Apparently, we have $0 \le A_0, A_\diamond, B \le \frac{\pi}{2}$. By the fact that $\delta_\diamond$ is the right-most solution to $G(t) = \beta t$, we have $G'(\delta_\diamond) \le \beta$. Hence, $\phi := B - A_\diamond \ge 0$. This, coupled with (2.9), implies that $G'(0) < \beta$, so that $\theta := B - A_0 > 0$.

Let $\mathcal{L}_0$ denote the line with slope β passing through the origin, and for $d \ge 0$, let $\mathcal{L}_d$ denote the line parallel to $\mathcal{L}_0$ at a distance d below $\mathcal{L}_0$. See Figure 2.5 for an illustration. Because $G'(0) < \beta$ and $G(\delta_\diamond) = \beta\delta_\diamond$, and by continuity of G, the graph of G intersects $\mathcal{L}_0$ at least twice.

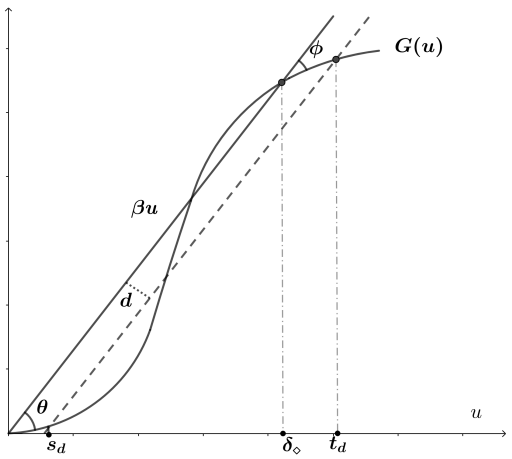

Figure 2.5: Example which satisfies Condition 2.9 in Theorem 7.

Choosing d small enough, it is therefore also the case that G intersects $\mathcal{L}_d$ at least twice. Let s_d and t_d denote the horizontal coordinates of the leftmost and rightmost intersection points, respectively. Note that $s_d \to 0$ as $d \to 0$ by that fact that $G'(0) < \beta$, and $t_d \to \delta_\diamond$ by the fact that $\delta_\diamond$ is the right-most solution to $G(t) = \beta t$. Moreover, as $d \to 0$, by simple geometry arguments, we have

$$s_d \sim \sin(A_0) \cdot \frac{d}{\sin(\theta)}, \quad t_d - \delta_\diamond \sim \sin(A_\diamond) \cdot \frac{d}{\sin(\phi)}.$$

Since $G'(0) < G'(\delta_\diamond)$, we have $\theta > \phi \geq 0$ and also $A_0 < A_\diamond$. It follows that, for d small enough, $\delta_\diamond < t_d - s_d$. Due to the fact, by construction,

$$\frac{G(t_d) - G(s_d)}{t_d - s_d} = \beta,$$

we have $t_d - s_d \leq \delta$, by definition of the latter.

2.7.5 Proof of Theorem 8

The Law of Large Numbers implies that the condition (2.5) holds. We thus turn to the remaining of the statement.

We note that $G(t) = \bar{\Psi}(\bar{\Psi}^{-1}(t) - \mu)$ is differentiable on $(0,1)$, with derivative

$$G'(t) = \frac{\psi(\bar{\Psi}^{-1}(t) - \mu)}{\psi(\bar{\Psi}^{-1}(t))}.$$

As $t \to 0$, we have $\bar{\Psi}^{-1}(t) \to \infty$, and by the fact that for all $a \in \mathbb{R}$, $\psi(x-a) \sim \psi(x)$ as $x \to \infty$, we have that G' is differentiable at 0, with derivative $G'(0) = 1$.

We also have that, for any fixed t, $G'(t) \to 0$ as $\mu \to \infty$, due to the fact that $\psi(x) \to 0$ as $x \to -\infty$. Hence, $\delta_\diamond \to 0$ as $\mu \to \infty$. Let $x_\diamond = \bar{\Psi}^{-1}(\delta_\diamond)$, so that $x_\diamond \to \infty$ as $\mu \to \infty$. Because $G(\delta_\diamond) = \beta\delta_\diamond$, we have $\bar{\Psi}(x_\diamond - \mu) = \beta\bar{\Psi}(x_\diamond)$. Because the right-hand side tends to 0, we must have $x_\diamond - \mu \to \infty$ as $\mu \to \infty$. Then using the fact that $\bar{\Psi}(x) \sim \frac{1}{\gamma}x^{-\gamma}(\log x)^c$ as $x \to \infty$, we must have

$$\frac{1}{\gamma}(x_\diamond - \mu)^{-\gamma}(\log(x_\diamond - \mu))^c \sim \beta\frac{1}{\gamma}x_\diamond^{-\gamma}(\log x_\diamond)^c,$$

or equivalently,

$$(1 - \mu/x_\diamond)^{-\gamma}\left(\frac{\log(x_\diamond - \mu)}{\log x_\diamond}\right)^c \to \beta,$$

as $\mu \to \infty$. This is seen to imply that $x_\diamond \sim a\mu$, where $a := (1 - \beta^{-1/\gamma})^{-1}$. Note that $a > 1$. We then have, as $\mu \to \infty$,

$$G'(\delta_\diamond) = \frac{\psi(x_\diamond - \mu)}{\psi(x_\diamond)} \sim \frac{(x_\diamond - \mu)^{-\gamma-1}(\log(x_\diamond - \mu))^c}{x_\diamond^{-\gamma-1}(\log x_\diamond)^c} \to \left(\frac{a}{a-1}\right)^{\gamma+1} > 1.$$

We conclude that, for μ large enough, $G'(\delta_\diamond) > 1 = G'(0)$.

2.8 Acknowledgement

Chapter 2, partially, is a version of the paper "A Scan Procedure for Multiple Testing", Chen, Shiyun; Ying Andrew; Arias-Castro, Ery. The manuscript has been submitted for publication in a major statistical journal. The dissertation author was the primary investigator and author

of this material.

Chapter 3

Online (Sequential) Multiple Testing

3.1 Abstract

We study an online multiple testing problem where the hypotheses arrive sequentially in a stream. The test statistics are independent and assumed to have the same distribution under their respective null hypotheses. We investigate two procedures LORD and LOND, proposed by Javanmard and Montanari [JM15], which are proved to control the FDR in an online manner. In some (static) model, we show that LORD is optimal in some asymptotic sense, in particular as powerful as the (static) Benjamini-Hochberg procedure to first asymptotic order. We also quantify the performance of LOND. Some numerical experiments complement our theory.

3.2 Introduction

In previous chapters, we mainly focus on the *offline* multiple testing problem, where we have an entire batch of hypotheses and the corresponding P-values, and analyzed several standard procedures (like the BH procedure) which can control the FDR below a pre-assigned level.

However, the fact that offline FDR-controlling techniques require aggregating the P-values

from all the tests and processing them jointly, makes it impossible to utilize them for a number of applications which are best modeled as an *online hypothesis testing* problem [FS08]. For example, in marketing research a sequence of A/B tests can be carried out in an online fashion, or in a pharmaceutical drug test a sequence of clinical trials are conducted over time. In such scenario, we consider that infinitely many hypotheses $\mathcal{H} = (\mathbb{H}_1, \mathbb{H}_2, \mathbb{H}_3, \dots)$ arrive sequentially in a stream with corresponding P-values $P_1, P_2, P_3, \dots$, and we are required to decide whether we accept or reject $\mathbb{H}_i$ only based on $P_1, \dots, P_i$. We propose to use the recent sequential multiple testing procedures of [JM15] which control the FDR in an online manner, and study the asymptotic optimality properties of these methods in the context of sparse mixture asymptotically generalized Gaussian model (see Definition 1) which the normal model often used as benchmark in various works on multiple testing.

3.2.1 The risk of an online multiple testing procedure

Similar to the offline multiple testing procedures, we would like to define the risk of online testing procedures. Consider a setting where we want to test an ordered infinite sequence of null hypotheses, denoted $\mathcal{H} = (\mathbb{H}_1, \mathbb{H}_2, \mathbb{H}_3, \dots,)$, where at each step i we have to decide whether to reject $\mathbb{H}_i$ having access to only previous decisions. The test that we use for $\mathbb{H}_i$ rejects for large positive values of a statistic X_i. Throughout, we assume that test statistics are all independent. Denote the collection of the first n hypotheses in the stream by $\mathcal{H}(n) = (\mathbb{H}_1, \dots, \mathbb{H}_n)$, and the vector of first n test statistics by $\mathbf{X}(n) = (X_1, \dots, X_n)$. Let Φ_i denote the survival function[1] of X_i and $\Phi(n) = (\Phi_1, \dots, \Phi_n)$. We assume that the corresponding P-values can be computed. The simplest such case is when $\mathbb{H}_i$ is a singleton, $\mathbb{H}_i = \{\Phi_i^{\text{null}}\}$, and the null distributions $\Phi_1^{\text{null}}, \Phi_2^{\text{null}}, \dots,$ are known. In that case, the i-th P-value is defined as $P_i = \Phi_i^{\text{null}}(X_i)$, which is the probability of exceeding the observed value of the statistic under its null distribution.

Let $\mathcal{H}^0(n) \subset [n] = \{1, \dots, n\}$ index the true null hypotheses in the first n hypotheses, and

[1] In this chapter, the survival function of a random variable Y is defined as $y \mapsto \mathbb{P}(Y \geq y)$.

Table 3.1: This table summarizes the outcome of applying an online multiple testing procedure $\mathcal{R}$ to first n null hypotheses.

	accept null	reject null	total
null true	$U_{\mathcal{R}}(n)$	$V_{\mathcal{R}}(n)$	$\lvert\mathcal{H}^0(n)\rvert$
null false	$T_{\mathcal{R}}(n)$	$S_{\mathcal{R}}(n)$	$\lvert\mathcal{H}^1(n)\rvert$
total	$W_{\mathcal{R}}(n)$	$R_{\mathcal{R}}(n)$	n

$\mathcal{H}^1(n) \subset [n] := \{1,\dots,n\}$ index the false null hypotheses in the first n hypotheses, meaning

$$\mathcal{H}^0(n) = \{i \in [n] : \Phi_i \in \mathbb{H}_i\}, \quad \mathcal{H}^1(n) = \{i \in [n] : \Phi_i \notin \mathbb{H}_i\}. \tag{3.1}$$

A multiple testing procedure $\mathcal{R}$, for each $n \geq 1$, takes in $\mathbf{X}(n)$ and returns a subset of $\{1,\dots,n\}$ indicating the null hypotheses that the procedure rejects among the first n in the sequence. Table 3.1 describes the outcome when applying some online testing rule to the first n hypotheses in the sequence and defines some necessary notations.

Using the similar definition in the offline setting [BH95], for a given procedure $\mathcal{R}$ and for each $n \in \mathbb{N}$, the online false discovery rate is defined as the expected value of the false discovery proportion

$$\mathrm{FDR}_{\mathcal{R}}(n) = \mathbb{E}_{\Phi}\big[\mathrm{FDP}_{\mathcal{R}}(n)\big], \quad \mathrm{FDP}_{\mathcal{R}}(n) = \frac{V_{\mathcal{R}}(n)}{R_{\mathcal{R}}(n) \vee 1}. \tag{3.2}$$

Similarly, the online false non-discovery rate (FNR) is defined as the expected value of the false non-discovery proportion[2]

$$\mathrm{FNR}_{\mathcal{R}}(n) = \mathbb{E}_{\Phi}\big[\mathrm{FNP}_{\mathcal{R}}(n)\big], \quad \mathrm{FNP}_{\mathcal{R}}(n) = \frac{T_{\mathcal{R}}(n)}{\lvert\mathcal{H}^1(n)\rvert \vee 1}, \tag{3.3}$$

where we denoted the cardinality of a set $\mathcal{A} \subset [n]$ by $\lvert\mathcal{A}\rvert$. In analogy with the risk of a test — which is defined as the sum of the probabilities of type I and type II error — we define the risk of an online multiple testing procedure $\mathcal{R}$ as the sum of the false discovery rate and the false

[2]This definition is different from that of [GW02].

non-discovery rate

$$\text{risk}_{\mathcal{R}}(n) = \text{FDR}_{\mathcal{R}}(n) + \text{FNR}_{\mathcal{R}}(n). \tag{3.4}$$

3.2.2 More related work

The literature on multiple testing is by now vast. Only more recently, though, have multiple testing procedures been proposed for the sequential setting. In the context of testing $J > 2$ (fixed) null hypotheses about J sequences of data streams of arbitrary size, [Bar14] proposes general stepup and stepdown procedures which provide control of the simultaneous generalized type I and II error rates. See also [BS14] for procedures controlling the type I and II FWER's, and [BS13] for procedures controlling the FDR and FNR (defined differently).

Another situation also considered in literature is where the hypotheses are ordered based on prior information on how promising each hypothesis is. In this context, [GWCT16] develops two rules (FowardStop and StrongStop) to choose the number of hypotheses to reject which are shown to control the FDR. A variation of StrongStop rule can also be applied in sequential model selection in regression model. [FBC15] proposes the Sequential stepup procedure (SeqStep) which also guarantees the FDR control under independence. [LB16a] develops a broader class of ordered hypotheses testing procedures under such setting, called *accumulation tests*, which generalize the existing two methods (FowardStop and SeqStep). [LF16] derives an improved version of Selective SeqStep, called Adaptive SeqStep. See [FST14, FTTT15, LTTT14] for more methods and applications in selective sequential model selection.

Still in the sequential setting, [FS08] develops an alpha-investing procedure which provides uniform control of mFDR (a weaker control than FDR control) in online testing under some condition. The alpha-investing rule spends some of the wealth to perform each test and earns more wealth each time a discovery is made. [AR14] provides a broader class of online procedures called generalized alpha-investing and also establish mFDR control. [JM15] proposes two procedures called LOND and LORD algorithms which control both FDR and mFDR in

online testing. We refer to Section 3.5.1 and 3.5.2 for more details of rules and discuss their asymptotic risk in our context. More generally, [JM18] studies generalized alpha-investing rules and obtains conditions for FDR control under a general dependence structure of test statistics. They also develop modified procedures for online control of the false discovery exceedance.

In the present chapter we study some asymptotic power properties of the LORD and LOND methods, complementing the results of [JM15]. This chapter is a continuation of our previous work in the static[3] setting [ACC17] in Chapter 1, where an asymptotic oracle risk bound for multiple testing is obtained, and both the method of [BH95] and the distribution-free method of [FBC15] are proved to achieve that bound. Various other oracle bounds and corresponding optimality results for multiple testing procedures are available in the literature; see, for example, [GW02, SC07, Sto07, BCFG11, NR12, MMB11, JJ12, JK14, BST15].

3.2.3 Content

The rest of the chapter is organized as follows. In Section 3.4.1 we consider the normal location model and derive the performance of LORD under this model. Generalizing this model, in Section 3.4.2 we consider a nonparametric Asymptotic Generalized Gaussian model. We analyze the asymptotic performance of the LORD and LOND procedures of [JM15] under this model in Section 3.5.1 and Section 3.5.2. We present some numerical experiments in Section 3.6. All proofs are gathered in Section 3.7.

3.3 Methods

We describe the LORD and LOND procedures of [JM15], which are the methods we study in this chapter. Recall that $\mathbb{H}_1, \mathbb{H}_2, \ldots$ are tested sequentially and that P_i denotes the P-value corresponding to the test of $\mathbb{H}_i$. These two procedures, and most others, work as follows: set a

[3] By 'static' we mean a setting where all the null hypotheses of interest are considered together. This is the more common setting considered in the multiple testing literature.

significance level α_i based on $P_1, \ldots, P_{i-1}$ (except for α_1 which is set beforehand) and reject $\mathbb{H}_i$ if $P_i \leq \alpha_i$. The LORD and LOND methods vary in how they set these thresholds, although they both start with a sequence of the form

$$\lambda_i \geq 0 \text{ such that } \sum_{i=1}^{\infty} \lambda_i = q, \tag{3.5}$$

where q denotes the desired FDR control level. In what follows, we stay close to the notation used in [JM15].

3.3.1 The LORD method

Based on a chosen sequence (3.5), the LORD algorithm — which stands for (significance) Levels based On Recent Discovery — sets the sequential significance levels $(\alpha_i)_{i=1}^{\infty}$ as follows:

$$\alpha_i = \lambda_{i-t_i}, \quad t_i = \max\{l < i : \mathbb{H}_l \text{ is rejected}\}, \tag{3.6}$$

with $t_i := 0$ for all i before the time of first discovery.

In [JM18] the LORD algorithm is shown to control the FDR at a level less than or equal to q in an online fashion, specifically,

$$\sup_{n \geq 1} \mathrm{FDR}_{\mathcal{R}}(n) \leq q, \tag{3.7}$$

if the P-values are independent. More generally, [JM18] study a class of monotone generalized alpha-investing procedures (which includes LORD as a special case) and prove that any rule in this class controls the cumulative FDR at each stage provided the P-values corresponding to true nulls are independent from the other P-values.

3.3.2 The LOND method

Based on a chosen sequence (3.5), the LOND algorithm — which stands for (significance) Levels based On Number of Discovery — sets the sequential significance levels $(\alpha_i)_{i=1}^{\infty}$ as follows:

$$\alpha_i = \lambda_i(D(i-1)+1). \tag{3.8}$$

where $D(n)$ denotes the number of discoveries in $\mathcal{H}(n) = (\mathbb{H}_1, \ldots, \mathbb{H}_n)$, with $D(0) := 0$.

In [JM15] the LOND is shown to control the FDR at level less than or equal to q everywhere in an online manner, the same as (3.7), if the P-values are independent.

3.4 Models

In this chapter we study the FNR of each of the LORD and LOND methods of [JM15] on the first n hypotheses as $n \to \infty$. As benchmark, we use the oracle that we considered previously in [ACC17] for the static setting defined by these n hypothesis testing problems. For the reader not familiar with that setting, at least in the models that we consider, this turns out to be asymptotically equivalent to applying the Benjamini-Hochberg (BH) method to the first n hypotheses. Note that the latter accesses all the first n hypotheses at once and is thus not constrained to be sequential in nature.

The static setting we consider is that of a location mixture model. We assume that we know the null distribution function Φ, assumed to be continuous for simplicity. We then assume that the test statistics are independent with respective distribution $X_i \sim \Phi_i = \Phi(\cdot - \mu_i)$, where $\mu_i = 0$ under the null $\mathbb{H}_i$ and $\mu_i > 0$ otherwise. Both minimax and Bayesian considerations lead one to consider a prior on the μ_i's where a fraction ε of the μ_i's are randomly picked and set equal to some $\mu > 0$, while the others are set to 0. The prior is therefore defined based on ε and μ, which together control the signal strength. The P-value corresponding $\mathbb{H}_i$ is $P_i := \bar{\Phi}(X_i)$, where

$\bar{\Phi} := 1 - \Phi$ is the null survival function.

3.4.1 The normal model

As an emblematic example of the distributional models that we consider in this chapter, let Φ denote the standard normal distribution. Assume as above that $X_i \sim \Phi$ under $\mathbb{H}_i$ and $X_i \sim \Phi(\cdot - \mu)$ otherwise. Thus, under the each null hypothesis, the corresponding test statistic is standard normal, while that statistic is normal with mean μ and unit variance otherwise. This is the model we consider in [ACC17] and the inspiration comes from a line of research on testing the global null $\bigcap_i \mathbb{H}_i$ in the static setting [Ing97, IS03, DJ04]. As in this line of work, we use the parameterization pioneered by [Ing97], namely

$$\varepsilon = n^{-\beta} \text{ and } \mu = \sqrt{2r \log n}. \tag{3.9}$$

In the static setting, we know from our previous work [ACC17] that any threshold-type procedure has risk tending to 1 as $n \to \infty$ when $r < \beta$ are fixed. We also know that the BH method with FDR control at $q \to 0$ slowly has risk tending to 0 when $r > \beta$ are fixed. In fact, these results are derived in the wider context of an asymptotically generalized Gaussian model, which we consider later. Thus $r = \beta$ is the static selection boundary.

Remark 13. [JM15] compared the power of their procedures in terms of lower bounds on the total discovery rate under the same mixture model but with a fixed mixture weight ε. In contrast, here we focus on the setting where $\varepsilon \to 0$, meaning that the fraction of false null hypotheses (i.e., true discoveries) is negligible compared to the total number of null hypotheses being tested.

3.4.2 Asymptotically generalized Gaussian model

Beyond the normal model, we follow [ACC17, DJ04] and consider other location models where the base distribution has a polynomial right tail in log scale, defined as AGG class of distributions. See definition Definition 1 in Chapter 1.

Remark 14. We note that the scale (e.g., standard deviation) is fixed, but this is really without loss of generality as both the LORD and LOND methods are scale invariant. This is because the P-values are scale invariant.

The model is the same as the one considered in Section 3.4.1 except that Φ is an AGG distribution with parameter $\gamma \geq 1$. As in our previous work [ACC17], we use the following parametrization

$$\varepsilon = n^{-\beta} \text{ and } \mu = (\gamma r \log n)^{1/\gamma}, \tag{3.10}$$

where $r \geq 0$ and $\beta \in (0,1)$ are always assumed fixed.

3.5 Performance analysis

In this section we analyze the performance of the LORD and LOND methods in the static setting described earlier. Recall that q denotes the desired FDR control level. Typically q is set to a small number, like $q = 0.10$. In this chapter we allow $q \to 0$ as $\varepsilon \to 0$, but slowly. Specifically, we always assume that

$$q = q(n) > 0 \text{ and } n^a q(n) \to \infty \text{ for all fixed } a > 0. \tag{3.11}$$

3.5.1 The performance of LORD

We first establish a performance bound for LORD. It happens that, despite required to control the FDR in an online fashion, LORD achieves the static selection boundary when desired FDR control is appropriately set.

Theorem 9 (Performance bound for LORD)**.** *Consider a static AGG mixture model with exponent* $\gamma \geq 1$ *parameterized as in* (3.10)*. Assume that we apply LORD with* $(\lambda_i)_{i=1}^{\infty}$ *defined as* $\lambda_i \propto i^{-\nu}$ *with* $\sum_{i=1}^{\infty} \lambda_i = q$*, where* $\nu > 1$ *and* q *satisfies* (3.11)*. If* $r > \nu\beta$*, the LORD procedure has* $\mathrm{FNR}(n) \to 0$

as $n \to \infty$. In particular, if $q \to 0$, then it has risk tending to 0.

Note that the latter part comes from the fact that the LORD procedure controls the FDR at the desired level q as established in [JM15] in the more demanding online setting. In essence, therefore, LORD (with a proper choice of ν above) achieves the static oracle selection boundary $r = \beta$.

