Contributions to multiple testing theory for high-dimensional data

Etienne Roquain

UNIVERSITÉ PIERRE ET MARIE CURIE

HABILITATION À DIRIGER DES RECHERCHES

Auteur :
Etienne ROQUAIN

Contributions to multiple testing theory for high-dimensional data

Laboratoire de Probabilités et Modèles Aléatoires

Rapporteurs :

Prof. Yoav Benjamini	-	Tel Aviv University
Dir. Stéphane Robin	-	Institut National de la Recherche Agronomique
Prof. Larry Wasserman	-	Carnegie Mellon University
Prof. Michael Wolf	-	University of Zurich

Soutenue le 21 septembre 2015 devant le jury composé de :

Prof. Yoav Benjamini	-	Tel Aviv University	-	Rapporteur
Prof. Gérard Biau	-	Université Pierre et Marie Curie	-	Examinateur
Prof. Lucien Birgé	-	Université Pierre et Marie Curie	-	Examinateur
Prof. Pascal Massart	-	Université Paris-Sud	-	Président
Dir. Catherine Matias	-	Centre National de la Recherche Scientifique	-	Examinatrice
Prof. Gilles Pagès	-	Université Pierre et Marie Curie	-	Examinateur
Dir. Patricia Reynaud-Bouret	-	Centre National de la Recherche Scientifique	-	Examinatrice
Dir. Stéphane Robin	-	Institut National de la Recherche Agronomique	-	Rapporteur

Acknowledgements

I am very grateful to the four referees Yoav Benjamini, Stéphane Robin, Larry Wasserman and Michael Wolf, who kindly accepted to review this manuscript. It has been a great pleasure and honor for me to have them as reviewers. I also feel extremely privileged that Gérard Biau, Lucien Birgé, Pascal Massart, Catherine Matias, Gilles Pagès and Patricia Reynaud-Bouret so kindly and spontaneously agreed to be part of the habilitation committee.

Going back to the earliest stages of my scientific life, I would like to thank very much my PhD advisors Gilles Blanchard and Sophie Schbath for their advice and guidance. I am now able to better appreciate how crucial this was for my career. Next, I would like to warmly thank my co-authors Sylvain Arlot, Sylvain Delattre, Thorsten Dickhaus, Kyung-In Kim, Pierre Neuvial, Fanny Villers, Mark van de Wiel, for their optimism when facing uncooperative equations. My sincere gratitude also goes to the "Statistique et Génome" team for making me feel welcome during my "délégation CNRS" in Évry.

I would like to warmly thank everyone who helped me during the writing of this manuscript, in particular Sylvain Arlot, Pierre Neuvial and Fanny Villers for their accurate comments and Mark van de Wiel for his help with the *R*-package "dnaCplusT". Also many thanks to my sister who corrected certain non-English sentences. I am also grateful to Tabea Rebafka and Lucien Birgé for their careful review of my multiple testing survey a few years ago.

Thank you to the members of the laboratory LPMA and LSTA: the administrative and technical staff for their efficiency, all the colleagues for many interesting discussions (scientific or not) during lunch and more specifically to Sonia Fourati, Eric Saias and Fanny Villers who have patiently endured my presence in their office!

My last thanks are to my family, wife and each of my three kids.

Foreword

This manuscript provides a mathematical study of the multiple testing problem in settings motivated by modern applications, for which the number of variables is much larger than the sample size. As we will see, this problem highly depends on the nature of the data and of the desired interpretation of the results.

Chapter 1 is a wide introduction to the multiple testing theme which is intended to be accessible for a possibly non-specialist reader and which includes a presentation of some high-dimensional genomic data. The necessary probabilistic materials are then introduced in Chapter 2, while Chapters 3, 4, 5 and 6 are guided by the findings [P2, P3, P4, P5, P6, P7, P8, P9, P10, P11, P12, P15, P16] listed page 81. Nevertheless, compared to the original papers, I have tried to simplify and unify the studies as much as possible. An effort has also be done for presenting self-contained proofs when possible. Let us also mention that, as an upstream work, I have proposed a survey paper in [P13]. Nevertheless, the overlap between that work and the present manuscript turns out to be only minor.

The publications [P1, P2, P3, P6, P7, P14] correspond to a research (essentially) carried out during my PhD period at the Paris-Sud University and INRA[1]. The papers [P15, P11] are related to my postdoctoral position at the VU University Amsterdam. I have elaborated the work [P4, P5, P8, P9, P10, P12, P13, P16] afterwards, as a "maître de conférences" at the Pierre et Marie Curie University in Paris.

Throughout this manuscript, we will see that while the multiple testing problem occurs in various practical and concrete situations, it relies on an astonishingly wide variety of theoretical concepts, as combinatorics, resampling, empirical processes, concentration inequalities, positive dependence, among others. This symbiosis between theory and practice explains the worldwide success of the multiple testing research field, which has become a prominent research area of contemporary statistics.

[1] Institut National de la Recherche Agronomique.

Contents

Notation

- m: number of null hypotheses to be tested (number of variables);
- n: sample size (number of individuals);
- $H_{0,i}$ (resp. $H_{1,i}$), $1 \leq i \leq m$, the null (resp. alternative) hypotheses to be tested;
- $\theta \in \{0,1\}^m$: underlying configuration of true/false nulls hypotheses, i.e., $\theta_i = 0$ if and only if $H_{0,i}$ is true;
- $\mathcal{H}_0(\theta)$ (resp. $\mathcal{H}_1(\theta)$): index set corresponding to true (resp. false) nulls;
- $m_0(\theta)$ (resp. $m_1(\theta)$): number of true (resp. false) nulls;
- π_0: depending on the context, π_0 can denote the value of $m_0(\theta)/m$ (non-asymptotic setting), the limit value of $m_0(\theta)/m$ (asymptotical setting), or the probability that $\theta_i = 0$ (random effects setting);
- $\{p_i(X),\, 1 \leq i \leq m\}$: family of p-values;
- $\widehat{\mathbb{G}}_m$: empirical distribution function of the p-values;
- τ_ℓ, $1 \leq \ell \leq m$: sequence of critical values;
- $R \subset \{1, \ldots, m\}$: multiple testing procedure;
- $\mathrm{SU}(\tau)$ (resp. $\mathrm{SD}(\tau)$, $\mathrm{SUD}_\lambda(\tau)$): step-up (resp. step-down, step-up-down) procedure with critical values τ_ℓ, $1 \leq \ell \leq m$.
- μ (resp. Γ): mean (resp. covariance matrix) of the observed random variable X when $X \in \mathbb{R}^m$ is multivariate Gaussian;
- Δ: common value of the alternative means when they are assumed to be all equal and positive;
- $\|y\|_q = \left(m^{-1} \sum_{i=1}^m |y_i|^q\right)^{1/q}$ for $q \in [1, \infty)$ and $\|y\|_\infty = \sup_{1 \leq i \leq m} |y_i|$, for $y \in \mathbb{R}^m$;
- $\overline{\Phi}(\cdot)$: upper-tail function of a standard Gaussian distribution, i.e., $\overline{\Phi}(z) = \mathbb{P}(Z \geq z)$, $Z \sim \mathcal{N}(0,1)$;
- $\mathcal{D}(Z)$: distribution of Z;
- B: number of resamples in randomized quantities.

Chapter 1

Introduction

This introduction is intended for statisticians who are not specialist of the multiple testing research field. It is deliberately informal and oriented towards simple illustrations. A rigorous formulation will be proposed in Chapter 2. The examples presented in this section have been inspired by several readings, e.g., [143, 90, 31] and by the talks of Christopher Genovese (2004) [58] and Yoav Benjamini (2013) [6].

1.1 From single hypothesis testing ...

Let us first introduce basic notions of single hypothesis testing. Hypothesis testing is a statistical inference that has been conceptualized in the early 20th century by Karl Pearson [102] and then Ronald A. Fisher [52, 53][1]. A test aims at deciding whether a prior hypothesis (null hypothesis) is true or not from repeated observations of a single phenomenon. More formally, a test is a 0/1 decision, based on a random variable X (possibly a sample) that should determine whether the distribution P of X satisfies the null hypothesis, generally denoted by H_0. In case of a rejection, a test chooses an alternative hypothesis, generally denoted by H_1. Each hypothesis formally corresponds to a certain family of possible distributions for P. Two types of errors can occur: rejecting H_0 ("1" decision) while it is true (type I error); or accepting H_0 ("0" decision) while it is false (type II error). The originality of the testing approach is that these two errors are not equivalent; a test primarily focus on bounding the probability of type I error by some $\alpha \in (0,1)$, called the level (of significance) of the test. From an intuitive point of view, this means that the "0" decision is favored: when the sample is uninformative, and H_0 being indifferently true or false, a test of level α cautiously chooses to accept H_0 with probability larger than $1-\alpha$. In this regards, a test is considered as "finding something" only when it rejects H_0.

The probably most common illustration is the case of a n-normal sample $X = (X_1, \dots, X_n)$ where the variables of the sample are i.i.d. and all follow a $\mathcal{N}(\mu, 1)$ distribution, for an unknown $\mu \in \mathbb{R}$. If the problem is to test H_0: "$\mu \leq 0$" against H_1: "$\mu > 0$", the classical Neyman-Pearson test of level α rejects the null H_0 when $n^{1/2}\overline{X}_n = n^{-1/2}\sum_{i=1}^n X_i$ exceeds $\overline{\Phi}^{-1}(\alpha)$, where $\overline{\Phi}(\cdot)$ denotes the upper-tail function of a standard Gaussian distribution. Since the choice of α is quite arbitrary (why considering $\alpha = 5\%$ and not $\alpha = 3.14259\%$?), an interesting way to measure the significance of a test is to consider the largest α for which the test rejects H_0 at level α, called the *p-value* of the test. Here, we can merely check that the p-value of the above test is $p(X) = \overline{\Phi}\left(n^{1/2}\overline{X}_n\right)$. It satisfies the following important stochastic domination property:

$$\text{when } P \text{ satisfies } H_0, \text{ for all } t \in [0,1], \quad \mathbb{P}_{X\sim P}(p(X) \leq t) \leq t, \tag{1.1}$$

[1] For an historical context, we refer the reader to [105] (among others).

which comes from the fact that under the null, the event "$p(X)$ is smaller than α" (i.e., "H_0 is rejected at level α") occurs with a probability smaller than α, because the test is of level α. By nature, the p-value gives the decision of the test at all possible levels. More intuitively, the p-value measures the "plausibleness" of H_0 measured by the test. A "small" p-value provides evidence against H_0 and thus tends to show that a "discovery" is made by the test.

1.2 ... to multiple hypothesis testing

To analyze a given complex phenomenon, a researcher rarely asks only one question; he/she looks at his/her data from different angles in order to gets a wider idea of what is the trueness behind the data. This raises the problem of assessing statistical significance for many features simultaneously, which can be set as the issue of *multiple hypothesis testing*. Historically, the earliest appearance of multiple inference seems to go back to Carlo E. Bonferroni [22], while John W. Tukey's ideas [136] were seminal for multiple inference, as reported in [7]. For additional historical notes and references, we refer the interested reader to [38], [73], [90], [126] and [31]. Compared to earlier work, contemporary multiple hypothesis testing deals with an higher resolution: several recent technological jumps have made the potential number of desired inferences grows from dozens to several thousands or even millions. This appears for instance in practical fields where a massive amount of data can be collected, as microarray analysis [138, 37], neuro-imaging [9, 101] and source detection [96], among others. As we will see in Section 1.3, this makes the issue of multiplicity even more crucial.

The problem of simultaneous significance is, by essence, frustrating: to take fully advantage of the data, it is tempting to perform many inferences simultaneously and to neglect the multiplicity issue. However, this is incorrect. Simple paradoxical situations can be employed to illustrate the latter. A first instance is that most clinical trials can be made significant; suppose that we have at hand data coming from clinical trials, with some characteristics for the patients (say, sex, age and geographical location). If the test using the whole sample is not significant, certainly a part of it should be significant, at least by looking carefully enough. For this, we can subdivide the sample in many subgroups by using the patients characteristics. At the end, we should find that, say, Scandinavian women aged 51-60 are significantly affected by a drug. This way to add "chances of winning" is referred to as the "Munchhausen's Statistical Grid" by Dr. Graham Martin, see Appendix I of [143]. A similar humorous story is provided in the comic strips at the beginning of each chapter (source: http://xkcd.com).

The above "data snooping" processes are inappropriate because they generate items that are declared wrongly significant, often referred to as a *false discoveries* or *false positives*. At this point, it is useful to recall the Young's False Positive Rules, see [143] page 7:

(i) With enough testing, false positives *will* occur.

(ii) Internal evidence will not contradict a false positive result.

(iii) Good investigators will come up with a possible explanation.

(iv) It only happens to the other persons.

An elementary probabilistic argument supporting item (i) is that, if m tests are performed simultaneously and if A_i is the event "make a false positive for the i-th null hypothesis", then the probability that this multiple decision makes at least one false positive is

$$\mathbb{P}\left(\bigcup_{i=1}^{m} A_i\right), \tag{1.2}$$

which is a potentially much larger probability than each individual errors $\mathbb{P}(A_i)$, $1 \leq i \leq m$. The value of the probability (1.2) is displayed in Figure 1.1 (left), in the independent case.

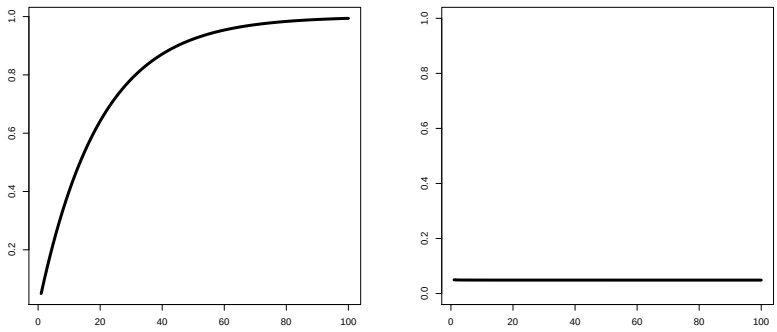


Figure 1.1: Graphical representations of $m \mapsto 1-(1-0.05)^m$ (left) and $m \mapsto 1-(1-0.05/m)^m$ (right).

The consequence is that the level at which each individual test is performed should be considerably reduced. This idea, going back to Tukey, is now commonly referred to as the "Higher criticism", see [32], or simply by "multiple testing correction" or "multiple testing adjustment". The most simple example is the Bonferroni correction, which is simply to take α/m instead of α in the individual tests. This stabilizes the probability (1.2), see Figure 1.1 (right), in the independent case. Hence, the probability to make at least one false discovery, called the family-wise error rate (FWER), is ensured to be below α. However, this approach might be unsatisfactory in practice, because applying the correction reduces considerably the quantities of discoveries, especially when m is large. Furthermore, with the development of new high-throughput technologies, for which m goes far beyond several thousands (see Section 1.3), this problem became an increasing cause of concern.

A renewed interest in multiple testing correction indisputably occured after the paper [10] of Yoav Benjamini and Yosef Hochberg (1995). They introduced a new simple procedure, baptized later "the BH procedure", which can be seen as filling the gap between an uncorrected procedure (too many discoveries) and the Bonferroni procedure (too few discoveries), while controlling a global error rate called the false discovery rate (FDR). The latter is defined informally as the average of the (random and unknown) quantity

$$\text{FDP} = \frac{\text{number of false discoveries}}{\text{number of discoveries}}.$$

The idea is that, in an exploratory research, making a few more false discoveries may be tolerated if this yields a substantial increase in the total number of discoveries. For instance, making 4 false discoveries out of 10 discoveries (FDP = 0.4) can be viewed as less acceptable than making 7 false discoveries out of 100 discoveries (FDP = 0.07). This more optimistic view on the total amount of false discoveries made the multiple testing correction more appealing for the practitioner, because it allows to make much more discoveries while still keeping an overall statistical guarantee.

The philosophical difference between FWER control (with Bonferroni procedure) and FDR control (with BH procedure) is illustrated on Figure 1.2 on a toy example. On a two dimensional grid, the signal lies in a disk (in gray, with a strength linearly decreasing from the center to the border of the disk) and the observations simply are i.i.d. Gaussian perturbations of the signal. For each method, the rejected items are marked by black dots. FWER control ensures that no detection will be made outside the disk (with probability at least 0.95), while FDR control ensures that the number of detections outside the disk out of the number of total detections is, on average, less than 0.05. A consequence is that BH procedure automatically adapts to the "amount of detectable signal" contained in the data: for weak signal and a small circle (topleft), BH behaves like a Bonferroni procedure, while for a strong signal

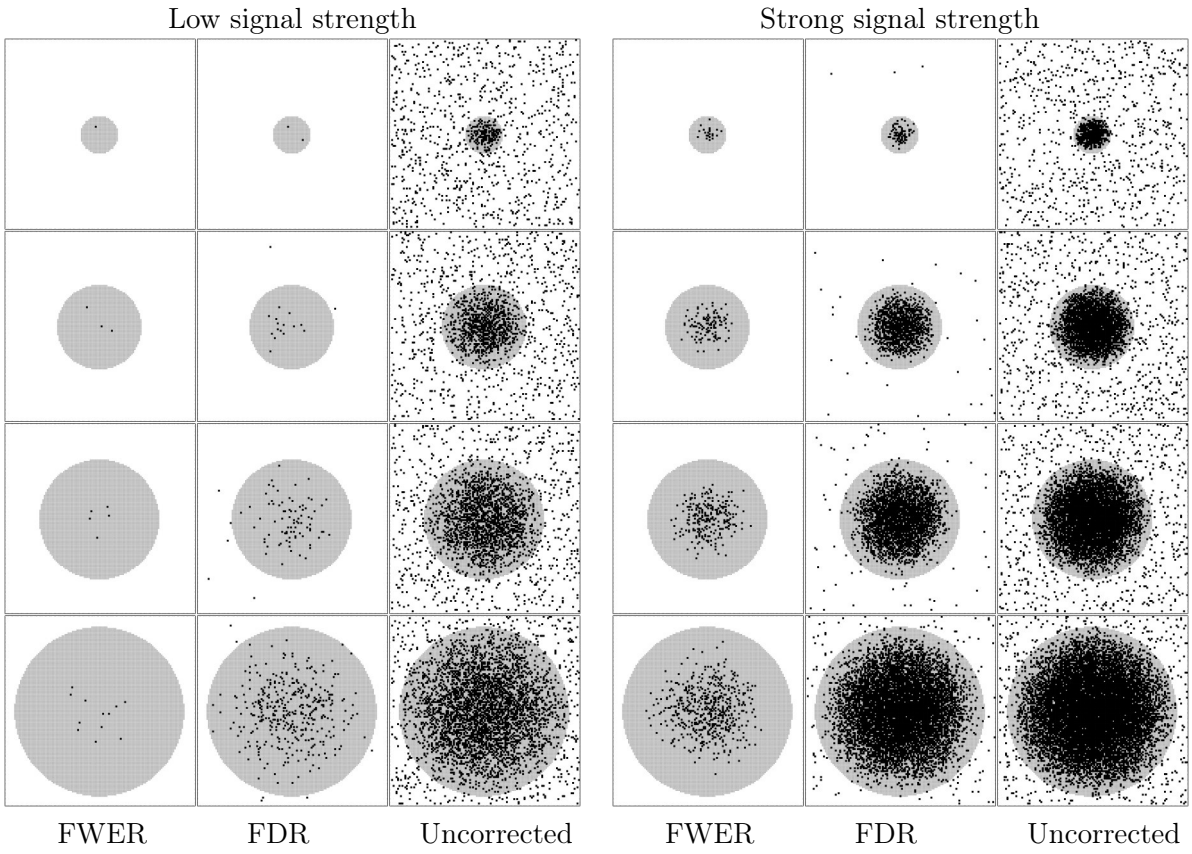


Figure 1.2: Discoveries for FWER (left), FDR (middle) or without correction (right) at level 0.05 in an independent two-dimensional setting where $m = 128^2$ items are tested. See text.

and a large circle (bottomright), BH acts more like a non-corrected procedure. Hence, FDR control allows (many) more true detections than FWER control, at the price of an amount of false positives that looks appropriate in all situations.

To get an idea of the impact of the FDR in the scientific landscape, Figure 1.3 provides the citations of the paper [10] between 1996 and 2013 as reported by "the web of science". In addition to the number of citations, which is considerable for a statistical paper, the most noteworthy fact might be lying in the variety of the impacted fields of research.

1.3 Multiple testing in genomic data

One of the most emblematic situations where a large number of tests should be performed simultaneously arises with the analysis of genomic data that comes from high-throughput technologies (microarrays or more recently next generation sequencing). In a typical exploratory research, the practitioner wants to relate a given type of cellular information (e.g., gene expression, copy number alterations or genotype) to some phenotype (e.g., a type of disease). This involves thousands or even hundreds of thousands underlying items for which a decision should be inferred (e.g., genes, probes or SNPs). While individual tests can be often easily built, the question of providing an overall error rate control is central in such high-dimensional contexts. Some examples of genomic data are given below.

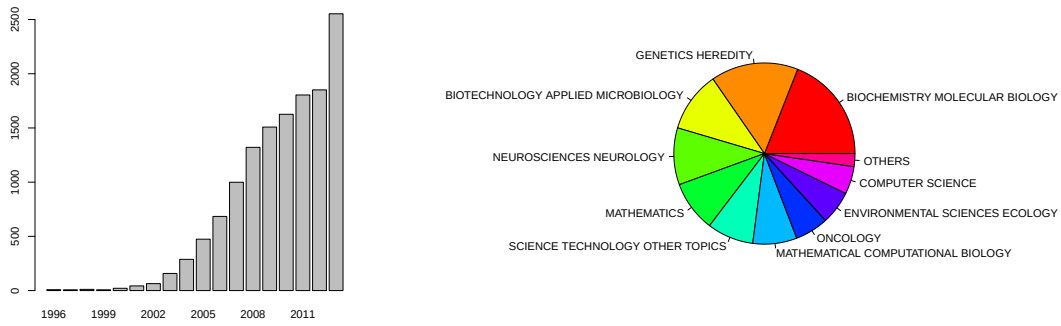


Figure 1.3: Statistics for the 13,427 papers citing [10] from 1996 to 2013 and according to "the web of science". Left: per year. Right: per research field.

Gene expressions Figure 1.4 (bottom) displays an instance of data resulting from microarray experiments. Each spot corresponds to a gene (or a portion of it), and its color codes for a level of expression. Typically, the latter is obtained via RNA measurements which have been transformed into a real number by using standard normalization steps. Hence, from a mathematical point of view, the data-set is a huge real-valued $n \times m$ matrix with $m \gg n$, that is, with many variables (m genes) and only few repetitions (n individuals). For instance, $m = 11,169$ and $n = 42$ in the data set of [97]. Additionally, there are potentially many dependencies between gene expressions, because some genes may activate/inhibit others (along so-called "pathways"). Many data source are available on the web, see for instance http://strimmerlab.org/data.html; http://www.ncbi.nlm.nih.gov/geo/; http://bioconductor.org.

DNA copy number alterations Typically, while normal cells have 2 copies of each chromosome, tumor cells often have chromosomal alterations at some positions of the genome. Such abberations can correspond either to a deletion (< 2 copies) or to a gain (> 2 copies). Studying how these copy number alterations (CNAs[2] in short) are related to some phenotype can be useful to study various type of cancers (e.g., to find the genome positions related to the cancer development process). The common technology used to measure CNAs is the array Comparative Genomic Hybridization (CGH) which is a special type of microarray. It compares the DNA quantity of the test sample to a reference sample at each location (probe) along the genome via a fluorescence device. After some transformations, this array can be loosely encoded as loss, normal and gain (< 2, $= 2$ or > 2). Such a resulting data set is displayed in Figure 1.4 (top); white (resp. black) codes for loss (resp. gain). Many CNA data are available on the web, see, e.g., http://cancergenome.nih.gov/. Again, the chromosomal aberrations are measured along the genome via a huge number of probes, which naturally entails a multiplicity issue. Among many available methods, Section 5.3 will provide a solution for analysing CNA data.

Genome-wide association studies Markedly, while most the bases of the genome are identical across a given human population, some specific (tiny) genomic regions vary among individuals. Typical such regions are the single nucleotide polymorphisms (SNPs), for which the variation is only carried by one basis. For each SNP, the pair of values (e.g., `(a,t)`) taken at that location (one value for each chromosome) is a characteristic of interest, that can be summarized as a copy number variation

[2] Also sometimes called copy number variations (CNVs). Here, we distinguish copy number alterations (CNAs) and copy number variations (CNVs). While the CNAs are expected to happen in some tissues (often tumor cells), the second are expected to occur in all the cells of the individuals (typically inherited from parents).

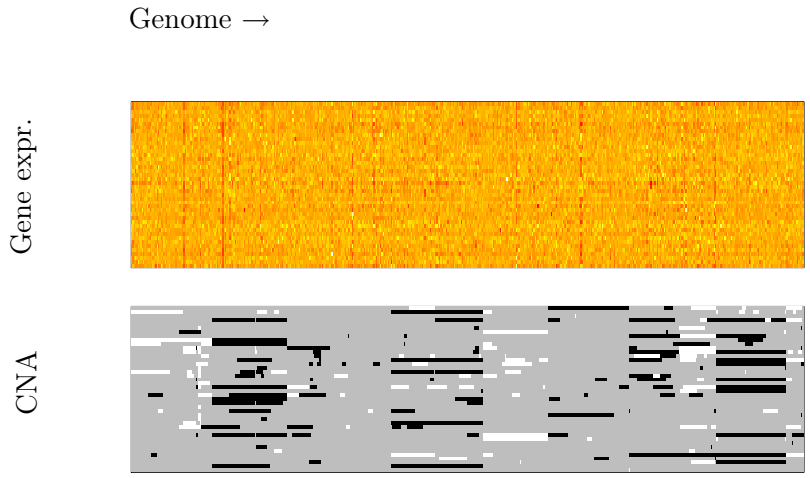


Figure 1.4: Two types of microarray data, see text. 500 probes taken along the genome. 42 individuals having a lymphoma cancer under specific conditions, see [97].

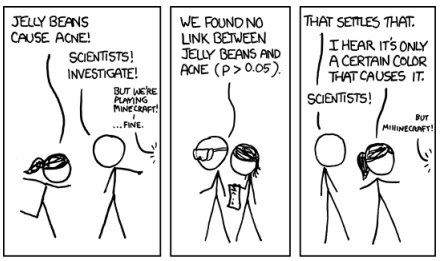
(CNV) of a reference basis (say, a). Across the genome, these CNVs can be used as a *genetic marker* to explain some *phenotype* (e.g., obesity, diabetes, anorexia). Nowadays, a massive amount of such data sets are provided by giant consortium of scientists, e.g., "The Wellcome Trust Case Control Consortium" http://www.wtccc.org.uk/, "The International HapMap Project" http://hapmap.ncbi.nlm.nih.gov/ or "The 1000 Genomes Project" http://www.1000genomes.org/, typically involving a number m of SNPs that can reach several millions. Moreover, underlying biological processes can imply dependencies between the genetic markers.

1.4 Big data?

In this section, we briefly situate our work with respect to the recent "big data" phenomenon. The most general definition of "big data" science is the treatment of massive amount of data. This admittedly vague definition makes the work investigated here part of this trendy concept. However, this might be wrong. A more accurate definition for "big data" is data that are so massive that we cannot upload it onto a standard computer. Hence, applying solely a statistical method is irrelevant and a task at the border of algorithmics, informatics and statistics should be investigated. This task is *not* explored in this manuscript.

Nevertheless, by somehow anticipating the future work that could be investigated in these "big data" research field, it is hard to believe that the multiple testing issue will disappear. Actually, it should be even more crucial. For more details, we refer the reader to the interview of Michael Jordan by Lee Gomes [66] and in particular to the discussion around "Why Big Data Could Be a Big Fail".

In conclusion, while the work investigated here *is not* made for "big data analysis" (with the accurate definition), we can safely argue that some of the tools developed here are likely to be useful for future work in that area.



...

Chapter 2

Probabilistic preliminaries

This chapter presents a general mathematical background which will be used throughout the manuscript.

2.1 General statistical setting

Let X be a random variable valued in an observation space $(\mathcal{X}, \mathfrak{X})$ and coming from an underlying measurable space $(\Omega, \mathfrak{F})$. The distribution of X on $(\mathcal{X}, \mathfrak{X})$ is denoted by P and is assumed to belong to some set $\mathcal{P}$, which is called the model. For each $P \in \mathcal{P}$, we assume that there exists a distribution on $(\Omega, \mathfrak{F})$ for which $X \sim P$; it is referred to as $\mathbb{P}_{X\sim P}$ or simply by $\mathbb{P}$ when unambiguous. The corresponding expectation operator is denoted $\mathbb{E}_{X\sim P}$ or $\mathbb{E}$ for short.

Let $m \geq 2$ be the number of null hypotheses to be tested. For $i \in \{1, \ldots, m\}$, let $H_{0,i}$ be a null hypothesis for P, that corresponds to some subset $\mathcal{P}_{0,i}$ of $\mathcal{P}$. The goal of a multiple testing decision is to infer from X whether P is in $\mathcal{P}_{0,i}$, for each $i \in \{1, \ldots, m\}$. Hence, the parameter of interest is the underlying "configuration"

$$\theta(P) \in \{0,1\}^m, \text{ where } \theta_i = 0 \text{ if and only if } P \in \mathcal{P}_{0,i}. \tag{2.1}$$

Conversely, for each configuration $\theta \in \{0,1\}^m$, we can define $\mathcal{P}_\theta$ the subset of the distributions of $\mathcal{P}$ that are compatible with the configuration θ, that is, $\mathcal{P}_\theta = \{P \in \mathcal{P} \,:\, \theta(P) = \theta\}$ (possibly empty). Hence, choosing $P \in \mathcal{P}$ is equivalent to first choose a configuration $\theta \in \{0,1\}^m$ and then to choose $P \in \mathcal{P}_\theta$. Separating θ from $P \in \mathcal{P}_\theta$ is sometimes convenient, so this distinction should be kept in mind. We also denote $\mathcal{H}_0(\theta) = \{1 \leq i \leq m : \theta_i = 0\}$, $m_0(\theta) = \sum_{i=1}^m (1 - \theta_i)$ and $\mathcal{H}_1(\theta) = \{1 \leq i \leq m : \theta_i = 1\}$, $m_1(\theta) = \sum_{i=1}^m \theta_i$ the set/number of true and false null coordinates, respectively.

As a first illustration, let $X = (X_1, X_2)$ be a vector of $m = 2$ independent variables with $X_i \sim \mathcal{N}(\mu_i, 1)$, $i \in \{1, 2\}$, and consider the test of $H_{0,i}$: "$\mu_i = 0$" against $H_{1,i}$: "$\mu_i \neq 0$", for any $i \in \{1, 2\}$. There is $2^m = 4$ possible configurations: $\theta = (0,0)$ ($\mu_1 = \mu_2 = 0$); $\theta = (1,0)$ ($\mu_1 \neq 0$, $\mu_2 = 0$); $\theta = (0,1)$ ($\mu_1 = 0$, $\mu_2 \neq 0$); $\theta = (1,1)$ ($\mu_1, \mu_2 \neq 0$). The aim of a multiple testing decision is to choose among these four configurations. This inference does not concern the values of the non-zero μ_i's (that is, does not concern which distribution $P \in P_\theta$ is followed by X), at least not directly.

In the sequel, we will focus on decisions based upon p-values. Hence, a basic assumption is that for each $i \in \{1, \ldots, m\}$, there is a random variable $p_i(X)$, called p-value, satisfying the following assumption

$$\forall P \in \mathcal{P}_{0,i}, \text{ we have } \forall t \in [0,1], \mathbb{P}_{X\sim P}(p_i(X) \leq t) \leq t, \tag{pvalueprop}$$

or, equivalently, $\forall \theta \in \{0,1\}^m$ with $\theta_i = 0$, $\forall P \in \mathcal{P}_\theta$, we have $\forall t \in [0,1]$, $\mathbb{P}_{X\sim P}(p_i(X) \leq t) \leq t$. We will sometimes denote $p_i(X)$ simply by p_i for short. Property (pvalueprop) means that each p-value of

$(p_i(X), i \in \mathcal{H}_0(\theta))$ must be stochastically lower-bounded by a uniform variable. Hence, (pvalueprop) can be seen as a generalization of Assumption 1.1 to the case of multiple null hypotheses. In many cases, the p-values under the null are exactly uniformly distributed, that is, for all $i \in \{1, \dots, m\}$,

$$\forall P \in \mathcal{P}_{0,i}, \text{ we have } \forall t \in [0,1], \mathbb{P}_{X\sim P}(p_i(X) \leq t) = t. \tag{pvaluepropunif}$$

This is slightly less general than (pvalueprop). Furthermore, a specificity of the (multiple) testing setting is that, while it requires a strong assumption on the distribution of $p_i(X)$ under the null, no such assumption is made in general under the alternative $P \notin \mathcal{P}_{0,i}$. However, in addition to (pvalueprop), specific dependency structures can be assumed for the p-value family, see Section 2.4. Alone, the assumption (pvalueprop) is often referred to as *general dependence*.

A canonical example is the case where we test whether the coordinates of the mean of a multivariate Gaussian vector are zero or not. Let X be a multivariate Gaussian vector of mean $\mu \in \mathbb{R}^m$ and covariance Γ, assumed to satisfied $\Gamma_{i,i} = 1$ for simplicity. Typically, two different multiple testing problems can be investigated:

- one-sided: $H_{0,i}$: "$\mu_i \leq 0$" against $H_{1,i}$: "$\mu_i > 0$" , $1 \leq i \leq m$;
- two-sided: $H_{0,i}$: "$\mu_i = 0$" against $H_{1,i}$: "$\mu_i \neq 0$" , $1 \leq i \leq m$;

Also, simpler one-sided nulls can be considered with $H_{0,i}$: "$\mu_i = 0$" against $H_{1,i}$: "$\mu_i > 0$" , $1 \leq i \leq m$, which implicitly assumes that μ has nonnegative coordinates. In the one-sided (resp. two-sided) case, classical p-values are given by $p_i(X) = \overline{\Phi}(X_i)$ (resp. $p_i(X) = 2\overline{\Phi}(|X_i|)$), $1 \leq i \leq m$. We can merely check that (pvalueprop) is satisfied. In addition, when $H_{0,i}$ is "$\mu_i = 0$" for all i (one-sided or two-sided), the p-values also satisfy the stronger condition (pvaluepropunif).

While the dependence parameter Γ is generally unknown, it can be known in some specific multiple testing situations. A simple example is provided by the regular Gaussian linear model with a full rank design matrix. In high dimension, this is also the case when testing marginal associations, see [44]. In some cases[1], Γ can also come from external experiments.

Finally, note that the Gaussian modeling can also be useful to approximate some non-Gaussian multiple testing situations, see [26].

2.2 Global p-value thresholding

In the above multivariate Gaussian example, Neyman-Pearson's lemma indicates that the best individual decision when testing $H_{0,i}$ against $H_{1,i}$ is to reject $H_{0,i}$ when p_i is smaller than a threshold. Hence, a compatible multiple testing decision will reject the nulls corresponding to p-values smaller than some thresholds. However, since the tests are performed simultaneously, these thresholds should be conveniently adjusted to take into account the multivariate aspect of the distribution of the p-value family.

Figure 2.1 provides a useful graphical scheme of the multiple testing problem in the p-value-based setting. The data are generated in the one-sided Gaussian framework for $m = 100$, $m_0 = 50$, $\mu_i \in \{0, 1\}$ and $\Gamma = I$ (independence between the tests). Also, for each $p_i = p_{(\ell)}$, the associated θ_i is marked by "0" if $\theta_i = 0$ (comes from a null) or by " $\times$ " if $\theta_i = 1$ (comes from an alternative). Since we aim at rejecting nulls corresponding to "small" p-values, it is natural to order them as follows:

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

[1]For instance, in Genome-wide association studies, the dependency structure can be related to the linkage disequilibrium phenomenon, see Chapter 9 in [31].

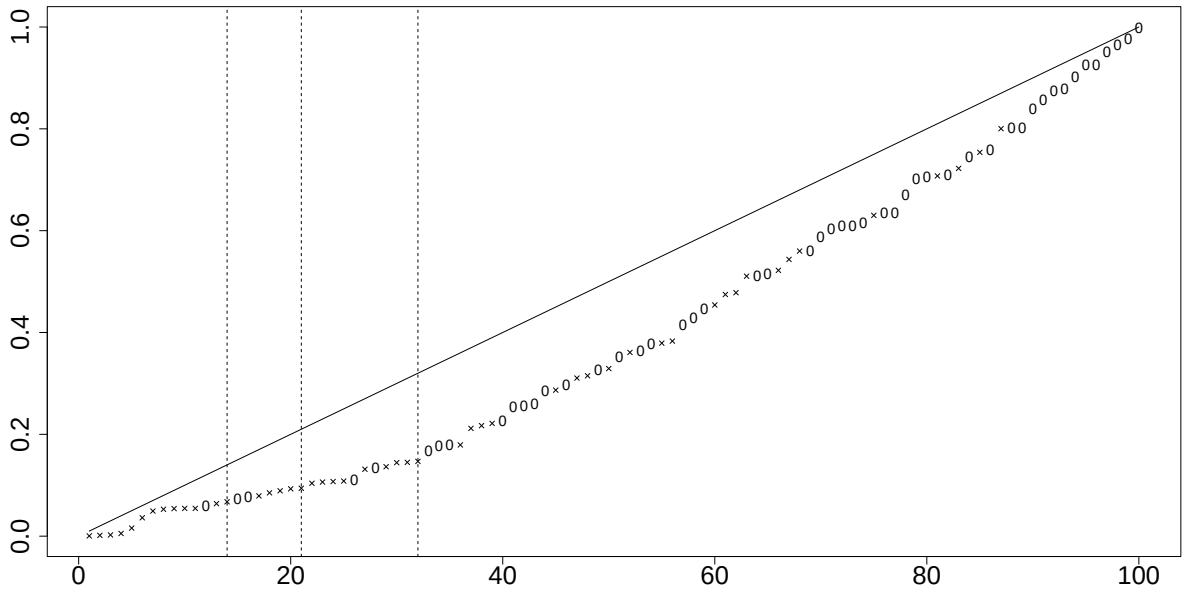


Figure 2.1: Pictorial representation of the multiple testing issue based on p-value ordering, see text.

and to reject the corresponding nulls until some "suitable" rank. For the realization reported in Figure 2.1, it is not totally clear what is the most desirable decision to make, *even if we knew the label*: while rejecting the 14 first null hypotheses (first dashed vertical line) certainly ensures no false discovery, maybe rejecting the 21 first nulls, or even the 32 first nulls (second and third dashed vertical lines) is better in order to include more true discoveries, up to make few additional false discoveries. This points out the need of choosing an appropriate criterion.

