Contributions to statistical inference from genomic data

Pierre Neuvial

UNIVERSITÉ TOULOUSE III PAUL SABATIER

Document de synthèse

présenté par

Pierre NEUVIAL

en vue de l'obtention de

l'Habilitation à Diriger des Recherches

Spécialité: Mathématiques Appliquées

Contributions à l'inférence statistique pour des données génomiques

---

Soutenue le 16 septembre 2020

devant le jury composé de:

Philippe Berthet	Professeur (Université Paul Sabatier), président
Anne-Laure Boulesteix	Professeur (Ludwig-Maximilians-Universität München), rapportrice
Elisabeth Gassiat	Professeur (Université Paris Saclay), rapportrice
Béatrice Laurent	Professeur (INSA de Toulouse), marraine
Franck Picard	Directeur de Recherche (CNRS), examinateur
Patricia Reynaud-Bouret	Directrice de Recherche (CNRS), examinatrice
Bertrand Thirion	Directeur de Recherche (INRIA), examinateur
Jean-Philippe Vert	Directeur de Recherche (Mines ParisTech et Google Brain), rapporteur

A new class of "high throughput" biomedical devices (...) routinely produce hypothesis-testing data for thousands of cases at once. This is not at all the situation envisioned in the classical frequentist testing theory of Neyman, Pearson, and Fisher.

— Bradley Efron, *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press, 2012.

- Where would you like to live?
- I'd like to live in Theory because "In theory, it works".

— Anonymous quote

Remerciements

Je remercie Anne-Laure Boulesteix, Elisabeth Gassiat, et Jean-Philippe Vert d'avoir accepté de rapporter ce mémoire. Vous êtes, chacun dans votre domaine, des modèles pour moi et je suis honoré que des scientifiques de votre stature jugent et apprécient mes travaux.

Je remercie également à Philippe Berthet, Béatrice Laurent-Bonneau, Franck Picard, Patricia Raynaud-Bouret, et Bertrand Thirion d'avoir accepté de faire partie du jury, et spécialement Béatrice d'avoir parrainé ce mémoire. Franck, Patricia, Bertrand, j'admire votre culture scientifique et l'originalité de vos recherches à l'interface entre disciplines. Philippe et Béatrice, vous êtes la fois des références pour moi en statistique mathématique, et des personnes d'une grande humanité. J'espère avoir un jour prochain l'occasion de travailler avec chacun des membres de ce jury.

Je souhaite remercier (à nouveau !) Emmanuel Barillot et Stéphane Boucheron d'avoir guidé mon travail de thèse, ainsi que Terry Speed et Christophe Ambroise de m'avoir l'un après l'autre accueilli en postdoc. Je salue mes ex-collègues d'Évry où j'ai passé cinq belles années – de l'équipe "Statistique et Génome" au LaMME, en particulier Julien Chiquet, Cyril Dalmasso et Guillem Rigaill.

À Toulouse, je remercie mes collègues de l'Institut de Mathématiques, pour nos discussions scientifiques ou non, et pour les bons moments partagés: je pense notamment à Cathy, Clément, Fanny, François, Gersende, Jean-Marc, Laurent, Manon, Mélisande, Nicolas, Pierre, Philippe (×2), Sébastien (×2), Xavier. À l'INRAE, un merci spécial à Nathalie, pour son accueil et pour m'avoir entraîné dès mon arrivée dans des projets passionnants. Un grand merci aussi à Céline, Delphine, Françoise, Marie-Laure, Nathalie, Nicole, Pierre, Tamara pour leur efficacité et leur soutien administratif ou technique, souvent salvateur.

Je remercie les (ex-)doctorants qui m'ont fait confiance pour leur mettre le pied à l'étrier, me permettant ainsi de découvrir différentes facettes de l'encadrement de thèse: Alia Dehman, Morgane Pierre-Jean, Benjamin Sadacca, Guillermo Durand, Nathanaël Randriamihamison. Un grand merci également aux co-encadrants avec qui j'ai eu le plaisir de partager ces aventures: Christophe Ambroise, Catherine Matias, Fabien Reyal, Etienne Roquain, Nathalie Vialaneix et Marie Chavent.

Je remercie tous mes coauteurs, à qui ce manuscrit doit bien sûr énormément. Un merci spécial à Gilles et Etienne (a.k.a. Blanchard et Roquain) pour nos échanges autour des tests multiples, en particulier dans le cadre du projet ANR SansSouci: j'apprends énormément à vos côtés. Merci aussi à Benjamin, Guillermo, Magali et Marie d'avoir embarqué avec nous pour ce voyage.

Je veux saluer ici trois collègues/amis/coauteurs (pas de mention inutile), côtoyés à Berkeley, et dont j'admire profondément la démarche et la rigueur scientifiques: Henrik Bengtsson, Antoine Chambaz, et Laurent Jacob. Travailler avec vous est à la fois très plaisant et extrêmement stimulant (même conditionnellement au nombre de "double espressos" ingurgités de conserve).

Enfin, je remercie mes proches pour leur soutien, en premier lieu mes parents, et mes beaux-parents. Un clin d'oeil à Romu dont les "coups de fil du lundi" m'ont aidé à trouver l'énergie pour terminer la rédaction de ce document. Agathe, merci d'avoir su parfois (souvent ?) adapter tes contraintes aux miennes, et pour ta capacité à évaluer mieux que moi le temps qui m'est nécessaire. Naël, Lou, Anouk: bien qu'en majorité vous ne croyiez plus au père Nöel, merci de n'avoir jamais douté de moi lorsque j'affirmais terminer la rédaction de ce mémoire "cet été". Merci d'avoir fait de moi un *Happy Daddy Researcher* !

Contents

Chapter 1

Introduction and overview

This chapter provides a brief account of my scientific path, followed by context and motivation for the research described in this manuscript. After a description of the recent transition from hypothesis-driven to data-driven research, we introduce specific types of genomic data and associated biomedical questions, which will serve as a basis for the leading applications in this document. Next, we highlight some challenges and opportunities raised by the analysis of such of high-throughput genomic data. Finally, we give an overview of the contributions described in this document.

Contents

1.1 Foreword: scientific path

My first professional experience was a one-year internship at Crédit Lyonnais in the Groupe de Recherche Opérationnelle (2000-2001), where I enjoyed being involved in methodological research projects while being in contact with their application [R2]. However, this was also the exciting period where the human genome was first sequenced and I decided to dive into the domain of statistics applied to genomics. For my masters thesis project in 2003 I had the opportunity to work with bacterial sequences at the Statistique et Génome lab headed by Bernard Prum in Évry [R1]. After graduating from ENSAE (2003) I obtained a 18 month position as a research engineer in biostatistics in the Bioinformatics group at Institut Curie, which had just been created by Emmanuel Barillot. This experience convinced me of the richness of the statistical questions raised by genomic data analysis. I decided to start a PhD in Statistics (2004-2008; co-supervised by Stéphane Boucheron at Université Paris 7 and Emmanuel Barillot at Institut Curie) in order to find an appropriate balance between theory and application. I was lucky enough to maintain this balance by being hired as a post doc in Terry Speed's group at the Statistics Department of UC Berkeley (2008-2010). When I got back to France I had the opportunity to join the Statistique et Génome lab in Évry for a one-year postdoc. In 2011, I obtained a tenured researcher position at CNRS, and stayed in Évry for five more years. I moved to Toulouse in 2016 where I joined the Institut de Mathématiques de Toulouse.

1.2 From hypothesis-driven to data-driven research

The number and size of available data sets of different types is both a consequence and a cause of scientific and technological breakthroughs. For example, the completion of the human genome sequencing in 2003 by an international consortium triggered the development of high-throughput molecular profiling technologies: microarrays, followed by massively parallel sequencing of group of cells ("bulk" sequencing), and now of individual cells ("single-cell" sequencing); see [35] for a review of sequencing technologies. An emblematic example of large-scale initiative to produce molecular data is the Cancer Genome Atlas (TCGA) project from the US National Cancer Institute. Quoting the web page of TCGA[1],

> *The Cancer Genome Atlas (TCGA), a landmark cancer genomics program, molecularly characterized over 20,000 primary cancer and matched normal samples spanning 33 cancer types. (...) TCGA generated over 2.5 petabytes of genomic, epigenomic, transcriptomic, and proteomic data. The data, which has already lead to improvements in our ability to diagnose, treat, and prevent cancer, will remain publicly available for anyone in the research community to use.*

This "data deluge" has been accompanied by a shift from hypothesis-driven research to data-driven research. The classical statistical approach to data analysis starts by defining a scientific hypothesis, collecting data, and performing inference. In contrast, the current practice in genomics starts by data collection and tends to become technology- or data-driven. Hypotheses are then defined based on this data, and inference is performed on these hypotheses[2]. An important consequence of this change of paradigm is the need for dedicated statistical methods. This point is illustrated in the next sections for the particular case of genomic data.

[1] Source: `https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga`, retrieved on October 8, 2019.

[2] See e.g. the discussion Collect Data First, Ask Questions Later from the podcast "The Effort Report".

1.3 Instances of biomedical questions

In this section we introduce three biomedical contexts and questions, together with associated genomic data. We describe some of the statistical questions raised by the analysis of these data. The leading applications considered in this document will be related to these questions.

1.3.1 Differential expression studies

Differential gene expression studies in cancerology aim at identifying genes whose mean activity differs significantly between two (or more) cancer populations. The activity of a gene is obtained by gene expression measurements from individuals of these populations. These measurements are typically assessed by microarray or sequencing experiments, in which millions of features are collected and summarized in thousands of gene expression measurements. In this document, we specifically consider a Leukemia data set studied in [91]. It consists of expression measurements for $12,625$ genes for biological samples from 79 individuals with B-cell acute lymphoblastic leukemia (ALL) [133]. 37 of these individuals harbor a specific mutation called BCR/ABL. The goal of this study is to understand the molecular differences at the gene expression level between the mutated and non-mutated individuals in the population[3]. Perhaps the most basic question to ask is: for which genes is there a difference in the mean expression level of the mutated and non-mutated population? This question can be addressed, after relevant data preprocessing, by performing a statistical test of equality in means for each gene, and to derive a list of "differentially expressed" genes (DEG) as those passing some significance threshold. This is a typical example of **multiple testing** situation, which requires the definition of dedicated risk measures and associated methods to control these risks.

Other challenges raised by the analysis of gene expression data in cancer samples include the discovery of new cancer subtypes, or the prediction of clinical phenotypes such as survival or the severity of the disease. In statistical terms, these biomedical questions can be cast as problems of **unsupervised or supervised classification**, or **prediction**.

1.3.2 DNA copy number studies

Each normal human cell has 23 pairs of chromosomes. For each of them, one chromosome has been inherited from each biological parent. Tumor cells harbor numerous structural alterations of their DNA including point mutations, translocations, small insertion or deletion events, larger scale copy number changes, or amplifications. Figure 1.1 illustrates the effect of some of these events in the lung tumor cell line NCI-H1395-4W. The left panel shows copy number changes at the scale of the entire genome for one particular cell, while the right panel shows copy number changes at the level of one single chromosome for a large number of cells[4]. In these cells, chromosome 5 has four copies in the first 50 Mb and three copies in the rest of the chromosome.

Such copy number alterations can affect genes and regulatory transcripts, which may result in major cellular modifications and are associated with diagnostic and prognostic factors [165, 82]. An important issue in cancer research is to therefore to estimate the underlying *copy number state* (to be defined more formally in Chapter 9) at each position along the genome of a tumor sample. Microarray and sequencing-based technologies have been used in the last two decades to quantify copy numbers at a large number of genomic loci [102, 168]. For instance, the copy number profile in Figure 1.1 was obtained from Affymetrix Genome-Wide Human SNP 6.0 Arrays, which contain more than 1.8 million genomic markers. In contrast to these "bulk" technologies which provide genomic information averaged

[3]This data set will be used in Chapters 2, 5 and 8.
[4]These data will be used in Chapter 9

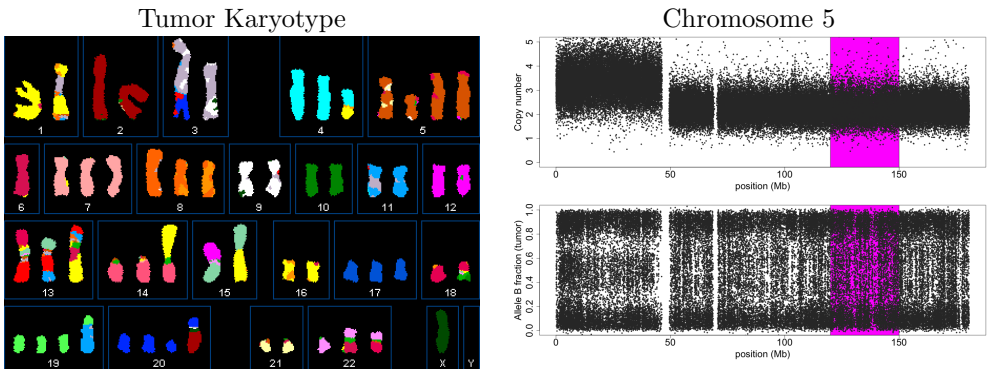


Figure 1.1: DNA copy numbers for the lung tumor cell line NCI-H1395-4W. Left: karyotype. Right: DNA copy number profile of Chromosome 5.

over all cells in a sample, single-cell sequencing technologies have been developed in the last few years to provide genomic information at the level of individual cells [33].

The analysis of DNA copy number data involves several steps. First, these data have to be preprocessed/normalized in order to make data points comparable across loci and possibly across samples. Then, they are segmented into regions of constant DNA copy number state, and these copy number states are labelled ("called") in a biologically-meaningful manner (e.g. normal, loss of one copy, etc). As for gene expression data, DNA copy number data also raise **unsupervised and supervised classification** questions, before or after **segmentation**.

1.3.3 Genome-Wide Association Studies

Genome-Wide Association Studies (GWAS) aim at identifying genomic markers associated with a phenotype of interest. This phenotype may be binary in the case of case-control studies, or continuous (e.g. the date of disease onset). In Chapter 10 we consider a GWA study with 615 patients infected by Human Immunodeficiency Viruses (HIV). One of the goal of this study was to detect genetic factors that influence the plasma viral load, that is, the level of HIV RNA in the patients blood. For each patient, more than $300,000$ genetic markers were assessed by an Illumina genotyping microarray experiment.

Like for differential expression analyses, state-of-the-art approaches start by testing the univariate association between each marker's genotype and the phenotype, and retaining the most significant markers using a multiple testing correction. This list is then interpreted at the genome scale using a *Manhattan plot* (as in Fig. 1.2, left panel, which is reproduced from [103, Figure 1]), where the ($-\log$-transformed) p-values of the markers are plotted against their genomic position in order to identify which genome regions show an enrichment in significantly associated markers. The right panel in Fig. 1.2 (which is reproduced from [103, Figure 1]) shows a measure of statistical dependence (called linkage disequilibrium, LD) between pairs of genomic markers, together with relevant genomic annotation.

Other types of genomic information that can be assessed via high-throughput genomic experiments include:

- DNA methylation, a chemical transformation of one of the four letters of the DNA alphabet that can alter the expression of neighbor genes [27];
- the interaction between a protein of interest and genomic regions, which can be quantified by Chromatin Immuno-Precipitation (ChIP) [37];
- the intensity of physical interaction between two genomic regions, which can be quantified by Chromosome conformation capture techniques such as 3C, 4C, or Hi-C [72]

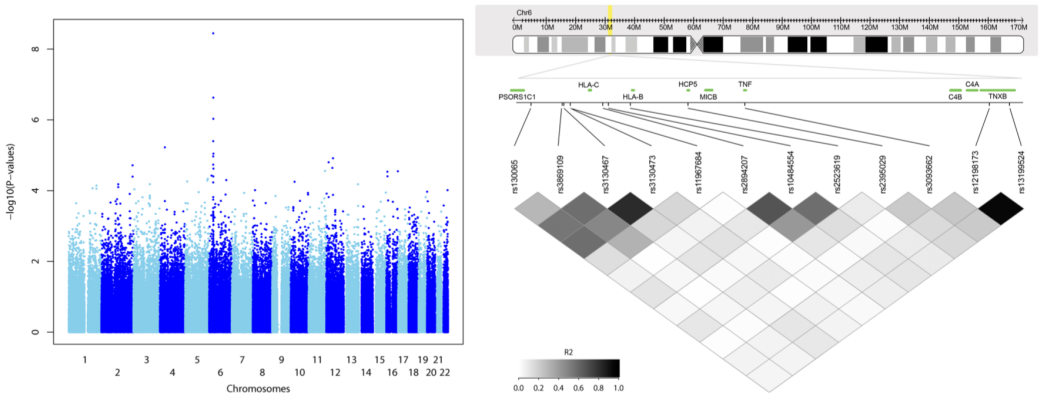


Figure 1.2: Example of a Manhattan plot (left panel) and a linkage disequilibrium plot (right panel) obtained from a GWA study on HIV. Credits: [103, Figure 1 and 2].

1.4 Statistical challenges and opportunities

At first glance, the above-described biomedical questions can be cast as classical statistical problems of testing, prediction, unsupervised and supervised classification, or segmentation. However, genomic data have specificities that imply that classical statistical tools can generally not be readily applied to them, thus requiring new statistical developments. We refer to [42, Chapter 1] for a detailed typology of genomic data. In this section, we focus on specific features of these data – namely, their high-dimensionality, sparsity, heterogeneity and structuration – and try to analyze some of the challenges and opportunities raised by each of these features for statisticians.

1.4.1 High-dimensionality and sparsity

The number p of variables (genes, loci) is typically of the order of 10^3 to 10^6, that is, several orders of magnitude larger than the number n of observations (biological samples, patients), which generally ranges from 1 to 10^3. It is also generally the case that these data are *sparse*, in the sense that only a small (unknown) number of (unknown) variables or of combinations of variables actually contain signal. This high-dimensional setting is in sharp contrast with the classical statistical setting where we generally have $n \gg p$ and always $n > p$. As a consequence, even the most basic statistical tools like linear regression, or statistical tests have to be revisited to cope with this context. High-dimensionality is not specific to the field of genomic data, as it is the consequence of the general sophistication of data acquisition technologies not only in biomedical sciences (genomic and imaging data), but also in other fields such as physics, environmental sciences, economy, finance, to name but a few. A number of mathematical, statistical and computational tools dedicated to high-dimensional data have been developed in the past 20-25 years. Two emblematic examples of these developments are the Least Absolute Shrinkage and Selection Operator (lasso) for penalized regression and variable selection [172] and the False Discovery Rate (FDR) for multiple testing [173][5].

A possibly distinctive feature of genomic data compared to other data types is that each new technological advance results in a increase of one to several orders of magnitude for p, but with n typically remaining constant or even getting temporary smaller because of

[5] According to `scholar.google.com` these papers have been cited 60,000 and 30,000 times, respectively, as of December 2019. Both are among the most cited statistics paper, and Yoav Benjamini has been awarded the 2019 Karl Pearson Prize from the International Statistics Institute for the FDR paper.

the prohibitive cost of new technologies. Moreover, the increase in the number of measured variables is unfortunately accompanied by an increasingly lower signal to noise ratio for each measurement. This can be illustrated with the example of sequencing technologies (from bulk to single cell sequencing) getting at the same time wider (larger p), sparser and noisier, see e.g. [83] whose provocative title is: "Sequencing technology does not eliminate biological variability".

Curse or blessing? The high-dimensional nature of genomic data is considered as a blessing by biomedical scientists (following the possibly misleading argument that "more data" implies "more knowledge"), and as a curse by mathematicians as nicely explained and illustrated in [46, Chapter 1] from several standpoints. In particular, n points in p-dimensional space with $p \gg n$ are typically isolated. This makes the notion of "neighborhood", which is fundamental in classical statistical inference, essentially useless in this setting. Moreover, the number of candidate models to explain a given phenomenon increases exponentially with the number of parameters of this model, which is typically indexed by p.

Fortunately, the sparsity of genomic data may also be seen as a blessing, because it implies that the biological signal is concentrated in a much lower-dimensional space than the p-dimensional observation space. For example, a typical assumption in prediction or variable selection tasks is that only a small number of variables, or of groups of variables, are actually relevant. Such assumptions are the basis of the methods recently developed for high-dimensional (genomic) data analysis. They are also assumptions under which statistical guarantees for such methods and efficient algorithms can be obtained, see e.g. [98, 46].

1.4.2 Heterogeneity and structure

In the classical statistical framework, inference is generally performed on a single $n \times p$ table. By *heterogeneous*, we mean the situation where several levels of biological information are available for the same set of observations. For example, a tumor biopsy can be analyzed to study mutations, DNA copy numbers, DNA methylation, gene or protein expression, or gene regulation. Moreover, these genomic data can also be complemented by other data types such as imaging data (e.g. phenotyping experiments, or histology) or clinical data coming from electronic health records (EHR). This setting is called multi-view data in the machine learning literature, see e.g. [71] for a survey.

Moreover, the measured variables are often *structured*, that is, linked by networks, by similarity relationships, or simply by a natural ordering along a chromosome. This may be illustrated by the above examples: in differential expression studies, sets of genes can be co-differentially expressed because they belong to the same gene network or pathway; in DNA copy number studies, neighboring loci on a chromosome are expected to have identical copy number state (as in Figure 1.1, right); in GWAS, a SNPs can have its genotype associated to a phenotype because its belongs to the same block of LD (see Figure 1.2) as a causal SNP.

Curse or blessing? The above-described characteristics of genomic data represent a challenge for the statistician, because standard statistical tools are likely not to be suited to address a new biomedical question, and because devising a tailored method requires some basic understanding of the question at hand. I believe that a relevant way to tackle this complexity is precisely to take advantage of the heterogeneity and structuration of genomic data by considering them as constraints that alleviate the curse of dimensionality, and guide statistical methods toward solutions that are more plausible from a biological standpoint.

1.5 Overview of contributions

In the above-described context, my research objective has been to contribute to the development, mathematical study, and practical application in interdisciplinary research projects of statistical tools that take advantage of the complexity and structure of genomic data. My contributions generally result of the quest for an acceptable trade-off between mathematical rigor, computational efficiency, and biological interpretability. It is often the case that the intersection between what is mathematically justified, algorithmically feasible, and biologically interpretable is empty for a given problem. In such situations, I strive to make one or more of these three sets bigger, or the gap between these objectives smaller. Depending on the scientific question and the state of the art for a given problem, my contributions can be of different types:

- proposing new statistical methods, associated algorithms and their implementation
- establishing statistical properties for new or existing methods
- evaluating the statistical and computational performance of new or existing methods

Organization of the manuscript. The remainder of this manuscript is organized in two main parts.

Part I summarizes my contributions to the field of multiple testing, and post-selection or post hoc inference. After an introduction to multiple testing (Chapter 2), I describe my contributions to the asymptotic properties of FDR controlling procedures (Chapter 3), to FDR thresholding for classification under sparsity assumptions (Chapter 4) and to post hoc inference (Chapter 5)

Part II gathers my other contributions to statistical inference for genomic data. As explained in a short introductory chapter (Chapter 6), which can be categorized in two broad themes:

- Inference from *heterogeneous* genomic data, where inference is performed by combining several data types, either corresponding to the same observations (Chapter 7) or to the same variables (Chapter 8);
- Inference from *ordered* genomic data, where one of the main statistical challenges is to segment a genome (or more precisely each of its chromosomes) into successive homogeneous regions. This is tackled via segmentation methods for DNA copy numbers (Chapter 9) and by constrained clustering methods for GWAS and Hi-C studies (Chapter 10).

Finally, some directions for future research are discussed in Chapter 11.

Part I

From multiple testing to post hoc inference

Chapter 2

Introduction to multiple testing

Significance testing was introduced in the early 1900s with the works of Fisher, Neyman and Pearson. The examples of differential expression analyses and GWAS taken in Chapter 1 require thousands to millions of statistical tests to be performed simultaneously. In such situations, it is desirable to control a global risk measure associated to the entire set of tested hypotheses. Large-scale multiple testing is concerned with the definition of such risk measures, the formalization of mathematical assumptions under which these risks are effectively controlled, and the construction of dedicated algorithms (called multiple testing procedures). This chapter provides an overview of large-scale multiple testing theory, with emphasis on concepts that will be useful to the reader for the next chapters of this part. This chapter draws from several reference books or surveys on multiple testing [54, 59, 60, 86, 128].

Remark: Throughout this part, the number of hypotheses tested is denoted by m whereas in Part II the number of variables will be denoted by p.

Contents

2.1 Statistical setting

We consider an observation $x \in \mathcal{X}$, that is, a realization of a random variable X valued in a measurable space $(\mathcal{X}, \mathfrak{X})$. Our statistical model is a family $\mathcal{P}$ of candidate probability distributions for the true distribution $\mathbb{P}$ of X on $(\mathcal{X}, \mathfrak{X})$. Given a subset $\mathcal{P}_0$ of $\mathcal{P}$, we consider the *null hypothesis* $H_0 : \mathbb{P} \in \mathcal{P}_0$. The corresponding *alternative hypothesis* is $H_1 = \mathbb{P} \in \mathcal{P} \setminus \mathcal{P}_0$ unless otherwise noted. Given an observation $x \in \mathcal{X}$, we aim at deciding whether $\mathbb{P}$ is compatible with the null hypothesis, that is, whether $\mathbb{P} \in \mathcal{P}_0$. In order to do so, following e.g. [156], we define a p-value as a random variable $p(X)$ on $[0,1]$ satisfying[1]:

$$\forall \mathbb{P} \in \mathcal{P}_0, \forall u \in [0,1],\ \mathbb{P}(p(X) \leq u) \leq u\,. \tag{2.1}$$

By construction, the p-value $p(x)$ is a quantitative measure of how unlikely observation x is if H_0 is true. This measure can be used to take a decision as to whether H_0 is true or not. Given $\alpha \in [0,1]$, a test of level α "rejects" H_0 if and only if $p(x) \leq \alpha$. In order to compare the decision of rejection/acceptance of the null hypothesis to the truth, we introduce the classical vocabulary of true/false positive/negative in Table 2.1.

	H_0 not rejected	H_0 rejected
H_0 true	true negative	false positive
H_0 false	false negative	true positive

Table 2.1: Qualification of the four possible outcomes of a statistical test.

2.1.1 Multiple testing setting

For $m \in \mathbb{N}$, we consider a collection of subsets $\mathcal{P}_{0,i}$ of $\mathcal{P}$ and the associated null hypotheses: $H_{0,i} : \mathbb{P} \in \mathcal{P}_{0,i}$ indexed by $\mathbb{N}_m = \{1, \ldots, m\}$. The corresponding alternative hypotheses are $H_{1,i} = \mathbb{P} \in \mathcal{P} \setminus H_{0,i}$. each $i \in \mathbb{N}_m$. From the collection $\mathcal{H} = (H_{0,i})_{i \in \mathbb{N}_m}$, we define a m-dimensional parameter of interest: $(\theta_i(\mathbb{P}))_{i \in \mathbb{N}_m}$, where $\theta_i(\mathbb{P}) = \mathbf{1}\{\mathbb{P} \in H_{0,i}\}$ for $i \in \mathbb{N}_m$. The sets of true null hypotheses and true alternative hypotheses are denoted by $\mathcal{H}_0(\mathbb{P})$ and $\mathcal{H}_1(\mathbb{P})$, respectively. When non ambiguous, we omit $\mathbb{P}$ in the notation and write θ_i, $\mathcal{H}_0$ and $\mathcal{H}_1$ for simplicity. We let $m_0(m) = \sum_{i \in \mathbb{N}_m}(1 - \theta_i) = |\mathcal{H}_0|$ and $m_1(m) = \sum_{i \in \mathbb{N}_m} \theta_i = |\mathcal{H}_1|$ be the number of true null hypotheses and of true alternative hypotheses, respectively. Finally, $\pi_0(m) = m_0(m)/m$ is the proportion of true null hypotheses.

Location model. We introduce a standard location model which will be used several times for illustration:

$$X_i = \mu_i + \varepsilon_i, \quad i \in \mathbb{N}_m\,, \tag{2.2}$$

where the ε_i are identically distributed, with a common marginal distribution that is assumed to be continuous, and the location parameter μ_i is null if and only if $\theta_i = 0$. A location model is characterized by the distribution of the test statistics conditionally on $\theta_i = 0$. The most common instances of location models are the Gaussian and Laplace (double exponential) models. Both are particular instances of the Subbotin location model, where the density of the test statistics conditionally on $\theta_i = 0$ is given by the ζ-Subbotin density:

$$d(x) = (L_\zeta)^{-1} e^{-|x|^\zeta/\zeta}, \text{ with } L_\zeta = \int_{-\infty}^{+\infty} e^{-|x|^\zeta/\zeta} dx = 2\Gamma(1/\zeta)\zeta^{1/\zeta - 1}\,, \tag{2.3}$$

for $\zeta \geq 1$. The Gaussian and Laplace models correspond to $\zeta = 2$ and $\zeta = 1$, respectively. In Section 4 we will also consider scaling models based on the Subbotin density.

[1] For notation simplicity, we also follow [156] in using the same letter $\mathbb{P}$ for probability measures on the observation space $(\mathcal{X}, \mathfrak{X})$ and on the implicitly defined domain of definition of X.

Multiple testing procedures. We define a multiple testing procedure as a mapping from an observation X to a subset of $\mathbb{N}_m$, corresponding to a set R_m of rejected hypotheses. For notation simplicity we will identify a multiple testing procedure with the associated rejection set R_m. We denote by $|R_m|$ the corresponding number of rejections, and often omit the subscript m when m is fixed. We assume to be given a set of p-values associated to $H_{0,i}$ for each $i \in \mathbb{N}_m$, that is, that there exists a random variable $p_i(X)$ on $[0,1]$ satisfying:

$$\forall \mathbb{P} \in \mathcal{P}_{0,i},\ \forall u \in [0,1], \mathbb{P}(p_i(X) \leq u) \leq u\,. \tag{2.4}$$

We consider multiple testing procedures that reject the hypotheses whose p-value is less than a (possibly random and data-driven) threshold. Such *p-value thresholding-based multiple testing procedures* (or thresholding procedures for short) may be written as

$$R_m = \{i \in \mathbb{N}_m,\ p_i \leq \hat{t}_m\}\,,$$

where $\hat{t}_m$ is called the threshold of the multiple testing procedure. More precisely, we consider three classes of procedures. In "single-step" procedures, the rejection of a given hypothesis does not depend on the rejection of other hypotheses. In contrast, "step-up" and "step-down" procedures are part of the broader class of stepwise multiple testing procedures, where the rejection of a given hypothesis may depend on the rejection of other hypotheses. Let us denote by $(p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)})$ the ordered p-values.

Definition 2.1 (Step-up and step-down multiple testing procedures)**.** *Let* $\mathbf{c} = (c_i)_{i \in \mathbb{N}_m}$ *be a non-decreasing set of values in* $[0,1]$.

- *The step-up procedure with critical values* $\mathbf{c}$ *is the procedure with threshold* $\hat{t} = c_{\hat{\imath}^{\uparrow}(\mathbf{c})}$, *where*
$$\hat{\imath}^{\uparrow}(\mathbf{c}) = \max\left\{i \in \mathbb{N}_m, p_{(i)} \leq c_i\right\}\,. \tag{2.5}$$
- *The step-down procedure with critical values* $\mathbf{c}$ *is the procedure with threshold* $\hat{t} = c_{\hat{\imath}^{\downarrow}(\mathbf{c})}$, *where*
$$\hat{\imath}^{\downarrow}(\mathbf{c}) = \max\left\{i \in \mathbb{N}_m, \forall j \leq i, p_{(j)} \leq c_j\right\}\,. \tag{2.6}$$

The threshold of a step-down procedure is the first crossing point between the ordered p-values and $\mathbf{c}$, while the threshold of a step-up procedure is the last such crossing point. By definition, for a fixed family of critical values $\mathbf{c}$, the associated step-up procedure rejects at least as many hypotheses than the associated step-down procedure.