Remark 15. Assume that, instead, we apply LORD with any decreasing sequence $(\lambda_i)_{i=1}^{\infty}$ satisfying $\sum_{i=1}^{\infty} \lambda_i = q$ and

$$i^{\nu}\lambda_i \to \infty, \text{ for any fixed } \nu > 1. \tag{3.12}$$

Then the conclusions of Theorem 9 remain valid. In particular, such a choice of sequence (e.g., $\lambda_i \propto (\log i)^2/i$) adapts to the (usually unknown) values of r and β. (We provide details in Section 3.7.)

3.5.2 The performance of LOND

We now turn to LOND and establish a performance bound under the same setting.

Theorem 10. *Consider a static AGG mixture model with exponent $\gamma \geq 1$ parametrized as in* (3.10). *Assume that we apply LOND with $(\lambda_i)_{i=1}^{\infty}$ defined as $\lambda_i \propto i^{-\nu}$ with $\sum_{i=1}^{\infty} \lambda_i = q$, where $\nu > 1$ and q satisfies* (3.11). *If $r > \beta + (\nu^{1/\gamma} - r^{1/\gamma})^{\gamma} + \nu - 1$, the LOND procedure has $\mathrm{FNR}(n) \to 0$ as $n \to \infty$. In particular, if $q \to 0$, then it has risk tending to 0.*

In essence, LOND (with a proper choice of ν above) has risk tending to 0 when $r - (1 - r^{1/\gamma})^{\gamma} > \beta$. This is the best upper bound that we were able to establish for the LOND algorithm. We do not know if it is optimal or not. In particular, it's quite possible that LOND also achieves the static selection boundary.

Remark 16. The analog of Remark 15 applies here as well. (Technical details are omitted.)

3.6 Numerical experiments

In this section, we perform some simulations to study the performance of LORD and LOND algorithms on finite data, and also to compare them with the (static) BH procedure. We consider the normal model and the double-exponential model. It is worth repeating that the BH procedure, which is a static procedure, requires knowledge of all P-values to determine the significance level for testing the hypotheses. Hence, it does not address the scenario in online testing. In contrast, the sequential methods decide the significance level at each step based on previous outcomes and are required to control the FDR at each step.

In our experiments, for both LORD and LOND, we choose the sequence $(\lambda_i)_{i=1}^{\infty}$ as

$$\lambda_i = Li^{-1.05}, \tag{3.13}$$

with L set to ensure $\sum_{i=1}^{\infty} \lambda_i = q$, where (as before) q denotes the desired FDR level.

3.6.1 Fixed sample size

In this first set of experiments, the sample size is chosen large at $n = 10^9$. We draw m observations from the alternative distribution $\Phi(\cdot - \mu)$, and the other $n - m$ from the null distribution Φ. All the models are parameterized as in (3.10). We choose a few values for the parameter β so as to exhibit different sparsity levels, while the parameter r takes values in a grid of spanning $[0, 1.5]$. Each situation is repeated 300 times and we report the average FDP and FNP for each procedure. The FDR control level is set at $q = 0.1$.

Normal model

In this model Φ is the standard normal distribution. The simulation results are reported in Figure 3.1 and Figure 3.2. In Figure 3.1 we report the FDP. We see that LOND becomes more conservative than LORD as r increases. In Figure 3.2 we report the FNP. We see that LOND

is clearly less powerful than LORD in the regime $\beta = 0.2$, but performs comparably to LORD in the regime $\beta = 0.6$. This is in line with the theory that LOND can at least achieve the line $r = \beta + (1 - r^{1/\gamma})^{\gamma}$, which is getting closer to $r = \beta$ with increasing values of β. We notice that both LORD and LOND are clearly less powerful than BH in finite samples, even at $n = 10^9$, even though our theory says that LORD achieves the same selection boundary as BH in the large-sample limit. Also, due to the limitation in choice of ν (here $\nu = 1.05$), the selection boundary that LORD can achieve is $r = \nu\beta$ by theory, which explains why LORD lags behind BH. Finally, we remark the transition of LORD from high FNP to low FNP happens in the vicinity of the theoretical threshold ($r = \beta$).

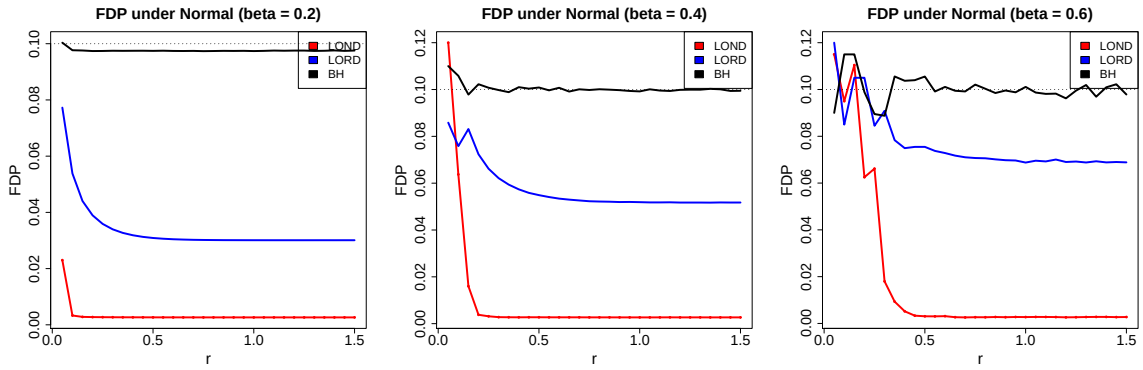

Figure 3.1: Simulation results showing the FDP for the BH, LORD and LOND methods under the normal model in three distinct sparsity regimes. The black horizontal line delineates the desired FDR control level ($q = 0.1$).

Double-exponential model

In this model Φ is the double-exponential distribution with variance 1. The simulation results are reported in Figure 3.3 (FDP) and Figure 3.4 (FNP). Here we observe that LOND becomes more conservative than LORD as r increases in terms of FDP. The LOND and LORD perform more comparably than in the normal setting in terms of FNP, especially when β is close to 1. This is again in line with our theoretical results. The BH method clearly outperforms the other two methods even though $n = 10^9$. Due to the limitation in choice of ν (here $\nu = 1.05$), the

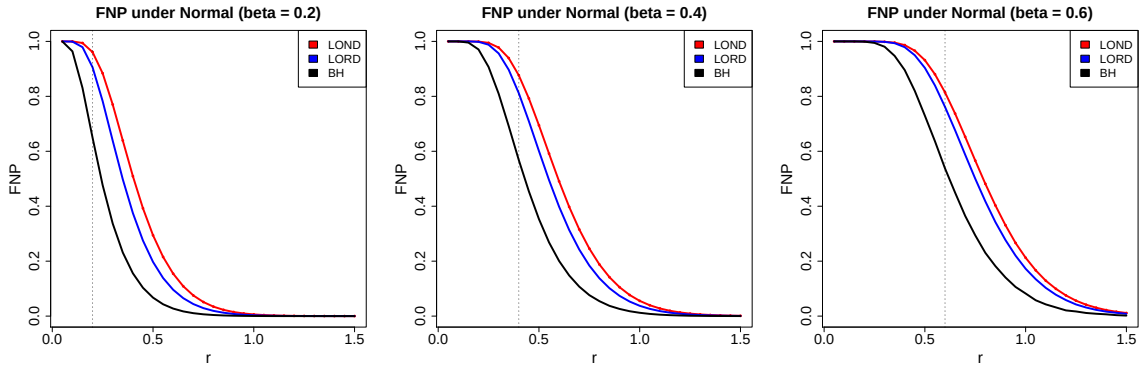

Figure 3.2: Simulation results showing the FNP for the BH, LORD and LOND methods under the normal model in three distinct sparsity regimes. The black vertical line delineates the theoretical threshold ($r = \beta$).

selection boundary that LORD can achieve is $r = \nu\beta$ by theory, which explains why LORD lags behind BH. The transition of three methods from FNP near 1 to FNP near 0 happens, again, in the vicinity of the theoretical threshold, but is much sharper here.

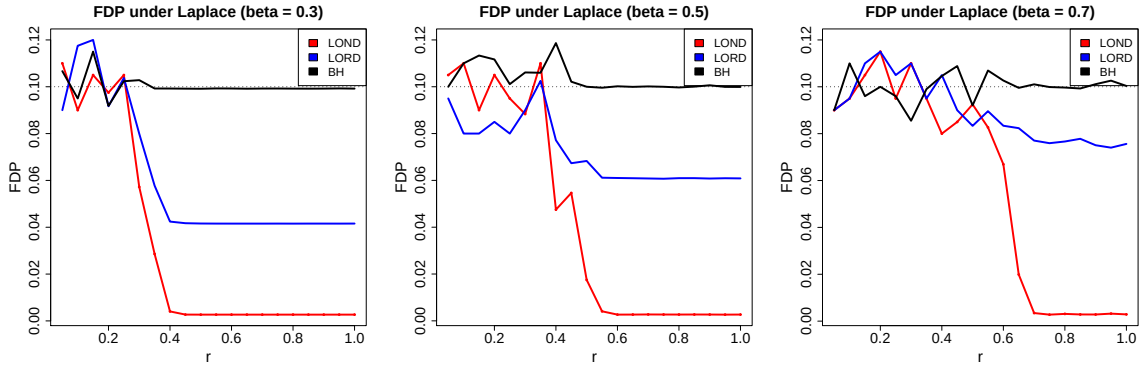

Figure 3.3: Simulation results showing the FDP for the BH, LORD and LOND methods under the double-exponential model in three distinct sparsity regimes. The black horizontal line delineates the desired FDR control level ($q = 0.1$).

3.6.2 Varying sample size

In this second set of experiments we examine the effect of various sample sizes on the risk of the LORD and LOND procedures under the standard normal model and the double-exponential model (with variance 1).

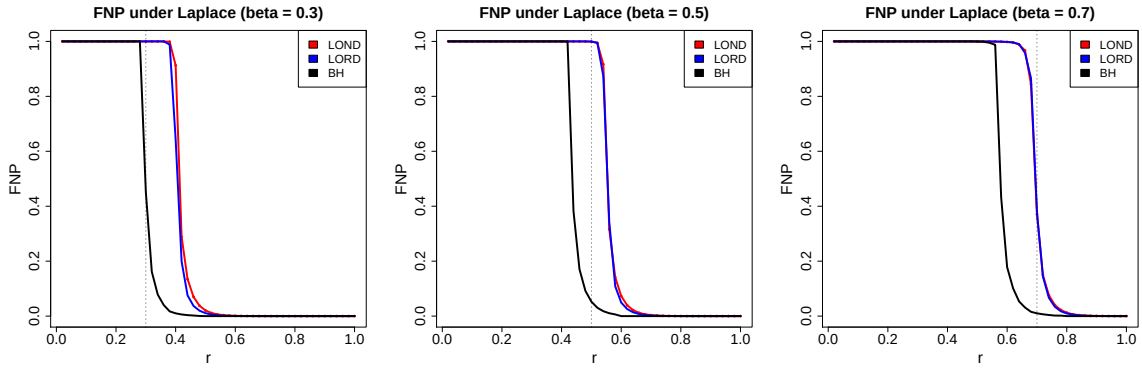

Figure 3.4: Simulation results showing the FNP for the BH, LORD and LOND methods under the double exponential model in three distinct sparsity regimes. The black vertical line delineates the theoretical threshold ($r = \beta$).

FNR of LORD with a fixed level

In this subsection, we present numerical experiments meant to illustrate the theoretical results we derived about asymptotic FNR of LORD. We fix $q = 0.1$, and choose a few values for the parameter β so as to exhibit different sparsity levels, while the parameter r takes values in a grid of spanning $[0, 1.5]$. We plot the average FNP of LORD procedure with different $n \in \{10^6, 10^7, 10^8, 10^9\}$. The simulation results are reported in Figure 3.5 and Figure 3.6. Each situation is repeated 200 times. We observe that in the normal model when $r > \beta$, the FNP decreases as n is getting larger. In the double-exponential model, as n increases, the FNP transition lines are getting closer the theoretical thresholds $r = \beta$, especially when $\beta = 0.7$.

Varying level

Here we explore the effect of letting the desired FDR control level q tend to 0 as n increases in accordance with (3.11). Specifically, we set it as $q = q_n = 1/\log n$. We choose n on a log scale, specifically, $n \in \{10^5, 10^6, 10^7, 10^8, 10^9\}$. Each time, we fix a value of (β, r) such that $r > \beta$.

In the first setting, we set $(\beta, r) = (0.4, 0.9)$ for normal model and $(\beta, r) = (0.4, 0.7)$ for double-exponential model. The simulation results are reported in Figure 1.5 and Figure 1.6. We

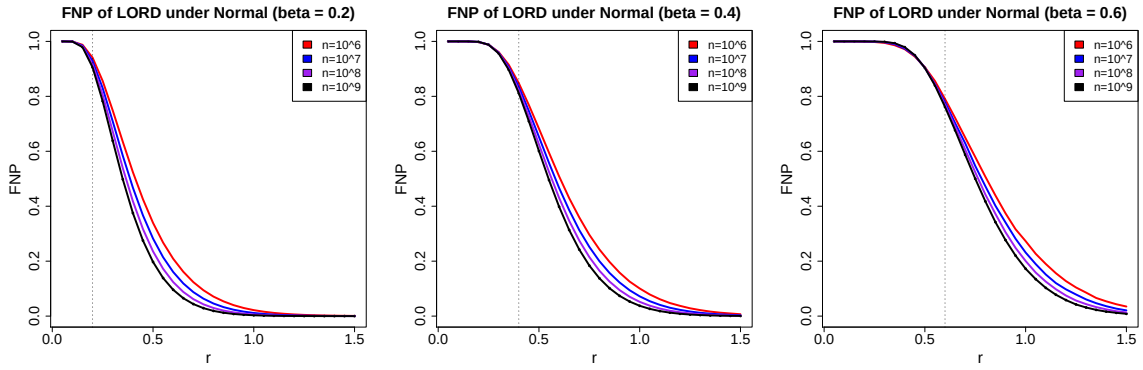

Figure 3.5: Simulation results showing the FNP for LORD under the normal model in three distinct sparsity regimes with different sample size. The black vertical line delineates the theoretical threshold ($r = \beta$).

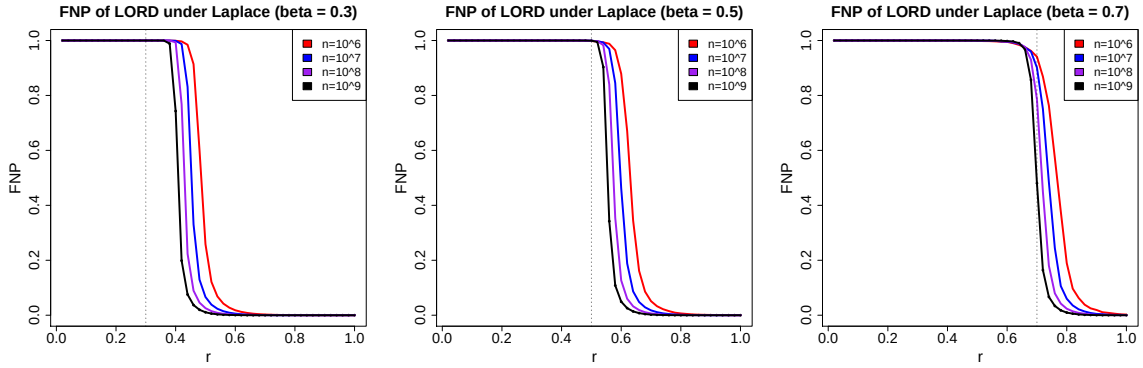

Figure 3.6: Simulation results showing the FNP for LORD under the double exponential model in three distinct sparsity regimes with different sample size. The black vertical line delineates the theoretical threshold ($r = \beta$).

see that, in both models, the risks of the two procedures decrease to zero as the sample size gets larger. LORD clearly dominates LOND (in terms of FNP). Both methods have FDP much lower than the level q_n, and in particular, LOND is very conservative.

In the second setting, we set $(\beta, r) = (0.7, 1.5)$ for normal model and $(\beta, r) = (0.7, 0.9)$ for double-exponential model. The simulation results are reported in Figure 3.9 and Figure 3.10. In this sparser regime, we can see that although LORD still dominates, the difference in FNP between two methods is much smaller than that in dense regime, especially in the double-exponential model. Both methods have FDP lower than the level q_n, and in particular, LOND is very conservative.

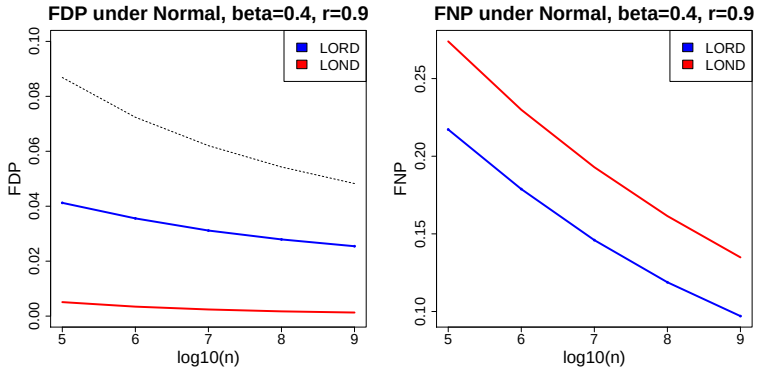

Figure 3.7: FDP and FNP for the LORD and LOND methods under the normal model with $(\beta, r) = (0.4, 0.9)$ and varying sample size n. The black line delineates the desired FDR control level $(q = q_n)$.

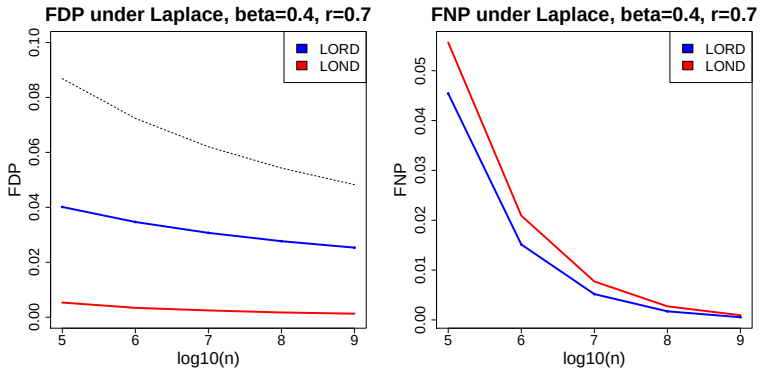

Figure 3.8: FDP and FNP for the LORD and LOND methods under the double-exponential model with $(\beta, r) = (0.4, 0.7)$ and varying sample size n. The black line delineates the desired FDR control level $(q = q_n)$.

3.7 Proofs

We prove our results in this section. Let Φ denote the CDF of null distribution. Without loss of generality, we assume throughout that $\Phi(0) = 1/2$. Let $F(t)$ denote the CDF of the P-values under alternatives so that

$$F(t) = \Phi(\mu - \Phi^{-1}(1-t)), \tag{3.14}$$

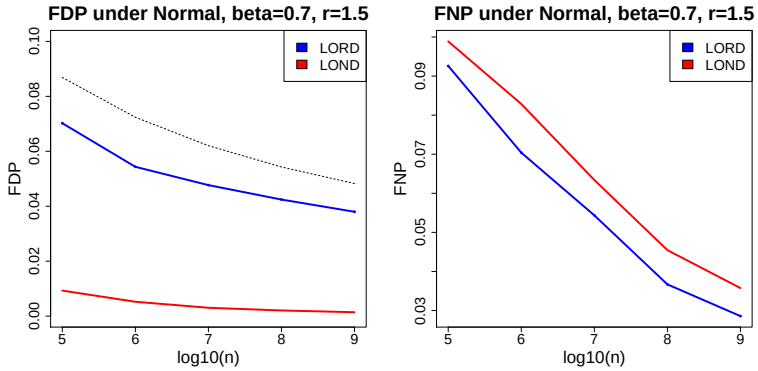

Figure 3.9: FDP and FNP for the LORD and LOND methods under the normal model with $(\beta, r) = (0.7, 1.5)$ and varying sample size n. The black line delineates the desired FDR control level ($q = q_n$).

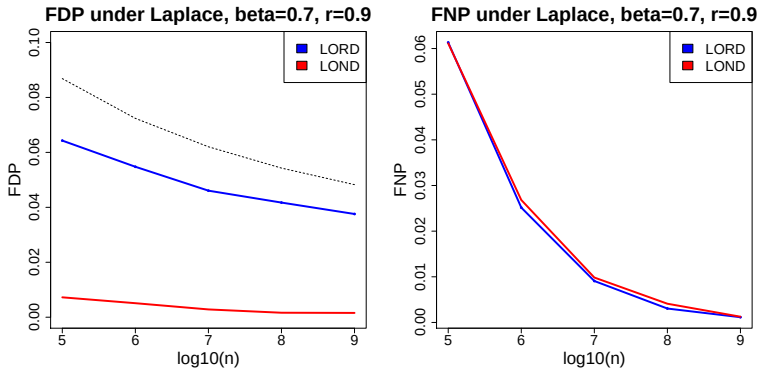

Figure 3.10: FDP and FNP for the LORD and LOND methods under the double-exponential model with $(\beta, r) = (0.7, 0.9)$ and varying sample size n.

where Φ^{-1} is the inverse function of Φ. Let

$$G(t) = (1-\varepsilon)t + \varepsilon F(t), \tag{3.15}$$

which is the CDF of the P-values from the mixture model. Let $\bar{F} = 1 - F$, which is the survival function of the P-values under alternatives. Note that

$$\bar{F}(t) = 1 - F(t) = 1 - \Phi(\mu - \xi) = \bar{\Phi}(\mu - \xi), \tag{3.16}$$

where $\xi := \Phi^{-1}(1-t)$, or equivalently, $t = \bar{\Phi}(\xi)$. Because Φ is as in Definition 1, when $\xi \to \infty$, we have

$$t = \bar{\Phi}(\xi) = \exp\left\{-\frac{\xi^\gamma}{\gamma}(1+o(1))\right\} \to 0, \tag{3.17}$$

which also implies, when $t \to 0$, that

$$\xi = \Phi^{-1}(1-t) \sim (\gamma \log(1/t))^{1/\gamma}. \tag{3.18}$$

3.7.1 Discovery times (LORD)

We apply LORD to the static setting under consideration. Denote τ_l as the time of l-th discovery (with $\tau_0 = 0$), and $\Delta_l = \tau_l - \tau_{l-1}$ as the time between the $(l-1)$-th and l-th discoveries. Assume a sequence satisfying (3.5) has been chosen. Given the update rule of (3.6), it can be seen that the inter-discovery times $\{\Delta_l : l \geq 1\}$ are IID.