2.3 Criteria and decisions

A general way to define a multiple testing procedure is to describe the index set of the null hypotheses that it rejects. Formally, a *multiple testing procedure* is a function

$$R : \omega \in \Omega \mapsto R(\omega) \subset \{1, \dots, m\}$$

such that, for all $i \in \{1, \dots, m\}$, the event $\{\omega \in \Omega : i \in R(\omega)\}$ is measurable, that is, lies in $\mathcal{F}$. As mentioned above, it is of interest to consider the class of *p-value thresholding-based* multiple testing procedures, which are of the form

$$R = \left\{1 \leq i \leq m \ : \ p_i(X) \leq \widehat{t}\,\right\}, \tag{2.2}$$

where $\widehat{t} \in [0, 1]$ is some random variable (itself possibly depending on the family of the $p_i(X)$'s). The default choice in this manuscript is to use (2.2) with a non-strict inequality. However, in some specific cases, using it with a strict inequality is more convenient, in which case we will explicitly mention it in the text. For instance, the form (2.2) includes the non-corrected procedure and the Bonferroni procedure that use $\widehat{t} = \alpha$ and $\widehat{t} = \alpha/m$, respectively. For short, we will sometimes say that $\widehat{t}$ is itself a multiple testing procedure.

2.3.1 Family-wise error rates

For a multiple testing procedure R, the set of the false discoveries corresponds to $R \cap \mathcal{H}_0(\theta)$. Various type I error rates have been proposed in the literature to measure the "size" of this set. We focus in this manuscript only on those commonly used. The probably earliest one is the *family-wise error rate* (FWER), which is defined as follows: for all $P \in \mathcal{P}$,

$$\mathrm{FWER}(R, P) = \mathbb{P}(|R \cap \mathcal{H}_0(\theta)| \geq 1). \tag{2.3}$$

Hence, when controlling the FWER at level α, it is asked that, with a probability larger than $1-\alpha$, we have $|R \cap \mathcal{H}_0(\theta)| = 0$, that is, no false discovery is made. As an illustration, the event "$|R \cap \mathcal{H}_0(\theta)| = 0$" occurs in Figure 2.1, provided that R only contains items "$\times$" (e.g., when R rejects the 14 first nulls). We merely check that the Bonferroni procedure controls the FWER under general dependence:

$$\mathbb{P}(\exists i \in \mathcal{H}_0(\theta) \ : \ p_i(X) \leq \alpha/m) \leq \sum_{i=1}^{m} (1-\theta_i)\mathbb{P}(p_i(X) \leq \alpha/m) \leq m_0\alpha/m \leq \alpha,$$

by combining (pvalueprop) and an union bound argument. When the union bound is accurate (e.g., under independence, when α is small and $m_0 \simeq m$), controlling the FWER with Bonferroni procedure is satisfactory in the sense that the error rate will be close to α. However, under "strong" dependence, Bonferroni procedure has an FWER that can be far below α and a more accurate control can be obtained by circumventing the union bound argument: for a thresholding based procedure R of the form (2.2),

$$\mathrm{FWER}(R, P) = \mathbb{P}\left(\inf_{i \in \mathcal{H}_0(\theta)} \{p_i(X)\} \leq \widehat{t} \right). \tag{2.4}$$

In other words, for controlling the FWER, we should consider $\widehat{t}$ according to the quantile of the distribution of the smallest p-value among those under the null (i.e., the distribution of the first point labeled "0" in Figure 2.1). This raises several questions related to the distribution of this infimum, especially the issue of obtaining a conservative estimate of $\widehat{t}$ while "learning" both the dependence and the set $\mathcal{H}_0(\theta)$ from the data. This issue will be investigated in Section 5.1.

While the FWER control has a clear interpretation, it might be too strict, especially when m is large. Several criteria aim at relaxing it. Going back to Figure 2.1, a fair idea seems to tolerate the first "0" in order to make the next 6 true discoveries. This motivates the introduction of the *k-family-wise error rate* (k-FWER) (see, e.g., [73, 115, 89]): for all $P \in \mathcal{P}$,

$$k\text{-FWER}(R, P) = \mathbb{P}(|R \cap \mathcal{H}_0(\theta)| \geq k), \tag{2.5}$$

where $k \in \{1, ..., m\}$ is a pre-specified "tolerance" parameter. Note that for $k = 1$, the k-FWER reduces to the FWER. When controlling the k-FWER at level α, it is asked that, with a probability larger than $1-\alpha$, the rejection set contains less than $k-1$ false discoveries, that is, $|R \cap \mathcal{H}_0(\theta)| \leq k-1$. As one might expect, the techniques involved while controlling the k-FWER are quite similar than those related to FWER, because, loosely, the infimum has simply to be replaced by the k-th infimum. Note however that it might add some difficulties when "learning" the set $\mathcal{H}_0(\theta)$, see [116].

2.3.2 False discovery rate

The main drawback of k-FWER is that the choice of k should be prescribed *a priori*, so seems arbitrary: for instance, choosing $k = 6$ provides that, with high probability, there is at most $k-1 = 5$ errors in R. While this seems fair for $|R| = 100$ (say), it is less relevant for $|R| = 10$ (say). The latter example

relies on a seminal idea: we should use a criterion that *relates the quantity of tolerated false discoveries to the number of discoveries*. To this end, the *false discovery proportion* (FDP) is defined as follows: for all $P \in \mathcal{P}$,

$$\mathrm{FDP}(R, P) = \frac{|R \cap \mathcal{H}_0(\theta)|}{|R| \vee 1}, \tag{2.6}$$

where "$\vee$" denotes the maximum operator (hence $\mathrm{FDP}(R, P) = 0$ if $|R| = 0$). The latter unobservable quantity is not an error rate, because it is random. An error rate can be built out of it by taking its expectation: the *false discovery rate* (FDR) is defined as follows: for all $P \in \mathcal{P}$,

$$\mathrm{FDR}(R, P) = \mathbb{E}[\mathrm{FDP}(R, P)]. \tag{2.7}$$

While the underlying idea leading to the FDR criterion has several early occurrences in statistical literature, see [125, 123, 128], it has been formalized in [10] and becomes popular afterwards. Now, deriving an FDR control of the form $\mathrm{FDR}(R, P) \leq \alpha$ means that, on average, there are less than $\alpha|R|$ errors among the discoveries of R. For instance, for $\alpha = 0.05$, $|R| = 1000$ means that R contains at most 20 false discoveries (on average). The FDR control has thus a clear and attractive interpretation. Another explanation of the FDR success is that there is a simple procedure that controls it (see Figure 2.2, left, for an illustration):

Algorithm 2.1. [BH procedure at level α]

- Order the p-values as $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$ and let $p_{(0)} = 0$;
- consider the rank $\widehat{\ell} = \max\{\ell \in \{0, 1, \ldots, m\} \ : \ p_{(\ell)} \leq \alpha\ell/m\}$;
- choose $\widehat{t} = \alpha\widehat{\ell}/m$ and rejects the nulls with a p-value smaller than the threshold $\widehat{t}$, or, equivalently, the $\widehat{\ell}$ nulls corresponding to the smallest p-values.

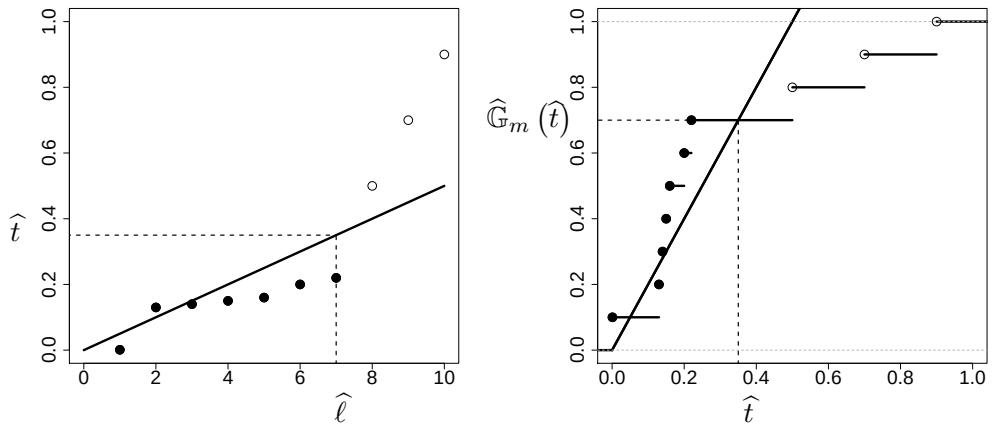


Figure 2.2: Left: illustration for Algorithm 2.1 for $m = 10$ p-values. The stopping rule appears as a "last right crossing point" between the ordered p-values $\ell \mapsto p_{(\ell)}$ and the solid line $\ell \mapsto \alpha\ell/m$. The p-values colored in black are those corresponding to rejected nulls. Right: equivalent formulation according to the empirical distribution function $\widehat{\mathbb{G}}_m$ of the p-values, see (2.8) below.

Theorem 2.2 ([10, 14]). *Assume* (pvalueprop). *For any $P \in \mathcal{P}$ satisfying* (Indep) *or* (wPRDS), *the BH procedure R defined by Algorithm 2.1 satisfies $FDR(R, P) \leq \alpha m_0(\theta)/m$. Moreover, by assuming* (pvaluepropunif), *the latter becomes an equality for any $P \in \mathcal{P}$ satisfying* (Indep).

The distributional assumptions (Indep) (independence) and (wPRDS) (positive dependence) will be further discussed in Section 2.4. To give more intuition behind Theorem 2.2, Figure 2.3 displays 9 realizations of the BH procedure together with the corresponding (unobservable) FDP value in the independence case. Theorem 2.2 ensures in this case that the FDR of the BH procedure is *equal* to $\alpha/2 = 0.125$. Hence, from a more intuitive point of view, this result implies that, if we repeat these experiments infinitely (for the same values of the parameters), the average of the realized values of FDP(BH) would converge to 0.125. We already see at this point that the variations of the FDP around its expectation is a key property for providing a correct interpretation of the BH procedure (here, the strong variability is only due to the fact that m is small).

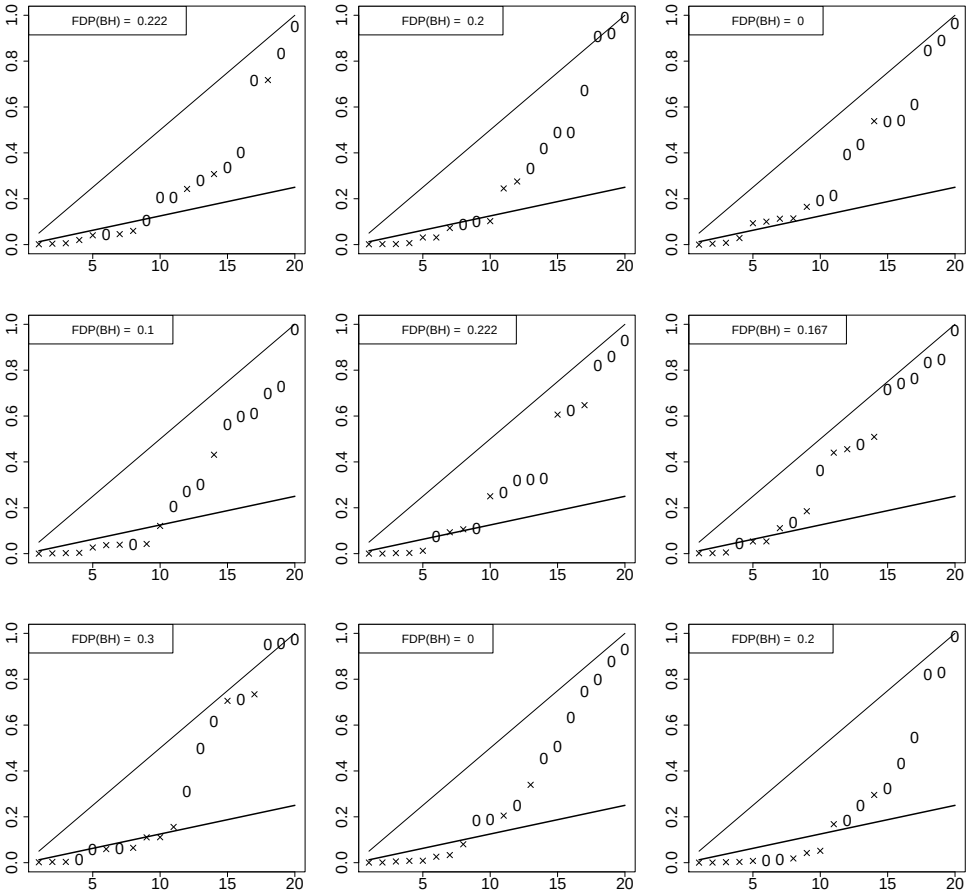


Figure 2.3: False discoveries of BH procedure ($\alpha = 0.25$) for 9 simulations in the Gaussian one-sided model with $m = 20$, $m_0 = 10$, means $\mu_i \in \{0, 2\}$ and $\Gamma = I$ (independence). Same labeling as Figure 2.1. The realized FDP is reported in the topleft box. The solid lines are $\ell \mapsto \ell/m$ and $\ell \mapsto \alpha\ell/m$.

We finally illustrate a *scale invariance property* of the BH procedure as m grows to infinity, which shows that it somehow avoids the "curse of dimensionality", provided that the proportion of signal stays the same. For this, we use the same setting as above, except that we consider increasing values of

m while $\pi_0 = m_0/m$ stays equal to 0.5, see Figure 2.4. We observe that, when m grows, the proportion of rejected hypotheses stays away[2] from zero. Specifically, this holds because the BH procedure is related to the asymptotic behavior of the empirical distribution function of the p-values through the equation (see Figure 2.2, right)

$$\widehat{t} = \max\{t \in [0,1] \ : \ \widehat{\mathbb{G}}_m(t) \geq t/\alpha\}. \tag{2.8}$$

As we will see in Sections 5.2, 6.2 and 6.3, the formulation (2.8) is convenient for asymptotical studies (as the number m of tests tends to infinity).

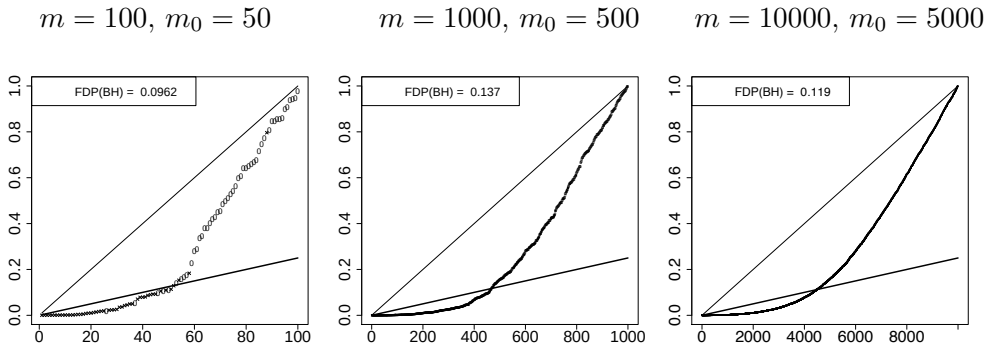


Figure 2.4: Scale invariance property of BH procedure, see text. Gaussian one-sided model with means $\mu_i \in \{0, 2\}$ and $\Gamma = I$. Same labeling as Figure 2.1.

2.3.3 Power issue

Following the Neyman-Pearson paradigm, provided that the chosen type I error rate is below α, we would like to maximize the quantity of true discoveries, that is the "power" of the multiple testing procedure. Formally, a simple and standard choice for the power is

$$\mathrm{Pow}(R) = C^{-1}\mathbb{E}\left[|R \cap \mathcal{H}_1(\theta)|\right], \tag{2.9}$$

the (rescaled[3]) averaged number of true discoveries made by R. Hence, similarly to the "Uniformly Most Powerful" theory for single testing, an *optimal* multiple procedure maximizes $\mathrm{Pow}(R)$ under the constraint that it controls the type I error rate at level α.

In multiple testing theory, assessing optimality is a difficult task and only few results exist, see [130, 133, 16, 114, 95]. While no general answer to that challenging problem will be provided in this manuscript, a partial solution is proposed in Section 4.2 via an optimal p-value weighting.

2.4 Model assumptions

The probabilistic properties of multiple testing procedures usually rely on specific assumptions on the model $\mathcal{P}$, or, equivalently, on the distribution P of X. We loosely separate in this section the assumptions concerning the dependency structure from the assumptions concerning the signal.

[2]Note that this would not be the case with an FWER controlling procedure, as the distribution of the infimum would tend to zero.

[3]The normalizing constant $C > 0$ can be for instance $m_1(\theta)$ or m.

2.4.1 Dependence assumptions

Dependence assumptions rarely correspond to realistic situations. However, they are often unavoidable to get accurate and rigorous controlling results, as in Theorem 2.2.

First, a strong but useful assumption on P is the independence between the individual tests: more specifically,

$$\begin{array}{c}(p_i(X), i \in \mathcal{H}_0(\theta)) \text{ is a family of mutually independent variables} \\ \text{and } (p_i(X), i \in \mathcal{H}_0(\theta)) \text{ is independent of } (p_i(X), i \in \mathcal{H}_1(\theta)).\end{array} \tag{Indep}$$

This assumption encompasses the classical case where the whole p-value family is assumed to be mutually independent:

$$(p_i(X), 1 \leq i \leq m) \text{ is a family of mutually independent variables.} \tag{Full-Indep}$$

However, (Indep) is not restricted to (Full-Indep) since the joint distribution of $(p_i(X), i \in \mathcal{H}_1(\theta))$ is let arbitrary in (Indep).

Second, an assumption weaker than (Indep) is the *positive regression dependency on each one from a subset* (PRDS) property, a notion that can be traced back to Erich L. Lehmann (1966) [88] in the bivariate case. First, let us define a subset $D \subset [0,1]^m$ as *nondecreasing* if for all $q, q' \in [0,1]^m$ such that $\forall i \in \{1, ..., m\}$, $q_i \leq q'_i$, we have $q' \in D$ when $q \in D$. Then, the *weak PRDS property* is as follows.

$$\begin{array}{c}\text{For any } i_0 \in \mathcal{H}_0(\theta) \text{ and any measurable nondecreasing set } D \subset [0,1]^m, \\ \text{the function } u \mapsto \mathbb{P}\left((p_i(X), 1 \leq i \leq m) \in D \,|\, p_{i_0}(X) \leq u\right) \\ \text{is nondecreasing on the set } \{u \in [0,1] \ : \ \mathbb{P}(p_{i_0}(X) \leq u) > 0\}.\end{array} \tag{wPRDS}$$

Assumption (wPRDS) is slightly different than the *PRDS property*, as defined in [14]:

$$\begin{array}{c}\text{For any } i_0 \in \mathcal{H}_0(\theta) \text{ and any measurable nondecreasing set } D \subset [0,1]^m, \\ \text{the function } u \mapsto \mathbb{P}\left((p_i(X), 1 \leq i \leq m) \in D \,|\, p_{i_0}(X) = u\right) \text{ is nondecreasing.}\end{array} \tag{PRDS}$$

To be completely rigorous, since the function $u \mapsto \mathbb{P}\left((p_i(X), 1 \leq i \leq m) \in D \,|\, p_{i_0}(X) = u\right)$ is defined up to some $p_{i_0}(X)$-negligible set, (PRDS) assumes that this function coincides $p_{i_0}(X)$-a.s. with a nondecreasing function. We can check that (wPRDS) is weaker than (PRDS) (see, e.g., Proposition 3.6 in [P6]). Hence, Theorem 2.2 also holds under (PRDS)[4].

As an illustration, in the one-sided Gaussian testing framework, the PRDS assumptions (regular and thus also weak) are satisfied whenever $\Gamma_{i,j} \geq 0$ for all i, j, see Section 3.1 in [14]. Hence, by applying Theorem 2.2, the FDR control holds for the BH procedure in that case. Note that the two-sided case is more delicate because the p-values lose the (wPRDS) property, even when $\Gamma_{i,j} \geq 0$ for all i, j (see [107]). However, obviously, the FDR control is maintained in the two-sided case when $\Gamma = I$ because (Indep) holds.

A stronger notion of positive dependence is the *multivariate total positivity of order* 2 (MTP2), as introduced by Tapan K. Sarkar (1969) [121]. It requires that the joint distribution of the p-values has a density f that satisfies

$$\text{for all } q, q' \in \mathbb{R}^m,\ f(q)f(q') \leq f(q \wedge q')f(q \vee q'), \tag{MTP2}$$

where " $\wedge$ " (resp. " $\vee$ ") denotes the infimum (resp. supremum) operator, that should be considered component-wise. Provably, (MTP2) implies (PRDS) and thus (wPRDS) (and also positive association of the p-values), see, e.g., Theorems 4.1 and 4.2 in [84]. In the one-sided Gaussian setting, (MTP2) is

[4] The original results in [14] are stated with the PRDS property.

satisfied if and only if the off-diagonal elements of $-\Gamma^{-1}$ are all nonnegative[5]. For instance, a special case of interest is the equi-correlated case:

$$\Gamma_{i,j} = \rho, \text{ for all } i \neq j, \text{ where } \rho \in [-(m-1)^{-1}, 1]. \tag{Gauss-ρ-equi}$$

When $\rho \geq 0$, the latter model can be realized via the well known decomposition

$$X_i - \mathbb{E}X_i = \rho^{1/2}\, W + (1-\rho)^{1/2}\, \xi_i, \tag{2.10}$$

where W and the ξ_i's are all i.i.d. $\mathcal{N}(0,1)$. In this case, W can be interpreted as an overall "disturbing" factor that affects equally all the measurements. Hence, a useful function is the distribution function of $p_i(X)$ conditionally on $W = w$ (under the null), which is given by

$$f(t,w) = \overline{\Phi}\left((\overline{\Phi}^{-1}(t) - \rho^{1/2}\, w)/(1-\rho)^{1/2} \right). \tag{2.11}$$

Finally, in the two-sided Gaussian setting but only when $m_0 = m$, (MTP2) holds if and only if there exists a diagonal matrix D with diagonal elements in $\{-1, 1\}$ such that the off-diagonal elements of $-D\Gamma^{-1}D$ are all nonnegative, see [85]. This includes the case (Gauss-ρ-equi) with $\rho \geq 0$, for which the distribution function of $p_i(X)$ conditionally on $W = w$ (under the null) is given by

$$f(t/2, w) + f(t/2, -w). \tag{2.12}$$

Note that establishing (MTP2) in the two-sided case when $m_0 < m$ requires in general some restrictions on $\mu \in \mathbb{R}^m$, see [45].

2.4.2 Signal assumptions

Loosely, we can consider two kinds of assumptions related to the "signal": first, assumptions on the *amount* of signal, which concerns m_0. Second, assumptions on the *strength* of the signal, that can be measured by the "distance" between the distributions of the alternative p-values and the uniform distribution (e.g., via the value of the alternative means when testing $\mathbb{E}X$).

For m_0, a standard assumption is to consider that, as m grows to infinity, m_0/m tends to some quantity[6] π_0 which lies in $(0,1)$, see Figure 2.4, and also Sections 5.2 and 6.2. This means that the proportion of signal is asymptotically of the order of m, which is optimistic but convenient. At the opposite side, the *sparsity* assumption supposes that m_0/m tends to 1 as m grows to infinity, which typically arises when the dimension of the data grows faster than the number of entities of interest. For instance, a common assumption is that $1 - m_0/m \sim m^{-\beta}$ for some $\beta \in (0,1]$. This case will be examined in Section 6.3.

Now, for the signal strength, a useful assumption is that it is maximal:

$$\text{for all } i \in \mathcal{H}_1(\theta), \quad p_i(X) = 0 \quad \text{a.s.} \tag{Dirac}$$

The alternative distribution given by (Dirac) can be seen as the most optimistic: in that case, the alternatives are perfectly separated from the nulls (a.s.). Hence, there is no multiple testing problem in that case. However, this special configuration remains interesting, because it often has the property to be the distribution under which the type I error rate is the largest, in which case it is called "least favorable configurations" (LFC). This generally requires that the multiple testing procedure at hand has special monotonic properties, see Section 3.3. Hence, perhaps surprisingly, (Dirac) turns out to be a useful assumptions for proving type I error rate controls.

[5] Note that Γ is invertible by the existence of the density.

[6] Here, π_0 is not denoting m_0/m but rather its limit when m grows to infinity.

2.4.3 Random effects relaxation

As introduced in [43], a Bayesian-like layer can be proposed for the general model of Section 2.1. It is usually referred to as the *random effects model*. It will be useful at several points of this manuscript, see Sections 3.3, 4.2 and 6.3.

Remember that from (2.1), part of the true underlying distribution P is contained in the vector $\theta = \theta(P) \in \{0,1\}^m$. In the general framework, θ is fixed and arbitrary valued in all the possible true/false configurations, hence the type I error rate control should be established in any of these configurations. A less constrained framework is to consider a "fairly averaged" configuration for θ, by assuming that the θ_i's have been generated previously and independently of the data as i.i.d. Bernoulli variables:

$$\theta \text{ is a random vector composed of } m \text{ i.i.d. Bernoulli variables } \mathcal{B}(1-\pi_0). \tag{Mixture}$$

Formally, it means that the parameters of this model become $\pi_0 \in [0,1]$ and $(P_\theta)_{\theta\in\{0,1\}^m}$ with $P_\theta \in \mathcal{P}_\theta$ for all $\theta \in \{0,1\}^m$ [7]. The model is thus built by first generating θ according to the distribution of (Mixture) and then $X \sim P_\theta$. Note that under Assumption (Mixture), $m_0(\theta)$ is random and follows a binomial distribution of parameters m and π_0.

The random effects assumption (Mixture) is often used with the independent assumption (Full-Indep) while assuming that the alternative p-values have the same distribution function F_1. In that case, unconditionally on θ, the $p_i(X)$, $1 \le i \le m$, are i.i.d. with common distribution function $G(t) = \pi_0\, t + \pi_1 F_1(t)$, $t \in [0,1]$, where $\pi_1 = 1-\pi_0$ stands for the probability of generating an alternative. The parameters of the model are thus simply $\pi_0 \in [0,1]$ and $F_1 \in \mathcal{F}$, where $\mathcal{F}$ is some space of distribution functions. For instance, in the one-sided Gaussian case, when the alternative means are all equal to some $\Delta > 0$, $F_1(t)$ is of the form $\overline{\Phi}(\overline{\Phi}^{-1}(t) - \Delta)$ and the parameters of the model are given by $\pi_0 \in [0,1]$ and $\Delta > 0$.

2.5 Classes of procedures

2.5.1 Step-wise procedures

The BH procedure is a thresholding-based procedure $\widehat{t}$ built by taking the "last crossing point" between the sequence of ordered p-values $p_{(1)} \le p_{(2)} \le \cdots \le p_{(m)}$ and the sequence $\tau_\ell = \alpha\ell/m$, $\ell = 1,\dots,m$, see Algorithm 2.1. However, many other choices are possible for the values τ_ℓ, $\ell = 1,\dots,m$. Namely, considering any sequence of non-decreasing constants τ_ℓ, $\ell = 1,\dots,m$ (with the convention $\tau_0 = 0$), we can consider the procedure $\widehat{t} = \tau_{\widehat{\ell}}$ with

$$\widehat{\ell} = \max\{\ell \in \{0,\dots,m\} \ : \ p_{(\ell)} \le \tau_\ell\}, \tag{2.13}$$

that is called the *step-up procedure with critical values* τ_ℓ, $\ell = 1,\dots,m$, which is denoted by $\mathrm{SU}(\tau)$. Hence, according to Algorithm 2.1, the BH procedure is step-up with critical values $\tau_\ell = \alpha\ell/m$, $1 \le \ell \le m$.

Next, if the ordered p-values and the sequence of the τ_ℓ's have several crossing points, the rightmost crossing point is not the only choice that can be made. For instance, the most-left crossing point is called "step-down": in this case,

$$\widehat{\ell} = \max\{\ell \in \{0,\dots,m\} \ : \ \forall \ell' \in \{0,\dots,\ell\},\ p_{(\ell')} \le \tau_{\ell'}\}, \tag{2.14}$$

[7] In particular, the $\mathcal{P}_\theta$'s are all assumed non empty and thus form a partition of the model $\mathcal{P}$.

and the corresponding procedure is called *step-down procedure with critical values* τ_ℓ, $\ell = 1, \dots, m$ and is denoted by $\mathrm{SD}(\tau)$. Figure 3.1 illustrates the difference between a step-down and a step-up procedure. While $\mathrm{SD}(\tau) = \mathrm{SU}(\tau)$ if there is only one crossing point, $\mathrm{SD}(\tau) \subset \mathrm{SU}(\tau)$ holds in general. A counterpart is that the additional constraints of the step-down algorithm can be useful to get additional controlling results, see, e.g., [113, 57] and Section 5.1.

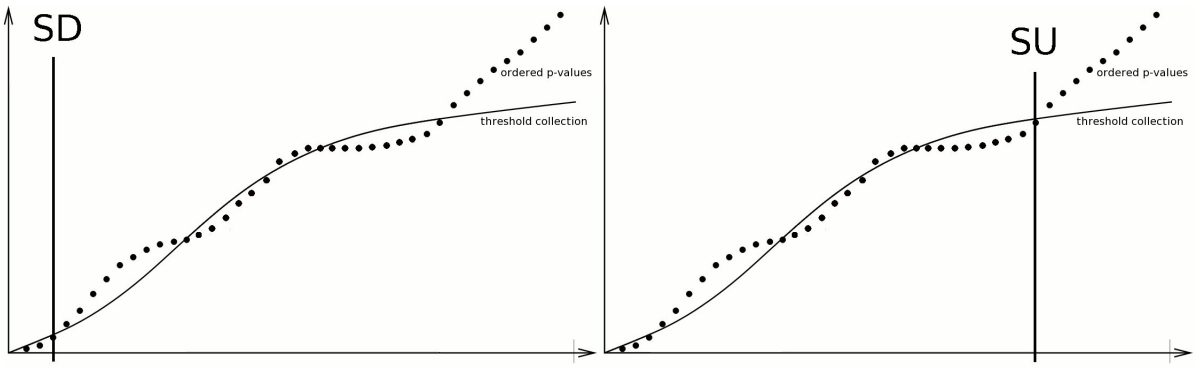


Figure 2.5: Pictorial view of a step-down (first crossing point) and a step-up (last crossing point) algorithms.

In a more general manner, we can think to consider intermediate crossing points. To this end, step-up-down procedures have been introduced by Tamhane et al. [135] (see also [118]): for an extra parameter $\lambda \in \{1, \dots, m\}$, if $p_{(\lambda)} > \tau_\lambda$, a step-up procedure is run in the left direction; if $p_{(\lambda)} \leq \tau_\lambda$, a step-down procedure is run in the right direction, see Figure 2.6. More formally, the step-up-down (SUD) procedure of order $\lambda \in \{1, ..., m\}$ and with critical values τ_ℓ, $\ell = 1, \dots, m$, is denoted by $\mathrm{SUD}_\lambda(\tau)$, and is defined as $\widehat{t} = \tau_{\widehat{\ell}}$, with

$$\widehat{\ell} = \begin{cases} \max\{\ell \in \{\lambda, \dots, m\} \ : \ \forall \ell' \in \{\lambda, \dots, \ell\},\ p_{(\ell')} \leq \tau_{\ell'}\} & \text{if } p_{(\lambda)} \leq \tau_\lambda; \\ \max\{\ell \in \{0, \dots, \lambda\} \ : \ p_{(\ell)} \leq \tau_\ell\} & \text{if } p_{(\lambda)} > \tau_\lambda. \end{cases} \tag{2.15}$$

We merely check that choosing $\lambda = m$ (resp. $\lambda = 1$) reduces (2.15) to the step-up case (2.13) (resp. to the step-down case (2.14)). Also, the set of rejected nulls is nondecreasing with respect to λ. From an historical point of view, the step-up-down procedures were initially introduced for improving the detection that at least λ out of m null hypotheses are false while controlling the FWER, see [135]. A more modern use of such procedures is made in [49] for finding an asymptotical optimal rejection curve (AORC) while controlling the FDR. In this manuscript, SUD procedures will be considered in Section 3.3.

2.5.2 Adaptive procedures

While establishing a type I error rate control, it can be of interest to try to improve the power of the procedure by incorporating in it a part ϑ of the underlying distribution P. For instance, in the Gaussian multivariate framework, we can study adaptation w.r.t. $\vartheta = \pi_0$ (proportion of true null), $\vartheta = (\mu_i)_{i \in \mathcal{H}_1}$ (values of the alternative means), or $\vartheta = \Gamma$ (dependence structure). Since ϑ is often unknown, a procedure using the true value of ϑ is generally referred to as *oracle*. Loosely, a procedure that aims at approaching the oracle is said *adaptive* with respect to ϑ, or ϑ-*adaptive* in short. Classical examples include procedures that explicitly use an estimator $\widehat{\vartheta}$ of ϑ, often referred to as "plug-in". In addition, with some abuses, the oracle version will be sometimes called "adaptive" in the manuscript. However, we should keep in mind that it is only usable when ϑ is known.

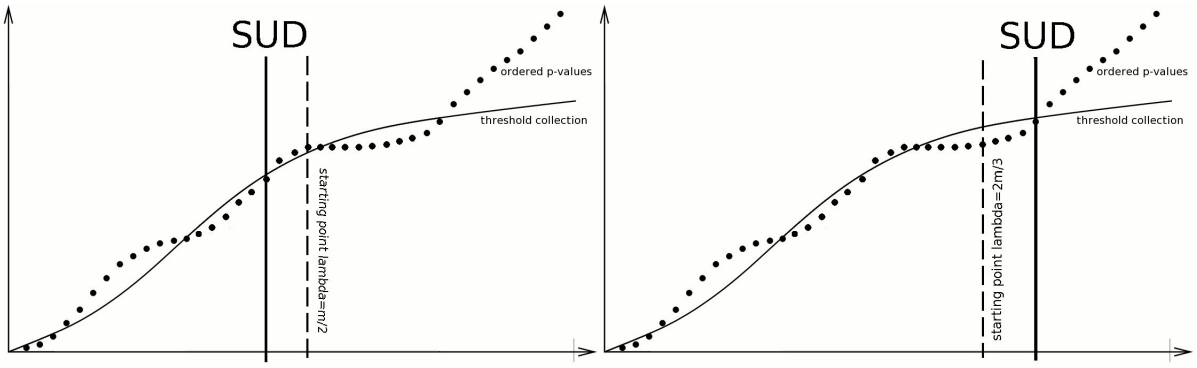
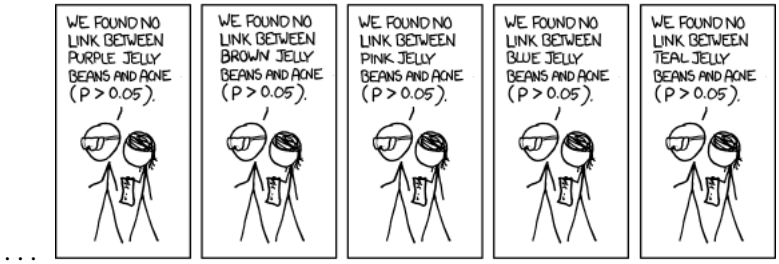


Figure 2.6: Pictorial view of step-up-down procedures for two choices of λ. Left: $\lambda = m/2$. Right: $\lambda = 2m/3$.

When studying adaptive procedures, the major issue is to show that the type I error rate control is maintained, while the power is (significantly) improved. Building and studying adaptive procedures is an important part of the work presented in this manuscript, see Chapters 4 and 5.



Chapter 3

Consolidating and extending the theory

In this chapter, we first aim at controlling the FDR, that is, given α, we want to build critical values τ_ℓ, $\ell = 1, \dots, m$, such that $\mathrm{FDR}(\mathrm{SUD}_\lambda(\tau), P) \leq \alpha$ under some assumptions on P. Section 3.1 presents a general methodology for controlling the FDR that allows several generalizations of former results in an unified manner. In addition, this method can be naturally extended to the case where a "continuous amount" of null hypotheses is tested, which is investigated in Section 3.2. Finally, in Section 3.3, we investigate a task that can be seen as the converse of FDR control: given the critical values τ_ℓ, $\ell = 1, \dots, m$, we aim at calculating $\mathrm{FDR}(\mathrm{SUD}_\lambda(\tau), P)$, or, more generally, the distribution of $\mathrm{FDP}(\mathrm{SUD}_\lambda(\tau), P)$. As an application, we contribute to the theory of "least favorable configurations" for the FDR criterion.

3.1 Two simple conditions for controlling the FDR [P6]

We present here contributions of the work [P6] that aims at identifying essential arguments implying FDR control.

3.1.1 Main idea

Let $\beta : \mathbb{R}^+ \to \mathbb{R}^+$ be a nondecreasing function with $\beta(0) = 0$, to be chosen further on. An elementary fact is that the FDR, defined by (2.7), can be upper-bounded as follows (by using the convention $0/0 = 0$):

$$\begin{aligned} \mathrm{FDR}(R, P) = \mathbb{E}\left[\frac{|R \cap \mathcal{H}_0|}{|R|}\right] &= \sum_{i \in \mathcal{H}_0(\theta)} \mathbb{E}\left[\frac{\mathbf{1}\{i \in R\}}{|R|}\right] \\ &\leq \sum_{i \in \mathcal{H}_0(\theta)} \mathbb{E}\left[\frac{\mathbf{1}\{p_i(X) \leq \alpha\beta(|R|)/m\}}{|R|}\right] \leq \alpha m_0(\theta)/m. \end{aligned} \tag{3.1}$$

Above, two conditions are assumed: the first inequality is implied by the following algorithmic condition, called *self-consistency*,

$$R \subset \{1 \leq i \leq m \ : \ p_i(X) \leq \alpha\beta(|R|)/m\}. \tag{SC(β)}$$

The second inequality is implied by a probabilistic condition, called *dependency control* condition: for any $i \in \mathcal{H}_0(\theta)$, the random variable couple $(U, V) = (p_i(X), |R|)$ satisfies

$$\forall c > 0, \quad \mathbb{E}\left[\frac{\mathbf{1}\{U \leq c\beta(V)\}}{V}\right] \leq c\,. \tag{DC(β)}$$

Lemma 3.1. *Let $\beta : \mathbb{R}^+ \to \mathbb{R}^+$ be any nondecreasing function with $\beta(0) = 0$ such that (DC(β)) holds (for the underlying distribution P) and R be any multiple testing procedure satisfying (SC(β)). Then $FDR(R, P) \leq \alpha m_0(\theta)/m$.*

Now, to prove FDR control, it is sufficient to check these two conditions (SC(β)) and (DC(β)). To this end, the choice of β is pivotal.

3.1.2 Study of the two conditions

Self-consistency An illustration of self-consistency is given in Figure 3.1 by following the p-value ordering representation: the blue area indicates the rejections numbers ℓ such that the procedure rejecting the ℓ smallest p-values satisfies (SC(β)). In particular, by considering the critical values $\tau_\ell = \alpha\beta(\ell)/m$, $1 \leq \ell \leq m$, it is easy to check that the corresponding step-down, step-up and even step-up-down procedures all satisfy (SC(β)). Also, we can easily check that the step-up procedure $\mathrm{SU}(\tau)$ is the less conservative of the multiple testing procedure R satisfying (SC(β)), that is, if R satisfies (SC(β)), then $R \subset \mathrm{SU}(\tau)$.