2.1.2 Multiple testing risks

The (unobserved) number of false positives of a multiple testing procedure is $|R_m \cap \mathcal{H}_0|$, which implicitly depends on the distribution $\mathbb{P}$ of X through $\mathcal{H}_0$. Historically, the first risk measure considered in a multiple testing context is the Family-Wise Error Rate (FWER). It is defined as the probability of (at least) one false rejection:

$$\mathrm{FWER}_{\mathbb{P}}(R_m) = \mathbb{P}(|R_m \cap \mathcal{H}_0| > 0)\,.$$

A natural generalization is $\text{k-FWER}_{\mathbb{P}}(R_m) = \mathbb{P}(|R_m \cap \mathcal{H}_0| \geq k)$ which allows at most k false rejections. A multiple testing procedure R_m is said to control k-FWER (strongly) at level $\alpha \in [0,1]$ if $\text{k-FWER}_{\mathbb{P}}(R_m) \leq \alpha$ for all $\mathbb{P} \in \mathcal{P}$. A much less demanding criterion called weak k-FWER control consists in ensuring that $\text{k-FWER}_{\mathbb{P}}(R_m) \leq \alpha$ for all $\mathbb{P} \in \cap_{i \in \mathbb{N}_m} \mathcal{P}_{0,i}$. In other words, a weak k-FWER controlling procedure simply corresponds to a test of level α of the global null hypothesis $\cap_{i \in \mathbb{N}_m} H_{0,i}$. Here, we focus on strong control. Another quantity of interest is the False Discovery Proportion (FDP), which is defined as the fraction of true null hypotheses among those rejected:

$$\mathrm{FDP}_{\mathbb{P}}(R_m) = \frac{|R_m \cap \mathcal{H}_0|}{|R_m| \vee 1}\,.$$

As the FDP is a random quantity, one possibility to define an associated risk is to focus on its expectation under $\mathbb{P}$, which is known as the False Discovery Rate (FDR), or its quantiles under $\mathbb{P}$ (see the FDX below). The FDR is defined by

$$\mathrm{FDR}_{\mathbb{P}}(R_m) = \mathbb{E}\left[\frac{|R_m \cap \mathcal{H}_0|}{|R_m| \vee 1}\right],$$

and a multiple testing procedure R_m is said to control FDR (strongly) at level $\alpha \in [0,1]$ if $\mathrm{FDR}_{\mathbb{P}}(R_m) \leq \alpha$ for all $\mathbb{P} \in \mathcal{P}$. A related quantity is the positive FDR [154], defined by

$$\mathrm{pFDR}_{\mathbb{P}}(R_m) = \mathbb{E}\left[\frac{|R_m \cap \mathcal{H}_0|}{|R_m|}\Big||R_m| > 0\right].$$

Finally, we also define the False Discovery Exceedance [125] as the tail probability of the FDP for a given $q \in [0,1]$:

$$\mathrm{FDX}(R_m) = \mathbb{P}\left[\frac{|R_m \cap \mathcal{H}_0|}{|R_m| \vee 1} \geq q\right].$$

Naturally, (strong) pFDR and FDX control can be defined similarly as for FDR.

2.1.3 Assumptions on the joint p-value distribution

We consider several types of assumptions on the joint distribution of the p-value family (or, equivalently, on the set $\mathcal{P}$ of possible distributions for $\mathbb{P}$). We introduce the notation $p_{\mathcal{A}} = (p_i)_{i \in \mathcal{A}}$ for $\mathcal{A} \subset \mathbb{N}_m$. Two extreme distributional assumptions are general dependence, where the p-value distribution is let arbitrary, and independence:

$$p_{\mathcal{H}(\mathbb{P})} \text{ is a family of mutually independent variables} \tag{indep}$$

Remark 2.2. As the statistical risks defined above only focus on type I errors (false positives), any result stated below under (indep) is also valid under the weaker assumption that $p_{\mathcal{H}_0(\mathbb{P})}$ is a family of mutually independent variables, and is also independent from $p_{\mathcal{H}_1(\mathbb{P})}$. In practice, $\mathcal{H}_0$ is unknown so we chose to use Assumption (indep) for simplicity.

Assumption (indep) is useful in theory because it simplifies the statistical study of multiple testing risks, but it is unrealistic in applications. A weaker assumption is Positive regression dependency on a subset of hypotheses (PRDS)[2]. The set $S \subset [0,1]^m$ is *non-decreasing* if for all $(q, q') \in ([0,1]^m)^2$ such that $\forall i \in \mathbb{N}_m, q_i \leq q'_i$, we have: $q \in S$ implies $q' \in S$. We assume that the p-value family is PRDS on the subset $\mathcal{H}_0(\mathbb{P})$ of true null hypotheses, that is:

$$\begin{cases} \text{for any } i_0 \in \mathcal{H}_0(\mathbb{P}) \text{ and any non-decreasing set } S \subset [0,1]^m, \\ \text{the function } u \mapsto \mathbb{P}((p_i)_{i \in \mathbb{N}_m} \in S | p_{i_0} \leq u) \text{ is non-decreasing} \end{cases} \tag{PRDS($\mathcal{H}_0$)}$$

Assumption (PRDS($\mathcal{H}_0$)) is weaker than independence, and it is considered as realistic in genomics [60], although it is generally not possible to check whether it holds for a particular application. A classical example of PRDS distribution of the p-value family is the following equi-correlated model:

Example 2.7. The test statistics are multivariate Gaussian, with a covariance matrix whose non-diagonal entries are all equal to $\rho \in [0,1]$.

[2]The results stated in this chapter under PRDS in fact hold for a slightly larger class of distribution called weak PDRS. We refer to [54] for references on this distinction, which is not essential for our purpose.

More generally, Assumption (PRDS($\mathcal{H}_0$)) holds whenever the test statistics are multivariate Gaussian, with a covariance matrix Σ whose entries are all non-negative. Example 2.7 will be used repeatedly for illustration in this part.

The main reason for the popularity of the PRDS assumption in multiple testing is that the Simes inequality [179] is valid under this assumption. We recall this inequality, together with Hommel's inequality [181], a more conservative inequality valid under general dependence. We denote by $p_{(1:I)} \leq \cdots \leq p_{(|I|:I)}$ the ordered p-values of I.

Proposition 2.3 (Simes and Hommel's inequalities)**.** *We have:*

$$\mathbb{P}\left(\exists i \in \mathcal{H}_0 \,:\, p_{(i:\mathcal{H}_0)} \leq \frac{\alpha i}{m}\right) \leq \frac{\alpha|\mathcal{H}_0|}{m} c(m)\,, \tag{2.8}$$

in the two following cases:

- $c(m) = 1$ *and the p-value family satisfies* (PRDS($\mathcal{H}_0$)) *(Simes' inequality);*
- $c(m) = C(m) := \sum_{i=1}^{m} 1/i$ *(Hommel's inequality).*

Moreover (2.8) *is an equality (with $c(m) = 1$) when the p_i, $i \in \mathcal{H}_0(P)$, are i.i.d. $U(0,1)$.*

In particular, the right-hand side of (2.8) is upper bounded by $\alpha c(m)$. The family of thresholds $(\alpha i/m)_{1\leq i\leq m}$ is called the Simes threshold family. The factor $C(m)$ is of the order of $\log(m)$ and quantifies "the price to pay" for allowing for general dependence. The last statement in the above Proposition illustrates the sharpness of the Simes inequality, in the sense that they cannot be uniformly improved on the class of PRDS distribution. The same holds for Hommel's and for arbitrary distributions, respectively[3]. The Simes inequality will is also the fundamental ingredient of the original post hoc procedures proposed by [81], as will be discussed in Section 5.

2.2 Family-Wise Error Rate control

FWER control under general dependence can be obtained using a simple union bound argument, which is sometimes referred to as the Bonferroni inequality. For any $I \subset \mathbb{N}_m$ and any $t > 0$,

$$\mathbb{P}\left(\exists i \in I, p_i \leq t\right) \leq \sum_{i\in I} \mathbb{P}\left(p_i \leq t\right) \leq t|I| \tag{2.9}$$

where the last inequality follows from the p-value property. Applying this inequality with $I = \mathcal{H}_0$ and $t = \alpha/m$ ensures that

$$\mathbb{P}\left(\exists i \in \mathcal{H}_0, p_i \leq \alpha/m\right) \leq \frac{m_0}{m}\alpha \leq \alpha \tag{2.10}$$

The left-hand side is the FWER of the Bonferroni procedure with threshold $t^{\mathrm{Bonf}} = \alpha/m$. Equation 2.10 demonstrates that the Bonferroni procedure controls FWER under general dependence, but also that it is conservative, in the sense that it controls FWER at level $\pi_0\alpha$ for a target level α. Therefore, it is possible to obtain more powerful FWER controlling procedures by bridging the gap between $\pi_0\alpha$ and α. The (Bonferroni-)Holm procedure [182] is a stepwise modification of the Bonferroni procedure which also controls FWER under general dependence, but possibly with an increased number of rejections. The Holm procedure is defined as the step-down procedure with critical values:

$$c_i = \frac{\alpha}{m-(i-1)}, \quad i \in \mathbb{N}_m\,. \tag{2.11}$$

[3]However, both of them can also be seen as conservative, as discussed and illustrated in Section 5.4.

Another way to construct FWER controlling procedures that are less conservative than the Bonferroni procedure is to make dependency assumptions on the family of p-values. For completeness we mention that under independence, FWER is controlled by the Šidák procedure [191], whose threshold is defined as $t^{\text{Sidak}} = 1-(1-\alpha)^{1/m}$. Because $1-(1-\alpha)^{1/m} \geq \alpha/m$, the Šidák procedure is less conservative than the Bonferroni procedure. However, when m is large, both thresholds are equivalent when α/m is small, so the Šidák procedure may only improve the number of rejections marginally, at the price of a much narrower applicability. Under (PRDS($\mathcal{H}_0$)), the FWER is controlled by the step-up procedure with the same critical values as the Holm procedure (2.11), which is called the Hochberg procedure. The decreased conservativeness of the Hochberg procedure compared to the Holm procedure comes at the price of a narrower applicability.

2.2.1 Connection to closed testing

The Holm and Hochberg procedures can be derived from a general method for obtaining FWER control, known as closed testing [186]. This method can also be viewed as the basis of the post hoc procedures proposed by [81], that inspired our work [J2] described in Section 5. Closed testing focuses on the larger multiple testing problem of testing *all possible intersections* of null hypotheses. Formally, let us define the closure of the family of hypotheses $\mathcal{H} = (H_{0,i})_{i\in\mathbb{N}_m}$ as the set $\bar{\mathcal{H}} = (H_I)_{I\subset\mathbb{N}_m, I\neq\emptyset}$, where H_I denotes the intersection hypothesis associated with I:

$$H_I = \bigcap_{i\in I} H_{0,i}\,.$$

We denote for short by $\bar{\mathbb{N}}_m = \mathcal{P}(\mathbb{N}_m)\setminus\{\emptyset\}$ the set of all non empty subsets of $\mathbb{N}_m$. The set of true null intersection hypotheses is then $\bar{\mathcal{H}}_0 = \left\{I \in \bar{\mathbb{N}}_m, H_I \text{ is true}\right\}$. Since $H_{\{i\}} = H_{0,i}$ for all i, we can write[4] $\mathcal{H} \subset \bar{\mathcal{H}}$ and $\mathcal{H}_0 \subset \bar{\mathcal{H}}_0 \subset \mathcal{P}(\mathcal{H}_0)$. Assume that a test $\phi_I \in \{0,1\}$ of H_I is available for any possible intersection hypothesis I, where $\phi_I = 1$ if and only if H_I is rejected. Such a test is called a *local test* of H_I. The closed testing procedure associated to the collection $(\phi_I)_{I\in\bar{\mathbb{N}}_m}$ is a multiple testing procedure (of size $\left|\bar{\mathbb{N}}_m\right| = 2^m-1$ instead of $|\mathbb{N}_m| = m$) for the closure $\bar{\mathcal{H}}$, which rejects

$$\bar{R} = \left\{I \in \bar{\mathbb{N}}_m, \forall J \supset I, \phi_J = 1\right\}. \tag{2.12}$$

A fundamental property of closed testing is given in the next Proposition.

Proposition 2.4. *If for all $I \in \bar{\mathbb{N}}_m$, the local test ϕ_I is a test of level α of H_I, then the closed testing procedure associated to the collection $(\phi_I)_{I\in\bar{\mathbb{N}}_m}$ controls FWER in $\bar{\mathcal{H}}$ at level α, that is,*

$$\mathbb{P}\left(\left|\bar{R}\cap\bar{\mathcal{H}}_0\right| > 0\right) \leq \alpha$$

As $\mathcal{H} \subset \bar{\mathcal{H}}$, $\bar{R}\cap\mathcal{H}$ can be seen as a multiple testing procedure for $\mathcal{H}$. By Proposition 2.4, this procedure controls FWER (in $\mathcal{H}$) at level α. To see why Proposition 2.4 holds, simply note that as $\mathcal{H}_0 \in \bar{\mathbb{N}}_m$, $\phi_{\mathcal{H}_0}$ is an α-level test of $H_{\mathcal{H}_0}$. Therefore, there exists an event of probability larger than $1-\alpha$ under which $H_{\mathcal{H}_0}$ is not rejected by the closed testing procedure. Under this event, any true null intersection hypothesis, which is by definition a subset of $\mathcal{H}_0$, is also not rejected by the closed testing procedure.

The closed testing method thus provides an elegant and generic construction of FWER-controlling procedures, where the properties of the procedure are inherited from the properties of the local tests. Closed testing can itself be seen as a consequence of the sequential rejection principle [93], which provides generic sufficient conditions to build stepwise FWER controlling procedures. A caveat to the practical application of closed testing is that it implies testing all 2^m-1 possible non-empty intersections between m hypotheses. However, depending on the form of the local test, it may be possible to avoid testing all 2^m-1 hypotheses explicitly. This is the case in the procedures listed in the next paragraph.

[4] Formally we should write $\forall i \in \mathcal{H}, \{i\} \in \bar{\mathcal{H}}$ but we identify i with $\{i\}$ in order to alleviate notation.

The closed testing principle may be used to recover some of the procedures described above. First, the Holm procedure coincides with the closed testing procedure associated with local Bonferroni tests, which are defined by $\phi_I = \mathbf{1}\{\forall i \in I, p_i \leq \alpha/|I|\}$ for the null hypothesis H_I. These local tests are of level α under general dependence by (2.9), so the closed testing principle ensures that the Holm procedure controls FWER under general dependence. Similarly, the Hochberg procedure can be seen as a closed testing procedure associated with local Simes tests, which are defined by $\phi_I = \mathbf{1}\left\{\forall i \in I, p_{(i)} \leq \alpha i/|I|\right\}$. These tests are of level α under (PRDS($\mathcal{H}_0$)) by (2.8), so the closed testing principle ensures that the Hochberg procedure controls FWER under (PRDS($\mathcal{H}_0$))[5]. The procedures described in this section are summarized in Table 2.2. Although the Holm procedure is always at least as powerful as the Bonferroni procedure, the latter is much more widely used, for example in bio-medical applications.

Setting	Procedures
General dependence	Bonferroni $\ll$ Holm
PRDS($\mathcal{H}_0$)	Hochberg $\ll$ Hommel
indep	Šidák

Table 2.2: Summary of FWER-controlling procedures

2.2.2 Adaptivity to dependence via randomization

The above-defined FWER controlling procedures rely on one of the assumptions on the joint distribution of the p-values formulated in Section 2.1.3, namely general dependence, (PRDS($\mathcal{H}_0$)) or (indep). By construction, these procedures are not *adaptive* to the dependency structure at hand in a specific context. To illustrate this point let us reformulate the FWER of a thresholding procedure $R = \{i \in \mathbb{N}_m, p_i \leq \hat{t}\}$:

$$\begin{aligned}\mathrm{FWER}_{\mathbb{P}}(R) &= \mathbb{P}\left(\exists i \in \mathcal{H}_0(\mathbb{P}), p_i \leq \hat{t}\right) \\ &= \mathbb{P}\left(\inf_{i \in \mathcal{H}_0(\mathbb{P})} p_i \leq \hat{t}\right).\end{aligned}$$

Denoting by $q_\alpha(\mathcal{A})$ the $(1-\alpha)$-quantile of the distribution of $\inf\{p_i : i \in \mathcal{A}\}$, an optimal choice is $\hat{t} = q_\alpha(\mathcal{H}_0(\mathbb{P}))$. Using this formulation, the above-described FWER-controlling procedures provide lower bounds for this quantile under specific dependency assumptions. Rather than making such assumptions, Westfall and Young [174] proposed to use procedures based on permutation, called `minP` and `maxT`, to build adaptive bounds for $q_\alpha(\mathbb{N}_m)$, where $q_\alpha(\mathbb{N}_m) \leq q_\alpha(\mathcal{H}_0(\mathbb{P}))$. They also introduced step-down versions of these procedures.

Romano and Wolf [140] introduced a randomization assumption under which these procedures are proved to control FWER, bypassing the need for a technical condition called subset pivotality as in the original results of Westfall and Young [174]. Here, we use a recent formulation of this randomization assumption due to Hemerik and Goeman [18], which is slightly weaker than the assumption of Romano and Wolf [140]. Specifically, we assume that there exists a finite group of transformations $\mathcal{G}$ acting onto the observation space $\mathcal{X}$, is such a way that the joint distribution of the transformed null p-values is invariant under the action of any $g \in \mathcal{G}$. Formally,

$$\forall P \in \mathcal{P},\ \forall g \in \mathcal{G},\ \left(p_{\mathcal{H}_0}(g'.X)\right)_{g' \in \mathcal{G}} \sim \left(p_{\mathcal{H}_0}(g'.g.X)\right)_{g' \in \mathcal{G}}, \tag{Rand}$$

where $g.X$ denotes X that has been transformed by g. We refer to [18] for examples of situations where (Rand) holds. Examples based on sign-flipping in the location model (2.2) and permutation testing for two-sample tests are given in Section 5.

[5]The closed testing principle may also be used to prove the FWER control of the Hommel procedure [177] under (PRDS($\mathcal{H}_0$)). This procedure is slightly more powerful but also more complicated to define than the Hochberg procedure, see [59, Definition 5.4].

2.2.3 Extension to k-FWER control

We refer to [86, 137] for a survey of k-FWER controlling procedures. In particular, [137] provides single step and step-down generalizations of the adaptive FWER controlling procedures under (Rand) to k-FWER control. Here, we simply note that a natural generalization of the Bonferroni procedure to k-FWER control may be obtained by elementary arguments. Indeed, the k-FWER of the procedure with threshold t may be written as

$$\begin{aligned}\mathbb{P}\left(\sum_{i\in\mathcal{H}_0}\mathbf{1}\left\{p_i\le t\right\}\ge k\right) &\le k^{-1}\mathbb{E}\left(\sum_{i\in\mathcal{H}_0}\mathbf{1}\left\{p_i\le t\right\}\right)\\ &= k^{-1}\sum_{i\in\mathcal{H}_0}\mathbb{P}\left(p_i\le t\right)\\ &\le \frac{|\mathcal{H}_0|}{k}t\,,\end{aligned}$$

where the first inequality holds by Markov's inequality, and the last inequality holds by the definition of the p-values. Therefore, the "Generalized Bonferroni" procedure with threshold $t^{k-\mathrm{Bonf}}=\alpha k/m$ controls k-FWER under general dependence. Similar to the improvement of the Bonferroni procedure for FWER control by the Holm procedure, it is possible to obtain a more powerful k-FWER controlling procedure by building a step-down procedure with critical values:

$$\frac{\alpha k}{m-(i-k)\mathbf{1}\left\{i>k\right\}}\,.$$

Just as for FWER control, the threshold of this procedure depends on the p-values, contrary to the deterministic threshold of the generalized Bonferroni procedure.

2.3 False Discovery Rate control

2.3.1 FDR control by the BH procedure

Benjamini and Hochberg [173] have proposed to use the step-up procedure associated with the Simes critical values $(\alpha i/m)_{1\le i\le m}$ in order to control FDR. The threshold of the Benjamini-Hochberg procedure at level α (or BH(α)) is thus defined as $\hat{t}^{\mathrm{BH}}(\alpha)=\alpha\hat{\imath}^{\mathrm{BH}}(\alpha)/m$, where

$$\hat{\imath}^{\mathrm{BH}}(\alpha)=\max\left\{i\in\mathbb{N}_m, p_{(i)}\le\frac{\alpha i}{m}\right\}\,. \tag{2.13}$$

The BH procedure controls FDR at level α under independence [173]. Benjamini and Yekutieli [161] later proved that the BH(α) procedure controls FDR at level $\pi_0\alpha$ under (PRDS($\mathcal{H}_0$)), which is a much weaker assumption than (indep). Moreover, the BH($\alpha/C(m)$) procedure controls FDR at level $\pi_0\alpha$ under general dependence [161], where $C(m)$ is Hommel's correction factor for dependence defined in Proposition 2.3. This procedure is called the BY procedure at level α. Although the BH procedure is based on Simes critical values, the results obtained in [161] are not a consequence of the Simes inequality. Sarkar et al. [108] noted that the validity of the Simes inequality is in fact a necessary condition for the BH procedure to control FDR.

2.3.2 Estimation of the proportion of true null hypotheses

The FWER control by the Bonferroni procedure and the control offered by the Simes or Hommel inequalities are based on the behavior on the true null hypotheses, and therefore imply the unknown factor π_0. Similarly, the FDR control provided by the BH procedure at the target level α is conservative by a factor $\pi_0\le 1$. Several procedures have been

proposed to fill the "conservativeness gap" between FDR control at $\pi_0\alpha$ and α. Such procedures incorporate an explicit or implicit estimation of the proportion π_0 of true null hypotheses [123, 97, 149, 151, 160]. One important class of such procedures is plug-in procedures, which apply the BH procedure at level $\alpha/\hat{\pi_0}$, where $\hat{\pi_0}$ is an estimator of π_0. We give the definition of one of the most widely used estimator of π_0 in applications is the Storey-λ estimator [151, 160]:

$$\hat{\pi}_0(\lambda) = \frac{1 + 1/m - \widehat{\mathbb{G}}_m(\lambda)}{1 - \lambda}\,. \tag{2.14}$$

The FDR control of plug-in procedures such as the Storey-λ typically only holds under (indep). Some asymptotic properties of such plug-in procedures under (indep) are studied in Section 3.

2.3.3 FDR control vs FWER control

As the FDP is between 0 and 1, we have FDR $= \mathbb{E}(\text{FDP}) \leq \mathbb{P}(\text{FDP} > 0) =$ FWER. Therefore, any procedure controlling FWER at a given level also controls FDR at the same level. In the multiple testing literature, FDR controlling procedures are often said to be "more powerful" than FWER controlling procedures, because FDR controlling procedures generally reject more hypotheses than FWER controlling procedures. As an illustration, for the Leukemia study described in Chapter 1, 20 genes are called differentially expressed at a FWER level of 0.05 (Bonferroni, Holm, Hochberg and Hommel procedures produced the same results for this data set); 163 genes (including these 20) are called differentially expressed at a FDR level of 0.05 by the BH procedure, 30 of them being also rejected by the BY procedure at the same level.

We believe that the above interpretation in terms of "power" is quite misleading, because it does not make sense to compare the number of rejections (or power) of two statistical procedures that aim at controlling different type I error risks. By design, FWER control is adapted to confirmatory analyses, where a statement on the rejected hypotheses needs to be made with high probability, whereas FDR control has been developed for exploratory analyses, where a statement in expectation is acceptable.

Perhaps surprisingly, it may be the case that a FWER controlling procedure rejects more hypotheses than a FDR controlling procedure. As noted by [60], in the very sparse situation where the expected number of false null hypotheses is $o(\log(m))$, the Bonferroni (or Holm) procedure may reject more hypotheses than the BY procedure, because the $\log(m)$ first critical values of the BY procedure are smaller than the Bonferroni threshold α/m. In this extreme example, if one is not willing to make any assumption on the dependency within the p-value family, it may make sense to use the Bonferroni (or Holm) procedure rather than the BY procedure in order to control FDR. This example is mainly an illustration of the conservativeness of the BY procedure, which is seldom used in practice.

2.3.4 FDR control vs FDX control

The great popularity of FDR control and of the BH procedure in particular can probably be explained in part by the fact that the FDP is an appealing and intuitive quantity. An obvious caveat from the statistical perspective is that controlling FDR $= \mathbb{E}(\text{FDP})$ does not tell much about the distribution of the underlying FDP. Here we illustrate this point numerically in the model of Example 2.7. One simulation run consists in $m = 1,000$ null hypotheses distributed as Gaussian equi-correlated with a given correlation ρ. Among those, $m_1 = 200$ (corresponding to $\pi_0 = 0.8$) are true alternatives, with marginal distribution $\mathcal{N}(2,1)$ while the other $m_0 = 800$ are true nulls with marginal distribution $\mathcal{N}(0,1)$. For each $\rho \in \{0, 0.1, 0.2, 0.3, 0.4\}$, we have performed $B = 1,000$ simulation runs, and apply the BH procedure at level $\alpha = 0.25$. The results are summarized in Figure 2.1, where the empirical distribution of the FDP actually achieved on the B replications is represented as

a "violin plot", ie a mirrored and rotated kernel density plot. The empirical FDR achieved is represented by a triangle, and the median FDP by a diamond. The oracle BH procedure at level α under independence ($\rho = 0$) achieves a FDR of $\pi_0\alpha = 0.2$, which is represented by a dashed horizontal line.

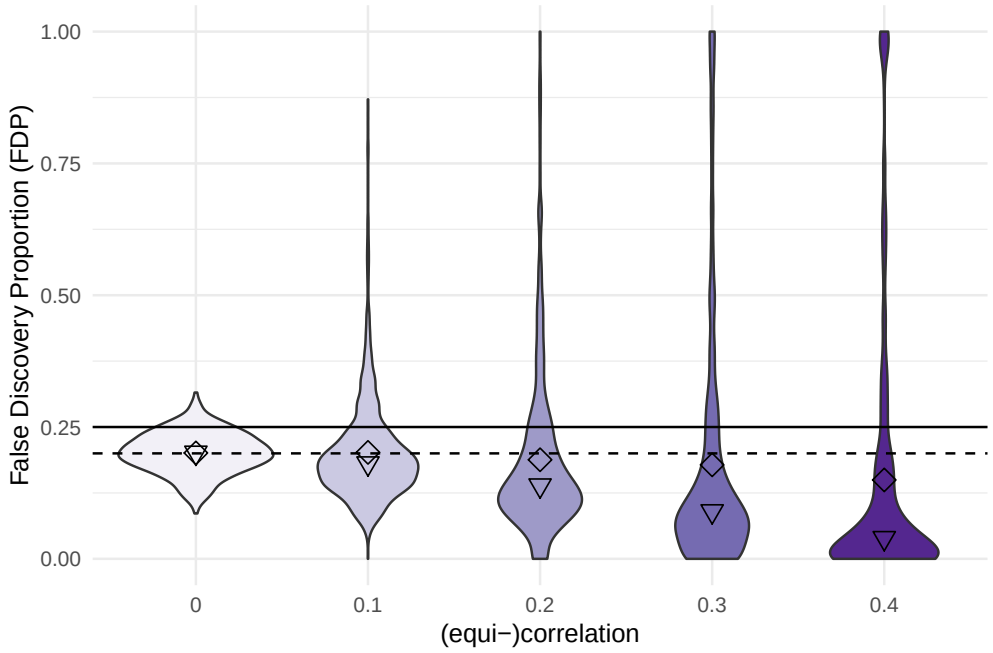


Figure 2.1: FDR control is not FDP control.

As expected in this PRDS setting, FDR is controlled in all situations by the BH procedure. When $\rho = 0$ (independence), the distribution of the FDP of the BH procedure has a Gaussian-like distribution centered at $\pi_0\alpha$ and well-concentrated[6]. In this regime, forgetting the random nature of the FDP of the BH procedure is not too problematic, as most of the mass is below the target level α. However, for larger values of ρ the FDP of the BH procedure has a strikingly different empirical distribution. First, FDR control is conservative, in the sense that the FDR achieved is substantially below $\pi_0\alpha = 0.2$, and the distribution of the FDP is heavily shifted toward smaller values. More problematic is the influence of ρ on the dispersion of the FDP achieved by the BH(0.25) procedure: for $\rho > 0$ the range of the FDP achieved by the BH(0.25) procedure is pretty much the entire $[0, 1]$ interval, with more than 20% of replications having a FDP larger than $\alpha = 0.25$. In such a situation of positively dependent tests, which is expected to be common in applications, we believe that FDR control is not very helpful, and can be misleading. Indeed, most practitioners assume that FDR $\leq \alpha$ guarantees that the proportion of false positives *in their particular experiment* is no more than α. This experiment shows that this may be far from true, even under moderate positive dependence. Although illustrated with the BH procedure for simplicity, this limitation is *intrinsic to FDR control*, and not specific of a particular FDR controlling procedure.

From this perspective, it seems that a more sensible objective than FDR control would be to control the FDX, that is, quantiles of the FDP instead of only controlling its expectation. A straightforward way to formalize the connection between FDR and FDP control has been proposed by [137]: by Markov's inequality, for any $q \in (0, 1)$,

$$\text{FDX} \leq \text{FDR}(R)/q \tag{2.15}$$

[6]Chapter 3 provides theoretical results supporting this observation.

As a consequence, any FDR controlling procedure can be turned in a FDX-controlling procedure. For example, under Assumption PRDS($\mathcal{H}_0$), the BH procedure at level $q\alpha$ yields FDR $\leq q\pi_0\alpha \leq q\alpha$, so that

$$\mathrm{FDX}\left(R^{\mathrm{BH}(q\alpha)})\right) \leq \alpha\,.$$

This simple remark is interesting in that it explicitly provides a quantification of the price to pay for moving from the control of FDR, the expected FDP, to the control of the corresponding tail probability: specifically, for the FDP to be less than some q with probability greater than $1-\alpha$, one needs to apply the FDR controlling procedure at the smaller level $q\alpha$. We believe this type of statement to be much more informative than FDR control. However, it is also more demanding, in the sense that $q\alpha$ will typically be very small if one desires a guarantee with small FDP (small q) and high-confidence (small α). Continuing the example of DGE studies, taking $\alpha = 0.1$ and $q = 0.5$, we can guarantee that

$$\mathrm{FDX}(R^{\mathrm{BH}(0.05)} \geq 0.5) \leq 0.1,$$

that is, with 90% confidence, the FDP among the 163 genes selected by the BH(0.05) procedure is less than 1/2. Or, taking $q = 0.1$ and $\alpha = 0.5$, we can guarantee that

$$\mathrm{FDX}(R^{\mathrm{BH}(0.05)}) \geq 0.1) \leq 0.5\,.$$

That is, the median FDP of the 163 genes selected by the BH(0.05) procedure is less than 1/10. Another approach to FDX control is to simply "augment" the set R of rejections of a FWER controlling procedure by any set A of hypotheses whose cardinality is such that $|A|/(|A|+|R|) \leq q$ [152]. Related works on the construction of confidence envelopes for the FDP process [149, 125, 3] are discussed in the next section.

2.4 From selective inference to simultaneous inference

In this section, we take one step back and look at existing methods in terms of their ability to address the problem of data snooping, in a multiple testing context (Section 2.4.1) and for linear models (Section 2.4.2). In both cases, we distinguish two broad types of approaches pertaining to *selective inference*, which is described by [56] as "the assessment of significance and effect sizes from a dataset after mining the same data to find these associations" and to *simultaneous inference* in the sense of [196], that is, inference valid for any possible look at the data.