To prove Theorem 9, we will use the following bound on the expected inter-discovery time.

Proposition 1. *Consider a static AGG mixture model with exponent $\gamma \geq 1$ parameterized as in* (3.10)*. Assume that $\beta \in (0,1)$ and $r \geq 0$ are both fixed. Assume that $r > \beta$ and let $\nu > 1$ be such that $\nu < r/\beta$. If we apply LORD with $(\lambda_i)_{i=1}^\infty$ defined as $\lambda_i \propto i^{-\nu}$ with $\sum_{i=1}^\infty \lambda_i = q$,*

$$\mathbb{E}(\Delta_l \wedge n) \leq 2n^\beta + C, \quad \textit{for all } l > 0, \tag{3.19}$$

for some $C > 0$ that does not depend on n. The same holds if we apply LORD with $(\lambda_i)_{i=1}^\infty$ satisfying (3.12) *and $\sum_{i=1}^\infty \lambda_i = q$.*

We prove this result. Recall the definition of G in (3.15) and note that $G \geq \varepsilon F$. By the

update rule of LORD algorithm, for all $m \geq 1$ we have

$$\mathbb{P}(\Delta_l > m) = \prod_{i=\tau_{l-1}+1}^{\tau_{l-1}+m} (1 - G(\alpha_i)) = \prod_{i=\tau_{l-1}+1}^{\tau_{l-1}+m} (1 - G(\lambda_{i-\tau_{l-1}})) \tag{3.20}$$

$$= \prod_{i=1}^{m} (1 - G(\lambda_i)) \leq \exp\Big\{ - \sum_{i=1}^{m} G(\lambda_i) \Big\} \leq \exp\Big\{ - \varepsilon \sum_{i=1}^{m} F(\lambda_i) \Big\}. \tag{3.21}$$

Let t^* be the value such that $\Phi^{-1}(1-t^*) = \mu$, i.e., $t^* = \Phi(-\mu) = n^{-r+o(1)}$ by the fact that Φ satisfies Definition 1. Then, for $t \geq t^*$, we get

$$\Phi^{-1}(1-t) \leq \Phi^{-1}(1-t^*) = \mu, \tag{3.22}$$

and then

$$F(t) = \Phi(\mu - \Phi^{-1}(1-t)) \geq \Phi(\mu - \Phi^{-1}(1-t^*)) = \Phi(\mu - \mu) = \Phi(0) = 1/2, \tag{3.23}$$

so that if $\lambda_i = Li^{-\nu} \geq t^*$, i.e., $i \leq n_1 := \lfloor (L/t^*)^{1/\nu} \rfloor = n^{r/\nu + o(1)}$, we have $F(\lambda_i) \geq \Phi(0) = 1/2$.

Remark 17. If instead $(\lambda_i)_{i=1}^{\infty}$ satisfies (3.12) then $i^{\nu}\lambda_i \to \infty$ as $i \to \infty$, so that exists a constant $L > 0$ such that $\lambda_i \geq Li^{-\nu}$ for all i, and this is all that we need to proceed.

Thus, for $m \leq n_1$,

$$\sum_{i=1}^{m} F(\lambda_i) \geq m/2, \tag{3.24}$$

and for $m > n_1$,

$$\sum_{i=1}^{m} F(\lambda_i) \geq \sum_{i=1}^{n_1} F(\lambda_i) \geq n_1/2. \tag{3.25}$$

Thus,

$$\mathbb{P}(\Delta_l > m) \leq \exp\{-\varepsilon(m \wedge n_1)/2\}. \tag{3.26}$$

Next we bound $\mathbb{E}(\Delta_l \wedge n)$. Due to the fact that $\{\Delta_l \wedge n > m\} = \{\Delta_l > m\}$ for $1 \leq m \leq n-1$,

and $\{\Delta_l \wedge n > m\} = \varnothing$ if $m \geq n$, we have

$$\mathbb{E}(\Delta_l \wedge n) = \sum_{m=0}^{\infty} \mathbb{P}(\Delta_l \wedge n > m) \tag{3.27}$$

$$= \sum_{m=1}^{n-1} \mathbb{P}(\Delta_l > m) + 1 \tag{3.28}$$

$$\leq \sum_{m=1}^{n-1} \exp\{-\varepsilon(m \wedge n_1)/2\} + 1. \tag{3.29}$$

We split the summation over $1 \leq m \leq n_1$ and $n_1 + 1 \leq m \leq n$ and derive the corresponding upper bound separately. For the first part,

$$\sum_{m=1}^{n_1} \exp\{-\varepsilon(m \wedge n_1)/2\} = \sum_{m=1}^{n_1} \exp\{-\varepsilon m/2\} \leq \frac{1}{\exp\{\varepsilon/2\} - 1} < \frac{2}{\varepsilon} = 2n^{\beta}. \tag{3.30}$$

For the second part,

$$\sum_{m=n_1+1}^{n-1} \exp\{-\varepsilon(m \wedge n_1)/2\} = \sum_{m=n_1+1}^{n-1} \exp\{-\varepsilon n_1/2\} \leq n \exp\{-\varepsilon n_1/2\} = o(1), \tag{3.31}$$

since $\varepsilon n_1 = n^{r/\nu - \beta + o(1)}$ and $\frac{r}{\nu} > \beta$. Combining the above two bounds, we obtain

$$\mathbb{E}(\Delta_l \wedge n) \leq 2n^{\beta} + o(1) + 1. \tag{3.32}$$

This establishes Proposition 1.

3.7.2 Proof of Theorem 9

Note the number of false nulls is $m = |\mathcal{H}^1(n)| = \varepsilon n \sim n^{1-\beta}$. The false non-discovery rate of LORD (denoted $\mathrm{FNR}(n)$) is as follows:

$$\mathrm{FNR}(n) = \mathbb{E}\left(\frac{\sum_{i=1}^{n} \mathbb{I}\{i \notin \mathcal{H}_0(n) : P_i \geq \alpha_i\}}{m}\right) \tag{3.33}$$

$$= \frac{\sum_{i=1}^{n} \mathbb{E}[\mathbb{E}(\mathbb{I}\{i \notin \mathcal{H}_0(n) : P_i \geq \alpha_i\} \mid \alpha_i)]}{m} \tag{3.34}$$

$$= \frac{\sum_{i=1}^{n} \mathbb{E}[\mathbb{P}(i \notin \mathcal{H}_0(n), P_i \geq \alpha_i \mid \alpha_i)]}{m} \tag{3.35}$$

$$= \frac{\sum_{i=1}^{n} \mathbb{E}[\varepsilon \bar{F}(\alpha_i)]}{\varepsilon n} \tag{3.36}$$

$$= \frac{\sum_{i=1}^{n} \mathbb{E}[\bar{F}(\alpha_i)]}{n}. \tag{3.37}$$

So it suffices to bound the RHS of the equation.

Let $D(n)$ be the number of discoveries in first n hypotheses $\mathcal{H}(n)$ by applying LORD with the sequence (λ_i). Let $\tilde{\Delta}_l = \Delta_l = \tau_l - \tau_{l-1}$, for $1 \leq l \leq D(n)$, and $\tilde{\Delta}_{D(n)+1} = n - \tau_{D(n)}$. Due to the fact that $0 \leq \tilde{\Delta}_l \leq (\Delta_l \wedge n)$, for $1 \leq l \leq D(n)+1$, we have for any fixed $\delta > 0$,

$$\mathbb{P}(\tilde{\Delta}_l \geq \frac{\mathbb{E}(\Delta_l \wedge n)}{\delta}) \leq \frac{\mathbb{E}(\tilde{\Delta}_l)}{\mathbb{E}(\Delta_l \wedge n)} \cdot \delta \leq \delta, \quad \text{for } 1 \leq l \leq D(n)+1, \tag{3.38}$$

by Markov Inequality. Note that $(\Delta_l \wedge n)$'s are IID. We define $M_n := \lceil \mathbb{E}(\Delta)/\delta \rceil$, where $\Delta \overset{d}{=} \Delta_l \wedge n$ for all $l > 0$.

For any $i \in \mathcal{H}(n)$, there exists only one $j = j(i) \in \{1, 2, \ldots, D(n)+1\}$ such that $i \in (\tau_{j-1}, \tau_j \wedge n]$, and

$$\mathbb{E}\left[\bar{F}(\alpha_i)\right] = \mathbb{E}\left[\bar{F}(\alpha_i) \cdot \mathbb{I}\{\tilde{\Delta}_{j(i)} \geq M_n\}\right] + \mathbb{E}\left[\bar{F}(\alpha_i) \cdot \mathbb{I}\{\tilde{\Delta}_{j(i)} < M_n\}\right] \tag{3.39}$$

$$\leq \delta + \mathbb{E}\left[\bar{F}(\alpha_i) \cdot \mathbb{I}\{\tilde{\Delta}_{j(i)} < M_n\}\right], \tag{3.40}$$

so that

$$\sum_{i=1}^{n} \mathbb{E}\left[\bar{F}(\alpha_i)\right] \leq n\delta + \mathbb{E}\left[\sum_{i=1}^{n} \bar{F}(\alpha_i) \cdot \mathbb{I}\{\tilde{\Delta}_{j(i)} < M_n\}\right]. \tag{3.41}$$

By Proposition 1, there is $C > 0$ not depending on n such that

$$\mathbb{E}(\Delta) \leq 2n^{\beta} + C, \quad \text{for all } l > 0. \tag{3.42}$$

And thus, there is some $L' > 0$ (constant in n) such that, for $1 \leq i \leq M_n$,

$$\lambda_i = Li^{-\nu} \geq L \cdot (M_n)^{-\nu} = L \cdot \left\lceil \mathbb{E}(\Delta)/\delta \right\rceil^{-\nu} \geq L \cdot \left\lceil (2n^{\beta} + C)/\delta \right\rceil^{-\nu} \geq L' n^{-\beta\nu}. \tag{3.43}$$

Remark 18. If instead $(\lambda_i)_{i=1}^{\infty}$ satisfies (3.12) then $i^{\nu}\lambda_i \to \infty$ as $i \to \infty$, so that exists a constant $L > 0$ such that $\lambda_i \geq Li^{-\nu}$ for all i, and this is all that we need to proceed.

Since $\bar{F}$ is a decreasing function, the second term in RHS of (3.41) can be bounded as

$$\mathbb{E}\left[\sum_{i=1}^{n} \bar{F}(\alpha_i) \cdot \mathbb{I}\{\tilde{\Delta}_{j(i)} < M_n\}\right] = \mathbb{E}\left[\sum_{j=1}^{D(n)+1} \sum_{i=\tau_{j-1}+1}^{\tau_j \wedge n} \bar{F}(\alpha_i) \cdot \mathbb{I}\{\tilde{\Delta}_j < M_n\}\right] \tag{3.44}$$

$$= \mathbb{E}\left[\sum_{j=1}^{D(n)+1} \sum_{i=1}^{\tilde{\Delta}_j} \bar{F}(\lambda_i) \cdot \mathbb{I}\{\tilde{\Delta}_j < M_n\}\right] \tag{3.45}$$

$$\leq \mathbb{E}\left[\sum_{j=1}^{D(n)+1} \sum_{i=1}^{\tilde{\Delta}_j} \bar{F}(L' n^{-\beta\nu}) \cdot \mathbb{I}\{\tilde{\Delta}_j < M_n\}\right] \tag{3.46}$$

$$\leq \mathbb{E}\left[\sum_{i=1}^{n} \bar{F}(L' n^{-\beta\nu})\right] \leq n \cdot \bar{F}(L' n^{-\beta\nu}). \tag{3.47}$$

Combining these bounds, we obtain

$$\mathrm{FNR}(n)(\mathcal{R}) = \frac{\sum_{i=1}^{n} \mathbb{E}[\bar{F}(\alpha_i)]}{n} \leq \delta + \bar{F}(L' n^{-\beta\nu}). \tag{3.48}$$

Since $L'n^{-\beta\nu} \to 0$ as $n \to \infty$, by equation (3.18) we have

$$\xi_n := \Phi^{-1}(1 - L'n^{-\beta\nu}) = (\gamma\beta\nu \log n)^{1/\gamma}(1 + o(1)), \tag{3.49}$$

so that

$$\mu - \xi_n = (\gamma r \log n)^{1/\gamma} - (\gamma\beta\nu \log n)^{1/\gamma}(1 + o(1)) \tag{3.50}$$

$$\sim (r^{1/\gamma} - (\beta\nu)^{1/\gamma})(\gamma \log n)^{1/\gamma} \to \infty, \quad \text{as } n \to \infty, \tag{3.51}$$

since $r > \beta\nu$. Therefore, $\bar{F}(L'n^{-\beta\nu}) = \bar{\Phi}(\mu - \xi_n) \to 0$ as $n \to \infty$. Hence,

$$\limsup_{n\to\infty} \mathrm{FNR}(n) \le \delta. \tag{3.52}$$

This being true for any $\delta > 0$, necessarily, $\mathrm{FNR}(n) \to 0$ as $n \to \infty$. This establishes Theorem 9.

3.7.3 Discovery times (LOND)

We apply LOND to the static setting under consideration. Denote τ_l as the time of l-th discovery (with $\tau_0 = 0$), and $\Delta_l = \tau_l - \tau_{l-1}$ as the time between the $(l-1)$-th and l-th discoveries. Assume a sequence satisfying (3.5) has been chosen. Given the update rule of (3.8), it can be seen that the inter-discovery times $\{\Delta_l : l \ge 1\}$ are i.i.d..

To prove Theorem 10, we will use the following bound on the expected discovery times.

Proposition 2. *Consider a static AGG mixture model with exponent $\gamma \ge 1$ parameterized as in* (3.10)*. Assume that $\beta \in (0,1)$ and $r \in [0,1]$ are both fixed. For any $\nu > 1$, if we apply LOND with $(\lambda_i)_{i=1}^{\infty}$ defined as $\lambda_i \propto i^{-\nu}$ with $\sum_{i=1}^{\infty} \lambda_i = q$,*

$$\mathbb{E}(\tau_l \wedge n) \le l \cdot n^{\beta + (\nu^{1/\gamma} - r^{1/\gamma})^{\gamma} + b_n}, \quad \textit{for all } l > 0, \tag{3.53}$$

where $b_n \to 0$ as $n \to \infty$.

We now prove this result. By the update rule of LOND algorithm, for all $l \geq 0$, and all $m \geq \tau_l + 1$, we have

$$\mathbb{P}(\tau_{l+1} > m \mid \tau_l) = \prod_{i=\tau_l+1}^{m} (1 - G((l+1)\lambda_i)) \leq \exp\{-\sum_{i=\tau_l+1}^{m} G((l+1)\lambda_i)\}. \tag{3.54}$$

Note τ_l is the time of l-th discovery (with $\tau_0 = 0$) by LOND. Let $\tilde{\tau}_l = \tau_l \wedge n$. If $\tilde{\tau}_l = n$, we have $\mathbb{E}(\tilde{\tau}_{l+1} \mid \tilde{\tau}_l) = n = \tilde{\tau}_l$. Otherwise, if $\tilde{\tau}_l = \tau_l < n$,

$$\begin{aligned}
\mathbb{E}(\tilde{\tau}_{l+1} \mid \tilde{\tau}_l) &= \tau_l + 1 + \sum_{m=\tau_l+1}^{\infty} \mathbb{P}(\tau_{l+1} \wedge n > m \mid \tau_l) && (3.55)\\
&= \tau_l + 1 + \sum_{m=\tau_l+1}^{n-1} \mathbb{P}(\tau_{l+1} > m \mid \tau_l) && (3.56)\\
&\leq \tau_l + 1 + \sum_{m=\tau_l+1}^{n} \exp\{-\sum_{i=\tau_l+1}^{m} G((l+1)\lambda_i)\} && (3.57)\\
&= \tilde{\tau}_l + 1 + \sum_{m=\tilde{\tau}_l+1}^{n} \exp\{-\sum_{i=\tilde{\tau}_l+1}^{m} G((l+1)\lambda_i)\}. && (3.58)
\end{aligned}$$

Next we bound $\mathbb{E}(\tilde{\tau}_{l+1} \mid \tilde{\tau}_l)$. Let t^* be the value such that $\Phi^{-1}(1-t^*) = \mu$, i.e., $t^* = \Phi(-\mu) = n^{-r+o(1)}$ by the fact that Φ satisfies Definition 1. Then, for $t \geq t^*$, we get

$$\Phi^{-1}(1-t) \leq \Phi^{-1}(1-t^*) = \mu, \tag{3.59}$$

and,

$$F(t) = \Phi(\mu - \Phi^{-1}(1-t)) \geq \Phi(\mu - \Phi^{-1}(1-t^*)) = \Phi(\mu - \mu) = \Phi(0) = 1/2, \tag{3.60}$$

so that if $(l+1)\lambda_i = (l+1)Li^{-\nu} \geq t^*$, i.e., $i \leq n_1 := \lfloor ((l+1)L/t^*)^{1/\nu} \rfloor = n^{r/\nu+o(1)}$, we have

$$F((l+1)\lambda_i) \geq \Phi(0) = 1/2.$$

We consider the following cases.

Case 1: $\tilde{\tau}_l < n_1 < n$**.** In this case, for $\tilde{\tau}_l + 1 \le m \le n_1$,

$$\sum_{i=\tilde{\tau}_l+1}^{m} G((l+1)\lambda_i) \ge \sum_{i=\tilde{\tau}_l+1}^{m} \varepsilon F((l+1)\lambda_i) \ge \varepsilon \cdot (m-\tilde{\tau}_l)/2, \tag{3.61}$$

and for $m > n_1$,

$$\sum_{i=\tilde{\tau}_l+1}^{m} G((l+1)\lambda_i) \ge \sum_{i=\tilde{\tau}_l+1}^{m} \varepsilon F((l+1)\lambda_i) \ge \sum_{i=\tilde{\tau}_l+1}^{m} \varepsilon F((l+1)\lambda_m) = (m-\tilde{\tau}_l)\varepsilon F((l+1)\lambda_m), \tag{3.62}$$

since $F(x)$ is non-decreasing.

We split the summation in (3.58) over $\tau_l + 1 \le m \le n_1$ and $n_1 + 1 \le m \le n$ and derive the corresponding upper bound separately. For the first part,

$$\sum_{m=\tilde{\tau}_l+1}^{n_1} \exp\{-\sum_{i=\tilde{\tau}_l+1}^{m} G((l+1)\lambda_i)\} \le \sum_{m=\tilde{\tau}_l+1}^{n_1} \exp\{-\varepsilon(m-\tilde{\tau}_l)/2\} = \sum_{m=1}^{n_1-\tilde{\tau}_l} \exp\{-\varepsilon m/2\} \tag{3.63}$$

$$\le \frac{1}{\exp\{\varepsilon/2\}-1} < \frac{2}{\varepsilon} = 2n^{\beta}. \tag{3.64}$$

For the second part,

$$\sum_{m=n_1+1}^{n} \exp\Big\{-\sum_{i=\tilde{\tau}_l+1}^{m} G((l+1)\lambda_i)\Big\} \le \sum_{m=n_1+1}^{n} \exp\{-(m-\tilde{\tau}_l)\varepsilon F((l+1)\lambda_m)\} \tag{3.65}$$

$$\le \sum_{m=n_1+1}^{n} \exp\{-(m-n_1)\varepsilon F((l+1)\lambda_n)\} \tag{3.66}$$

$$\le \sum_{m=1}^{n-n_1} \exp\{-m\varepsilon F((l+1)\lambda_n)\} \tag{3.67}$$

$$\le \frac{1}{\exp\{\varepsilon F((l+1)\lambda_n)\}-1} \tag{3.68}$$

$$< \frac{1}{\varepsilon F((l+1)\lambda_n)} \le \frac{1}{\varepsilon F(\lambda_n)}. \tag{3.69}$$

Case 2: $n_1 \le \tilde{\tau}_l < n$**.** For this case, we don't need to split the summation, since

$$\sum_{m=\tilde{\tau}_l+1}^{n} \exp\{-\sum_{i=\tilde{\tau}_l+1}^{m} G((l+1)\lambda_i)\} \le \sum_{m=\tilde{\tau}_l+1}^{n} \exp\{-(m-\tilde{\tau}_l)\varepsilon F((l+1)\lambda_m)\} \tag{3.70}$$

$$\le \sum_{m=1}^{n-\tilde{\tau}_l} \exp\{-m\varepsilon F((l+1)\lambda_n)\} \tag{3.71}$$

$$< \frac{1}{\varepsilon F((l+1)\lambda_n)} \le \frac{1}{\varepsilon F(\lambda_n)}. \tag{3.72}$$

Case 3: $n_1 \ge n$**.** Since $\tilde{\tau}_l < n \le n_1$, we have that

$$\sum_{m=\tilde{\tau}_l+1}^{n} \exp\{-\sum_{i=\tilde{\tau}_l+1}^{m} G((l+1)\lambda_i)\} \le \sum_{m=\tilde{\tau}_l+1}^{n} \exp\{-(m-\tilde{\tau}_l)\varepsilon/2\} \tag{3.73}$$

$$\le \sum_{m=1}^{n-\tilde{\tau}_l} \exp\{-m\varepsilon/2\} < \frac{2}{\varepsilon} = 2n^{\beta}. \tag{3.74}$$

Combining all the cases, we obtain

$$\mathbb{E}(\tilde{\tau}_{l+1} \mid \tilde{\tau}_l) \le \tilde{\tau}_l + 1 + \frac{2}{\varepsilon} + \frac{1}{\varepsilon F(\lambda_n)}, \tag{3.75}$$

where $F(\lambda_n) = \bar{\Phi}(\xi_n - \mu)$, and $\xi_n := \Phi^{-1}(1-\lambda_n)$. Since $\lambda_n = Ln^{-\nu} \to 0$ as $n \to 0$, by equation (3.18), we have $\xi_n \sim (\gamma\nu \log n)^{1/\gamma}$, so that

$$\xi_n - \mu = (\gamma\nu \log n)^{1/\gamma}(1+o(1)) - (\gamma r \log n)^{1/\gamma} \tag{3.76}$$

$$\sim (\nu^{1/\gamma} - r^{1/\gamma})(\gamma \log n)^{1/\gamma} \to \infty, \quad \text{as } n \to \infty, \tag{3.77}$$

by the fact that $\nu > 1 \ge r$. By Definition 1,

$$F(\lambda_n) = \bar{\Phi}(\xi_n - \mu) = \exp\Big\{-\frac{(\xi_n-\mu)^{\gamma}}{\gamma}(1+o(1))\Big\} = n^{-(\nu^{1/\gamma}-r^{1/\gamma})^{\gamma}+o(1)}. \tag{3.78}$$

Thus, when n is large enough,

$$\mathbb{E}(\tilde{\tau}_{l+1} \mid \tilde{\tau}_l) \le \tilde{\tau}_l + 1 + \frac{2}{\varepsilon} + \frac{1}{\varepsilon F(\lambda_n)} \le \tilde{\tau}_l + n^{\beta+(\nu^{1/\gamma}-r^{1/\gamma})^\gamma+o(1)}, \quad \text{for all } l > 0, \tag{3.79}$$

where the $o(1)$ is uniform in l, and this further implies that

$$\mathbb{E}(\tilde{\tau}_{l+1}) = \mathbb{E}[\mathbb{E}(\tilde{\tau}_{l+1} \mid \tilde{\tau}_l)] \le \mathbb{E}(\tilde{\tau}_l) + n^{\beta+(\nu^{1/\gamma}-r^{1/\gamma})^\gamma+o(1)}, \quad \text{for all } l > 0, \tag{3.80}$$

so that

$$\mathbb{E}(\tau_l \wedge n) \le l \cdot n^{\beta+(\nu^{1/\gamma}-r^{1/\gamma})^\gamma+o(1)}, \quad \text{for all } l > 0. \tag{3.81}$$

3.7.4 Proof of Theorem 10

It suffices to consider the case where $r \in [0,1]$ since the observations from $\mathbb{H}_0$ almost never get substantially larger than $(\gamma \log n)^{1/\gamma}$. For $r \in [0,1]$, if $r - (1-r^{1/\gamma})^\gamma > \beta$, we can choose $\nu > 1$ close to 1 and $\eta > 0$ close to 0 such that $r > \rho := \beta + (\nu^{1/\gamma} - r^{1/\gamma})^\gamma + \nu - 1 + \eta$. By Proposition 2, when n is large enough,

$$\mathbb{E}(\tau_l \wedge n) \le l \cdot n^{\beta+(\nu^{1/\gamma}-r^{1/\gamma})^\gamma+\eta}, \quad \text{for all } l > 0. \tag{3.82}$$

Fix $\delta > 0$ and let $n_2 := \lceil n^{\beta+(\nu^{1/\gamma}-r^{1/\gamma})^\gamma+\eta}/\delta \rceil$. Note $n_2 = o(n)$, since $1 \ge r > \beta + (\nu^{1/\gamma} - r^{1/\gamma})^\gamma + \eta$.