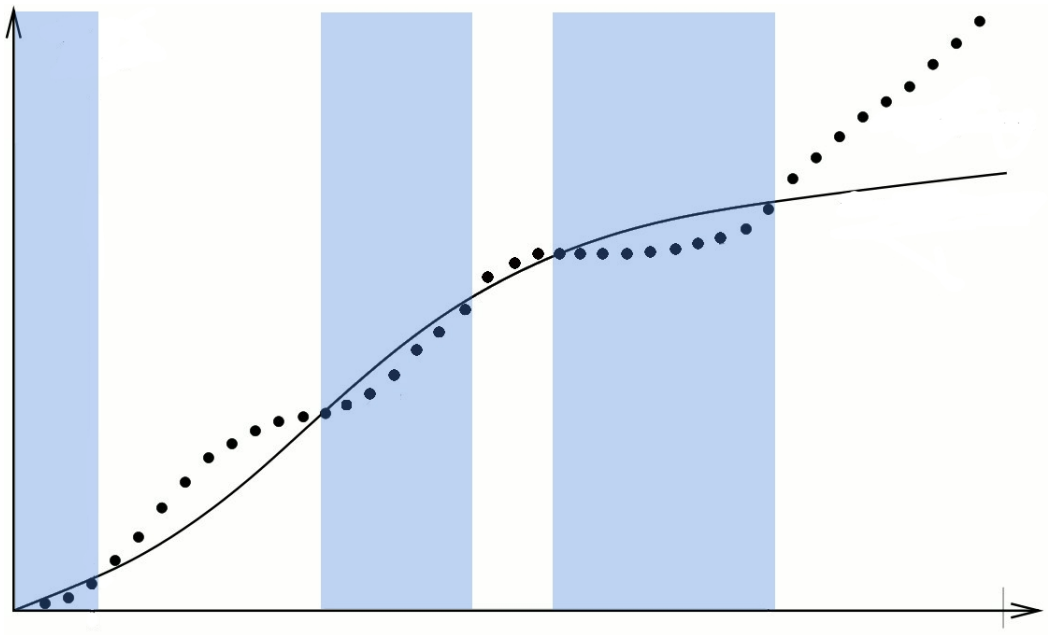


Figure 3.1: Pictorial view of self-consistent stopping rules (light blue), containing step-down, step-up and step-up-down rules. Ordered p-values $p_{(\ell)}$ (dots) and critical values $\tau_\ell = \alpha\beta(\ell)/m$ (solid line) in function of ℓ.

Dependency control condition In [P6], the following is proved:

(i) assuming (pvalueprop) and (wPRDS) and if $|R|$ is a component-wise nonincreasing function of the p-value family, then (DC(β)) holds with $\beta(x) = x$;

(ii) assuming (pvalueprop) (and no assumption on the dependence), then (DC(β)) holds with β of the following special form:

$$\beta(x) = \int_0^x u d\nu(u), \ x \geq 0, \tag{3.2}$$

where ν is any distribution on $(0, \infty)$.

Note that $\beta(x) \leq x$ for all $x \geq 0$ whenever β is of the form (3.2).

3.1.3 Applications

Application 1: extending classical FDR controls Since the BH procedure corresponds to the step-up procedure with critical values $\tau_\ell = \alpha\beta(\ell)/m$ and $\beta(x) = x$, if we assume (wPRDS), then both (SC(β)) and (DC(β)) hold and Lemma 3.1 implies that the BH procedure controls the FDR. In particular, this entails the classical result stated in Theorem 2.2 (with $\leq$). Now, under arbitrary dependence, Lemma 3.1 provides an FDR control for any step-up procedure with critical values $\tau_\ell = \alpha\beta(\ell)/m$, with β of the form (3.2). Several choices of β are possible by tuning the parameter ν, see Figure 3.2. In particular, this allows to recover several other results of the literature:

- $\beta(\ell) = \ell/\left(\sum_{1\leq i\leq m} 1/i\right)$ with $\nu(\{k\})$ proportional to $1/k$, $k \in \{1,\dots,m\}$. The resulting step-up procedure is classically called the BY procedure, see [14];
- $\beta(\ell) = \ell(\ell+1)/2m$ with ν uniformly distributed on $\{1,\dots,m\}$, see [120];
- $\beta(\ell) = \lfloor \gamma m \rfloor \mathbf{1}\{\ell \geq \lfloor \gamma m \rfloor\}$ with $\nu = \delta_{\lfloor \gamma m \rfloor}$ for some $\gamma \in (0,1)$, see [94].

On the intuitive point of view, ν can be interpreted as a *prior* distribution on the number of rejections, as argue in [18], which presented a previous version of this work in another context. To this respect, the BY procedure "favors" the small number of rejections.

Finally, for $\tau_\ell = \alpha\beta(\ell)/m$ with β of the form (3.2), we have $\tau_\ell \leq \alpha\ell/m$ and thus SU(τ) is more conservative than the BH procedure. Hence, there is a price to pay to get an FDR control valid under arbitrary dependence. This seems fair since the critical values τ_ℓ are not "learning" the dependence so they are adjusted to the "worst dependent case". Note that procedures learning the dependency structure will be investigated in Chapter 5.

Application 2: shape constraints A property similar to (SC(β)) appeared independently in [49], named as (T2) therein. It uses "$=$" instead of "$\subset$". Actually, using "$\subset$" can be useful to accommodate external constraints on the shape of the rejection set R that prevent the equality into (SC(β)). In our paper [P6], we mentioned the case of a convex constraint in a two dimensional hypothesis testing framework for instance. Actually, another illustration came very recently in [68], in which the null hypotheses are ordered according to some *a priori* preference. Precisely, while the null hypotheses are not necessarily assumed to be nested, the decision is constrained to be of the form $R = \{1,2,\dots,\widehat{k}\}$, for some $\widehat{k}$ to be chosen (by convention, $R = \emptyset$ if $\widehat{k} = 0$). Note the difference with a step-up procedure, for which the ordered is necessarily prescribed by the p-values. Here, the order is previously decided, and allows situations for which $p_1(X) = 0.9$, $p_2(X) = 0.1$ and $p_{10}(X) = 10^{-9}$ for instance. We merely check that taking $R = \{1,2,\dots,\widehat{k}\}$ of maximum size such that (SC(β)) holds gives

$$\widehat{k} = \max\left\{1 \leq k \leq m \ : \ \max_{1\leq i\leq k}\{p_i(X)\} \leq \alpha\beta(k)/m\right\}.$$

Hence, Lemma 3.1 implies an FDR control for this type of stopping rule (under appropriate dependence assumptions).

Application 3: weighting In multiple testing literature, there has been some interest in adding weights to null hypotheses to favor some tested items more than others. Two different ways to use weighting have been proposed:

- *weighted FDR* [11, 9]: let Λ be some measure on $\{1,\dots,m\}$, called a *volume measure*. The relative importance of false discoveries can be adjusted according to Λ by replacing $\mathrm{FDP}(R,P)$

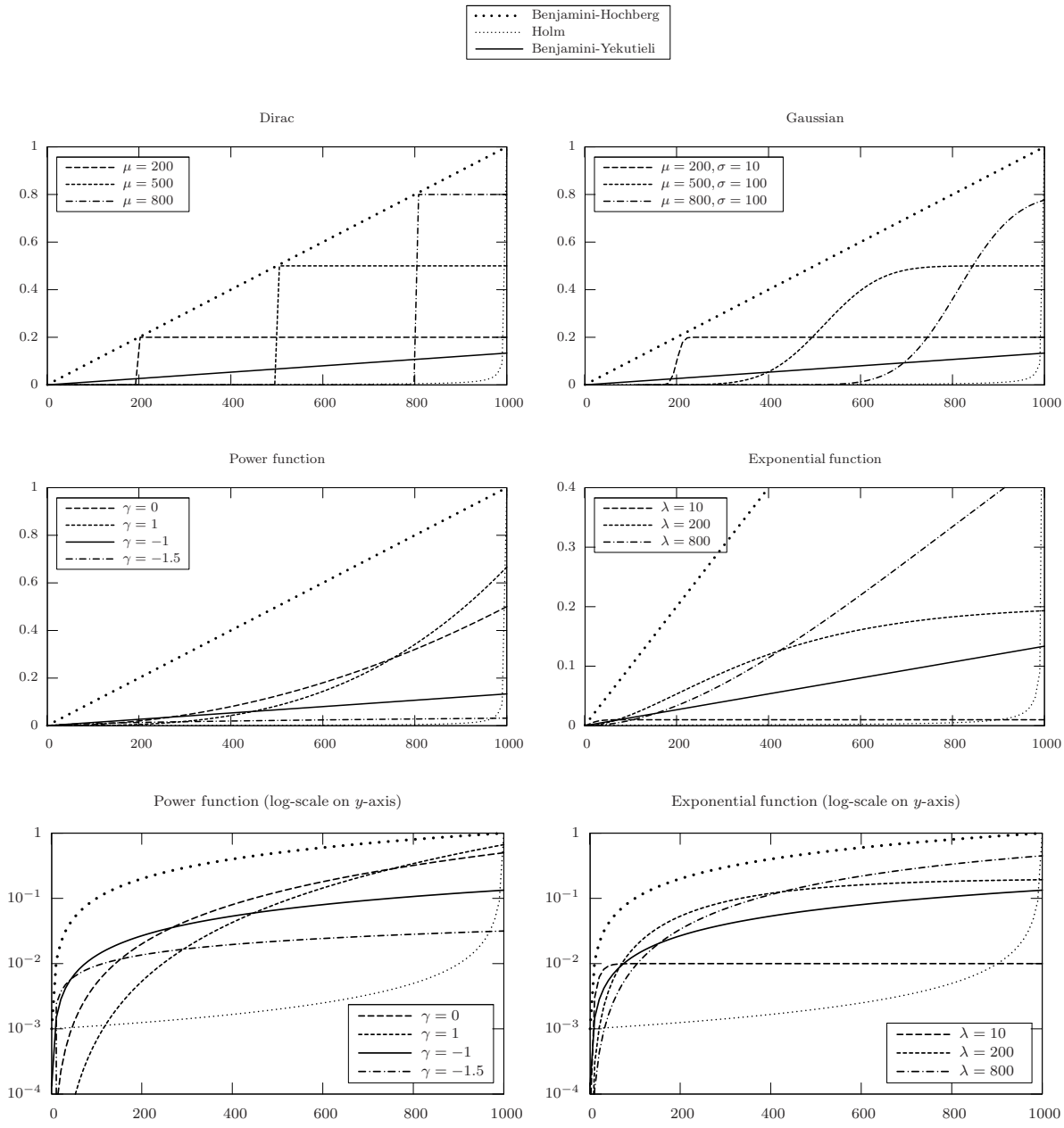


Figure 3.2: Plot of $m^{-1}\beta$ associated to different ν according to expression (3.2): *Dirac*: $\nu = \delta_\mu$, with $\mu > 0$. *Gaussian*: ν is the distribution of $\max(X, 1)$, where $X \sim \mathcal{N}(\mu, \sigma^2)$. *Power*: $d\nu(u) = u^\gamma \mathbf{1}\{u \in [1, m]\}du / \int_1^m t^\gamma dt$, $\gamma \in \mathbb{R}$. *Exponential*: $d\nu(u) = (1/\lambda)\exp(-u/\lambda)\mathbf{1}\{u \in [0, m]\}dr$, with $\lambda > 0$. On each graph: Holm's choice is $m^{-1}\beta(x) = 1/(m - x + 1)$ (small dots); BH choice $\beta(x) = x$ (large dots); BY choice $\beta(x) = x / \left(\sum_{1 \le i \le m} 1/i\right)$ (solid).

by

$$\mathrm{FDP}(R,P) = \frac{\Lambda(R \cap \mathcal{H}_0(\theta))}{\Lambda(R)} \mathbf{1}\{\Lambda(R \cap \mathcal{H}_0(\theta)) > 0\}, \tag{3.3}$$

that is, by replacing the standard counting measure $|\cdot|$ by the volume measure Λ.

- *weighted procedure* [11, 60]: any multiple testing procedure can be "weighted" by replacing each p-value $p_i(X)$ by its weighted counterpart $p'_i(X) = p_i(X)/w_i$, for some *weight vector* $(w_i)_{1\le i\le m} \in (\mathbb{R}^+)^m$ (by convention $0/0 = 0$). The condition (SC(β)) simply becomes

$$R \subset \{1 \le i \le m \,:\, p_i(X) \le \alpha w_i \beta(\Lambda(R))/m\}. \tag{SC(w,β)}$$

Note the difference between the two types of weighting: the first one is a modification of the criterion (so changes the aim and the final interpretation) while the second one is a manner to use a wider range of procedures. The presented methodology directly accomodate to these two weighting types[1]. As in (3.1), under (SC(w,β)) and (DC(β)), the (Λ-weighted) FDR is upper bounded by

$$\alpha \sum_{i \in \mathcal{H}_0(\theta)} \Lambda(\{i\}) w_i / m. \tag{3.4}$$

Hence, to control the Λ-weighted FDR at level $\alpha m_0/m$, the choice $w_i = 1/\Lambda(\{i\})$ seems to be appropriate. In particular, for Λ being the counting measure, taking any weight vector $(w_i)_{1\le i\le m}$ with $\sum_{i=1}^m w_i = m$ can be used to control the (original) FDR. This defines a the so-called family of "weighted BH procedures", which will be useful in Section 4.2. Many other choices are possible by tuning the bound (3.4), that potentially combines the two types of weighting.

3.1.4 A conclusion

This "two conditions-based" methodology has the benefit to prove a great variety of FDR controls, so avoid to reinvent a proof devoted to each particular configuration.

To provide a summarized illustration of the potential benefit that brings this methodology, let us consider the following (exaggeratingly-)complex procedure: the step-up-down procedure of order $\lambda = \lfloor m/3 \rfloor$ using the critical values $\tau_\ell = \alpha\ell(\ell+1)(2\ell+1)/(3m^2(m+1))$ and the p-value $p'_i = p_i/2$ for half of the tested nulls and $p'_i = 2p_i$ for the rest. Then, it controls the FDR at level $\alpha m_0/m$ under general dependence, that is, under (pvalueprop). While starting to write-down a devoted proof seems difficult and tedious, (3.4) indicates a simple way to prove this result: first, (DC(β)) holds for $\beta(\ell) = \ell(\ell+1)(2\ell+1)/(3m(m+1))$ because it is of the form (3.2) (by choosing $\nu(\{k\})$ proportional to k). Second, (SC(w,β)) holds with suitably chosen weights $w_i \in \{1/2, 2\}$ (satisfying $\sum_{i=1}^m w_i = m$) and with $\Lambda(\{i\}) = 1$ for all i. This shows the result in an admittedly short manner.

Finally, let us emphasize that the task of adding complexity in FDR controlling procedures is not only investigated for generality: it is often the first step when one wants to accurately *adapt* to a specific feature of the underlying distribution P of the data, see Chapter 4.

3.2 Extension to a continuous space [P4]

[1] Note that it is the original setting of the paper [P6].

3.2.1 Motivation

The standard framework of multiple testing is to consider a *finite* number of tests. However, when the data can be modeled as depending on an underlying continuously-indexed parameter, it can be desirable to make a decision for each $t \in [0,1]$ (say). We provide below two simple examples:

(i) A first example is the case where we observe

$$X(t) = \mu(t) + \varepsilon(t), \quad t \in [0,1], \tag{3.5}$$

where ε is a Gaussian process that is assumed to have continuous paths with $\mathbb{E}(\varepsilon(t)) = 0$ and $\mathrm{Var}(\varepsilon(t)) = 1$ for all $t \in [0,1]$, while $\mu \in [0,1] \mapsto \mu(t) = \mathbb{E}[X(t)]$ is some (measurable) mean function. This model is the analogue of the Gaussian framework of Section 2.1 in a continuous setting. The process ε corresponds to the "noise". For instance, it can be chosen as a (normalized) Ornstein-Uhlenbeck process $\varepsilon(t) = e^{-ct}(W_{e^{2ct}-1} + \varepsilon(0))$, where W is a Wiener process and $\varepsilon(0) \sim \mathcal{N}(0,1)$ is independent of W. For the latter process, the covariance function in (s,t) is $e^{-c|t-s|}$ hence $c \in (0,\infty)$ corresponds intuitively to the "strength" of the local dependence ($c \to \infty$ would be independence while $c \to 0$ would give a constant process). In this context, it is of interest to test the null hypothesis $H_{0,t}$: "$\mu(t) \leq \mu_0(t)$" against $H_{1,t}$: "$\mu(t) > \mu_0(t)$" for each location $t \in [0,1]$, where μ_0 contains some benchmark values. This gives rise to the p-value process

$$p_t(X) = \overline{\Phi}(X(t)), \quad t \in [0,1]. \tag{3.6}$$

(ii) As a second instance, let us consider a Poisson process $X = (N_t)_{t\in[0,1]}$ with intensity $\lambda : [0,1] \to \mathbb{R}^+ \in L^1(d\Lambda)$, where Λ denotes the Lebesgue measure. In practice, this model can be used to describe the (non-overlapping) word occurrence process in DNA sequences, see [122, 109]. Doing so, a goal can be *region detection*, that is, finding the locations "t" that correspond to regions "significantly richer" than a reference. Another example is the modeling of the read process in next generation sequences (NGS), where regions with many reads are of interest, see [104]. A setting is therefore to test the null hypothesis $H_{0,t}$: "$\lambda(t) \leq \lambda_0(t)$" against $H_{1,t}$: "$\lambda(t) > \lambda_0(t)$" in each location $t \in [0,1]$, where λ_0 is some benchmark intensity. This gives rise to the p-value process

$$p_t(X) = G_t\left(N_{(t+\eta)\wedge 1} - N_{(t-\eta)\vee 0}\right), \quad t \in [0,1]. \tag{3.7}$$

where for any $k \in \mathbb{N}$, $G_t(k)$ denotes $\mathbb{P}(Z \geq k)$ for Z following a Poisson distribution of parameter $\delta_{t,\eta} = ((t+\eta)\wedge 1 - (t-\eta)\vee 0)(\lambda_0(t) + L_\eta)$. Here, η can be chosen arbitrary, and L_η is an upper bound on $\sup_{s,t:|s-t|\leq\eta} |\lambda(t) - \lambda(s)|$.

Let us mention another option which avoids the bias term L_η into the p-value process: consider the windows $I_\eta(t) = [0,1] \cap [t-\eta, t+\eta]$ for $t \in [0,1]$ for some bandwidth $\eta \in (0,1)$ (to be chosen by the user). Now, (3.7) is a p-value process for testing $H_{0,t}$: "$\lambda(s) \leq \lambda_0(s)$ for all $s \in I_\eta(t)$" against $H_{1,t}$: "$\exists s \in I_\eta(t) : \lambda(s) > \lambda_0(s)$" in each location $t \in [0,1]$, by simply choosing $\delta_{t,\eta} = \int_{I_\eta(t)} \lambda_0(s)ds$.

3.2.2 Continuous versions of FDR, step-up and PRDS

The problems raised above both concern the simultaneous test of a "continuous amount" of null hypotheses. Interestingly, the methodology described in Section 3.1 has already paved the way for the case where the set of null hypotheses is continuous.

To start with, the setting described in Section 2.1 can be extended straightforwardly by supposing that $\mathcal{H}_0(\theta)$, the set containing the elements $t \in [0,1]$ corresponding to a true null $H_{0,t}$, is a measurable

set of $[0,1]$ (with respect to the Borel σ-field). Also, as it was already proposed in [103, 9], the FDR can be generalized to the continuous case directly by using the form (3.3), where Λ is some finite positive measure on $[0,1]$, for instance the Lebesgue measure. Nevertheless, other multiple testing notions are more difficult to generalize and require to solve specific issues, as we discuss below.

The first difficulty arises when considering a continuously-indexed p-value family $(p_t(X), t \in [0,1])$, as in (3.6) or (3.7) in the above examples. Since our reasoning in (3.1) to establish FDR control relied on a Fubini-type argument, the application $(\omega, t) \mapsto p_t(X(\omega))$ should be assumed *(jointly) measurable* (with respect to the product σ-field). A primary consequence is that it precludes the possibility that the p-values of $(p_t(X), t \in [0,1])$ are mutually independent (see, e.g., [108] page 36). This is a major difference with the case where a finite number of null hypotheses are tested, for which independence is generally considered as the standard case. By contrast, this measurability condition requires regularity conditions on the paths of the observed process. It is satisfied in particular in the common case where $p_t(X(\omega))$ is a càdlàg process, as in the two leading examples (i) and (ii) of Section 3.2.1 (see [P4] for more details).

Second, we loose in general the definition of the step-up procedure via p-value ordering. We considered the following alternative definition:

$$\begin{aligned} R(\omega) &= \{t \in [0,1] \ : \ p_t(X(\omega)) \leq \alpha\, \widehat{r}(X(\omega))\} \\ \widehat{r}(X(\omega)) &= \max\{r \geq 0 \ : \ \Lambda(\{t \in [0,1] \ : \ p_t(X(\omega)) \leq \alpha r\}) \geq r\}. \end{aligned} \tag{3.8}$$

Again, we should check that $\omega \mapsto \widehat{r}(X(\omega))$ is measurable, which comes from the joint measurability of $p_t(X(\omega))$. While the formulation is more abstract than in the finite case, we have shown that we can recover a traditional p-value ordered definition when the p-value process is a.s. piecewise constant (as for Example (ii) of Section 3.2.1).

Third, the weak PRDS property should be adapted to the continuous context. To this end, the p-value process $(p_t(X), t \in [0,1])$ is said *finite dimensional weak PRDS on* $\mathcal{H}_0(\theta)$ if for any finite subset $S \subset [0,1]$, the finite p-value family $(p_t(X), t \in S)$ is weak PRDS on $\mathcal{H}_0(\theta) \cap S$ in the sense of (wPRDS). As a consequence, this notion of positive dependence is a property on the finite dimensional distribution of the p-value process, which is convenient. For instance, in the context of Example (i) of Section 3.2.1, the p-value process (3.6) is finite dimensional weak PRDS (on any subset) as soon as the covariance function C of ε is such that $C(s,t) \geq 0$ for all $s, t \in [0,1]$ (as for an Ornstein-Uhlenbeck process for instance). We have also established the finite dimensional weak PRDS property for the p-value process (3.7) of Example (ii), see the appendix of [P4].

3.2.3 Result

We can prove the following result.

Theorem 3.2. *In the above continuous multiple testing setting, the (generalized) step-up procedure R defined by (3.8) controls the (continuous) FDR at level α, provided that the p-value process is finite dimensional weak PRDS on $\mathcal{H}_0(\theta)$.*

The proof relies on an argument similar to (3.1). The crucial step is to establish that the (generalized) condition (DC(β)) holds if the p-value process is finite dimensional weak PRDS, which requires an appropriate finite approximation of (generalized) step-up procedures. A consequence of Theorem 3.2 is that FDR controlling procedures can be derived both in Examples (i) and (ii) of Section 3.2.1. Figure 3.3 provides an illustration for Example (ii) in the case where $\lambda(t)$ is a truncated triangular signal.

In conclusion, continuous testing is possible with FDR control, to the price of more abstract concepts (e.g., for step-up procedure). However, one could argue that a practitioner would maybe prefer to

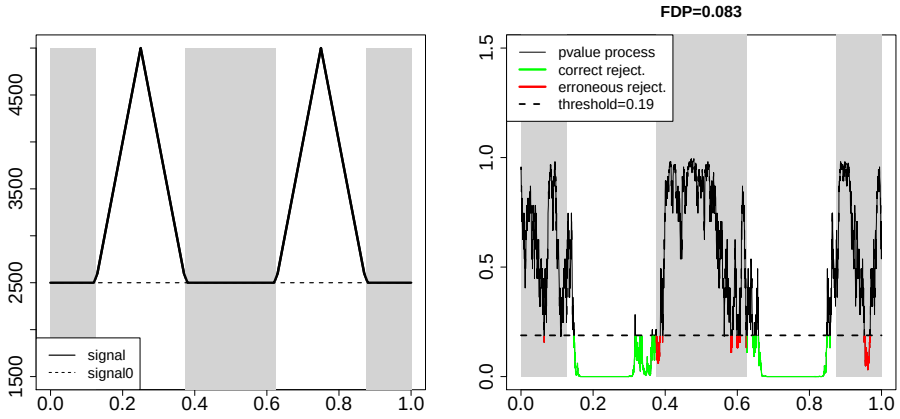


Figure 3.3: Left: $\lambda(t)$ (solid) and λ_0 (dashed) versus $t \in [0,1]$. Right: p-value process $p_t(X)$ defined by (3.7) versus $t \in [0,1]$. $\eta = 0.015$, $\alpha = 0.4$. The grey areas indicate regions where the null hypotheses are true.

discretize the set of null hypotheses and apply a standard BH procedure on the so-obtained finite number of nulls. It would certainly lead to a similar rejection set at the end. Nevertheless, an important property of a statistical modeling is to respect the deep nature of the data: if it is felt to be continuous, it is more appropriate to consider a continuum of null hypotheses. This work has sowed seeds in that direction.

3.3 Exact formulas for FDP with applications to LFCs [P16, P5]

In order to evaluate the type I/II error rate of a given procedure of interest, it is common to use a simulation study. However, the noisy variation due to Monte-Carlo approximation can be undesirable, especially when one wants to infer whether a probability is below a small quantity, say 0.05. Exact formulas provide a suitable alternative, although we should take care of the combinatorial complexity when m grows.

This section gathers some results of the papers [P16, P5], that investigate exact formulas for the distribution of the FDP (2.6) of any step-up-down procedure (2.15), under independence of the p-values, both with and without random effects (Mixture). This work follows a series of studies [132, 118, 48, 27], in particular [51] and [30], tackling the cases $m_0 = m$ and (Dirac), respectively. Several applications of these formulas are provided in [P16, P5], and we will present some of them at the end of the section.

3.3.1 Exact formulas

Our formulas rely on the following multidimensional distribution functions of order statistics: for any $t_1, \dots, t_\ell \in [0,1]$, let us consider the probability

$$\mathbb{P}(U_{(1)} \leq t_1, \dots, U_{(\ell)} \leq t_\ell), \tag{3.9}$$

depending on the joint distribution of $(U_i)_{1 \leq i \leq \ell}$ (the variables $U_{(1)} \dots U_{(\ell)}$ being increasingly ordered). When $(U_i)_{1 \leq i \leq \ell}$ is a sequence of i.i.d. random variables uniformly distributed on $(0,1)$, the value of (3.9) is denoted by $\Psi_\ell(t_1, \dots, t_\ell)$; when $(U_i)_{1 \leq i \leq \ell}$ is a sequence of independent random variables,

with $(U_i)_{1\le i\le \ell_0}$ uniformly distributed on $(0,1)$, and $(U_i)_{\ell_0+1\le i\le \ell}$ having the distribution function F on $[0,1]$ (two-population model), the value of (3.9) is denoted by $\Psi_{\ell,\ell_0,F}(t_1,\dots,t_\ell)$. In the latter, $\ell\ge 0$, $0\le \ell_0\le \ell$, and $\Psi_0(\cdot)=\Psi_{0,0,F}(\cdot)\equiv 1$ by convention. The function $\Psi_\ell(\cdot)$ can be evaluated by using Steck's recursion [127, p. 366–369], while computing $\Psi_{\ell,\ell_0,F}(\cdot)$ can be done by using another more complex recursion, as we have shown in Proposition 1 of [P5]. The next result is the main contribution of this section.

Theorem 3.3. *Let $R=SUD_\lambda(\tau)$ be a step-up-down procedure of order $\lambda\in\{1,...,m\}$ with critical values τ_ℓ, $\ell=1,\dots,m$. Assume that the p-value family $(p_i(X),1\le i\le m)$ satisfies (Full-Indep) and (pvaluepropunif) and that the p-values of $(p_i(X),i\in\mathcal{H}_1)$ have a common distribution function F. Then, the following holds:*

(i) Under assumption (Mixture), for any $\pi_0\in[0,1]$, $0\le\ell\le m$, $0\le j\le\ell$,

$$\mathbb{P}(|R\cap\mathcal{H}_0|=j,|R|=\ell)=\begin{cases}\mathcal{P}_{m,\pi_0,F}(\tau\wedge t_\lambda,\ell,j) & \text{for } \ell<\lambda,\\ \widetilde{\mathcal{P}}_{m,\pi_0,F}(\tau\vee t_\lambda,\ell,j) & \text{for } \ell\ge\lambda.\end{cases} \tag{3.10}$$

where $\mathcal{P}_{m,\pi_0,F}(t_1\cdots t_m,\ell,j)$ and $\widetilde{\mathcal{P}}_{m,\pi_0,F}(t_1\cdots t_m,\ell,j)$ are given by

$$\binom{m}{j}\binom{m-j}{\ell-j}\pi_0^j\pi_1^{\ell-j}(t_\ell)^j(F(t_\ell))^{\ell-j}\Psi_{m-\ell}(1-G(t_m),\dots,1-G(t_{\ell+1})); \tag{3.11}$$

$$\binom{m}{j}\binom{m-j}{\ell-j}\pi_0^j\pi_1^{\ell-j}(1-G(t_{\ell+1}))^{m-\ell}\Psi_{\ell,j,F}(t_1,\dots,t_\ell), \tag{3.12}$$

respectively, by letting $G(t)=\pi_0 t+(1-\pi_0)F(t)$.

(ii) Without assumption (Mixture), for any $m_0\in\{0,\dots,m\}$, $0\le\ell\le m$, $0\vee(\ell-m+m_0)\le j\le m_0\wedge\ell$,

$$\mathbb{P}(|R\cap\mathcal{H}_0|=j,|R|=\ell)=\begin{cases}\mathcal{Q}_{m,m_0,F}(\tau\wedge t_\lambda,\ell,j) & \text{for } \ell<\lambda,\\ \widetilde{\mathcal{Q}}_{m,m_0,F}(\tau\vee t_\lambda,\ell,j) & \text{for } \ell\ge\lambda.\end{cases} \tag{3.13}$$

where $\mathcal{Q}_{m,m_0,F}(t_1\cdots t_m,\ell,j)$ and $\widetilde{\mathcal{Q}}_{m,m_0,F}(t_1\cdots t_m,\ell,j)$ are given by

$$\binom{m_0}{j}\binom{m-m_0}{\ell-j}(t_\ell)^j(F(t_\ell))^{\ell-j}\Psi_{m-\ell,m_0-j,\overline{F}}(1-t_m,\dots,1-t_{\ell+1}); \tag{3.14}$$

$$\binom{m_0}{j}\binom{m-m_0}{\ell-j}(1-t_{\ell+1})^{m_0-j}(1-F(t_{\ell+1}))^{m-m_0-\ell+j}\Psi_{\ell,j,F}(t_1,\dots,t_\ell), \tag{3.15}$$

respectively, by letting $\overline{F}(t)=1-F(1-t)$.

The above formulas provide the full (joint) distribution of $(|R\cap\mathcal{H}_0|,|R|)$ and therefore also the distribution of the FDP (2.6). Obviously, this also implies explicit expressions for the FDR (2.7) by taking the expectation. These computations were found to be numerically tractable up to m of the order of several hundreds.

Let us provide a brief sketch of proof for these formulas. First, by definition of SUD procedures, it is sufficient to look at SD and SU procedures separately. Second, a crucial property is exchangeability: item (i) is obtained by using that the variable couples (p_i,θ_i), $1\le i\le m$, are i.i.d. under (Mixture), while item (ii) uses that both p_i, $i\in\mathcal{H}_0(\theta)$, and p_i, $i\in\mathcal{H}_1(\theta)$, are i.i.d. (the common distribution

being uniform and F respectively). For instance, for a SD procedure and under assumption (Mixture), the combinatoric argument is as follows:

$$\begin{aligned}
&\mathbb{P}[|\mathcal{H}_0(\theta) \cap \mathrm{SD}(\tau)| = j, |\mathrm{SD}(\tau)| = \ell] \\
&= \binom{\ell}{j}\binom{m}{\ell}\mathbb{P}[\mathrm{SD}(\tau) = \{1, ..., \ell\}, \mathcal{E}_{j,\ell}] \\
&= \binom{\ell}{j}\binom{m}{\ell}\mathbb{P}\left[\forall \ell' \le \ell, \sum_{i=1}^{\ell} \mathbf{1}\{p_i \le \tau_{\ell'}\} \ge \ell', \forall i \ge \ell+1, p_i > \tau_{\ell+1}, \mathcal{E}_{j,\ell}\right] \\
&= \binom{\ell}{j}\pi_0^j\pi_1^{\ell-j}\mathbb{P}\left[\forall \ell' \le \ell, \sum_{i=1}^{\ell} \mathbf{1}\{p_i \le \tau_{\ell'}\} \ge \ell' \,\middle|\, \mathcal{E}_{j,\ell}\right]\binom{m}{\ell}(1-G(\tau_{\ell+1}))^{m-\ell} \\
&= \widetilde{\mathcal{P}}_{m,\pi_0,F}(\tau_1 \cdots \tau_m, \ell, j),
\end{aligned}$$

where $\mathcal{E}_{j,\ell}$ denotes the event "$\theta_1 = ... = \theta_j = 0,\ \theta_{j+1} = ... = \theta_\ell = 1$". This leads to (3.12).

Going back to Theorem 3.3, Assumption (Mixture) allows a markedly simple expression for the distribution of the FDP of a step-up procedure, only depending on functions $\Psi_{m-\ell}(\cdot)$, $1 \le \ell \le m$. As a matter of fact, formula (3.10) relies on the striking fact that, conditionally on $|\mathrm{SU}(\tau)| = \ell$, the number of false discoveries $|\mathrm{SU}(\tau) \cap \mathcal{H}_0(\theta)|$ follows a binomial distribution of parameters ℓ and $\pi_0\tau_\ell/G(\tau_\ell)$. For a step-down procedure, we can show that such a property is not true in general. However, further investigations show that the FDR of a step-down procedure can still be derived only in function of the Ψ_ℓ's (and not the $\Psi_{\ell,j,F}$'s), see Theorem 3.2 in [P16].

3.3.2 Application to least favorable configurations

Now that we have at hand these new formulas, we can examine the issue of assessing whether the configuration (Dirac) (i.e., $F \equiv 1$) maximizes the FDR, or, in other words, whether this configuration is a "least favorable configuration" (LFC) for the FDR. For a *step-up* procedure $\mathrm{SU}(\tau)$, this holds provided that $\ell \mapsto \tau_\ell/\ell$ is nondecreasing, see [14]. By using the new formulas, Theorem 4.1 in [P16] contributes to solve the question for a *step-down* procedure: under (Mixture), it shows that $F \equiv 1$ is an LFC for the FDR of $\mathrm{SD}(\tau)$ (over the set of concave F) if the following function is nondecreasing:

$$\ell \in \{1, ..., m\} \mapsto \sum_{i=0}^{m-\ell} \frac{\tau_\ell}{\ell+i}\binom{m-\ell}{i}\left(\frac{1-\tau_{\ell+i+1}}{1-\tau_\ell}\right)^{m-\ell-i}\Psi_i\left(\left(\frac{\tau_{\ell+j}-\tau_\ell}{1-\tau_\ell}\right)_{1\le j\le i}\right).$$

For instance, after some calculations, we can check that the latter condition is satisfied for the critical values $\tau_\ell = \alpha\ell/m$, $1 \le \ell \le m$.

Now, still for $\tau_\ell = \alpha\ell/m$, $1 \le \ell \le m$, since $F \equiv 1$ is an LFC for the FDR both for $\mathrm{SD}(\tau)$ and $\mathrm{SU}(\tau)$, it is natural to conjecture that this assertion can be extended to the case of a general step-up-down procedure $\mathrm{SUD}_\lambda(\tau)$ which is neither step-up nor step-down (i.e., $\lambda \notin \{1, m\}$). While Theorem 1 in [67] validates the latter asymptotically (when $m \to \infty$), our exact formulas can be used to disprove the conjecture numerically.

Let us consider the one-sided Gaussian model fulfilling (Full-Indep) and (Mixture). The alternative means are supposed constant, with common value Δ. Figure 3.4 exhibits a case where the FDR is not maximal in the case $F \equiv 1$ (corresponding to "$\Delta = \infty$" on the graph). This striking fact has been studied in detail in [P5]: the amplitude of the violation has been upper bounded by a quantity that

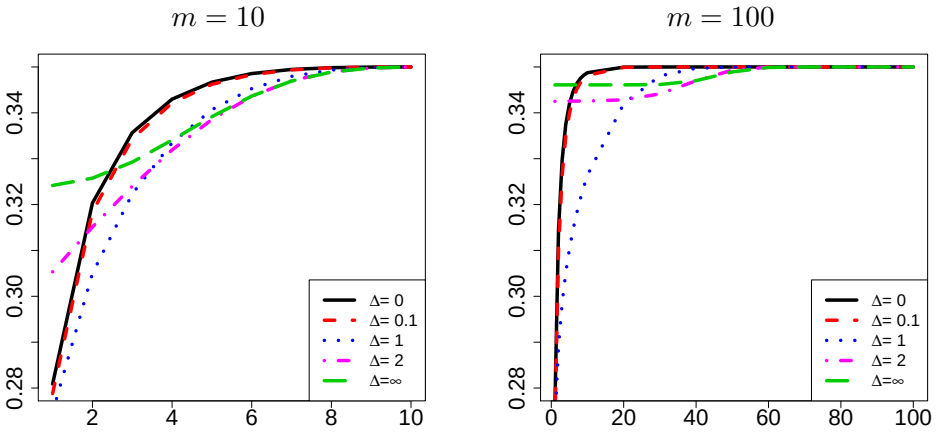


Figure 3.4: FDR of $\mathrm{SUD}_\lambda(\tau)$ for $\tau_\ell = \alpha\ell/m$ as a function of the order $\lambda \in \{1, \dots, m\}$. $\alpha = 0.5$. One-sided Gaussian model under (Full-Indep) and (Mixture) with an alternative mean Δ and $\pi_0 = 0.7$.