2.4.1 Post hoc inference via multiple testing

As noted by Goeman and Solari [81] and more recently by Katsevich and Ramdas [3], controlling multiple testing risks such as FDR – and even FDX – provides statistical guarantees on a specific set of hypotheses selected by the procedure. Therefore, there is a substantial gap between the statistical guarantees provided by state-of-the-art multiple testing procedures and the actual needs of practitioners. To illustrate this important point, let us go back to our example of differential expression analyses. The state-of-the-art approach to this problem consists in performing one statistical test of no difference between means for each gene, and to derive a list R of "significant" genes according to a multiple testing criterion, usually the FDR. This list is then typically refined and/or interpreted using *prior knowledge* on the problem at hand. For example, as smaller effects are generally of a less relevant from a biological perspective, a typical practice is then to only retain those genes whose "fold change" (that is, the ratio of mean expression levels between the two groups) exceeds a prescribed level [153]. Another example is to retain only those genes that belong to a specific biological pathway of interest.

In the above examples, multiple testing procedures provide no statistical guarantee as to the number or proportion of false positives in these user-refined gene lists. This illustrates a gap between theory and practice: multiple testing procedures have been built as tools for statistical *inference*, but they are commonly used as tools for *exploratory data analysis* (EDA) tools. From a statistician's perspective, these examples are instances of data snooping, which is inherent to current data-driven research. The statistical community should not only warn investigators about the caveats of this type of practice, but also develop inference methods dedicated to this new paradigm. Two recent research areas have been developed within the multiple testing community: False Coverage Rate (FCR) control in the framework of selective inference, and post hoc inference in the framework of simultaneous inference.

The False Coverage Rate (FCR) is a multiple testing criterion inspired by FDR, but specifically dedicated to inference on confidence intervals for *selected* parameters [132]. The FCR is defined as the average proportion of covering intervals among those selected. Benjamini and Yekutieli [132] introduced a generic procedure controlling FCR under (indep) or (PRDS($\mathcal{H}_0$)). This procedure consists in adjusting the level of each selected marginal confidence interval by a scaling factor depending on the selection. Importantly, it can be applied to any selection rule. This work has then been generalized to the case where families of hypotheses are selected, instead of individual ones [57].

An important contribution to the field of simultaneous inference is the work of Goeman and Solari [81] to construct procedures that provide confidence statements for the number (or proportion) of false positives in such arbitrary, possibly data-driven subsets of hypotheses. Formally, the aim is to find a functional V_α satisfying

$$\mathbb{P}\Big(\forall S \subset \mathbb{N}_m,\ |S \cap \mathcal{H}_0(P)| \leq V_\alpha(S)\Big) \geq 1-\alpha\,, \tag{PH$_\alpha$}$$

The bound $V_\alpha(S)$ is an $(1-\alpha)$-upper confidence bound on the number of false positives in S. The aim formalized in (PH$_\alpha$) can be equivalently formulated in terms of a $1-\alpha$ lower confidence bound on the number of true positives in S; we focus on V_α for conciseness. A bound V_α satisfying (PH$_\alpha$) is called a *post hoc* bound Goeman and Solari [81], as the set of selected hypotheses may be defined by an investigator after "seeing the data". An earlier contribution in this direction is the work of Genovese and Wasserman [125], where it was noted that a bound satisfying (PH$_\alpha$) could be derived from a multiple testing procedure controlling k-FWER. The derivation of confidence envelopes for the process (FDP$(t), t \in [0,1]$) under (indep) [149, 138, 3] or arbitrary dependence [129] is also related to post hoc inference, because these envelopes are uniform in t. A more precise comparison between these approaches is given in Chapter 5.

General post hoc bounds have been obtained by Goeman and Solari [81] using a construction based on closed testing. This construction takes advantage of so-called non-consonant rejections in the closed testing procedure. It inherits the elegance and genericity of closed testing, but also its heavy computational complexity (see Section 2.2.1). To address this issue, Goeman and Solari [81] have introduced computational shortcuts to obtain a computationally efficient post-hoc procedure in the particular case of Simes local tests, thus under Assumption (PRDS($\mathcal{H}_0$)).

2.4.2 Inference after model selection

In this section we follow the notation of [65] and focus on the case of a linear model of the form

$$y = \mu + \epsilon\,, \tag{2.16}$$

where y is a n-dimensional observation, μ a fixed vector in $\mathbb{R}^n$ and $\epsilon \sim \mathcal{N}(0, \sigma I_n)$. We consider a $n \times p$ fixed design matrix X, whose columns correspond to explanatory variables

for μ. It is not necessarily assumed that there exists $\beta \in \mathbb{R}^p$ such that $X\beta = \mu$. A model corresponds to a subset of selected variables in $\{1, \ldots, p\}$. For a *fixed* model $M \subset \{1, \ldots, p\}$, assuming that $X_M^\top X_M$ is invertible, we define $\beta_M = X_M^\dagger \mu$ as the target of inference, where $X_M^\dagger = (X_M^\top X_M)^{-1} X_M^\top$. The ordinary least-squares (OLS) estimate of β_M is $\hat{\beta}_M = X_M^\dagger y$. By definition, $\hat{\beta}_M$ is an unbiased estimator of β_M. We use the "full model indexing" convention defined in [65]: for $j \in M$, we denote by $j \cdot M \in \{1, \ldots, |M|\})$ the rank of j in M, and $\beta_{j \cdot M} = \mathbb{E}(\hat{\beta}_{j \cdot M})$ is the coordinate of the vector β_M corresponding to the j-th predictor in X. A candidate confidence interval on $\beta_{j \cdot M}$, centered at $\hat{\beta}_M$, is denoted by:

$$CI_{j \cdot M}(K) = \left[\hat{\beta}_{j \cdot M} \pm K \hat{\sigma} \left\| v_{j \cdot M} \right\| \right] , \tag{2.17}$$

where $v_{j \cdot M}^\top$ is the $(j \cdot M)$-th row of $X_M^\dagger$, which is such that $\hat{\beta}_{j \cdot M} = v_{j \cdot M}^\top y$. Here, $\hat{\sigma}$ is assumed to be an estimator of σ that is independent of the estimates $\hat{\beta}_{j \cdot M}$ for all M, and such that $\hat{\sigma}^2 \sim \sigma^2 \chi_r^2 / r$ some $r > 0$, as discussed e.g. in [65]. Let $K = t_{r, 1-\alpha/2}$ be the $1 - \alpha/2$ quantile of Student's t-distribution with r degrees of freedom. Then $CI_{j \cdot M}(K)$ is a valid $(1 - \alpha)$-confidence interval on $\beta_{j \cdot M}$, in the sense that

$$\mathbb{P} \left(\beta_{j \cdot M} \in CI_{j \cdot M}(K) \right) \geq 1 - \alpha \, .$$

In practice however, the model M is not fixed, as it is typically the result $\widehat{M}$ of a *data-dependent selection* step. This extra layer of randomness implies that $CI_{j \cdot \widehat{M}}(K)$ is not a valid $(1 - \alpha)$-confidence interval anymore. Constructing valid inference after model selection is recognized as a difficult problem, see e.g. [47, 107, 126, 127, 136]. The numerous contributions to this problem can be categorized into sample splitting, inference conditional on model selection, and inference uniformly over all model selections.

Sample splitting : single sample splitting consists in using part of the data for model selection, and the rest for inference. This idea is probably the oldest method for inference after model selection [187, 176]. More recently, this idea has been extended to high-dimensional contexts (see e.g. [58, 45, 5]). In order to address the fact that the results depend on the split, multi-sample splitting methods [99] have been proposed, that consist in successively applying single sample splitting to B different splits, and then aggregating the resulting p-values. An intrinsic limitation of this kind of methods is the loss of power inherent to splitting.

Inference conditional on model selection : in the context of penalized linear regression, an asymptotic "covariance test" has been introduced by Lockhart, Taylor, Tibshirani, and Tibshirani [61]. For each k along the regularization path of the lasso [172], it tests for signal evidence in the k^{th} predictor in the model, conditionally on the event that $k - 1$ predictors are already present in the model.

This theory has been extended an consolidated, notably with a test for more generic sequential selection rules [40], a non-asymptotic "spacing test" [39], and a test allows inference to be made at any fixed value of the regularization parameter [36] among other contributions. The power of (a studentized version of) the spacing test is also studied in [14]. By construction, the resulting confidence intervals are *conditional* on the outcome of a model selection step of the form $CI_{j \cdot \widehat{M}}(K)$, and satisfy guarantees of the form:

$$\mathbb{P} \left(\beta_{j \cdot M} \in CI_{j \cdot \widehat{M}}(K) | M = \widehat{M} \right) \geq 1 - \alpha \, .$$

A recent simulation-based study suggests that the expected length of the associated confidence intervals is typically large and can even be infinite [21].

Inference uniformly over all model selections : the post hoc adjustment method of Scheffé [196] yields confidence intervals for predictors that are valid for all possible

model selections in low-dimensional linear models. This method was introduced in the specific context of analysis of variance (ANOVA) for group mean comparison [196], which can be cast as a specific linear model (2.16). Scheffé's method consists in finding a constant K such that the associated confidence intervals (2.17) are simultaneously valid for all possible contrasts involving group means[7]. [196]'s method can be seen as a method to control the FWER, where individual tests would correspond to (infinitely many) contrasts. This method is known as "post hoc analysis" and the type of guarantee it provides is very similar in spirit to the post hoc goals of Section 2.4.1, hence the name "post hoc inference" coined by [81] for the latter approach.

Elaborating on this idea of casting the problem of inference after model selection as a problem of simultaneous inference, Berk, Brown, Buja, Zhang, Zhao, et al. [65] have proposed so-called Post-Selection Inference (PoSI) confidence intervals of the form (2.17), which are simultaneously valid for all possible model selections, that is,

$$\mathbb{P}\left(\forall M \in \mathcal{M}, \forall j \in M, \beta_{j \cdot M} \in CI_{j \cdot M}(K)\right) \geq 1 - \alpha\,,$$

where $\mathcal{M}$ denotes the set of all possible models, corresponding to all $2^p - 1$ non-empty subsets of $\{1, \ldots, p\}$. The corresponding constant K is called the PoSI constant. It depends on the design matrix, which makes the associated confidence possibly sharper than the ones obtained by the Scheffé method, but also much more computationally demanding. Extensions to this work have removed the Gaussian homoscedasticity assumption [2] and generalized PoSI inference to the problem of prediction after model selection [1]. Using a reduction of the PoSI problem to a simultaneous inference problem, Kuchibhotla, Brown, Buja, Cai, George, and Zhao [4] have recently introduced confidence intervals of the following form:

$$CI^*_M = \left\{\theta \in \mathbb{R}^{|M|}, \left\|\widehat{\Sigma}_M(\hat{\beta}_M - \theta)\right\|_\infty \leq C_{xy}(\alpha) + C_{xx}(\alpha)\left\|\theta\right\|_1\right\},$$

where $\widehat{\Sigma}_M = n^{-1}X_M^\top X_M$ is the empirical covariance matrix associated with the submodel indexed by M and $C_{xy}(\alpha)$ and $C_{xx}(\alpha)$ denote $(1-\alpha)$ joint quantiles of $\left\|n^{-1}(\sum_{i=1}^n X_i Y_i - \mathbb{E}(X_i Y_i))\right\|_\infty$ and $\left\|n^{-1}(\sum_{i=1}^n X_i^\top X_i - \mathbb{E}(X_i^\top X_i))\right\|_\infty$. These confidence intervals enjoy the desired uniform coverage property:

$$\mathbb{P}\left(\forall M \in \mathcal{M}, \beta_M \in CI^*_M\right) \geq 1 - \alpha\,.$$

When compared to the original PoSI intervals, these confidence intervals enjoy two remarkable properties: they have a much reduced computational cost of $O(p^2)$, and their volume scales with the size of the selected model and not with the largest possible model size considered.

2.5 Contributions

In the next chapters of this part, we describe several contributions to the multiple testing and post hoc inference literature:

- Chapter 3 is motivated by the frequent misinterpretation of FDR controlling procedures in (genomic) applications as "controlling FDP", which is not a well-defined property as the FDP is a random variable. This chapter is a synthesis of my contributions [J24] and [J13], which were done during my PhD thesis and my post-doc.

[7] In the context of analysis of variance Scheffé [194], a *contrast* is simply a linear combination of the group means such that the coefficients of the combination sum to 0. This naturally covers the test for equality of means of two subgroups.

- Chapter 4 is a synthesis of a joint work Etienne Roquain [J17]] in the framework of classification in sparse high-dimensional models, toward the characterization of the asymptotic properties of FDR controlling procedures viewed as binary classification procedures.

- Chapter 5 describes our contributions to the field of post hoc inference. The contributions described in this chapter [J2, J3, P1] were made in collaboration with Gilles Blanchard and Etienne Roquain in the context of the JCJC ANR project SansSouci (2016-2020). The works described in the last section were performed by Guillermo Durand during the last year of his PhD thesis [31]. All of the procedures described in this chapter are implemented in the R package `sansSouci` [S1], which is available from `https://github.com/pneuvial/sanssouci` and described at `https://pneuvial.github.io/sanssouci`.

Before describing these contributions, we briefly mention a contribution to the PoSI literature introduced in Section 2.4.2, which is not described in the rest of this document. Let us denote by $K(X, \mathcal{M})$ the PoSI constant associated with $n \times p$ design matrix X, where $\mathcal{M}$ denotes the set of all $2^p - 1$ non-empty submodels of $\{1, \dots, p\}$. As mentioned in Section 2.4.2, the computation of $K(X, \mathcal{M})$ for a generic X is a priori of exponential complexity as it involves a maximization over all possible submodels $M \in \mathcal{M}$. This motivates attempts to find upper bounds on the PoSI constant. If X is orthogonal, then $K(X, \mathcal{M}) = \Omega(\sqrt{\log(p)})$ [65]. Moreover, it was shown in an intermediary version of [64] using a cardinality-based argument that the PoSI constant restricted to the set $\mathcal{M}_s$ of s-sparse models satisfies $K(X, \mathcal{M}_s) = O(\sqrt{s \log(p/s)})$, with no assumption on the design X. In Bachoc, Blanchard, and Neuvial [J6], we have proposed a new upper bound on $K(X, \mathcal{M}_s)$ in the case of design matrices satisfying a Restricted Isometry Property (RIP) condition. This upper bound is an explicit function of the RIP constant δ of the design matrix, and it can be seen as an interpolation between the orthogonal setting and the generic sparse setting. In particular, when X is such that $\delta \to 0$, our results imply that $K(X, \mathcal{M}_s) = O(\sqrt{\log(p)} + \delta\sqrt{s \log(p/s)})$. This corresponds to the intuition that for such design matrices, any subset of s columns of X is "approximately orthogonal". Moreover, we have shown that this upper bound is asymptotically optimal in many settings by constructing a matching lower bound.

Chapter 3

Asymptotics of FDR controlling procedures

We introduce a mathematical formalism to study the concentration of the FDP around the FDR asymptotically as the number of hypotheses tested tends to infinity. This formalism is applied to derive central limit theorems for the FDP of a wide class of FDR controlling procedures including the BH procedures and π_0-adaptive procedures. This chapter is a synthesis of my contributions [J24] and [J13], which has required the adaptation of the original results of [J24] to the "random effects setting" used in [J13]. Moreover, I have chosen to focus here on the results of [J24] obtained for plug-in procedures, because the procedures studied in [J13] also fall into this category. I have also added a section dedicated to FDR control under dependency.

References:

[**J13**] P. Neuvial. "Asymptotic Results on Adaptive False Discovery Rate Controlling Procedures Based on Kernel Estimators". *Journal of Machine Learning Research* 14 (2013), pp. 1423–1459

[**J24**] P. Neuvial. "Asymptotic properties of false discovery rate controlling procedures under independence". *Electron. J. Statist.* 2 (2008). With corrigendum in EJS 2009(3):1083, pp. 1065–1110

Contents

3.1 FDP as a stochastic process of a random threshold

3.1.1 Random effects setting

In the setting defined in Chapter 2, the collection $\mathcal{H}$ of hypotheses is deterministic and the p-values $(p_i)_{i\in\mathbb{N}_m}$ are such that $p_i \sim \mathcal{U}[0,1]$ when $\theta_i = 0$. During my PhD thesis, I studied the asymptotic properties of FDR controlling procedures in this setting [T1, J24]. In this chapter however, we consider the "random effects" setting introduced by [162]. This setting, also known as the two-group mixture model, has been widely used in the multiple testing literature, see, e.g., [115, 149, 154], which makes it easier to replace my contributions in the context of this literature. In the present chapter, $\mathcal{H} = (\mathcal{H}_{0,i}, i \in \mathbb{N}_m)$ is a sequence of random indicators, independently and identically distributed as $\mathcal{B}(1-\pi_0)$, where $\pi_0 \in (0,1)$. Conditional on $\mathcal{H}$, the p-values satisfy $p_i|\theta_i \sim G_{\theta_i}$, where G_{θ_i} denotes the cumulative distribution function of the p-values under θ_i. Thus the p-values are independently, identically distributed as $G = \pi_0 G_0 + (1-\pi_0)G_1$, and $m_0(m)$ follows the binomial distribution $\mathrm{Bin}(m, \pi_0)$. We make the following assumptions on the conditional distributions of the p-values: G_0 is uniform (that is, $G_0(u) = u$ for any $u \in [0,1]$), and G_1 is concave. This last assumption is quite natural because the concavity of G_1 is equivalent to the fact that the likelihood ratios of the test statistics under the null versus under the alternative is non-decreasing. We refer to [J13, section 2.1] for a discussion on this assumption.

We consider an asymptotic setting where $m \to +\infty$. Therefore, in this chapter the phrase "multiple testing procedure" refers to a collection $R = (R_m)_{m\in\mathbb{N}}$ of rejection sets R_m rather than a single one as in chapter 2. Consequently, a thresholding procedure is here defined by a collection $\hat{t} = (\hat{t}_m)_{m\geq 1}$ of thresholds such that for all $m \in \mathbb{N}$, $R_m = \{i \in \mathbb{N}_m, p_i \leq \hat{t}_m\}$. Following [149], we note that the False Discovery Proportion can be viewed as a stochastic process. Let $\widehat{\mathbb{G}}_{0,m}$ and $\widehat{\mathbb{G}}_{1,m}$ denote the (unobserved) empirical cumulative distribution function of the p-values under the null and alternative hypotheses:

$$\begin{cases} \widehat{\mathbb{G}}_{0,m}(t) &= m_0(m)^{-1} \sum_{i=1}^m (1-\theta_i)\mathbf{1}\left\{p_i \leq t\right\} \\ \widehat{\mathbb{G}}_{1,m}(t) &= (m - m_0(m))^{-1} \sum_{i=1}^m \theta_i \mathbf{1}\left\{p_i \leq t\right\} \end{cases}.$$

With this notation, and letting $\pi_{0,m} = m_0(m)/m$, $\widehat{\mathbb{G}}_m = \pi_{0,m}\widehat{\mathbb{G}}_{0,m} + (1-\pi_{0,m})\widehat{\mathbb{G}}_{1,m}$ is the (observable) empirical cumulative distribution function of the p-values. Moreover, for any $t \in [0,1]$, $|R_m(t)| = \sum_{i=1}^m \mathbf{1}\left\{p_i \leq t\right\} = \widehat{\mathbb{G}}_m(t)$ and $|R_m(t) \cap \mathcal{H}_0| = \sum_{i=1}^m (1-\theta_i)\mathbf{1}\left\{p_i \leq t\right\} = \pi_{0,m}\widehat{\mathbb{G}}_{0,m}(t)$, so that

$$\mathrm{FDP}_m(t) = \frac{\pi_{0,m}\widehat{\mathbb{G}}_{0,m}(t)}{\widehat{\mathbb{G}}_m(t) \vee \frac{1}{m}} \tag{3.1}$$

is the False Discovery Proportion achieved at the deterministic threshold t. The asymptotic properties of the stochastic process $(\mathrm{FDP}_m(t))_{0\leq t\leq 1}$ were analyzed by Genovese and Wasserman [149]. The FDR achieved at threshold t, $\mathrm{FDR}_m(t) = \mathbb{E}\left(\mathrm{FDP}_m(t)\right)$, may be written as $\mathrm{FDR}_m(t) = \mathrm{p}(t)\left(1-(1-G(t))^m\right)$, where $\mathrm{p}(t) = \pi_{0,m}t/G(t)$ is the positive False Discovery Rate at t defined above in (2.1.2). Using the functional Delta method [169], they proved that the FDP_m process converges to pFDR at a rate $m^{-1/2}$, and built confidence envelopes for the FDP process using this result. However, this result is not sufficient to characterize the behavior of the FDP *actually achieved by a given multiple testing procedure*, that is, by the random variable $\mathrm{FDP}_m(\hat{t}_m)$. We are interested in the asymptotic behavior of this variable and, in particular, its fluctuations around the asymptotic FDR achieved by the procedure with threshold $\hat{t}_m$.

Following [149], we note that the threshold $\hat{t}^{\mathrm{BH}}(\alpha)$ of the BH procedure defined in Section 2.3 may also be written as

$$\hat{t}^{\mathrm{BH}}(\alpha) = \sup\left\{u \in [0,1], \widehat{\mathbb{G}}_m(u) \geq u/\alpha\right\}. \tag{3.2}$$

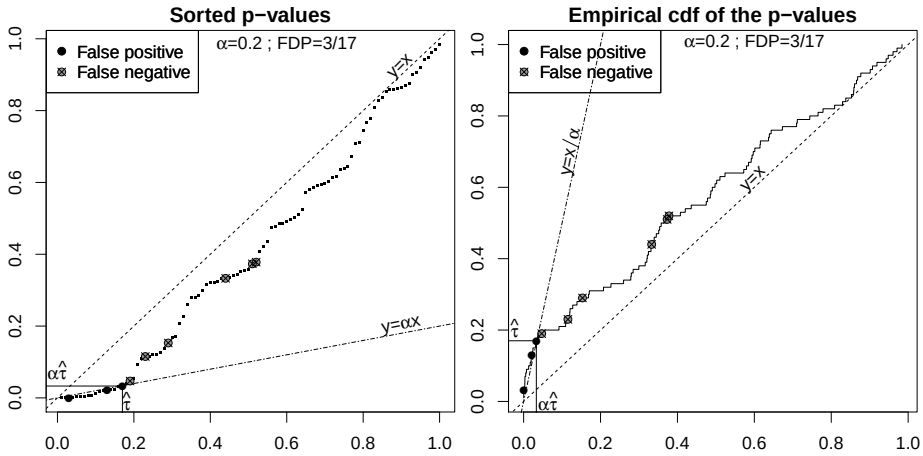


Figure 3.1: Dual interpretations of the BH threshold. Left: ordered p-values; right: empirical cumulative distribution function of the p-values.

Therefore, defining

$$\mathcal{U}(F,\alpha) = \sup\{u \in [0,1], F(u) \geq u/\alpha\}\,, \tag{3.3}$$

we have $\hat{t}^{\mathrm{BH}}(\alpha) = \mathcal{U}(\widehat{\mathbb{G}}_m, \alpha)$. Figure 3.1 illustrates the duality between the classical definition of the BH procedure, and its interpretation using threshold functions. Following [96], $u \mapsto u/\alpha$ is called the *rejection curve* of the BH procedure (also known as *the Simes line* [179]).

3.1.2 Two types of plug-in procedures

The plug-in procedures introduced in Chapter 2 are a class of procedures that aim at filling the "conservativeness gap" between the level $\pi_0\alpha$, which is the FDR control actually provided by the BH procedure, and the target FDR level α. Plug-in procedures mimic the Oracle BH(α/π_0) procedure by applying the BH procedure at level $\alpha/\hat{\pi_0}$, where $\hat{\pi_0}$ is an estimator of π_0. Therefore, the threshold of such procedures may be written as

$$\hat{t}_m(\alpha) = \mathcal{U}(\widehat{\mathbb{G}}_m, \alpha/\hat{\pi}_0)\,. \tag{3.4}$$

We consider two types of such estimators of π_0: estimators that are based on the empirical cumulative distribution function $\widehat{\mathbb{G}}_m$ of the p-values, and estimators that are based on the estimation of the density $g(1)$ of the p-values at 1. We begin with the Storey-λ procedure introduced in Section 2.3. The threshold of this procedure may be written as

$$\hat{t}^{0,\lambda}_m(\alpha) = \mathcal{U}\left(\widehat{\mathbb{G}}_m, \alpha/\Pi_\lambda(\widehat{\mathbb{G}}_m)\right)\,, \tag{3.5}$$

where $\Pi_\lambda : F \mapsto (1-F(\lambda))/(1-\lambda)$ only depends on λ[1]. More generally, we define plug-in procedures of order 0 (PI-0) as those whose threshold can be written as

$$\hat{t}^{0}_m(\alpha) = \mathcal{U}\left(\widehat{\mathbb{G}}_m, \alpha/\Pi(\widehat{\mathbb{G}}_m)\right)\,. \tag{3.6}$$

for a given $\Pi : D[0,1] \to [0,1]$, where $D[0,1]$ is the set of càdlàg functions on $[0,1]$, that is, the set of all real-valued functions on $[0,1]$ that are right-continuous at each point of

[1]The exact translation of the definition in (2.14) incorporates an additional $1/m$ term in the numerator of Π_λ. This term is required to ensure finite sample FDR control, but the two associated FDR controlling procedures are asymptotically equivalent, in a sense formally defined in [J24].

$[0,1)$ and have left limits in each point of $(0,1]$. In particular, the BH procedure can be viewed as a PI-0 procedure with a constant $\Pi = 1$. The BKY procedure [123] is another example of PI-0 procedure studied in [J24]. One limitation of PI-0 estimators in general and the Storey-λ estimator in particular, is that these estimators are generally not consistent estimators of π_0. For $\lambda \in (0,1)$, the expectation of the Storey-λ estimator is

$$\Pi_\lambda(G) = \pi_0 + (1-\pi_0)\frac{1-G_1(\lambda)}{1-\lambda}\,, \tag{3.7}$$

As G_1 is concave, $\Pi_\lambda(G)$ is a non-increasing function of λ, which is therefore lower-bounded by $\overline{\pi_0} = \pi_0 + (1-\pi_0)g_1(1) = g(1)$. Therefore, estimators of $g(1)$ are relevant candidates to estimate π_0. The second class of estimators considered here is a general class of kernel estimators of $g(1) = \overline{\pi_0}$, which we have introduced in [J13].

Definition 3.1.

1. *A kernel of order $\ell \in \mathbb{N}$ is a function $K : \mathbb{R} \to \mathbb{R}$ such that the functions $u \mapsto u^j K(u)$ are integrable for any $j = 0 \dots \ell$, and satisfy $\int_{\mathbb{R}} K = 1$, and $\int_{\mathbb{R}} u^j K(u)du = 0$ for $j = 1 \dots \ell$.*

2. *The kernel estimator of a density g at x_0 based on m independent, identically distributed observations $x_1, \dots x_m$ from g is defined by*

$$\hat{g}_m(x_0) = \frac{1}{mh}\sum_{i=1}^{m} K\left(\frac{x_i - x_0}{h}\right)\,,$$

where $h > 0$ is called the bandwidth of the estimator, and K is a kernel.

The threshold of the corresponding FDR controlling procedures may be written as

$$\hat{t}^1_m(\alpha) = \mathcal{U}\left(\widehat{\mathbb{G}}_m, \alpha/\hat{g}_m(1)\right)\,. \tag{3.8}$$

We call procedures whose threshold can be written as (3.8) *plug-in procedures of order 1* (PI-1), because their threshold depends on the observations through both $\widehat{\mathbb{G}}_m$ and $\hat{g}_m(1)$, where g is the derivative (of order 1) of G. Note that Storey's estimator may technically be viewed as a kernel estimator of order 0, with kernel function $K(t) = \mathbf{1}\{[-1,0]\}(t)$.

3.1.3 Criticality

The notion of criticality has been introduced by [115] in the context of the BH procedure. It is necessary to define this notion here because it is tightly connected to the asymptotic properties of FDR controlling procedures. Letting $\alpha^\star = \lim_{u\to 0} u/G(u)$, if $\alpha < \alpha^\star$, the number of discoveries made by the BH procedure is stochastically bounded as the number of tested hypotheses increases. Conversely, if $\alpha > \alpha^\star$, the proportion of discoveries converges in probability to a positive value. This property is quite intuitive when we recall that according to (3.2), the threshold of the BH procedure is the largest right-crossing point between the empirical cumulative distribution function $\widehat{\mathbb{G}}_m$ of the p-values and the line with slope $1/\alpha$. An illustration of the critical value of the BH procedure in the Laplace two-sided model is given in the right panel of Figure 3.2. The line $y = x/\alpha^\star$ corresponds to the tangent to G at the origin. For $\alpha < \alpha^\star$, the BH(α) procedure asymptotically makes no rejections as G does not cross the Simes line $y = x/\alpha$ in the interior of the interval $[0,1]$. For $\alpha > \alpha^\star$, the BH(α) asymptotically rejects all the p-values less than $\tau^\star$, where $G(\tau^\star) = \tau^\star/\alpha$. The *critical value* $\alpha^\star$ only depends on the distribution function G of the p-values.

Chi and Tan [101] demonstrated that this property is not specific to the BH procedure. For any multiple testing procedure, for $\alpha < \underline{\alpha}^\star := \pi_0\alpha^\star$, there exists a positive constant $c(\alpha)$ such that almost surely, for m large enough, the events $\{|R_m \cap \mathcal{H}_0|/|R_m| \le \alpha\}$ and $\{|R_m| \ge$

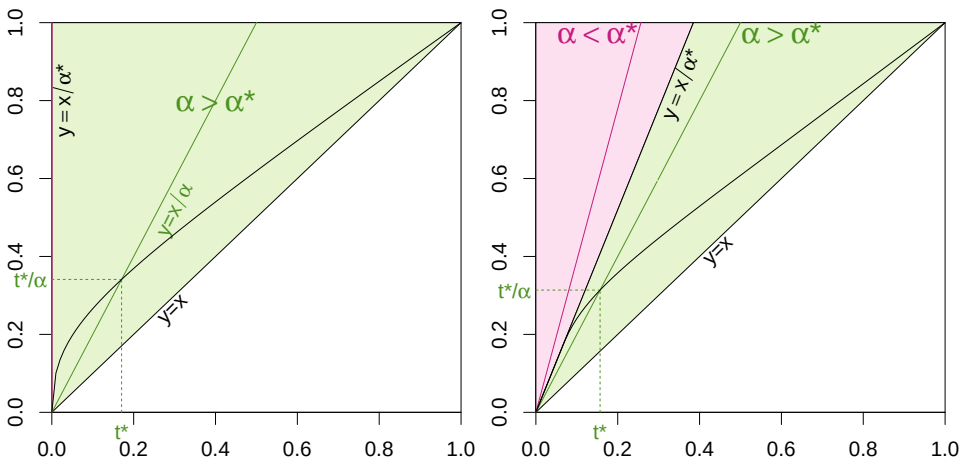


Figure 3.2: Critical value of the BH procedure: illustration of the results of [115] in the Gaussian (left; no criticality, $\alpha^\star = 0$) and Laplace/double exponential (right: criticality; $\alpha^\star > 0$) location models.

$c(\alpha) \log m\}$ are incompatible [101, proof of Proposition 3.2]. This restriction is intrinsic to the multiple testing problem, in the sense that it holds regardless of the considered multiple testing procedure. The bound $\underline{\alpha}^\star$ can be interpreted as the critical value of the above-mentioned Oracle BH procedure. Obviously, when $\underline{\alpha}^\star = 0$, there is no criticality, e.g. the BH procedure has positive asymptotic power for any $\alpha > 0$. For instance, this is the case in the Gaussian location model, but not in the Laplace location model, as illustrated in Figure 3.2. More examples of critical and non-critical distributional settings can be found in [J13, Section 5].