For $n_2 \le i \le n$, let $\zeta_i := i\delta n^{-\beta-(\nu^{1/\gamma}-r^{1/\gamma})^\gamma-\eta}$, we get

$$\mathbb{P}(D(i) < \zeta_i) = \mathbb{P}(\tau_{\lceil\zeta_i\rceil} > i) \le \mathbb{P}(\tau_{\lceil\zeta_i\rceil} \ge i) = \mathbb{P}(\tau_{\lceil\zeta_i\rceil} \wedge n \ge i) \tag{3.83}$$

$$\le \frac{\mathbb{E}(\tau_{\lceil\zeta_i\rceil} \wedge n)}{i} \le \frac{\lceil\zeta_i\rceil \cdot n^{\beta+(\nu^{1/\gamma}-r^{1/\gamma})^\gamma+\eta}}{i} \tag{3.84}$$

$$< \frac{(\zeta_i+1)\cdot n^{\beta+(\nu^{1/\gamma}-r^{1/\gamma})^\gamma+\eta}}{i} \tag{3.85}$$

$$= \delta + \frac{n^{\beta+(\nu^{1/\gamma}-r^{1/\gamma})^\gamma+\eta}}{i} < 2\delta. \tag{3.86}$$

By Rule (3.8) defining the LOND algorithm,

$$\mathbb{E}[\bar{F}(\alpha_i)] = \mathbb{E}[\bar{F}(\lambda_i(D(i-1)+1))] \le \mathbb{E}[\bar{F}(\lambda_i D(i))], \tag{3.87}$$

due to the fact that $D(i-1)+1 \ge D(i)$ and that $\bar{F}(x)$ is a non-increasing function, so that LOND's false non-discovery rate (denoted $\mathrm{FNR}(n)$) is bounded as follows

$$\mathrm{FNR}(n) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\bar{F}(\alpha_i)] \le \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\bar{F}(\lambda_i D(i))]. \tag{3.88}$$

For $1 \le i \le n_2$,

$$\frac{1}{n}\sum_{i=1}^{n_2}\mathbb{E}[\bar{F}(\lambda_i D(i))] \le \frac{n_2}{n}. \tag{3.89}$$

And for $n_2+1 \le i \le n$,

$$\mathbb{E}[\bar{F}(\lambda_i D(i))] = \mathbb{E}[\bar{F}(\lambda_i D(i))\cdot\mathbb{I}\{D(i) < \zeta_i\}] + \mathbb{E}[\bar{F}(\lambda_i D(i))\cdot\mathbb{I}\{D(i) \ge \zeta_i\}] \tag{3.90}$$

$$\le 2\delta + \bar{F}(\lambda_i\zeta_i), \tag{3.91}$$

and since $\nu > 1$, we have

$$\lambda_i\zeta_i = L\delta\cdot i^{1-\nu}\cdot n^{-\beta-(\nu^{1/\gamma}-r^{1/\gamma})^\gamma-\eta} \ge L\delta\cdot n^{1-\nu}\cdot n^{-\beta-(\nu^{1/\gamma}-r^{1/\gamma})^\gamma-\eta} = \lambda_n\zeta_n, \tag{3.92}$$

which implies that,

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\bar{F}(\lambda_i D(i))] \leq \frac{n_2}{n} + \frac{n-n_2}{n}(2\delta + \bar{F}(\lambda_n\zeta_n)) \leq 2\delta + \bar{F}(\lambda_n\zeta_n) + o(1). \tag{3.93}$$

Since $\lambda_n\zeta_n = L\delta n^{-\rho} \to 0$ as $n \to \infty$, by equation (3.18)

$$\xi_n := \Phi^{-1}(1-\lambda_n\zeta_n) = (\gamma\rho\log n)^{1/\gamma}(1+o(1)), \tag{3.94}$$

then

$$\mu - \xi_n = (\gamma r\log n)^{1/\gamma} - (\gamma\rho\log n)^{1/\gamma}(1+o(1)) \tag{3.95}$$

$$\sim (r^{1/\gamma} - \rho^{1/\gamma})(\gamma\log n)^{1/\gamma} \to \infty, \quad \text{as } n \to \infty, \tag{3.96}$$

since $r > \rho$. Therefore, $\bar{F}(\lambda_n\zeta_n) = \bar{\Phi}(\mu - \xi_n) \to 0$ as $n \to \infty$. Hence,

$$\limsup_{n\to\infty} \mathrm{FNR}(n) \leq 2\delta. \tag{3.97}$$

This being true for any $\delta > 0$, necessarily, $\mathrm{FNR}(n) \to 0$ as $n \to \infty$.

3.8 Acknowledgement

Chapter 3, partially, is a version of the paper "Sequential Multiple Testing", Chen Shiyun; Arias-Castro, Ery. The manuscript has been submitted for publication in a major statistical journal. The dissertation author was the primary investigator and author of this material.

Chapter 4

Contextual Online False Discovery Rate Control

4.1 Abstract

In online multiple testing problem where hypotheses are performed sequentially, controlling the false discovery rate (FDR) is an important challenge for making meaningful inferences. In this chapter, we consider a setting where an ordered (possibly infinite) sequence of hypotheses arrives in a stream, and for each hypothesis we observe a P-value along with a set of features specific to that hypothesis. The decision whether or not to reject the current hypothesis must be made immediately at each timestep, before the next hypothesis is observed. This model provides a general way of leveraging the side (contextual) information in the data to make more discoveries while controlling the FDR. We propose a new class of powerful online testing procedures, where the rejection thresholds are learned sequentially by incorporating contextual information and previous results. We prove that any rule in this class controls the online FDR under some standard assumptions. We then focus on a subclass of these procedures, based on weighting the rejection thresholds, to derive a practical algorithm that learns a parametric weight function in

an online fashion to make more discoveries. We also theoretically prove, in a stylized setting, that our proposed procedures would lead to an increase of statistical power over a popular online testing procedure proposed by [JM18]. Finally, we demonstrate the superior performance of our procedure, by comparing it to the state-of-the-art online multiple testing procedures, on both synthetic data and real data from different applications.

4.2 Introduction

As in the previous chapter, we have briefly reviewed the online testing procedures in the literature. However, most of these procedures take only P-values as input but ignore additional information that is often available in modern applications. In addition to the P-value P_i, each hypothesis H_i could also have a feature vector $X_i \in \mathcal{X} \subseteq \mathbb{R}^d$, which encodes the contextual [1] information related to the tested hypothesis. The feature vector X_i only carries indirect information about the likelihood of the hypothesis H_i to be false, but the relationship is not fully known ahead of time. For example, when conducting an A/B test for a logo size change in a website, contextual information such as text, layouts, images and colors in this specific page can be useful in making a more informative decision. Another example, when testing whether a mutation is correlated with the trait, contextual information about both the mutation such as its location, epigenetic status, etc., and the trait could provide valuable information which can increase the power of these tests.

The problem of using side information has been considered in offline setting. [Sto02] proposed an adaptive FDR-control procedure based on estimating the proportion of true nulls from data. [BH97, GRW06, Dob16, DFKO15] considers reweighting the P-values by applying the prior information. [IKZH16] proposed *Independent Hypothesis Weighting* procedure, which clusters the similar hypotheses into groups and assigns different weights to each group. [HZZ10]

[1] Also sometimes referred to as *prior* or *side* information.

utilized the idea of both grouping and estimating the true null proportions within each group. Some more procedures in [FBC15, GWCT16, LB16a, LF16] incorporate a prior ordering as the auxiliary information to focus on more promising hypotheses near the top of the ordering. *Structure-adaptive Benjamini-Hochberg Algorithm* (SABHA) [LB16b] and *Adaptive P-value Thresholding* (AdaPT) [LF18] are two FDR-controlling adaptive methods which derive decision rules dependent on the feature vectors. The offline testing algorithm mostly related to our results is the *NeuralFDR* procedure proposed by [XZZT17], which uses a neural network to incorporate the side information and parametrize the decision threshold.

In this chapter we still focus on the online setting, where the P-values and contextual features are not available at the onset, and a decision about a hypothesis should be made when it is presented. To the best of our knowledge, this online testing problem has not been considered before. Our main contributions in this chapter are as follows.

1. **Incorporating Contextual Information.** We propose a new broad class of powerful online testing rules, referred to as *contextual generalized alpha-investing* (CGAI) rules, which incorporates the available contextual features in the testing process. Formally, we assume each hypothesis H is characterized by a tuple (P,X) where $P \in (0,1)$ is the P-value, and X is the contextual feature vector from some generic space $\mathcal{X} \subseteq \mathbb{R}^d$. We consider a sequence of hypotheses $(H_1, H_2, \dots)$ that arrive sequentially in a stream at each timestep $t = 1, 2, \dots$, with corresponding $((P_1,X_1),(P_2,X_2),\dots)$. Our testing rule generates a sequence of significance levels $(\alpha_1, \alpha_2, \dots)$ at each time based on previous decisions and contextual information seen so far. The test for each hypothesis H_t takes the form $\mathbb{I}\{P_t \le \alpha_t\}$. Assuming the independence of the P-values, and the mutual independence between the P-values and the contextual features under null hypotheses, we show that any monotone rule from this class controls FDR below a preassigned level at any time. We also show that a variant of FDR (mFDR) can be controlled under a weaker assumption on P-values.

2. **Context Weighting.** We focus on a subclass of CGAI rules, referred to as *context-weighted generalized alpha-investing* (CwGAI) rules, for designing a practical online FDR control procedure. In particular, we take a parametric function $\omega(;\theta)$ with parameters θ, and at time t use $\omega(X_t;\theta)$ as a weight on α_t generated through the GAI rules, with the intuition that larger weights should reflect an increased willingness to reject the null. Since the parameter set θ is unknown, a natural idea here will be to learn it in an online fashion to maximize the number of empirical discoveries. This gives rise to a new class of online testing rules that incorporates the context weights through a learnt parametric function.

3. **Statistical Power Analysis.** We then look into the effect of including the context weight in discovering the true positives. Considering a general model of random weighting, and under the assumption that weights are positively associated with false null hypotheses, we derive a natural sufficient condition under which the weighting improves the power in an online setting, while still controlling the FDR. This is the first result that demonstrates the benefits of appropriate weighting in the online setting. Prior to this such results were only known in the offline setting [GRW06].

4. **A Practical Procedure.** To design a practical online FDR control procedure with good performance, we model the context weight using a parametric function $\omega(;\theta)$ of a neural network (multilayer perceptron), and train it in an online fashion to maximize the number of empirical discoveries. Our experiments on synthetic and real datasets show that our procedure makes substantially more correct decisions compared to state-of-the-art online testing procedures.

The rest of the chapter is organized as follows. In Section 4.3, we review some related online FDR control rules. In Section 4.4, we propose a new broad class of online testing procedures which can incorporate contextual information of the hypothesis, and present our theoretical guarantee of online FDR control for this class. In Section 4.5, we focus on a subclass

of the above one, which is based on using contextual features to weight the rejection thresholds. In Section 4.6, we theoretically show that power can be increased by using weighted procedures in the online setting. In Section 4.7, we design a practical algorithm for online multiple testing with contextual information, and demonstrate its superior performance on synthetic and real datasets.

4.3 Related Online FDR Control Rules

Similar to the setting in Chapter 3, we consider an ordered (possibly infinite) sequence of hypotheses arriving in a stream, denoted by $\mathcal{H} = (H_1, H_2, H_3, \dots)$, and we have to decide at each timestep t whether to reject H_t having only access to previous decisions. Here $H_t \in \{0,1\}$ represents if tth hypothesis is a true *null* ($H_t = 0$) or *alternative* ($H_t = 1$). Each hypothesis is associated with a P-value P_t. A *valid* P-value (P_t) is stochastically larger than the uniform distribution under the true null hypothesis (if $H_t = 0$), i.e.,

$$\Pr[P_t \le u] \le u, \text{ for all } u \in [0,1]. \tag{4.1}$$

The only requirement of the P-values under alternatives is that they should be stochastically smaller than the uniform distribution, such that they can be differentiated from those of the nulls. We still let $\mathcal{H}^0$ index the true null hypotheses and let $\mathcal{H}^1$ index the false null hypotheses in $\mathcal{H}$.

An online multiple testing procedure $\mathcal{R}$, also called the *decision rule*, provides a sequence of significance level $(\alpha_t)_{t=1}^{\infty}$ and makes the corresponding decisions:

$$R_t = \mathbb{I}\{P_t \le \alpha_t\} = \begin{cases} 1 & P_t \le \alpha_t \quad \Rightarrow \text{reject } H_t, \\ 0 & \text{otherwise} \quad \Rightarrow \text{accept } H_t. \end{cases} \tag{4.2}$$

For any time T, denote the first T hypotheses in the stream by $\mathcal{H}(T) = (H_1, \dots, H_T)$.

Using the same notation as in Table 3.1 (by replacing n with T and omitting $\mathcal{R}$ here), the online false discovery proportion and rate till time T are defined as:

$$\text{FDP}(T) := \frac{V(T)}{R(T) \vee 1}, \qquad \text{FDR}(T) := \mathbb{E}[\text{FDP}(T)], \tag{4.3}$$

where $R(T) \vee 1 = \max\{R(T), 1\}$. The online true discovery proportion and rate till time T are defined as:

$$\text{TDP}(T) := \frac{S(T)}{|\mathcal{H}^1(T)| \vee 1}, \qquad \text{TDR}(T) := \mathbb{E}[\text{TDP}(T)]. \tag{4.4}$$

The true discovery rate is also referred to as the *power* of a testing procedure. Note that $\text{TDR}(T) = 1 - \text{FNR}(T)$.

In this context, the goal of an online testing procedure is to provide a sequence of significance levels $(\alpha_t)_{t\in\mathbb{N}}$ such that the online FDR can be controlled at a desired level q at any time $T \in \mathbb{N}$, i.e,

$$\sup_{T\in\mathbb{N}} \text{FDR}(T) \leq q. \tag{4.5}$$

Note that none of these four metrics can be computed without the underlying true labels (ground truth).

A variant of FDR studied in early online multiple testing works [FS08] is the *marginal FDR*, defined as:

$$\text{mFDR}(T)_\eta = \frac{\mathbb{E}[V(T)]}{\mathbb{E}[R(T)] + \eta},$$

with a special case of $\text{mFDR}(T) = \frac{\mathbb{E}[V(T)]}{\mathbb{E}[R(T)]+1}$ when $\eta = 1$. Note that FDR and mFDR can be very different, and controlling mFDR is not equivalent to controlling FDR at a similar level [JM18]. We will also provide a theoretical guarantee on mFDR control in a contextual setting under some weaker assumptions on P-values.

Generalized Alpha-Investing Rules. [FS08] proposed the first class of online multiple testing rules (referred to as alpha-investing rules) to control the mFDR . [AR14] further extended

this class to generalized alpha-investing (GAI) rules, again to control the mFDR.

Let $\mathcal{F}^t = \sigma(R_1, \ldots, R_t)$ denote the sigma-field of decisions till time t. Any GAI rule generates the significance level α_t at each time t based on past decisions of the rule till time $t-1$:

$$\alpha_t = \alpha_t(R_1, \ldots, R_{t-1}), \tag{4.6}$$

which means that $\alpha_t \in \mathcal{F}^{t-1}$.

Specifically, it starts with an initial wealth of $W(0) > 0$, and records the available wealth $W(t)$ over time. At each time t, an amount of ϕ_t, which is the penalty of testing the tth hypothesis at level α_t, will be deducted from the remaining wealth. If the tth hypothesis is rejected, i.e, $R_t = 1$, then an extra wealth of amount ψ_t is rewarded to the current wealth. This can be explicitly stated as:

$$W(0) = w_0, \quad 0 < w_0 < q \tag{4.7}$$

$$W(t) = W(t-1) - \phi_t + R_t \cdot \psi_t, \tag{4.8}$$

where w_0 and the sequences $\alpha_t, \phi_t, \psi_t \in \mathcal{F}^{t-1}$ are user-defined. The wealth $W(t)$ is required to be always non-negative, and thus $\phi_t \le W(t-1)$. Once the wealth ever equals zero, the procedure has to set $\alpha_t = 0$ from then on and hence is not allowed to make any further rejections. An additional restriction is needed for the goal to control the FDR, in that the reward ψ_t has to be bounded whenever a rejection takes place. Formally, the constraints are:

$$\phi_t \le W(t-1), \tag{4.9}$$

$$\psi_t \le \min\{\phi_t + b_t, \frac{\phi_t}{\alpha_t} + b_t - 1\}. \tag{4.10}$$

[JM15, JM18] defined b_t as a user-defined constant $b_0 = q - w_0$ and proved the FDR control for monotone GAI rules under the independence of P-values. The monotonicity of a rule is defined

as:

$$\text{If } \tilde{R}_i \le R_i \text{ for all } i \le t-1, \text{ then } \alpha_t(\tilde{R}_1, \ldots, \tilde{R}_{t-1}) \le \alpha_t(R_1, \ldots, R_{t-1}). \tag{4.11}$$

Recently, [RYWJ17] demonstrated that setting

$$b_t = q - w_0 \mathbb{I}\{\rho_1 > t-1\}$$

could potentially lead to larger statistical power. Here, ρ_k is the time of kth rejection (discovery). [RYWJ17] refer to this class of rules as GAI++ rules. Unless otherwise specified, we use this improved setting of b_t throughout this chapter.

Level based On Recent Discovery (LORD) Rules. Here we revisit the LORD rules which have been analyzed in Chapter 3, as it is a subclass of the GAI rules that represents the current state-of-the-art in this area. There are several variants of the LORD rules. The one we consider here is consistent with the definition of the GAI rule, thus is slightly different from that in Chapter 3. Formally, we choose any sequence of non-increasing nonnegative constants $\gamma = (\gamma_t)_{t=1}^{\infty}$ with $\sum_{t=1}^{\infty} \gamma_t = 1$. At each time t, let τ_t be most recent discovery time before t, i.e.,

$$\tau_t := \max\{i \in \{1, \ldots, t-1\} : R_i = 1\},$$

with $\tau_t = 0$ for all t before the first discovery. The LORD rule defines α_t, ϕ_t, ψ_t as follows:

$$\begin{aligned}
&W(0) = w_0, \\
&\phi_t = \alpha_t = \begin{cases} \gamma_t w_0 & \text{if } t \le \rho_1 \\ \gamma_{t-\tau_t} b_0 & \text{if } t > \rho_1, \end{cases} \\
&\psi_t = b_0.
\end{aligned}$$

Throughout this chapter, we stick to one version (though much of the discussion in this chapter also holds for the other versions), and we set $b_0 = w_0 = q/2$, in which case, the above rule could be simplified as $\phi_t = \alpha_t = \gamma_{t-\tau_t} b_0$.

As with any GAI rule, [RYWJ17] defined the LORD++ by replaceing b_0 with $b_t = q - w_0 \mathbb{I}\{\rho_1 > t-1\}$ and showed it achieves a power increase while still controlling the online FDR at same level q. It can be easily observed that both LORD and LORD++ rules satisfy the monotonicity condition from (4.11).

SAFFRON Procedure. This is a very recently proposed procedure by [RZWJ18]. The main difference between SAFFRON (Serial estimate of the Alpha Fraction that is Futilely Rationed On true Null hypotheses) and the previously discussed LORD/LORD++ procedures comes in that SAFFRON is an adaptive method, based on adaptively estimating the proportion of true nulls. SAFFRON can be viewed as an online extension of Storey's adaptive version of the BH procedure in the offline setting. SAFFRON does not belong to the GAI class, whose extension to the contextual online setting is the main focus of this chapter.

4.4 Contextual Online FDR Control

While these online FDR-controlling procedures are widely used, a major shortcoming of them is that they ignore side (prior) information which is often available during the testing process. For example, in genetic association studies, each hypothesis tests the correlation between a gene variant and a trait. In addition to the P-values returned from the tests, we also have contextual features for each variant (e.g. its location, epigenetics etc.) and trait which could inform how likely the variant is truly associated with a trait.

To deal with such situation, we now assume that a P-value $P_t \in (0,1)$ and a vector of contextual features $X_t \in \mathcal{X} \subseteq \mathbb{R}^d$ are observed for each hypothesis H_t. At each step t, we have to decide whether to reject H_t having access to previous decisions and contextual information

seen so far. The overall goal is to control the online FDR under a given level q at any time, and improve the number of true discoveries by using the contextual information.

Under the alternative, we denote the density distribution (PDF) of P-values as $f_1(p \mid X)$ (depending on the feature vector $X \in \mathcal{X}$) and the corresponding cumulative distribution (CDF) as $F_1(p \mid X)$. Here $f_1(p \mid X)$ can be any arbitrary unknown positive function, as long as the P-values are stochastically smaller than those under the null. Note that $f_1(p \mid X)$ is not identifiable from the data as we never observe H_t's directly. This can be illustrated through a simple example described in Section 4.8.