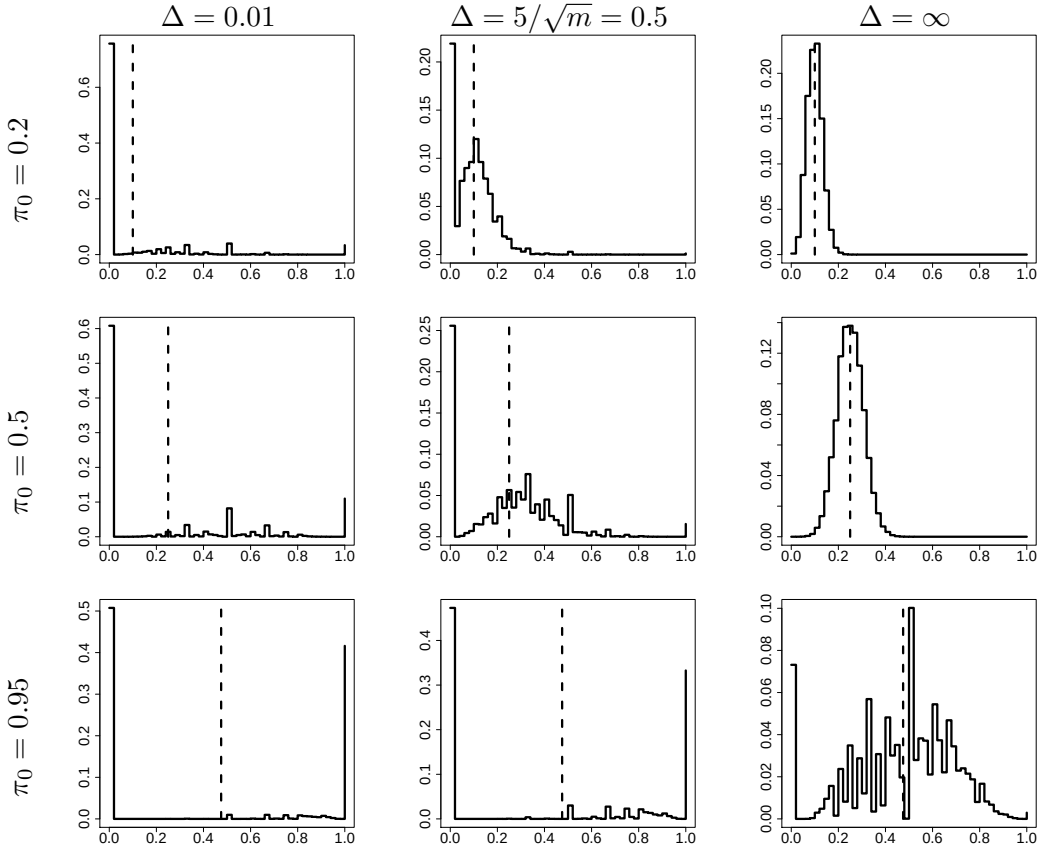
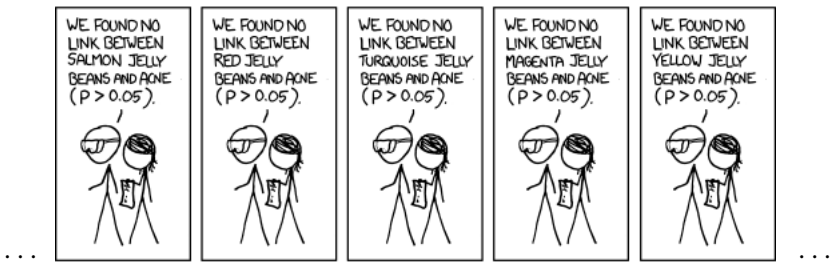


Figure 3.5: Exact probability $\mathbb{P}(\mathrm{FDP}(\mathrm{BH}, P) \in [i/50, (i+1)/50))$ for $0 \le i \le 50$. The value of $\mathrm{FDR}(\mathrm{BH}, P) = \pi_0\alpha$ is displayed by the vertical dashed line. $\alpha = 0.5$, $m = 100$. One-sided Gaussian model under (Full-Indep) and (Mixture) with an alternative mean Δ.

converges to 0 at a specific rate as m grows to infinity. Let us also mention that [P5] provides an asymptotical study of step-up-down procedures which is of independent interest.

Finally, let us go back to our initial motivation with Figure 3.5. It displays the distribution of $\mathrm{FDP}(\mathrm{BH}, P)$ on the basis of the formulas of Theorem 3.3 under (Mixture) for some values of Δ and π_0. Let us emphasize that this distribution is *exact* and does not rely on any Monte-Carlo approximation. This puts forward the lack of concentration of the FDP around the FDR when either Δ is small (low signal strength) or π_0 is close to 1 (sparsity). A consequence is that, even if the BH procedure has a FDR below α, its true underlying FDP is not necessarily near $\pi_0\alpha$. An elementary explanation is that only few rejections are made in that case (i.e., $|R|$ is small): this prevents the FDP to be close[2] to some arbitrary quantity.

[2]let us also mention that we will see in Section 5.2 that such a lack of concentration can also be met for a large rejection number if there are important dependencies between the p-values. This is much more problematic.



Chapter 4

Adaptive procedures under independence

This chapter presents two studies that aim at building adaptive procedures in the context of independence between the p-values. While the first study looks at adaptation w.r.t. the proportion of true null hypotheses, the second one deals with an adaptation w.r.t. the marginal distribution of the p-values under the alternative. Both work aim at building FDR controlling procedures *more powerful* than the BH procedure.

4.1 Adaptation to the proportion of true nulls [P7]

4.1.1 Background

In Theorem 2.2, the FDR control is provided at level $\pi_0\alpha$ instead of α, for $\pi_0 = m_0/m$. Hence, when π_0 is not close to 1, this gap entails an inevitable loss of power. If π_0 is known, a simple way to solve this issue is to use the BH procedure at level α/π_0 instead of α, that is, to use the step-up procedure with critical values $\pi_0^{-1}\alpha\ell/m$, $\ell \in \{1, \dots, m\}$. The latter is classically called the *oracle BH procedure*. By Theorem 2.2, it controls the FDR under an independence or PRDS assumption. However, π_0 is often unknown and we should estimate π_0, or more precisely

$$\vartheta = \pi_0^{-1}$$

(with the notation of Section 2.5.2) and then incorporate this estimator in the critical values in a way that still provides FDR control.

The question of estimating π_0 is a wide and rich research field of independent interest, which started with the work of Tore Schweder and Emil Spjøtvoll (1982) [123] and has undertaken developments of various natures, including histogram estimators [25], kernel estimators [99], Fourier transform [82, 81] and optimality results [23, 100]. The problem of finding π_0-adaptive procedures is slightly different in the sense that we should focus on estimators simple enough to have a convenient behavior when used in combination with a step-up procedure. This issue has been initially proposed in [12], and has been followed by many work, either asymptotic [59, 134, 98, 49, 99], or nonasymptotic, e.g., [17, 13, 119]. Our contribution [P7] lies in the last category and is the topic of this section.

4.1.2 One-stage adaptive procedures

A first class of adaptive procedures is the class of *one-stage adaptive procedures*. Such procedures perform a single round of step-wise adjustment and are based on particular deterministic critical

values that render them adaptive. A contribution of [P7] was to introduce a new one-stage adaptive procedure. For $\lambda \in (0,1)$, consider the step-up procedure SU(τ) using the critical values

$$\tau_\ell = \min\left((1-\lambda)\frac{\alpha\ell}{m-\ell+1}, \lambda\right), \quad 1 \le \ell \le m, \tag{4.1}$$

which we denote by BR-1S-λ (or simply BR-1S). We have proved the following FDR control.

Theorem 4.1. *Let $\lambda \in (0,1)$. Assume that the p-value family satisfies* (pvalueprop) *and* (Full-Indep)*, then BR-1S-λ controls the FDR at level α.*

Compared to the critical values of the BH procedure, those in (4.1) include an implicit "step-wise" estimation of π_0^{-1} by the quantity $(1-\lambda)m/(m-\ell+1)$ (up to the λ-capping), where ℓ is the "current" number of rejections. For comparison purpose, we have displayed these critical values on Figure 4.1, together with those coming from the asymptotically optimal rejection curve (AORC): $\alpha\ell/(m-(1-\alpha)\ell)$, $1 \le \ell \le m$, which defines a one-stage step-up procedure which is shown to be asymptotically optimal, see [49]. Here, the BR-1S critical values, although dominated, are admittedly close to the AORC critical values (before the capping). The capping by λ into (4.1) is required to prevent the estimator of π_0^{-1} to be too large and thus to provide the desired FDR control. Also note that the AORC does not control the FDR as it is, because the largest critical value is equal to 1; hence the corresponding step-up procedure always rejects all the nulls. Several modifications have been proposed in [49] in order to control the FDR asymptotically, the simplest one being to take $\min\big(\alpha\ell/(m-(1-\alpha)\ell), \eta^{-1}\alpha\ell/m\big)$, $1 \le \ell \le m$, for some parameter $\eta \in (0,1)$. It is denoted by FDR09-η in Figure 4.1. Let us also mention that the step-down version of the AORC controls the FDR nonasymptotically (up to add "+1" in the denominator) as proved by [57]. We also refer the reader to [50] for recent developments on AORC modifications.

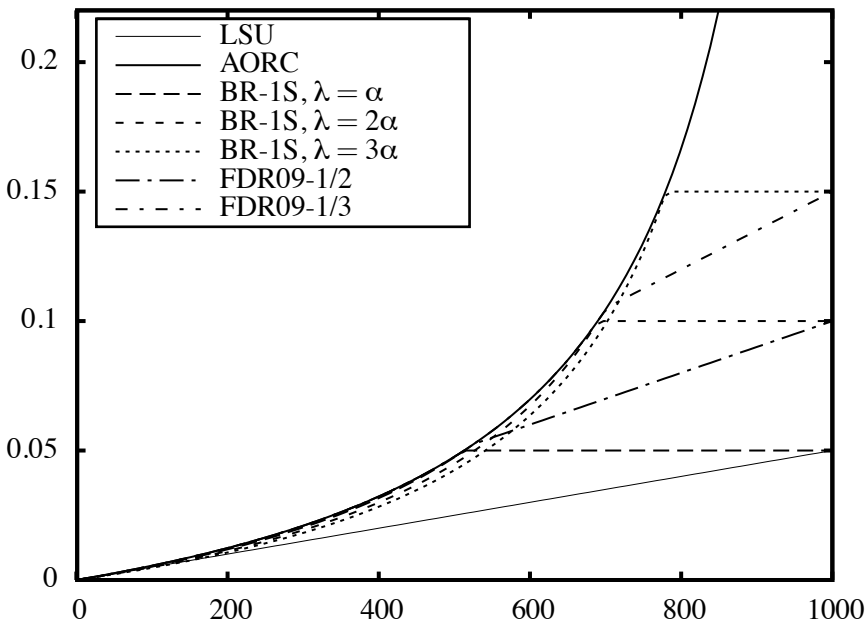


Figure 4.1: Comparison of critical values (4.1) to that of the BH (denoted LSU), the AORC and FDR09-η, see text. $m = 1000$, $\alpha = 0.05$.

4.1.3 Two-stage adaptive procedures

A second class of adaptive procedures is the class of *plug-in adaptive procedures*, which use the critical values

$$\tau_\ell = \widehat{\vartheta}\,\alpha\ell/m, \quad 1 \le \ell \le m, \tag{4.2}$$

where $\widehat{\vartheta}$ is an estimator of π_0^{-1}, taken as a function of the p-value family $(p_i(X), 1 \le i \le m)$. For instance, this estimator can be obtained via a first round of multiple testing procedure R_0, in which case the plug-in procedure is said *two-stage*. The next result is an unified reformulation of the results presented in [13].

Theorem 4.2. *Assume that the p-value family satisfies* (pvalueprop) *and* (Full-Indep). *Consider an estimator $\widehat{\vartheta}$ which is a coordinate-wise nonincreasing function of the p-value family $(p_i(X), 1 \le i \le m)$. Then the step-up procedure $SU(\tau)$ with τ given by* (4.2) *has an FDR smaller than or equal to αb_m, where*

$$b_m = \max_{1 \le u \le m} \left\{ \frac{u}{m} \mathbb{E}_{DU(m,u-1)} \left(\widehat{\vartheta} \right) \right\}, \tag{4.3}$$

in which $DU(m, j)$ denotes the configuration where the j first p-values are i.i.d. $U(0, 1)$ and the $m - j$ others are equal to 0. Furthermore, the following estimators are such that $b_m \le 1$ and thus entails an FDR bounded by α:

$$\begin{aligned}
&\textit{[Storey-}\lambda\textit{]} && \widehat{\vartheta}_1 = \frac{(1-\lambda)m}{\sum_{i=1}^m \mathbf{1}\{p_i > \lambda\} + 1}\,; \\
&\textit{[Quant-}k_0/m\textit{]} && \widehat{\vartheta}_2 = \frac{(1-p_{(k_0)})m}{m - k_0 + 1}\,; \\
&\textit{[BKY06-}\lambda\textit{]} && \widehat{\vartheta}_3 = \frac{(1-\lambda)m}{m - |R_0| + 1}\,, \ \textit{where } R_0 \textit{ is the standard BH procedure at level } \lambda\,; \\
&\textit{[BR-2S-}\lambda\textit{]} && \widehat{\vartheta}_4 = \frac{(1-\lambda)m}{m - |R_0'| + 1}\,, \ \textit{where } R_0' \textit{ is BR-1S-}\lambda,
\end{aligned}$$

in which $\lambda \in (0, 1)$ and $k_0 \in \{1, \dots, m\}$ are fixed parameters.

The first part of Theorem 4.2 can be proved by using the methodology developed in Section 3.1. In addition, the monotonic property of $\widehat{\vartheta}$ implies that the $DU(m, m_0 - 1)$ configuration is least favorable, which leads to the bound (4.3). The second part of Theorem 4.2 mainly uses that for $k \ge 2$, $q \in (0, 1]$, a binomial random variable Y with parameter $(k - 1, q)$ satisfies $\mathbb{E}((1 + Y)^{-1}) = 1/(kq)$, as already shown in [13].

The estimators defining "Storey-λ", "Quant-$\frac{k_0}{m}$" and "BKY06-λ" have been introduced in [132], [12] and [13][1], respectively. "BR-2S-λ" is a new two-stage adaptive procedure which uses in the first stage the new adaptive procedure "BR-1S-λ".

4.1.4 Robustness to dependence

Which adaptive procedures should be used in practice? To address this issue, a possible angle is to find a procedure which is both powerful under independence and robust to dependence, for instance in the one-sided Gaussian ρ-equicorrelated case (Gauss-ρ-equi). First, standard calculations when $\rho = 1$ suggest to use $\lambda = \alpha$ in the above procedures. Second, an extensive simulation study making varying the alternative mean, π_0 and ρ shows that the adaptive plug-in procedure Storey-α seems to be a good power/robustness tradeoff and so we recommend it in priority. This recommendation is noticeably

[1] More precisely, the estimator in [13] does not use the "+1" in the denominator.

different than the standard choice $\lambda = 1/2$ of [131], which substantially lacks of robustness in our simulations (as already reported in [13]), because of the variance of the corresponding π_0^{-1}-estimator.

In addition, let us note that an asymptotic study of some of these adaptive procedures has been made in [98]. Interestingly, while connections between BKY06 and BR-1S are proposed, it is also proved that BR-1S is asymptotically more powerful than BKY06 for π_0 large enough.

Finally, let us mention that we have also[2] provided in this work an adaptive two-stage procedure that controls the FDR at level α under (PRDS). It uses the critical values

$$\tau_\ell = \frac{\alpha \ell}{2m} \left(1 - (2|R_0|/m - 1)_+^{1/2}\right)^{-1}, \ 1 \leq \ell \leq m,$$

where R_0 is the BH procedure at level $\alpha/4$. Although this is among the first results that provide adaptiveness under dependence, it is shown to improve the standard BH procedure only over a narrow range where π_0 has very small values. The reason is that the underlying π_0^{-1}-estimator is built upon Markov's inequality, which is not very accurate. Interestingly, a way to improve the power of our procedure has been recently proposed in [71], by incorporating the values of the pairwise correlation between the test statistics (e.g., the value of ρ under (Gauss-ρ-equi)).

4.2 Adaptation to the alternative structure [P15]

4.2.1 Motivation

In this section, we consider the case where the null hypotheses are not "equally treated" in the data, because the alternative p-value marginal distributions are *heterogeneous*. This is the case when the sample size available to test each null varies across the null hypotheses. Let us provide below two such examples.

- Adequate yearly progress (AYP) data[3] collect the academic performance of California high schools and have been presented as reference data for the multiple testing problem by Bradley Efron [41, 42]. These data report the academical performances of students together with several characteristics, as the socioeconomic status. One issue is then to detect the schools where the performance difference between economically advantaged and disadvantaged students is significantly larger than a reference. This can be studied by using a multiple testing procedure, for which the m tested items are the schools. In [24], it is pointed out that "the number of scores reported by each school varies from less than a hundred to more than ten thousands". Hence, ignoring the school size would inevitably advantage the large schools, because they have more experiments at hand to detect a signal of a given strength.

- In typical microarray experiments, we want to find the genes that are differentially expressed between two groups (see Section 1.3). Now, assume that the groups are built according to a binary covariate which characterizes the gene (e.g., the DNA copy number alteration "normal" or "amplified"), which makes the group sizes different across the genes. Then, a two-sample test will be more able to discover an effect when the group sizes are balanced (e.g., n vs n more favorable than 2 vs $2n-2$). Here, again, performing a multiple test that ignores this information would inevitably give more "detection chance" to genes related to balanced groups, just because the corresponding individual tests are more efficient, and not because the signal is stronger.

[2] Additional results were provided in [P7], including a procedure controlling the FDR under general dependence which is based on β-functions (3.2).

[3] Data mandated by the "No Child Left Behind Act of 2001"; available at http://www.cde.ca.gov/ta/ac/ay/

4.2.2 Optimal p-value weighting

The examples above show that it is desirable to incorporate the heterogeneous structure in our multiple testing decision. For this, a natural approach is the *p-value weighting*: before applying the multiple testing procedure, we replace each initial p-value $p_i(X)$ by its weighted counterpart $p'_i(X) = p_i(X)/w_i$, for some *weight vector* $w = (w_i)_{1\leq i\leq m} \in (\mathbb{R}^+)^m$, with the conventions $0/0 = 0$, $1/0 = \infty$. One major issue is then to find the weight vector $w^\star$ that is the most suitable for the heterogeneous structure, which is called the *optimal weight vector*.

For FWER control and by using the weighted Bonferroni procedure, the optimal weighting has been found in [117, 142, 110]. For the FDR, however, while it is shown that "informative" weighting can improve the BH procedure [60], no such optimal weighting were derived. The goal of our study [P15] is to provide such an optimality result.

For simplicity, let us consider the one-sided Gaussian framework under assumptions (pvaluepropunif), (Full-Indep) and (Mixture)[4]. In that case, the p-values are independent and for each $i = 1, \dots, m$, the variable $p_i(X)$ has (unconditionally) for distribution function

$$G_i(t) = \pi_0 t + \pi_1 \overline{\Phi}(\overline{\Phi}^{-1}(t) - \mu_i), \;\; t \in [0,1], \tag{4.4}$$

for some overall vector $\mu = (\mu_i)_{1\leq i\leq m}$ of candidate (positive) alternative means. The optimization problem can be set as follows (with the notation of Section 2.5.2): given

$$\vartheta = \mu,$$

what is the optimal weight vector $w^\star$ that maximizes the power (2.9) of the weighted BH procedure? Unfortunately, this maximization problem seems intractable, because the BH threshold is both data dependent and defined by a self-referring condition. The solution we have chosen is based on the following simple observation, based on [117, 142, 110]:

Lemma 4.3. *Consider the set* $\mathcal{W} = \{w = (w_i)_{1\leq i\leq m} \in (\mathbb{R}^+)^m : \sum_{i=1}^m w_i = m\}$ *containing all the possible weight vectors. For each* $u \in [0,1]$*, the function*

$$w \in \mathcal{W} \mapsto Pow_u(w) = (\pi_1/m) \sum_{i=1}^m \overline{\Phi}\left(\overline{\Phi}^{-1}(w_i \alpha u) - \mu_i\right) \tag{4.5}$$

attains its maximum in $w^\star(u)$ *given by*

$$w_i^\star(u) = (\alpha u)^{-1} \overline{\Phi}\left(\frac{\mu_i}{2} + \frac{c(u)}{\mu_i}\right), \;\; 1 \leq i \leq m, \tag{4.6}$$

where $c(u)$ *is the unique element of* $\mathbb{R}$ *such that* $\sum_{i=1}^m w_i^\star(u) = m$.

Above, $\mathrm{Pow}_u(w)$ is the power of the procedure $R = \{1 \leq i \leq m : p_i(X) \leq w_i \alpha u\}$, which intuitively corresponds to the "weighted BH procedure at rejection proportion u". Figure 4.2 displays the shape of $w_i^\star(u)$, $1 \leq i \leq m$. Interestingly, for a given u, the weighting $w^\star(u)$ favors the nulls in a "moderate regime" of alternative means. An explanation is that, in order to maximize $\mathrm{Pow}_u(w)$, this weighting "sacrifizes" the μ_i's "too small" because they will not entail a small p-value (with high probability) and thus would not contribute into $\mathrm{Pow}_u(w)$ anyway. At the opposite side, this weighting does not "help" the μ_i's "too large" because they will generate a small p-value (with high probability) and thus would contribute into $\mathrm{Pow}_u(w)$ anyway. Additionally, observe that, as expected, the range of this moderate regime is strongly affected by the value of u. This indicates that a strategy focusing on a prescribed single weight vector $w^\star(u_0)$ (say $u_0 = 1$) is certainly suboptimal. Hence, we propose to use simultaneously all the weighting vectors with the following new procedure:

[4] Different alternative distributions can be used; sufficient conditions are given in [P15]

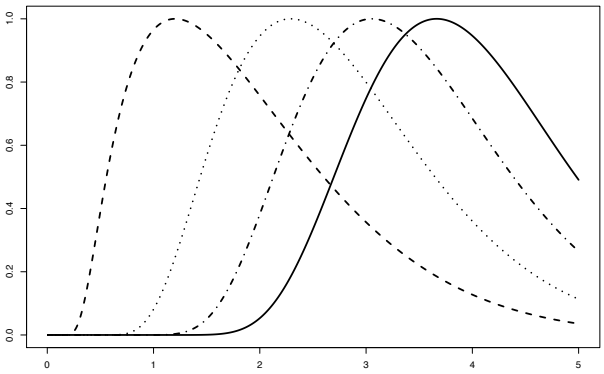


Figure 4.2: Plot $w_i^\star(u)$ in function of $\mu_i = 5i/m$, $1 \le i \le m$ for several values of u; $u = 1/m$ (solid), $u = 10/m$ (dashed-dotted), $u = 100/m$ (dotted), $u = 1$ (dashed). $m = 1000$, $\alpha = 0.05$. Each curve is normalized to have a maximum equal to 1.

Algorithm 4.4. [Optimal multi-weighted BH procedure at level α]

- Compute the weighting vectors $w^\star(\ell/m)$, for $\ell \in \{1, \dots, m\}$ according to (4.6);
- For each $\ell \in \{1, \dots, m\}$, compute $q_\ell(X)$ the ℓ-th smallest $w^\star(\ell/m)$-weighted p-value, i.e., $q_\ell(X) = p'_{(\ell)}(X)$ where $p'_{(1)}(X) \le \cdots \le p'_{(m)}(X)$ are the ordered values of $p'_i(X) = p_i(X)/w_i^\star(\ell/m)$, $1 \le i \le m$. Let also $q_0 = 0$ by convention;
- consider the integer $\widehat{\ell} = \max\{\ell \in \{0, 1, \dots, m\} \ : \ q_\ell \le \alpha\ell/m\}$;
- finally let $R = \left\{1 \le i \le m : p_i(X) \le w_i^\star(\widehat{\ell}/m)\alpha\widehat{\ell}/m\right\}$.

The main innovation of this procedure is that the p-values are ordered in several ways, sequentially and according to the weight vector that suits the best to each considered rejection number. It is therefore a good candidate to outperform all the weighted BH procedures.

4.2.3 Results

Unfortunately, we can show that, in general, the multiple testing procedure defined by Algorithm 4.4 looses the finite sample FDR control at level α (even in this independent setting). This comes from the additional variations generated by the term $w^\star(\widehat{\ell}/m)$. Instead, we have proved that replacing the weighting $w_i^\star(u)$ by $w_i^\star(u)/(1+\alpha w_i^\star(1))$ into Algorithm 4.4 ensures the finite sample FDR control. Also, we have stated the finite sample optimality of Algorithm 4.4 (up to some remainder terms) by using concentration inequalities. For the sake of simplicity we report below only the asymptotic counterparts of these results in the case where the signal strengths, i.e., the μ_i's, are "grouped"[5].

Let us consider the case where the values of μ are structured into $D \ge 2$ clusters $A_d^{(m)}$, $1 \le d \le D$, that is, where for all d and $i \in A_d^{(m)}$, $\mu_i = \Delta_d$, for some fixed positive values Δ_d, $1 \le d \le D$, not depending on m. Further assume that $|A_d^{(m)}|/m$ is converging to some value as m tends to infinity, for all d. Here, note that the clusters are assumed to be known and given a priori. For instance, in AYP data of Section 4.2.1, the clusters can be built according to small, medium and large schools.

To study the asymptotic behavior of the weighted BH procedure in that grouped context, it is natural to introduce the class of *grouped weight vector sequences* $(w^{(m)})_m$, which contains weight

[5]In [P15], the asymptotic optimality is also shown for μ_i of the form $\psi(i/m)$ for some continuous function ψ.

vectors with values $w_i^{(m)}$, $1 \leq i \leq m$ that are constant on each cluster (also these values are supposed to converge as m tends to infinity). Then the following result holds.

Theorem 4.5. *In the above grouped setting, let us denote by $R_m^\star$ the optimal multi-weighted BH procedure defined by Algorithm 4.4 and by $BH_m(w^{(m)})$ the $w^{(m)}$-weighted BH procedure, for some given weight vector $w^{(m)}$. Then, as m grows to infinity, we have*

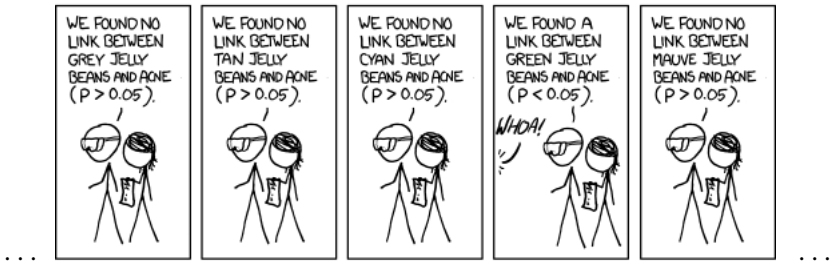
(i) $\lim_m FDR(R_m^\star) \leq \pi_0\, \alpha$;

(ii) $\lim_m Pow(R_m^\star) \geq \max_{(w^{(m)})_m} \left\{ \lim_m Pow\left(BH_m\left(w^{(m)}\right)\right)\right\}$, *where the maximum is taken over all the grouped weight vector sequences.*

Let us underline that the potential benefit of the new procedure has been evaluated by using simulations in [P15]: the obtained graphs are satisfactory: compared to the (uniformly weighted) BH procedure, the new optimal procedure can make more than $0.15 \times m_1$ additional true discoveries. Hence, taking into account the heterogeneity of the alternatives is definitely useful to increase power.

In addition, let us go back to the real data microarray motivation (second example in Section 4.2.1). We have applied the optimal multi-weighted procedure on the data [97] (displayed in Figure 1.4) and it overall increases the number of rejections w.r.t. a standard BH procedure, see [P15] for more details. Nevertheless, in this data analysis, the μ_i's are chosen from the sample sizes by using an estimator of a "global effect size". Strictly speaking, this makes the μ_i's data-depend which is not allowed in our theoretical investigations. Hence, the corresponding statistical analysis is not fully theoretically supported. More generally, the investigations made in this section mainly concern the *oracle* part of the adaptation, because the optimal weighting depends on the true alternative means, which are often unknown.

Since our work, some advances have been done to incorporate data-driven weighting in the case of grouped nulls: while a way to incorporate the possible heterogeneity structure of the proportion of true nulls has been proposed in [76], the issue left open in our work has been investigated in [144]. The latter work provides one-stage and two-stage data-driven weighted procedures and both an asymptotical FDR control and an asymptotical improvement over the BH procedure are proved. However, obtaining an asymptotical optimal procedure with data-driven weights seems yet to be an unsolved issue.



Chapter 5

Adaptation to the dependence structure

This chapter deals with adaptation with respect to the *dependence structure*. In Section 5.1, we first present the method of Joseph P. Romano and Michael Wolf (2005) [115], which provides a general solution to the problem of adaptive FWER control via randomized p-values. This method, which can be used with any randomization tool, is then applied with a sign-flipping technique. By contrast, the case of FDR/FDP control is more puzzling. Section 5.2 contributes to address the oracle part of the problem (i.e., when the dependence is known), by exploring a general heuristic proposed by Joseph P. Romano and Michael Wolf (2007) [116]. Finally, in Section 5.3, the adaptation to the unknown dependence structure is investigated by exploring another type of solution, that combines the dependence structure estimation (via clustering) with the hierarchical FWER control. The basic tool is the sequential rejection principle developed in [63].

5.1 Adaptive FWER control [P3]

This section presents a first[1] contribution of our work [P3], which is to provide a new application of the general approach of [115] (called RW's method below) in combination with a specific randomization technic, often referred to as "symmetrization". It can be seen as a variation of Example 5 in [115] which was investigated in the permutation case. Sections 5.3 and 6.1 will rely on similar arguments.

5.1.1 Reformulating Romano-Wolf's general method

In a nutshell, the FWER control corresponds to control the infimum of p-values; more formally, for a multiple testing procedure R rejecting nulls with a p-value $p_i(X)$ such that $p_i(X) < \widehat{t}$,[2] we have

$$\mathrm{FWER}(R,P) = \mathbb{P}(|R \cap \mathcal{H}_0(\theta)| \geq 1) = \mathbb{P}(\exists\, i \in \mathcal{H}_0(\theta) \,:\, p_i(X) < \widehat{t}) = \mathbb{P}\left(\inf_{i\in\mathcal{H}_0(\theta)}\{p_i(X)\} < \widehat{t}\right).$$

Hence, the ideal threshold $\widehat{t}$ is $\sup\left\{t \,:\, \mathbb{P}\left(\inf_{i\in\mathcal{H}_0(\theta)}\{p_i(X)\} < t\right) \leq \alpha\right\}$, the α-quantile[3] of the distribution of $\inf_{i\in\mathcal{H}_0(\theta)}\{p_i(X)\}$. While the latter depends on the joint distribution of the p-values under the null, it also primarily involves the unknown set $\mathcal{H}_0(\theta)$. Hence, a step of $\mathcal{H}_0(\theta)$ "localization" is needed.

[1] A second contribution of [P3] will be presented in Section 6.1.

[2] Here, by contrast with (2.2), the inequality is taken strict.

[3] Note that this corresponds to use the standard definition of the quantile function on the test statistic scale: for any one-to-one decreasing function ψ, we have $\widehat{t} = \psi^{-1}(\widehat{s})$, where $\widehat{s} = \inf\left\{s \,:\, \mathbb{P}\left(\sup_{i\in\mathcal{H}_0(\theta)}\{\psi(p_i(X))\} \leq s\right) \geq 1-\alpha\right\}$.

Let us consider $\widehat{t} = \widehat{t}_{\mathcal{H}_0(\theta)}$ as a member of a subset-indexed threshold collection $\{\widehat{t}_{\mathcal{C}}, \mathcal{C} \subset \{1,\dots,m\}\}$, with the property

$$\forall P \in \mathcal{P}, \text{ for } \mathcal{C} = \mathcal{H}_0(\theta(P)), \quad \mathbb{P}\left(\inf_{i\in\mathcal{C}}\{p_i(X)\} < \widehat{t}_{\mathcal{C}}\right) \leq \alpha. \tag{5.1}$$

While an union bound would yield $t_{\mathcal{C}} = \alpha/|\mathcal{C}|$, more accurate thresholds learning the dependence structure can be provided via randomized tests, as we will investigated in Section 5.1.3 in the Gaussian case. Now, the RW method provides an FWER control from (5.1), by only assuming a *monotonic* assumption on the family of thresholds: namely,

$$\forall \mathcal{C}, \mathcal{C}' \subset \{1,\dots,m\} \text{ such that } \mathcal{C} \subset \mathcal{C}', \text{ we have } \widehat{t}_{\mathcal{C}} \geq \widehat{t}_{\mathcal{C}'} \text{ (pointwise)}. \tag{5.2}$$

Let us mention that a benefit of the RW approach is that (5.2) replaces the quite undesirable "subset pivotality condition", roughly supposing the presence of an "overall least favorable null distribution", see [143] and [31].

Single step procedure By using successively (5.2) and (5.1), an FWER controlling procedure can be easily derived by taking $\mathcal{C} = \{1,\dots,m\}$, that is,

$$R_1 = \{1 \leq i \leq m \ : \ p_i(X) < \widehat{t}_{\{1,\dots,m\}}\}.$$

Indeed, the FWER of R_1 is upper bounded by the quantity

$$\mathbb{P}\left(\inf_{i\in\mathcal{H}_0(\theta)}\{p_i(X)\} < \widehat{t}_{\{1,\dots,m\}}\right) \leq \mathbb{P}\left(\inf_{i\in\mathcal{H}_0(\theta)}\{p_i(X)\} < \widehat{t}_{\mathcal{H}_0(\theta)}\right) \leq \alpha.$$

The procedure R_1 is generally called *single step*, because only $\widehat{t}_{\{1,\dots,m\}}$ is used. However, the latter can be too conservative w.r.t. the ideal threshold $\widehat{t}_{\mathcal{H}_0(\theta)}$, especially when the set $\mathcal{H}_0(\theta)$ is "small" or when $\widehat{t}_{\mathcal{C}}$ is inaccurately designed when $\mathcal{C}$ exceeds $\mathcal{H}_0(\theta)$[4].

Step-down procedure To improve over R_1, consider the following iterative approach: take the set R_1^c corresponding to the nulls that are not rejected by R_1 and then apply the single step procedure in restriction to R_1^c, i.e., use the threshold $\widehat{t}_{R_1^c} \geq \widehat{t}_{\{1,\dots,m\}}$. This gives a new rejection set $R_2 \supset R_1$. Now, iteratively repeat this operation and stop the first time that no new null is rejected, say for $R_{\widehat{k}}$. Finally reject $R_{\widehat{k}}$.

Quite strikingly, this way to increase the number of discoveries maintains the FWER control, as the following result shows.

Theorem 5.1 ([115]). *Consider any threshold collection $\{\widehat{t}_{\mathcal{C}}, \mathcal{C} \subset \{1,\dots,m\}\}$ satisfying (5.1) and (5.2). Consider the iterative sequence of rejection sets starting with $R_0 = \emptyset$ and such that for all $k \geq 1$,*

$$R_k = \{1 \leq i \leq m \ : \ p_i(X) < \widehat{t}_{R_{k-1}^c}\},$$

with the stopping rule $\widehat{k} = \min\{k \geq 1 \ : \ R_k = R_{k-1}\}$. Then the multiple testing procedure $R_{\widehat{k}}$ controls the FWER at level α.

[4]This can be due to randomization technics generating a "bad $\mathcal{H}_0$ distribution". An instance is given in Section 6.1.4 in a Gaussian setting.

Markedly, a "one-line proof" can be built for Theorem 5.1 by using the following acceptance functional

$$A(\mathcal{C}) = \{1 \le i \le m \ : \ p_i \ge \widehat{t}_{\mathcal{C}}\},$$

which is nondecreasing, that is, $\forall \mathcal{C}, \mathcal{C}'$ such that $\mathcal{C} \subset \mathcal{C}'$, $A(\mathcal{C}) \subset A(\mathcal{C}')$, by (5.2). Now, the argument is that on the event $\Omega_0 = \{\mathcal{H}_0 \subset A(\mathcal{H}_0)\}$, we have for all $\mathcal{C} \subset \{1, \dots, m\}$,

$$\mathcal{H}_0 \subset \mathcal{C} \text{ implies } A(\mathcal{H}_0) \subset A(\mathcal{C}) \text{ and thus } \mathcal{H}_0 \subset A(\mathcal{C}).$$

Hence, on the event Ω_0, all the procedures R_k described in Theorem 5.1 do not make false discoveries. The proof is finished because Ω_0 holds with probability at least $1 - \alpha$ by (5.1).

As a first instance, we apply Theorem 5.1 with the threshold family defined by $t_{\mathcal{C}} = \alpha/|\mathcal{C}|$, which satisfies both (5.1) and (5.2). The resulting procedure turns out to be the standard procedure of [75], that is, the step-down procedure SD(τ) with critical values $\tau_\ell = \alpha/(m - \ell + 1)$, $1 \le \ell \le m$. This equivalence is illustrated on Figure 5.1. The left picture is the classical step-down representation (the filled points represent p-values that correspond to the rejected nulls). The right picture illustrates the algorithm of Theorem 5.1 by displaying the $\widehat{k} = 3$ thresholds $\alpha/10$ (step 1), $\alpha/7$ (step 2) and $\alpha/5$ (step 3) (for $i \in \{1, 2\}$, the points filled with the symbol "i" are rejected at the i-th step of the algorithm). Both pictures use the same realization of the p-values for $m = 10$ and $\alpha = 0.5$.

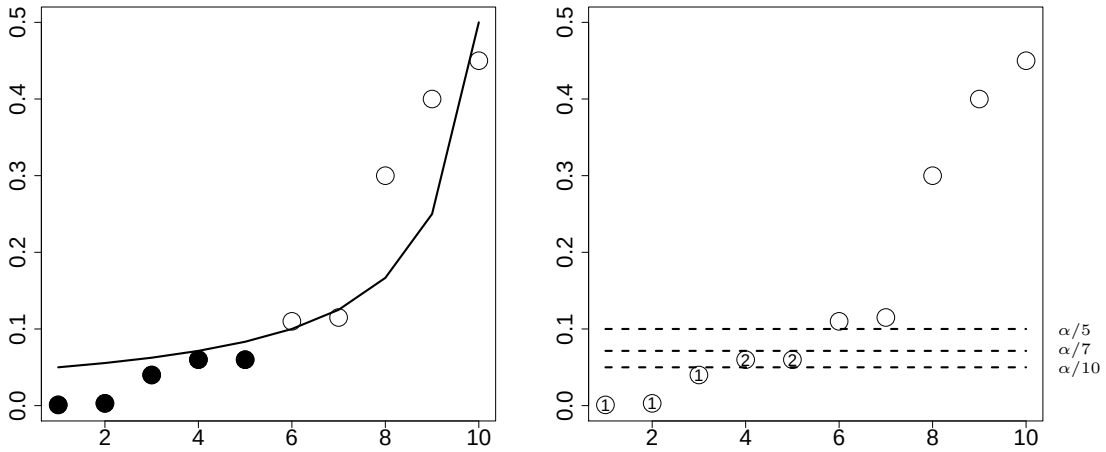


Figure 5.1: Illustration of the two equivalent definitions of Holm's procedure, see text.

However, notice that Holm's procedure does not take into account the dependencies. As an extreme instance, if all the p-values are equal, Holm's procedure ignores it and use $t_{\mathcal{C}} = \alpha/|\mathcal{C}|$ instead of $t_{\mathcal{C}} = \alpha$. This results in a huge loss of power. This raises the question of finding a better threshold collection $\{\widehat{t}_{\mathcal{C}}, \mathcal{C} \subset \{1, \dots, m\}\}$ that incorporates the dependence.

5.1.2 Oracle adaptive FWER control

In this section, we consider the case where the dependencies between the p-values is *known*. From now, let us focus on the two-sided Gaussian setting defined in Section 2.1, in which $X \sim \mathcal{N}(\mu, \Gamma)$ and $p_i = 2\overline{\Phi}(|X_i|)$. For any distribution Q on $\mathbb{R}^m$, we let

$$q_\alpha(\mathcal{C}, Q) = \sup \left\{ t \in [0,1] \ : \ \mathbb{P}_{Z \sim Q} \left(\inf_{i \in \mathcal{C}} \{2\overline{\Phi}(|Z_i|)\} < t \right) \le \alpha \right\} \tag{5.3}$$

the α-quantile of the distribution of $\inf_{i\in\mathcal{C}}\{2\overline{\Phi}(|Z_i|)\}$ whenever $Z \sim Q$.