The critical value of a generic multiple testing procedure may be defined as the largest target FDR level for which the corresponding asymptotic power is null [J13, Definition 11]. The following Lemma gives a lower bound on the critical value of plug-in procedures.

Lemma 3.2 (Criticality of plug-in procedures [J13])**.** *Let α_m be a sequence of (possibly data-dependent) levels that converges in probability to $\alpha_\infty \in (0,1)$ as $m \to +\infty$. If $\alpha_\infty < \alpha^\star$, then the threshold $\hat{t}_m(\alpha_m)$ of the BH(α_m) procedure converges in probability to 0 as $m \to +\infty$. If the convergence of α_m to α_∞ holds almost surely, then the convergence of $\hat{t}_m(\alpha_m)$ to 0 holds almost surely as well.*

This property holds in particular for both PI-0 and PI-1 procedures. In the next sections, we give results specific to each type of procedure.

3.2 Results for PI-0 procedures

The asymptotic distribution of the FDP of a wide class of procedures including PI-0 procedures has been obtained in [J24] in the original setting of Benjamini and Hochberg [173]. A fundamental result to derive such asymptotic distributions is given in the next Theorem.

Theorem 3.3 (Adapted from [J24])**.** *Consider a multiple testing procedure with threshold function $\mathcal{T}$. Assume that:*

(i) $\mathcal{T}(G) > 0$, and is Hadamard-differentiable[2] at G;

[2]Here, we consider Hadamard-differentiability on the set of càdlàg functions on [0,1], tangentially to the set of continuous functions on $[0,1]$. We refer to [169] for a formal definition of Hadamard differentiability.

(ii) $(\pi_{0,m}\widehat{\mathbb{G}}_{0,m}, (1-\pi_{0,m})\widehat{\mathbb{G}}_{1,m})$ *converges in distribution to* $(\pi_0 G_0, (1-\pi_0)G_1)$ *at rate* r_m *as* $m \to +\infty$*, where* $r_m \to 0$*;*

Then the FDP of the procedure, $\mathrm{FDP}_m(\mathcal{T}(\widehat{\mathbb{G}}_m))$*, converges in distribution to* $\pi_0\tau^\star/G(\tau^\star)$ *at rate* r_m *as* $m \to +\infty$*.*

The limit distribution is an explicit function of the limit distribution $\mathbb{Z}$ of $(\pi_{0,m}\widehat{\mathbb{G}}_{0,m}, (1-\pi_{0,m})\widehat{\mathbb{G}}_{1,m})$ and the Hadamard derivative at G taken at $\mathbb{Z}$ [J24, Theorem 3.2]. The proof relies on the functional Delta method [169]. The interest of this result is that it breaks down the asymptotic analysis of the FDP of a multiple testing procedure with threshold function $\mathcal{T}$ into (i) the Hadamard differentiability of the mapping $\mathcal{T}$, which only depends on the procedure, and (ii) the asymptotic behavior of the empirical distribution functions of the p-values, which only depends on the model assumptions. We will go back to this modularity in Section 3.4.

In particular, as for (ii), $(\pi_{0,m}\widehat{\mathbb{G}}_{0,m}, (1-\pi_{0,m})\widehat{\mathbb{G}}_{1,m})$ classically satisfies a Donsker-type theorem (with $r_m = m^{-1/2}$) in the random effects setting, as noted by [149, Theorem 4.1]. Moreover, in the case of a PI-0 procedure associated to Π, (i) holds as soon as Π is itself Hadamard-differentiable at G, tangentially to $C[0,1]$, and satisfies $\alpha > \Pi(G)\alpha^\star$ [J24, Proposition 7.11 and Corollary 7.12]. Under these assumptions, Theorem 3.3 yields the convergence in distribution as $m \to +\infty$ of $\sqrt{m}\left(\mathrm{FDP}_m(\hat{t}^0_m(\alpha)) - \pi_0\alpha/\Pi(G)\right)$, where $\hat{t}^0_m(\alpha) = \mathcal{U}(\widehat{\mathbb{G}}_m, \alpha/\Pi(\widehat{\mathbb{G}}_m))$ is the threshold of the PI-0 procedure associated to Π. The Hadamard-differentiability of Π is a technical condition which is typically verified for classical PI-0 procedures such as the Storey-λ procedure or the BKY procedure. The important assumption is the condition: $\alpha > \Pi(G)\alpha^\star$. Recalling that by the definition of Π, the PI-0 procedure consists in applying the BH procedure at level $\alpha_m = \alpha/\Pi(\widehat{\mathbb{G}}_m)$, the above result combined with Lemma 3.2 implies that $\Pi(G)\alpha^\star$ is the critical value of the PI-0 procedure associated to Π.

As both the BH and the Storey-λ procedures can be viewed as PI-0 procedures, the asymptotic distribution of the FDP of these procedures are obtained below as consequences of [J24, Theorem 4.12]; these results were also stated in the Appendix of [J13].

Corollary 3.4 (BH procedure – adapted from [J24])**.** *Let* $\hat{t}_m(\alpha) = \mathcal{U}(\widehat{\mathbb{G}}_m, \alpha)$ *denote the threshold of the BH*(α) *procedure, and* $t_\infty(\alpha) = \mathcal{U}(G, \alpha)$*. For any* $\alpha \geq \alpha^\star$*, we have*

$$\sqrt{m}\left(\mathrm{FDP}_m(\hat{t}_m(\alpha)) - \pi_0\alpha\right) \rightsquigarrow \mathcal{N}\left(0, (\pi_0\alpha)^2\left(\frac{1}{\pi_0 t_\infty(\alpha)} - 1\right)\right) \tag{3.9}$$

This central limit theorem provides a theoretical explanation for the Gaussian shape of the empirical distribution of the FDP observed in the numerical experiments reported in Section 2.3.4 the particular case of independence (corresponding to $\rho = 0$ in the Gaussian equi-correlated model of Example 2.7).

Corollary 3.5 (Storey-λ procedure – adapted from [J24])**.** *For any* $\lambda \in [0,1)$*, and* $\alpha \in [0,1]$*, let* $\Pi_\lambda : F \mapsto (1-F(\lambda))/(1-\lambda)$*. Let* $\hat{t}^{0,\lambda}_m(\alpha) = \mathcal{U}(\widehat{\mathbb{G}}_m, \alpha/\Pi(\widehat{\mathbb{G}}_m))$ *be the threshold of the Storey-*λ *procedure at level* α*, and* $t^\lambda_\infty(\alpha) = \mathcal{U}(G, \alpha/\Pi(G))$*. Then, for any* $\alpha > \Pi_\lambda(G)\alpha^\star$*, we have:*

$$\sqrt{m}\left(\mathrm{FDP}_m(\hat{t}^{0,\lambda}_m(\alpha)) - \pi_0\alpha/\Pi_\lambda(G)\right) \rightsquigarrow \mathcal{N}\left(0, \sigma^2_\lambda\right), \tag{3.10}$$

where

$$\sigma^2_\lambda = \left(\frac{\pi_0\alpha}{\Pi_\lambda(G)}\right)^2\left\{\frac{1}{\pi_0 t^\lambda_\infty(\alpha)} + 2\frac{t^\lambda_\infty(\alpha)\wedge\lambda}{t^\lambda_\infty(\alpha)(1-G(\lambda))} - \frac{1}{1-G(\lambda)}\right\}$$

Note that Corollary 3.5 with $\lambda = 0$ recovers Corollary 3.4. Moreover, the asymptotic properties of the BH Oracle procedure can be obtained by applying Corollary 3.4 at level α/π_0.

3.3 Results for PI-1 procedures

Corollary 3.5 establishes that the FDP of the Storey-λ procedure converges in distribution at the parametric rate $m^{-1/2}$ to $\pi_0\alpha/\Pi_\lambda(G)$. As noted above, we have $\Pi_\lambda(G) > g(1) \geq \pi_0$. In this section, we show that PI-1 procedures, which are based on estimators of $g(1)$, may provide an asymptotic FDP control closer to the target FDR level α. However, this tighter control comes at the price of slower convergence rates for the FDP. Indeed, PI-0 estimators involve estimators of π_0 that depend on the observations only through $\widehat{\mathbb{G}}_m$, and therefore achieve parametric convergence rates. Conversely, PI-1 estimators involve estimators of π_0 that rely on the estimation of $g(1)$, and such estimators are known to yield non-parametric convergence rates that depend on the regularity of g at 1. For a kernel estimator of $g(1)$, the convergence rate is typically of $m^{-k/(2k+1)}$ when g is k times differentiable at 1 [100]. We state in Proposition 3.6 the asymptotic properties of a particular kernel estimator: the Storey-λ estimator with $\lambda = 1 - h_m$ tending to 1 as $m \to +\infty$. This estimator is denoted by $\hat{\pi}_{0,m}(1-h_m)$.

Proposition 3.6 (Asymptotic properties of $\hat{\pi}_{0,m}(1-h_m)$ — [J13])**.** *Assume that g is k times differentiable at 1 for $k \geq 1$, with $g^{(l)}(1) = 0$ for $1 \leq l < k$.*

1. *If $g^{(k)}(1) \neq 0$, then the asymptotically optimal bandwidth for $\hat{\pi}_{0,m}(1-h_m)$ in terms of Mean Squared Error (MSE) is of order $m^{-1/(2k+1)}$, and the corresponding MSE is of order $m^{-2k/(2k+1)}$.*

2. *Let η_m be any sequence such that $\eta_m \to 0$ and $m^{k/(2k+1)}\eta_m \to +\infty$ as $m \to +\infty$. Then, letting $h_m(k) = m^{-1/(2k+1)}\eta_m^2$, we have, as $m \to +\infty$:*

$$m^{k/(2k+1)}\eta_m \left(\hat{\pi}_{0,m}(1-h_m(k)) - g(1)\right) \rightsquigarrow \mathcal{N}(0, g(1)) \tag{3.11}$$

This result cannot be derived from classical results on kernel estimators (e.g. [100]) as such results typically require that the order of the kernel matches the regularity k of the density, whereas the kernel of Storey's estimator, $K(t) = \mathbf{1}\{[-1,0]\}(t)$, is of order 0. The proof is based on a formulation of $\hat{\pi}_{0,m}(1-h_m)$ as a sum of m independent random variables that satisfy the Lindeberg-Feller conditions for the Central Limit Theorem [180].

In order to derive the asymptotic distribution of the FDP of PI-1 procedures, we begin by stating a slightly more general result which holds for any plug-in procedure of the form $\mathrm{BH}(\alpha/\hat{\pi}_{0,m})$, where $\hat{\pi}_{0,m}$ is a generic estimator of π_0 whose convergence rate is slower than $m^{-1/2}$.

Theorem 3.7 (Plug-in procedures with non-parametric rates — [J13])**.** *Let $\hat{\pi}_{0,m}$ be an estimator of π_0 whose asymptotic distribution is given by*

$$\sqrt{mh_m}\left(\hat{\pi}_{0,m} - \pi_{0,\infty}\right) \rightsquigarrow \mathcal{N}(0, s_0^2)$$

for some s_0, with $h_m = o(1/\log\log m)$ and $mh_m \to +\infty$ as $m \to +\infty$. Denote by $\hat{t}_m(\alpha)$ the threshold of the associated plug-in procedure. Then, for any $\alpha > \pi_{0,\infty}\alpha^\star$, The asymptotic distribution of the FDP achieved by the BH$(\alpha/\hat{\pi}_{0,m})$ procedure is given by

$$\sqrt{mh_m}\left(\mathrm{FDP}_m(\hat{t}_m(\alpha)) - \frac{\pi_0\alpha}{\pi_{0,\infty}}\right) \rightsquigarrow \mathcal{N}\left(0, \left(\frac{\pi_0\alpha s_0}{\pi_{0,\infty}^2}\right)^2\right).$$

Corollary 3.8 (Asymptotic distribution of the FDP of PI-1 procedures [J13])**.** *Assume that g is k times differentiable at 1 for $k \geq 1$. Define $h_m(k) = m^{-1/(2k+1)}\eta_m^2$, where $\eta_m \to 0$ and $m^{k/(2k+1)}\eta_m \to +\infty$ as $m \to +\infty$. Denote by $\hat{\pi}_{0,m}^k$ one of the following two estimators of π_0:*

- *Storey's estimator $\hat{\pi}_{0,m}^{\mathrm{Sto}}(1-h_m(k))$; in this case, it is further assumed that $g^{(l)}(1) = 0$ for $1 \leq l < k$;*

- *A kernel estimator of $g(1)$ associated with a k^{th} order kernel with bandwidth $h_m(k)$.*

Then $\alpha_0^\star = g(1)\alpha_{BH}^\star$ is the critical value of the BH$(\alpha/\hat{\pi}_{0,m}^k)$ procedure, and for any $\alpha > \alpha_0^\star$,

$$m^{k/(2k+1)}\eta_m\left(\mathrm{FDP}_m(\widehat{\tau}_m^0(\alpha)) - \frac{\pi_0\alpha}{g(1)}\right) \rightsquigarrow \mathcal{N}\left(0, \frac{\pi_0^2\alpha^2}{g(1)^3}\right).$$

3.4 Extensions to other dependency settings

The results obtained in [J24] give a characterization of the asymptotic distribution of the FDP actually achieved by a wide class of multiple testing procedures. In particular, these results make it possible to go beyond FDR control by constructing asymptotic confidence intervals for $\mathrm{FDP}_m(\hat{t}_m)$. The obtained central limit theorems are stated here in the random effects setting of [162], and in [J24] in the original setting of [173]. In both cases, the null hypotheses are assumed to be independent, which is a strong assumption, as discussed in Chapter 2. By Theorem 3.3, the results stated above for the BH and PI(0) procedures may be extended to any multiple testing setting where the asymptotic distribution of $(\pi_{0,m}\widehat{\mathbb{G}}_{0,m}, (1-\pi_{0,m})\widehat{\mathbb{G}}_{1,m})$ is characterized. For example, Wu [114, Theorem 1] obtained a functional central limit theorem for $(\pi_{0,m}\widehat{\mathbb{G}}_{0,m}, (1-\pi_{0,m})\widehat{\mathbb{G}}_{1,m})$ under the assumption that $\mathcal{H}$ is stationary and $\pi_{0,m}$ itself satisfies a central limit theorem [114, Condition 1]. These assumptions cover the case of Ising models in $\mathbb{Z}^2$.

Two other examples of uses of this formalism to obtain asymptotic results under dependence are Delattre and Roquain [80, 30]. Delattre and Roquain [80] consider the Gaussian ρ_m-equi-correlated model introduced in Example 2.7, additionally assuming that the equi-correlation parameter ρ_m satisfies $\rho_m \to 0$ as $m \to +\infty$. In this model, the empirical process $(\widehat{\mathbb{G}}_{0,m}, \widehat{\mathbb{G}}_{1,m})$ can be shown to converge in distribution at rate $r_m = \min(m, 1/\rho_m)^{-1/2}$[80, Lemma 3.3]. This rate is different from the standard convergence rate $m^{-1/2}$ holding under independence. A consequence of the results obtained in [J24, J13] is that the FDP of the BH procedure converges to the corresponding false discovery rate (FDR) at the same rate $\min(m, 1/\rho_m)^{-1/2}$ [80, Theorem 2.1]. Using our above-described formalism, the same convergence rates could also be obtained for all of the procedures studied in [J24], including PI(0) procedures.

Delattre and Roquain [30] have obtained a functional central limit theorem for the process $(\pi_{0,m}\widehat{\mathbb{G}}_{0,m}, (1-\pi_{0,m})\widehat{\mathbb{G}}_{1,m})$ under a more general dependency model, where the test statistics are Gaussian, with a covariance matrix $\Gamma^{(m)}$ asymptotically lying in a neighborhood of the identity matrix I_m[3]. The convergence rate of $(\pi_{0,m}\widehat{\mathbb{G}}_{0,m}, (1-\pi_{0,m})\widehat{\mathbb{G}}_{1,m})$ naturally depends on the rate at which the elements of $\Gamma^{(m)}$ vanish. Interestingly, the asymptotic distribution of the FDP of the BH procedure cannot be deduced directly from the results of [J24] because the *joint* convergence of $(\pi_{0,m}\widehat{\mathbb{G}}_{0,m}, (1-\pi_{0,m})\widehat{\mathbb{G}}_{1,m})$ is required. Delattre and Roquain [30] have cleverly noted that the Hadamard derivative of the functional Ψ such that $\mathrm{FDP}_m(\hat{t}) = \Psi(\pi_{0,m}\widehat{\mathbb{G}}_{0,m}, (1-\pi_{0,m})\widehat{\mathbb{G}}_{1,m})$ at $(\pi_0 G_0, (1-\pi_0)G_1)$ only depends on the first of its two coordinates. This property makes it possible to extend the original result of [J24] by deriving the asymptotic distribution of FDP of the BH procedure from only the marginal convergence of $\widehat{\mathbb{G}}_{0,m}$ and $\widehat{\mathbb{G}}_{1,m}$. To our knowledge however, the result obtained in Delattre and Roquain [30] cannot be extended to PI(0) procedures because the above property of the mapping Ψ is specific to the BH procedures. Indeed, our calculations show that for PI(0) procedures, the Hadamard derivative is given by

$$\alpha\frac{H_0}{\mathcal{T}(G)} - \alpha\frac{\dot{\Pi}_G(H_0+H_1)}{\Pi(G)},$$

where the second term (induced by the estimation of π_0) does depend on both H_0 and H_1.

[3] We refer to the "vanish-second-order" condition in [30] for a formal definition of this condition.

Chapter 4

FDR thresholding for classification under sparsity

Albeit motivated by pure testing considerations, the BH procedure has been shown to enjoy remarkable properties as an estimation procedure in Gaussian location [122] and Laplace scaling [124] problems. More specifically, it turns out to be adaptive to the amount of signal contained in the data, which has been referred to as "adaptation to unknown sparsity". More recently, Bogdan, Chakrabarti, Frommlet, and Ghosh [78] have studied FDR thresholding with respect to the mis-classification risk, and proved that FDR thresholding is asymptotically optimal (as the number m of tests goes to infinity) with respect to that risk in the case of a Gaussian scaling model. A natural question is whether this asymptotic optimality is specific to the Gaussian scaling model or whether it also holds in more general settings. We have generalized and extended the work of [78] by studying the asymptotic properties of the BH procedure with respect to the mis-classification risk in more general models that encompass Subbotin location and scale models.

References:

[**J17**] P. Neuvial and E. Roquain. "On false discovery rate thresholding for classification under sparsity". *Annals of Statistics* 40.5 (2012), pp. 2572–2600

Contents

4.1 Settings

4.1.1 Random effects model

We consider the random effects model described in Section 3.1, with additional assumptions on the distribution of the test statistics. In particular, we study the specific case of Subbotin location and scale models where the density d of the test statistics conditionally on $\theta_i = 0$ is the ζ-Subbotin density defined in Section 2.1.1: $d(x) = (L_\zeta)^{-1} e^{-|x|^\zeta/\zeta}$, for $\zeta \geq 1$, where the normalization constant L_ζ is defined in Equation (2.3). The density of the test statistics conditionally on $\theta_i = 1$ is either a shift or a scaling of d. In the Subbotin location model, we have $d_{1,m}(x) = d(x-\mu_m)$, for some (unknown) location parameter $\mu_m > 0$. The corresponding p-values are obtained using the transformation $p_i = \overline{D}(X_i)$, and the cumulative distribution function of the p-values conditionally on $\theta_i = 1$ is given by $G_{1,m}(t) = \overline{D}(\overline{D}^{-1}(t) - \mu_m)$. In the Subbotin scale model, we have $d_{1,m}(x) = d(x/\sigma_m)/\sigma_m$, for some (unknown) scale parameter $\sigma_m > 1$, and the p-values are obtained as $p_i = 2\overline{D}(|X_i|)$, which yields $G_{1,m}(t) = 2\overline{D}(\overline{D}^{-1}(t/2)/\sigma_m)$. This setting covers the Gaussian location model studied in [122] and the Laplace scale model studied in [124] for estimation, and the Gaussian scale model studied in [78] for classification.

4.1.2 Classification risk

We are interested in classification rules that predict the labels $(\theta_i)_{i\in\mathbb{N}_m}$ from the sample of test statistics $(X_i)_{i\in\mathbb{N}_m}$. We define the classification risk of such a rule as the expected proportion of misclassified items. In our setting where the distribution of the X_i under the null distribution is known, a classification procedure can be identified with a threshold $\hat{t}_m \in [0,1]$, that is, a measurable function of the p-value family $(p_i, i \in \{1,...,m\})$. The corresponding procedure chooses label 1 whenever the p-value is smaller than $\hat{t}_m$. The corresponding classification risk may be written in terms of p-values:

$$R_m^c(\hat{t}_m) = m^{-1}\sum_{i=1}^m \mathbb{P}(p_i \leq \hat{t}_m, \theta_i = 0) + m^{-1}\sum_{i=1}^m \mathbb{P}(p_i > \hat{t}_m, \theta_i = 1)$$

In particular, for a deterministic threshold $t_m \in [0,1]$, we have $R_m^c(t_m) = \pi_{0,m}t_m + \pi_{1,m}(1 - G_{1,m}(t_m))$. Classically, the rule minimizing R_m^c corresponds to the threshold $t_m^B = f_m^{-1}(\tau_m)$, where $\tau_m = \pi_{0,m}/(1-\pi_{0,m})$ is called the mixture parameter of the model. This rule is called the Bayes rule. It depends on the unknown value of the model parameters. Our goal is to find a classification rule whose risk is "comparable" to $R_m^c(t_m^B)$, in the following sense. A classification rule is said to be *asymptotically optimal at rate* $r_m = o(1)$ if and only if there exists $D > 0$ such that for large m,

$$R_m^c(\hat{t}_m) - R_m^c(t_m^B) \leq D \times R_m^c(t_m^B) \times r_m\,. \tag{4.1}$$

4.1.3 Power of the Bayes rule under sparsity

Following [78, 122, 124], we consider a sparse situation, in the sense that $\tau_m \to +\infty$ as $m \to +\infty$. Typically, we consider scenarios where $\tau_m = m^\beta$, with $0 < \beta \leq 1$. To compensate for this weakening of the signal, the distance between distribution of the test statistics under the alternative is allowed to grow as $m \to +\infty$, in such a way that $\mu_m \to +\infty$ and $\sigma_m \to +\infty$ in the location and scale models, respectively. Formally, we make the general assumption that the power $C_m = G_{1,m}(t_m^B)$ of the Bayes rule is bounded away from 0 and 1. This assumption makes the classification problem "just solvable" under the sparsity constraint.

4.2 Asymptotic optimality of FDR thresholding

Although we are not directly interested in the control of the FDR or the FDP in this chapter, the BH procedure can be viewed as a classification rule with threshold $\hat{t}_m = \hat{t}_m^{\mathrm{BH}}(\alpha)$. We have proved in [J17] that there exists a choice of the target level α_m, such that the BH procedure at level α_m is asymptotically optimal (with an explicit rate) in terms of the above classification risk. Our analysis of the risk of the BH procedure relies on two key elements: a finite sample Oracle inequality on the excess risk of the BH procedure [J17, Corollary 4.3], and an argument of concentration of the threshold of the BH procedure at level α_m. We give a short explanation of the second argument. As already noted above, we have $\hat{t}_m^{\mathrm{BH}}(\alpha) = \sup\left\{u \in [0,1], \widehat{\mathbb{G}}_m(u) \geq u/\alpha\right\}$ (3.2). Therefore, the threshold of the BH procedure at level α_m is close to the threshold of its deterministic counterpart where $\widehat{\mathbb{G}}_m$ is replaced by G:

$$t_m^\star(\alpha) = \sup\left\{u \in [0,1], G(u) \geq u/\alpha\right\} .$$

The classification procedure with threshold $t_m^\star(\alpha)$ controls the pFDR. This suggests that a relevant choice for the target (p)FDR level is the level α_m^{opt} satisfying

$$t_m^\star(\alpha_m^{opt}) = t_m^B \tag{4.2}$$

The main result can be stated as follows for the location and scale models with Subbotin distribution:

Theorem 4.1 ([J17], Corollary 4.4)**.** *Take $\zeta > 1$, $\gamma = 1 - \zeta^{-1}$ for the location case and $\zeta \geq 1$, $\gamma = 1$ for the scale case. Consider a ζ-Subbotin density* (2.3) *in the sparsity regime $\tau_m = m^\beta$, $0 < \beta \leq 1$ and assume that C_m is bounded away from 0 and 1. Letting $r_m = \alpha_m + (\log(\alpha_m^{-1}/(\log m)^\gamma))_+/(\log m)^\gamma$, the following holds:*

(i) The pFDR threshold $t_m^\star(\alpha_m)$ is asymptotically optimal if and only if

$$\alpha_m \to 0 \text{ and } \log \alpha_m = o\left((\log m)^\gamma\right), \tag{4.3}$$

in which case it is asymptotically optimal at rate r_m .

(ii) The FDR threshold $\hat{t}_m^{FDR}$ at a level α_m satisfying (4.3) *is asymptotically optimal at rate r_m.*

(iii) Choosing $\alpha_m \propto 1/(\log m)^\gamma$, pFDR and FDR thresholding are both asymptotically optimal at rate $r_m = 1/(\log m)^\gamma$.

When applied to the Gaussian scale model, Theorem 4.1 recovers the asymptotic optimality results obtained by [78], and extends these results by providing explicit convergence rates. Moreover, our result shows that these theoretical properties are not specific to the Gaussian scale model, but carry over to Subbotin location and scale models. These models include the Gaussian location model studied in [122], and the Laplace scale model studied in [124]. Theorem 4.1 suggests an asymptotic choice of $\alpha_m \propto 1/(\log m)^\gamma$ for FDR thresholding, but does not prescribe an explicit value for α_m for a given m. However the level satisfying (4.2) may be explicitly written as a function of the model parameters:

$$\alpha_m^{opt} = \left(1 + m^{-\beta} C_m / G_{1,m}^{-1}(C_m)\right)^{-1} . \tag{4.4}$$

Therefore, one possible choice for α_m in a given model is $\alpha_m = \alpha_m^{opt}(\beta_0, C_0)$, where (β_0, C_0) is an *a priori* value for the unknown model parameters (β, C_m).

4.3 Numerical experiments

We report the results of some of the numerical experiments performed in [J17] the Gaussian location model in order to illustrate the adaptation to unknown sparsity by the FDR classification rule, and to discuss the relevance of the choice of α_m proposed above. We quantify the quality of a thresholding procedure using the relative excess risk

$$\mathcal{E}_m(\hat{t}) = (R_m^c(\hat{t}) - R_m^c(t_m^B))/R_m^c(t_m^B).$$

Figure 4.1 compares the relative excess risks of the Bayes procedure and the BH procedure for several choices of α_m in the Gaussian location model. For each procedure (in rows) and each value of m (in columns), the behavior of the relative excess risk is studied as the (unknown) true model parameters (β, C_m) vary in $[0,1] \times [0,1]$. The threshold of the Bayes procedure depends on the model parameters (β, C_m); the results reported in the first row "Bayes0" correspond to the choice $(\beta, C_m) = (\beta_0, C_0)$. For FDR thresholding, we report the results for the choice $\alpha_m = \alpha_m^{opt}(\beta_0, C_0)$. The values for β_0 and C_0 are arbitrarily chosen as the midpoints of the corresponding intervals, i.e. $(\beta_0, C_0) = (1/2, 1/2)$.

Bayes0 performs well when the sparsity parameter β is correctly specified, and its performance is fairly robust to C_m. However, it performs poorly when β is misspecified, and increasingly so as m increases. The results is markedly different for FDR: FDR thresholding is less adaptive to C_m than Bayes0, but much more adaptive to the sparsity parameter β, as illustrated by the fact that the configurations with low relative excess risk span the whole range of β. For FDR thresholding at $\alpha_m = \alpha_m^{opt}(\beta_0, C_0)$, the fraction of configurations (β, C_m) for which $\mathcal{E}_m \leq 0.1$ increases as m increases. This illustrates the asymptotic optimality of FDR thresholding for classification under sparsity in the Gaussian location model. Similar conclusions hold for the Gaussian scale and the Laplace scale models, see the Supplementary Materials of [J17]).

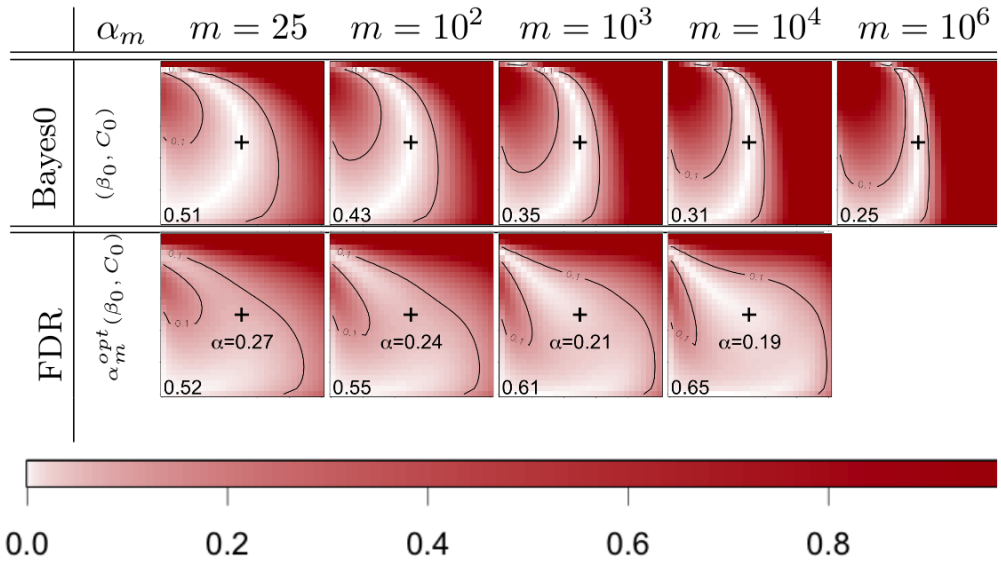


Figure 4.1: Adaptation to sparsity by FDR thresholding in the Gaussian location model. relative excess risks $\mathcal{E}_m$ for various thresholding procedures (rows) and different values of m (columns). In each panel, the corresponding risk is plotted as a function of $\beta \in [0,1]$ (horizontal axis) and $C_m \in [0,1]$ (vertical axis). Colors range from white (low risk) to dark red (high risk), as indicated by the color bar at the bottom. Black lines represent the level set $\mathcal{E}_m = 0.1$. The point $(\beta, C_m) = (\beta_0, C_0)$ is marked by "+". We chose $\beta_0 = 1/2$ and $C_0 = 1/2$. See main text for details.

4.4 Extensions

We have also shown in the Supplementary Materials of [J17] that these results can be extended to a more general model where the density of the test statistics under the alternative is log-concave. Following our work [J17], several papers have studied the performance of the BH procedure for classification under sparsity in other settings. In Frommlet and Bogdan [66], the test statistics are independently and identically distributed as a mixture between $\mathcal{N}(0,1)$ (corresponding to a null hypothesis) and a convolution between a measure ν and the $\mathcal{N}(0,\sigma^2/n)$ distribution (corresponding to the alternative distribution), with a known σ^2. Other contributions have studied monotone polynomial tail distributions including the Student's t, Pareto, and Inverse Gamma distributions [70, 6, 55][1]. Another interesting contribution in the same area of research is [34], where the Gaussian scale models with so-called one-group shrinkage priors. In all these contributions, the authors establish the asymptotic optimality of the BH procedure (without studying the convergence rates).