Definition 2 (Contextual Online FDR Control Problem). Given a (possibly infinite) sequence of $(P_t, X_t)_{t\in\mathbb{N}}$ where $P_t \in (0,1)$ and $X_t \in \mathcal{X}$, the goal is to generate a significance levels$(\alpha_t)_{t\in\mathbb{N}}$ as a function of prior decisions and contextual features

$$\alpha_t = \alpha_t(R_1, \ldots, R_{t-1}, X_1, \ldots, X_t), \tag{4.12}$$

and a corresponding set of decisions

$$R_t = \mathbb{I}\{P_t \leq \alpha_t(R_1, \ldots, R_{t-1}, X_1, \ldots, X_t)\} \tag{4.13}$$

such that $\sup_T \ \mathrm{FDR}(T) \leq q$.

We now define a contextual extension of GAI rules, that we refer to as *Contextual Generalized Alpha-Investing* (contextual GAI or CGAI) rules. In the presence of contextual information, we consider the sigma-field of decisions till time t as $\mathcal{F}^t = \sigma(R_1, \ldots, R_t)$, and the sigma-field of features till time t as $\mathcal{G}^t = \sigma(X_1, \ldots, X_t)$. A contextual GAI rule is defined through three functions, $\alpha_t, \phi_t, \psi_t \in \sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t)$, that are all computable at time t, with the same conditions (4.7), (4.8), (4.9), (4.10) satisfied. We set $b_t = q - w_0 \mathbb{I}\{\rho_1 > t-1\}$ as proposed by [RYWJ17].

Similar to GAI rules, we consider a monotonicity property to a contextual GAI rule as

follows:

Monotoncity: If $\tilde{R}_i \le R_i$ for all $i \le t-1$, then

$$\alpha_t(\tilde{R}_1,\ldots,\tilde{R}_{t-1},X_1,\ldots,X_t) \le \alpha_t(R_1,\ldots,R_{t-1},X_1,\ldots,X_t), \text{ for any fixed } \mathbf{X}^t = (X_1,\ldots,X_t). \quad (4.14)$$

A contextual GAI rule satisfying the monotonicity condition is referred to as a *monotone contextual GAI* rule.

Our first result establishes the online FDR control for any monotone contextual GAI rule under some standard conditions.

We start by presenting the following lemma, which is an intermediate result for the proof of FDR later. Recall that $R_t = \mathbb{I}\{P_t \le \alpha_t\}$, where $\alpha_t \in \sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t)$ is a coordinatewise non-decreasing function of $R_1,\ldots,R_{t-1}$ for any fixed $\mathbf{X}^t = (X_1,\ldots,X_t)$. Due to the marginal super-uniformity (4.1), we immediately have that for independent P-values under the null

$$\Pr[P_t \le \alpha_t \mid \sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t)] \le \alpha_t. \quad (4.15)$$

Lemma 5 states a more general result about super-uniformity of independent P-values under the null. The proof is based on a *leave-one-out* technique which is common in the multiple testing. A variant of this result was also used by [RYWJ17] and [JM18] in their analysis of GAI rules. The main distinction for us comes in that we consider the sigma-field at each time t as $\sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t)$ including the information of contextual features till time t, instead of just $\mathcal{F}^{t-1}$.

Lemma 5 (Super-uniformity)**.** *Let $g : \{0,1\}^T \to \mathbb{R}$ be any coordinatewise non-decreasing function such that $g(\mathbf{R}) > 0$ for any vector $\mathbf{R} \ne (0,\ldots,0)$. Then for any $t \le T$ such that $t \in \mathcal{H}^0$, we have*

$$\mathbb{E}\left[\frac{\mathbb{I}\{P_t \le \alpha_t(R_1,\ldots,R_{t-1},X_1,\ldots,X_t)\}}{g(R_1,\ldots,R_T)\vee 1}\middle|\sigma(\mathcal{F}^{t-1}\cup\mathcal{G}^t)\right]$$
$$\le \mathbb{E}\left[\frac{\alpha_t(R_1,\ldots,R_{t-1},X_1,\ldots,X_T)}{g(R_1,\ldots,R_T)\vee 1}\middle|\sigma(\mathcal{F}^{t-1}\cup\mathcal{G}^t)\right].$$

The following theorem establishes the FDR control for any monotone contextual GAI rule under the independence among all P-values, and between P-values and contextual features under the null hypotheses.

Theorem 11 (FDR Control)**.** *Consider a sequence of* $(P_t, X_t)_{t\in\mathbb{N}}$ *of P-values and contextual features. If the P-values* P_t*'s are independent, and additionally* P_t*'s are independent of all* $(X_t)_{t\in\mathbb{N}}$ *under the null, then for any monotone contextual generalized alpha-investing rule, we have the online FDR control at any time, i.e.,*

$$\sup_{T\in\mathbb{N}} \mathrm{FDR}(T) \leq q. \tag{4.16}$$

Note that a standard assumption in hypothesis testing is that the P-values under the null are uniformly distributed in $(0,1)$, which does not depend on the contextual features. That means the mutual independence of P-values and contextual features is valid under such standard case.

Turning to the mFDR, we can also prove a guarantee for the mFDR control under a weaker condition than that in Theorem 11. In particular, one can relax the independence assumptions to a weaker conditional super-uniformity assumption. Our next theorem proves mFDR control for any contextual GAI rule (not necessarily monotone) under this weaker condition.

Theorem 12 (mFDR Control)**.** *Under the same setting as Theorem 11, if the P-values* P_t*'s are super-uniform conditional on all past discoveries and contextual features so far, meaning that* (4.15) *is satisfied, then for any contextual generalized alpha-investing rule, we have the online mFDR control,*

$$\sup_{T\in\mathbb{N}} \mathrm{mFDR}(T) \leq q. \tag{4.17}$$

Remark 19. For arbitrary dependent P-values and contextual features, the FDR control can also be obtained by using a modified LORD rule defined in [JM18], under a special case where the contextual features are transformed into weights satisfying some conditions. See Section 4.6 for more details about weakening the assumptions.

4.5 Context-weighted Generalized Alpha-Investing Rules

The contextual GAI rules form a very general class of online multiple testing rules. In this section, we focus on a subclass of these rules, which we refer to as *Context-weighted Generalized Alpha-Investing* (context-weighted GAI or CwGAI) rules. Specifically, it considers α_t to be a product of two functions with the first one of previous decisions and second one based on the current contextual feature,

$$\alpha_t = \tilde{\alpha}_t(R_1, \ldots, R_{t-1}) \cdot \omega(X_t; \theta), \tag{4.18}$$

where $\omega(X_t; \theta)$ is a parametric *weight function* with parameters $\theta \in \Theta$. Since CwGAI is a subclass of CGAI rules, the above FDR and mFDR control theorems are valid for this class too. Applying this idea of context-weighting to the LORD++ (resp. LORD) give rise to a new class of testing procedures that we refer to as CwLORD++ (resp. CwLORD).

Our reasons for considering this subclass include: 1. We obtain a simpler form of α_t by separating the contextual features from that of previous outcomes, making it easier to design functions that satisfy the monotonicity requirement of the GAI rules. 2. It is convenient to model the weight function by any parametric function, and 3. we can learn the parameters of the weight function empirically by maximizing the number of discoveries. This forms the basis of a practical algorithm for the contextual online FDR control that we describe in Section 4.7.1. Note that the GAI rules are context-weighted GAI rules when the weight function equals 1. We illustrate the relationship among various classes of testing rules in Figure 4.1.

The idea of weighting P-values using prior information has been widely studied in offline multiple testing setup [GRW06, IKZH16, LB16b, LF18, XZZT17, RBWJ17]. Intuitively, the weights represent the extent of a prior belief about whether the hypothesis is a true null or not. A larger weight $\omega_t > 1$ provides more belief of a hypothesis being an alternative which makes the procedure to reject it more aggressively, while a smaller weight $\omega_t < 1$ indicates a higher

likelihood of a true null which makes the procedure reject it more conservatively.

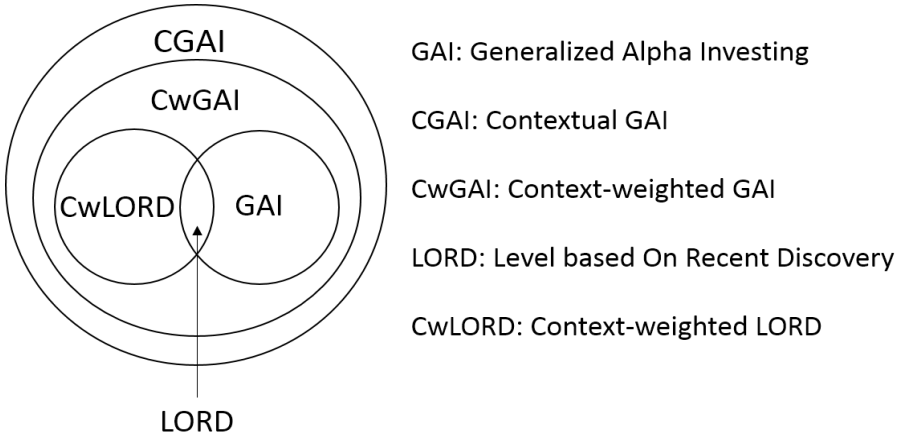

Figure 4.1: Relationship among various testing rules. One could replace LORD with LORD++ (CwLORD with CwLORD++).

Weighting in Online vs. Offline Setting with FDR Control. In the offline setting, prior weights are usually rescaled to have unit mean, and then existing offline FDR-controlling algorithm is applied to the weighted P-values P_i/ω_i instead of P_i [GRW06]. However, in the online setting, the weights are computed at each timestep without knowing the total number of hypothesis or contextual information, thus cannot be rescaled to have unit mean in advance. Instead, as represented in (4.18), we consider weighting the significance levels α_t's, as was also used by [RBWJ17]. Note that weighting P-values is equivalent to weighting significance levels in terms of the decision rules with the same significance levels, i.e., given the same α_t's, we have $\{P_t/\omega_t \le \alpha_t\} \equiv \{P_t \le \alpha_t\omega_t\}$ for all t. The subtle difference is that when α_t's are weighted, the penalty ϕ_t's and rewards ψ_t's are also adjusted according to the GAI constraints. For example, as dictated by (4.10), if we overstate our prior belief in the hypothesis being alternative by assigning a large $\omega_t > 1$, the penalty will need to be more or the reward will need to be less.

Context-weighted LORD++. We then apply the context weighting to the LORD++ rules. Given a weight function, $\omega : \mathcal{X} \times \Theta \to \mathbb{R}$, we define the context-weighted LORD++ (CwLORD++)

testing rule as follows.

$$
\begin{aligned}
&W(0) = w_0,\\
&\phi_t = \alpha_t = \min\{\gamma_{t-\tau_t} b_t \cdot \omega(X_t;\theta), W(t-1)\},\\
&\psi_t = b_t = q - w_0 \mathbb{I}\{\rho_1 > t-1\}.
\end{aligned}
$$

We can see here when α_t's are weighted, and the penalty ϕ_t's are also adjusted accordingly.

4.6 Statistical Power of Weighted Online Rules

In this section, we would like to present theoretical support of power increased through proper weighting in the online setting. Missing details are collected in Section 4.9.

The benefits of weighting in terms of increasing the power was first studied by [GRW06] in the offline setting, who showed that a weighted BH procedure improves the power over the unweighted BH procedure if the weighting is *informative*, which roughly means that the weights are positively associated with the alternative hypotheses.

We consider a mixture model where each null hypothesis is false with a fixed probability π_1, and the P-values corresponding to different hypotheses are all independent. While the mixture model is *idealized*, it does offer a natural ground for comparing the power of various testing procedures [GRW06, JM18].

Mixture Model. For any $t \in \mathbb{N}$, let

$$
\begin{aligned}
H_1, \ldots, H_t &\overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\pi_1),\\
X_1, \ldots, X_t &\overset{\text{i.i.d.}}{\sim} \mathcal{L}(\mathcal{X}),\\
P_t \mid (H_t = 0, X_t) &\sim \text{Uniform}(0,1),\\
P_t \mid (H_t = 1, X_t) &\sim F_1(p \mid X_t).
\end{aligned}
$$

where $0 < \pi_1 < 1$ and $\mathcal{L}(\mathcal{X})$ is a probability distribution on the contextual feature space $\mathcal{X}$. Let $F = \int F_1(p \mid X) \mathrm{d}\mathcal{L}(\mathcal{X})$ be the marginal distribution of the P-values under alternative, which we assume is stochastically smaller than the uniform distribution. Marginally, the P-values are i.i.d. from the CDF $G(a) = (1-\pi_1)U(a) + \pi_1 F(a)$, where $U(a)$ is the CDF of Uniform(0,1).

We consider the general weighting as in [GRW06] where weight ω is a random variable and conditionally independent of P given H. We assume that weight has different marginal distributions under null and alternative,

$$\omega \mid H_t = 0 \sim Q_0, \quad \omega \mid H_t = 1 \sim Q_1, \tag{4.19}$$

for some unknown probability distributions Q_0, Q_1 that are continuous on $(0, \infty)$. Note that this framework of weights is very general. It includes the contextual weighting scenario as a special case, where we assume two different implicit weight functions of contextual features X as $\omega_0(X)$ and $\omega_1(X)$, and the marginal distributions of weights can be computed as:

$$\omega \mid H_t = 0 \sim \int \omega_0(X) \mathrm{d}\mathcal{L}(\mathcal{X}) := Q_0,$$
$$\omega \mid H_t = 1 \sim \int \omega_1(X) \mathrm{d}\mathcal{L}(\mathcal{X}) := Q_1.$$

Under (4.19), the marginal distribution of ω is $Q = (1-\pi_1)Q_0 + \pi_1 Q_1$. For $j = 0, 1$, let $u_j = \mathbb{E}[\omega \mid H_t = j]$ be the means of Q_0 and Q_1 respectively. We also assume that the weighting is *informative*, based on the definition from [GRW06] in the offline setting,

$$\textbf{Informative-weighting: } u_0 < 1, u_1 > 1, u = \mathbb{E}[\omega] = (1-\pi_1)u_0 + \pi_1 u_1 = 1. \tag{4.20}$$

Remark 20. We would like to point that $\mathbb{E}[\omega] = 1$ is an assumption on the weight distribution that may not hold in a practical setting. So for empirical experiments, we will use an instantiation of the CwLORD++ (see Section 4.7.1) that does not require this assumption. But for the theoretical

comparison of the power of different procedures, we make this assumption, under which we can use the P-value reweighting akin to the offline setting.

Let the weights ω_t's be random variables that are i.i.d. from this mixture model with marginal distribution Q. Taking the LORD++ rule from Section 4.3, we define a weighted LORD++ rule as follows.

Definition 3 (Weighted LORD++). Given a sequence of P-values, $(P_1, P_2, \dots)$ and weights $(\omega_1, \omega_2, \dots)$, apply the LORD++ with level q to the weighted P-values $(P_1/\omega_1, P_2/\omega_2, \dots)$.

We want to emphasize that the weighted LORD++ rule actually reweights the P-values (as done in the offline weighted BH procedure [GRW06]) and then applies the original LORD++ to these reweighted P-values. So this is slightly different from the idea of CwLORD++, which reweights the significance levels and then applies it to the original P-values. To understand the difference, we look at their definitions.

In LORD++, the penalty is $\phi_t = \alpha_t = \gamma_{t-\tau_t} b_t < W(t-1)$, which is always less than current wealth due to the construction of γ_t's and w_0. So in weighted LORD++, each weighted P-value P'_t is tested at level $\alpha_t = \gamma_{t-\tau_t} b_t$, which means that the original P-value P_t is tested at level $\alpha'_t = \gamma_{t-\tau_t} b_t \omega_t$. On the other hand, in CwLORD++, we take a penalty of the form $\hat{\phi}_t = \hat{\alpha}_t = \min\{\gamma_{t-\tau_t} b_t \omega_t, W(t-1)\}$. We take the minimum of reweighted significance level and the current wealth, to prevent the penalty $\gamma_{t-\tau_t} b_t \omega_t$ from exceeding the current wealth which would violate a tenet of the alpha-investing rules.

It is obvious that the actual significance levels used in weighted LORD++ are greater than those in CwLORD++, i.e.,

$$\alpha'_t = \gamma_{t-\tau_t} b_t \omega_t \geq \min\{\gamma_{t-\tau_t} b_t \omega_t, W(t-1)\} = \hat{\alpha}_t.$$

That implies the power of weighted LORD++ is equal to or greater than the power of CwLORD++, whereas the FDR of weighted LORD++ may also be higher than that of CwLORD++. From

Theorem 11, we know that we have the FDR control with the CwLORD++, however the condition does not hold for the weighted LORD++ (as weighted LORD++ is not strictly a contextual GAI rule). We now show that the above weighted LORD++ can still control the online FDR at any given level q under the condition $\mathbb{E}[\omega] = 1$, which we do in the following proposition.

Proposition 3. *Suppose that the weight distribution satisfies the informative-weighting assumption in* (4.20)*. Suppose that P-values P_t's are independent, and are conditionally independent of the weights ω_t's given H_t's. Then the weighted LORD++ rule can control the online FDR at any given level q, i.e.,*

$$\sup_{T \in \mathbb{N}} \mathrm{FDR}(T) \leq q. \tag{4.21}$$

Weakening the Assumptions. In most applications, the independence between P-values and weights needed in Proposition 3 is not guaranteed. [JM18] defined a modified LORD rule to achieve the FDR control under dependent P-values. We would like to extend the FDR control results to the dependent weighed P-values. In particular, as long as the following condition is satisfied, i.e., for each weighted P-value P_t/ω_t marginally, we have

$$\Pr[P_t/\omega_t \leq u \,|\, H_t = 0] \leq u, \quad \text{for all } u \in [0,1], \tag{4.22}$$

then the upper bound of the FDR stated in Theorem 3.7 in [JM18] is valid for weighted P-values.

Specifically, if the modified LORD rule in Example 3.8 of [JM18] is applied to the weighted P-values under the assumption in (4.22), then the FDR can also be controlled below level q. The proofs of this extension are almost the same as those in [JM18] and are omitted here.

Comparison of Power. Next, we establish conditions under which the weighting leads to increased power for LORD. We start with a particular version of LORD from [JM18], denoted as **LORD***, which tightly lower bounds the average power of any versions of LORD (LORD++)

under the mixture model,

$$\textbf{LORD*: } W(0) = w_0 = b_0 = q/2,\ \phi_t = \alpha_t = b_0\gamma_{t-\tau_t},\ \psi_t = b_0. \tag{4.23}$$

As shown by [JM18], the average power of the LORD* almost surely equals

$$\liminf_{T\to\infty} \mathrm{TDP}(T) = \Big(\sum_{m=1}^{\infty}\prod_{j=1}^{m}(1-G(b_0\gamma_j))\Big)^{-1}, \tag{4.24}$$

where $G(a) = (1-\pi_1)U(a) + \pi_1 F(a)$ as defined earlier.

We then analyze the power of the weighted LORD++. Define $D(a) = \Pr[P/\omega \le a]$ as the marginal distribution of weighted P-values. Under the assumptions on the weight distribution from (4.19), the marginal distribution of weighted P-value equals,

$$\begin{aligned}
D(a) &= \Pr[P/\omega \le a] = \int \Pr[P/\omega \le a \mid \omega = w]\,\mathrm{d}Q(w)\\
&= \int \sum_{h\in\{0,1\}} \Pr[P/\omega \le a \mid \omega = w, H = h] g(h \mid w)\,\mathrm{d}Q(w)\\
&= \int \sum_{h\in\{0,1\}} \Pr[P/w \le a \mid H = h] g(h \mid w)\mathrm{d}Q(w)\\
&= \int \sum_{h\in\{0,1\}} ((1-h)aw + hF(aw)) g(h \mid w)\,\mathrm{d}Q(w)\\
&= \int \sum_{h\in\{0,1\}} ((1-h)aw + hF(aw))\,\mathrm{d}Q(w \mid h) g(h)\\
&= \sum_{h\in\{0,1\}} \int ((1-h)aw + hF(aw))\,\mathrm{d}Q(w \mid h) g(h)\\
&= (1-\pi_1)\int aw\mathrm{d}Q(w|h=0) + \pi_1 \int F(aw)\,\mathrm{d}Q(w|h=1)\\
&= (1-\pi_1)\mu_0 a + \pi_1 \int F(aw)\mathrm{d}Q_1(w).
\end{aligned} \tag{4.25}$$

Theorem 13. *Let $D(a) = \Pr[P/\omega \le a]$ be the marginal distribution of weighted P-values as in* (4.25)*. Then, the average power of the weighted LORD++ rule is almost surely bounded as*

follows:

$$\liminf_{T\to\infty} \mathrm{TDP}(T) \geq \Big(\sum_{m=1}^{\infty}\prod_{j=1}^{m}(1-D(b_0\gamma_j))\Big)^{-1}. \tag{4.26}$$

The proof of Theorem 13 uses the similar technique as the proof of the statistical power of LORD in [JM18]. The main distinction is that we replace the marginal distribution of P-values by the marginal distribution of weighted P-values.

Similarly, we define the weighted LORD* as follows,

Definition 4 (Weighted LORD*). Given a sequence of P-values, $(P_1, P_2, \dots)$ and weights $(\omega_1, \omega_2, \dots)$, apply the LORD* with level q to the weighted P-values $(P_1/\omega_1, P_2/\omega_2, \dots)$.

From the proof of Theorem 13, the average power of the weighted LORD* achieves the lower bound and almost surely equals

$$\liminf_{T\to\infty} \mathrm{TDP}(T) = \sum_{m=1}^{\infty}\prod_{j=1}^{m}(1-D(b_0\gamma_j))\Big)^{-1}. \tag{4.27}$$

Assume the CDF of the P-values under alternative F is differentiable and let $f = F'$ be the PDF. Due to the fact that P-values under alternative are stochastically dominated by uniform distribution, we assume that there exists some $a_0 > 0$ such that $f(a) > 1$ for all $0 \leq a < a_0$. The following theorem is based on comparing this power on weighted LORD* from (4.27) with the power on LORD* from (4.24).

Theorem 14 (Power Comparison)**.** *Suppose that the parameters in LORD** (4.23) *satisfy* $b_0\gamma_1 < a_0$*, and the weight distribution satisfies* $\Pr[\omega < a_0/(b_0\gamma_1) \mid H_t = 1] = 1$ *for every* $t \in \mathbb{N}$ *and the informative-weighting assumption in* (4.20)*. Then, the average power of weighted LORD* is greater than or equal to that of LORD* almost surely, i.e, the power lower bound of weighted LORD (LORD++) is greater than or equal to that of LORD (LORD++) almost surely.*

Remark 21. Intuitively, the condition implies that to achieve higher power while controlling the FDR, the weights given to alternative hypotheses cannot be too large. Let us discuss this point in

the context of weighted LORD++ and context-weighted LORD++.