Conveniently, in the Gaussian case, (5.1) is satisfied by taking $\widehat{t}_{\mathcal{C}}$ equal to $q_\alpha(\mathcal{C}, Q_0)$, with $Q_0 = \mathcal{N}(0,\Gamma)$ and where $q_\alpha(\mathcal{C},\cdot)$ is defined by (5.3). Since this threshold family also satisfies (5.2), applying Theorem 5.1 directly yields an adaptive FWER controlling procedure.

To evaluate the potential benefit of the above threshold collection, let us consider the case of equi-correlation (Gauss-ρ-equi). First, by exchangeability, $t_{\mathcal{C}}$ only depends on $|\mathcal{C}|$ so that RW's procedure reduces to a simple step-down algorithm SD(τ) in a sense of (2.14). Second, by contrast with Holm's procedure, for each ℓ, the critical value τ_ℓ is equal to the α-quantile of the distribution of $\inf_{1\le i\le m-\ell+1}\{2\overline{\Phi}(|Z_i|)\}$, with $Z \sim \mathcal{N}(0,\Gamma)$, so take into account Γ. For instance, if $\rho = 1$, we have $\tau_\ell = \alpha$ for all ℓ, which means that no multiple testing correction is performed (as desired in that situation). For $\rho \in [0,1)$, by using (2.12), we can obtain τ_ℓ as the value $t \in [0,1]$ solving the equation :

$$\int_0^1 (1 - f(t/2, w) - f(t/2, -w))^{m-\ell+1}\phi(w)dw = 1-\alpha,$$

where f is defined by (2.11) and ϕ is the standard Gaussian density. Figure 5.2 displays the resulting critical values for different values of ρ. While they appear close to Holm's critical values for $\rho = 0$, we notice that they increase as ρ increases (to end up at $\tau_\ell = \alpha$ when $\rho = 1$, as mentioned above).

Outside the equi-correlated case, the thresholds $t_{\mathcal{C}}$ can simply be adjusted from Γ via Monte-Carlo simulations. Overall, this indicates that procedures less conservative than Holm's procedure (or Bonferroni procedure) can be used to control the FWER by incorporating appropriately Γ (i.e., ρ under equi-correlation). However, note that this adaptive procedure is only *oracle* because it uses explicitly Γ (i.e., the true value of the parameter $\vartheta = \Gamma$ with the notation of Section 2.5.2).

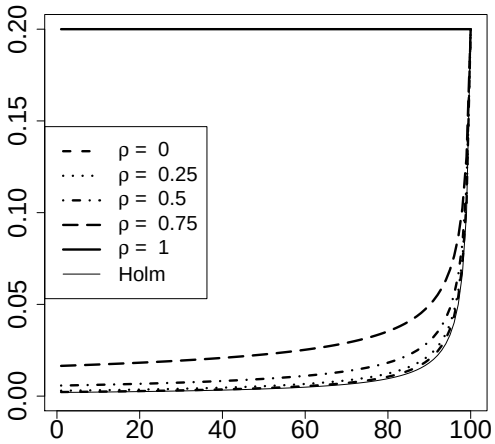


Figure 5.2: Illustration of oracle RW's critical values for FWER control in the two-sided (Gauss-ρ-equi) case. $m = 100$, $\alpha = 0.2$.

5.1.3 Randomized adaptive procedure

To "learn" the distribution of a multivariate vector without using the knowledge of the covariance Γ, a classical requirement is that $n \ge 2$ copies $X^{(1)}, \dots, X^{(n)}$ of the vector are at hand. Let us assume that the data are of the following form

$$X = (X^{(1)} \dots X^{(n)}) = \begin{pmatrix} X_1^{(1)} & \dots & X_1^{(n)} \\ \vdots & & \vdots \\ X_m^{(1)} & \dots & X_m^{(n)} \end{pmatrix}, \tag{5.4}$$

where the vectors $X^{(1)}, \dots, X^{(n)}$ are i.i.d. of common (multivariate) distribution $\mathcal{N}(\mu, \Gamma)$. Since we deal with high dimensional data, note that m is typically much larger than n. The p-values are given by $p_i(X) = 2\overline{\Phi}(|S_i(X)|)$, where the statistic $S(X)$ is

$$S(X) = n^{1/2}(\overline{X}_i)_{1\le i\le m} = n^{-1/2}\left(\sum_{j=1}^{n} X_i^{(j)}\right)_{1\le i\le m}. \tag{5.5}$$

Note that the sample induces a different scaling on the mean of the alternatives: the joint distribution of the test statistics $S_i(X), 1 \le i \le m$, is now $\mathcal{N}(n^{1/2}\mu, \Gamma)$, instead of $\mathcal{N}(\mu, \Gamma)$ as above.

Now, we turn to find thresholds $\widehat{t}_{\mathcal{C}}$ satisfying (5.1) by using a randomization technic. Let us consider the algebraic group $\mathcal{G} = (\{-1, 1\}^n, \times)$ of "sign vector" acting on the vector X in the following way: for any sign vector $\varepsilon \in \{-1, 1\}^n$,

$$X^{\langle\varepsilon\rangle} = (\varepsilon_1 X^{(1)} \dots \varepsilon_n X^{(n)}) = \begin{pmatrix} \varepsilon_1 X_1^{(1)} & \dots & \varepsilon_n X_1^{(n)} \\ \vdots & & \vdots \\ \varepsilon_1 X_m^{(1)} & \dots & \varepsilon_n X_m^{(n)} \end{pmatrix}. \tag{5.6}$$

A key property to establish (5.1), called *the randomization hypothesis* in [115], is that the joint distribution of the test statistics under the null are invariant under the transformations of $\mathcal{G}$, that is, for any sign vector $\varepsilon \in \{-1, 1\}$,

$$\mathcal{D}\left((S_i(X^{\langle\varepsilon\rangle}))_{i\in\mathcal{H}_0}\right) = \mathcal{D}\left((S_i(X))_{i\in\mathcal{H}_0}\right). \tag{5.7}$$

Note that (5.7) is also true for any random sign vector. In the sequel, it will be appropriate to consider random sign vectors uniformly distributed on $\mathcal{G}$. Now, a remarkable fact is that, while (5.1) holds with $\widehat{t}_{\mathcal{C}} = q_\alpha(\mathcal{C}, Q_0)$ and $Q_0 = \mathcal{N}(0, \Gamma)$ from the previous section, (5.1) still holds when replacing the unknown distribution $Q_0 = \mathcal{N}(0, \Gamma)$ by the observable randomized distribution

$$\widehat{Q}(X) = \frac{1}{B+1}\left(\delta_{S(X)} + \sum_{b=1}^{B} \delta_{S(X^{\langle\varepsilon^{(b)}\rangle})}\right), \tag{5.8}$$

where $\varepsilon^{(1)}, \dots, \varepsilon^{(B)}$ are i.i.d. uniformly distributed on $\mathcal{G}$, for some $B \ge 1$.

Lemma 5.2. *Consider the threshold collection given by $\widehat{t}_{\mathcal{C}} = q_\alpha(\mathcal{C}, \widehat{Q}(X))$, where $\widehat{Q}(X)$ is given by (5.8) and $q_\alpha(\mathcal{C}, \cdot)$ by (5.3). Then the two requirements of RW's method (5.1) and (5.2) both hold.*

Proof. First note that (5.2) is trivial and let us prove (5.1). Denote $\psi(X) = \inf_{i\in\mathcal{H}_0(\theta)}\{p_i(X)\}$ for short. By definition of q_α, for any distribution Q, any t such that $t < q_\alpha(\mathcal{H}_0, Q)$ satisfies $Q((-\infty, t]) \le \alpha$. Hence, we have

$$\begin{aligned}\mathbb{P}\left(\psi(X) < q_\alpha(\mathcal{H}_0, \widehat{Q}(X))\right) &\le \mathbb{P}\left(1 + \sum_{b=1}^{B} \mathbf{1}\{\psi(X^{\langle\varepsilon^{(b)}\rangle}) \le \psi(X)\} \le \lfloor\alpha(B+1)\rfloor\right). \\ &\le \mathbb{P}\left(\sum_{b=0}^{B} \mathbf{1}\{Y_b \le Y_0\} \le \lfloor\alpha(B+1)\rfloor\right),\end{aligned} \tag{5.9}$$

where we let $Y_0 = \psi(X)$ and $Y_b = \psi(X^{\langle\varepsilon^{(b)}\rangle})$, $b = 1, \dots, B$. Now,

$$\begin{aligned}(Y_0, Y_1, \dots, Y_B) &= (\psi(X), \psi(X^{\langle\varepsilon^{(1)}\rangle}), \dots, \psi(X^{\langle\varepsilon^{(B)}\rangle})) \\ &\sim (\psi(X^{\langle\varepsilon^{(0)}\rangle}), \psi(X^{\langle\varepsilon^{(0)}\varepsilon^{(1)}\rangle}), \dots, \psi(X^{\langle\varepsilon^{(0)}\varepsilon^{(B)}\rangle})) \\ &\sim (\psi(X^{\langle\varepsilon^{(0)}\rangle}), \psi(X^{\langle\varepsilon^{(1)}\rangle}), \dots, \psi(X^{\langle\varepsilon^{(B)}\rangle})),\end{aligned}$$

by using (5.7), the group properties of $\mathcal{G}$, and by considering $\varepsilon^{(0)}$ uniformly distributed on $\mathcal{G}$ (and being independent of all the other variables). This entails that the vector $(Y_0, Y_1, \dots, Y_B)$ is exchangeable (but not independent in general). Thus, the random variable $\sum_{b=0}^{B} \mathbf{1}\{Y_b \le Y_0\}$ is the rank of Y_0 within the exchangeable sample $(Y_0, Y_1, \dots, Y_B)$ and is thus stochastically lower bounded[5] by a variable uniformly distributed on $\{0, \dots, B\}$. Hence, the probability (5.9) is smaller than or equal to $\lfloor \alpha(B+1) \rfloor / (B+1) \le \alpha$. □

Interestingly, combining Lemma 5.2 with Theorem 5.1 entails a new adaptive FWER control, with an easily computable thresholding collection, not relying on the knowledge of Γ. It is worth to note that, while the FWER control holds whatever B is, the achieved FWER is below $\lfloor \alpha(B+1) \rfloor / (B+1)$, so the control is inaccurate when B is too small. Typical choices are $B = 1,000$ or $B = 10,000$.

The new threshold collection $\widehat{t}_{\mathcal{C}} = q_\alpha(\mathcal{C}, \widehat{Q}(X))$ is more appropriate than the Holm threshold collection. For instance, if all the lines in (5.4) are equal (perfect equi-correlation), the threshold collection is

$$q_\alpha(\mathcal{C}, \widehat{Q}(X)) = \sup \left\{ t \in [0,1] \ : \ \frac{1}{B+1} \left(\mathbf{1}\{2\overline{\Phi}(|S_1(X)|) < t\} + \sum_{b=1}^{B} \mathbf{1}\{2\overline{\Phi}(|S_1(X^{\langle \varepsilon^{(b)} \rangle})|) < t\} \right) \le \alpha \right\},$$

which would correspond to the threshold of a single randomized test. More generally, simulations made in [P3] have shown that the new procedure is close to the oracle procedure of Section 5.1.2 in a particular setting with local dependencies. Finally, let us note that when the signal strength varies in a "wide dynamic range", this procedure can be accelerated by reducing the number of steps in RW's step-down method, see Section 6.1.4.

5.2 Adaptive FDP control [P9]

As explained in Chapter 1, using the FWER control might be too stringent for an exploratory research and a user might prefer to control the FDR (i.e., the mean of the FDP), or, more precisely, might prefer to ensure that the FDP is near or below the targeted level α with a large probability. The BH procedure presented in Algorithm 2.1 is widely used to that respect. However, we would like to emphasize the following trivial fact:

$$\text{FDR} = \mathbb{E}[\text{FDP}] \le \alpha \text{ does not provide FDP} \le \alpha \text{ or even FDP} \simeq \alpha \text{ (with high probability).}$$

In other words, FDR control does not imply FDP control. This is the starting point of our work [P9], from which we report some results below.

5.2.1 BH procedure and FDP control

Under weak dependence[6], it is well known that the BH procedure has an FDP that converges in probability to $\pi_0 \alpha$, where π_0 is the limit of the proportion of true nulls m_0/m (under mild conditions implying $\pi_0 < 1$), see, e.g., our Lemma 4.1 in [P9]. This tends to show that the BH procedure provides a correct control of the FDP when the dependence is weak.

However, as already noted in [86], the distribution of the FDP of BH procedure can be affected by strong dependence. We illustrate further this phenomenon in Figure 5.3. In the case

[5] Note that ties may appear into the vector $(Y_0, Y_1, \dots, Y_B)$. For instance, the same sign vector can be generated twice with a positive probability.

[6] The notion of weak dependence used here is simply the convergence in probability of the empirical distribution function of the p-values towards a deterministic distribution function on $[0,1]$.

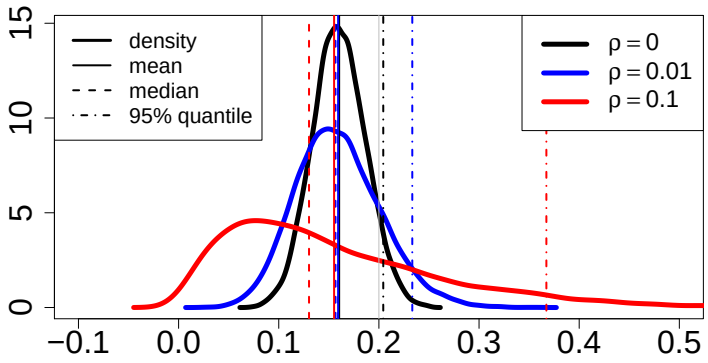


Figure 5.3: Fitted density of the FDP of BH procedure when increasing the dependence. $\alpha = 0.2$, $m = 1000$, $m_0 = 800$, 10^4 simulations, model (Gauss-ρ-equi).

(Gauss-ρ-equi) (with $\rho = 0.1$), the latter suggests that the BH procedure is much too optimistic for a 95%-quantile control of the FDP. This suggests that FDR control is not necessarily meaningful for the underlying FDP when the latter varies too much. This can lead to an erroneous interpretation of the quantity of false discoveries.

One way to solve this issue is to control the $(1-\zeta)$-quantile of the FDP distribution at level α, that is, to find a multiple testing procedure R such that for all $P \in \mathcal{P}$,

$$\mathbb{P}(\mathrm{FDP}(R,P) > \alpha) \leq \zeta. \tag{5.10}$$

Here, ζ is an extra parameter that has to be fixed by the user (small in practice). In this section, we will also consider the asymptotic FDP control:

$$\limsup_m \{\mathbb{P}(\mathrm{FDP}(R,P) > \alpha)\} \leq \zeta, \tag{5.11}$$

where $P = P^{(m)}$ lies in an appropriate sequence of models.

The FDP control has been introduced in [59, 103, 89] and it has received a considerable attention in the last decades, see for instance [111, 112, 70, 37, 27, 69] (among others).

5.2.2 Study of RW's heuristic

The heuristic proposed in [116], called RW's heuristic from now on, builds FDP controlling procedure from a family of k-FWER controlling procedures R_k, $1 \leq k \leq m$, by choosing[7] $\widehat{k}$ such that $(\widehat{k}-1)/|R_{\widehat{k}}| \leq \alpha$. In the context of step-wise procedures (see Section 2.5.1), this heuristic can be reformulated as follows: choose critical values τ_ℓ, $1 \leq \ell \leq m$, such that for all ℓ, we have that $V'_m(\tau_\ell) \geq \lfloor \alpha \ell \rfloor + 1$ with a probability less than ζ, where for $t \in [0,1]$,

$$V'_m(t) \succeq \sum_{i=1}^{m} (1-\theta_i)\mathbf{1}\{p_i(X) \leq t\}, \tag{5.12}$$

where the symbol $\succeq$ means stochastically larger (or equal). Above, RW's critical values depend on the distribution of the majoring process $V'_m(t)$. For instance, in the Gaussian case, we can build $V'_m(t)$

[7] In the original formulation, $\widehat{k}$ was constrained to be chosen such that any k' with $k' < \widehat{k}$ should also satisfy $(k'-1)/|R_{k'}| \leq \alpha$, which corresponds to a step-down approach. Here the step-up approach is also considered.

upon the distribution of the vector $\mathcal{N}(0,\Gamma)$ (i.e., under the "full null"). Other choices for $V'_m(t)$ can be done by assuming $m_0 = m - \ell + 1$ in (5.12), for instance under (Gauss-ρ-equi) (adaptive RW's critical values), see [P9] for more details.

The rationale behind RW's heuristic is that, by using a step-down or a step-up procedure R with such critical values and $\widehat{\ell} = |R|$ rejections, if the random variable $\widehat{\ell}$ was deterministic, we would have

$$\mathbb{P}(\mathrm{FDP}(R,P) > \alpha) = \mathbb{P}\left(|R \cap \mathcal{H}_0(\theta)| > \alpha\,\widehat{\ell}\right) \leq \mathbb{P}(V'_m(\tau_{\widehat{\ell}}) \geq \lfloor \alpha\widehat{\ell} \rfloor + 1) \leq \zeta, \tag{5.13}$$

because for all ℓ, the probability $\mathbb{P}(V'_m(\tau_\ell) \geq \lfloor \alpha\ell \rfloor + 1)$ is below ζ. Obviously, (5.13) is not rigorous because it neglects the fluctuations due to the randomness of $\widehat{\ell}$.

This heuristic has been theoretically justified (for the step-down form) in settings where the p-values under the null are independent of the p-values under the alternative (full independence in [70]; alternative p-values all equal to 0 in [116]). However, while the FDP is particularly interesting under dependence, these situations all rely on an independence assumption. Therefore, a first task of [P9] is to study the precise behavior of this method in "simple" dependent cases. We found the two following results:

- under weak-dependence[8], we have proved that RW's method controls the FDP asymptotically, in the sense of (5.11). The idea is that the reasoning behind RW's heuristic can be made rigorous in that case, because the fluctuations of $\widehat{\ell}/m$ asymptotically disappear. However, note that the interest of this result is of limited scope, because the simple BH procedure also asymptotically controls the FDP in that case.

- under (Gauss-ρ-equi), for the (adaptive) critical values derived from RW's heuristic, we have identified parameters (namely $\alpha = 0.1$; $\zeta = 0.05$; $\pi_0 = 0.9$; $\rho = 0.2$, $m \in \{200, 1000, 5000\}$) for which the probability that the FDP of SD(τ) exceeds α is (slightly but indubitably) above ζ. This annihilates any hope of finding a general finite sample proof of FDP control for RW's heuristic, even in the step-down case and for a very simple form of positive dependence.

5.2.3 New FDP controlling procedures under strong dependence

Finite sample control As the above counter-example shows, establishing (5.10) for RW's method is not possible in the finite sample case. Hence, this heuristic should be modified in order to be valid. Importantly, many existing results in FDP controlling methodology can be recast as modifications of RW's heuristic: the "diminution" principle starts with an upper bound on the exceedance probability and then reduces the critical values to make the bound below ζ, see [111, 112, 69]; the "augmentation" principle starts with an FWER (=1-FWER) controlling procedure R_1 at level ζ and then enlarges the rejection set in a way that maintains the FDP below α whenever R_1 makes no false discovery, see [139, 47]; the "simultaneous" k-FWE control is another strategy which takes RW's critical values but with ζ replaced by ζ/m, see [61].

Two new modifications have been proposed in our work, which incorporate part of the dependencies and that provably control the FDP for a fixed value of m. Also, in our simulations, the proposed procedures are shown to improve the state-of-the-art methods in most standard situations. While we refer the reader to the paper [P9] for more details, let us mention that there is still some loss when performing this modifications and thus there is overall a price to pay for getting rigorous *finite sample* FDP control.

[8] For this result, we need a functional central limit theorem for the empirical distribution function of the p-values.

Asymptotical control in a factor model Let us consider the relaxed FDP control (5.11) in a particular sequence of strongly dependent multiple testing models. For the sake of simplicity, we restrict our attention to models that can be written as follows:

$$X_i = \mu_i + c_i W + \zeta_i, \ \ 1 \le i \le m, \tag{facmodel}$$

where c_i, $1 \le i \le m$, are i.i.d., ζ_i, $1 \le i \le m$, are i.i.d., W is a random variable and $(c_i)_{1\le i\le m}$, $(\xi_i)_{1\le i\le m}$ and W are independent. In model (facmodel), we consider the (one-sided) testing problem

$$H_{0,i}\text{: ``}\mu_i = 0\text{'' against } H_{1,i}\text{: ``}\mu_i > 0\text{'' }, \ 1 \le i \le m,$$

and we assume that $\mu_i = \Delta\theta_i$ for $1 \le i \le m$, with $\Delta > 0$ and $(\theta_i)_{1\le i\le m} \in \{0,1\}^m$, for convenience. Hence, p-values can be built by taking $p_i(X) = \overline{F}(X_i)$ where $\overline{F}$ is the upper-tail distribution function of $c_1 W + \zeta_1$. Note that we implicitly assume here that the distributions of c_1, W and ζ_1 are known.

Model (facmodel) induces a strong dependence between the X_i's, which is carried by the factor W. The latter is classically referred to as "one factor model" in the literature (see, e.g., [87, 55, 44]). Here, the c_i's are unknown and taken random with a prescribed distribution. From an intuitive point of view, (facmodel) is modeling situations where some of the measurements X_i's have been deteriorated by unknown nuisance factors $c_i W$, $1 \le i \le m$. For instance, we can simultaneously deteriorate the measurements of some unknown (random) sub-group of $\{1, \dots, m\}$ by taking

$$X_i = \mu_i + \rho^{1/2}\delta_i W + (1-\rho)^{1/2}\xi_i, \ \ 1 \le i \le m, \tag{partial-Gauss-ρ-equi}$$

where $\rho \in [0,1]$ is a parameter, where W and the ξ_i's are all i.i.d. $\mathcal{N}(0,1)$ and independent of the δ_i's i.i.d. with common distribution $\mathcal{B}(a)$, $a \in [0,1]$. Obviously, model (Gauss-ρ-equi) can be recovered by taking $a = 1$ in (partial-Gauss-ρ-equi).

Now, in (facmodel), we can modify RW's heuristic in an asymptotical manner, which leads to consider critical values satisfying

$$F_0(\tau_\ell, q_\zeta) = \alpha\ell/m, \quad 1 \le \ell \le m, \quad \text{with } q_\zeta \text{ s.t. } \mathbb{P}(W \ge q_\zeta) \le \zeta, \tag{5.14}$$

where $F_0(t,w)$ is the distribution function of $p_i(X)$ conditionally on $W = w$ (under the null). The rationale behind this is that, by the law of large number (taken conditionally on W), the probability $\mathbb{P}(V_m(\tau_\ell) > \alpha\ell)$ is asymptotically "close" to $\mathbb{P}(F_0(\tau_\ell, W) > \alpha\ell/m) = \mathbb{P}(F_0(\tau_\ell, W) > F_0(\tau_\ell, q_\zeta)) \le \mathbb{P}(W \ge q_\zeta) \le \zeta$, if we assume that $F_0(t,\cdot)$ is increasing.

The following result states that this asymptotic view of RW's heuristic entails a rigorous asymptotic FDP control under appropriate assumptions.

Theorem 5.3. *Consider model* (facmodel) *with* m_0/m *tending to* $\pi_0 \in (0,1)$ *and additionally assume:*

(*i*) c_1 *is a random variable with a finite support in* $\mathbb{R}^+$*;*

(*ii*) *the distribution function of* W *is continuous;*

(*iii*) *the function* $x \in \mathbb{R} \mapsto \overline{F}_\xi(x) = \mathbb{P}(\xi_1 \ge x)$ *is continuous increasing and is such that, for all* $y \in \mathbb{R}$, *as* $x \to +\infty$, $\overline{F}_\xi(x-y)/\overline{F}_\xi(x)$ *tends to* $+\infty$ *if* $y > 0$ *and to* 0 *if* $y < 0$.

Consider the step-up procedure associated to the critical values τ_ℓ, $\ell = 1, \dots, m$ *satisfying* (5.14). *Then the asymptotic FDP control* (5.11) *holds for* $R = SU(\tau)$.

In Theorem 5.3, (i) can typically be interpreted as a positive dependence assumption. As for (iii), it is satisfied for several distributions, for instance when ξ_1 has a Subbotin density (including the Gaussian case), defined by (6.24) further on.

The proof of Theorem 5.3 relies on a very simple idea: if $\tau_{\hat{\ell}}$ converges to some (non-deterministic) random variable T, the exceedance probability $\mathbb{P}(V_m(\tau_{\hat{\ell}}) > \alpha\hat{\ell})$ should be asymptotically close to $\mathbb{P}(F_0(T, W) > F_0(T, q_\zeta))$. Now, since $F_0(t, \cdot)$ is increasing, the latter is below $\mathbb{P}(W \geq q_\zeta) \leq \zeta$.

Let us now apply Theorem 5.3 in model (partial-Gauss-ρ-equi) for $\rho \in [0, 1)$. This gives rise to a new step-up FDP controlling procedure with critical values given by the equation $(1-a)f(\tau_\ell, 0) + af(\tau_\ell, q_\zeta) = \alpha\ell/m$, $1 \leq \ell \leq m$, where f is given by (2.11) and $q_\zeta = \overline{\Phi}^{-1}(\zeta)$. In particular, in model (Gauss-ρ-equi), this gives the explicit new critical values

$$\tau_\ell = \overline{\Phi}\left(\rho^{1/2}\overline{\Phi}^{-1}(\zeta) + (1-\rho)^{1/2}\overline{\Phi}^{-1}(\alpha\ell/m)\right), \quad 1 \leq \ell \leq m. \tag{5.15}$$

Note that the latter reduce to the BH critical values when $\rho = 0$, but are markedly different when $\rho > 0$, as illustrated on Figure 5.4.

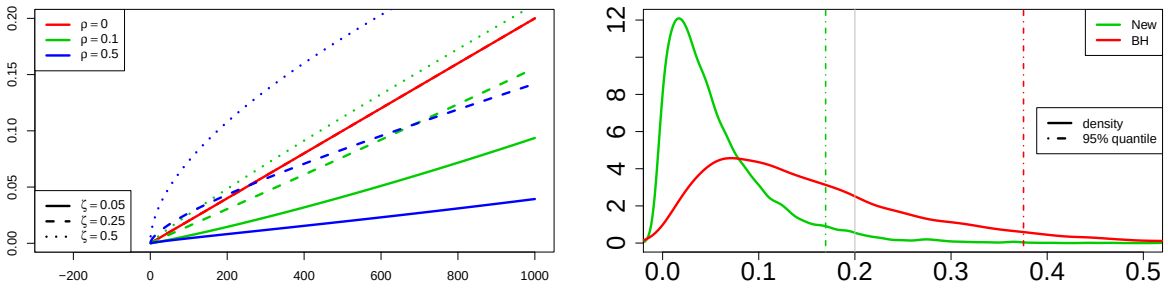


Figure 5.4: Left: plot of the critical values (5.15) in function of ℓ. Right: same as Figure 5.3 but only for $\rho = 0.1$ and by adding the new step-up procedure using (5.15).

5.2.4 Qualitative comparison with FWER and FDR controls

To give further intuition behind Theorem 5.3, let us now provide a qualitative comparison of FWER, FDR and FDP controls under strong dependence by coming back to the two-dimensional representation scheme used in Figure 1.2 of Chapter 1. While the signal is carried by a disk (with a signal stronger at the center of the disk), the errors follow the factor model (partial-Gauss-ρ-equi). Hence, each realization corresponds to a specific value of the factor W. Figure 5.5 displays[9] the observed discoveries of several procedures: the (oracle and single step) FWER controlling procedure of Section 5.1.1, the FDR controlling BH procedure and the new FDP controlling procedure of Theorem 5.3. First note that the procedures are particularly sensitive to the sign of W when $|W|$ is large: when $W = -2.1$, all the procedures look conservative. By contrast, the case $W = 1.6$ allows many discoveries but also entails situations where the quantity of false discoveries is "too large". Next, the picture seems in accordance with the respective roles of FWER, FDR and FDP controls: the FWER controlling procedure ensures no false discovery (except for one realization); the FDR controlling procedure yields FDP values with an empirical mean close to α, but it does not prevent the FDP from taking values exceeding α, which can arise when W is a bit large. By contrast, the $(1-\zeta)$-FDP quantile controlling procedure is more cautious: from an intuitive point of view, it is

[9] Since there are only 6×2 realizations here, this picture is only illustrative and the conclusion is only qualitative.

"prepared" to face the fluctuations of W in the range $(-\infty, q_\zeta)$. This means that, roughly, the only realizations for which the FDP can exceed α arise when $W \geq q_\zeta$.

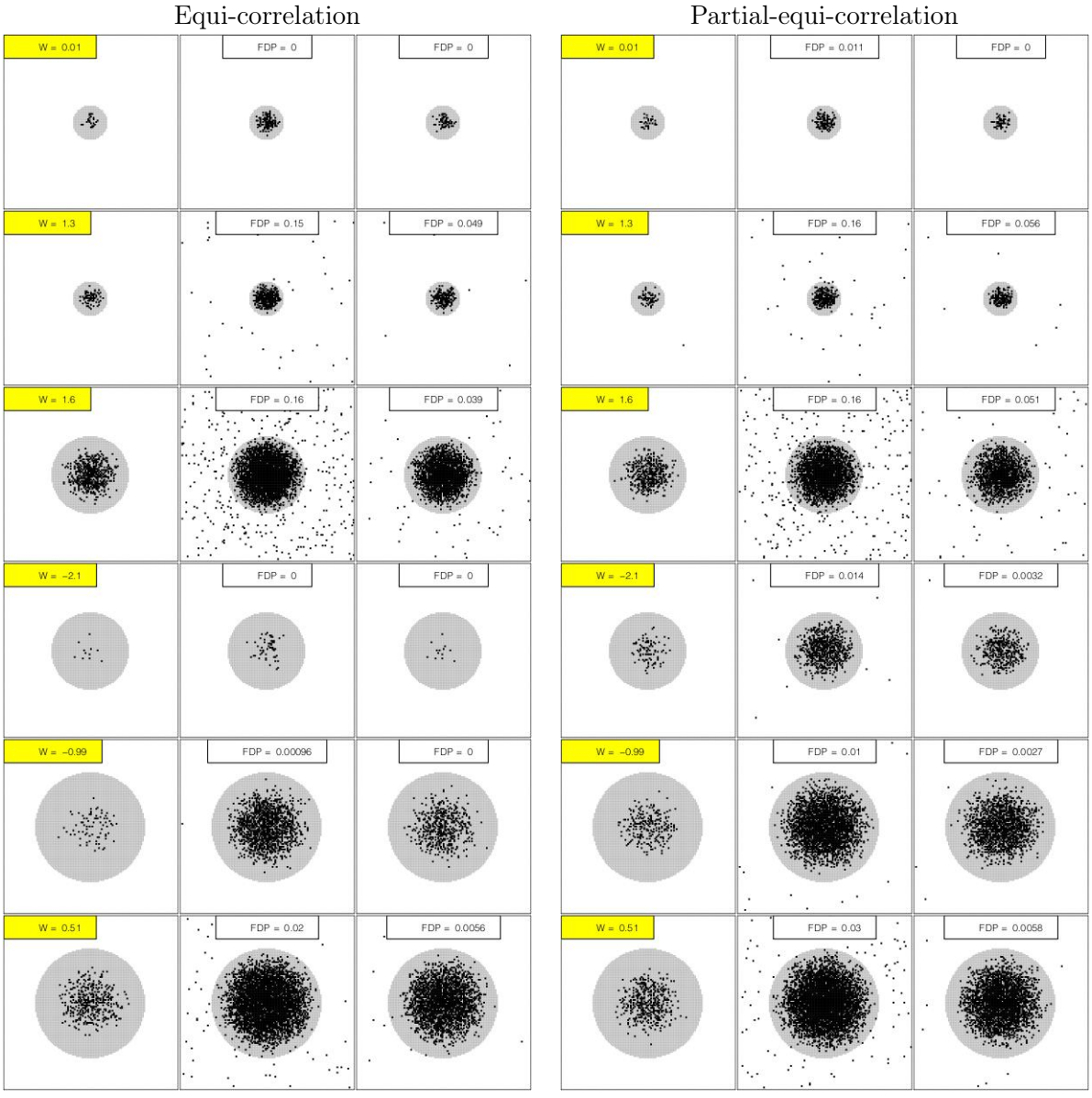


Figure 5.5: For each model (Gauss-ρ-equi) or (partial-Gauss-ρ-equi) (with $a = 1/2$): discoveries on 6 realizations for FWER (left), FDR (middle) or 0.95-FDP quantile (right) at level $\alpha = 0.05$ in a two-dimensional setting where $m = 128^2$ items are tested. W is the factor of (facmodel) (same value along each line) and $\rho = 0.3$, see text.

5.3 Adaptation to a clustered dependence structure [P11]

At the opposite side of model (facmodel), let us consider the case where strong local dependencies appear between the tests. In that case, we might ask whether testing groups of items is not more appropriate than testing the individual items. This raises the delicate issue of choosing an appropriate testing resolution, not necessarily related to the precision of the underlying measuring device.

5.3.1 Motivation

This problem is motivated by the analysis of CNA data (described in Section 1.3) and more specifically the analysis of the data of [28], concerning breast cancer with 96 ER-positive individuals and 49 ER-negative individuals. The goal is to identify the parts of the genome for which the chromosomal aberration due to the breast cancer is significantly different between the two samples. The data have been discretized[10] via standard pre-processing steps (segmentation and calling, see [137]) which give an array with -1 (deletion), 0 (normal) or 1 (gain). Finally, we reduce the dimension of the array with the method of [138] that collapses the (essentially) equal probe profiles.

We obtained $m = 383$ "regions" of the genome, see Figure 5.6 (b). We consider the latter as our input data. In particular, the region level is the basic testing resolution that we will consider. Markedly, even with the collapsing, Figure 5.6 (a) shows that the local dependency between the regions can still be high. This suggests that the regions are maybe not the only relevant items to test. A cluster family, as the one appearing on the left part of Figure 5.6 (b), might also be worth to examine.

This is a motivation for combining clustering and testing. The latter task has been only scarcely studied in the literature, at least from a theoretical point of view: in [9], the clusters are inferred from independent experiments, which is an assumption that we do not want to consider here. In [103], the clustering is made on the basis of the p-values, which inter-relates the testing and clustering phases and renders meaningful clusters from a testing perspective only. Another option is sample splitting. However, the potential loss of power is huge, while the output can depend substantially on the splitting.

More generally, the problem emerging behind making testing on clusters coming from the same data set is the validity of testing random null hypotheses $H_{0,i}(X)$. Note that even if the approach is theoretically valid, the final interpretation of the test is also an issue. The main idea of our work is to use independent parts of the data X to make the two phases: while the clustering will use the CGH array without the group labels and will be permutation invariant, the testing will be made by using the group labels and permutation techniques.

5.3.2 Model and p-values

Let us model the CGH array and the phenotype information by an i.i.d. sample

$$X = (X^{(1)}, \dots, X^{(n)}) \quad \text{where } X^{(1)} = (Z^{(1)}, Y^{(1)}) \in \{-1, 0, 1\}^m \times \{1, 2\}.$$

Here, while $Z_i^{(j)} \in \{-1, 0, 1\}$ codes for the chromosomal aberration status (deletion, normal or gain, respectively) for region i and individual j, $Y^{(j)}$ denotes the group label of individual j (say, 1 for ER-positive, 2 for ER-negative). The $m \times n$ array formed by the $Z_i^{(j)}$'s is denoted by Z. It does not contain the label information.

We assume that there exists a clustering method, only using Z (and not Y), which formally corresponds to a family $\mathcal{A}(Z) = \{A_d(Z), 1 \le d \le D(Z)\}$ of non-empty and contiguous parts of $\{1, \dots, m\}$

[10] While the raw data are continuous, we choose here to pay the variability of the discretization in order to use data under a form better reflecting the underlying nature of the considered biological process.

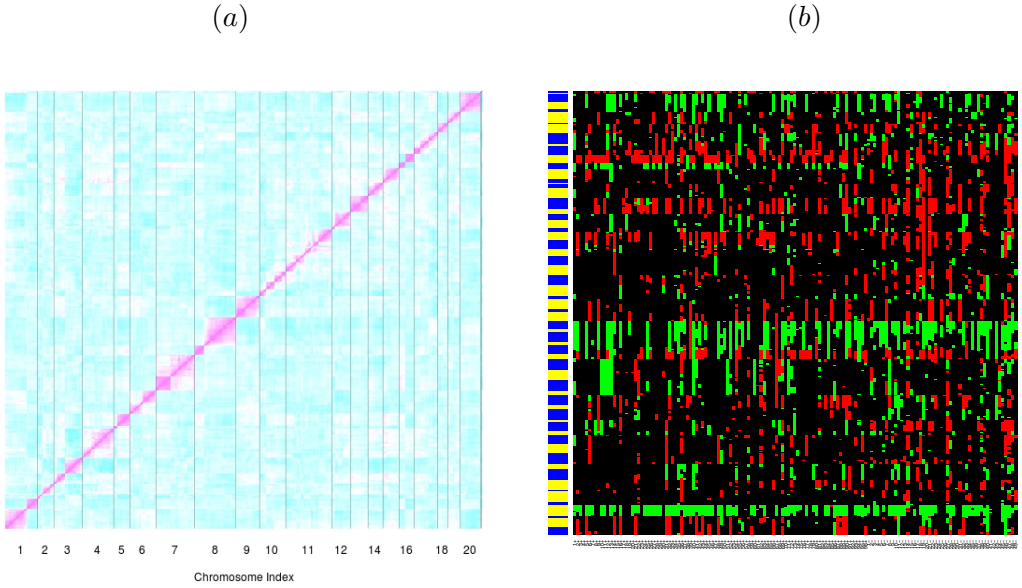


Figure 5.6: Left: Kendall's τ heatmap for the (transformed) data of [28], see text. Regions are plotted according to the chromosomal position; colors represent correlations from -1 (cyan) to 1 (pink). Right: status -1 (red), 0 (black) or 1 (green) of the individuals (X-axis) in function of the genome region (Y-axis), as resulting from the pre-processing steps (see text). The regions are ordered according to chromosomal position from bottom to top. The clustering at the most-left part of the right display (depicted alternately in blue and yellow) comes from a clustering algorithm via a log linear model, see [P11].

that form a partition of $\{1,\dots,m\}$. From a technical point of view, we should also assume that $D(Z)$ and $\mathbf{1}\{i \in A_d(Z)\}$ are measurable for all i, d. Now, let us make the following assumption, which will be crucial in the sequel:

$$\begin{gathered}\mathcal{A}(\cdot) \text{ is permutation invariant, that is, for any realization of the array } Z,\\ \text{for any permutation } \sigma \text{ of } \{1,\dots,n\},\ \mathcal{A}(Z^\sigma) = \mathcal{A}(Z),\end{gathered} \tag{PermutInv}$$

where Z^σ denotes the matrix formed by the $Z_i^{\sigma(j)}$'s, that is, the matrix Z in which the permutation σ has been applied to the columns. A clustering method satisfying (PermutInv) has been proposed in [P11], by using a log linear model on the distribution $Z^{(1)}$ and a likelihood maximization algorithm. More generally, any (reasonable) model-based clustering will satisfies (PermutInv) because it will rely on the fact that $(Z^{(1)},\dots,Z^{(n)})$ is an i.i.d. sample.