[1] Two of these manuscripts seem to have merged into the third one [55].

Chapter 5

Post hoc inference via multiple testing

We introduce a generic framework to perform simultaneous inference, based on the control of a risk called the Joint Error Rate. This framework provides a unified view of post hoc inference methods. We propose two types of JER controlling procedures: procedures that are valid under arbitrary dependence under a randomization assumption, and procedures that are valid under independence but dedicated to the case of locally-structured signals.

References:

[**P1**] G. Blanchard, P. Neuvial, and E. Roquain. *On agnostic post hoc approaches to false positive control*. Book chapter in revision for Handbook of Multiple Comparisons; available from `https://hal.archives-ouvertes.fr/hal-02320543`. Oct. 2019

[**J2**] G. Blanchard, P. Neuvial, and E. Roquain. "Post Hoc Confidence Bounds on False Positives Using Reference Families". *Annals of Statistics* 48.3 (2020), pp. 1281–1303

[**J3**] G. Durand, G. Blanchard, P. Neuvial, and E. Roquain. "Post hoc false positive control for structured hypotheses". *Scandinavian Journal of Statistics* (2020)

[**S1**] G. Blanchard, G. Durand, P. Neuvial, and E. Roquain. *sansSouci: Post Hoc Multiple Testing Inference*. R package version 0.8.0. 2019

Contents

Goeman and Solari [81] have introduced the appealing notion of post hoc inference as a means to close the gap between the exploratory nature of data-driven research, and to address the relative inflexibility and difficulty of interpretation of the control of classical multiple testing risks such as the FDR. We refer so Section 2.3.4 and Figure 2.1 for an illustration of the latter point. Their proposed procedure rely on the availability of (i) a valid local test for all possible intersection hypotheses, and (ii) a computational shortcut to bypass the exponential complexity of the calculation of the bound.

In this chapter, we propose another construction of post hoc inference procedures, where the post hoc guarantee is obtained, by interpolation, from the control of a new multiple testing risk called the Joint Error Rate (JER). After taking a closer look at the construction by Goeman and Solari [81] in Section 5.1, we introduce the notion of JER and its properties regarding post hoc inference (Section 5.2), and make a number of connections with previous approaches. In Section 5.4 we introduce a relatively generic way to obtain sharp JER control by adapting to the unknown dependency in the data. We have tried in this presentation to emphasize the ideas and practical implications of the results of [J2]. Finally, in Section 5.5 we describe an extension of this work to the specific case of structured hypotheses (e.g. in time or space)

5.1 State of the art and motivation

To motivate this chapter, we go back to the example of differential gene expression analysis of leukemia samples, introduced in Chapter 1. The state-of-the-art approach to this problem consists in performing one statistical test of no difference between means for each gene, and to derive a list of "significant" genes according to a multiple testing criterion, usually the FDR. As argued in Section 2.4.1, a common practice in the biomedical literature is then to only retain those genes whose "fold change" (that is, the ratio of mean expression levels between the two groups) exceeds a prescribed level [153]. This is illustrated by Figure 5.1, where each gene is represented as a point in the $(\log(\text{fold change}), -\log(p))$ plan. This representation is called a "volcano plot" in the biomedical literature. In this example, 163

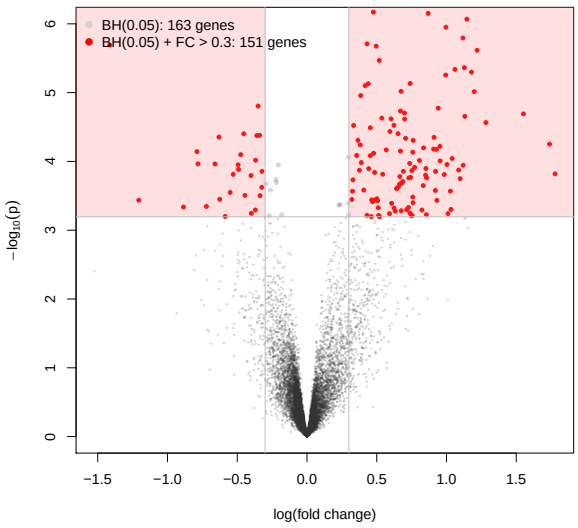


Figure 5.1: Volcano plot.

genes were selected by the BH procedure at level 0.05. 151 of these 163 genes have an absolute log fold change larger than 0.3, but because FDR is an expected *proportion*, it

is not stable by selection, so that we have no statistical guarantee on these 151 genes. Following [81], our goal in this chapter is to provide statistical guarantees on such user-defined selections, in the form of (PH_α), which was already stated in Section 2.4.1.

$$\mathbb{P}\Big(\forall S \subset \mathbb{N}_m,\ |S \cap \mathcal{H}_0(P)| \leq V_\alpha(S)\Big) \geq 1-\alpha\,. \tag{PH_α}$$

We recall that for $P \in \mathcal{P}$, $\mathcal{H}_0(P)$ denotes the set of (indices of) true null hypotheses satisfied by P, that is, $\mathcal{H}_0(P) = \{i \in \mathbb{N}_m\ :\ P \in H_{0,i}\}$, and $\mathcal{H}_1(P) = \mathbb{N}_m \setminus \mathcal{H}_0(P)$. In words, $V_\alpha(S)$ is an $(1-\alpha)$-upper confidence bound on the number of false positives in S. Using the closed testing machinery, Goeman and Solari [81] have proposed a generic bound:

$$V_\alpha^{\mathrm{GS}}(S) = \max_{J \not\subset R^{\mathrm{ct}}} |S \cap J| = \max\big\{|I| : I \subset S, I \notin R^{\mathrm{ct}}\big\} \tag{5.1}$$

where the second equality holds because R^{ct} (the set of intersection hypotheses rejected by the closed testing procedure, defined in (2.12)) is stable by the superset operation. This bound exploits *non-consonant rejections* of the closed testing procedure. A non-consonant rejection is a subset $I \subset \mathbb{N}_m$ such that H_I is rejected by closed testing, but no elementary hypothesis $H_i = \mathcal{H}_{0,i}$ for $i \in I$ is rejected by closed testing. Non-consonant rejections are a waste from the FWER control perspective, because FWER control only retains elementary hypotheses, that is, elements of $R^{\mathrm{ct}} \cap \mathcal{H}$. However, they can be useful for post hoc inference, as we now illustrate by reproducing Figure 1 of [81] in Figure 5.2, with $m = 3$ and where the hypotheses rejected by the local test are crossed. In this particular example, all of the hypotheses rejected by the local test are also rejected by the closed testing procedure. The rejection of $H_{\{2,3\}} = \mathcal{H}_{0,2} \cap \mathcal{H}_{0,3}$ is non-consonant, because neither $H_{\{2\}} = \mathcal{H}_{0,2}$ or $H_{\{3\}} = \mathcal{H}_{0,3}$ is rejected by the closed testing procedure. The bound V_α^{GS} tells us not only that $V_\alpha^{\mathrm{GS}}(\{1\}) = 0$ (which we already knew from the classical FWER control), but also that $V_\alpha^{\mathrm{GS}}(\{2,3\}) = 1$, that is, at least $\mathcal{H}_2$ or $\mathcal{H}_3$ is a true positive.

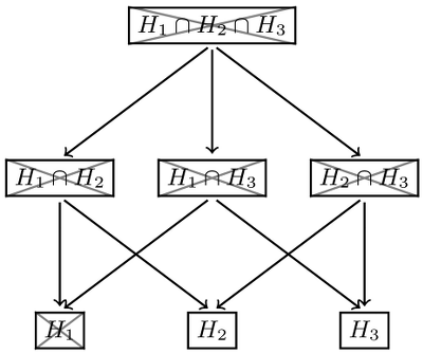


Figure 5.2: [81, Figure 1]: "Intersection hypotheses formed by elementary hypotheses H_1, H_2 and H_3. Rejected hypotheses have been marked with a cross. The rejection of $H_2 \cap H_3$ is a non-consonant rejection."

Goeman and Solari [81] have introduced shortcuts to obtain a computationally efficient post-hoc procedure in the particular case of Simes local tests, thus valid under Assumption ($\mathrm{PRDS}(\mathcal{H}_0)$). An improved shortcut in linearithmic time (that is, $O(m\log(m))$) complexity has recently been proposed [22]. This shortcut is exact, i.e. its output is the bound (5.1) and not an upper bound of it.

Figure 5.3 illustrates the application of this shortcut with $\alpha = 0.1$. In the left plot, this bound is applied to the set of 151 genes previously selected, and also to the two subsets consisting of genes with positive and negative log fold change. Note that this application to three sets is perfectly valid, because the bound in (5.8) is post hoc, ie it holds jointly an all sets of genes of interest. With probability larger than 0.9, we have the following guarantees:

79 of the 151 genes are truly differentially expressed; among the subset of 27 genes which are more expressed in BCR/ABL samples, at least 1 is a true discovery, while among the complementary subset of 123 genes which are less expressed in BCR/ABL samples, at least 62 are true discoveries.

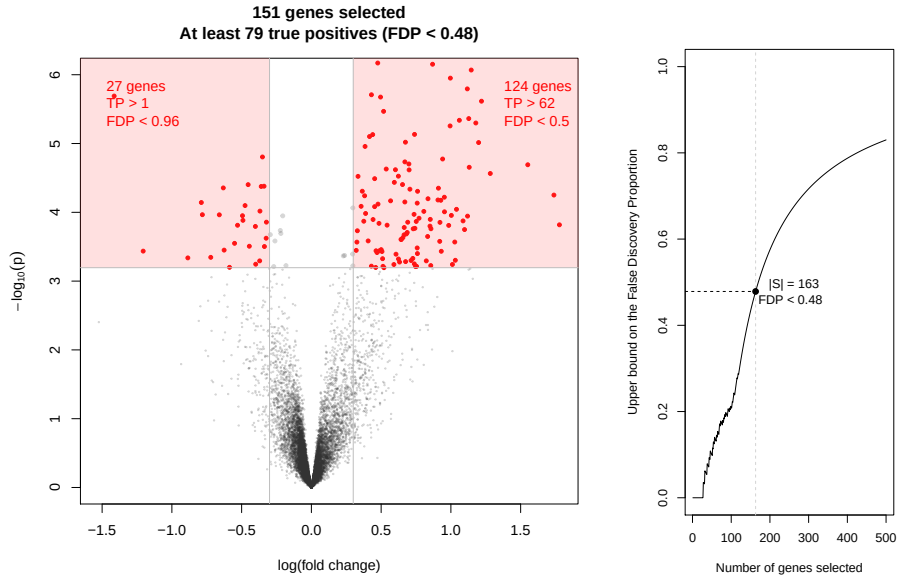


Figure 5.3: Left: volcano plot using post hoc inference; right: corresponding FDP confidence envelope.

The right plot of Figure 5.3 illustrates the use of this shortcut to build a $(1-\alpha)$-level confidence envelope for the FDP among the most significant items. For example, we are 90% confident that the FDP of the 163 genes that were selected by the BH procedure at level 0.05 is less than 0.48. This statement can be obtained to the one that we obtained directly from FDX control (as a simple by-product of Markov's inequality) in Section 2.3.4, that is, $\mathrm{FDX}(R^{\mathrm{BH}(0.05)} \geq 0.5) \leq 0.1$. The fundamental difference is that post hoc bounds yield confidence statements on arbitrary subsets of $\mathbb{N}_m$, not only $S = R^{\mathrm{BH}(0.05)}$.

5.2 Joint Error Rate

5.2.1 Definition and properties

Our proposed construction of post hoc inference procedures relies on a risk measure that we term *Joint (Family-Wise) Error Rate* (JER). In this section we give the definition of this risk measure and explain how in can be used to build post hoc confidence bounds.

Definition 5.1. *Consider a family* $\mathfrak{R} = (R_k, \zeta_k)_{k \in \mathbb{N}_K}$*, for some integer* $K \geq 1$*, with* $R_k \subset \mathbb{N}_m$ *and* $\zeta_k \in \mathbb{N}$*. We define the Joint Family-Wise Error Rate (JER) of* $\mathfrak{R}$ *as*

$$\mathrm{JER}(\mathfrak{R}) = \mathbb{P}\big(\exists k \in \mathbb{N}_K, |R_k \cap \mathcal{H}_0| > \zeta_k\big) \tag{5.2}$$

The family $\mathfrak{R}$ *is said to control JER at level* $\alpha \in [0,1]$ *if:*

$$\mathbb{P}\big(\forall k \in \mathbb{N}_K, |R_k \cap \mathcal{H}_0| \leq \zeta_k\big) \geq 1 - \alpha. \tag{5.3}$$

In the above definition, we allow $\mathfrak{R}$ to be a data-dependent family with $R_k(X), \zeta_k(X)$, but omit the dependence in X to ease notation. If we denote by $\mathcal{A}(\mathfrak{R}) = \{A \subset \mathbb{N}_m :$

$\forall k, |R_k \cap A| \leq \zeta_k\}$, then the functional $V^*_{\mathfrak{R}}$ defined by

$$V^*_{\mathfrak{R}}(S) = \max\{|S \cap A| : A \in \mathcal{A}(\mathfrak{R})\}\,, \;\; S \subset \mathbb{N}_m \tag{5.4}$$

satisfies (PH_α) under (5.3). Moreover, it is optimal in the sense that it is the smallest upper bound on $|S \cap \mathcal{H}_0|$ which is valid for any S under (5.3), because $\mathcal{H}_0$ could be any subset A of $\mathbb{N}_m$ satisfying $\forall k, |R_k \cap A| \leq \zeta_k$. However, it can be shown that the problem of computing $V^*_{\mathfrak{R}}(S)$ given an arbitrary reference family $\mathfrak{R}$ and $S \subset \mathbb{N}_m$, is NP-hard [J2, Proposition 2.3]. To overcome this computational complexity, we introduce the following coarser bound:

$$\overline{V}_{\mathfrak{R}}(S) = |S| \wedge \min_{k \in \mathbb{N}_K} |S \cap R^c_k| + \zeta_k\,, \;\; S \subset \mathbb{N}_m\,. \tag{5.5}$$

It is easy to see that $\overline{V}_{\mathfrak{R}}$ satisfies (PH_α) when $\mathfrak{R}$ controls JER. Indeed, we note that on the event of probability greater than $1-\alpha$ on which (5.3) holds, we have for any $k \in \mathbb{N}_K$ and any $S \subset \mathbb{N}_m$,

$$\begin{aligned} |S \cap \mathcal{H}_0| &= |S \cap R^c_k \cap \mathcal{H}_0| + |S \cap R_k \cap \mathcal{H}_0| \\ &\leq |S \cap R^c_k| \qquad\quad + |R_k \cap \mathcal{H}_0| \\ &\leq |S \cap R^c_k| \qquad\quad + \zeta_k\,, \end{aligned}$$

where we have used (5.3) in the last inequality. Proposition 5.2 below formalizes the general link between JER control and the associated post hoc bounds:

Proposition 5.2 ([J2])**.** *Let $\alpha \in [0,1]$. The family $\mathfrak{R}$ controls JER at level α if and only if $\overline{V}_{\mathfrak{R}}$ or $V^*_{\mathfrak{R}}$ satisfies* (PH_α).

5.2.2 Connection to confidence envelopes and FDX control

The JER framework makes it possible to draw interesting connections with the concepts on confidence envelopes [149, 125, 129, 130], and closed testing [81]. Some of these connections are also made in a recent paper by Katsevich and Ramdas [3]. This point is developed in detail in Section S-1 of the supplement of the paper [J2]. Here, we only summarize the main points regarding these connections.

Confidence envelopes and quantile bounding functions. Uniform upper confidence bounds on the empirical distribution function of null p-values are deduced by [149, 129, 130, 3] from probabilistic guarantees of the form

$$\mathbb{P}\left(\forall t \in [0,1] \;:\; |\mathcal{H}_0 \cap \widetilde{R}_t| \leq B(t)\right) \geq 1-\alpha\,,$$

where $\widetilde{R}_t = \{i : p_i \leq t\}$ denotes a p-value level set. The above guarantee assumption can be interpreted as a specific JER control, and the bounds derived in [130] can be obtained by $\overline{V}$ in (5.5) under natural assumptions.

Augmentation. Augmentation procedures [125] are based on the control of the k-FWER on a specific set R_k, and an interpolation from this control to all possible rejection sets. The augmentation bound of [125] may be written as $V^{\mathrm{aug}}(S) = \min(|S \cap R^c_k| + (k-1), |S|)$. This is a particular case of the bound $\overline{V}$ in (5.5), when the reference family only consists of a single element ($R_k, \zeta_k = k-1$). As noted by [125], the augmentation approach of [152] can itself be seen as the particular case when $k = 1$. Therefore, the bound $\overline{V}$ in (5.5) can be seen as an extension of the augmentation principle.

Inversion and closed testing. Inversion procedures [125] are based on family of local tests for all intersection hypotheses, which also form the basis over which [81] is built. The generic post hoc bound obtained in [81] is in fact equivalent to the inversion bound that can be derived from Genovese and Wasserman [125, Equation (10)], even though the latter relies only on local intersection tests, and not on closed testing.

Moreover, the inversion procedure of [125] can be seen as a particular case of the bound V^*, for the choice $\zeta_k = |R_k| - 1$ which corresponds to a FWER guarantee in each R_k provided by local intersection tests. Conversely, JER control based on a reference family can be embedded as a particular case of the local intersection test framework, by defining the local test of an intersection hypothesis H_I as $\phi_I = 0$ if and only if $\forall k \in \mathbb{N}_K,\ |R_k \cap I| \leq \zeta_k$.

To summarize these connections, with the formalism of JER control, the inversion procedure of [125], and therefore the post hoc bound of Goeman and Solari [81], can be seen as a particular case of the bound V^*, while the bound $\overline{V}$ can be seen as an extension of the augmentation principle [125].

Structural assumption on the reference family

In the next sections we further study JER control, under additional assumptions on the reference family $\mathfrak{R}$:

- in Sections 5.3 and 5.4, we make two further assumptions: (i) $\zeta_k = k - 1$, (ii) $\mathfrak{R}$ is nested, in the sense that for any $k \leq k'$, $R_k \subset R_{k'}$. In this case, it can be shown [J2, Proposition 2.5] that $\overline{V}_{\mathfrak{R}}$ coincides with the optimal bound $V^*_{\mathfrak{R}}$. In the light of the above connections, this property can be seen as an extension of the equivalence between inversion and augmentation procedures stated in [125, Theorem 5].
- In Section 5.5 we take a drastically different point of view in order to address the case where the hypotheses are linearly ordered (e.g. in time or space). We assume that the (R_k) are given and deterministic, and that any two elements of the reference family are either disjoint or nested. The corresponding ζ_k are then random and have to be calibrated from the data in order for (5.3) to be satisfied.

5.3 JER control via Simes inequality

We assume in this section that $\mathfrak{R}$ is nested. Under this assumption, it can be shown that the bounds $\overline{V}$ and V^* coincide, i.e. that the bound $\overline{V}$ is optimal. Moreover, we assume that $\zeta_k = k - 1$. In this setting, it is natural to consider thresholding-based reference families based on p-value level sets, that is, families of the form

$$R_k = \{i \in \mathbb{N}_m \,:\, p_i \leq t_k\}\,,\ \ k \in \{1, \ldots, K\}, \tag{5.6}$$

where $t_k \in \mathbb{R}$, $1 \leq k \leq K$. In view of Proposition 5.2, the main challenge in order to obtain post hoc bounds is to identify a suitable reference family $\mathfrak{R}$ that controls JER at some prescribed level α.

5.3.1 Obtaining Simes' shortcut by a particular JER control

JER control is related to $p_{(k:\mathcal{H}_0)}$, the k-th smallest value in the set $\{p_i, i \in \mathcal{H}_0(P)\}$:

$$\mathrm{JER}(\mathfrak{R}) = \mathbb{P}\Big(\exists k = 1, \ldots, K \wedge m_0,\ p_{(k:\mathcal{H}_0)} \leq t_k\Big). \tag{5.7}$$

The most natural example of a thresholding reference family is the Simes threshold family, defined by $R_k(\alpha) = \{i \in \mathbb{N}_m : p_i \leq \alpha k/m\}$ and $\zeta_k = k-1$ for $k \in \mathbb{N}_m$ (here, $K = m$). A straightforward consequence of the Simes and Hommel inequalities (Proposition 2.3)

is that the Simes family controls JER at level α under (PRDS($\mathcal{H}_0$)), while the Hommel family $R_k(\alpha/C(m))$ controls JER at level α under general dependence. We recall that $C(m) = \sum_{i \in \mathbb{N}_m} i^{-1}$. The corresponding post hoc bound according to (5.5) is given by

$$\overline{V}_{\mathfrak{R}(\alpha)}(S) = |S| \wedge \min_{1 \le k \le |S|} \left\{ \sum_{i \in S} \mathbf{1}\left\{ p_i > \frac{\alpha k}{m} \right\} + k - 1 \right\} \tag{5.8}$$

for $\alpha \in (0,1)$ and $S \subset \mathbb{N}_m$. Proposition 5.2 entails the following result:

Corollary 5.3. *For any $\alpha \in (0,1)$,*

1. *$\overline{V}_{\mathfrak{R}(\alpha)}$ satisfies* (PH_α) *under PRDS($\mathcal{H}_0$);*

2. *$\overline{V}_{\mathfrak{R}(\alpha/C(m))}$ satisfies* (PH_α) *under general dependence.*

This result illustrates the potential of JER control as a generic device to obtain post hoc bounds: we have obtained an easy-to-compute post-hoc bound as an elementary consequence of the Simes-Hommel inequality. It turns out that the bound (5.8) is the bound obtained by the Simes shortcut in [81]. We refer to [J2, Section S-1.4] for a proof.

In our construction, the problem of finding a JER controlling family replaces the two-step construction of [81]: (i) find a valid (and usually computationally prohibitive) post hoc bound under a specific probabilistic assumption, and (ii) identify shortcuts to calculate this bound efficiently. We find the formulation (5.8) to be simple and natural: for each k, adding $k-1$ to the number of p-values in S larger than $\alpha k/m$ (i.e., not rejected by the generalized Bonferroni procedure for k-FWER control) is a natural upper bound on the number of false positives in S.

5.3.2 Limitations of Simes-based post hoc bounds

We report a small simulation study in the same Gaussian ρ-equi-correlated model as in Section 2.3.4, which illustrates the possible conservativeness of the Simes inequality. We consider a "white" setting, that is, all null hypotheses are true, $|\mathcal{H}_0| = m = 1,000$. In Table 5.1, we quantify the conservativeness of the Simes inequality as the ratio of its size (that is, the level actually achieved by the left-hand side in inequality (2.8)) to the target level α. Here, the size is estimated from $1,000$ replications. The fact that the size is very close to the target level for $\rho = 0$ illustrates the sharpness of the Simes inequality under independence. However, for $\rho = 0.4$, the achieved level is only 42% of the target level. Similarly, the Hommel inequality is sharp in the sense that there exists a worst-case p-value

Equi-correlation level: ρ	0	0.1	0.2	0.4	0.8
Size (achieved level)	0.99	0.85	0.72	0.42	0.39

Table 5.1: Conservativeness of Simes inequality in the Gaussian equi-correlated model. Here, $|\mathcal{H}_0| = m = 1,000$ and $\alpha = 0.2$.

distribution such that it is an equality, but it is typically quite conservative when applied to a specific p-value distribution. By construction, the bounds obtained in Corollary 5.3 inherit the properties of the Simes and Hommel inequalities. In particular, the Simes bound (5.8) is sharp under independence and conservative under (PRDS($\mathcal{H}_0$)).

Instead of *assuming* a specific dependency setting, we propose in the next section a construction of JER controlling families (and associated post hoc bounds) that take into account the dependency between the tested hypotheses. This idea can be seen as a generalization of the notion of adaptivity to dependence by randomization for FWER control, that we reviewed in Section 2.2.2.

5.4 Adaptive JER control from a reference family

As in the preceding section, we assume in this section that $\mathfrak{R}$ is nested, with $\zeta_k = k - 1$, and we consider thresholding-based reference families.

5.4.1 General construction

We consider a reference family $\mathfrak{R}(\lambda)$ of the form (5.6), based on thresholds $t_k(\lambda)$, $1 \leq k \leq K$, for some functions $t_k : \lambda \in [0,1] \mapsto t_k(\lambda)$. Our goal is to choose $\lambda = \lambda(\alpha)$ in such a way that JER control (5.3) is satisfied. For example, in the above simulation example, we would like when $\rho = 0.4$ to use the Simes reference family with $\lambda(\alpha) = \alpha/0.42$, even if the data-generating model was unknown.

Definition 5.4 (Threshold template)**.** *A template is a family of functions $t_k(\lambda)$, $\lambda \in [0,1]$, $k \in \mathbb{N}_K$, such that $K \in \mathbb{N}_m$ and for all $k \in \mathbb{N}_K$, $t_k(0) = 0$ and $t_k(\cdot)$ is non-decreasing and left-continuous on $[0,1]$. The parameter K is called the size of the template.*

A template can be seen as a spectrum of curves, parametrized by λ. The theoretical results below are given in terms of a generic, fixed template $t_k(\lambda)$, $\lambda \in [0,1]$, $k \in \mathbb{N}_K$. When λ is fixed, we refer to $t_k(\lambda)$, $k \in \mathbb{N}_K$, as thresholds. In our examples and numerical illustrations we will work with the following two templates:

- Linear template: $t_k(\lambda) = \lambda k/m$, $t_k^{-1}(y) = ym/k$;
- Beta template: $t_k(\lambda) = \lambda$-quantile of $\beta(k, m-k+1)$, $t_k^{-1}(y) = \mathbb{P}(\beta(k, m-k+1) \leq y)$.

An illustration for these templates is provided in Figure 5.4.

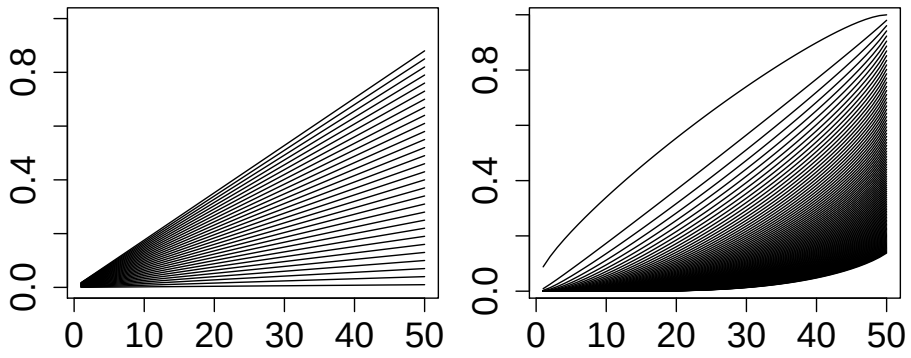


Figure 5.4: Curves $k \mapsto t_k(\lambda)$ for a wide range of λ values. Left: linear template. Right: beta template.

Definition 5.5 (λ-calibration)**.** *Given a threshold template $t_k(\lambda)$, $\lambda \in [0,1]$, $1 \leq k \leq K$, a functional $\lambda(\alpha, A)$, $\alpha \in (0,1)$, $A \subset \mathbb{N}_m$, is called a λ-calibration if it is non-increasing in A and satisfies: $\forall \alpha \in (0,1)$, $\mathrm{JER}(\mathfrak{R}(\lambda(\alpha, \mathcal{H}_0))) \leq \alpha$.*

The question of how to associate a valid λ-calibration to a given template is addressed in the next section. Given such a valid λ-calibration, $\lambda(\alpha, \mathcal{H}_0)$ is still not accessible because $\mathcal{H}_0$ is unknown. However, as λ is non-decreasing in A, we have that $\lambda(\alpha, \mathcal{H}_0) \geq \lambda(\alpha, \mathbb{N}_m)$, so that $\mathrm{JER}(\mathfrak{R}(\lambda(\alpha, \mathbb{N}_m))) \leq \alpha$ as well. Therefore, the template $t_k(\lambda(\alpha, \mathbb{N}_m))$ already gives us an operative reference family for JER control. In order to overcome the possible conservativeness this template due to the fact that $\mathcal{H}_0 \subsetneq \mathbb{N}_m$, we define the following step-down procedure:

Algorithm 1 General step-down algorithm

$j \leftarrow 0$
$A^{(0)} \leftarrow \mathbb{N}_m$
repeat
  $j \leftarrow j + 1$
  $\lambda_j \leftarrow \lambda(\alpha, A^{(j-1)})$
  $A^{(j)} \leftarrow \{i \in \mathbb{N}_m \,:\, p_i(X) \geq t_1(\lambda_j)\}$
until $A^{(j)} = A^{(j-1)}$
return $\hat{A} = A^{(j)}$

By construction, this algorithm leads to a tighter JER control than its single-step version based on $t_k(\lambda(\alpha, \mathcal{H}))$. The proof that it yields JER control at the target level α can be obtained using the classical step-down methodology laid down by [141].

5.4.2 Randomization-based JER control under general dependence

In [J2] we provide valid λ-calibrations both for the case where the observation follows a translation model with *known dependence*, and for the case of general dependence under the randomization assumption (Rand) introduced in Section 2.2. Here, we focus on the second case, which is more practically relevant. Under (Rand), there exists a group $\mathcal{G}$ of transformations acting on the observation set in such a way that the joint distribution of the transformed null p-values is invariant under the action of any element of $\mathcal{G}$. This assumption is satisfied[1] in particular in two important settings:

- location models of the form (2.2) with symmetric noise distribution: $\mathcal{G} = \{-1, 1\}^n$ is the group of signs, which acts on the observation set by the element-wise product;
- two-sample multiple testing problems: $\mathcal{G}$ is the symmetric group of order n (where n is the sample size as in the Leukemia study), which acts on the observation set by permutation of the sample labels. This is the case in the differential gene expression study that we have been using as a working example throughout this part.

Given a $\mathcal{G}$ satisfying (Rand), we consider a (random) $B-$tuple $(g_1, g_2, \ldots, g_B)$ of $\mathcal{G}$ (for some $B \geq 2$), where g_1 is the identity element of $\mathcal{G}$ and $g_2, \ldots, g_B$ have been drawn (independently of the other variables) as i.i.d. variables, each being uniformly distributed on $\mathcal{G}$. Let us consider a deterministic template $t_k(\cdot)$, $1 \leq k \leq K$, and, for short, denote for all $A \subset \mathbb{N}_m$,

$$\Psi(X, A) = \min_{1 \leq k \leq K \wedge |A|} \left\{ t_k^{-1} \left(p_{(k:A)}(X) \right) \right\}.$$

Now introduce the (data-dependent) functional

$$\lambda(\alpha, A) = \max \left\{ \lambda \geq 0 \,:\, B^{-1} \sum_{j=1}^{B} \mathbf{1} \left\{ \Psi(g_j.X, A) < \lambda \right\} \leq \alpha \right\}. \tag{5.9}$$

In practice, we can compute this functional easily as $\lambda(\alpha, A) = \Psi_{(\lfloor \alpha B \rfloor + 1)}$ where $\Psi_{(1)} \leq \Psi_{(2)} \leq \cdots \leq \Psi_{(B)}$ denote the ordered sample $(\Psi(g_j.X, A), 1 \leq j \leq B)$. We have the following result, where $\widehat{A}$ denotes the output of Algorithm 5.4.1.

Theorem 5.6 (λ-calibration for unknown dependence)**.** *Consider any p-value family satisfying* (Rand)*, a deterministic template and the associated reference family* $\mathfrak{R}(\lambda)$*. Then the (data-dependent) functional* $\lambda(\cdot, \cdot)$ *defined by* (5.9) *is a* λ*-calibration in the sense of Definition 5.5, and both* $\mathfrak{R}(\lambda(\alpha, \mathbb{N}_m))$ *and* $\mathfrak{R}(\lambda(\alpha, \widehat{A}))$ *control the JER at level* α*.*

[1] See [J2, Section 2.1 and Appendix S-4] for proofs.