In weighted LORD++, we can always assign large weights to make the reweighted P-values small enough to be rejected, in order to achieve high power. But this can lead to a loss in the FDR control which is why we need the restriction of $\mathbb{E}[\omega] = 1$ for proving the FDR control in Proposition 3. Therefore, it is natural to have weights not too large.

If we consider reweighting the significance levels as in CwLORD++, assigning a large weight will not affect the FDR control (we prove that the FDR is controlled for any choice of weights in Theorem 11). However, the price is paid in terms of power. When we use a large weight, the penalty ϕ_t increases and therefore the wealth might go quickly down to zero. Once the wealth is exhausted, the significance levels afterwards must all be zero and thus preventing any further discoveries.

To further interpret the weight condition in Theorem 14, let us take a concrete example and consider the hypotheses $(H_1, \ldots, H_T)$ concerning the means of normal distributions (referred to as *normal means model*). This model corresponds to getting test statistics $Z_t \sim \mathcal{N}(\mu, 1)$. So the two-sided P-values are $P_t = 2\Phi(-|Z_t|)$, where Φ is the CDF of standard normal distribution. Suppose under the null hypothesis $\mu = 0$, and under the alternative hypothesis $0 < \mu \leq 4$. Then from simple computation we obtain that $a_0 > 0.022$ for any μ such that $0 < \mu \leq 4$. In fact, a_0 increases as μ decreases. Typically, we set $q = 0.05$, so $b_0 = q/2 = 0.025$. Using the sequence of hyperparameters $\{\gamma_t\}_{t\in\mathbb{N}}$ (where $\gamma_t = 0.0722\log(t \vee 2)/(t\exp(\sqrt{\log t}))$) as suggested by [JM18] for this normal means model, we compute that $\gamma_1 \approx 0.117$ when the number of total hypotheses $T = 10^5$. Therefore, we get $a_0/(b_0\gamma_1) \approx 7.52 > 1$. As long as the weight ω is bounded by 7.52 with probability 1 under the alternative hypotheses, the condition needed for Theorem 14 is satisfied.

4.7 Experiments with Context-weighted GAI

In this section, we propose a practical procedure for the contextual online FDR control based on context-weighted GAI rules, which sets $\alpha_t(R_1,\ldots,R_{t-1},X_1,\ldots,X_t) = \tilde{\alpha}_t(R_1,\ldots,R_{t-1})\cdot \omega(X_t;\theta)$, and present numerical experiments to illustrate the performance with this procedure. Technically, we can use any parametric function $\omega(X_t;\theta)$ (with parameter set $\theta \in \Theta$) to model the weight function. In this chapter, we choose a deep neural network (multilayer perceptron) due to its expressive power, as noted in a recent offline FDR control result by [XZZT17]. Given this, a natural goal will be to find $\theta \in \Theta$ that maximizes the number of empirical discoveries (or discovery rate), while controlling the FDR. Note that if the function $\alpha_t(R_1,\ldots,R_{t-1})$ is monotone (such as with LORD or LORD++) with respect to the R_i's, the function $\alpha_t(R_1,\ldots,R_{t-1})\cdot\omega(X_t;\theta)$ is also monotone with respect to the R_i's.

Given a stream $((P_t,X_t))_{t\in\mathbb{N}}$, our algorithm (see Algorithm 1) processes the stream in batches, in a single pass. Let $b \geq 1$ denote the batch size. Let θ_j be the parameter obtained before batch j is processed, thus θ_j is only based on all previous P-values and contextual features which are assumed to be independent of all future batches. For each batch, the algorithm fixes the parameters to compute the significance levels for hypothesis in that batch, so the FDR control is guaranteed by our theory (under appropriate assumptions) for each batch and thus for the whole stream. Define, the Empirical Discovery Rate for batch j as follows:

$$\mathrm{EDR}_j = \frac{\sum_{i=jb+1}^{(j+1)b} \mathbb{I}\{P_i \leq \alpha_i(X_i;\theta)\}}{b}. \tag{4.28}$$

Since the above function is not differentiable, we use the sigmoid function σ to approximate the indicator function, and define

$$\mathrm{EDR}_j = \frac{\sum_{i=jb+1}^{(j+1)b} \sigma(\lambda(\alpha_i(X_i;\theta) - P_i))}{b}. \tag{4.29}$$

Algorithm 1: Online FDR Control using a Context-Weighted GAI Procedure

Model the weight function as a multi-layer perceptron (MLP);

Input: A sequence of P-value, contextual feature vector pairs $((P_1, X_1), (P_2, X_2), \dots)$, a monotone GAI rule (such as LORD++) denoted by GAI with desired FDR control level q, batch size b, learning rate η

Output: Neural network model parameter set $\theta \in \Theta$

Randomly initialize the parameter set θ_0; batch index $j = 0$

repeat

  for $i = 1$ *to* b **do**

    Consider the pair (P_{jb+i}, X_{jb+i}) (ith entry in the jth batch)

    Let $\tilde{\alpha} \leftarrow \alpha_{jb+i}(R_1, \dots, R_{jb+i-1})$ (computed based on GAI)

    Apply GAI on P_{jb+i} with significance level $\tilde{\alpha} \cdot \omega(X_{jb+i}; \theta_j)$

  end

  Use the decisions in the jth batch to update the empirical discovery proportion (EDR_j)

  Compute the gradient with respect to the parameter set $\frac{\partial \mathrm{EDR}_j}{\partial \theta}$

  Update the parameter set: $\theta_{j+1} \leftarrow \theta_j + \eta \frac{\partial \mathrm{EDR}_j}{\partial \theta}$

  $j \leftarrow j + 1$

until *convergence or end-of-stream*;

Return θ_j

Here λ is a large positive hyperparameter. With this, the parameter set θ can now be optimized by using standard gradient methods in an online fashion. Note that we are only maximizing empirical discovery rate subject to the empirical FDR control, and the training does not require any ground truth labels on the hypothesis. In fact, the intuition behind Algorithm 1 is very similar to the *policy gradient descent* method which is popular in reinforcement learning, which aims to find the best policy that optimizes the rewards. In an online multiple testing problem, we can regard the number of discoveries as reward and parametric functions as policies.

In all our experiments, we use a multilayer perceptron to model the weight function, which is constructed by 10 layers and 10 nodes with ReLU as the activation function in each layer, and exponential function of the output layer, since the weight has to be non-negative. In the following, we use the context-weighted LORD++ (CwLORD++) to denote the testing rule obtained from Algorithm 1 by using the LORD++ as the monotone GAI rule.

4.7.1 Experiments

In this section, we present results for numerical experiments with both synthetic and real data to compare the performance of our proposed CwLORD++ testing rule with the current state-of-the-art online testing rule LORD++ [RYWJ17]. The synthetic data experiments are based on the normal means model, which is commonly used in hypothesis testing literature. Our real data experiments focus on a diabetes prediction problem and gene expression data analyses. The primary goal of this section is to show the increased power (under FDR control) in online multiple testing setup that can be obtained with our proposed contextual weighting in many real world problems.

Synthetic Data Experiments

For the synthetic data experiments, we consider the hypotheses $\mathcal{H}(T) = (H_1, \ldots, H_T)$ coming from the normal means model. The setup is as follows: for $t \in [T] := \{1, \ldots, T\}$, under the null hypothesis, $H_t : \mu_t = 0$, versus under the alternative, $\mu_t = \mu(X_t)$ is a function of X_t. We observe test statistics $Z_t = \mu_t + \varepsilon_t$, where ε_t's are independent standard normal random variables, and thus the two-sided P-values are $P_t = 2\Phi(-|Z_t|)$. For simplicity, we consider a linear function $\mu(X_t) = \langle \beta, X_t \rangle$ for β unknown to the testing setup. We choose the dimension of the features X_t's as $d = 10$ in all following experiments.

We set the total number of hypotheses as $T = 10^5$. We generate each d-dimensional vector X_t i.i.d. from $\mathcal{N}(0, \sigma^2 I_d)$ with $\sigma^2 = 2\log T$. The choice of the σ^2 is to put the signals in a detectable (but not easy) region. This is because under the global null hypothesis where $Z_t \sim \mathcal{N}(0,1)$ for all $t = 1, \ldots, T$, we have $\max_{t \in [T]} Z_t \sim \sqrt{2\log T}$ with high probability. Here, β is a deterministic parameter vector of dimension $d = 10$, we generate the ith coordinate in β as $\beta_i \sim \text{Uniform}(-2, 2)$ and fix β throughout the following experiments. Let π_i denote the fraction of non-null hypotheses. For LORD++, we choose the sequence of hyperparameters $\{\gamma_t\}$ (where $\gamma_t = 0.0722 \log(t \vee 2)/(t \exp(\sqrt{\log t}))$) as suggested by [JM18].

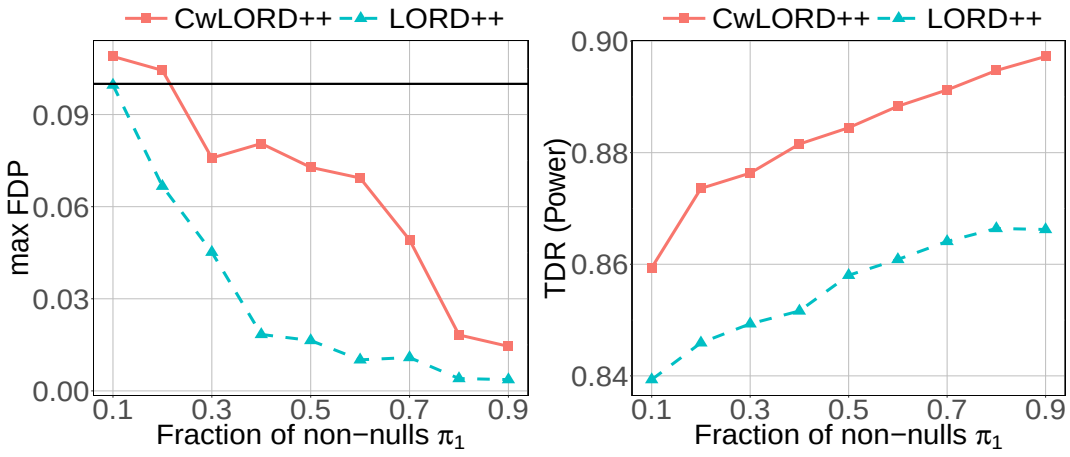

Figure 4.2: Simulation results showing average of max FDP and TDR (power) for our proposed CwLORD++ and LORD++ as we vary the fraction of non-nulls (π_1) under the normal means model. The nominal FDR control level $q = 0.1$.

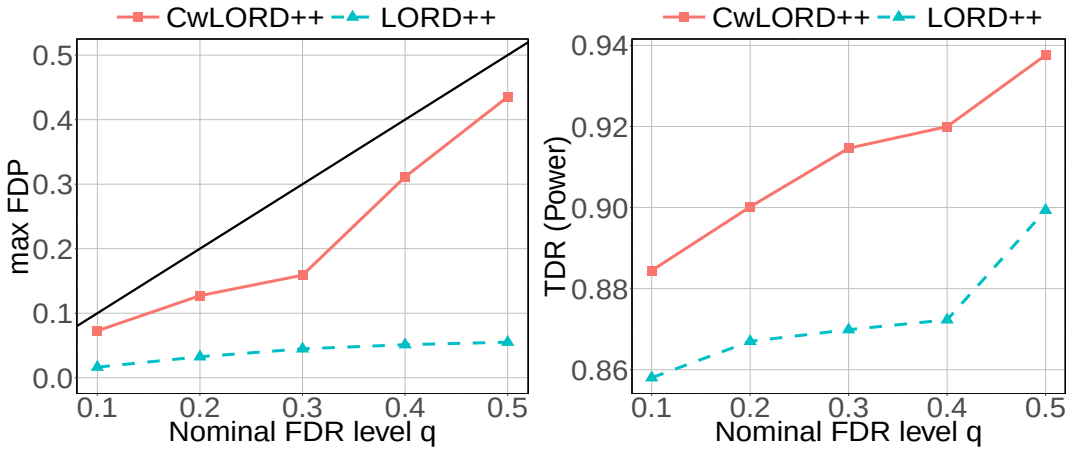

Figure 4.3: Simulation results showing average of max FDP and TDR (power) for our proposed CwLORD++ and LORD++ as we vary the nominal FDR levels (q) under the normal means model. The fraction of non-nulls is set $\pi_1 = 0.5$.

In Figure 4.2 and 4.3, we report the maximum FDP and the statistical power of the two compared procedures as we vary the fraction of non-nulls π_1 and desired level q. The average is taken over 20 repeats.

In the first set of experiments (Figures 4.2), we set $q = 0.1$, and vary the fraction of non-nulls π_1 from 0.1 to 0.9. We can see that FDP of both rules (CwLORD++ and LORD++) are almost always under the set level $q = 0.1$ and are decreasing with the increasing fraction of

non-nulls. As expected, the power increases with increasing π_1, and the power of CwLORD++ uniformly dominates that of LORD++.

Note that we take the average of maximum FDP over 20 repeats, which is an estimate for $\mathbb{E}[\sup \mathrm{FDP}]$. Due to the fact that $\mathbb{E}[\sup \mathrm{FDP}] \geq \sup \mathbb{E}[\mathrm{FDP}] = \sup \mathrm{FDR}$, the reported average of maximum FDP is probably higher than the true maximum FDR. In Figure 4.2, we see that the average of maximum FDP is almost always controlled under the black line, which means the true maximum FDR should be even lower than that level. When π_1 is really small (like 0.1), the number of non-nulls is too sparse to make a high proportion of true discoveries, which leads to a higher average of maximum FDP that also has a higher variance in Figure 4.2.

In the second set of experiments (Figures 4.3), we vary the nominal FDR level q from 0.1 to 0.5. The fraction of non-nulls is set as 0.5. Again we observe while both rules have FDR controlled under nominal level (the black line), and our proposed CwLORD++ is more powerful than the LORD++ with respect to the true discovery rate. On average, we notice about 3-5% improvement in the power with CwLORD++ when compared to LORD++.

Diabetes Prediction Problem

In this section, we apply our online multiple testing rules to a real-life application of diabetes prediction. Machine learning algorithms are now commonly used to construct predictive health scores for patients. In this particular problem, we want to test whether the patients are at risk of developing diabetes. A high predicted risk score can trigger an intervention (such as medical follow-up, medical tests), which can be expensive and sometimes unnecessary, and therefore it is important to control the fraction of false alerts. That is, for each patient i, we form the null hypothesis H_i as the "patient will not develop diabetes" versus its alternative. The dataset was released as part of a Kaggle competition [2], which contains de-identified medical records of 9948 patients (labeled as $1, 2, \ldots$). For each patient, we have a response variable Y that indicates

[2] http://www.kaggle.com/c/pf2012-diabetes

if the patient is diagnosed with Type 2 diabetes mellitus, along with information on medications, lab results, immunizations, allergies, and vital signs. In the following, we train a predictive score based on the available records, and then will apply our online multiple testing rule rules to control the FDR on test set. Our overall methodology is similar to that used by [JM18] in their FDR control experiments on this dataset. We construct the following features for each patient.

1. Biographical information: Age, height, weight, BMI (Body Mass Indicator), etc.
2. Medications: We construct TF-IDF vectors from the medication names.
3. Diagnosis information: We derive 20 categories from the ICD-9 codes and construct an one-hot encoded vector.
4. Physician specialty: We categorize the physician specialties and create features that represents how many times a patient visited certain specialist.

We regard the biographical information of patients as treated as contextual features. The choice of using biographical information as context is loosely based on the idea of *personalization* which is common in machine learning applications. In theory, one could use other features too as context.

Second, we split the dataset into four parts **Train1**, comprising 40% of the data, **Train2**, 20% of the data, **Test1**, 20% of the data and **Test2**, 20% of the data. The **Train** sets are used for training a machine learning model (**Train1**) and for computing the null distribution of test statistics (**Train2**), which allows us to compute the P-values in the **Test** sets. We first learn the neural network parameters in the CwLORD++ procedure in an online fashion by applying it to the P-values in **Test**1, and then evaluate the performance of both LORD++ and CwLORD++ on **Test2**. This process is explained in more details below.

We note that our experimental setup is not exactly identical to that of [JM18], since we are using a slightly different set of features and data cleaning for the logistic regression model. We also split the data to four subsets instead of three as they did, which gives less training data

for the predictive model. Our main focus, is to compare the power of LORD++ and CwLORD++, for a reasonable feature set and machine learning model.

Training Process. We start by training a logistic model similar to [JM18].[3] Let x_i denote the features of patient i. We use all the features to model the probability that the patient does not have diabetes through a logistic regression model as

$$\Pr[Y_i = 0 \mid x = x_i] = \frac{1}{1 + \exp(\langle \beta, x_i \rangle)}. \tag{4.30}$$

The parameter β is estimated from the **Train1** set.

P-values Computation. Let S_0 be the subset of the patients in **Train2** set with the labels as $Y = 0$, and let $n_0 = |S_0|$. For each $i \in S_0$, we compute its predictive score as $v_i = 1/(1 + \exp(\langle \beta, x_i \rangle))$. The empirical distribution of $\{v_i : i \in S_0\}$ serves as the null distribution of the test statistic, which allows for computation of the P-values. Explicitly, for each j in either **Test1** or **Test2** sets, we compute $v_j^{\text{Test}} = 1/(1 + \exp(\langle \beta, x_j \rangle))$, and construct the P-value P_j by

$$P_j = \frac{1}{n_0} \left| \{ i \in S_0 : v_i \le v_j^{\text{Test}} \} \right|. \tag{4.31}$$

Smaller P-value indicates that the patient has higher risk of developing diabetes. We use the P-values computed on the patients in **Test1** to train the weight function in the CwLORD++, and the P-values on the patients in **Test2** to the compare performance of the CwLORD++ and LORD++. Note that the training of the neural network does not utilize the labels of the hypothesis in the **Test1** set. Since the dataset does not have the timestamps of the hypotheses, we consider an ordering of the hypotheses in the ascending order of the corresponding P-values, and use this ordering for both LORD++ and CwLORD++. Since the **Train** and **Test** sets are exchangeable, the null P-values is uniform in expectation and asymptotically uniform under mild conditions.

[3] Even though the chosen logistic model is one of the best performing models on this dataset in the Kaggle competition, in this chapter, we do not actively optimize the prediction model in the training process.

Online Hypothesis Testing Process and Results. We set the desired FDR control level at $q = 0.2$. The set of hyperparameters $\{\gamma_t\}$ is again chosen as above from [JM18]. For the patients in **Test1** set, we use their biographical information of patients as contextual features in the training process for the CwLORD++ for learning the neural network parameters. We apply the LORD++ and CwLORD++ procedures to the **Test2** set and compute the false discovery proportion and statistical power. Note that for both CwLORD++ and LORD++, the P-values are identically computed for patients in **Test2** set. This generates a sequence of P-values $(P_i)_{i\in T_2}$. Now, while the LORD++ is applied to this P-value sequence directly, the CwLORD++ is applied to the sequence of (P_i, X_i) where X_i is the biographical information of patient $i \in T_2$. For the CwLORD++, the neural network parameters are fixed in this testing over the **Test2** set.

We repeat the whole process and average the results over 30 random splittings of the dataset. Table 4.1 presents our final result.

We can use biographical information again as contextual features in training CwLORD++ because the P-values under the null are uniformly distributed, no matter which features (including biographical information or not) are used in logistic modeling. This guarantees that the P-values under the null are independent to any features, which is the only condition we need to have the FDR control, assuming the P-values themselves are mutually independent.

Notice that while the FDR is under control for both procedures, the statistical power of the CwLORD++ is substantially (about 51%) more than the LORD++. This improvement illustrates the benefits of using contextual features for improving the power with the FDR control in a typical machine learning setup. A possible reason for the observed increase in power with the CwLORD++ is that in addition to using the labeled data in the Train1 set for training a supervised model, the CwLORD++ uses in an unsupervised way (i.e., without considering labels) some features of the data in the **Test1** set in its online training process with the intent of maximizing the discoveries.

In order to further probe some of these improvements, we repeated the experiment with

Table 4.1: Results from diabetes dataset with nominal FDR control level $q = 0.2$.

	FDR	Power
LORD++	0.147	0.384
CwLORD++	0.176	0.580

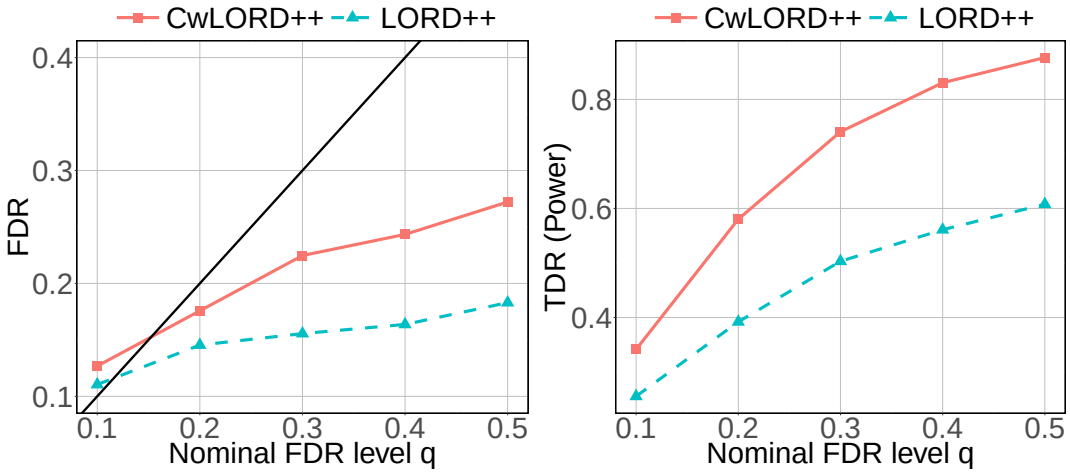

Figure 4.4: FDR and TDR results on diabetes dataset as we vary the nominal FDR level q. Note that the power of CwLORD++ uniformly dominates that of LORD++, with an average improvement in power of about 44%.

different nominal FDR levels ranging from 0.1 to 0.5. The results (see Figure 4.4) demonstrate that our CwLORD++ procedure achieves more true discoveries than the LORD++ procedure while controlling the FDR under the same level. The FDR is controlled exactly under the desired level starting around $q \geq 0.15$, while it is close to the desired level even when q is as small as 0.1. This phenomenon can also be observed in [JM18], where the FDR is 0.126 for LORD under the target level $q = 0.1$. This is probably because both the experiments here and in [JM18] do not adjust for the dependency among the P-values, which violates the theoretical assumption behind the FDR control proof, and is more of a concern when target q level is small.