Conditionally on $\mathcal{A}(Z)$, we consider the problem of testing the independence between the chromosomal aberration and the group label. According to Section 5.3.1, we would like to make testing both on the region and cluster levels. Hence, p-values should be defined for both levels. We consider the following null hypotheses: for $i \in \{1,\dots,m\}$ and $A \in \mathcal{A}(Z)$,

$$H'_{0,i} : \text{“}Z_i^{(1)} \text{ is independent of } Y^{(1)} \text{ conditionally on } \mathcal{A}(Z)\text{”}; \tag{5.16}$$

$$H'_{0,A} : \text{“}\left(Z_i^{(1)}\right)_{i\in A} \text{ is independent of } Y^{(1)} \text{ conditionally on } \mathcal{A}(Z)\text{”}. \tag{5.17}$$

Let us consider some test statistic $S(Z_i, Y)$ for testing the individual regions (5.16), for instance a standard χ^2 statistic. Appropriate p-values can be built by generating random i.i.d. uniform permu-

tations $\sigma_1, \dots, \sigma_B$ of the labels, by following an approach similar to Section 5.1.3. More precisely, we let for $i \in \{1, \dots, m\}$ and $A \in \mathcal{A}(Z)$,

$$p_i(X) = (B+1)^{-1} \left(1 + \sum_{b=1}^{B} \mathbf{1}\{S(Z_i, Y^{\sigma_b}) \geq S(Z_i, Y)\} \right); \tag{5.18}$$

$$p_A(X) = (B+1)^{-1} \left(1 + \sum_{b=1}^{B} \mathbf{1}\{\max_{i \in A}\{S(Z_i, Y^{\sigma_b})\} \geq \max_{i \in A}\{S(Z_i, Y)\}\} \right), \tag{5.19}$$

respectively[11], where Y^{σ_b} denotes $(Y^{(\sigma_b(j))})_{1 \leq j \leq n}$, that is, the label vector whose components have been permuted according to σ_b. Then, similarly to Lemma 5.2, we have the following result.

Lemma 5.4. *Let us consider the above model with a clustering method $\mathcal{A}(Z)$ satisfying (PermutInv). Consider the p-values $p_i(X)$ (5.18) testing $H'_{0,i}$ (5.16), $1 \leq i \leq m$, and the p-values $p_A(X)$ (5.19) testing $H'_{0,A}$ (5.17), $A \in \mathcal{A}(Z)$, for $B \geq 1$ i.i.d. (and uniformly distributed) permutations $\sigma_1, \dots, \sigma_B$ of $\{1, \dots, m\}$. Then, for any distribution of $X^{(1)} = (Z^{(1)}, Y^{(1)})$, we have*

(i) for $i \in \{1, \dots, m\}$ such that $H'_{0,i}$ holds, we have $\mathbb{P}(p_i(X) \leq t \,|\, \mathcal{A}(Z)) \leq t$ for all $t \in [0,1]$;

(ii) for $A \in \mathcal{A}(Z)$ such that $H'_{0,A}$ holds, we have $\mathbb{P}(p_A(X) \leq t \,|\, \mathcal{A}(Z)) \leq t$ for all $t \in [0,1]$.

In other words, Lemma 5.4 states that the fundamental property (pvalueprop) is satisfied both on the region and cluster levels. The proof is a variation of the proof of Lemma 5.2 that appropriately combines (PermutInv) with the conditioning.

Proof. Let us prove (ii) (the proof for (i) is similar). Let $\mathcal{A}$ be any realization of $\mathcal{A}(Z)$ and consider any $A \in \mathcal{A}$. Let $\psi(Z_A, Y) = \max_{i \in A}\{S(Z_i, Y)\}$ where $Z_A = (Z_i^{(j)})_{i \in A, 1 \leq j \leq n}$. From (5.19) and following the proof of Lemma 5.2, it is sufficient to prove that the vector $(\psi(Z_A, Y), \psi(Z_A, Y^{\sigma_1}), \dots, \psi(Z_A, Y^{\sigma_B}))$ is exchangeable conditionally on $\mathcal{A}(Z) = \mathcal{A}$. For this, let us note that for any deterministic permutation σ, we have under $H'_{0,A}$,

$$\begin{aligned}
\mathcal{D}((Z_A, Y) \,|\, \mathcal{A}(Z) = \mathcal{A}) &= \mathcal{D}(Z_A \,|\, \mathcal{A}(Z) = \mathcal{A}) \otimes \mathcal{D}(Y \,|\, \mathcal{A}(Z) = \mathcal{A}) \\
&= \mathcal{D}(Z_A \,|\, \mathcal{A}(Z) = \mathcal{A}) \otimes \mathcal{D}(Y^{\sigma} \,|\, \mathcal{A}(Z^{\sigma}) = \mathcal{A}) \\
&= \mathcal{D}(Z_A \,|\, \mathcal{A}(Z) = \mathcal{A}) \otimes \mathcal{D}(Y^{\sigma} \,|\, \mathcal{A}(Z) = \mathcal{A}) \\
&= \mathcal{D}((Z_A, Y^{\sigma}) \,|\, \mathcal{A}(Z) = \mathcal{A}),
\end{aligned}$$

where we used (PermutInv). This entails that, if σ_0 is a random permutation (uniformly distributed and independent of the remaining variables), conditionally on $\mathcal{A}(Z) = \mathcal{A}$, we have

$$\begin{aligned}
(\psi(Z_A, Y), \psi(Z_A, Y^{\sigma_1}), \dots, \psi(Z_A, Y^{\sigma_B})) &\sim (\psi(Z_A, Y^{\sigma_0}), \psi(Z_A, Y^{\sigma_1 \circ \sigma_0}), \dots, \psi(Z_A, Y^{\sigma_B \circ \sigma_0})) \\
&\sim (\psi(Z_A, Y^{\sigma_0}), \psi(Z_A, Y^{\sigma_1}), \dots, \psi(Z_A, Y^{\sigma_B})),
\end{aligned}$$

because $(\sigma_0, \sigma_1 \circ \sigma_0, \dots, \sigma_B \circ \sigma_0) \sim (\sigma_0, \sigma_1, \dots, \sigma_B)$. This implies the result. □

[11] More precisely, in [P11], the test chosen for cluster A rejects $H'_{0,A}$ if $\min_{i \in A}\{p_i(X)\}$ is small enough. Here, we choose (5.19) for clarity.

5.3.3 Method

Now that we have at hand two resolutions of tests, how should we combine the corresponding p-values? Again, we face the problem of choosing a convenient global error rate. In this multi-resolution setting, an overall FDR or FDP control over the two levels seems not appropriate, because it would consider equally a false positive coming from a cluster and a false positive coming from a region. Also, whithin the cluster level, the FDP device might be too sophisticated, because the clustering stage is expected to generate only few clusters. Finally, rejecting a region and not the cluster containing it would make the final decision somewhat difficult to interpret, which tends to show the presence of an underlying hierarchy in the desired decision.

To solve this issue, we consider the problem of controlling a "joint" family-wise error rate (FWER), both on the cluster and region levels. Hence, in this setting, a multiple testing procedure is a family $\mathcal{R}$ of rejected items, each item being either a region $i \in \{1, \dots, m\}$ or a cluster $A \in \mathcal{A}(Z)$. The FWER of $\mathcal{R}$ can then be rewritten as

$$\mathrm{FWER}(\mathcal{R}) = \mathbb{P}\left(\left\{ \exists i \in \mathcal{R} \ : \ H'_{0,i} \text{ is true} \right\} \cup \left\{ \exists A \in \mathcal{R} \ : \ H'_{0,A} \text{ is true} \right\} \ \middle| \ \mathcal{A}(Z) \right), \tag{5.20}$$

where $H'_{0,i}$ and $H'_{0,A}$ are defined by (5.16) and (5.17), respectively. By denoting $\mathcal{H}_0$ the set of true nulls for regions and clusters, that is,

$$\mathcal{H}_0 = \left\{ i \in \{1, \dots, m\} \ : \ H'_{0,i} \text{ is true} \right\} \cup \left\{ A \in \mathcal{A}(Z) \ : \ H'_{0,A} \text{ is true} \right\},$$

the quantity $\mathrm{FWER}(\mathcal{R})$ can be rewritten as $\mathbb{P}(|\mathcal{R} \cap \mathcal{H}_0| \geq 1 \mid \mathcal{A}(Z))$, as in the initial definition of Chapter 2. To control (5.20) at level α, we follow the method proposed in [64] (itself generalizing an approach of [93]), that allows to produce an hierarchical testing of null hypotheses following a *tree structure*. Although the tree structure is given *a priori* in [64], their argument can be generalized to our context of "cluster→region" structure, because the p-value property (pvalueprop) is true conditionally on the clusters, thanks to Lemma 5.4. The seminal idea is the very general "sequential rejection principle" of [63], which provides sufficient conditions under which an iterative procedure controls the FWER[12]. In our two-level context, this gives rise to the following algorithm, depending on thresholding functions $\mathcal{T}_A(\mathcal{R})$ and $\mathcal{T}_i(\mathcal{R})$ to be chosen later.

Algorithm 5.5. - Step 0: compute the p-values given by (5.18) and (5.19) and set $\mathcal{R}_0 = \emptyset$;

- Step k ($k \geq 1$): compute $\mathcal{T}_A(\mathcal{R}_{k-1})$ and $\mathcal{T}_i(\mathcal{R}_{k-1})$ for $1 \leq i \leq m$, $A \in \mathcal{A}(Z)$ and put

$$\mathcal{R}_k = \{A \in \mathcal{A}(Z) \ : \ p_A(X) \leq \mathcal{T}_A(\mathcal{R}_{k-1})\} \cup \{1 \leq i \leq m \ : \ p_i(X) \leq \mathcal{T}_i(\mathcal{R}_{k-1})\}.$$

If $\mathcal{R}_k = \mathcal{R}_{k-1}$, stop and reject $\mathcal{R} = \mathcal{R}_{k-1}$. Otherwise, go to step $k+1$.

Now, Theorem 1 of [63] provides two sufficient conditions on $\mathcal{T}_A(\mathcal{R})$ and $\mathcal{T}_i(\mathcal{R})$ to ensure that the above procedure controls the FWER (5.20) at level α. The first condition is that $\mathcal{T}_A(\mathcal{R})$ and $\mathcal{T}_i(\mathcal{R})$ are nondecreasing when the set $\mathcal{R}$ grows. The second condition is the Bonferroni-Shaffer inequality:

$$\sum_{A \in \mathcal{A}(Z)} \mathcal{T}_A(\mathcal{H}_0^c) + \sum_{A \in \mathcal{A}(Z)} \sum_{i \in A} \mathcal{T}_i(\mathcal{H}_0^c) \leq \alpha. \tag{5.21}$$

While several threshold choices are possible for satisfying (5.21), it seems appropriate to weigh the clusters equally, because small clusters might be as relevant as large ones in our application. We propose the following thresholds:

$$\begin{cases} \mathcal{T}_A(\mathcal{R}) = \frac{\alpha}{D(Z) - |\{A' \in \mathcal{A}(Z)\} \,:\, A' \subset \mathcal{R}\}|} & \text{for all } A \in \mathcal{A}(Z) \\ \mathcal{T}_i(\mathcal{R}) = \frac{\mathcal{T}_A(\mathcal{R}) \mathbf{1}\{A \in \mathcal{R}\}}{|A| - |\{i' \in A \,:\, i' \in \mathcal{R}\}|} & \text{for all } i \in \{1, \dots, m\}, \end{cases} \tag{5.22}$$

[12] As a matter of fact, this principle generalizes the step-down argument stated in Theorem 5.1.

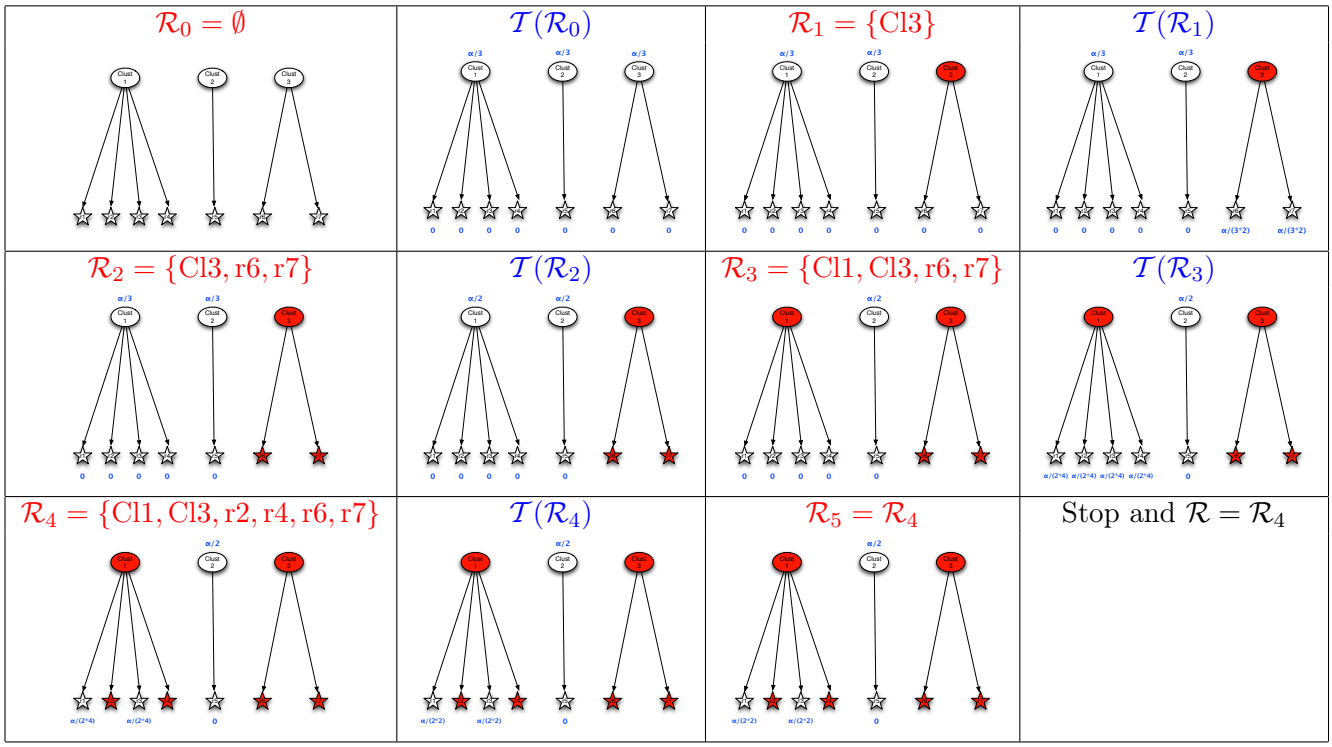


Figure 5.7: Toy example illustrating Algorithm 5.5 with the thresholding functions (5.22).

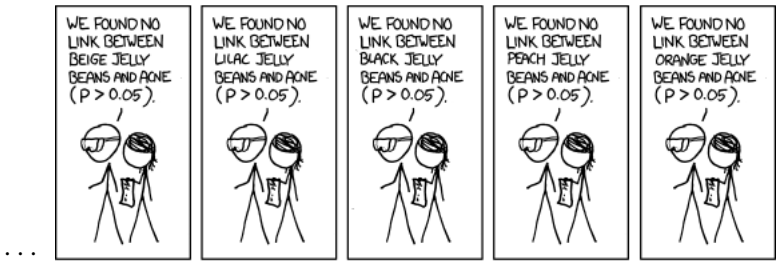
with the conventions $0/0 = 0$ and $1/0 = +\infty$ (also remember that $D(Z) = |\mathcal{A}(Z)|$ is the number of clusters).

Overall, by combining Lemma 5.4 with Theorem 1 of [63], we obtain the following result.

Theorem 5.6. *Let us consider the model of Section 5.3.2 with a clustering method $\mathcal{A}(Z)$ satisfying* (PermutInv)*. Consider the multiple testing procedure $\mathcal{R}$ defined by Algorithm 5.5, used with the thresholding functions* (5.22)*. Then the (conditional) FWER of $\mathcal{R}$ given by* (5.20) *is smaller or equal to α.*

This method is illustrated in Figure 5.7 with a "toy" hierachical structure (3 clusters and 7 regions). Note that for a trivial clustering $\mathcal{A}(Z) = \{\{i\}, 1 \leq i \leq m\}$, it simply reduces to Holm's procedure. For a general $\mathcal{A}(Z)$, the procedure proposes another detection level with the clusters, which is a substantial advantage. Typically, even if no region is significant (which can be the case when the number of replicates is small), Algorithm 5.5 can still detect something at the cluster level. A counterpart is that we should test conditional null hypotheses, which leads to a different interpretation than the traditional unconditional testing. Namely, an interpretation can be that, when repeating the experiments, among experiments that lead to the same observed clustering $\mathcal{A}(Z)$, no cluster/region under the null will be rejected by $\mathcal{R}$ with probability at least $1 - \alpha$. Hence, an important issue is to choose a "stable" clustering in the first round, for instance, a clustering (essentially) invariant when a small part of the individuals is removed. We refer the reader to [P11] for more details on these issues and for outputs of the method on real-data sets.

Finally, let us mention that this hierarchical testing (combined with the clustering of Figure 5.6 (b)) has been implemented in the *R*-package "dnaCplusT", by Kyung In Kim and Mark A. van de Wiel (it is available on the homepage of Mark A. van de Wiel).



Chapter 6

Connections with other statistical issues

This chapter aims at building bridges between multiple testing theory and other statistical problems. We focus here on confidence regions, functional central limit theorems and classification.

6.1 Multivariate confidence regions [P2, P3]

In this section, we first emphasize that, while some links exist between controlling the FWER and building confidence regions, the two tasks are not equivalent and building confidence regions is a more general aim, which has an interest in its own right. Then, we present a result of [P2] which provides confidence regions for the mean of a Gaussian multivariate vector whose dependence structure is unknown. Here, the regions are adaptive, which means that they learn implicitly the dependence structure. To this end, we use randomization techniques close to the spirit of the bootstrap and that turn out to be the same sign-flipping operation as in Section 5.1.3, except that the data are *empirically recentered*. Finally, we discuss the performance of the FWER controlling procedures resulting from these confidence regions.

6.1.1 Confidence regions and FWER control: same goal?

For a single null hypothesis, it is well known that single testing is intrinsically related to confidence regions. In the case of a multiple inference, similar correspondences can be outlined by using the FWER criterion.

For the sake of simplicity, let us consider a two-sided multiple testing problem where we test $H_{0,i}(\mu^0)$: "$\mu_i = \mu_i^0$" against $H_{1,i}(\mu^0)$: "$\mu_i \neq \mu_i^0$" , $1 \leq i \leq m$, for the mean μ of a statistic $T(X) \in \mathbb{R}^m$, and for some arbitrary reference vector $\mu^0 \in \mathbb{R}^m$. The following result is adapted from Theorem 1.1 of [31]:

Proposition 6.1. *The following correspondence holds between a family of FWER controlling procedures and a multivariate confidence region:*

(i) Let $\{R(\mu^0), \mu^0 \in \mathbb{R}^m\}$ be a family of multiple testing procedures such that for any $\mu^0 \in \mathbb{R}^m$, the FWER of $R(\mu^0)$ is controlled at level α (for testing the nulls $H_{0,i}(\mu^0)$ against $H_{1,i}(\mu^0)$, $1 \leq i \leq m$). Then the set

$$C(X) = \{z \in \mathbb{R}^m \ : \ R(z) = \emptyset\} \tag{6.1}$$

is a $(1-\alpha)$-confidence region for μ.

(ii) Conversely, let $C(X)$ be a $(1-\alpha)$-confidence region for μ. Then, for all $\mu^0 \in \mathbb{R}^m$, the procedure

$$R(\mu^0) = \{i \in \{1,\dots,m\} \,:\, \{z \in \mathbb{R}^m : z_i = \mu_i^0\} \cap C(X) = \emptyset\} \tag{6.2}$$

has an FWER controlled at level α (for testing the nulls $H_{0,i}(\mu^0)$ against $H_{1,i}(\mu^0)$, $1 \le i \le m$).

Proof. For (i), we have for all $\mu^0 \in \mathbb{R}^m$, $\mathbb{P}_{\mu=\mu^0}(\mu^0 \notin C(X)) = \mathbb{P}_{\mu=\mu^0}(R(\mu^0) \neq \emptyset) \le \alpha$. For (ii), we have for all $\mu^0 \in \mathbb{R}^m$ and $\mu \in \mathbb{R}^m$, $\mathbb{P}_\mu(\exists i \in R(\mu^0) \text{ s.t. } \mu_i = \mu_i^0) \le \mathbb{P}_\mu(\mu \notin C(X)) \le \alpha$. □

With words, item (i) states that a confidence region can be built upon the fact that, with probability at least $1-\alpha$, none of the nulls $H_{0,i}(\mu)$, $1 \le i \le m$, are rejected by the FWER controlling procedure $R(\mu)$. Conversely, item (ii) means that, in order to control the FWER, we can reject the null $H_{0,i}(\mu^0)$ for each i such that all the vectors $z \in \mathbb{R}^m$ with $z_i = \mu_i^0$ fall outside the confidence region $C(X)$.

As a simple illustration of Proposition 6.1, given a test statistic $T(X) \in \mathbb{R}^m$, there is a one to one correspondence between a confidence region of the form $C(X) = \{z \in \mathbb{R}^m \,:\, \|T(X) - z\|_\infty \le r(X)\}$ and the family of multiple testing procedure $R(\mu^0) = \{1 \le i \le m \,:\, |T(X)_i - \mu_i^0| > r(X)\}$. However, in general, the correspondence outlined in Proposition 6.1 is not one to one (this can be easily seen by considering $m = 2$ and $\|\cdot\|_2$ instead of $\|\cdot\|_\infty$ in the example above).

In addition, in general, building a confidence region on the basis of (6.1) requires to test each point z of $\mathbb{R}^m$, which is intractable (even with a discretization when m is large). Hence, except in some specific situations, multivariate confidence regions do not come directly from FWER controlling procedures.

Finally, while Proposition 6.1 involves *single-step* FWER controlling procedures, RW's method described in Section 5.1.1 shows that a *step-down* FWER controlling procedure can be deduced from a family of $(1-\alpha)$-confidence regions of the form:

$$C(X, \mathcal{C}) = \left\{ z \in \mathbb{R}^m \,:\, \sup_{i \in \mathcal{C}} |T(X)_i - z_i| \le r(X, \mathcal{C}) \right\}, \ \mathcal{C} \subset \{1,\dots,m\}, \tag{6.3}$$

where $r(X, \mathcal{C})$ is non-decreasing w.r.t. $\mathcal{C}$. For this, $\{\widehat{t}_{\mathcal{C}}, \mathcal{C} \subset \{1,\dots,m\}\}$ should be chosen according[1] to $r(X, \mathcal{C})$ and we easily check that (5.1) and (5.2) both hold, so that an algorithm similar to the one of Theorem 5.1 can be employed.

6.1.2 Oracle confidence regions

Consider the n-sample Gaussian multivariate setting defined in Section 5.1.3 (and the notation therein), see (5.4). We search a confidence region for the multivariate parameter $\mu \in \mathbb{R}^m$ of the following form:

$$C(X) = \left\{ z \in \mathbb{R}^m \,:\, \|\sqrt{n}(\overline{X} - z)\|_q \le r(X) \right\}, \tag{6.4}$$

where $q \in [1, \infty]$ and $r(X)$ is some possibly data-dependent threshold. Sometimes, for convenience, (6.4) will be rewritten as the set of z such that $\|S(X - z)\|_q \le r(X)$, where $S(\cdot)$ is the "empirical mean operator" defined by (5.5) and $X - z$ denotes the recentered sample $(X^{(1)} - z, \dots, X^{(n)} - z)$. The goal is thus to find $r(X) = r_\alpha(X)$ such that the following holds:

$$\mathbb{P}\left(\|\sqrt{n}(\overline{X} - \mu)\|_q \le r_\alpha(X) \right) \ge 1 - \alpha. \tag{6.5}$$

Let us recall that the point of view developed here is non-asymptotic, in the sense that (6.5) must hold when n and m are kept fixed (hence it covers the case where m is much larger than n).

[1] Note that the quantities are formulated here on the "test statistic scale", rather than on the "p-value scale".

A first idea is to apply the inequality $\|y\|_q \leq \|y\|_\infty$, $y \in \mathbb{R}^m$, and an union bound to get the Bonferroni threshold

$$r_{\alpha,\mathrm{Bonf}} = \overline{\Phi}^{-1}(\alpha/(2m)), \tag{6.6}$$

which satisfies (6.5). However, this region becomes too conservative when the dependence is high and for a large m. At the opposite side, the ideal (unknown) threshold $r_\alpha(X)$ is equal to $r_{\alpha,\mathrm{Quant}}(Q_0)$ where $Q_0 = \mathcal{N}(0,\Gamma)$ and where for any distribution Q on $\mathbb{R}^m$, we denote

$$r_{\alpha,\mathrm{Quant}}(Q) = \inf\left\{r \geq 0 \ : \ \mathbb{P}_{Z\sim Q}\left(\|Z\|_q \leq r\right) \geq 1-\alpha\right\}. \tag{6.7}$$

Figure 6.1 displays an illustration of the confidence region (6.4) with the oracle threshold when $m = 2$ and under (Gauss-ρ-equi), for $q \in \{1, 2, \infty\}$. Interestingly, while the threshold $r_{\alpha,\mathrm{Quant}}(Q_0)$ decreases with ρ for $q = \infty$, it is increasing with ρ when $q \in \{1, 2\}$. Hence, even in the simple equi-correlated case and $m = 2$, various behaviors can appear in the presence of dependence, which motivates the use of confidence regions that learn the dependence structure.

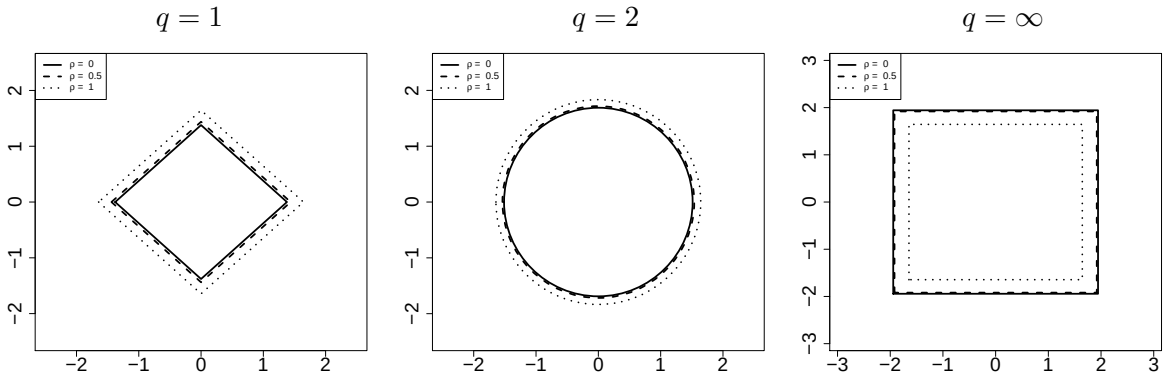


Figure 6.1: Illustration of ℓ_q-oracle confidence regions in $\mathbb{R}^2$ in the two-sided ρ-equi-correlated case. $m = 2$, $\alpha = 0.1$. Here we assume $\overline{X} = 0$.

6.1.3 Randomized confidence regions

To build adaptive confidence regions, we follow an idea similar to Section 5.1.3, by replacing into $r_{\alpha,\mathrm{Quant}}(Q_0)$ the distribution Q_0 by a randomized substitute. However, here, the distribution of $(S_i(X-\mu))_{i\in\mathcal{C}}$ should be approached on the whole set $\mathcal{C} = \{1,\dots,m\}$ and not only on $\mathcal{C} = \mathcal{H}_0$. Thus a different "recentered" approach should be used.

We follow the spirit of the bootstrap: if P_n is the empirical distribution of the sample $X = (X^{(1)},\dots,X^{(n)})$ and P is the distribution of $X^{(1)}$ (here $\mathcal{N}(\mu,\Gamma)$), then the general resampling heuristic of Bradley Efron [39, 40] is illustrated as follows ($\psi(X)$ being any statistic):

$$\begin{array}{ccc} \text{Real world} & \Big\| & \text{Bootstrap world} \\ P \rightsquigarrow X & & P_n \rightsquigarrow X^\star \\ \downarrow & & \downarrow \\ \psi(X) & & \psi(X^\star) \end{array}$$

The rationale behind the above scheme is that the "real world", where the truth is unknown, can be mirrored by a "bootstrap world", where everything is known. The single modification is that the unknown distribution P has to be replaced by P_n, while all the other operations are unchanged. For instance, the sampling process ($P \rightsquigarrow$) becomes a resampling process ($P_n \rightsquigarrow$) and the statistic $\psi(X)$

should be replaced by $\psi(X^\star)$. The heuristic is then that the distribution of any function of P and X, say $F(P,X)$, should be close to the distribution of $F(P_n,X^\star)$, taken conditionally on X.

Let us transpose this heuristic in our context. Let $\varepsilon=(\varepsilon_1,\dots,\varepsilon_n)$ be a sign[2] vector taken uniformly distributed on $\{-1,1\}^n$ as in Section 5.1.3, and let us assign the weight $W_j=\varepsilon_j+1\in\{0,2\}$ to each observation $X^{(j)}$. This gives the following heuristic:

$$\begin{aligned}\mathcal{D}\left(\|S(X-\mu)\|_q\right) &\simeq \mathcal{D}\left(\left\|S\left((X-\overline{X})^{\langle W\rangle}\right)\right\|_q \mid X\right)\\ &= \mathcal{D}\left(\left\|S\left((X-\overline{X})^{\langle \varepsilon\rangle}\right)\right\|_q \mid X\right),\end{aligned} \tag{6.8}$$

where for any vector $y\in(\mathbb{R}^m)^n$ and $z\in\mathbb{R}^n$, $y^{\langle z\rangle}=(z_1X^{(1)}\dots z_nX^{(n)})$ is defined as in formula (5.6). In addition, remember that the function $S(\cdot)$ is the "empirical mean operator" defined by (5.5) and $X-\overline{X}$ denotes the empirically recentered sample $(X^{(1)}-\overline{X},\dots,X^{(n)}-\overline{X})$.

When n grows to infinity, approximations of the type (6.8) can be validated by using the results on the exchangeable weighted bootstrap, e.g., Theorem 3.6.13 in [141]. Meanwhile, the particular "Rademacher" weighting $W_j=\varepsilon_j+1\in\{0,2\}$, $1\le j\le n$, shares some similarities with the subsampling process (resampling of the data without replacement), which is known to be appropriate for estimating quantiles under weak asymptotical conditions, see, e.g., Theorem 2.2.1 in [106].

In a non-asymptotical framework, however, there are only few results validating the use of (6.8) in the literature (see [56, 3] and the references therein). In the latter, typically, explicit remainder terms can be derived from concentration inequalities. In that spirit, we can show the following result (which is a simplified version of Proposition 3.4 in [P2]):

Theorem 6.2. *Let $\widehat{Q}(\cdot)$ be the resampling distribution operator as in (5.8) for some chosen $B\ge 1$, let $r_{\alpha,Quant}(\cdot)$ be the quantile operator defined by (6.7) and $r_{\alpha,Bonf}$ be the Bonferroni threshold (6.6). Let $\alpha_0\in(0,\alpha)$ and $\delta\in(0,1)$. Then, with the above notation, the following threshold satisfies (6.5):*

$$r_\alpha(X)=r_{\alpha_0(1-\delta),Quant}(\widehat{Q}(X-\overline{X}))+\left(1\wedge(Cn^{-1/2})\right)\ r_{\alpha-\alpha_0,Bonf}, \tag{6.9}$$

where $C=\{2\log(2/(\alpha_0\delta-(B+1)^{-1})_+)\}^{1/2}$.

The main idea of the proof (given below) is that the unknown threshold $r^\star_\alpha=r_{\alpha,\mathrm{Quant}}(\widehat{Q}(X-\mu))$ satisfies (6.5) by the symmetry of $X^{(1)}-\mu$ and the same exchangeability argument as in the proof of Lemma 5.2. Then, roughly, we just have to replace μ by $\overline{X}$ in $r^\star_\alpha$. The role of the remainder term is to compensate the "cost" of this operation. Let us also mention that Theorem 6.2 can be extended outside the Gaussian case by only assuming symmetry, that is, $X^{(1)}-\mu\sim\mu-X^{(1)}$, see [P2].

Proof. A crucial but elementary inequality is as follows: for any $\varepsilon\in\{-1,1\}^n$,

$$\left\|S\left((X-\mu)^{\langle\varepsilon\rangle}\right)\right\|_q\le\left\|S\left((X-\overline{X})^{\langle\varepsilon\rangle}\right)\right\|_q+|\overline{\varepsilon}_n|\times\|S(X-\mu)\|_q.$$

Let us now consider the "unfortunate" event

$$\mathcal{U}=\left\{X\ :\ r_{\alpha_0(1-\delta),\mathrm{Quant}}(\widehat{Q}(X-\overline{X}))+Cn^{-1/2}r_{\alpha-\alpha_0,\mathrm{Bonf}}<r_{\alpha_0,\mathrm{Quant}}(\widehat{Q}(X-\mu))\right\},$$

for which our threshold is "too small". Let $\varepsilon^0=1$ and $\varepsilon,\varepsilon^{(1)},\dots,\varepsilon^{(B)}$ be i.i.d. uniformly distributed on $\{-1,1\}^n$ and independent of a variable U uniformly distributed on $\{0,1,\dots,B\}$ (all these variables

[2] Other resampling choices are possible, see [P2].

being taken independent of X). Then, by definition of the quantile function, we have on the event $\mathcal{U}$,

$$\begin{aligned}\alpha_0 &\leq \mathbb{P}\left(\left\|S\left((X-\mu)^{\langle\varepsilon^{(U)}\rangle}\right)\right\|_q \geq r_{\alpha_0,\mathrm{Quant}}(\widehat{Q}(X-\mu)) \mid X\right)\\ &\leq \mathbb{P}\left(\left\|S\left((X-\mu)^{\langle\varepsilon^{(U)}\rangle}\right)\right\|_q > r_{\alpha_0(1-\delta),\mathrm{Quant}}(\widehat{Q}(X-\overline{X})) + Cn^{-1/2}r_{\alpha-\alpha_0,\mathrm{Bonf}} \mid X\right)\\ &\leq \alpha_0(1-\delta) + (B+1)^{-1}\left(1 + B\times\mathbb{P}\left(|\overline{\varepsilon}_n|\|S(X-\mu)\|_q > Cn^{-1/2}r_{\alpha-\alpha_0,\mathrm{Bonf}} \mid X\right)\right),\end{aligned}$$

and thus, on the event $\mathcal{U}$,

$$\mathbb{P}\left(|\overline{\varepsilon}_n| \times \|S(X-\mu)\|_q > Cn^{-1/2}r_{\alpha-\alpha_0,\mathrm{Bonf}} \mid X\right) \geq \alpha_0\delta - (B+1)^{-1},$$

which entails $\|S(X-\mu)\|_q > r_{\alpha-\alpha_0,\mathrm{Bonf}}$ by using Hoeffding's inequality, see [74]. Finally, we obtain

$$\begin{aligned}\mathbb{P}\left(\|\sqrt{n}(\overline{X}-\mu)\|_q > r_\alpha(X)\right) &\leq \mathbb{P}\left(\|\sqrt{n}(\overline{X}-\mu)\|_q > r_{\alpha_0,\mathrm{Quant}}(\widehat{Q}(X-\mu))\right) + \mathbb{P}(\mathcal{U})\\ &\leq \alpha_0 + \mathbb{P}(\|S(X-\mu)\|_q > r_{\alpha-\alpha_0,\mathrm{Bonf}}) \leq \alpha_0 + \alpha - \alpha_0,\end{aligned}$$

and the result is proved. □

First note that Theorem 6.2 corroborates the former asymptotical results, because when n grows to infinity and m is kept fixed, the remainder term disappears. When both m and n increase, however, there is a competition between the two terms of (6.9): while the first term is adaptive to dependence, so potentially improves Bonferroni threshold, the remainder term might be largely over-estimated and can deteriorate the global confidence region.

To illustrate the latter remark, the quality of the threshold (6.9) has been evaluated through a simulation study in [P3]. The setting considered there involves a two-dimensional spatial process, obtained by convolution between a Gaussian white noise and a pseudo-Gaussian convolution filter. The parameters are $n = 1000$; $m = 128^2$; $B = 1000$; $\alpha = 0.05$. The main conclusions are listed below:

- The remainder term is largely overestimated, because the "raw" threshold $r_{\alpha,\mathrm{Quant}}(\widehat{Q}(X-\overline{X}))$ looks indistinguishable from the ideal threshold $r_{\alpha,\mathrm{Quant}}(Q_0)$ on the plots.

- The threshold (6.9) nevertheless overcomes Bonferroni threshold when the dependence (here the bandwidth of the convolution filter) is large enough.

This study provides, to our knowledge, the first nonasymptotic approximation result on resampled quantile with an unknown distribution mean. However, we suspect that the remainder term can be made significantly smaller or, possibly, even completely removed in some cases.

6.1.4 Application to adaptive FWER control

Let us go back to the two-sided testing problem of $H_{0,i}$: "$\mu_i = 0$" against $H_{1,i}$: "$\mu_i \neq 0$" , $1 \leq i \leq m$. By Proposition 6.1, (single-step) FWER controlling procedures can be derived from the confidence regions of the previous section. In addition, since Theorem 6.2 still holds by replacing the supremum norm by the supremum norm over an arbitrary subset $\mathcal{C}$ of $\{1,\ldots,m\}$, we have at hand a family of confidence regions of the form (6.3). By Theorem 5.1, this gives rise to a new step-down FWER controlling procedure (called "recentered"). Below, it is qualitatively compared to the procedure developed in Section 5.1.3 (called "uncentered").

Dependence in μ and "bad H_0": the main advantage of using the empirically recentered distribution $\widehat{Q}(X-\overline{X})$ instead of the uncentered distribution $\widehat{Q}(X)$ (see (5.8)) is that $\widehat{Q}(X-\overline{X})$ is *translation invariant*: replacing X by $X-\mu$ leads to the same final threshold. As a result, the distribution of (6.9) does not depend on μ anymore. By contrast, the uncentered threshold is affected by the value of μ, which has some consequences for the resulting step-down FWER controlling procedure, that we now explain in a qualitative and informal discussion. The threshold of the k-th step of the step-down algorithm (see Theorem 5.1) is based on the distribution of

$$\sup_{i\notin R_{k-1}}\left|n^{-1/2}\sum_{j=1}^{n}\varepsilon_j X_i^{(j)}\right| = \sup_{i\notin R_{k-1}}\left|n^{1/2}\overline{\varepsilon}_n\mu_i + n^{-1/2}\sum_{j=1}^{n}\varepsilon_j(X_i^{(j)}-\mu_i)\right|,$$

where $\varepsilon\in\{-1,1\}^n$ is random. Hence, the uncentered supremum includes an undesirable term $n^{1/2}\overline{\varepsilon}_n\mu_i$. This term will unfavorably affect the supremum when μ_i takes large values outside R_{k-1}. At first step ($k=1$), we have $R_{k-1}=\emptyset$. Hence, a deterioration of the threshold can occur if μ has some "moderate" non zero coordinates. Roughly speaking, the uncentered resampling does generate a "bad H_0" distribution for the coordinates under moderate alternative, which inevitably affects the supremum distribution in the first iteration of the step-down algorithm. Hopefully for the uncentered approach, this phenomenon vanishes along the step-down iterations, because the large means are precisely weeded out after each step. Eventually, the above annoying term will become small at the $\widehat{k}$-th step.