A related idea has been proposed independently by Hemerik, Solari, and Goeman [7] to build confidence envelopes for the False Discovery Proportion. To illustrate Theorem 5.6, we report (in Section 5.4.3) numerical experiments in a Gaussian location model where we use sign-flipping-based λ-calibration, and (in Section 5.4.4) the results of permutation-based λ-calibration in two-sample testing for the Leukemia data set.

5.4.3 Numerical experiments

We report experiments performed with the linear template in the two-sided Gaussian location model under equi-correlation, in the case of an *unknown dependence*. We let n be the sample size. The observations $(X_{i,j})_{i\in\mathbb{N}_m} \in \mathbb{R}^m$, $j \in \mathbb{N}_n$ are distributed as ρ-equi-correlated, and the test statistics for $i \in \mathbb{N}_m$ is defined as $T(X_{i,j}, 1 \le j \le n) = n^{-1/2}\sum_{j=1}^n X_{i,j}$. We use sign-flipping to approximate the joint distribution of the test statistics under the null hypothesis. Specifically, the sign-flipped observation associated to a vector of signs $s \in \{-1,1\}^n$ is defined as

$$(s.X)_{i,j} = s_j X_{i,j},\ i \in \mathbb{N}_m,\ j \in \mathbb{N}_n.$$

The location parameter is set to $\mu_i = n^{-1/2}\overline{\mu}\,\mathbf{1}\{i \in \mathcal{H}_1\}$, where $\overline{\mu} > 0$ quantifies the signal-to-noise ratio. In Figure 5.5, the target JER level α and the level $\pi_0\alpha$ are represented by horizontal lines. Each color corresponds to a different λ-calibration (or absence of calibration for the Simes reference family):

Simes	Single step	Step down	Oracle
α	$\lambda(\alpha, \mathbb{N}_m)$	$\lambda(\alpha, \widehat{A})$	$\lambda(\alpha, \mathcal{H}_0)$

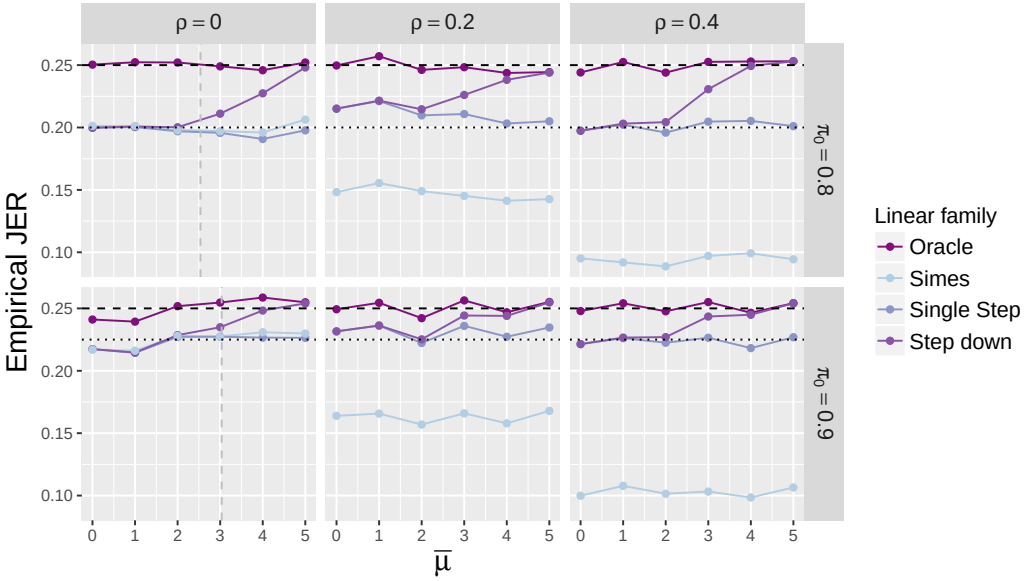


Figure 5.5: JER control based on the linear template for equi-correlated test statistics: λ-calibration by sign-flipping in the location model.

The JER is controlled at the target level α in all situations, as expected from Theorem 5.6. Oracle calibration yields exact JER control, up to sampling fluctuations. As discussed above, the Simes reference family with parameter α yields JER equal to $\pi_0\alpha$ under independence ($\rho = 0$), while it is more conservative under positive dependence $\rho > 0$. Single-step λ-calibration addresses this conservativeness by adapting to the (unknown) dependence: it yields JER control at $\pi_0\alpha$ in all settings considered. Finally, as the signal-to-noise ratio $\overline{\mu}$ gets larger, the step-down λ-calibration yields a JER closer to the nominal level α in

non-sparse situations ($\pi_0 \in \{0.8, 0.9\}$). In a sparse situation ($\pi_0 = 0.99$), corresponding to $m_1 = 10$ true alternative hypotheses, the single-step procedure is already quite sharp and essentially indistinguishable from its Oracle counterpart, so this setting has been omitted from Figure 5.5.

5.4.4 Illustration on the Leukemia data set

Figure 5.6 compares confidence envelopes obtained from the Simes reference family (long-dashed purple curve), to the permutation-based λ-calibration derived from Theorem 5.6 using $B = 1,000$ permutation of the sample labels for the linear template with $K = m$ (dashed red curve) and the beta template with $K = 50$ (solid green curve). Note that Assumption (Rand) holds in this two-sample framework.

While $(1-\alpha)$-lower confidence bounds on the number of true positives of the form $\left\{\left(k, |S_k| - \overline{V}(S_k)\right) : k \in \mathbb{N}_m\right\}$ are displayed in the left panel, $(1-\alpha)$-upper confidence bounds on the proportion of false positives $\left\{\left(k, \overline{V}(S_k)/|S_k|\right) : k \in \mathbb{N}_m\right\}$ are shown in the right panel. The vertical line in Figure 5.6 corresponds to the 163 genes selected by the BH procedure

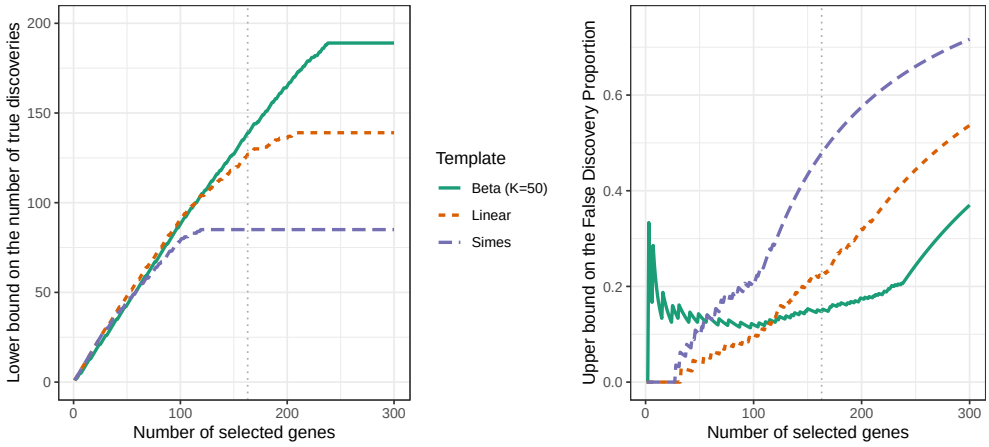


Figure 5.6: Confidence bounds on the number of true positives (left) and on the proportion of false positives (right) for several reference families: Simes reference family (long-dashed purple curve), linear template after λ-calibration (dashed red curve), and beta template after λ-calibration (solid green curve).

at level 5%. The Simes bound ensures that the FDP of this subset is not larger than 0.48. As noted above concerning the BH procedure, we have a priori no guarantee that this bound is valid, because such multiple two-sample testing situations have not been shown to satisfy the PRDS assumption under which the Simes inequality is valid[2]. In contrast, the λ-calibrated bounds built by permutation are by construction valid here. Moreover, both are much sharper than the Simes bound while the λ-calibrated bound using the linear template is twice smaller, ensuring FDP< 0.23, and even smaller for the beta template with $K = 50$. The bound obtained by λ-calibration of the linear template is uniformly sharper that the original Simes bound (5.8), which corresponds to $\lambda = \alpha$. This illustrates the adaptivity to dependence achieved by λ-calibration. The bound obtained from the beta template is less sharp for p-value level sets S_k of cardinal less than $k = 120$, and then sharper. This is consistent with the shape of the threshold functions displayed in Figure 5.4.

To further illustrate the power of λ-calibration to obtain adaptivity to dependence in the classical problem of differential expression studies, Figure 5.7 illustrates the application of the bound derived from the linear template with (single-step) λ-calibration using $B = 1,000$

[2]In this particular case, λ-calibration with the linear template yields $\lambda(\alpha) > \alpha$, which a posteriori implies that the Simes inequality was indeed valid.

random permutations, with $\alpha = 0.1$ (that is, the same bound as the dashed red line in Figure 5.6). Again, the obtained guarantees are substantially more informative than the ones obtained from the Simes reference family with no λ-calibration (left plot of Figure 5.3).

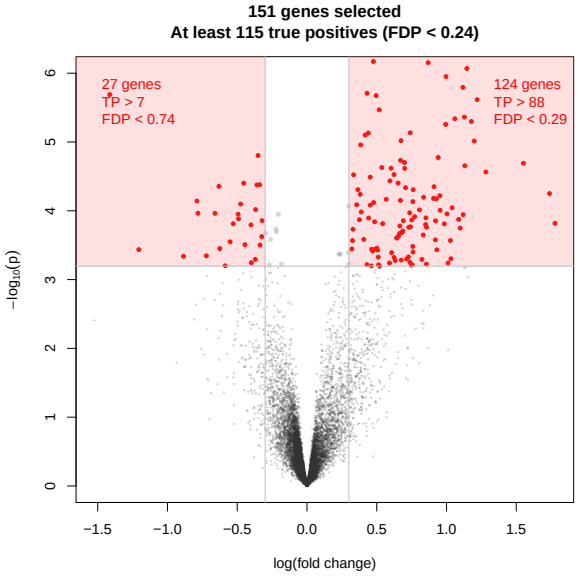


Figure 5.7: Volcano plot using adaptive post hoc inference.

5.5 Spatially-structured hypotheses

In this section we describe a recent work [J3] that illustrates the flexibility of the JER framework introduced in Section 5.2. This contribution stems from a use-case described in [50]. This paper focuses on applications where the hypotheses $(\mathcal{H}_{0,i})_{1\leq i\leq m}$ to be tested are linearly ordered (e.g. in time or space). In such situations, of particular interest are *interval hypotheses*, that is, intersection hypotheses of the form $H_{i:j} = \bigcup_{i\leq k\leq j} \mathcal{H}_{0,k}$. A natural idea in the JER framework is to include such sets in the reference family, in order for the associated post hoc bounds to be particularly sharp on such sets. These candidate sets are deterministic and given *a priori*. The corresponding (possibly random) $\zeta_{i:j}$ achieving JER control have to be calibrated from the data or from probabilistic inequalities. This is in sharp contrast to Sections 5.3 and 5.4, where the ζ_k are deterministic and given a priori, while the R_k are random, and chosen in such a way that JER is controlled. The JER framework described in Section 5.2 accommodates both settings.

5.5.1 Theoretical results

A convenient assumption to accommodate the case of linearly ordered hypotheses is that the reference family satisfies a *forest* assumption, that is, that any two elements of $\{R_k\}_{k\in\mathcal{K}}$ are either disjoint or nested:

Definition 5.7. *A reference family $\mathfrak{R} = (R_k, \zeta_k)_{k\in\mathcal{K}}$ is said to have a forest structure if following property is satisfied:*

$$\forall k, k' \in \mathcal{K}, \quad R_k \cap R_{k'} \in \{R_k, R_{k'}, \varnothing\}, \tag{Forest}$$

If a reference family satisfies (Forest), it can be shown that there exists a partition of the original hypotheses into interval hypotheses called *leaves*, such that each element of the reference family may be written as the disjoint union of consecutive leaves [J3, Lemma 5].

When the reference family is not nested, we do not necessarily have $V^* = \overline{V}$. In particular, under (Forest), the bound $\overline{V}$ may not adequately capture the hierarchical structure of the forest. A natural extension of $\overline{V}$ to the non-nested case is given by the bounds:

$$\widetilde{V}^q_{\mathfrak{R}}(S) = \min_{Q \subset \mathcal{K}, |Q| \leq q} \left(\sum_{k \in Q} \zeta_k \wedge |S \cap R_k| + \left| S \setminus \bigcup_{k \in Q} R_k \right| \right), \quad 1 \leq q \leq K, \quad S \subset \mathbb{N}_m \,. \tag{5.10}$$

By construction, we have $\widetilde{V}^1_{\mathfrak{R}} = \overline{V}$, $\widetilde{V}^q_{\mathfrak{R}}$ is non-increasing in q, and $\widetilde{V}^q_{\mathfrak{R}} \leq V^*$. The main result of [J3] (Theorem 3.6) is that under (Forest), we have

$$V^* = \widetilde{V}^\ell_{\mathfrak{R}} \,,$$

where ℓ is the number of leaves in a partition associated to the reference family. The proof of this result is constructive and therefore provides an algorithm to calculate V^*. The complexity of the computation of $V^*(S)$ for a given S with this algorithm is in $O(dm)$, where d is the depth of the forest. The above construction is generic and yields post hoc bounds tailored to the Forest structure as soon as the reference family controls JER. In particular, this is the case as soon as for each $k \in \mathbb{N}_K$

$$\mathbb{P}\big(|R_k \cap \mathcal{H}_0| > \zeta_k\big) < \alpha/K,$$

that is, as soon as ζ_k is an over-estimator of the proportion of true null hypotheses in R_k at level α/K. A simple way to achieve this without any dependence assumption is to calibrate ζ_k using a Bonferroni test, i.e. to choose

$$\zeta_k = \sum_{i \in R_k} \mathbf{1}\left\{ p_i > \frac{\alpha}{K\,|R_k|} \right\}.$$

While this calibration can be improved by using a Holm-Bonferroni test instead of a Bonferroni test, we can expect the associated post hoc bounds to be conservative because ζ_k is deducted from a FWER control on the hypotheses in R_k, which is more demanding than the mere estimation of the number of true nulls in R_k.

In [J3], we consider an alternative calibration where ζ_k is given by a procedure to directly *estimate* the $\mathcal{H}_0 \cup R_K$. Specifically, we use an estimator inspired by the Storey-λ estimator (2.14). Formally, we define

$$\zeta_k = |R_k| \wedge \min_{t \in [0,1)} \left\lfloor \frac{C}{2(1-t)} + \left(\frac{C^2}{4(1-t)^2} + \frac{\sum_{i \in R_k} \mathbf{1}\{p_i > t\}}{1-t} \right)^{1/2} \right\rfloor^2, \quad k \in \mathcal{K}, \tag{5.11}$$

where $C = \sqrt{\frac{1}{2} \log\left(\frac{K}{\alpha}\right)}$. The [195] inequality with the optimal constant of [175] yields that the associated reference family controls JER under (indep) [J3, Proposition 4.1].

5.5.2 Application to the Leukemia data set

We illustrate these results with a preliminary application of the above bounds to the Leukemia data set. Our biological motivation is the fact that gene expression activity can be clustered along the genome. The m individual hypotheses are naturally partitioned into 23 subsets, each corresponding to a given chromosome. Within each chromosome, we consider sets of $s = 10$ successive genes. Hence, we focus on a reference family with the following elements

$$R_{c,k} = \{(k-1)s + 1, \ldots, \min(ks, m_c)\}, \quad k \in \mathbb{N}_{K_c}, \quad c \in \{1, \ldots, 23\},$$

where, in chromosome c, m_c denotes the number of genes, $K_c = \lceil m_c/s \rceil$ the number of corresponding regions. In addition, for each (c,k) we obtain $\zeta_{c,k}(X)$ by (5.11), with $R_k = R_{c,k}$ and $\alpha = \alpha_c/K_c$. This choice accounts for a union bound over all chromosomes. In this genomic example, (indep) may not hold, so we have no formal guarantee that this bound is valid. Therefore, the results obtained below are merely illustrative of the approach and may not have biological relevance.

We report the results for chromosome $c = 19$, which contains $m_c = 626$ genes. In this particular case, we obtain trivial bounds $\zeta_{c,k}(X) = |R_{c,k}|$ for all $k \in \mathbb{N}_{K_c}$. However, non-trivial bounds can be obtained by enriching the reference family by recursive binary aggregation of the neighboring $R_{c,k}$, as described in detail in [J3]. The total number of elements in the enriched family is less than $2K_c$. In our example, it turns out that (5.10) combined with (5.11) yields 6 true discoveries in the interval $R_{17:24}$ and 1 true discovery in the interval $R_{53:54}$, where we have denoted

$$R_{u:v} = \bigcup_{u \le k \le v} R_{c,k}.$$

This is illustrated by Figure 5.8 where the individual p-values are displayed (on the $-\log_{10}$ scale) as a function of their order on chromosome 19. The sets $R_{17:24}$ and $R_{53:54}$ are highlighted in orange, with the corresponding number of true discoveries marked in each region. We obtain a non-trivial bound not because of the large effect of any individual

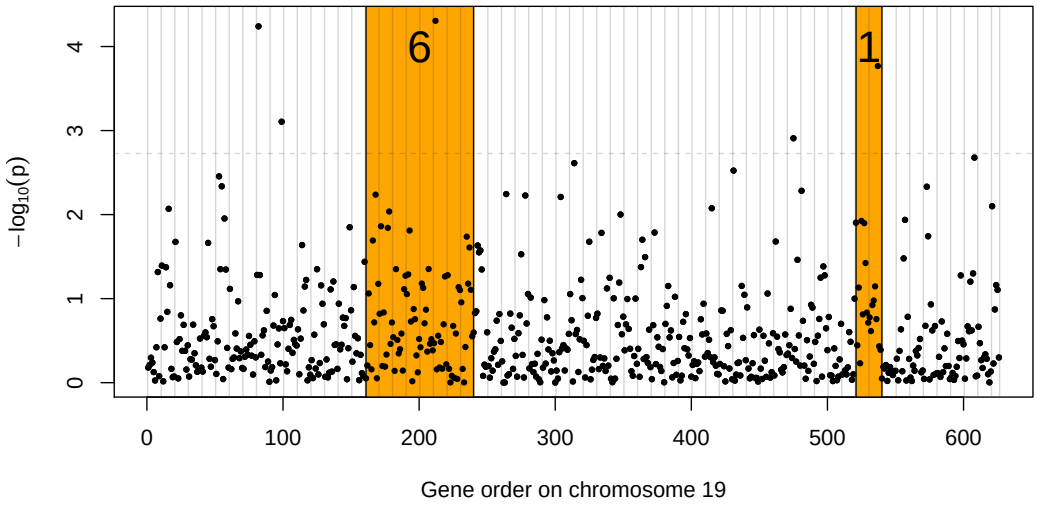


Figure 5.8: Evidence of locally-structured signal on chromosome 19 detected by the bound (5.11).

gene, but because of the presence of sufficiently many moderate effects. In particular, in the rightmost orange region in Figure 5.8, the distribution of $-\log_{10}(p)$ is shifted away from 0 when compared to the rest of chromosome 19. In comparison, we obtain trivial bounds $\overline{V}_{\mathfrak{R}}(R_{53:54}) = |R_{53:54}| = 2s$ and $\overline{V}_{\mathfrak{R}}(R_{17:24}) = |R_{17:24}| = 8s$ from (5.5) both for the linear or the beta template. These numerical results illustrate the interest of these bounds tailored to situations where the signal is expected to be spatially structured.

Part II

Inference from heterogeneous and ordered genomic data

Chapter 6

Overview of contributions

This introductory chapter provides an overview of the contributions detailed in Chapters 7 to 10 and their scientific context. We also provide a short description of contributions that are not detailed further in this part. As announced in Chapter 1, these contributions can be categorized in two broad themes.

- Inference from heterogeneous genomic data: by "heterogeneous", we mean that several levels of biological information are available for the same variables (typically, genes). We focus on the question of identifying relevant features in situations where the sampling units are the same across levels of information or where prior knowledge on the dependency between variables is available in the form of gene networks;
- Inference from ordered genomic data: in this case, an important challenge is to summarize the data locally into biologically meaningful units. This can be addressed by segmenting a genome (or more precisely each of its chromosomes) into successive homogeneous regions.

A guiding principle for my research has been to try to take advantage of these data characteristics in order to build relevant trade-offs between biological interpretation and statistical and computational performance.

Remark: Throughout this part, the number of variables is denoted by p, whereas the number of hypotheses tested was denoted by m in Part I.

6.1 Inference from heterogeneous genomic data

6.1.1 Association between different levels of biological information

We focus on the case where different levels of biological information are observed for the same variables (e.g., genes) and the same observations (e.g. biological samples, or patients). For instance, within the Cancer Genome Atlas (TCGA) project, molecular profiling of each cancer sample is typically performed at least at 3 different biological levels: DNA copy numbers, DNA methylation, and gene expression. After summarizing each of these types of information at the gene scale (an important bioinformatic issue which is not covered in this document except in Section 9.2 for DNA copy numbers), the data is in the form of a $3 \times n \times p$-dimensional array, where n is the number of observations and p the number of genes. A natural question in this setting is to identify variables for which the different levels of biological information are *associated*, in a statistical sense yet to be defined more formally for a specific biomedical question.

In Chapter 7, we tackle the question of identifying genes whose DNA copy number is associated with their expression level, accounting for DNA methylation. The approach proposed in that chapter is a *marginal* (i.e., gene by gene) feature selection method, which

explicitly exploits some of the possible links between these levels of biological information. This approach provides statistical guarantees on the obtained estimates [J14, J8, S5]. This work was initiated while I was a post doc at UC Berkeley, jointly with Antoine Chambaz, then a visiting professor in the Biostatistics Department in the group of Mark van der Laan. It stems from the theory of Targeted Minimal Loss Estimation (TMLE), which has been initiated by van der Laan and Rubin [131].

6.1.2 Incorporating prior knowledge in the form of networks

Genes operate not individually but jointly via biological interaction networks, which are partly known. In order to better understand the mechanisms of cancer development, as well as the response of cancer cells to treatments, biomedical scientists need mathematical tools that exploit this knowledge of functional links between genes. Below we describe two contributions in this field. Only the first one is detailed in a specific chapter.

Chapter 8: Statistical tests on graphs [J16, S6]. One of the most classical application of multiple testing to genomics is the search for differentially expressed genes, as illustrated in Part I. Besides gene expression data, we have at our disposal prior knowledge on functional links between genes, for example via gene regulation networks. However, most existing methods either rely on *marginal* (that is, gene by gene) tests of differential expression, or reduce the biological information at hand to unstructured gene lists, disregarding the *network* information. In Chapter 8 we describe a method to perform multivariate tests of differential expression on graphs [J16, S6]. This method relies on a dimension reduction of gene expression data driven by the graph topology. This work was done while I was a post-doc at Berkeley, in collaboration with Laurent Jacob (also a post-doc at Berkeley at that time) and Sandrine Dudoit.

Identification of deregulated genes and potential regulators [J10, C1]. The mechanisms of regulation of normal cells are known to be altered in cancer cells. We have proposed a method to identify transcription factors involved in such gene deregulations [C1]. Transcription factors (TFs) are proteins that control or regulate gene expression, that is, the transcription of DNA into RNA. As any other protein, TFs are coded by genes, which justifies using their expression to quantify their activity. The method proposed in [C1] consists in three steps. First, a reference gene regulatory network that connects transcription factors to their downstream targets is inferred from gene expression data in a steady (or normal) state. This was done by adapting of the LICORN method [J25][1]. In a second step, the behavior of genes in tumor samples is then compared to this reference network in order to detect deregulated target genes [J10], as detailed in the next paragraph. Finally, the ability of each transcription factor to explain these deregulations is quantified via a linear model. The performance of this three-step strategy has been illustrated by numerical experiments on a TGCA breast cancer data set. These experiments show that the information about deregulation is complementary to the expression data, as the combination of the two improves the performance of the supervised classification of cancer samples.

The second step relies on a statistical method to identify deregulated genes from a reference regulation network, and gene expression data [J10]. We have proposed a model based on a regulatory process in which all genes are allowed to be deregulated, and the hidden variables correspond to the status (under/over/normally expressed) of the genes. The model parameters are estimated via a tailored Expectation-Maximization (EM) [185] algorithm. The resulting method infers posterior probabilities of gene deregulation for each (gene, observation) pair, whereas more classical approaches only produce a gene deregulation score common to all observations. This distinctive feature, which is relevant to biologists,

[1] LICORN has been developed by Mohamed Elati, and I participated to this development during my PhD thesis.

comes at the price of a more tricky model inference step, due to the large number of possible states for the latent variables. This difficulty is bypassed by taking advantage of the fact that the Maximization (M) step of the EM algorithm only requires the marginal distributions of the latent variables to be known. We have shown by factorizing the model likelihood that these marginal distributions may be estimated at the Expectation (E) step of the EM by an algorithm of "Belief Propagation" [155].

This work was done in the context of a collaboration with former colleagues at Évry: Julien Chiquet and Etienne Birmelé at LaMME for the statistical aspects, and Mohamed Elati (Institute of Systems & Synthetic Biology, iSSB). This collaboration started with the support of CNRS (PEPS BMI 2013, PI Etienne Birmelé), and continued thanks to the support of an INSERM grant (2015-2019, PI Mohamed Elati) called LIONS for "Large-scale Integrative approach to unravel the complex relationships between differentiatiON and tumorigenesiS". The work was mainly carried out by Thomas Picchetti [J10] during his PhD thesis and then by Magali Champion during her post-doc [C1].

6.2 Inference from ordered genomic data

Genetic information is coded in long strings of DNA organized in chromosomes. High-throughput sequencing such as RNAseq, DNAseq, ChipSeq and Hi-C makes it possible to study biological phenomena along the entire genome at a very high resolution [52]. In most cases, neighboring positions are expected to be statistically dependent. Using this *a priori* information is one way of addressing the complexity of genome-wide analyses. For instance, it is common practice to partition each chromosome into regions, because such regions hopefully correspond to biological relevant or interpretable units (such as genes or binding sites) and because statistical modelling and inference are simplified at the scale of an individual region. In simple cases, such regions are given (for example genes in gene expression analyses). However, in more complex cases as in Chapters 9 and 10, the regions of interest are unknown and need to be learned from the data.

Recovering the "best" partition of p loci for a given number of classes is a segmentation problem, also known as "multiple change point problem". Such segmentation problems are combinatorial in nature, as the number of possible segmentations of p loci for K change points (or breakpoints) for a given $K = 1 \dots p-1$ is $\binom{p-1}{K} = O(p^K)$. When $K = o(p)$, the best segmentation for all $k = 1, \dots, K$ in terms of ℓ^2 loss can be recovered efficiently in quadratic time and space complexity using dynamic programming. However, this complexity is typically too large for genomic applications, where $p \sim 10^4 - 10^6$. Therefore, an important challenge is to build statistically sound methods with sufficiently low algorithmic complexity and that provides biologically interpretable results.

Chapter 9: Statistical inference from DNA copy number data. The simplest scenario where the signals to be segmented are piecewise-constant is described in Chapter 9, which is dedicated to the analysis of DNA copy numbers in cancer studies. In this case, segmentation can be cast as a least squares minimization problem [139, 67]. I have started working on the statistical analysis of DNA copy numbers in cancers before my PhD thesis, when I got hired as a research engineer at the Curie Institute. At that time and during my PhD thesis, I developed data normalization methods [J28, S7], contributed to the development of data analysis pipelines [J27, s6], visualization methods [J26, s5] and to a review of clustering methods [J12]. I also worked on the development of a method to distinguish new primary tumors from true recurrences on the basis of copy-number profiles from a patient suffering from a second breast cancer [J23]. Copy numbers turned out to also have a central part in my postdoc with Terry Speed at UC Berkeley, thanks to a long-standing collaboration with Henrik Bengtsson (now an Associate Professor at UCSF). Together we developed a normalization method [J21, s3], and contributed to the development of another one [J18, s2] as well as segmentation method [J19, s1]. We also coauthored a

book chapter on the statistical analysis of DNA copy numbers [B1], and participated to larger research projects in the context of the Cancer Genome Atlas project (TCGA) that funded my post doc [J20, J22, J15] I have continued working on this topic since I came back to France [J11, S4, S2], in particular with Guillem Rigaill and during the PhD thesis of Morgane Pierre-Jean (2013-2016) [38], which I co-supervised with Catherine Matias.

Chapter 9 presents some statistical insights I gained from working with DNA copy number data. It is mostly written as a review chapter, with contributions posterior to my PhD thesis mentioned along the way in varying levels of detail: normalization methods in Section 9.2, a review of segmentation method in Section 9.3, and a performance evaluation framework in Section 9.4. Several software contributions are also highlighted.

Chapter 10: Adjacency-constrained clustering for Hi-C and GWAS. This chapter tackles the more general situation of data described via a similarity measure, as in the case for linkage disequilibrium in GWAS, and contact maps in Hi-C studies. The segmentation problem consists in finding a common partition of rows and columns of a matrix of similarity between objects, such that the signal is mostly concentrated in the diagonal blocks resulting from the partition. This type of segmentation problems can be tackled by kernel-based segmentation methods [118, 29]. However, the quadratic time complexity of these methods cannot be improved without making additional assumptions on the kernel [16]. Indeed, for a generic kernel, even computing the loss (e.g., the least square error) of any given segmentation into a fixed number of segments has a computational cost of $O(p^2)$.

Chapter 10 summarizes a series of works initiated by the PhD thesis of Alia Dehman (2012-2015) [44], which I co-supervised with Christophe Ambroise. Motivated by the idea of taking into account the structure induced by linkage disequilibrium (LD) in GWAS, this thesis lead to the development of a method to detect blocks of LD associated to a given phenotype [J9] (Section 10.4). An important component of this method is an algorithm called `adjclust` dedicated to the above-described problem of segmenting a similarity matrix (Section 10.3), and its application to the detection of LD blocks. Since I moved to Toulouse in 2016, we started with Nathalie Vialaneix (INRA MIA-Toulouse) to work on the problem of detecting Topologically Associated Domains (TAD) from Hi-C data, for which this algorithm is also relevant [J5, S3]. This collaboration was supported by a grant from the Mission for Interdisciplinarity (MITI) at CNRS (SCALES project, 2017-2019). A major difference between GWAS and Hi-C from a statistical perspective is that while the LD similarity is a kernel, it is not necessarily the case for Hi-C similarity matrices, possibly leading to reversals in the constrained HAC. This point, which is briefly mentioned in Section 10.2 has been studied by Nathanaël Randriamihamison [J1] during the first year (2018-2019) of his PhD thesis. This thesis is co-supervised by Nathalie Vialaneix, Marie Chavent (Université de Bordeaux and Inria) and myself and funded by a joint INRA/Inria doctoral program.

Chapter 7

Targeted minimal loss estimation

Looking for genes whose DNA copy number is "associated with" their expression level in a cancer study can help pinpoint candidates implied in the disease and enhance our understanding of its molecular bases. Genomic covariates may play an important role in the biological process and should therefore be taken into account. For instance, DNA methylation is known to regulate gene expression. To quantify the association between DNA copy number and expression, accounting for such relevant genomic covariates, we propose to define a new parameter of interest, and build a method to infer it upon the targeted minimum loss-based inference principle (TMLE).

Considering associations between DNA copy numbers and expression levels in genes is not new [159, 113, 90, 87, 84]. In contrast to these earlier contributions, ours does explicitly exploit that DNA copy number measurements feature both a reference level and a continuum of other levels, instead of discretizing them or considering them as purely continuous. Moreover, we naturally handle multi-dimensional, continuous covariates without discretization. We do not need to assume that they are normally distributed, nor that their true effect of DNA copy number on gene expression is linear.