Gene Expression Data

Our final set of experiments are on gene expression datasets. In particular, we use the Airway RNA-Seq and GTEx datasets[4] as also studied by [XZZT17]. For both experiments, we use the original ordering of hypotheses as provided in the datasets. Since we don't know the ground truth, we only report the empirical FDR and the empirical discovery rate number in the experiments.

In the Airway RNA-Seq data, the goal is to identify the glucocorticoid responsive (GC) genes that modulate cytokine function in the airway smooth muscle cells. The dataset contains $n = 33469$ genes. The P-values are obtained in regular two-sample analysis of gene expression levels. Log counts of each gene serves as the contextual feature in this case. Figure 4.5 reports the empirical FDR and the discovery number. We see that our CwLORD++ procedure make about 10% more discoveries than the LORD++ procedure.

In the GTEx study, a major question is to quantify the expression Quantitative Trait Loci (eQTLs) in human tissues. In the eQTL analysis, the association of each pair of single nucleotide polymorphism (SNP) and the nearby gene is tested. The P-value is obtained under the null hypothesis which the SNP genotype is not correlated with the gene expression. The GTEx dataset contains 464,636 pairs of SNP-gene combination from chromosome 1 in a brain tissue (interior caudate). Besides the P-values from the correlation test, contextual features may affect whether a SNP is likely to be an eQTL, and thus we can discover more of the true eQTLs if we utilize them in tests. For the tests, we consider three contextual feature studied by [XZZT17]: 1) the log of the distance (GTEx-dist) between the SNP and the gene; 2) the average expression (GTEx-exp) of the gene across individuals (measured in log rpkm); and 3) the evolutionary conservation measured by the standard PhastCons scores (GTEx-PhastCons). We apply the LORD++ to solely the P-values, and the CwLORD++ to the P-value, contextual feature vector pairs. Figure 4.6 report the empirical FDR and the discovery number with respect to different contextual features.

[4]Datasets are at: `https://www.dropbox.com/sh/wtp58wd60980d6b/AAA4wA60ykP-fDfS5BNsNkiGa?dl=0`.

In GTEx experiments, we use each contextual feature in the CwLORD++, which increases the discovery number by 5.5% (using GTEx-dist), 2.6% (using GTEx-dist), and 2.9% (using GTEx-PhastCons) correspondingly compared to the LORD++ procedure.

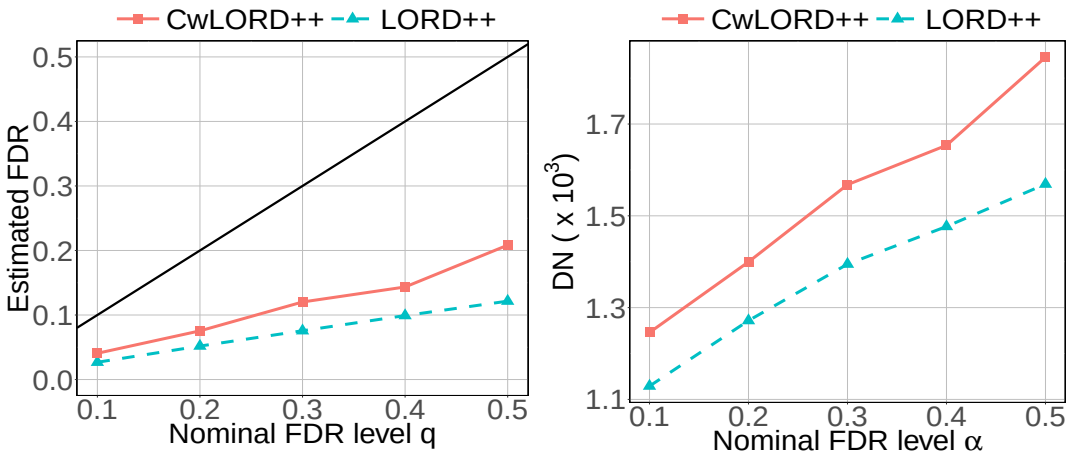

Figure 4.5: Results on Airway RNA-Seq dataset.

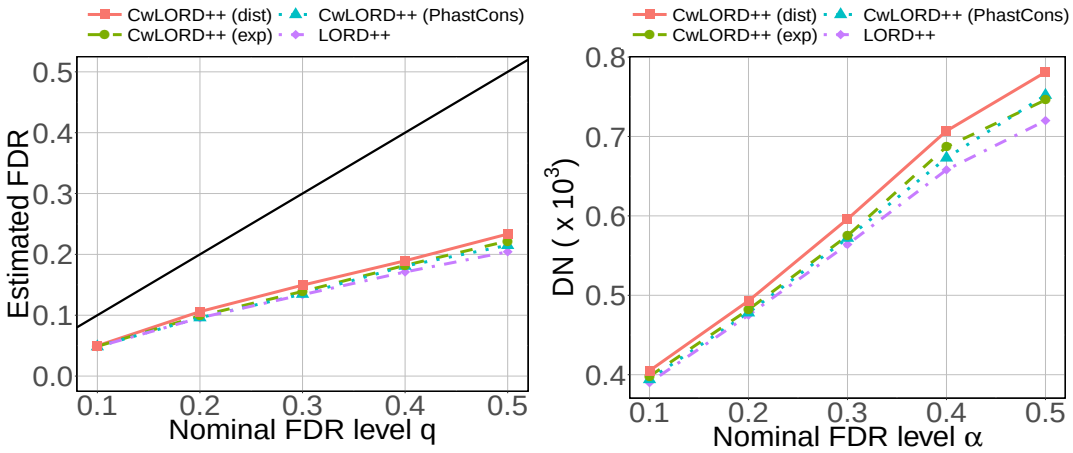

Figure 4.6: Results on GTEx experiments.

4.8 Proofs of Section 4.4

Identifiability of $f_1(p \mid X)$. We present a simple example from [LF18] that illustrates why $f_1(p \mid X)$ (distribution of p under the alternate) is not identifiable. Consider the following

mixture model:

$$H_t \mid X_t \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\pi_1),$$
$$P_t \mid H_t, X_t = \begin{cases} \text{Uniform}(0,1) & \text{if } H_t = 0, \\ f_1(p \mid X_t) & \text{if } H_t = 1. \end{cases}$$

Now consider the conditional mixture density $f(p \mid X) = (1-\pi_1) + \pi_1 f_1(p \mid X)$. Note that the H_t's are not observed. Thus, while f is identifiable from the data, π_1 and f_1 are not: for example, $\pi_1 = 0.5$, $f_1(p \mid X) = 2(1-p)$ and $\pi_1 = 1$, $f_1(p \mid X) = 1.5 - p$ result in exactly the same mixture density $f(p \mid X)$.

Lemma 6 (Lemma 5 Restated)**.** *Let $g : \{0,1\}^T \to \mathbb{R}$ be any coordinatewise non-decreasing function such that $g(\mathbf{R}) > 0$ for any vector $\mathbf{R} \neq (0,\ldots,0)$. Then for any index $t \leq T$ such that $t \in \mathcal{H}^0$, we have*

$$\mathbb{E}\left[\frac{\mathbb{I}\{P_t \leq \alpha_t(R_1,\ldots,R_{t-1},X_1,\ldots,X_t)\}}{g(R_1,\ldots,R_T) \vee 1}\middle| \sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t)\right]$$
$$\leq \mathbb{E}\left[\frac{\alpha_t(R_1,\ldots,R_{t-1},X_1,\ldots,X_T)}{g(R_1,\ldots,R_T) \vee 1}\middle| \sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t)\right].$$

Proof. Let $\mathbf{P} = (P_1,\ldots,P_T)$ be the sequence of P-values, and $\mathbf{X} = (X_1,\ldots,X_T)$ be the sequence of the contextual feature vectors until sometime T. We define a "leave-one-out" vector of P-value as $\widetilde{\mathbf{P}}^{-t} = (\widetilde{P}_1,\ldots,\widetilde{P}_T)$, which was obtained from $\mathbf{P}$ by setting $P_t = 0$, i.e.,

$$\widetilde{P}_i = \begin{cases} P_i & \text{if } i \neq t, \\ 0 & \text{if } i = t. \end{cases} \tag{4.32}$$

Let $\mathbf{R} = (R_1,\ldots,R_T)$ be the sequence of decisions on the input $\mathbf{P}$ and $\mathbf{X}$, and $\widetilde{\mathbf{R}}^{-t} = (\widetilde{R}_1,\ldots,\widetilde{R}_T)$ be the sequence of decisions by applying the same rule on the input $\mathbf{P}^{-t}$ and $\mathbf{X}$. Note here we just set one P-value as zero but do not change the contextual feature vectors.

By the construction of P-values, we have that $R_i = \widetilde{R}_i$ for $i < t$, and hence

$$\alpha_i(R_1, \ldots, R_{i-1}, X_1, \ldots, X_i) = \alpha_i(\widetilde{R}_1, \ldots, \widetilde{R}_{i-1}, X_1, \ldots, X_i), \quad \text{for all } i \le t. \tag{4.33}$$

We also know that $\widetilde{R}_t = 1$ always holds due to the fact $\widetilde{P}_t = 0 \le \alpha_t$. Therefore, if the event $\{P_t \le \alpha_t(R_1, \ldots, R_{t-1}, X_1, \ldots, X_t)\}$ occurs, we have $R_t = \widetilde{R}_t$ and thus $\mathbf{R} = \widetilde{\mathbf{R}}^{-t}$.

From the above arguments, we conclude that

$$\frac{\mathbb{I}\{P_t \le \alpha_t(R_1, \ldots, R_{t-1}, X_1, \ldots, X_t)\}}{g(\mathbf{R}) \vee 1} = \frac{\mathbb{I}\{P_t \le \alpha_t(R_1, \ldots, R_{t-1}, X_1, \ldots, X_t)\}}{g(\widetilde{\mathbf{R}}^{-t}) \vee 1}. \tag{4.34}$$

Due to the fact that $t \in \mathcal{H}^0$ ($H_t = 0$), P_t is independent to all contextual features $\mathbf{X}$ by assumption (as P_i's and X_i's are independent under the null), which gives that P_t is independent of $\sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t)$. And since P_t is independent of $\widetilde{\mathbf{R}}^{-t}$, we have,

$$\begin{aligned}
&\mathbb{E}\left[\frac{\mathbb{I}\{P_t \le \alpha_t(R_1, \ldots, R_{t-1}, X_1, \ldots, X_t)\}}{g(\mathbf{R}) \vee 1} \middle| \sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t)\right] \\
&= \mathbb{E}\left[\frac{\mathbb{I}\{P_t \le \alpha_t(R_1, \ldots, R_{t-1}, X_1, \ldots, X_t)\}}{g(\widetilde{\mathbf{R}}^{-t}) \vee 1} \middle| \sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t)\right] \\
&\le \mathbb{E}\left[\frac{\alpha_t(R_1, \ldots, R_{t-1}, X_1, \ldots, X_t)}{g(\widetilde{\mathbf{R}}^{-t}) \vee 1} \middle| \sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t)\right] \qquad (4.35) \\
&\le \mathbb{E}\left[\frac{\alpha_t(R_1, \ldots, R_{t-1}, X_1, \ldots, X_t)}{g(\mathbf{R}) \vee 1} \middle| \sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t)\right] \qquad (4.36)
\end{aligned}$$

where inequality (4.35) follows by taking expectation with respect to P_t and using conditional super-uniformity (4.15), and inequality (4.36) is derived by the following observation.

Since $\widetilde{P}_t = 0 \le \alpha_t$, we have $\widetilde{R}_t = 1 \ge R_t$. Due to the monotonicity of the significance levels, we have

$$\alpha_i(\widetilde{R}_1, \ldots, \widetilde{R}_{i-1}, X_1, \ldots, X_i) \ge \alpha_i(R_1, \ldots, R_{i-1}, X_1, \ldots, X_i), \quad \text{for all } i > t, \tag{4.37}$$

ensuring $\widetilde{R}_i \ge R_i$ for all i, and thus $g(\widetilde{\mathbf{R}}^{-t}) \ge g(\mathbf{R})$ by the non-decreasing assumption on the

function g. $\square$

4.8.1 Proof of Theorem 11

Note that the number of false discoveries is $V(T) = \sum_{t=1}^{T} R_t \mathbb{I}\{t \in \mathcal{H}^0\}$ and the amount of wealth is $W(T) = w_0 + \sum_{t=1}^{T} (-\phi_t + R_t \psi_t)$.

We can derive the following expression by using the tower property of conditional expectation

$$\begin{aligned}
\mathbb{E}\left[\frac{V(T)+W(T)}{R(T)\vee 1}\right] &= \sum_{t=1}^{T} \mathbb{E}\left[\frac{R_t \mathbb{I}\{t \in \mathcal{H}^0\} + \frac{w_0}{T} - \phi_t + R_t \psi_t}{R(T)\vee 1}\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\frac{\frac{w_0}{T} + R_t\left(\psi_t + \mathbb{I}\{t \in \mathcal{H}^0\}\right) - \phi_t}{R(T)\vee 1}\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{E}\Big[\frac{\frac{w_0}{T} + R_t\left(\psi_t + \mathbb{I}\{t \in \mathcal{H}^0\}\right) - \phi_t}{R(T)\vee 1}\Big|\sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t)\Big]\right] \qquad (4.38)
\end{aligned}$$

We split the analysis in two cases based on whether $H_t = 0$ or $H_t = 1$.

- Case 1: Suppose that $t \in \mathcal{H}^0$. By applying Lemma 5, we have

$$\begin{aligned}
\mathbb{E}\left[\frac{R_t}{R(T)\vee 1}\middle|\sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t)\right] &= \mathbb{E}\left[\frac{\mathbb{I}\{P_t \le \alpha_t\}}{R(T)\vee 1}\middle|\sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t)\right] \\
&\le \mathbb{E}\left[\frac{\alpha_t}{R(T)\vee 1}\middle|\sigma(\mathcal{F}^{t-1} \cup \mathcal{G}^t)\right] \qquad (4.39)
\end{aligned}$$

  Since $\psi_t \le \frac{\phi_t}{\alpha_t} + b_t - 1$, we further obtain

$$
\begin{aligned}
&\mathbb{E}\left[\mathbb{E}\left[\frac{\frac{w_0}{T}+R_t\left(\psi_t+\mathbb{I}\{t\in\mathcal{H}^0\}\right)-\phi_t}{R(T)\vee 1}\middle|\sigma\left(\mathcal{F}^{t-1}\cup\mathcal{G}^t\right)\right]\right]\\
&\leq\mathbb{E}\left[\mathbb{E}\left[\frac{\frac{w_0}{T}+R_t\left(\frac{\phi_t}{\alpha_t}+b_t\right)-\phi_t}{R(T)\vee 1}\middle|\sigma\left(\mathcal{F}^{t-1}\cup\mathcal{G}^t\right)\right]\right]\\
&=\mathbb{E}\left[\mathbb{E}\left[\frac{\frac{w_0}{T}+R_t b_t+\frac{\phi_t}{\alpha_t}\left(R_t-\alpha_t\right)}{R(T)\vee 1}\middle|\sigma\left(\mathcal{F}^{t-1}\cup\mathcal{G}^t\right)\right]\right]\\
&\leq\mathbb{E}\left[\mathbb{E}\left[\frac{\frac{w_0}{T}+R_t b_t}{R(T)\vee 1}\middle|\sigma\left(\mathcal{F}^{t-1}\cup\mathcal{G}^t\right)\right]\right],
\end{aligned}
$$

where the last inequality follows by applying (4.39).

- Case 2: Suppose that $t\notin\mathcal{H}^0$. Using the fact that $\psi_t\leq\phi_t+b_t$, we have

$$
\begin{aligned}
&\mathbb{E}\left[\mathbb{E}\left[\frac{\frac{w_0}{T}+R_t\left(\psi_t+\mathbb{I}\{t\in\mathcal{H}^0\}\right)-\phi_t}{R(T)\vee 1}\middle|\sigma\left(\mathcal{F}^{t-1}\cup\mathcal{G}^t\right)\right]\right]\\
&\leq\mathbb{E}\left[\mathbb{E}\left[\frac{\frac{w_0}{T}+R_t\left(\phi_t+b_t\right)-\phi_t}{R(T)\vee 1}\middle|\sigma\left(\mathcal{F}^{t-1}\cup\mathcal{G}^t\right)\right]\right]\\
&=\mathbb{E}\left[\mathbb{E}\left[\frac{\frac{w_0}{T}+R_t b_t+\left(R_t-1\right)\phi_t}{R(T)\vee 1}\middle|\sigma\left(\mathcal{F}^{t-1}\cup\mathcal{G}^t\right)\right]\right]\\
&\leq\mathbb{E}\left[\mathbb{E}\left[\frac{\frac{w_0}{T}+R_t b_t}{R(T)\vee 1}\middle|\sigma\left(\mathcal{F}^{t-1}\cup\mathcal{G}^t\right)\right]\right].
\end{aligned}
$$

Combining the bound on $\mathbb{E}\left[\mathbb{E}\left[\frac{\frac{w_0}{T}+R_t\left(\psi_t+\mathbb{I}\{t\in\mathcal{H}^0\}\right)-\phi_t}{R(T)\vee 1}\middle|\sigma\left(\mathcal{F}^{t-1}\cup\mathcal{G}^t\right)\right]\right]$ from both cases in (4.38) and using the definition of b_t, we obtain that,

$$
\begin{aligned}
\mathbb{E}\left[\frac{V(T)+W(T)}{R(T)\vee 1}\right]&\leq\sum_{t=1}^{T}\mathbb{E}\left[\frac{\frac{w_0}{T}+R_t b_t}{R(T)\vee 1}\right]=\mathbb{E}\left[\frac{w_0+\sum_{t=1}^{T}R_t b_t}{R(T)\vee 1}\right]\\
&\leq\mathbb{E}\left[\frac{w_0+\sum_{t=1}^{T}R_t q-w_0\mathbb{I}\{T\geq\rho_1\}}{R(T)\vee 1}\right]=\mathbb{E}\left[\frac{w_0+qR(T)-w_0\mathbb{I}\{T\geq\rho_1\}}{R(T)\vee 1}\right]\leq q.
\end{aligned}
$$

This concludes the proof of the theorem.

4.8.2 Proof of Theorem 12

The conditional super-uniformity implies that under null

$$\mathbb{E}\Big[R_t\Big|\sigma(\mathcal{F}^{t-1}\cup\mathcal{G}^t)\Big]\le\alpha_t.$$

Now using a proof technique similar to Theorem 11, for any $T\in\mathbb{N}$, we get

$$\begin{aligned}
\mathbb{E}[V(T)] &\le \mathbb{E}[V(T)+W(T)] \\
&= \sum_{t=1}^{T}\mathbb{E}\left[R_t\mathbb{I}\{t\in\mathcal{H}^0\}+\frac{w_0}{T}-\phi_t+R_t\psi_t\right] \\
&= \sum_{t=1}^{T}\mathbb{E}\left[\frac{w_0}{T}+R_t(\psi_t+\mathbb{I}\{t\in\mathcal{H}^0\})-\phi_t\right] \\
&= \sum_{t=1}^{T}\mathbb{E}\left[\mathbb{E}\Big[\frac{w_0}{T}+R_t(\psi_t+\mathbb{I}\{t\in\mathcal{H}^0\})-\phi_t\Big|\sigma(\mathcal{F}^{t-1}\cup\mathcal{G}^t)\Big]\right] \\
&\le \sum_{t=1}^{T}\mathbb{E}\left[\mathbb{E}\Big[\frac{w_0}{T}+R_tb_t\Big|\sigma(\mathcal{F}^{t-1}\cup\mathcal{G}^t)\Big]\right] \\
&= \mathbb{E}\Big[w_0+\sum_{t=1}^{T}R_tb_t\Big] = \mathbb{E}\Big[w_0+qR(T)-w_0\mathbb{I}\{T\ge\rho_1\}\Big] \\
&\le q\,\mathbb{E}\Big[R(T)\vee 1\Big],
\end{aligned}$$

where for the second inequality we used an analysis similar to that used in the first case in the proof of Theorem 11. Therefore, for any $T\in\mathbb{N}$,

$$\mathrm{mFDR}(T)=\frac{\mathbb{E}[V(T)]}{\mathbb{E}[R(T)\vee 1]}\le q. \tag{4.40}$$

This concludes the proof of the theorem.

4.9 Proofs of Section 4.6

Proposition 4 (Proposition 3 Restated)**.** *Suppose that the weight distribution satisfies the informative weighting assumption in* (4.20)*. Suppose that P-values P_t's are independent, and are conditionally independent of the weights ω_t's given H_t's. Then the weighted LORD++ rule can control the online FDR at any given level q, i.e.,*

$$\sup_{T\in\mathbb{N}} \mathrm{FDR}(T) \le q. \tag{4.41}$$

Proof. We start with a frequently used estimator of FDR that is defined as:

$$\widehat{\mathrm{FDP}}(T) := \frac{\sum_{t=1}^{T} \alpha_t}{R(T)\vee 1}. \tag{4.42}$$

As established in Section 4 in [RYWJ17], the LORD++ applied to any sequence of P-values will ensure that $\sup_T \widehat{\mathrm{FDP}}(T) \le q$. We apply the LORD++ with the sequence of P-values defined as $P' = (\frac{P_1}{\omega_1}, \frac{P_2}{\omega_2}, \frac{P_3}{\omega_3} \dots)$. Let $P'_t = P_t/\omega_t$ for any $t \in \mathbb{N}$. Then it follows that,

$$\sup_{T\in\mathbb{N}} \widehat{\mathrm{FDP}}(T) = \sup_{T\in\mathbb{N}} \frac{\sum_{t=1}^{T} \alpha_t}{R(T)\vee 1} = \sup_{T\in\mathbb{N}} \frac{\sum_{t=1}^{T} \alpha_t}{(\sum_{t=1}^{T} \mathbb{I}\{P'_t \le \alpha_t\})\vee 1} \le q. \tag{4.43}$$

We denote the sigma-field of decisions based on the weighted P-values P' till time t as $\mathcal{C}^t = \sigma(R_1, \dots, R_t)$. By using the "leave-one-out" method used in Lemma 5, the FDR of the

weighted LORD++ at any time T can be written as,

$$\begin{aligned}
\mathrm{FDR}(T) &= \mathbb{E}\left[\frac{\sum_{t=1}^{T}\mathbb{I}\{t\in\mathcal{H}^0 : P'_t \le \alpha_t\}}{(\sum_{t=1}^{T}\mathbb{I}\{P'_t\le\alpha_t\})\vee 1}\right] \\
&= \sum_{t=1}^{T}\mathbb{E}\left[\frac{\mathbb{I}\{t\in\mathcal{H}^0 : \frac{P_t}{\omega_t}\le\alpha_t\}}{R(T)\vee 1}\right] \\
&= \sum_{t=1}^{T}\mathbb{E}\left[\mathbb{E}\Big[\frac{\mathbb{I}\{t\in\mathcal{H}^0 : \frac{P_t}{\omega_t}\le\alpha_t\}}{R(T)\vee 1}\Big|\mathcal{C}^{t-1}\Big]\right] \\
&= \sum_{t=1}^{T}\mathbb{E}\left[\mathbb{E}\Big[\frac{\mathbb{I}\{t\in\mathcal{H}^0 : \frac{P_t}{\omega_t}\le\alpha_t\}}{R^{-t}(T)\vee 1}\Big|\mathcal{C}^{t-1}\Big]\right],
\end{aligned}$$

where $R^{-t}(T) = \sum_{i=1}^{T}\mathbb{I}\{P_i/\omega_i\le\alpha_i\}$ is obtained by setting $P_t = 0$, while keeping all ω_t's unchanged. The last equality holds due to the fact that $R^{-t}(T) = R(T)$ given the event $\{P_t/\omega\le\alpha_t\}$.