This informal discussion has been illustrated on a devoted simulation framework in [P3] (see Section 4.4 there), for which the nonzero μ_i's are taken exponentially increasing.

Hybrid approach The recentered approach does not suffer from the above "bad H_0" phenomenon and thus already reaches its maximum performance after few iterations. However, it relies on a largely overestimated remainder term, which makes it far less powerful than the uncentered approach at the end of the step-down iterations. An interesting fact is that we can provably maintain the FWER control by combining these two approaches in the following way: given two parameters $\alpha_0\in(0,\alpha)$ and $\delta\in(0,1)$,

- first make the recentered thresholding coming from (6.9) (in its single step version, and with parameters α,α_0,δ), and consider its rejection set R_0;
- apply the step-down algorithm of Theorem 5.1 that takes R_0 as starting rejection set and that uses the uncentered thresholding collection of Lemma 5.2, but with α replaced by α_0.

The interest of this "hybrid" approach is that it can reduce the number of iterations of step-down algorithm when the non zero μ_i's lies in a wide range (up to some negligible loss in the level by taking α_0 close to α).

Let us finally mention that, from the practical point of view, the most relevant procedure is probably the step-down method based upon the threshold collection using the "raw" recentered threshold $r_{\alpha,\mathrm{Quant}}(\widehat{Q}(X-\overline{X}))$, without any remainder term. Even if not theoretically justified, we believe that it should be close to the optimal, at least for n larger than a "moderate" value.

A related work Let us mention that a recent study has brought a complement to our work, see [26]. The symmetry assumption has been removed by using Gaussian approximations for maxima of non-Gaussian sums. Also, the Gaussian multiplier bootstrap method has been used instead of our sign-flipping operation (i.e., with our notation, they use $\varepsilon_1\sim\mathcal{N}(0,1)$ combined with the uncentered

bootstrap). Then, they showed that the resulting RW's method controls asymptotically the FWER control by making moment assumptions and by assuming that $m = m_n$ depends on n in such a way that $(\log m_n)^7 = O(n^{1-c})$ for some $c \in (0,1)$ that can be arbitrary small.

6.2 Asymptotical study of FDP and a new central limit theorem [P8, P10]

The aim of this section is to find the asymptotic distribution of the FDP of the BH procedure, denoted below FDP_m, when there are weak dependencies between the individual tests. To achieve this goal, we establish a new functional central limit theorem (FCLT) for the empirical distribution function of the p-values, which is a result of independent interest. The considered asymptotic is in the number of hypotheses m, which is compatible with the high-dimensional data setting.

For instance, in model (Gauss-ρ-equi) with $\rho = \rho_m \to 0$ (one-sided testing), our result will imply that the convergence rate of FDP_m to $\pi_0\alpha$ is of the order $\{\min(m, 1/\rho_m)\}^{1/2}$ so potentially much slower than the standard rate $m^{1/2}$ holding under independence.

While the work [P8] was restricted to the particular model (Gauss-ρ-equi), this section deals with the more general framework of [P10].

6.2.1 Setting and aim

Let us consider the one-sided Gaussian setting of Section 2.1, with the random effect relaxation (Mixture) and a constant mean alternative $\Delta > 0$. Namely, while θ_i, $1 \le i \le m$, are i.i.d. $\mathcal{B}(1-\pi_0)$, the distribution of X conditionally on θ is a multivariate Gaussian vector of mean $\Delta\theta$ and covariance matrix $\Gamma^{(m)}$, where $\Gamma^{(m)}_{i,i} = 1$ for $1 \le i \le m$. The model parameters are thus π_0, Δ and $\Gamma^{(m)}$. In this setting, also note that π_0 is the limit (a.s.) of m_0/m as m tends to infinity.

Now, we search a rate $r_m = r_m(\Gamma^{(m)}, \Delta, \pi_0, \alpha) \to +\infty$ such that

$$r_m \ (\mathrm{FDP}_m - \pi_0\alpha) \rightsquigarrow \mathcal{N}(0,1)\,, \tag{6.10}$$

under some simple conditions on the covariance matrix $\Gamma^{(m)}$. Also, we seek for a rate r_m which can be written *explicitly* in function of the entries of the matrix $\Gamma^{(m)}$.

Our approach is based on a fundamental result due to Pierre Neuvial [98] (see also [59, 46]), who shows that the variable FDP_m can be written as an Hadamard differentiable functional of the null and alternative p-value e.d.f.'s:

$$\widehat{\mathbb{F}}_{0,m}(t) = \frac{1}{m_0(\theta)}\sum_{i=1}^{m}(1-\theta_i)\mathbf{1}\{\overline{\Phi}(X_i) \le t\};\ \widehat{\mathbb{F}}_{1,m}(t) = \frac{1}{m_1(\theta)}\sum_{i=1}^{m}\theta_i\mathbf{1}\{\overline{\Phi}(X_i) \le t\}\ ,t \in [0,1]. \tag{6.11}$$

Hence, by applying the functional delta method (see, e.g., Section 20.2 of [140]), the rate in (6.10) is directly related to the one given by the FCLT on the p-value e.d.f.'s (6.11).

6.2.2 Partial functional delta method for FDP_m

The functional delta method is a classical tool when studying asymptotic properties of statistics that can be written as a "smooth" function of converging processes. Classical examples include the convergence of quantiles, or of Mann-Whitney statistics. This section shows that FDP_m is another example, by using (and slightly extending) the method of [98]. Let us also recall that the notion of differentiability that suits to functional delta method is Hadamard differentiability (see, e.g., Section 20.2 in [140]).

Let us consider the linear space $D(0,1)$ of càd-làg functions on $[0,1]$ and the linear space $C(0,1)$ of continuous functions on $[0,1]$. By using (2.8), the processes (6.11) and $\widehat{\mathbb{G}}_m(t) = m^{-1}\sum_{i=1}^m \mathbf{1}\{\overline{\Phi}(X_i) \le t\} = \frac{m_0}{m}\widehat{\mathbb{F}}_{0,m}(t) + \frac{m_1}{m}\widehat{\mathbb{F}}_{1,m}(t)$, we can rewrite FDP_m as follows:

$$\mathrm{FDP}_m = \alpha\frac{\frac{m_0}{m}\widehat{\mathbb{F}}_{0,m}(\mathcal{T}(\widehat{\mathbb{G}}_m))}{\mathcal{T}(\widehat{\mathbb{G}}_m)} = \Psi\left(\frac{m_0}{m}\widehat{\mathbb{F}}_{0,m}, \frac{m_1}{m}\widehat{\mathbb{F}}_{1,m}\right), \tag{6.12}$$

where we used the following functionals (with the conventions $0/0 = 0$ and $\sup\{\emptyset\} = 0$):

$$\mathcal{T}(H) = \sup\{t \in [0,1] \ : \ H(t) \ge t/\alpha\} \text{ for } H \in D(0,1); \tag{6.13}$$

$$\Psi(H_0, H_1) = \alpha\frac{H_0(\mathcal{T}(H_0+H_1))}{\mathcal{T}(H_0+H_1)}, \text{ for } (H_0,H_1) \in D(0,1)^2. \tag{6.14}$$

Let us denote the expectations of $\widehat{\mathbb{F}}_{0,m}(t)$ (resp. $\widehat{\mathbb{F}}_{1,m}(t)$, $\widehat{\mathbb{G}}_m(t)$) by $F_0(t) = t$ (resp. $F_1(t) = \overline{\Phi}(\overline{\Phi}^{-1}(t) - \Delta)$, $G(t) = \pi_0 F_0(t) + \pi_1 F_1(t)$). Note that G is a strictly concave function such that $\lim_{t\to 0} G(t)/t = +\infty$ and thus also $\mathcal{T}(G) \in (0,1)$. In addition, by Corollary 7.12 of [98], $\mathcal{T}$ is Hadamard differentiable on the space $D(0,1)$ (endowed with the supremum norm) at G, tangentially to the set $C(0,1)$. As a consequence, standard calculations show that Ψ is Hadamard differentiable at $(\pi_0 F_0, \pi_1 F_1)$ on the space $D(0,1)^2$ (endowed with the supremum norm) tangentially to $C(0,1)^2$, with derivative

$$\dot{\Psi}_{(\pi_0 F_0, \pi_1 F_1)}(H_0, H_1) = \alpha\frac{H_0(\mathcal{T}(G))}{\mathcal{T}(G)}, \text{ for } (H_0,H_1) \in C(0,1)^2. \tag{6.15}$$

Now, by using (6.12), the functional delta method provides the asymptotic behavior of FDP_m from the one of $(\frac{m_0}{m}\widehat{\mathbb{F}}_{0,m}, \frac{m_1}{m}\widehat{\mathbb{F}}_{1,m})$.

Proposition 6.3. *Consider the setting and notation of Section 6.2.1, with F_0, F_1 and G defined above and denote by $t^\star$ the unique $t \in (0,1)$ such that $G(t) = t/\alpha$. Assume that the two following distribution convergences hold (w.r.t. the Skorokhod topology and the corresponding Borel σ-field [3]):*

$$a_m\left(\frac{m_0}{m}\widehat{\mathbb{F}}_{0,m} - \pi_0 F_0\right) \rightsquigarrow \mathbb{Z}_0; \quad a_m\left(\frac{m_1}{m}\widehat{\mathbb{F}}_{1,m} - \pi_1 F_1\right) \rightsquigarrow \mathbb{Z}_1, \tag{6.16}$$

for some positive sequence $(a_m)_m$ tending to infinity and where $\mathbb{Z}_0$ and $\mathbb{Z}_1$ are two processes valued a.s. in $C(0,1)$. Then we have

$$a_m(FDP_m - \pi_0\alpha) \rightsquigarrow \alpha\frac{\mathbb{Z}_0(t^\star)}{t^\star}. \tag{6.17}$$

Proposition 6.3 is a "partial" functional delta method in the sense that (6.16) does not involve the joint convergence of $(\frac{m_0}{m}\widehat{\mathbb{F}}_{0,m}, \frac{m_1}{m}\widehat{\mathbb{F}}_{1,m})$. This simplification is possible because the derivative $\dot{\Psi}_{(\pi_0 F_0, \pi_1 F_1)}(H_0, H_1)$ given by (6.15) only depends on H_0.

[3] Here, it is important to note that $\widehat{\mathbb{F}}_{0,m}$ and $\widehat{\mathbb{F}}_{1,m}$ are not measurable when endowing $D(0,1)$ with the Borel σ-field coming from the $\|\cdot\|_\infty$-topology (the so-called ball σ-field), see Section 18 of [15]. However, the functional delta method does use the $\|\cdot\|_\infty$-topology. This can appear as incompatible at first sight. As a matter fact, this is not, because $\mathbb{Z}_0$ and $\mathbb{Z}_1$ are a.s. in $C(0,1)$ and because converging to a continuous function w.r.t. the Skorokhod distance is equivalent to the uniform convergence, see the proof of Proposition S.1.1 in the supplement of [P10] for more details.

6.2.3 A new functional central limit theorem

Establishing (6.10) with Proposition 6.3 now requires a functional central limit theorem (FCLT) of the type (6.16). Consider $Y \sim \mathcal{N}(0, \Gamma^{(m)})$ and let us study the associated empirical distribution function $\widehat{\mathbb{F}}_m(t) = m^{-1}\sum_{i=1}^m \mathbf{1}\{\overline{\Phi}(Y_i) \le t\}$, $t \in [0,1]$. For this, write $a_m(\widehat{\mathbb{F}}_m(t) - t) = W_m + Z_m$, for some sequence a_m and where

$$W_m(t) = a_m\left(\widehat{\mathbb{F}}_m(t) - t - \phi(\overline{\Phi}^{-1}(t))\overline{Y}_m\right); \;\; Z_m(t) = \phi(\overline{\Phi}^{-1}(t))\, a_m\overline{Y}_m\,, \tag{6.18}$$

where ϕ denotes the standard Gaussian density. Classically, a consequence of Mehler's formula (see, e.g., [54]) is that the two processes W_m and Z_m are not correlated. *Our main idea is to focus on the case where the effect of the dependence is (asymptotically) only carried by the second process Z_m.* To this end, observe that $\mathrm{Var}(a_m\overline{Y}_m) = a_m^2(m^{-1} + \gamma_m)$, where

$$\gamma_m = m^{-2}\sum_{i\neq j}\Gamma^{(m)}_{i,j}. \tag{6.19}$$

Letting $a_m = \left(m^{-1} + |\gamma_m|\right)^{-1/2}$ and assuming the convergence

$$m\gamma_m \to \theta, \quad \text{for some } \theta \in [-1, +\infty] \tag{6.20}$$

(which always holds up to take a subsequence), the limit of the covariance function of Z_m is $(t,s) \mapsto \frac{1+\theta}{1+|\theta|}\phi(\overline{\Phi}^{-1}(t))\phi(\overline{\Phi}^{-1}(s))$. Meanwhile, we can see that under the condition

$$\frac{a_m^2}{m^2}\sum_{i\neq j}\left(\Gamma^{(m)}_{i,j}\right)^2 \to 0, \tag{vanish-secondorder}$$

the limit of the covariance function of the process W_m is the same as under independence (up to the rescaling), that is, $(t,s) \mapsto \frac{1}{1+|\theta|}\left(t\wedge s - ts - \phi(\overline{\Phi}^{-1}(t))\phi(\overline{\Phi}^{-1}(s))\right)$. This entails the following limit for the covariance function of the whole process $a_m(\widehat{\mathbb{F}}_m(t) - t)$:

$$K(t,s) = \frac{1}{1+|\theta|}(t\wedge s - ts) + \frac{\theta}{1+|\theta|}\phi(\overline{\Phi}^{-1}(t))\phi(\overline{\Phi}^{-1}(s)). \tag{6.21}$$

Note that Assumption (vanish-secondorder) can be rewritten as $\frac{a_m}{m}\|\Gamma^{(m)} - I_m\| \to 0$, where $\|\cdot\|$ denotes the Frobenius matrix norm. Hence, this assumption roughly means that $\Gamma^{(m)}$ lies asymptotically in a neighborhood of I_m.

Obviously, convergence of the covariance function is not sufficient to obtain an FCLT. We should establish the convergence of finite dimensional laws and show the tightness in the Skorokhod space. This requires some of the following assumptions on $\Gamma^{(m)}$:

$$\frac{r_m^{4+\varepsilon_0}}{m^2}\sum_{i\neq j}\left(\Gamma^{(m)}_{i,j}\right)^4 \to 0, \quad \text{for some } \varepsilon_0 > 0; \tag{H_1}$$

$$\frac{r_m^{2+\varepsilon_0}}{m^2}\sum_{i\neq j}\left(\Gamma^{(m)}_{i,j}\right)^2 = o(1), \quad \text{for some } \varepsilon_0 > 0; \tag{H_2}$$

$$m\gamma_m^{1+\varepsilon_0} \to \infty, \quad \text{for some } \varepsilon_0 > 0. \tag{H_3}$$

Theorem 6.4. *Let us consider* $Y \sim \mathcal{N}(0, \Gamma^{(m)})$ *and the associated empirical distribution function* $\widehat{\mathbb{F}}_m(t) = m^{-1}\sum_{i=1}^m \mathbf{1}\{\overline{\Phi}(Y_i) \le t\}$, $t \in [0,1]$. *Consider the sequence* γ_m *defined by* (6.19) *and* $a_m = \big(m^{-1} + |\gamma_m|\big)^{-1/2}$. *Assume that* $\Gamma^{(m)}$ *satisfies either* {(vanish-secondorder), (H_1) *and* (6.20)} *or* {(H_2) *and* (H_3)}. *Then, there exists a continuous Gaussian process* $(\mathbb{Z}_t)_{t\in[0,1]}$ *with covariance function defined by* (6.21) *and such that the following convergence holds (w.r.t. the Skorokhod topology and the corresponding Borel* σ*-field)*

$$a_m(\widehat{\mathbb{F}}_m - I) \rightsquigarrow \mathbb{Z}, \ \text{as } m \to \infty, \tag{6.22}$$

where $I(t) = t$ *denotes the identity function.*

There are two underlying regimes in (6.22):

(i) if $m\gamma_m \to \theta < +\infty$, we have $a_m \propto m^{1/2}$ and the process $m^{1/2}(\widehat{\mathbb{F}}_m - I)$ converges to a (continuous Gaussian) process with covariance function given by $(t,s) \mapsto t \wedge s - ts + \theta\, \phi(\overline{\Phi}^{-1}(t))\phi(\overline{\Phi}^{-1}(s))$. Hence, the limit process is a standard Brownian bridge when $\theta = 0$, but has a covariance function smaller (resp. larger) if $\theta < 0$ (resp. $\theta > 0$).

(ii) if $m\gamma_m \to \theta = +\infty$, we have $a_m \sim (\gamma_m)^{-1/2} \ll m^{1/2}$ and $(\gamma_m)^{-1/2}(\widehat{\mathbb{F}}_m - I)$ converge to the process $\phi(\overline{\Phi}^{-1}(\cdot))Z$ for $Z \sim \mathcal{N}(0,1)$. Hence the "Brownian" part asymptotically disappears.

The regimes (i) and (ii) are illustrated in Figure 6.2: as $m\gamma_m$ grows, the influence of the "Brownian" part decreases while that of the (randomly rescaled) function $\phi(\overline{\Phi}^{-1}(\cdot))$ increases. Also, the scale of the Y-axis indicates that $m^{1/2}$ is not a suitable rate for large values of $m\gamma_m$.

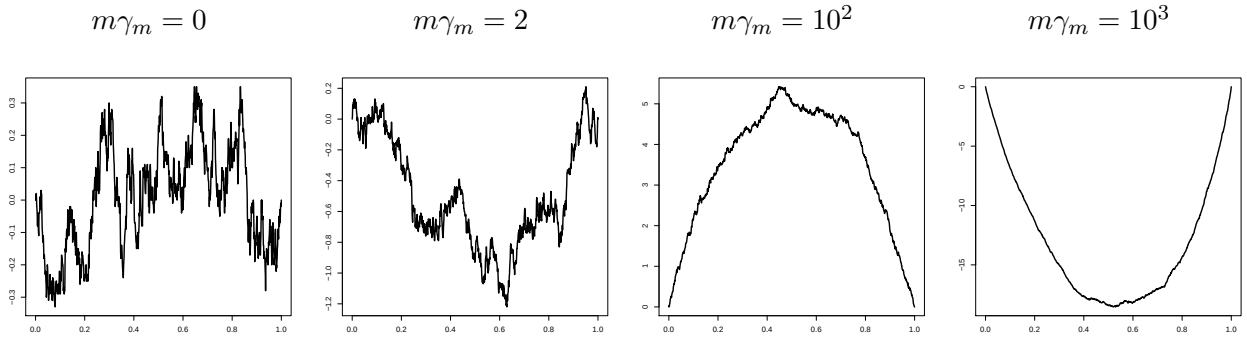


Figure 6.2: Plot of $t \mapsto m^{1/2}(\widehat{\mathbb{F}}_m(t) - t)$ for 4 realizations of Y. These realizations have been generated in model (Gauss-ρ-equi) and for $m = 10^4$.

Let us mention that Theorem 6.4 has an interest in its own right. While the literature on FCLT is colossal since the first Donsker theorem [34, 33, 36] (see for instance reviews in [2, 29, 35]), only few results deal with the non-stationary case. We can mention the work [129] making assumptions of type $\max_{i\neq j}\Gamma^{(m)}_{i,j}$ "small enough" and the work [5] assuming that $\Gamma^{(m)}_{i,j} \le r(|i-j|)$ for all i,j and for a function r independent of m and vanishing at infinity, so still relying on a stationary structure. By contrast, our result covers factor models or sample correlation matrix, see [P10] and the next section for more details. Nevertheless, a price to pay is that our assumptions exclude the case of short-range stationary correlations, e.g., $\Gamma^{(m)}$ tridiagonal 1/2-1-1/2.

6.2.4 Application to FDP convergence

By combining Theorem 6.4 with Proposition 6.3, we obtain the following result:

Corollary 6.5. *Consider the setting and notation of Section 6.2.1 and denote by* $t^\star = t^\star(\Delta, \pi_0, \alpha)$ *the unique* $t \in (0,1)$ *such that* $\pi_0 t + (1-\pi_0)\overline{\Phi}(\overline{\Phi}^{-1}(t) - \Delta) = t/\alpha$. *Assume that* $\Gamma^{(m)}$ *satisfies either* {(vanish-secondorder) *and* (H_1)} *or* {(H_2) *and* (H_3)}. *Then we have the convergence* (6.10) *with the rate* r_m *given by*

$$r_m(\Gamma^{(m)}, \Delta, \pi_0, \alpha) = \frac{1}{\pi_0 \alpha} \left\{ \frac{1}{\pi_0 m} \left(\frac{1}{t^\star} - \pi_0 \right) + \left(\frac{\phi(\overline{\Phi}^{-1}(t^\star))}{t^\star} \right)^2 m^{-2} \sum_{i \neq j} \Gamma^{(m)}_{i,j} \right\}^{-1/2}, \tag{6.23}$$

where ϕ *denotes the standard Gaussian density.*

Below, we provide some examples for which the assumptions of Corollary 6.5 are satisfied:

- Equi-correlation: $\Gamma^{(m)}$ is of the form (Gauss-ρ-equi) with $\rho = \rho_m \to 0$. The rate is $r_m \propto \{\min(m, 1/\rho_m)\}^{1/2}$ (as announced at the beginning of the section);
- Alternate equi-correlation: $\Gamma^{(m)}_{i,j} = (-1)^{i+j}\rho_m$ for $i \neq j$ with $m^{1+\delta}\rho_m^2 \to 0$ for some $\delta > 0$. The rate is $r_m \propto m^{1/2}$;
- Signed 1-factor model: $\Gamma^{(m)}_{i,j} = \xi^{(m)}_i \xi^{(m)}_j \rho_m$ for $i \neq j$ where $\xi^{(m)}$ is m-vector of signs with $\overline{\xi}^{(m)}_m \sim m^{-D}$, where $0 \leq D \leq 1/2$ and assuming $m^{(1+\delta)\wedge(D(4+\delta))}\rho_m^2 \to 0$ for some $\delta > 0$. The rate is $r_m \propto \min\left(m^{1/2}, m^D \rho_m^{-1/2}\right)$. Note that $D = 0$ encompasses equi-correlation.
- Long-range stationary correlations: $\Gamma^{(m)}_{i,j} = |j-i|^{-D}$ for $i \neq j$ and $D \in (0,1)$. The rate is $r_m \propto m^{D/2}$;
- Sample correlation matrix: consider Z a $n_m \times m$ matrix with i.i.d. standard Gaussian entries and let $\Gamma^{(m)} = D^{-1} S D^{-1}$ where $S = n_m^{-1} Z^T Z$ is the $m \times m$ sample covariance matrix of the columns of Z (not recentered) and D is the $m \times m$ diagonal matrix with diagonal $(S_{1,1}^{1/2}, \cdots, S_{m,m}^{1/2})$. Assuming $m^{1+\delta}/n_m \to 0$ for some $\delta > 0$, the rate is $r_m \propto m^{1/2}$ (and all the convergences hold in probability);

In conclusion, (6.23) shows that the convergence rate gets slower when the correlations between the individual statistical tests are positive and increase. This corroborates what is observed on Figure 5.3 in Section 5.2: correlations can deteriorate the concentration of FDP_m around $\pi_0\alpha$[4]. By contrast, perhaps surprisingly, our study shows that negative correlations help to increase the convergence rate r_m.

[4] Also observe on Figure 5.3 that $\rho = 0.1$ falls outside the Gaussian regime. Hence, the asymptotic distribution is not attained for this "too large" value of ρ. We suspect that the FDP cannot be approximated by a linear function in that case, which results in a poor accuracy when using the delta method.

6.3 BH procedure as an optimal classifier [P12]

This section provides connections between FDR control and classification, by studying optimal classification properties of the BH procedure. More precisely, we show that its mis-classification risk mimics the Bayes risk for a specific choice of the level α_m, as m tends to infinity. A crucial assumption is sparsity, that is, that the label "0" (no signal) is generated with a probability near to 1. Such an optimal property is often referred to as *adaptation to the unknown sparsity*, see [32, 1]. This section presents results coming from our work [P12], itself extended[5] the work of [20]. In addition, to put this issue in perspective, a comparison with the detection problem is also made.

6.3.1 ζ-Subbotin location model

Let $(X_i, \theta_i) \in \mathbb{R} \times \{0,1\}$, $1 \le i \le m$, be m i.i.d. random pairs such that

(i) we observe $X_1, \dots, X_m$ and not the "labels" $\theta_1, \dots, \theta_m$;

(ii) the distribution of X_1 conditionally on $\theta_1 = 0$ has the density $d(\cdot)$ given by (6.24);

(iii) the distribution of X_1 conditionally on $\theta_1 = 1$ has the density $d(\cdot - \Delta_m)$, for an unknown location parameter $\Delta_m > 0$.

Above, $d(\cdot)$ denotes the ζ-Subbotin density

$$d(x) = (L_\zeta)^{-1} e^{-|x|^\zeta/\zeta},\ x \in \mathbb{R}, \text{ with } L_\zeta = \int_{-\infty}^{+\infty} e^{-|x|^\zeta/\zeta} dx. \tag{6.24}$$

Throughout the section, the "shape" parameter ζ is supposed to be known and to belong to $(1, \infty)$. Note that taking $\zeta = 2$ gives the standard Gaussian density.

The goal is to make an inference on the labels $\theta_1, \dots, \theta_m$, on the basis of the sample $X_1, \dots, X_m$. In this section, we will study properties of such inferences when m is large. Hence, we assume that the model parameters $\pi_{0,m} = \mathbb{P}(\theta_1 = 0)$ and Δ_m both depend on m. Specifically, we assume that the signal is rare:

$$1 - \pi_{0,m} = m^{-\beta}, \quad 0 < \beta < 1. \tag{Sparsity}$$

In addition, to balance the sparsity, we will assume that Δ_m tends to infinity[6], typically $\Delta_m \propto (\log m)^{1/\zeta}$.

In the above model, $X_1, \dots, X_m$ are i.i.d. with a common mixture density equal to $x \in \mathbb{R} \mapsto (1 - m^{-\beta}) d(x) + m^{-\beta} d(x - \Delta_m)$. The induced distribution on $X = (X_1, \dots, X_m)$ is denoted by $P^{(m)}_{\beta, \Delta_m}$ in the sequel.

Finally, note that the above setting corresponds to a typical multiple testing framework, with Assumption (Mixture). However, it is presented above in a way that is intended to underline the similarity with the transductive classification model in machine learning theory. More specifically, this model is close to the semi-supervised novelty detection model, where the user has at hand both unlabeled data and training data *coming from only one nominal class*, see [19] and references therein. Here, while the unlabeled sample would be the X_i's, the training data would correspond to an infinite sample under the null (implying the knowledge of $d(\cdot)$).

[5] The work [20] was seminal but restricted to a Gaussian scale model and to an asymptotic result of the type of Theorem 6.9 (i) below. Afterwards, our study [P12] additionally provides non-asymptotic oracle inequalities, deals with general Subbotin location/scale models and supplies a finite sample choice of α_m.

[6] As we will see, even if Δ_m tends to infinity, the signal can still be undetectable. Hence, this setting is sometimes referred to as "rare/weak" signal by some authors.

6.3.2 Boundaries for detection and classification risks

In the above model, the two following questions can be asked:

- is there some signal?
- if so, where is the signal?

Detection The first question is classically referred to as *detection*, see the work [32] of David Donoho and Jiashun Jin (2004) and the series of work [77, 78, 79, 80] by Yuri Ingster and co-authors. It corresponds to a single testing of

$$H_0\text{: ``}X \sim \mathcal{N}(0, I_m)\text{''} \text{ against } H_1^{(m)}\text{: ``}X \sim P_{\beta,\Delta_m}^{(m)}\text{''}. \tag{6.25}$$

This testing problem is strongly connected to adaptive testing, for which one null hypothesis is tested against of family of alternatives hypotheses. Loosely, the family of alternatives is explored here through the mixture $P_{\beta,\Delta_m}^{(m)}$.

Obviously, the larger Δ_m, the easier the testing problem. This has been formalized in terms of a separation rate which roughly is the minimum rate at which Δ_m should grows to make H_0 and $H_1^{(m)}$ asymptotically separated, see [4]. As a matter fact, the constant matters in the rate. To emphasize this, let us define the detection risk of a single test $\psi_m(X) \in \{0,1\}$ by

$$\mathcal{R}_m^D(\psi_m) = \mathbb{P}_{H_0}(\psi_m(X) = 1) + \mathbb{P}_{H_1^{(m)}}(\psi_m(X) = 0). \tag{6.26}$$

Also define $\rho^D : (1/2, 1) \to (0, 1)$ by

$$\rho^D(\beta) = \begin{cases} (2^{1/(\zeta-1)} - 1)^{\zeta-1}(\beta - 1/2) & \text{if } 1/2 < \beta \le 1 - 2^{-\zeta/(\zeta-1)}; \\ (1 - (1-\beta)^{1/\zeta})^{\zeta} & \text{if } 1 - 2^{-\zeta/(\zeta-1)} \le \beta < 1. \end{cases} \tag{6.27}$$

The following result holds.

Proposition 6.6 ([77, 32]). *Consider the model of Section 6.3.1 with (Sparsity) and $\Delta_m = (\zeta r \log m)^{1/\zeta}$, for some unknown parameters $(\beta, r) \in (1/2, 1) \times (0, 1)$ and the function given by (6.27). Consider the detection risk defined by (6.26). Then the following holds:*

- *if $r > \rho^D(\beta)$, H_0 and $H_1^{(m)}$ separate asymptotically, that is, there exists a sequence of tests $(\psi_m)_m$ (possibly depending on β, r) such that $\mathcal{R}_m^D(\psi_m) \to 0$ as m tends to infinity;*
- *if $r < \rho^D(\beta)$, H_0 and $H_1^{(m)}$ merge asymptotically, that is, for all sequence of tests $(\psi_m)_m$ (possibly depending on β, r), we have $\mathcal{R}_m^D(\psi_m) \to 1$ as m tends to infinity.*

Proposition 6.6 hence puts forward a striking "threshold effect" related to the graph of the function ρ^D, which is referred to as the *detection boundary*.

Classification Compared to assessing whether there exists some signal, it is more demanding to search where the signal is, because an inference should be done for all the labels θ_i, $1 \le i \le m$. For a (measurable) classification rule $\widehat{h}_m : \mathbb{R} \to \{0,1\}$, depending on $X_1, ..., X_m$, the mis-classification risk is defined by

$$R_m^{\mathcal{C}}(\widehat{h}_m) = \mathbb{E}\left(m^{-1} \sum_{i=1}^m \mathbf{1}\{\widehat{h}_m(X_i) \neq \theta_i\} \right) / (1 - \pi_{0,m}). \tag{6.28}$$

Above, the rescaling by $1 - \pi_{0,m}$ makes $R_m^{\mathcal{C}}(h_m^0) = 1$ for the trivial procedure $h_m^0 \equiv 0$ that decides always 0 regardless of the data. The following result can certainly be considered as classical, see, e.g., [83].

Proposition 6.7. *Consider the model of Section 6.3.1 with (Sparsity) and* $\Delta_m = (\zeta r \log m)^{1/\zeta}$*, for some unknown parameters* $(\beta, r) \in (0,1) \times (0,1)$*. Consider the mis-classification risk defined by (6.28). Then the following holds:*

- *if* $r > \beta$*, the classification can be made perfect, that is, there exists a sequence of classification rules* $(\widehat{h}_m)_m$ *(possibly depending on* β, r*) such that* $\mathcal{R}_m^{\mathcal{C}}(\widehat{h}_m) \to 0$ *as* m *tends to infinity;*
- *if* $r < \beta$*, the classification problem is impossible, that is, for all sequence of classification rules* $(\widehat{h}_m)_m$ *(possibly depending on* β, r*), we have* $\liminf_m \{\mathcal{R}_m^{\mathcal{C}}(\widehat{h}_m)\} \geq 1$ *as* m *tends to infinity.*

In other words, Proposition 6.7 states that the classification boundary is $\rho^{\mathcal{C}}(\beta) = \beta$ for $\beta \in (0,1)$. The classification and detection boundaries are both displayed in Figure 6.3 according to a phase diagram in the $\beta \times r$ space (for the Gaussian case, i.e., for $\zeta = 2$).

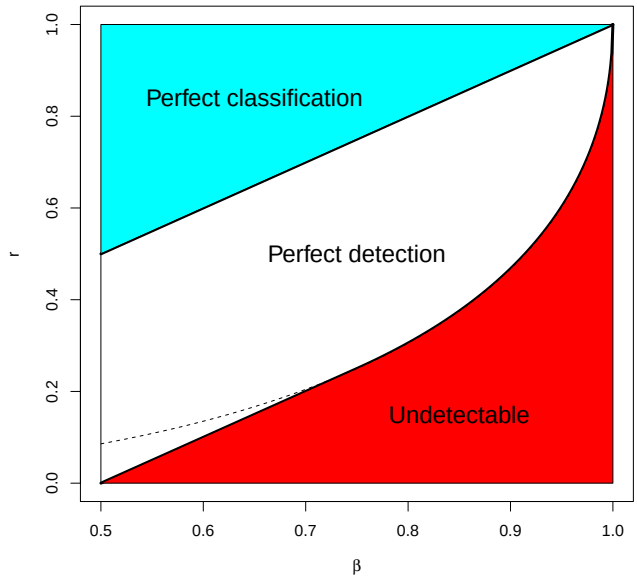


Figure 6.3: Phase diagram in the sparsity×signal space for the classification and detection problems in the Gaussian case. Classification boundary $\rho^{\mathcal{C}}(\beta) = \beta$ and detection boundary $\rho^{D}(\beta)$ given by (6.27) for $\beta \in (1/2, 1)$ (solid lines). The dashed line corresponds to the detection boundary achieved by the BH detection rule, see text.

Optimal procedure Now that the phase diagram is established, an issue is to find a procedure that *actually achieves* the boundary defined by ρ, that is, a procedure such that for all (β, r) satisfying $r > \rho(\beta)$, the corresponding risk is tending to 0 as m grows to infinity. Since such procedure is not using the true value of β and r, it is said *adaptive* to the unknown parameters β and r.

For the detection problem, the BH detection rule corresponds to rejecting the null H_0 of (6.25) when at least one rejection is made by the BH procedure. Donoho and Jin (2004) have proved that the BH procedure *does not* attain the boundary ρ^D, see [32] . As a matter of fact, the boundary attained by the BH procedure is slightly larger on the range $1/2 < \beta < 3/4$, see the dashed line in Figure 6.3. Roughly, the idea is that in this regime "not too sparse" the signal can be too weak to make one of the X_i's large, which makes the BH procedure missing the signal. By contrast, the moderate sparsity can

be taken into account by considering the cardinals of p-value level sets. This idea, that can be traced back to Tukey, has been formalized in [32] with the higher-criticism (HC) procedure. They proved that HC attains the boundary ρ^D on the full range $\beta \in (1/2, 1)$.

In the rest of the section, the optimality issue is investigated for the classification problem.

6.3.3 Optimality results for the BH classifier

For convenience, let us consider the p-value standardization $p_i(X) = \overline{D}(X_i)$, for $1 \le i \le m$, where $\overline{D}(u) = \int_u^\infty d(x)dx$ is the upper-tail cumulative distribution of X_1 conditionally on $\theta_1 = 0$.

Classically, since $d(x - \Delta_m)/d(x)$ is nondecreasing in x, the solution that minimizes the mis-classification risk (6.28) is a thresholding rule of the type $h^\star_m(x) = \mathbf{1}\{\overline{D}(x) \le t^\star_m\}$, generally referred to as Bayes' rule. The BH classifier (at level α_m) is defined by $\widehat{h}^{BH}_m(x) = \mathbf{1}\{\overline{D}(x) \le \widehat{t}^{BH}_m(\alpha_m)\}$, where $\widehat{t}^{BH}_m(\alpha_m)$ is defined by Algorithm 2.1 as usual.

Our first main result states that *the BH rule attains the classification boundary* when α_m is appropriately chosen.

Theorem 6.8. *Consider the BH rule $\widehat{h}^{BH}_m$ at a level α_m chosen such that*

$$\alpha_m \to 0\,, \quad \frac{\log \alpha_m}{(\log m)^{1-1/\zeta}} \to 0, \quad \text{as } m \to \infty. \tag{6.29}$$

Consider the model of Section 6.3.1 with (Sparsity) and $\Delta_m = (\zeta r \log m)^{1/\zeta}$ for an unknown parameter couple $(\beta, r) \in (0,1)^2$, and the mis-classification risk defined by (6.28). Then, whenever $r > \beta$, we have $\mathcal{R}^{\mathcal{C}}_m(\widehat{h}^{BH}_m) \to 0$ as m tends to infinity.