References:

[J8] A. Chambaz and P. Neuvial. "tmle.npvi: targeted, integrative search of associations between DNA copy number and gene expression, accounting for DNA methylation". *Bioinformatics* 31.18 (2015), pp. 3054–6

[J14] A. Chambaz, P. Neuvial, and M. J. van der Laan. "Estimation of a Non-Parametric Variable Importance Measure of a Continuous Exposure". *Electron. J. Statist.* 6 (2012), pp. 1059–1099

[S5] A. Chambaz and P. Neuvial. *Targeted Learning of a Non-Parametric Variable Importance Measure of a Continuous Exposure*. R package version 0.10.0. 2015

Contents

7.1 Defining the parameter of interest

Let $O = (W, X, Y)$ be a generic observation, where W, X and Y are respectively the covariates (*e.g.*, DNA methylation), DNA copy number and expression of a gene of interest in a randomly picked biological sample. Let x_0 be a reference value for X, corresponding to the normal state of 2 DNA copies. We assume that the probability to observe $X = x_0$ is bounded away from 0 and 1. We assume without loss of generality that $x_0 = 0$. In the absence of additional solid knowledge regarding the law of O, we decide to focus on the following non-parametric variable importance measure Ψ. It is a mapping from $\mathcal{M}$, the set of all laws compatible with the definition of O, to $\mathbb{R}$ given by

$$\Psi(P) = \underset{\beta \in \mathbb{R}}{\arg\min}\, \mathbb{E}_P \left\{ \left(Y - \mathbb{E}_P(Y|X = 0, W) - \beta X \right)^2 \right\}. \tag{7.1}$$

This parameter is a measure of the *importance* of X relative to Y, accounting for W. Moreover, it is a *non-parametric* measure as it is defined regardless of a semi-parametric model of the form $Y = \beta(X - x_0) + \eta(W) + U$, with unspecified η and U such that $\mathbb{E}_P(U|X, W) = 0$). Therefore, the parameter of interest, $\Psi(P_0)$, is universally defined no matter what properties the unknown true data-generating distribution P_0 enjoys, or does not enjoy.

Interpretation in terms of risk excess. For all $P \in \mathcal{M}$, $\Psi(P)$ may alternatively be written [J14, Proposition 1] as:

$$\Psi(P) = \frac{\mathbb{E}_P\{X(\theta(P)(X, W) - \theta(P)(0, W))\}}{\mathbb{E}_P\{X^2\}}, \tag{7.2}$$

where $\theta(P)(X, W) = \mathbb{E}_P(Y|X, W)$. Therefore, the functional Ψ may be interpreted as a generalization of the notion of *risk excess* to a continuous X. Indeed, when $X \in \{0, 1\}$, we have $\Psi(P) = \mathbb{E}_P\{(\theta(P)(x_1, W) - \theta(P)(0, W))h(P)(W)\}$, where $h(P)(W) = P(X = x_1|W)/\mathbb{E}_P\{X^2\}$. That is, $\Psi(P)$ is a weighted version of the classical risk excess: $\mathbb{E}_P\{\theta(P)(x_1, W) - \theta(P)(0, W)\}$, where the weights are given by h.

7.2 Targeted Minimal Loss Estimation

General principle. As Ψ is known, a substitution estimator $\psi_n = \Psi(P_n)$ may be derived from any estimator P_n of the law P_0 of the observations. We propose to construct one such estimator by relying on the theory of Targeted Minimal Loss Estimation (TMLE), which is described in detail in [88, 24]. This theory offers a generic principle of iterative estimation which may be summarized as follows. The first step consists in building a substitution estimator $\psi_n^0 = \Psi(P_n^0)$ from an initial estimator P_n^0 of P_0. The estimator ψ_n^0 is then used to build an updated estimator of P_0, denoted by P_n^1, which itself induces an update of the substitution estimator: $\psi_n^1 = \Psi(P_n^1)$.

The updating step is at the core of the TMLE approach. It relies on the notion of *efficient influence curve*, which is pivotal in semi-parametric estimation [170]. It can be shown that the parameter Ψ defined in (7.1) is pathwise differentiable: for every $P \in \mathcal{M}$, there exists a function $\nabla\Psi_P$ of O such that for any bounded $s : O \mapsto s(O)$ and all $|\varepsilon| < \|s\|_\infty^{-1}$, if we characterize $P_\varepsilon \in \mathcal{M}$ by setting $dP_\varepsilon/dP = 1 + \varepsilon s$ then

$$\Psi(P_\varepsilon) = \Psi(P) + \varepsilon \mathbb{E}_P\{\nabla\Psi_P(O)s(O)\} + o(\varepsilon).$$

The map $\nabla\Psi$ is called the efficient influence curve of Ψ. Informally, and as it names suggests, $\nabla\Psi_P$ quantifies the influence of the law P on the estimation of the parameter of interest, and indicates in which *direction* an estimator of P should be refined in order for the quadratic error of the associated substitution estimator to be improved.

Algorithm. A closed-form expression of $\nabla\Psi$ can be derived [J14, Proposition 1] for the parameter Ψ defined in (7.1). The expression of $\nabla\Psi$ involves finite- and infinite-dimensional *features* of P, notably including:

$$\begin{cases} \theta_P(X,W) & = \mathbb{E}_P(Y|X,W) \\ g_P(W) & = P(X=0|W) \\ \mu_P(W) & = \mathbb{E}_P(X|W) \\ \sigma_P^2 & = \mathbb{E}_P\{X^2\} \end{cases}.$$

Moreover, it can be shown that $\nabla\Psi$ is *double-robust*, in the sense that for any $(P,P') \in \mathcal{M}^2$, we have $\Psi(P') = \Psi(P)$ as soon as *either* ($\mu(P') = \mu(P)$ and $g(P') = g(P)$) *or* $\theta(P')(0,\cdot) = \theta(P)(0,\cdot)$.

Let us assume that we observe n independent random variables $O_1, \ldots, O_n$ drawn from P_0. By (7.2), building an initial substitution estimator $\Psi(P_n^0)$ of $\Psi(P_0)$ requires the estimation of θ_{P_0}, of $\sigma^2_{P_0}$, and of the marginal distribution of (W,X) under P_0. It can be shown that the latter can itself be obtained from estimates of g_{P_0} and μ_{P_0}, and the marginal distribution of W under P_n^0, which itself is simply estimated by its empirical counterpart. Therefore, using Monte-Carlo estimation, we can obtain an initial estimator of $\Psi(P_n^0)$ from initial estimates $(\theta_n^0, g_n^0, \mu_n^0, \sigma_n^0)$.

The k^{th} update step (for $k \geq 0$) consists in estimating the efficient influence curve $\nabla\Psi_{P_n^k}$ based on the features θ_n^k, μ_n^k, g_n^k and σ_n^k. This estimation step defines a one-dimensional model $\{P_n^k(\varepsilon) : |\varepsilon| < \|s\|_\infty^{-1}\} \subset \mathcal{M}$ by setting $dP_n^k(\varepsilon)/dP_n^k = 1 + \varepsilon s$ with $s = \nabla\Psi_{P_n^k}$. This model is called a *fluctuation* of the law P_n^k. An updated estimator of P_0 is then defined by $P_n^{k+1} = P_n^k(\varepsilon_n^k)$, where ε_n^k is the maximum likelihood estimator of the fluctuation model. Finally, updated estimations of the features and of the parameter: ψ_n^{k+1} are obtained similarly to the initialization step, i.e. via Monte-Carlo estimation of the marginal law of (X,W). The series of updates is interrupted if the total variation distance $d_{TV}(P_n^k, P_n^{k+1})$, $|\Psi(P_n^k) - \Psi(P_n^{k+1})|$, or $|\sum_{i=1}^n \nabla\Psi_{P_n^k}(O_i)|$ is small. Finally, the TMLE ψ_n^* is defined as $\psi_n^* = \lim_{k\to\infty} \psi_n^k$, assuming that the limit exists, or more generally as $\psi_n^{k_n}$ for a conveniently chosen sequence $\{k_n\}_{n\geq 0}$.

This estimation method has been implemented in the R package `tmle.npvi` [S5] which is available from CRAN[1]. The implementation is described in detail in [J8]. This implementation incorporates estimators of the features (θ, g, μ, σ) based on classical generalized linear models, as well as estimators relying on super-learning, a generic method of aggregation of estimators proposed by [120].

7.3 Inference

Convergence of the iterative procedure. Lemmas 3 and 4 in [J14] ensure that if the sequence (parametrized by the iteration index k) $P_n\nabla\Psi(P_n^k)^2$ is bounded away from 0, then the above iterative procedure converges, in the sense that the sequence $(\varepsilon_n^k)_{k>0}$ tends to 0. If moreover the series $\sum_k |\varepsilon_n^k|$ is convergent, then the sequence of estimators $\psi_n^k = \Psi(P_n^k)$ is also convergent.

Asymptotic properties of the proposed estimator. The TMLE is consistent as soon as the estimators of the features θ, μ, σ and g are convergent, and that one of the estimator of $\theta(0,\cdot)$ or the estimators of μ and of g is consistent [J14, Proposition 2]. This remarkable property is inherited from the double robustness of the efficient influence curve $\nabla\Psi$, combined with the fact that the TMLE solves the efficient influence curve equation (*i.e.*, $P_n\nabla\Psi(P_n^{k_n}) \approx 0$). The TMLE satisfies a central limit theorem [J14, Proposition 3]

[1] `http://cran.r-project.org/package=tmle.npvi`

under additional conditions on the convergence rates of these features:

$$\sqrt{n}(\psi_n^* - \Psi(P_0)) \rightsquigarrow \mathcal{N}\left(0, \mathrm{Var}_{P_0}(\nabla\Psi(P_0)(O))\right) \tag{7.3}$$

Again thanks to the double-robustness of $\nabla\Psi$, the fact that *either* the estimator of θ *or* both the estimators of μ and σ^2 converge at rate $n^{-1/2}$ is sufficient to guarantee the asymptotic normality of the TMLE. Finally, if the estimators of all features are consistent, then the TMLE is asymptotically efficient, and we can construct an estimate of its asymptotic variance.

Contrast to NP estimation. An obvious substitution estimator of $\Psi(P_0)$ is

$$\psi_n^{\mathrm{NP}} = \underset{\beta\in\mathbb{R}}{\arg\min}\, \mathbb{E}_{P_n}\left\{\left(\hat{\theta}_n(X,W) - \hat{\theta}_n(x_0,W) - \beta(X - x_0)\right)^2\right\},$$

an expression derived from (7.1) by substituting the empirical measure P_n for P_0 and the Nadaraya-Watson estimator $\hat{\theta}_n(X,W)$ of $E_{P_0}(Y|X,W)$ for it. Under regularity conditions of order ℓ on the true conditional expectation, the optimal bandwidth h_n for $\hat{\theta}_n$ satisfies $h_n = cn^{-1/(2\ell+d)}$ where $d = 2$ is the dimension of (X,W) [100]. Now it is possible to characterize a $P_0 \in \mathcal{M}$ such that $E_{P_0}\{\psi_n^{\mathrm{NP}} - \Psi(P_0)\} = c'h_n^\ell + o(h_n)$. In particular, ψ_n^{NP} cannot achieve $\sqrt{n}$-consistency under that P_0. We see that ψ_n^{NP} suffers from the fact that the bias-variance trade-off, which is at the core of the construction of $\hat{\theta}_n$, is optimized for the sake of estimating the infinite-dimensional parameter $E_{P_0}(Y|X,W)$ whereas we are eventually interested in estimating the one-dimensional parameter $\Psi(P_0)$.

7.4 Simulations

Because association patterns between copy number, expression and methylation are generally non-linear, setting up a realistic simulation model is a difficult task. We design here a simulation strategy based on perturbations of real observed data structures. Specifically, we focus on the gene `EGFR` in a glioblastoma multiforme (GBM) data set from The Cancer Genome Atlas (TCGA) project [116, 111]. This gene is known to be altered in GBM. Figure 7.1(a) represents W = DNA methylation, X = DNA copy number, and Y = gene expression data for one particular gene, `EGFR`, which is known to be altered in GBM. For this gene, the association between copy number and expression is non-linear, and high methylation levels are associated with low expression levels.

Our simulation strategy implements the following constraints:

- There are generally up to three copy number classes: normal regions, and regions of copy number gains and losses;
- In normal regions, expression is negatively correlated with methylation;
- In regions of copy number alteration, copy number and expression are positively correlated.

Figure 7.1(b) summarizes a simulation run with $n = 200$ independent copies of the synthetic observed data structure based on two specific GBM samples for the `EGFR` gene. Such a simulation study provides an arguably realistic use case where closed-form expressions for the features of interest $\theta(P)$, $\mu(P)$, $g(P)$, and $\sigma^2(P)$ can be derived, and the value of $\Psi(P)$ can be estimated accurately. We estimated this value by $\psi_B(P)$ using $B = 10^5$ Monte-Carlo samples. Gray rectangles represent 95%-accuracy intervals $[\psi_B(P) \pm \xi_{0.975}s_B(P)/\sqrt{n}]$ and $[\psi_B(P) \pm \xi_{0.975}s_B(P)/\sqrt{B}]$ for the true parameter $\Psi(P)$ based on n observed data structures (light gray) and $B = 10^5$ observed data structures (dark gray). Here, ξ_a is the $1-a$ quantile of the standard Gaussian distribution, and $s_B(P) = \mathrm{Var}_{P_B}(\nabla\Psi(P)(O))$ is a

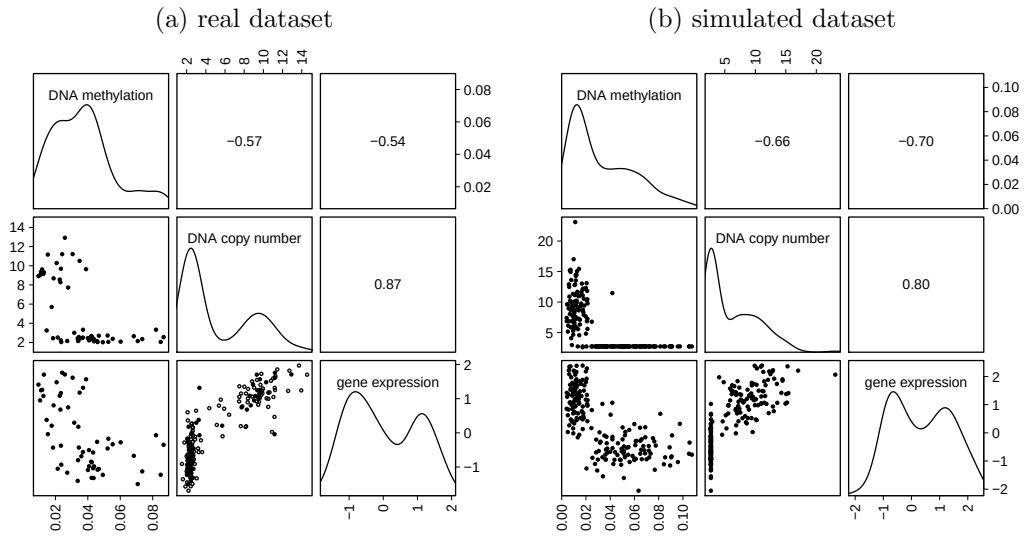


Figure 7.1: Illustrating DNA methylation, DNA copy number, and gene expression data. In both graphics, we represent kernel density estimates (diagonal panels), pairwise plots (lower panels), and report the pairwise Pearson correlation coefficients (upper panels). **(a).** Real dataset corresponding to the `EGFR` gene in 187 GBM tumor samples. For 130 among the 187 samples, only DNA copy number and gene expression data were available (circles in lower middle plot). **(b).** Simulated dataset consisting of $n = 200$ independent copies of the synthetic observed data structure described in the main text. Note that the constant O_2^X is added to each value of X so that graphics corresponding to real and simulated data can be more easily compared.

Monte-Carlo estimator of the asymptotic variance appearing in (7.3), which can be formed because $\nabla\Psi(P)(O)$ is available in closed form in our simulation (see [J14, Lemma 7]).

For each of $B' = 10^3$ simulation runs, we record the parameter estimate obtained after k iterations of the TMLE procedure: $\psi_{n,b}^k = \Psi(P_{n,b}^k), 1 \leq b \leq B'$. The results of this simulation are summarized by Figure 7.2 in the case where the features are learned using algorithms based on generalized linear models. This figure provides the empirical distribution of the parameter estimate across the B' simulation runs. These results illustrate some of the fundamental characteristics of the TMLE estimator and related confidence intervals:

Convergence and robustness. The substantial bias in the initial estimation is diminished (if not perfectly corrected) at the first updating step of the TMLE procedure, illustrating the robustness of the targeted estimator. The empirical distributions of $\{\psi_{n,b}^k : b \leq B'\}$ for $k = 1, 2, 3$ are not (visually) markedly different, an empirical indication that the TMLE procedure converges quickly.

Asymptotic normality, and coverage. The asymptotic normality of the TMLE estimator was confirmed by Lilliefors and Kolmogorov-Smirnov tests. We used the asymptotic variance estimate $(s_{n,b}^k)^2 = \mathrm{Var}_{P_{n,b}} \nabla\Psi(P_{n,b}^k)(O)$ to form confidence intervals. Contrary to s_B^2 above, which requires the knowledge of the simulation parameters, this estimator can be obtained only based on the n observations. While our theoretical results *do not* guarantee that it is safe to estimate the limit variance by $(s_{n,b}^k)^2$, the associated confidence intervals provide empirical coverage close to the nominal coverage.

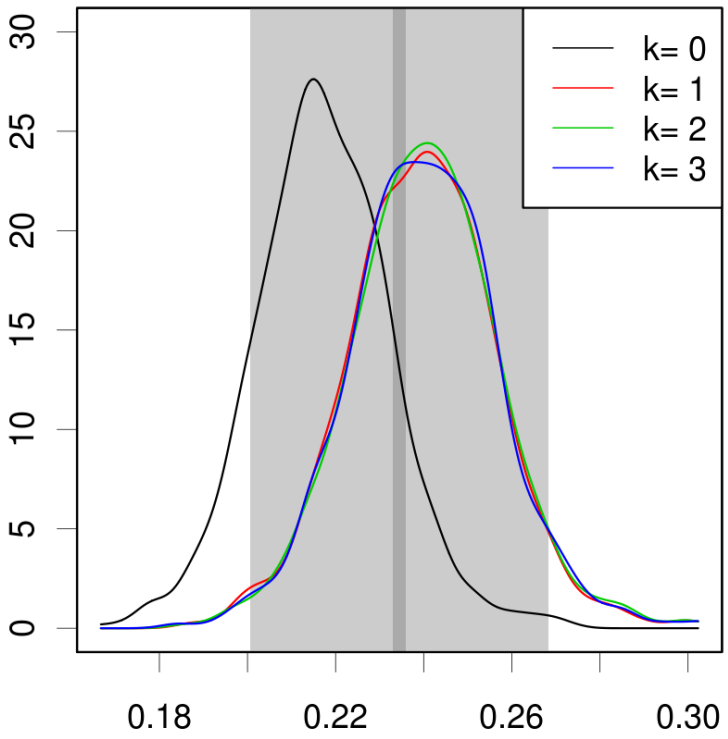


Figure 7.2: Empirical distribution of $\{\psi_{n,b}^k : b \leq B'\}$ based on $n = 200$ independent observed data structures for $k = 0$ (initial estimator) and k iterations of the updating procedure ($k = 1, 2, 3$), as obtained from $B' = 10^3$ independent replications of the simulation study. Gray rectangles represent 95%-accuracy intervals for the true parameter based on n observed data structures (light gray) and $B = 10^5$ observed data structures (dark gray).

7.5 An application to TCGA data

As an illustration, we study a breast cancer data set from The Cancer Genome Atlas (TCGA) Network [76]. We downloaded DNA methylation (W), DNA copy number (X), and expression (Y) of 11,314 genes for $n = 463$ patients[2]. The dimension of W is the number of CpG loci in the gene's promoter region, which can vary from one gene to another. Conveniently, our implementation handles multi-dimensional covariates.

In order to quantify the influence of DNA methylation on the strength of association between DNA copy number and gene expression in this data set, we compare our proposed non-parametric variable importance measure $\Psi(P)$ to its counterpart neglecting W. The latter is a different mapping from $\mathcal{M}$ to $\mathbb{R}$ given by

$$\mathcal{F}(P) = \arg\min_{\beta \in \mathbb{R}} \mathbb{E}_P \left\{ (Y - \beta X)^2 \right\} = \frac{\mathbb{E}_P\{XY\}}{\mathbb{E}_P\{X^2\}}.$$

A natural estimator of $\mathcal{F}(P)$ is given by $f_n = \sum_{i=1}^n X_i Y_i / \sum_{i=1}^n X_i^2$. Our theoretical results show that (f_n, ψ_n) satisfies a central limit theorem. Furthermore, it is possible to estimate the corresponding asymptotic covariance matrix, hence the asymptotic variances of ψ_n and of $(\psi_n - f_n)$.

We have performed the bilateral test of "$\Psi(P) = \mathcal{F}(P)$" against "$\Psi(P) \neq \mathcal{F}(P)$" for each of the 10,246 genes without missing data. Figure 7.3 presents the $(-\log_{10})$ p-values of this test against the gene's position. A pattern emerges of regions featuring very small p-values, among which chromosomes 1q, 8, 16q. The pattern is not correlated to the marginal

[2]The data are available from the TCGA at: https://tcga-data.nci.nih.gov/docs/publications/brca_2012.

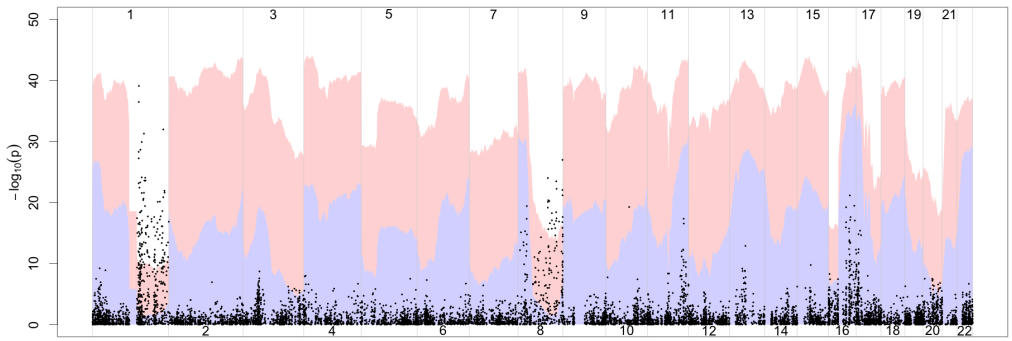


Figure 7.3: Each dot corresponds to the genomic position and $(-\log_{10})$ p-value of "$\Psi(P) = \mathcal{F}(P)$" against "$\Psi(P) \neq \mathcal{F}(P)$" for one of the 10,246 genes without missing data. The chromosomes are delimited by vertical grey lines. The background image represents, gene by gene, the proportions of the 463 samples for which $X < 0$ (blue), $X > 0$ (red), and $X = 0$ (white).

distribution of X, represented in the background. We also compute the partial correlation of X and Y given W for each gene (not shown). No pattern emerges. This suggests that the approach we propose may be useful to identify novel regions worthy of interest.

For this data set, the typical run time of the method for a single gene with the default options of the package is 10 seconds on a standard laptop. Obviously, this analysis can be parallelized very easily, as each gene is treated independently of all the other ones.

Chapter 8

Graph-structured two sample tests

In this chapter we consider multivariate two-sample tests of means, where the location shift between the two populations is expected to be related to a known graph structure. An important application of such tests is the detection of differentially expressed genes between two patient populations, as shifts in expression levels are expected to be coherent with the structure of graphs reflecting gene properties such as biological process, molecular function, regulation, or metabolism. For a fixed graph of interest, we demonstrate that accounting for graph structure can yield more powerful tests under the assumption of smooth distribution shift on the graph. We also investigate the identification of non-homogeneous subgraphs of a given large graph, which poses both computational and multiple hypothesis testing problems.

References:

[**J16**] L. Jacob, P. Neuvial, and S. Dudoit. "More Power via Graph-Structured Tests for Differential Expression of Gene Networks". *Annals of Applied Statistics* 6.2 (2012), pp. 561–600

[**S6**] L. Jacob and P. Neuvial. *DEGraph: Two-sample tests on a graph.* Bioconductor R package version 1.37.0. 2012

Contents

8.1 The two-sample Hotelling test for multivariate data

Most approaches to the joint analysis of gene expression data and gene graph data involve two distinct steps. Firstly, tests of differential expression are performed separately for each gene. Then, these univariate (gene-level) testing results are extended to the level of gene sets, *e.g.*, by assessing the over-representation of DE genes in each set based on p-values for Fisher's exact test[1] (or a χ^2 approximation thereof) adjusted for multiple testing [146] or based on permutation adjusted p-values for weighted Kolmogorov-Smirnov-like statistics [142].

There are two major caveats to this general two-step approach:

- The first step is based on *marginal* tests of association between one gene and class labels, so the information of the joint distribution of gene expression across genes is lost for the second step;

- The sampling units for the tests in the second step are the genes (instead of the observations/patients), which are expected to have strongly correlated expression measures. As noted by [117], this renders the interpretation of the tests at the second step problematic and may lead to a large loss of power or Type I error control when sets of genes have correlated expression.

The first point is illustrated by Figure 8.1 where the gene expression levels of two genes (PIAS2 and UBE2C) are displayed for the 79 samples of the Leukemia data set. Although neither of these genes is differentially expressed (the p-value of the t-test for differential expression of PIAS2 and UBE2C are 0.11 and 0.12, respectively , the pair (PIAS2, UBE2C) can be called differentially expressed using a bi-variate Hotelling test (as formally defined below) with p-value 0.023.

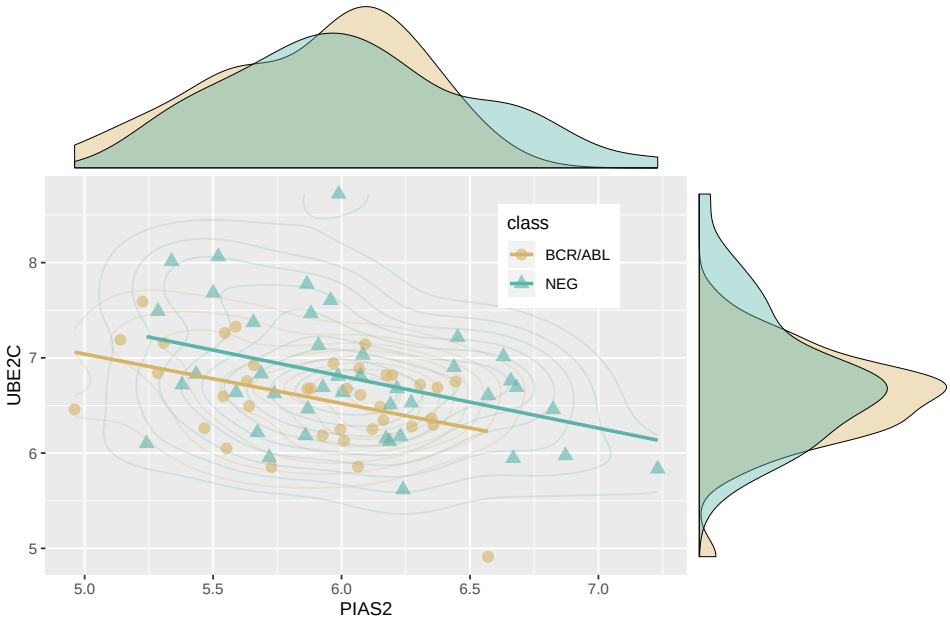


Figure 8.1: Expression measurements of the PIAS2 and UBE2C genes in two patient populations (BCR/ABL and NEG) from the Leukemia data set. The marginal distributions are similar across groups for its genes, but the scatter plot reveals differential expression in two dimensions.

[1] Sometimes referred to as hypergeometric test in the bioinformatics literature.

This example suggests that direct multivariate differential expression testing of gene sets could be more appropriate than posterior aggregation of marginal gene-level tests. We now give the definition of Hotelling's T^2-test, the most classical multivariate test of location shift. Let us consider random samples (x_1) and (x_2) of size n_1 and n_2 drawn from two p-dimensional Gaussian distributions, $\mathcal{N}(\mu_i, \Sigma)$, $i = 1, 2$. Assuming that $p < n_1 + n_2 - 1$ and Σ is invertible, Hotelling's T^2-test statistic is defined by

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{x}_1 - \bar{x}_2)^\top \hat{\Sigma}^{-1} (\bar{x}_1 - \bar{x}_2),$$

where $\bar{x}_i, i = 1, 2$ denote the sample means, and $\hat{\Sigma}$ the pooled sample covariance matrix. Up to the scaling factor $n_1 n_2/(n_1 + n_2)$, it corresponds to the squared Mahalanobis norm ($\Delta^2(x, S) = x^\top S^{-1} x$) of the sample mean shift $\bar{x}_1 - \bar{x}_2$ associated to the empirical covariance matrix $\hat{\Sigma}$). Under the null hypothesis $\mathbf{H}_0 : \mu_1 = \mu_2$ of equal means, NT^2 follows a (central) F-distribution $F_0(p, n_1 + n_2 - p - 1)$, where $N = \frac{n_1+n_2-p-1}{(n_1+n_2-2)p}$. In general, NT^2 follows a non-central F-distribution $F(\frac{n_1 n_2}{n_1+n_2} \Delta^2(\delta, \Sigma); p, n_1 + n_2 - p - 1)$, where the non-centrality parameter is, up to the same scaling factor as above, the Mahalanobis norm of the mean shift $\delta = \mu_2 - \mu_1$. We refer to $\Delta^2(\delta, \Sigma)$ as the distribution shift. In the remainder of this chapter, unless otherwise specified, T^2-statistics are assumed to follow the nominal F-distribution, *e.g.*, for critical value and power calculations.

Hotelling's test is known to be uniformly most powerful invariant for multivariate normal distributions against global-shift alternatives. However, it is not directly applicable to multivariate differential expression testing of gene sets, for two main reasons:

- first, the typical size of gene sets is of the same order as the typical sample size (dozens of genes/samples). Hotelling's test it is only defined when $p < n_1 + n_2 - 1$ and Σ is invertible. Even under these conditions, its power naturally decreases as p increases due to the increasingly poor conditioning of $\hat{\Sigma}^{-1}$ [171]. It is expected that genes from a particular gene set will have correlated expression levels. Continuing the example of Figure 8.1, the total expression level of the two genes (PIAS2 + UBE2C) has a much more significant association to the class labels (t-test p-value equal to 0.006) than the pair (PIAS2, UBE2C)

- second, such direct multivariate tests on unstructured gene sets do not take advantage of information on gene regulation or other relevant biological properties. An increasing number of regulation networks are becoming available, *e.g.*, Gene Ontology (GO; `http://www.geneontology.org`), Kyoto Encyclopedia of Genes and Genomes (KEGG; `http://www.genome.jp/kegg`) or NCI Pathway Integration Database (NCI graphs; `http://pid.nci.nih.gov`). Such networks specify, for example, which genes activate or inhibit the expression of other genes. If it is known that a particular gene in a tested gene set activates the expression of another, then one expects the two genes to have coherent (differential) expression patterns, *e.g.*, higher expression of the first gene in resistant patients should be accompanied by higher expression of the second gene in these patients.