Since $\alpha_t\in\mathcal{C}^{t-1}$, and P_t, ω_t are independent of $R^{-t}(T)$ and $\mathcal{C}^{t-1}$, we can take the expectation of the numerator inside the brackets and obtain that

$$\begin{aligned}
\Pr[P_t/\omega_t\le\alpha_t \mid \mathcal{C}^{t-1}, H_t = 0] &= \int \Pr[P_t/\omega_t\le\alpha_t \mid \mathcal{C}^{t-1}, \omega_t = w, H_t = 0]\,\mathrm{d}Q(w\mid H_t = 0) \\
&= \int w\alpha_t\,\mathrm{d}Q(w\mid H_t = 0) \\
&= u_0\alpha_t,
\end{aligned}$$

where $u_0 = \mathbb{E}[\omega \mid H_t = 0]$. Plugging this in the bound on $\mathrm{FDR}(T)$ from above gives,

$$\begin{aligned}
\mathrm{FDR}(T) &= \sum_{t=1}^{T}\mathbb{E}\left[\mathbb{E}\Big[\frac{u_0\alpha_t}{R^{-t}(T)\vee 1}\Big|\mathcal{C}^{t-1}\Big]\right] \\
&\le \sum_{t=1}^{T}\mathbb{E}\left[\mathbb{E}\Big[\frac{\alpha_t}{R^{-t}(T)\vee 1}\Big|\mathcal{C}^{t-1}\Big]\right] \qquad (4.44) \\
&\le \sum_{t=1}^{T}\mathbb{E}\left[\mathbb{E}\Big[\frac{\alpha_t}{R(T)\vee 1}\Big|\mathcal{C}^{t-1}\Big]\right] \qquad (4.45) \\
&= \mathbb{E}\left[\frac{\sum_{t=1}^{T}\alpha_t}{R(T)\vee 1}\right] = \mathbb{E}\left[\widehat{\mathrm{FDP}}(T)\right] \le q,
\end{aligned}$$

where inequality (4.44) is due to the assumption that $u_0 < 1$, (4.45) follows by the fact that $R^{-t}(T) \geq R(T)$ due to monotonicity of LORD++, and the last equality is based on (4.43). □

4.9.1 Proof of Theorem 13

Since we are interested in lower bounds, we consider a version of LORD (as also considered in [JM18]) which that is based on the following rule,

$$\textbf{LORD*: } W(0) = w_0 = b_0 = q/2, \quad \phi_t = \alpha_t = b_0\gamma_{t-\tau_t}, \quad \psi_t = b_0.$$

Note that since $b_t = q - w_0\mathbb{I}\{\rho_1 > t-1\} > b_0$ in LORD++, the test level in LORD++ is at least as large to the test level in LORD*. Therefore, for any P-value sequence the power of LORD* from is also a lower bound on the power of LORD++. In the rest of this proof, we focus on LORD* for the weighted P-value sequence $\{P_1/\omega_1, P_2/\omega_2, \dots\}$. The bound established below is in fact *tight* for LORD* under this P-value sequence.

Denote by ρ_i as the time of the ith discovery (rejection), with $\rho_0 = 0$, and $\Delta_i = \rho_i - \rho_{i-1}$ as the ith time interval between the $(i-1)$st and ith discoveries. Let $r_i := \mathbb{I}\{\rho_i \in \mathcal{H}^1\}$ be the reward associated with inter-discovery Δ_i. Since the weighted P-values are i.i.d, it can be seen that the times between successive discoveries are i.i.d. according to the testing procedure LORD*, and the process $R(T) = \sum_{l=1}^{T} R_l$ is a *renewal process* [CCCC67]. In fact, for each i, we have

$$\Pr[\Delta_i \geq m] = \Pr[\cap_{l=\rho_{i-1}}^{\rho_{i-1}+m}\{P_l/\omega_l > \alpha_l\}] \tag{4.46}$$

$$= \prod_{l=\rho_{i-1}}^{\rho_{i-1}+m}(1 - D(\alpha_l)) \tag{4.47}$$

$$= \prod_{l=\rho_{i-1}}^{\rho_{i-1}+m}(1 - D(b_0\gamma_{l-\rho_{i-1}})) \tag{4.48}$$

$$= \prod_{l=1}^{m}(1 - D(b_0\gamma_l)). \tag{4.49}$$

The above expression is same for every i. Therefore,

$$\mathbb{E}[\Delta_i] = \sum_{m=1}^{\infty} \Pr[\Delta_i \geq m] = \sum_{m=1}^{\infty} \prod_{l=1}^{m} (1 - D(b_0\gamma_l)). \tag{4.50}$$

Applying the strong law of large numbers for renewal-reward processes [CCCC67], we obtain that the following statement holds almost surely,

$$\lim_{T\to\infty} \frac{1}{T} \sum_{i=1}^{R(T)} r_i = \frac{\mathbb{E}(r_i)}{\mathbb{E}(\Delta_1)} = \pi_1 (\sum_{m=1}^{\infty} \prod_{l=1}^{m} (1 - D(b_0\gamma_l)))^{-1}. \tag{4.51}$$

Let $|\mathcal{H}^1(T)|$ be the number of true alternatives till time T. Since $\lim_{T\to\infty} |\mathcal{H}^1(T)|/T = \pi_1$ almost surely, we have

$$\lim_{T\to\infty} \frac{1}{|\mathcal{H}^1(T)|} \sum_{i\in\mathcal{H}^1(T)} R_i = \lim_{T\to\infty} \frac{1}{|\mathcal{H}^1(T)|} \sum_{i=1}^{R(T)} r_i = (\sum_{m=1}^{\infty} \prod_{l=1}^{m} (1 - D(b_0\gamma_l)))^{-1}. \tag{4.52}$$

Now by using the definition of $\mathrm{TDP}(T)$, almost surely, we have that for any weighted LORD++,

$$\liminf_{T\to\infty} \mathrm{TDP}(T) \geq (\sum_{m=1}^{\infty} \prod_{j=1}^{m} (1 - D(b_0\gamma_j)))^{-1}.$$

As discussed above, this bound translates into a lower bound for weighted LORD++. Furthermore, by using the Fatou's lemma [Car00], we can extend the same result for $\mathrm{TDR}(T)$ almost surely,

$$\liminf_{T\to\infty} \mathrm{TDR}(T) = \liminf_{T\to\infty} \mathbb{E}[\mathrm{TDP}(T)] \geq \mathbb{E}[\liminf_{T\to\infty} \mathrm{TDP}(T)] \geq (\sum_{m=1}^{\infty} \prod_{j=1}^{m} (1 - D(b_0\gamma_j)))^{-1}.$$

4.9.2 Proof of Theorem 14

We compare the average power bound of the weighted LORD* and LORD*. It is equivalent to comparing $D(a)$ and $G(a)$ for $a = b_0\gamma_l$, with $l = 1, \ldots, \infty$. Since $u = (1 - \pi_1)u_0 +$

$\pi_1 u_1 = 1$, we have $(1-\pi_1)u_0 = 1-\pi_1 u_1$. This means that

$$\begin{aligned} D(a) - G(a) &= (1-\pi_1)u_0 a + \pi_1 \int F(aw)\,\mathrm{d}Q_1(w) - (1-\pi_1)a - \pi_1 F(a) \\ &= (1-\pi_1)(u_0-1)a + \pi_1\Big(\int F(aw)\,\mathrm{d}Q_1(w) - F(a)\Big) \\ &= \pi_1(1-u_1)a + \pi_1\Big(\int F(aw)\,\mathrm{d}Q_1(w) - F(a)\Big). \end{aligned}$$

So we just need to compare $(\mu_1 - 1)a$ and $\int F(aw)\,\mathrm{d}Q_1(w) - F(a)$, for any $a = b_0\gamma_l$, with $l = 1,\ldots,\infty$. Due to the fact that $\{\gamma_l\}$ is a non-increasing sequence, we have $a = b_0\gamma_l \le b_0\gamma_1$. Since $b_0\gamma_1 < a_0$ and $\Pr[\omega < a_0/(b_0\gamma_1) \mid H = 1] = 1$ by assumption, then $\Pr[\max(a, aw) < a_0 \mid H = 1] = 1$.

For any fixed $a = b_0\gamma_l > 0$, we have

$$\begin{aligned} \frac{\int F(aw)\,\mathrm{d}Q_1(w) - F(a)}{a} &= \int \frac{F(aw)-F(a)}{a}\,\mathrm{d}Q_1(w) \\ &= \int \frac{F(aw)-F(a)}{(w-1)a}(w-1)\,\mathrm{d}Q_1(w) \\ &= \int f(\xi)(w-1)\,\mathrm{d}Q_1(w) \qquad (4.53) \\ &\ge \int (w-1)\,\mathrm{d}Q_1(w) \qquad (4.54) \\ &= \mathbb{E}[W \mid H = 1] - 1 = u_1 - 1, \end{aligned}$$

for some $\xi \in (\min(a, aw), \max(a, aw))$. Note we assume Q_1 is a continuous distribution, so $\Pr[w = 1 \mid H = 1] = 0$. The equality (4.53) is achieved by applying the Intermediate Value Theorem, and the inequality (4.54) is obtained by the fact that $\Pr[\xi < a_0 \mid H = 1] = 1$, i.e., $\Pr[f(\xi) > 1 \mid H = 1] = 1$.

Therefore, we prove that $\int F(aw)\,\mathrm{d}Q_1(w) - F(a) \ge u_1 - 1$, which implies that $D(a) \ge G(a)$ for $a = b_0\gamma_l$, with any $l = 1,\ldots,\infty$.

4.10 Acknowledgement

Chapter 4, partially, is a version of paper “Contextual Online False Discovery Rate Control”, Chen, Shiyun; Kasiviswanathan, Shiva. The manuscript has been submitted to a major machine learning conference. The dissertation author was the primary investigator and author of this material.

Bibliography

[ABDJ06] Felix Abramovich, Yoav Benjamini, David L Donoho, and Iain M Johnstone. Adapting to unknown sparsity by controlling the false discovery rate. *The Annals of Statistics*, 34(2):584–653, 2006.

[ABR10] S Arlot, G Blanchard, and E Roquain. Some non-asymptotic results on resampling in high dimension, II: Multiple tests. *The Annals of Statistics*, 38(1):83–99, 2010.

[ACC17] Ery Arias-Castro and Shiyun Chen. Distribution-free multiple testing. *Electronic Journal of Statistics*, 11(1):1983–2001, 2017.

[ACW17] Ery Arias-Castro and Meng Wang. Distribution-free tests for sparse heterogeneous mixtures. *Test*, 26(1):71–94, 2017.

[ACY19] Ery Arias-Castro and Andrew Ying. Detection of sparse mixtures: Higher criticism and scan statistic. *Electronic Journal of Statistics*, 13(1):208–230, 2019.

[AR14] Ehud Aharoni and Saharon Rosset. Generalized α-investing: definitions, optimality results and application to public databases. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):771–794, 2014.

[Bar14] Jay Bartroff. Multiple hypothesis tests controlling generalized error rates for sequential data. *arXiv preprint arXiv:1406.5933*, 2014.

[BCFG11] Małgorzata Bogdan, Arijit Chakrabarti, Florian Frommlet, and Jayanta K Ghosh. Asymptotic bayes-optimality under sparsity of some multiple testing procedures. *The Annals of Statistics*, pages 1551–1579, 2011.

[BH95] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[BH97] Yoav Benjamini and Yosef Hochberg. Multiple hypotheses testing with weights. *Scandinavian Journal of Statistics*, 24(3):407–418, 1997.

[BH00] Yoav Benjamini and Yosef Hochberg. 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, 2000.

[BH07] Yoav Benjamini and Ruth Heller. False discovery rates for spatial signals. *Journal of the American Statistical Association*, 102(480):1272–1281, 2007.

[BS13] Jay Bartroff and Jinlin Song. Sequential tests of multiple hypotheses controlling false discovery and nondiscovery rates. *arXiv preprint arXiv:1311.3350*, 2013.

[BS14] Jay Bartroff and Jinlin Song. Sequential tests of multiple hypotheses controlling type i and ii familywise error rates. *Journal of statistical planning and inference*, 153:100–114, 2014.

[BST15] Cristina Butucea, Natalia A Stepanova, and Alexandre B Tsybakov. Variable selection with hamming loss. *arXiv preprint arXiv:1512.01832*, 2015.

[Car00] Neal L Carothers. *Real analysis*. Cambridge University Press, 2000.

[CCCC67] David Roxbee Cox, David Roxbee Cox, David Roxbee Cox, and David Roxbee Cox. *Renewal theory*, volume 1. Methuen London, 1967.

[CdCS06] Marcia Caldas de Castro and Burton H Singer. Controlling the false discovery rate: a new application to account for multiple and dependent tests in local statistics of spatial association. *Geographical Analysis*, 38(2):180–208, 2006.

[Chi07] Zhiyi Chi. On the performance of fdr control: constraints and a partial solution. *The Annals of Statistics*, pages 1409–1431, 2007.

[CJ10] T Tony Cai and Jiashun Jin. Optimal rates of convergence for estimating the null density and proportion of nonnull effects in large-scale multiple testing. *The Annals of Statistics*, 38(1):100–145, 2010.

[DFKO15] Edgar Dobriban, Kristen Fortney, Stuart K Kim, and Art B Owen. Optimal multiple testing under a gaussian prior on the effect sizes. *Biometrika*, 102(4):753–766, 2015.

[Dic14] Thorsten Dickhaus. *Simultaneous statistical inference*. Springer, 2014.

[DJ04] David Donoho and Jiashun Jin. Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32(3):962–994, 2004.

[Dob16] Edgar Dobriban. A general convex framework for multiple testing with prior information. *arXiv preprint arXiv:1603.05334*, 2016.

[DR06] Cécile Durot and Yves Rozenholc. An adaptive test for zero mean. *Mathematical Methods of Statistics*, 15(1):26–60, 2006.

[DvdL07] Sandrine Dudoit and Mark J van der Laan. *Multiple testing procedures with applications to genomics*. Springer Science & Business Media, 2007.

[Efr04] Bradley Efron. Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association*, 99(465):96–104, 2004.

[Eic79] F Eicker. The asymptotic distribution of the suprema of the standardized empirical processes. *The Annals of Statistics*, 7(1):116–138, 1979.

[FBC15] Rina Foygel-Barber and Emmanuel J Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015.

[FS08] Dean P Foster and Robert A Stine. 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, 2008.

[FST14] William Fithian, Dennis Sun, and Jonathan Taylor. Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*, 2014.

[FTTT15] William Fithian, Jonathan Taylor, Robert Tibshirani, and Ryan Tibshirani. Selective sequential model selection. *arXiv preprint arXiv:1512.02565*, 2015.

[GDS03] Youngchao Ge, Sandrine Dudoit, and Terence P Speed. Resampling-based multiple testing for microarray data analysis. *Test*, 12(1):1–77, 2003.

[GRW06] Christopher R Genovese, Kathryn Roeder, and Larry Wasserman. False discovery control with p-value weighting. *Biometrika*, 93(3):509–524, 2006.

[GW02] Christopher Genovese and Larry Wasserman. Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):499–517, 2002.

[GW04] Christopher Genovese and Larry Wasserman. A stochastic process approach to false discovery control. *Annals of Statistics*, pages 1035–1061, 2004.

[GWCT16] Max Grazier G'Sell, Stefan Wager, Alexandra Chouldechova, and Robert Tibshirani. Sequential selection procedures and false discovery rate control. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(2):423–444, 2016.

[Het84] Thomas P. Hettmansperger. *Statistical inference based on ranks*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, 1984.

[HZZ10] James X Hu, Hongyu Zhao, and Harrison H Zhou. False discovery rate control with groups. *Journal of the American Statistical Association*, 105(491):1215–1227, 2010.

[IKZH16] Nikolaos Ignatiadis, Bernd Klaus, Judith B Zaugg, and Wolfgang Huber. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature methods*, 13(7):577, 2016.

[Ing97] Yuri I Ingster. Some problems of hypothesis testing leading to infinitely divisible distributions. *Mathematical Methods of Statistics*, 6(1):47–69, 1997.

[IS03] Yu. I. Ingster and I. A. Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*, volume 169 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 2003.

[JC07] Jiashun Jin and T Tony Cai. Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons. *Journal of the American Statistical Association*, 102(478):495–506, 2007.

[JJ12] Pengsheng Ji and Jiashun Jin. Ups delivers optimal phase diagram in high-dimensional variable selection. *The Annals of Statistics*, 40(1):73–103, 2012.

[JK14] Jiashun Jin and Tracy Ke. Rare and weak effects in large-scale inference: methods and phase diagrams. *arXiv preprint arXiv:1410.4578*, 2014.

[JM15] Adel Javanmard and Andrea Montanari. On online control of false discovery rate. *arXiv preprint arXiv:1502.06197*, 2015.

[JM18] Adel Javanmard and Andrea Montanari. Online rules for control of false discovery rate and false discovery exceedance. *The Annals of statistics*, 46(2):526–554, 2018.

[LB16a] Ang Li and Rina Foygel Barber. Accumulation tests for fdr control in ordered hypothesis testing. *Journal of the American Statistical Association*, (just-accepted):1–38, 2016.

[LB16b] Ang Li and Rina Foygel Barber. Multiple testing with the structure adaptive benjamini-hochberg algorithm. *arXiv preprint arXiv:1606.07926*, 2016.

[LF16] Lihua Lei and William Fithian. Power of ordered hypothesis testing. *arXiv preprint arXiv:1606.01969*, 2016.

[LF18] Lihua Lei and William Fithian. Adapt: an interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):649–679, 2018.

[LTTT14] Richard Lockhart, Jonathan Taylor, Ryan J Tibshirani, and Robert Tibshirani. A significance test for the lasso. *Annals of statistics*, 42(2):413, 2014.

[MMB11] Nicolai Meinshausen, Marloes H Maathuis, and Peter Bühlmann. Asymptotic optimality of the westfall–young permutation procedure for multiple testing under dependence. *The Annals of Statistics*, 39(6):3369–3391, 2011.

[Nau65] Joseph I Naus. The distribution of the size of the maximum cluster of points on a line. *Journal of the American Statistical Association*, 60(310):532–538, 1965.

[NR12] Pierre Neuvial and Etienne Roquain. On false discovery rate thresholding for classification under sparsity. *The Annals of Statistics*, 40(5):2572–2600, 2012.

[PGVW07] M Perone Pacifico, Christopher Genovese, I Verdinelli, and Larry Wasserman. Scan clustering: A false discovery approach. *Journal of Multivariate Analysis*, 98(7):1441–1469, 2007.

[PPGVW04] M Perone Pacifico, C Genovese, I Verdinelli, and L Wasserman. False discovery control for random fields. *Journal of the American Statistical Association*, 99(468):1002–1014, 2004.

[PRBR17] Franck Picard, Patricia Reynaud-Bouret, and Etienne Roquain. Continuous testing for poisson process intensities: A new perspective on scanning statistics. *arXiv preprint arXiv:1705.08800*, 2017.

[PvdL04] Katherine S Pollard and Mark J van der Laan. Choice of a null distribution in resampling-based multiple testing. *Journal of Statistical Planning and Inference*, 125(1):85–100, 2004.

[RBWJ17] Aaditya Ramdas, Rina Foygel Barber, Martin J Wainwright, and Michael I Jordan. A unified treatment of multiple testing with prior knowledge using the p-filter. *arXiv preprint arXiv:1703.06222*, 2017.

[Roq11] Etienne Roquain. Type i error rate control in multiple testing: a survey with proofs. *Journal de la Société Française de Statistique*, 152(2):3–38, 2011.

[RRJW17] Maxim Rabinovich, Aaditya Ramdas, Michael I Jordan, and Martin J Wainwright. Optimal rates and tradeoffs in multiple testing. *arXiv preprint arXiv:1705.05391*, 2017.

[RW07] Joseph P Romano and Michael Wolf. Control of generalized error rates in multiple testing. *The Annals of Statistics*, pages 1378–1408, 2007.

[RYWJ17] Aaditya Ramdas, Fanny Yang, Martin J Wainwright, and Michael I Jordan. Online control of the false discovery rate with decaying memory. In *Advances In Neural Information Processing Systems*, pages 5650–5659, 2017.

[RZWJ18] Aaditya Ramdas, Tijana Zrnic, Martin Wainwright, and Michael Jordan. Saffron: an adaptive algorithm for online control of the false discovery rate. *arXiv preprint arXiv:1802.09098*, 2018.

[SC07] Wenguang Sun and T Tony Cai. Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association*, 102(479):901–912, 2007.

[Sto02] John D Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002.

[Sto07] John D Storey. The optimal discovery procedure: a new approach to simultaneous significance testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3):347–368, 2007.

[STS04] John D Storey, Jonathan E Taylor, and David Siegmund. 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, 2004.

[SZY11] DO Siegmund, NR Zhang, and B Yakir. False discovery rate for scanning statistics. *Biometrika*, 98(4):979–985, 2011.

[WY93] Peter H Westfall and S Stanley Young. *Resampling-based multiple testing: Examples and methods for p-value adjustment*, volume 279. John Wiley & Sons, 1993.

[XZZT17] Fei Xia, Martin J Zhang, James Y Zou, and David Tse. Neuralfdr: Learning discovery thresholds from hypothesis features. In *Advances in Neural Information Processing Systems*, pages 1541–1550, 2017.

[YB99] Daniel Yekutieli and Yoav Benjamini. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference*, 82(1):171–196, 1999.