Theorem 6.8 shows that the behavior of BH rule is appropriate when the classification task is hopeless or can be made perfect. However, we might argue that the interesting cases lie in between. Hence, our second main result explores the optimality property of the BH rule *exactly on the classification boundary*. For this, we follow [20] and consider $C_m = \overline{D}(\overline{D}^{-1}(t^\star_m) - \Delta_m)$, which corresponds to the power of the Bayes rule. We can easily see that (Sparsity) and the assumption

$$C_m = C \in (0,1) \text{ for all } m \ge 2 \tag{BP}$$

entail that $\Delta_m \sim (\zeta\beta \log m)^{1/\zeta}$ and that $R^{\mathcal{C}}_m(h^\star_m) \sim 1 - C$. In particular, under (Sparsity) and (BP), the part of the phase diagram which is explored lies on the boundary $r = \beta$, see Figure 6.3. As a result, considering the pair of parameters (β, C) instead of (β, r) somewhat "distorts" the sparsity×signal space and "zooms in" the classification boundary to focus only on interesting balanced situations. In this framework, the following result can be proved:

Theorem 6.9. *Consider the Bayes rule $h^\star_m$ and the BH rule $\widehat{h}^{BH}_m$ at level α_m. Consider the model of Section 6.3.1 with (Sparsity) and (BP) for an unknown parameter couple $(\beta, C) \in (0,1)^2$, and the mis-classification risk defined by (6.28). Then the following holds:*

(i) Choosing α_m satisfying (6.29) ensures $R^{\mathcal{C}}_m(\hat{h}^{BH}_m) \sim R^{\mathcal{C}}_m(h^\star_m)$ as m tends to infinity;

(ii) Choosing additionally $\alpha_m \propto 1/(\log m)^{1-1/\zeta}$ ensures

$$R^{\mathcal{C}}_m(\hat{h}^{BH}_m) = R^{\mathcal{C}}_m(h^\star_m) \left(1 + O\left(\frac{1}{(\log m)^{1-1/\zeta}}\right)\right). \tag{6.30}$$

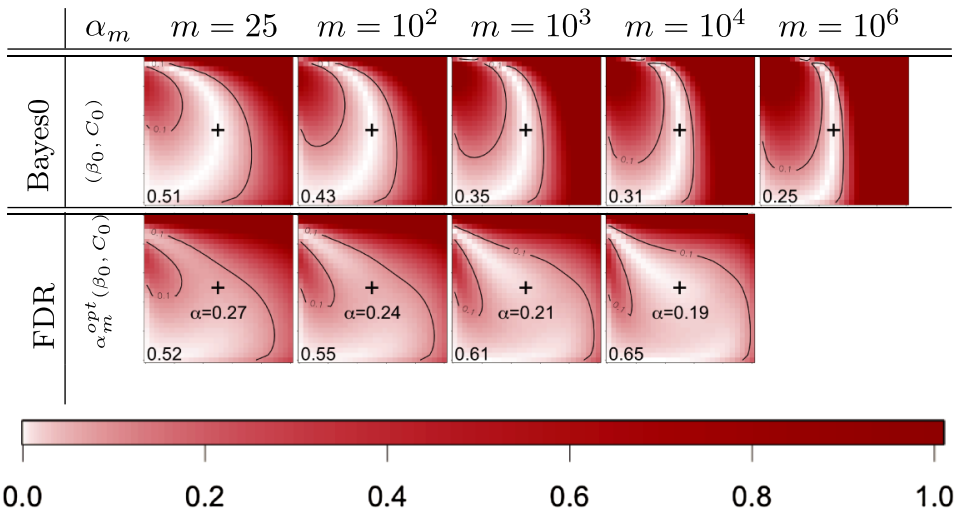


Figure 6.4: Heatmaps in the $\beta \times C$ space of the relative excess risks for the procedures "Bayes0" (top) and "FDR" (bottom) and for increasing values of m (from left to right), see text. For each plot: the black curve represents the level set where the relative excess risk is equal to 0.1 ; the bottom-left number is the fraction of configurations (β, C) inside the black curve; the point $(\beta_0, C_0) = (1/2, 1/2)$ is marked by "+". Finally, for the FDR plots, α is $\alpha_m^{opt}(1/2, 1/2)$.

Hence, while Theorem 6.8 states that the BH rule enjoys the same "thresholding effect" as the Bayes rule in the $\beta \times r$ space, Theorem 6.9 (i) shows that it has an equivalent risk in the finer $\beta \times C$ space. Here, the explicit convergence rate in Theorem 6.9 (ii) comes from a careful analysis of the role of α_m in our non-asymptotic oracle inequality. Namely, in the sketch of proof provided below, we found an "optimal" way to choose α_m from the parameters β and C, that we denote $\alpha_m^{opt}(\beta, C)$.

To get a computable BH rule, we cannot use $\alpha_m^{opt}(\beta, C)$ because β and C are unknown. However, an idea is to use the BH rule with $\alpha_m = \alpha_m^{opt}(\beta_0, C_0)$, for some prior values β_0, C_0 of β, C. In Figure 6.4, the performance of a classification procedure $\hat{h}_m$ is evaluated according to the relative excess risk $(R_m^{\mathcal{C}}(\hat{h}_m) - R_m^{\mathcal{C}}(h_m^\star))/R_m^{\mathcal{C}}(h_m^\star))$, when (β, C) is varying in the sparsity×signal square $(0, 1)^2$. Two procedures are considered: first, the BH rule with $\alpha_m = \alpha_m^{opt}(\beta_0, C_0)$ and $(\beta_0, C_0) = (1/2, 1/2)$, denoted by "FDR". Second, for comparison, we have also considered the plug-in Bayes rule $h_m^\star(\beta_0, C_0)$, in which the unknown values (β, C) are replaced by $(\beta_0, C_0) = (1/2, 1/2)$. It is denoted by "Bayes0". Increasing values of m are considered from $m = 25$ to[7] $m = 10^6$. Interestingly, while Bayes0 performs well when $\beta \simeq \beta_0$, it performs poorly when β is mis-specified, and increasingly so as m increases. By constrast, for the FDR method, the configurations with low relative excess risk span (essentially) the whole range of β. Overall, this illustrates the adaptation w.r.t. the sparsity parameter β of the BH classification rule on the classification boundary.

Sketch of proof for Theorem 6.8 and Theorem 6.9. Let $\widehat{\mathbb{G}}_m$ (resp. G) denote the empirical (resp. theoretical) distribution function of the p-values. The first argument is that, since $\widehat{\mathbb{G}}_m$ concentrates around G, the BH threshold $\widehat{t}_m^{BH}(\alpha_m) = \max\{t \in [0, 1] : \widehat{\mathbb{G}}_m(t) \geq t/\alpha_m\}$ (see (2.8)) should be close to the deterministic quantity $t_m^{BH}(\alpha_m) = \max\{t \in [0, 1] : G(t) \geq t/\alpha_m\}$. This guides an optimal

[7] Since we use here the time consuming exact calculations of Section 3.3 to compute the BH risk, the case $m = 10^6$ is not reported for FDR.

choice of α_m, denoted $\alpha_m^{opt}(\beta, C)$, which is such that

$$t_m^{BH}(\alpha_m^{opt}(\beta, C)) = t_m^\star. \tag{6.31}$$

Interestingly, in the latter relation, $\alpha_m^{opt}(\beta, C)$ can be interpreted as a correction factor that cancels the difference between $t \mapsto \overline{D}(\overline{D}^{-1}(t) - \Delta_m)/t$ and $t \mapsto \frac{\partial}{\partial t}\overline{D}(\overline{D}^{-1}(t) - \Delta_m)$ (see [P12] for more details).

The second argument is the following finite-sample oracle inequality, which (essentially) uses the log concavity of $d(\cdot)$, Bennett's inequality and exact formulas for the distribution of $\widehat{t}_m^{BH}$ (coming from [51]): for $\alpha_m \in (0, 1/2)$, letting $\tau_m = \pi_{0,m}/\pi_{1,m} > 1$, $\varepsilon, \nu \in (0, 1)$, and for any $m \geq 2$ such that $(\zeta \log \tau_m)^{1-1/\zeta} \geq \frac{d(0)}{C_m(1-\nu)}(\log(\alpha_m^{-1}/q_m^{opt}) - \log(\nu\pi_{0,m}(1-\varepsilon)))$, we have

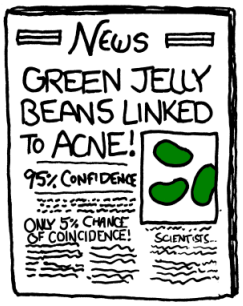
$$\begin{aligned} R_m^{\mathcal{C}}(\hat{h}_m^{BH}) - R_m^{\mathcal{C}}(h_m^\star) \leq & \left(\frac{\alpha_m}{1-\alpha_m} + d(0)\frac{(\log(\alpha_m^{-1}/q_m^{opt}) - \log(\nu\pi_{0,m}(1-\varepsilon)))_+}{(\zeta \log \tau_m)^{1-1/\zeta}} \right) \\ & + \frac{\alpha_m/(m\pi_{1,m})}{(1-\alpha_m)^2} + e^{-m\pi_{1,m}\nu\varepsilon^2 C_m/4}, \end{aligned} \tag{6.32}$$

where $q_m^{opt} = 1/\alpha_m^{opt} - 1 > 1$. Note that $\log \tau_m \sim \beta \log m$ under (Sparsity) which makes the RHS of (6.32) becomes small. Namely, since $\log(\alpha_m^{-1}/q_m^{opt}) \leq \log(\alpha_m^{-1})$, the RHS of (6.32) tends to zero when α_m satisfies (6.29) and thus Theorem 6.8 and Theorem 6.9 (i) follows. Finally, for Theorem 6.9 (ii), we use that $q_m^{opt} \propto (\log \tau_m)^{1-1/\zeta}$ under (BP), which comes from (6.31).

An outlook In a specific classification context, this work showed that controlling the FDR is strikingly linked to minimize the standard mis-classification risk. This illustrates that the BH thresholding allows to adapt to the quantity of signal in the data. Remember that this fact was qualitatively observed in Figure 1.2 of Chapter 1.

Interestingly, to get the optimality, the choice of α_m is not let arbitrary. It should be taken tending to zero (slowly enough) to produce an appropriate classifier under sparsity. This is markedly different from situations where the BH procedure is used for model selection, see [1, 8], for which we can choose $\alpha_m \sim \alpha \in (0, 1/2)$ to derive the optimality.

...



Conclusion

Presented results This manuscript presented an overview of some modern multiple testing problems and proposed some solutions, that relied on the work listed page 81.

In a nutshell, while the FWER control and the FDP-based controls had both their own interest and respective interpretation, the methodologies that came into play were not the same. FWER control required a probabilistic bound on the infimum of the p-values under the null (or, equivalently, on the supremum of the tests statistics under the null). As a consequence, via an appropriate step-down algorithm, the problem was reduced to control a *single* statistic (infimum or supremum). This fact was extensively used in the resampling-based approaches of Sections 5.1 and 6.1.

By contrast, such a reduction was not possible for the FDR/FDP control, because the involved procedures was defined by crossing points between the ordered p-values and the critical values. Instead, the error rates of such procedures depended on the multivariate distribution function of the ordered p-values, as we have seen in Section 3.3 by using combinatorics.

Markedly, this FDR control could be put in a very simple manner by using two simple sufficient conditions in Section 3.1, which paved the way for the "continuous" testing of Section 3.2 and the issue of *increasing power* via π_0-estimators and weighted p-values in Section 4.1 and Section 4.2, respectively. A counterpart is that strong dependence assumptions were required in the analysis. A second way to simplify the FDR/FDP controlling issue was to consider an asymptotic situations where the number m of hypotheses grows to infinity. While Section 6.2 computed the asymptotic distribution of the FDP via the Delta method for a general (weak) dependent setting, Section 5.2 showed that an asymptotic FDP control was possible under dependence, either under weak dependence, or in restriction to a particular family of positively (strongly) dependent factor models. In addition, as an overall summary, FWER, FDR and FDP controls was compared in Section 5.2.4.

Many mathematical difficulties came while establishing FDR/FDP control. An alternative strategy was used in Section 5.3 in the case where the p-values share local correlations. The idea was to combine clustering and multi-scale FWER control, which allows an extra "cluster scale" for signal detection. However, we emphasized that the interpretation of the testing phase holds *conditionally* on the clustering. Hence, the final interpretation of the tests changes.

Open problems While this manuscript solves some issues, it left some others open:

- Two-sided testing: the inferences proposed for FDR/FDP control were often investigated in the one-sided case. A reason is that the p-values are increasing functions of the errors in that case, hence positively correlated errors lead to positively correlated p-values, which is a desirable property to achieve FDR/FDP control. By contrast, for two-sided testing, such a positive dependence property is lost. How to make a correct FDR/FDP inference for that specific dependency structure? A direction could be to derive upper-bounds, by dividing the "two-sided rejection space" into a collection of "one-sided rejection spaces" plus some remainder terms.

- Data driven weighting: the optimal multi-weighted procedure described in Algorithm 4.4 used weight vectors (4.6) which depend on the unknown alternative distribution of the data. Could

we estimate these weight vectors and incorporate them into Algorithm 4.4 to provide both FDR control and optimality? To investigate this issue, a convenient setting could be the case where the nulls are grouped in several homogeneous blocks, because the p-value mixture distribution (4.4) of each block can be well approximated when the size of the block tend to infinity.

- Calibration of β: in Section 3.1, several ways to choose β are proposed according to (3.2). Under dependence between the p-values, is it possible to choose β according to some resampling/permutation scheme and to still provide an FDR control while improving power? The key point seems to choose an appropriate "prior" distribution ν on the number of rejections. A possible direction is to choose ν by sample splitting.

- In Section 6.1, the adaptive confidence region relied on a largely over-estimated remainder term, which was due to the empirical recentering of the data. What is the minimum value of this term? Maybe it is simply 0.

- FDP control under unknown dependence: the task of Section 5.2 was investigated under a previously known dependence (typically Γ known in the Gaussian framework). Is it possible to provide a rigorous FDP control under unknown dependence via a resampling scheme? A possible direction is to consider a simple n-sample factor model $X_i^{(j)} = \mu_i + c_i W^{(j)} + \zeta_i^{(j)}$, $1 \leq i \leq m$, $1 \leq j \leq n$ and to combine the sign-flipping randomization of Section 5.1.3 with the asymptotical analysis (when m tends to infinity) of Section 5.2.3.

- FDP control with rate: under weak dependence, we mentioned in Section 5.2 the property $\mathbb{P}(\mathrm{FDP}(\mathrm{BH}) > \alpha) \to 0$ as m grows to infinity. What is the rate of this convergence? This problem seems to have been ignored so far in the literature, even under independence.

- p-value correction: in (facmodel), remember that the observations $X_i = \mu_i + c_i W + \zeta_i$ are "disturbed" by the terms $\vartheta_i = c_i W$, which model the dependence structure. Obviously, in this model, it is desirable to remove the ϑ_i's and to consider the test statistics $X_i^\star = X_i - \vartheta_i = \mu_i + \zeta_i$ rather than the original X_i's. Are there estimates $\widehat{\vartheta}_i$'s of the ϑ_i's such that the new plug-in test statistics $\widehat{X}_i = X_i - \widehat{\vartheta}_i$ improve the multiple testing inference? If so, some signal assumptions seem to be required to avoid the case where the signal vector $(\mu_i)_i$ looks "similar" to the disturbance vector $(c_i W)_i$.

Obviously, many other research directions are possible, for instance relying on martingale proofs [134, 91, 72], post-hoc inference [92, 65], full bayesian approaches [124, 62] and variable selection [8, 21].

All these avenues are both exciting and challenging for future work.

Publications

[P1] Arlot, S., Blanchard, G., and Roquain, E. (2007). Resampling-based confidence regions and multiple tests for a correlated random vector. In *Learning theory*, volume 4539 of *Lecture Notes in Comput. Sci.*, pages 127–141. Springer, Berlin.

[P2] Arlot, S., Blanchard, G., and Roquain, E. (2010a). Some nonasymptotic results on resampling in high dimension. I. Confidence regions. *Ann. Statist.*, 38(1):51–82.

[P3] Arlot, S., Blanchard, G., and Roquain, E. (2010b). Some nonasymptotic results on resampling in high dimension. II. Multiple tests. *Ann. Statist.*, 38(1):83–99.

[P4] Blanchard, G., Delattre, S., and Roquain, E. (2014). Testing over a continuum of null hypotheses with False Discovery Rate control. *Bernoulli*, 20(1):304–333.

[P5] Blanchard, G., Dickhaus, T., Roquain, E., and Villers, F. (2014). On least favorable configurations for step-up-down tests. *Statist. Sinica*, 24(1):1–23.

[P6] Blanchard, G. and Roquain, E. (2008). Two simple sufficient conditions for FDR control. *Electron. J. Stat.*, 2:963–992.

[P7] Blanchard, G. and Roquain, E. (2009). Adaptive false discovery rate control under independence and dependence. *J. Mach. Learn. Res.*, 10:2837–2871.

[P8] Delattre, S. and Roquain, E. (2011). On the false discovery proportion convergence under Gaussian equi-correlation. *Statist. Probab. Lett.*, 81(1):111–115.

[P9] Delattre, S. and Roquain, E. (2015). New procedures controlling the false discovery proportion via Romano-Wolf's heuristic. *Ann. Statist.*, 43(3):1141–1177.

[P10] Delattre, S. and Roquain, E. On empirical distribution function of high-dimensional Gaussian vector components with an application to multiple testing. *Bernoulli*. To appear.

[P11] Kim, K. I., Roquain, E., and van de Wiel, M. A. (2010). Spatial clustering of array CGH features in combination with hierarchical multiple testing. *Stat. Appl. Genet. Mol. Biol.*, 9(1):Art. 40.

[P12] Neuvial, P. and Roquain, E. (2012). On false discovery rate thresholding for classification under sparsity. *Ann. Statist.*, 40(5):2572–2600.

[P13] Roquain, E. (2011). Type I error rate control for testing many hypotheses: a survey with proofs. *J. Soc. Fr. Stat.*, 152(2):3–38.

[P14] Roquain, E. and Schbath, S. (2007). Improved compound Poisson approximation for the number of occurrences of any rare word family in a stationary Markov chain. *Adv. in Appl. Probab.*, 39(1):128–140.

[P15] Roquain, E. and van de Wiel, M. (2009). Optimal weighting for false discovery rate control. *Electron. J. Stat.*, 3:678–711.

[P16] Roquain, E. and Villers, F. (2011). Exact calculations for false discovery proportion with application to least favorable configurations. *Ann. Statist.*, 39(1):584–612.

Bibliography

[1] Abramovich, F., Benjamini, Y., Donoho, D. L., and Johnstone, I. M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.*, 34(2):584–653.

[2] Arcones, M. A. (1994). Limit theorems for nonlinear functionals of a stationary Gaussian sequence of vectors. *Ann. Probab.*, 22(4):2242–2274.

[3] Arlot, S. (2007). *Resampling and Model Selection*. PhD thesis, University Paris-Sud 11.

[4] Baraud, Y. (2002). Non-asymptotic minimax rates of testing in signal detection. *Bernoulli*, 8(5):577–606.

[5] Bardet, J.-M. and Surgailis, D. (2013). Moment bounds and central limit theorems for Gaussian subordinated arrays. *J. Multivariate Anal.*, 114:457–473.

[6] Benjamini, Y. (2013). Are most research findings really false? Special public lecture of the conference on Multiple Comparison Procedures (MCP) in the Statistical Sciences Research Institute of Southampton.

[7] Benjamini, Y. and Braun, H. (2002). John W. tukey's contributions to multiple comparisons. *The Annals of Statistics*, 30(6):pp. 1576–1594.

[8] Benjamini, Y. and Gavrilov, Y. (2009). A simple forward selection procedure based on false discovery rate control. *Ann. Appl. Stat.*, 3(1):179–198.

[9] Benjamini, Y. and Heller, R. (2007). False discovery rates for spatial signals. *J. Amer. Statist. Assoc.*, 102(480):1272–1281.

[10] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, 57(1):289–300.

[11] Benjamini, Y. and Hochberg, Y. (1997). Multiple hypotheses testing with weights. *Scand. J. Statist.*, 24(3):407–418.

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

[13] Benjamini, Y., Krieger, A. M., and Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3):491–507.

[14] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, 29(4):1165–1188.

[15] Billingsley, P. (1968). *Convergence of probability measures*. John Wiley & Sons Inc., New York.

[16] Richard M. Bittman, Joseph P. Romano, Carlos Vallarino, and Michael Wolf. Optimal testing of multiple hypotheses with common effect direction. *Biometrika*, 96(2):399–410, 2009.

[17] Black, M. A. (2004). A note on the adaptive control of false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 66(2):297–304.

[18] Blanchard, G. and Fleuret, F. (2007). Occam's hammer. In *Learning theory*, volume 4539 of *Lecture Notes in Comput. Sci.*, pages 112–126. Springer, Berlin.

[19] Blanchard, G., Lee, G., and Scott, C. (2010). Semi-supervised novelty detection. *J. Mach. Learn. Res.*, 11:2973–3009.

[20] Bogdan, M., Chakrabarti, A., Frommlet, F., and Ghosh, J. K. (2011). Asymptotic bayes-optimality under sparsity of some multiple testing procedures. *Ann. Statist.*, 39(3):1551–1579.

[21] Bogdan, M., van den Berg, E., Su, W., and Candes, E. (2013). Statistical estimation and testing via the sorted L1 norm. *ArXiv e-prints*.

[22] Bonferroni, C. (1935). *Il calcolo delle assicurazioni su gruppi di teste*. Tipografia del Senato.

[23] Cai, T. T. and Jin, J. (2010). Optimal rates of convergence for estimating the null density and proportion of nonnull effects in large-scale multiple testing. *Ann. Statist.*, 38(1):100–145.

[24] Cai, T. T. and Sun, W. (2009). Simultaneous testing of grouped hypotheses: finding needles in multiple haystacks. *J. Amer. Statist. Assoc.*, 104(488):1467–1481.

[25] Celisse, A. and Robin, S. (2010). A cross-validation based estimation of the proportion of true null hypotheses. *Journal of Statistical Planning and Inference*, 140(11):3132 – 3147.

[26] Chernozhukov, V., Chetverikov, D., and Kato, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786–2819.

[27] Chi, Z. and Tan, Z. (2008). Positive false discovery proportions: intrinsic bounds and adaptive control. *Statist. Sinica*, 18(3):837–860.

[28] Chin, K., DeVries, S., Fridlyand, J., Spellman, P. T., Roydasgupta, R., Kuo, W. L., Lapuk, A., Neve, R. M., Qian, Z., Ryder, T., Chen, F., Feiler, H., Tokuyasu, T., Kingsley, C., Dairkee, S., Meng, Z., Chew, K., Pinkel, D., Jain, A., Ljung, B. M., Esserman, L., Albertson, D. G., Waldman, F. M., and Gray, J. W. (2006). Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell*, 10:529–541.

[29] Dedecker, J. and Prieur, C. (2007). An empirical central limit theorem for dependent sequences. *Stochastic Process. Appl.*, 117(1):121–142.

[30] Dickhaus, T. (2008). *False Discovery Rate and Asymptotics*. PhD thesis, Heinrich-Heine-Universität Düsseldorf.

[31] Dickhaus, T. (2014). *Simultaneous statistical inference*. Springer, Heidelberg. With applications in the life sciences.

[32] Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.*, 32(3):962–994.

[33] Donsker, M. D. (1952). Justification and extension of Doob's heuristic approach to the Komogorov-Smirnov theorems. *Ann. Math. Statistics*, 23:277–281.

[34] Doob, J. L. (1949). Heuristic approach to the Kolmogorov-Smirnov theorems. *Ann. Math. Statistics*, 20:393–403.

[35] Doukhan, P., Lang, G., Surgailis, D., and Teyssière, G., editors (2010). *Dependence in probability and statistics*, volume 200 of *Lecture Notes in Statistics*. Springer-Verlag, Berlin.

[36] Dudley, R. M. (1966). Weak convergences of probabilities on nonseparable metric spaces and empirical measures on Euclidean spaces. *Illinois J. Math.*, 10:109–126.

[37] Dudoit, S. and van der Laan, M. J. (2008). *Multiple testing procedures with applications to genomics*. Springer Series in Statistics. Springer, New York.

[38] Duncan, D. B. (1955). Multiple range and multiple F tests. *Biometrics*, 11:1–42.

[39] Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.*, 7(1):1–26.

[40] Efron, B. (2003). Second thoughts on the bootstrap. *Statist. Sci.*, 18(2):135–140. Silver anniversary of the bootstrap.

[41] Efron, B. (2007). Doing thousands of hypothesis tests at the same time. *Metron - International Journal of Statistics*, LXV(1):3–21.

[42] Efron, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.*, 23(1):1–22.

[43] Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.*, 96(456):1151–1160.

[44] Fan, J., Han, X., and Gu, W. (2012). Estimating false discovery proportion under arbitrary covariance dependence. *Journal of the American Statistical Association*, 107(499):1019–1035.

[45] Fang, Z. and Hu, T. (1997). Developments on MTP_2 properties of absolute value multinormal variables with nonzero means. *Acta Math. Appl. Sinica (English Ser.)*, 13(4):376–384.

[46] Farcomeni, A. (2007). Some results on the control of the false discovery rate under dependence. *Scand. J. Statist.*, 34(2):275–297.

[47] Farcomeni, A. (2009). Generalized augmentation to control the false discovery exceedance in multiple testing. *Scand. J. Stat.*, 36(3):501–517.

[48] Ferreira, J. A. and Zwinderman, A. H. (2006). On the Benjamini-Hochberg method. *Ann. Statist.*, 34(4):1827–1849.

[49] Finner, H., Dickhaus, T., and Roters, M. (2009). On the false discovery rate and an asymptotically optimal rejection curve. *Ann. Statist.*, 37(2):596–618.

[50] Finner, H., Gontscharuk, V., and Dickhaus, T. (2012). False discovery rate control of step-up-down tests with special emphasis on the asymptotically optimal rejection curve. *Scandinavian Journal of Statistics*, 39(2):382–397.

[51] Finner, H. and Roters, M. (2002). Multiple hypotheses testing and expected number of type I errors. *Ann. Statist.*, 30(1):220–238.

[52] Fisher, R. A. (1925). *Statistical methods for research workers.* Edinburgh Oliver & Boyd.

[53] Fisher, R. A. (1935). *The Design of Experiments.* Oliver and Boyd, Edinburgh.p.

[54] Foata, D. (1981). Some Hermite polynomial identities and their combinatorics. *Adv. in Appl. Math.*, 2(3):250–259.

[55] Friguet, C., Kloareg, M., and Causeur, D. (2009). A factor model approach to multiple testing under dependence. *J. Amer. Statist. Assoc.*, 104(488):1406–1415.

[56] Fromont, M. (2003). *Quelques problèmes de sélection de modèles : construction de tests adaptatifs, ajustement de pénalités par des méthodes de bootstrap.* PhD thesis, University Paris-Sud 11.

[57] Gavrilov, Y., Benjamini, Y., and Sarkar, S. K. (2009). An adaptive step-down procedure with proven FDR control under independence. *Ann. Statist.*, 37(2):619–629.

[58] Genovese, C. R. (2004). A tutorial on false discovery control. Talk at Hannover Workshop.

[59] Genovese, C. R. and Wasserman, L. (2004). A stochastic process approach to false discovery control. *Ann. Statist.*, 32(3):1035–1061.

[60] Genovese, C. R., Roeder, K., and Wasserman, L. (2006). False discovery control with p-value weighting. *Biometrika*, 93(3):509–524.

[61] Genovese, C. R. and Wasserman, L. (2006). Exceedance control of the false discovery proportion. *J. Amer. Statist. Assoc.*, 101(476):1408–1417.

[62] Ghosal, S. and Roy, A. (2011). Predicting false discovery proportion under dependence. *J. Amer. Statist. Assoc.*, 106(495):1208–1218.

[63] Goeman, J. and Solari, A. (2010). The sequential rejection principle of familywise error control. *Ann. Statist.*, 38(6):3782–3810.

[64] Goeman, J. J. and Finos, L. (2012). The inheritance procedure: multiple testing of tree-structured hypotheses. *Stat. Appl. Genet. Mol. Biol.*, 11(1):Art. 11, 20.

[65] Goeman, J. J. and Solari, A. (2011). Multiple testing for exploratory research. *Statist. Sci.*, 26(4):584–597.

[66] Gomes, L. (2014). Machine-Learning Maestro Michael Jordan on the Delusions of Big Data and Other Huge Engineering Efforts. *IEEE Spectrum*.

[67] Gontscharuk, V. (2010). *Asymptotic and Exact Results on FWER and FDR in Multiple Hypothesis Testing*. PhD thesis, Heinrich-Heine-Universität Düsseldorf.

[68] Grazier G'Sell, M., Wager, S., Chouldechova, A., and Tibshirani, R. (2013). Sequential Selection Procedures and False Discovery Rate Control. *ArXiv e-prints*.

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

[70] Guo, W. and Romano, J. (2007). A generalized Sidak-Holm procedure and control of generalized error rates under independence. *Stat. Appl. Genet. Mol. Biol.*, 6:Art. 3, 35 pp. (electronic).

[71] He, L. and Sarkar, S. K. (2013). On improving some adaptive BH procedures controlling the FDR under dependence. *Electronic Journal of Statistics*, 7:2683–2701.

[72] Heesen, P. and Janssen, A. (2014). Inequalities for the false discovery rate (FDR) under dependence. *ArXiv e-prints*.

[73] Hochberg, Y. and Tamhane, A. C. (1987). *Multiple comparison procedures*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York.

[74] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30.

[75] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.*, 6(2):65–70.

[76] Hu, J. X., Zhao, H., and Zhou, H. H. (2010). False discovery rate control with groups. *J. Amer. Statist. Assoc.*, 105(491):1215–1227.

[77] Ingster, Y. I. (1998). Minimax detection of a signal for l^n-balls. *Math. Methods Statist.*, 7(4):401–428 (1999).

[78] Ingster, Y. I. (2002). Adaptive detection of a signal of growing dimension. II. *Math. Methods Statist.*, 11(1):37–68.

[79] Ingster, Y. I., Pouet, C., and Tsybakov, A. B. (2009). Classification of sparse high-dimensional vectors. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, 367(1906):4427–4448.

[80] Ingster, Y. I., Tsybakov, A. B., and Verzelen, N. (2010). Detection boundary in sparse regression. *Electron. J. Stat.*, 4:1476–1526.

[81] Jin, J. (2008). Proportion of non-zero normal means: universal oracle equivalences and uniformly consistent estimators. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 70(3):461–493.

[82] Jin, J. and Cai, T. T. (2007). Estimating the null and the proportional of nonnull effects in large-scale multiple comparisons. *J. Amer. Statist. Assoc.*, 102(478):495–506.

[83] Jin, J. and Ke, T. (2014). Rare and Weak effects in Large-Scale Inference: methods and phase diagrams. *ArXiv e-prints*.

[84] Karlin, S. and Rinott, Y. (1980). Classes of orderings of measures and related correlation inequalities. I. Multivariate totally positive distributions. *J. Multivariate Anal.*, 10(4):467–498.

[85] Karlin, S. and Rinott, Y. (1981). Total positivity properties of absolute value multinormal variables with applications to confidence interval estimates and related probabilistic inequalities. *Ann. Statist.*, 9(5):1035–1049.

[86] Korn, E. L., Troendle, J. F., McShane, L. M., and Simon, R. (2004). Controlling the number of false discoveries: application to high-dimensional genomic data. *J. Statist. Plann. Inference*, 124(2):379–398.

[87] Leek, J. T. and Storey, J. D. (2008). A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences*, 105(48):18718–18723.

[88] Lehmann, E. L. (1966). Some concepts of dependence. *Ann. Math. Statist.*, 37:1137–1153.

[89] Lehmann, E. L. and Romano, J. P. (2005a). Generalizations of the familywise error rate. *Ann. Statist.*, 33:1138–1154.

[90] Lehmann, E. L. and Romano, J. P. (2005b). *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, New York, third edition.

[91] Liang, K. and Nettleton, D. (2012). Adaptive and dynamic adaptive procedures for false discovery rate control and estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1):163–182.

[92] Meinshausen, N. (2006). False discovery control for multiple tests of association under general dependence. *Scand. J. Statist.*, 33(2):227–237.

[93] Meinshausen, N. (2008). Hierarchical testing of variable importance. *Biometrika*, 95(2):265–278.

[94] Meinshausen, N., Meier, L., and Bühlmann, P. (2009). p-values for high-dimensional regression. *J. Amer. Statist. Assoc.*, 104(488):1671–1681.

[95] Meinshausen, N., Maathuis, M. H., and Bühlmann, P. (2011). Asymptotic optimality of the Westfall-Young permutation procedure for multiple testing under dependence. *Ann. Statist.*, 39(6):3369–3391.

[96] Miller, C. J., Genovese, C. R., Nichol, R. C., Wasserman, L., Connolly, A., Reichart, D., Hopkins, A., Schneider, J., and Moore, A. (2001). Controlling the false-discovery rate in astrophysical data analysis. *The Astronomical Journal*, 122(6):3492–3505.

[97] Muris, J., Ylstra, B., Cillessen, S., Ossenkoppele, G., Kluin-Nelemans, J., Eijk, P., Nota, B., Tijssen, M., de Boer, W., van de Wiel, M., van den Ijssel, P., Jansen, P., de Bruin, P., van Krieken, J., Meijer, G., Meijer, C., and Oudejans, J. (2007). Profiling of apoptosis genes allows for clinical stratification of primary nodal diffuse large B-cell lymphomas. *Br. J. Haematol.*, 136:38–47.

[98] Neuvial, P. (2008). Asymptotic properties of false discovery rate controlling procedures under independence. *Electron. J. Stat.*, 2:1065–1110.

[99] Neuvial, P. (2013). Asymptotic results on adaptive false discovery rate controlling procedures based on kernel estimators. *J. Mach. Learn. Res.*, 14:1423–1459.

[100] Nguyen, V. H. and Matias, C. (2014). On efficient estimators of the proportion of true null hypotheses in a multiple testing setup. *Scandinavian Journal of Statistics*. To appear.

[101] Pantazis, D., Nichols, T. E., Baillet, S., and Leahy, R. M. (2005). A comparison of random field theory and permutation methods for statistical analysis of meg data. *NeuroImage*, 25:383–394.

[102] Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50:157–175.

[103] Perone Pacifico, M., Genovese, C. R., Verdinelli, I., and Wasserman, L. (2004). False discovery control for random fields. *J. Amer. Statist. Assoc.*, 99(468):1002–1014.

[104] Picard, F. (2014). *A statistical tour of genomic data*. Habilitation à diriger des recherches, Université Lyon I.

[105] Plackett, R. L. (1983). Karl pearson and the chi-squared test. *International Statistical Review / Revue Internationale de Statistique*, 51(1):pp. 59–72.

[106] Politis, D. N., Romano, J. P., and Wolf, M. (1999). *Subsampling*. Springer Series in Statistics. Springer-Verlag, New York.

[107] Reiner-Benaim, A. (2007). FDR control by the BH procedure for two-sided correlated tests with implications to gene expression data analysis. *Biom. J.*, 49(1):107–126.

[108] Revuz, D. and Yor, M. (1991). *Continuous martingales and Brownian motion*, volume 293 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin.

[109] Robin, S. (2002). A compound Poisson model for word occurrences in DNA sequences. *J. Roy. Statist. Soc. Ser. C*, 51(4):437–451.

[110] Roeder, K. and Wasserman, L. (2009). Genome-wide significance levels and weighted hypothesis testing. *Statist. Sci.*, 24(4):398–413.

[111] Romano, J. P. and Shaikh, A. M. (2006a). On stepdown control of the false discovery proportion. In *Optimality*, volume 49 of *IMS Lecture Notes Monogr. Ser.*, pages 33–50. Inst. Math. Statist., Beachwood, OH.

[112] Romano, J. P. and Shaikh, A. M. (2006b). Stepup procedures for control of generalizations of the familywise error rate. *Ann. Statist.*, 34(4):1850–1873.

[113] Joseph P. Romano, Azeem M. Shaikh, and Michael Wolf. Control of the false discovery rate under dependence using the bootstrap and subsampling. *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research*, 17(3):417–442, December 2008.

[114] Joseph P. Romano, Azeem M. Shaikh, and Michael Wolf. Consonance and the closure method in multiple testing. *Int. J. Biostat.*, 7(1):Art. 12, 27, 2011.

[115] Romano, J. P. and Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *J. Amer. Statist. Assoc.*, 100(469):94–108.

[116] Romano, J. P. and Wolf, M. (2007). Control of generalized error rates in multiple testing. *Ann. Statist.*, 35(4):1378–1408.

[117] Rubin, D., Dudoit, S., and van der Laan, M. (2006). A method to increase the power of multiple testing procedures through sample splitting. *Stat. Appl. Genet. Mol. Biol.*, 5:Art. 19, 20 pp. (electronic).

[118] Sarkar, S. K. (2002). Some results on false discovery rate in stepwise multiple testing procedures. *Ann. Statist.*, 30(1):239–257.

[119] Sarkar, S. K. (2008a). On methods controlling the false discovery rate. *Sankhya, Ser. A*, 70:135–168.

[120] Sarkar, S. K. (2008b). Two-stage stepup procedures controlling FDR. *Journal of Statistical Planning and Inference*, 138(4):1072–1084.

[121] Sarkar, T. K. (1969). *Some lower bounds of reliability.* ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)–Stanford University.

[122] Schbath, S. (1995). Compound poisson approximation of word counts in DNA sequences. *ESAIM: Probability and Statistics*, 1:1–16.

[123] Schweder, T. and Spjøtvoll, E. (1982). Plots of P-values to evaluate many tests simultaneously. *Biometrika*, 69(3):493–502.

[124] Scott, J. G. and Berger, J. O. (2006). An exploration of aspects of Bayesian multiple testing. *J. Statist. Plann. Inference*, 136(7):2144–2162.

[125] Seeger, P. (1968). A note on a method for the analysis of significances en masse. *Technometrics*, 10(3):586–593.

[126] Shaffer, J. P. (2012). Erich Lehmann's contributions to multiple decision making. In *Selected works of E. L. Lehmann*, Sel. Works Probab. Stat., pages 609–616. Springer, New York.

[127] Shorack, G. R. and Wellner, J. A. (1986). *Empirical processes with applications to statistics.* Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York.

[128] Soric, B. (1989). Statistical "discoveries" and effect-size estimation. *Journal of the American Statistical Association*, 84(406):pp. 608–610.

[129] Soulier, P. (2001). Moment bounds and central limit theorem for functions of Gaussian vectors. *Statist. Probab. Lett.*, 54(2):193–203.

[130] Spjøtvoll, E. (1972). On the optimality of some multiple comparison procedures. *Ann. Math. Statist.*, 43:398–411.

[131] Storey, J. and Tibshirani, R. (2003). SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarrays. In *The analysis of gene expression data*, Stat. Biol. Health, pages 272–290. Springer, New York.

[132] Storey, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 64(3):479–498.

[133] Storey, J. D. (2007). The optimal discovery procedure: a new approach to simultaneous significance testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 69(3):347–368.

[134] Storey, J. D., Taylor, J. E., and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 66(1):187–205.

[135] Tamhane, A. C., Liu, W., and Dunnett, C. W. (1998). A generalized step-up-down multiple test procedure. *Canad. J. Statist.*, 26(2):353–363.

[136] Tukey, J. W. (1953). The problem of multiple comparisons. In *The Collected Works of John W. Tukey VIII. Multiple Comparisons: 1948-1983 1-300.* Chapman and Hall, New York.

[137] van de Wiel, M. A., Kim, K., Vosse, S., Van Wieringen, W., Wilting, S., and Ylstra, B. (2006). CGHcall: an algorithm to call aberrations for multiple array CGH tumor profiles. *Bioinformatics*, 23:892–894.

[138] van de Wiel, M. A. and van Wieringen, W. N. (2007). CGHregions: Dimension Reduction for Array CGH Data with Minimal Information Loss. *Cancer Inform*, 3:55–63.

[139] van der Laan, M. J., Dudoit, S., and Pollard, K. S. (2004). Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Stat. Appl. Genet. Mol. Biol.*, 3:Art. 15, 27 pp. (electronic).

[140] van der Vaart, A. W. (1998). *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

[141] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York. With applications to statistics.

[142] Wasserman, L. and Roeder, K. (2006). Weighted hypothesis testing. Technical report, Dept. of statistics, Carnegie Mellon University.

[143] Westfall, P. H. and Young, S. S. (1993). *Resampling-Based Multiple Testing*. Wiley. Examples and Methods for *P*- Value Adjustment.

[144] Zhao, H. and Zhang, J. (2014). Weighted p-value procedures for controlling FDR of grouped hypotheses. *J. Statist. Plann. Inference*, 151/152:90–106.