In the next sections, we propose multivariate test statistics for identifying differential expression patterns (or, more generally, shifts in distribution) that are coherent with a given graph structure. In a nutshell, we propose to take advantage of the prior information encoded in gene networks or pathways to perform graph-structured dimension reduction, combined with a multivariate test in the low-dimensional space. The example of Figure 8.1 suggests that projecting the two-dimensional data (PIAS2, UBE2C) in the one-dimensional space (PIAS2 + UBE2C) may lead to increased differential expression power.

8.2 Graph-structured dimension reduction

Let us consider a network of p genes, represented by a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with $|\mathcal{V}| = p$ nodes and edge set $\mathcal{E}$. Let $\delta \in \mathbf{R}^p$ denote the mean shift, *i.e.*, the vector of differences in mean expression measures for these p genes between the two populations of interest. Suppose we expect the shift δ to be coherent with the graph $\mathcal{G}$, in the sense that it has low energy $E_{\mathcal{G}}(\delta)$ for a particular energy function $E_{\mathcal{G}}$ defined on $\mathcal{G}$. Then, we wish to build a space of lower dimension $k \ll p$ capturing most of the low energy functions. To this end, we start by finding the function that has the lowest possible energy, then the function that has lowest possible energy in the orthogonal space of the first one, up to the kth function with lowest energy in the orthogonal subspace of the first $k-1$ functions. That is, for each $i \leq k$, we define

$$u_i = \begin{cases} \underset{f \in \mathbf{R}^p}{\arg\min}\, E_{\mathcal{G}}(f) \\ \text{such that } u_i \perp u_j,\ j < i. \end{cases} \tag{8.1}$$

If $E_{\mathcal{G}}$ is a positive semi-definite quadratic form $E_{\mathcal{G}}(\delta) = \delta^\top Q_{\mathcal{G}} \delta$, for some positive semi-definite matrix $Q_{\mathcal{G}} = U \Lambda U^\top$, where U is an orthogonal matrix and Λ a diagonal matrix with elements λ_i, $i = 1, \ldots, p$, then the solution to Equation (8.1) is given by the k eigenvectors of $Q_{\mathcal{G}}$ corresponding to the smallest k eigenvalues. It is easy to check that these eigenvalues are the energies of the corresponding functions u_i, *i.e.*, $E_{\mathcal{G}}(u_i) = \lambda_i$.

Any positive semi-definite matrix can be chosen for $Q_{\mathcal{G}}$, with different choices of $Q_{\mathcal{G}}$ leading to different notions of coherence of the expression shift with the network. A classical choice is the *graph Laplacian* $\mathcal{L}$. Suppose $\mathcal{G}$ is an *undirected* graph with adjacency matrix A, with $a_{ij} = 1$ if and only if $(i,j) \in \mathcal{E}$ and $a_{ij} = 0$ otherwise, and degree matrix $D = \mathrm{Diag}\,(A\mathbf{1})$, where $\mathbf{1}$ is a unit column-vector, $\mathrm{Diag}(x)$ is the diagonal matrix with diagonal x for any vector x, and $D_{ii} = d_i$. The Laplacian matrix of $\mathcal{G}$ is then typically defined as $\mathcal{L} = D - A$ or $\mathcal{L}_{norm} = I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$ for the normalized version, leading to energies $\sum_{(ij)\in\mathcal{E}} (\delta_i - \delta_j)^2$ and $\sum_{(i,j)\in\mathcal{E}} \left(\delta_i/\sqrt{d_i} - \delta_j/\sqrt{d_j}\right)^2$, respectively. Note that, in this case, the Laplacian matrix $\mathcal{L}$, energy E, and basis functions u_i extend the classical Fourier analysis of functions on Euclidean spaces to functions on graphs, by transferring the notions of Laplace operator, Dirichlet energy, and Fourier basis, respectively [166].

In the specific case of gene regulation networks, it may be relevant to model negative associations between genes by considering a signed version of the adjacency matrix, where $a_{ij} = 1$ if gene i activates gene j, and -1 if it inhibits gene j. A signed version of the graph Laplacian is then $\mathcal{L}_{\text{sign}} = D - A$, where $D = \mathrm{Diag}\,(|A|\mathbf{1})$ is the degree matrix and $|A|$ denotes the entry-wise absolute value of A. Note that the signed Laplacian is always positive definite, see e.g. [32, Chapter 5]. As an example, let us consider a simple four-node graph whose signed adjacency matrix is given by

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & -1 \\ 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix}, \tag{8.2}$$

The eigenvectors of the corresponding signed Laplacian are represented in Figure 8.2. Following our principle to build a lower dimension space, we use the first few eigenvectors of $Q_{\mathcal{G}}$ to obtain orthonormal functions with low energy. The first eigenvector, corresponding to the smallest energy (eigenvalue of zero), can be viewed as a "constant" function on the graph: its absolute value is identical for all nodes, but nodes connected by an edge with negative weight take on values of opposite sign. By contrast, the last eigenvector, corresponding to the highest energy, is such that nodes connected by positive edges take on values of opposite sign and nodes connected by negative edges take on values of the same sign.

Figure 8.3 illustrates the projection of two vectors onto the first two dimensions of the graph decomposition. The first vector is smooth along the graph, in the sense that its

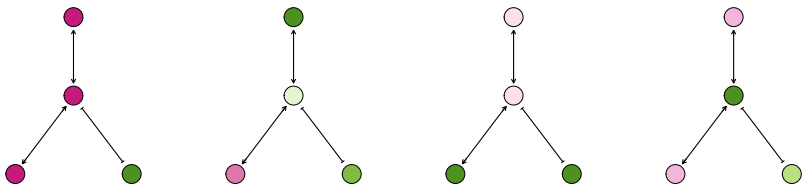


Figure 8.2: Eigenvectors of the graph with adjacency matrix defined in (8.2).

entries are more coherent with the graph structure. This is reflected by the magnitude of the coefficients of Therefore, it is essentially preserved by the projection in the first two dimensions. Conversely, the second vector is not smooth along the graph, so it is not preserved by the projection.

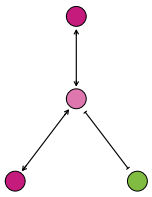
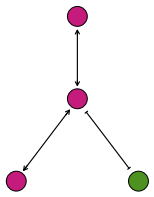
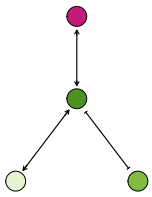
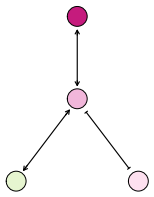
Smooth shift		Non-smooth shift	
Original vector	$2d$ projection	Original vector	$2d$ projection
			
(1.45, -0.04, -0.21, 0.20)	(1.45, -0.04, 0, 0)	(0.2, -0.41, +0.42, 1.16)	(0.2, -0.41, 0, 0)

Figure 8.3: Example of projection of two vectors along first two dimensions of the graph with adjacency matrix defined in (8.2). Left: a smooth vector along the graph; right: a non-smooth vector along the graph. Bottom: coefficients of the eigen decomposition of each vector.

While we have introduced the idea in the context of gene regulation networks and testing for differential expression, the same dimensionality reduction principle applies to any multivariate testing problem for which the variables have a known structure, as represented by a graph.

8.3 Graph-structured two-sample tests

In the remainder of this chapter, we denote by $\tilde{f} = U^\top f$ the coefficients of a vector $f \in \mathbf{R}^{|\mathcal{V}|}$ after projection on a basis U (typically the eigenvectors of a $Q_\mathcal{G}$ matrix).

For any orthonormal basis U and, in particular, for our graph-based basis, direct calculation shows that

$$T^2 = \tilde{T}^2 \triangleq \frac{n_1 n_2}{n_1 + n_2} (\bar{x}_1 - \bar{x}_2)^\top U \left(U^\top \hat{\Sigma} U \right)^{-1} U^\top (\bar{x}_1 - \bar{x}_2),$$

i.e., the statistic T^2 in the original space and the statistic $\tilde{T}^2$ in the new graph-based space are identical. More generally, for $k \leq p$, the statistic in the original space after filtering out dimensions above k is the same as the statistic $\tilde{T}^2_k$ restricted to the first k coefficients in the new space defined by U:

$$\begin{aligned}
\tilde{T}^2_k &\triangleq \frac{n_1 n_2}{n_1 + n_2} (\bar{x}_1 - \bar{x}_2)^\top U_{[k]} \left(U_{[k]}^\top \hat{\Sigma} U_{[k]} \right)^{-1} U_{[k]}^\top (\bar{x}_1 - \bar{x}_2) \\
&= \frac{n_1 n_2}{n_1 + n_2} (\bar{x}_1 - \bar{x}_2)^\top U 1_k U^\top \left(U 1_k U^\top \hat{\Sigma} U 1_k U^\top \right)^{+} U 1_k U^\top (\bar{x}_1 - \bar{x}_2),
\end{aligned}$$

where A^+ denotes the generalized inverse of a matrix A, the $p \times k$ matrix $U_{[k]}$ denotes the restriction of U to its first k columns, and 1_k is a $p \times p$ diagonal matrix, with ith diagonal element equal to one if $i \leq k$ and zero otherwise. Note that, as retaining the first k dimensions corresponds to a *non-invertible* transformation, this filtering indeed has an effect on the test statistic, that is, we have $\tilde{T}_k^2 \neq \tilde{T}^2$ in general. Lemma 1 stated below shows that gains in power can be achieved by filtering, under the assumption of a smooth shift along the graph. We let $\tilde{\delta} = U^\top \delta$ and $\tilde{\Sigma} = U^\top \Sigma U$ denote, respectively, the mean shift and covariance matrix in the new space. Given $k \leq p$, let $\Delta_k^2(\delta, \Sigma) = \delta_{[k]}^\top \left(\Sigma_{[k]}\right)^{-1} \delta_{[k]}$ denote the distribution shift restricted to the first k dimensions of δ and Σ, *i.e.*, based on only the first k elements of δ, $(\delta_i : i \leq k)$, and the first $k \times k$ diagonal block of Σ, $(\sigma_{ij} : i, j \leq k)$.

Lemma 8.1. *For any level α and any $1 < l \leq p - k$, there exists $\eta(\alpha, k, l) > 0$ such that*

$$\Delta_{k+l}^2(\tilde{\delta}, \tilde{\Sigma}) - \Delta_k^2(\tilde{\delta}, \tilde{\Sigma}) < \eta(\alpha, k, l) \Rightarrow \beta_{\alpha,k}(\Delta_k^2(\tilde{\delta}, \tilde{\Sigma})) > \beta_{\alpha,k+l}(\Delta_{k+l}^2(\tilde{\delta}, \tilde{\Sigma})),$$

where $\beta_{\alpha,k}(\Delta^2)$ is the power of Hotelling's T^2-test at level α in dimension k for a distribution shift Δ^2, according to the nominal F-distribution $F(\frac{n_1 n_2}{n_1+n_2}\Delta^2; k, n_1 + n_2 - k - 1)$.

Under the assumption that the distribution shift is smooth, *i.e.*, lies mostly in the first few graph-based coefficients, so that $\Delta_k^2(\tilde{\delta}, \tilde{\Sigma})$ is nearly maximal for a small value of k, Lemma 1 states that performing Hotelling's test in the new space restricted to its first k components yields more power than testing in the entire new space. Equivalently, the test is more powerful in the original space after filtering than in the original unfiltered space. The increase in shift $\eta(\alpha, k, l)$ required to maintain power when increasing dimension can be evaluated numerically for any (α, k, l). Note that this result holds because retaining the first k new components is a *non-invertible* transformation.

Corollary 8.2 states that if the distribution shift lies in the first k new coefficients, then testing in this subspace yields strictly more power than using additional coefficients:

Corollary 8.2. *If $\forall\ 1 < l \leq p - k,\ \Delta_k^2(\tilde{\delta}, \tilde{\Sigma}) = \Delta_{k+l}^2(\tilde{\delta}, \tilde{\Sigma})$, then*

$$\beta_{\alpha,k}(\Delta_k^2(\tilde{\delta}, \tilde{\Sigma})) > \beta_{\alpha,k+l}(\Delta_{k+l}^2(\tilde{\delta}, \tilde{\Sigma})).$$

In particular, if there exists $k < p$ such that $\tilde{\delta}_j = 0\ \forall\ j > k$ (*i.e.*, the mean shift is smooth) and $\tilde{\Sigma}$ is block-diagonal such that $\tilde{\sigma}_{ij} = 0\ \forall\ i < k, j > k$, then gains in power are obtained by testing in the first k new components. Although non-necessary, this condition is plausible when the mean shift lies at the beginning of the spectrum (i.e., has low energy), as the coefficients which do not contain the shift are not expected to be correlated with the ones that do contain it. Figure 8.4 illustrates, under different settings, the increase in distribution shift necessary to maintain a given power level against the number of added coefficients. Under the assumption of block-diagonal covariance, it is also possible to directly relate the energy of the mean shift vector to the gain in power, see Corollary 2 in [J16, Supplement A].

8.4 Numerical experiments

The empirical performance of the graph-structured test has been assessed in numerical experiments where the distribution shift Δ^2 satisfies the above-defined smoothness assumptions. More specifically, we assume that this shift lies in the first k_0-dimensional eigen space of the graph. As expected in this favorable setting, Figure 8.5 shows (in the case of a block diagonal covariance structure and for $k_0 = 3$) that our proposed T^2-statistic in the first k_0 graph-based coefficients (dashed red lines) compares favorably to several alternatives: the standard Hotelling T^2-statistic in the original space and T^2-statistic in the first k_0 principal components (left panel), the statistics of [171] (BS), [92] (CQ), and [109] (SD) (middle panel), and the Adaptive Neyman statistics of [167] (right panel).

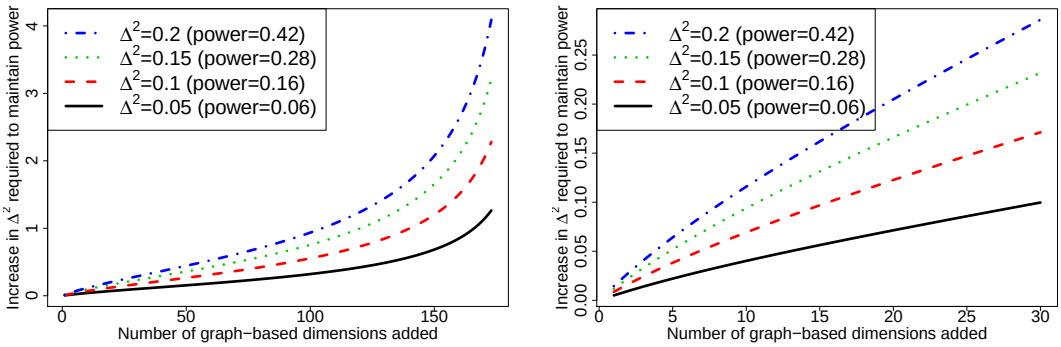


Figure 8.4: Left: Increase in distribution shift required for Hotelling's T^2-test to maintain a given power when increasing the number of tested new coefficients: $\Delta^2_{k+l} - \Delta^2_k$ vs. l such that $\beta_{\alpha,k+l}(\Delta^2_{k+l}) = \beta_{\alpha,k}(\Delta^2_k)$. Power $\beta_{\alpha,k+l}(\Delta^2_{k+l})$ computed under the non-central F-distribution $F\left(\frac{n_1 n_2}{n_1+n_2}\Delta^2_{k+l}; k+l, n_1+n_2-(k+l)-1\right)$, for $n_1 = n_2 = 100$ observations, $k = 5$, and $\alpha = 10^{-2}$. Each line corresponds to the fixed shift Δ^2_k and power $\beta_{\alpha,k}(\Delta^2_k)$ pair indicated in the legend. Right: Zoom on the first 30 dimensions.

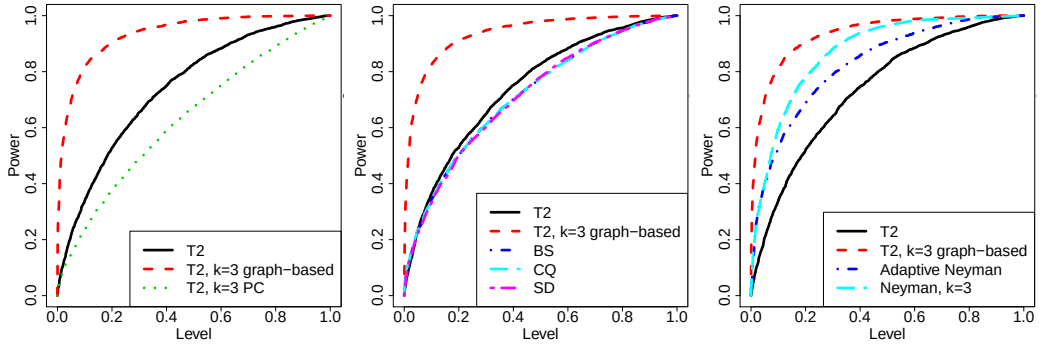


Figure 8.5: Synthetic data: Receiver Operating Characteristic (ROC) curves for the detection of a smooth shift under block-diagonal covariance structure.

Further numerical experiments reported in [J16] quantify empirically the robustness of our proposed method to a possible mis-specification of k_0 (as k_0 is typically unknown in practice), and to a possible mis-specification of the graph, where the graph used to perform dimension reduction and testing is a noisy version of the graph used for data generation.

8.5 Differential subgraph discovery

The methods described in this chapter are designed to test the differential expression of known graphs. An important related question is whether we can identify non-homogeneous subgraphs of a known graph. This poses a huge combinatorial problem even for moderately large graphs, as the number of (connected) subgraphs of size k of a graph of size p can be exponential in p and k. Exhaustive search is therefore not feasible in practice, especially for differential expression on gene networks, where p is typically in the dozens or hundreds of genes. To address this bottleneck, we describe in [J16] a branch and bound type algorithm which makes it possible to avoid testing all possible subgraphs. This algorithm relies on a pruning strategy based on an upper bound on the value of the test statistic for any subgraph containing a given set of nodes [J16, Supplement A, Lemma 2]. Two variants of this

algorithm are proposed: an exact one, and a quicker, approximate one. The approximation in the latter consists in focusing on the subgraphs whose sample mean shift in the first k components in the graph space has a sufficiently large Euclidean norm. Just like for volcano plots, this approximation is justified in practice by the fact that significant subgraphs corresponding to small mean shifts are typically less relevant from a biological perspective. Finally, we have devised a strategy based on class label permutations to estimate the expected number of type I errors and thereby account for the multiple testing situation created by testing a large number of subgraphs.

Chapter 9

Statistical inference from DNA copy number data

After a brief description of DNA copy number signals, this chapter reviews some statistical insights gained from working with DNA copy number data. We highlight the importance of data pre-processing, computational and statistical trade-offs for the design of segmentation methods, and the importance of performance evaluation and benchmarking studies in this context.

References:

[**J11**] M. Pierre-Jean, G. J. Rigaill, and P. Neuvial. "Performance evaluation of DNA copy number segmentation methods". *Briefings in Bioinformatics* 4 (2015), pp. 600–615

[**J18**] M. Ortiz-Estevez, A. Aramburu, H. Bengtsson, P. Neuvial, and A. Rubio. "CalMaTe: A Method and Software to Improve Allele-Specific Copy Number of SNP Arrays for Downstream Segmentation". *Bioinformatics* 28.13 (July 2012), pp. 1793–1794

[**J19**] A. B. Olshen, H. Bengtsson, P. Neuvial, P. T. Spellman, R. A. Olshen, and V. E. Seshan. "Parent-specific copy number in paired tumor-normal studies using circular binary segmentation". *Bioinformatics* 27.15 (Aug. 2011), pp. 2038–2046

[**J21**] H. Bengtsson, P. Neuvial, and T. P. Speed. "TumorBoost: Normalization of allele-specific tumor copy numbers from a single pair of tumor-normal genotyping microarrays". *BMC Bioinformatics* 11.1 (2010), p. 245

[**B1**] P. Neuvial, H. Bengtsson, and T. P. Speed. "Statistical analysis of Single Nucleotide Polymorphism microarrays in cancer studies". *Handbook of Statistical Bioinformatics*. Ed. by H. H.-S. Lu, B. Schölkopf, and H. Zhao. Springer Handbooks of Computational Statistics. Springer, 2011

[**S2**] M. Pierre-Jean, G. Rigaill, and P. Neuvial. *jointseg: Joint segmentation of multivariate (copy number) signals*. R package version 1.0.2. 2019

[**S4**] M. Pierre-Jean and P. Neuvial. *acnr: Annotated Copy-Number Regions*. R package version 1.0.0. 2017

Contents

9.1 Copy-number signals

As noted in Section 1.3, an important issue in cancer research is to estimate the underlying *copy number state* at each position along the genome of a tumor sample. Formally, we define the copy number state of a tumor at a given genomic locus j as a pair of non-negative numbers $(\underline{\gamma}_j, \overline{\gamma}_j)$, where $\underline{\gamma}_j \leq \overline{\gamma}_j$, which are respectively the smaller and the larger of the two parental copy numbers at this locus. Figure 9.1 illustrates the most common copy number states observed in cancer samples: normal, that is one copy from each parent (1,1), gain (1,2), deletion (0,1), copy-neutral LOH[1] (0,2). By definition we have $\underline{\gamma}_j \leq \overline{\gamma}_j$, and

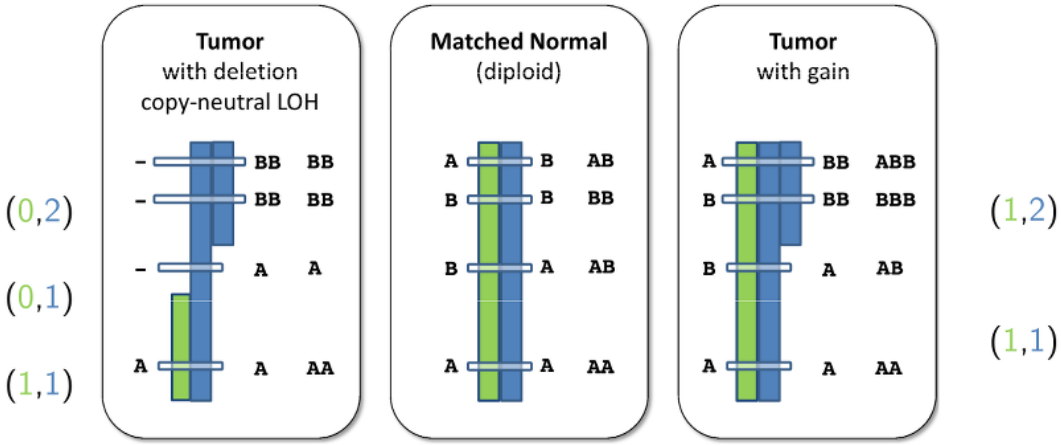


Figure 9.1: Illustration of the main DNA copy-number states in three cancer samples: normal, gain, deletion, copy-neutral LOH. Illustration by Henrik Bengtsson.

$\gamma_j = \underline{\gamma}_j + \overline{\gamma}_j$ is the total copy number. The quantities $\underline{\gamma}_j$ and $\overline{\gamma}_j$ are called minor and major copy numbers, respectively. Note that $\underline{\gamma}_j$, $\overline{\gamma}_j$, and γ_j need not be whole numbers, especially because of the possible presence of normal cells in the tumor sample.

Minor and major copy numbers are not directly observed from genotyping microarray or sequencing assays, but they can be estimated from the data collected by these assays. These data can be summarized by a pair of vectors $(\theta_j^A, \theta_j^B)_j$ where the index j refers to a genomic position along a chromosome. In Figure 9.1 four such positions are represented by horizontal segments. These positions correspond to (bi-allelic) Single Nucleotide Polymorphisms (SNPs), that is, genomic positions where the DNA sequence varies at a substantial rate across individuals of some population. For most SNPs only two letters (out of {A, C, G, T}) variants are observed. These versions of the genetic sequence are called *alleles* and arbitrarily denoted by A and B in Figure 9.1. The values θ_j^A and θ_j^B correspond to the signal intensity for allele A and B at SNP j.

9.2 Preprocessing: allelic ratio normalization

These "raw" signals have to be preprocessed (or "normalized") in order to make them comparable (i) across samples for each locus, and (ii) across loci for each sample. While (i) is a common issue in all genomic analyses, we focus here on (ii), which is particularly crucial for copy-number data, where the locus of a given sample have to be segmented into genomic regions.

[1]Loss of heterozygosity (LOH) is the loss of the contribution of one parent in a genomic region. It includes the case of a deletion, but also of copy-neutral LOH, corresponding to two copies from the same parent.

The TCGA project of molecular characterization of cancers has performed thousands of DNA copy number experiments for which both tumor sample and a normal sample (e.g. blood cells) from the same individual are available. For a given individual, we are therefore observing at each locus j $(\theta_{ij}^A, \theta_{ij}^B)$, where $i \in \{N, T\}$ for the normal and the tumor sample, respectively. Let us define $\theta_{ij} = \theta_{ij}^A + \theta_{ij}^B$, and $\beta_{ij} = \theta_{ij}^B / \theta_{ij}$. The copy number data for one such tumor/normal pair are generally summarized using

- total copy numbers in the tumor sample: $c_j = 2\theta_{Tj}/\theta_{Nj}$;
- allelic ratios (a.k.a. allele B fraction) in the normal sample: $b_{Nj} = \theta_{Nj}^B/\theta_{Nj}$;
- allelic ratios in the tumor sample: $b_{Tj} = \theta_{Tj}^B/\theta_{Tj}$.

Note that in the above definition of c_j, the total intensity θ_{Tj} at locus j in the tumor sample is scaled by the corresponding quantity in the matching normal sample. The goal of this normalization is to correct for systematic and locus-specific biases in these intensities. These signals are displayed in the first three rows of Figure 9.2 for Chromosome 2 (left) and Chromosome 10 (right) of one TCGA ovarian cancer sample. The last row of Figure 9.2 displays

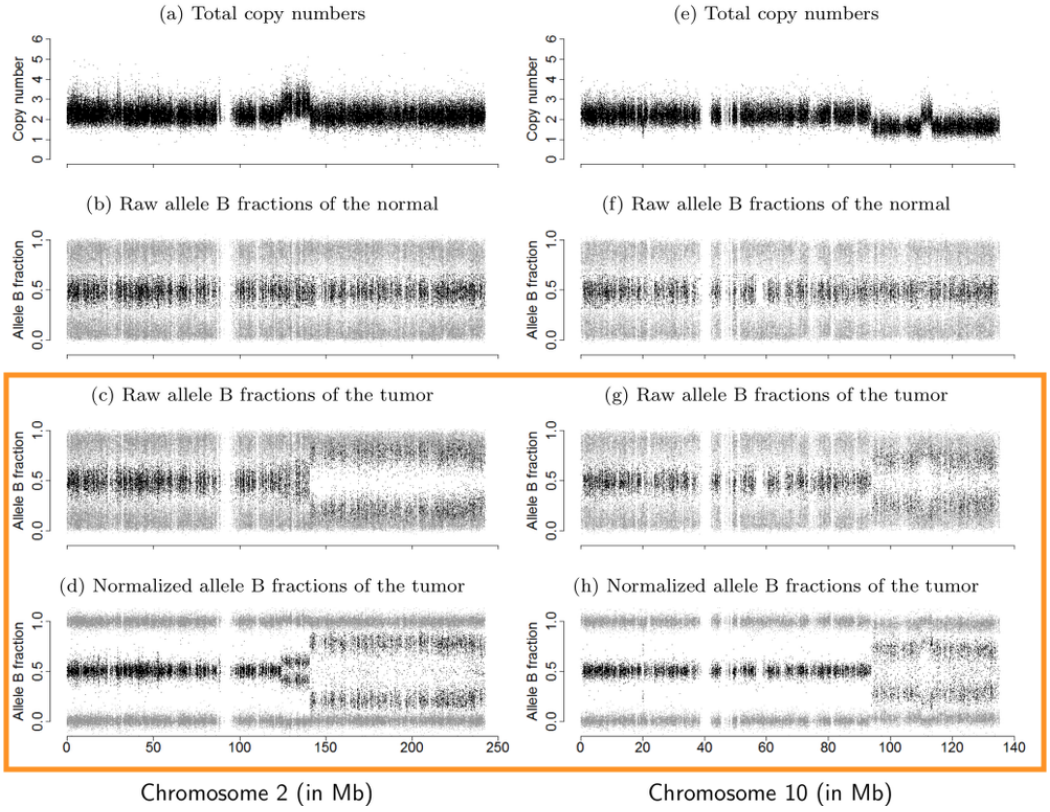


Figure 9.2: Copy-number signals in two genomic regions (left and right columns). Orange box: allelic signals after Tumorboost normalization (last row) have a much higher signal to noise ratio than before Tumorboost normalization (third row).

the tumor allelic ratios after processing by the TumorBoost normalization method that we developed [J21, s3] in the context of the Aroma Project. As illustrated by the comparison between the (c) and (d) panels and between the (g) and (h) panels, this method provides a substantial denoising of the input data, with changes in the allelic ratio distribution being much more visible after normalization. The method takes advantage of two observations. First, the variability of allelic ratios at a given position is systematic, i.e. highly reproducible

between the tumor and the paired normal sample. Second, one expects the true allelic ratios to be either 0, 1/2 or 1 in the normal samples. Therefore, the difference between the normal allelic ratio and the expected true one at a given locus provides a good estimate of the noise in the tumor allelic ratio, which can then be subtracted to estimate the tumor allelic ratio.

We have extended this method to the case where no paired normal is available in a collaboration with researchers at the University of Navarra [J18, s2]. We also stress the importance of the choice of a reference to estimated total copy numbers in this unpaired case, as illustrated in [B1]. While the methods described in this section are very simple from a mathematical point of view, their impact in the quality of the data is of major importance for downstream statistical analyses and results interpretation.

9.3 Detecting change points from copy-number signals

9.3.1 Copy numbers as a two-dimensional piecewise constant signal

The left panels of Figure 9.3 illustrate the fact that total DNA copy numbers (c) are piecewise constant while allelic ratios (b) can display several modes in a given region. The panel (d) at the bottom right displays the decrease in heterozygosity (also called mirrored B allele fraction [110]). It is a symmetrized version of b: $d = |b - 1/2|$ where only heterozygous SNPs (black dots) are displayed, because they carry all the information regarding copy number changes. Both c and d can be modeled as piecewise constant, and classical segmentation methods to detect a change in the mean of a signal can be used to perform segmentation on these signals.

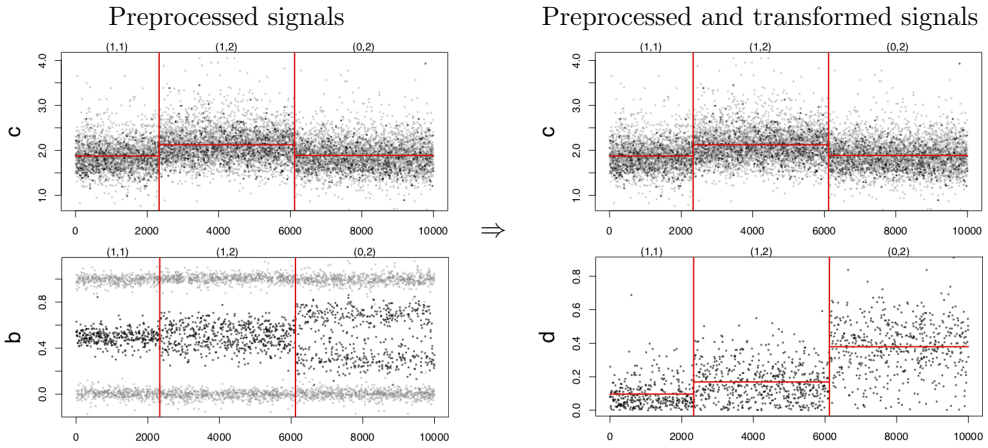


Figure 9.3: Copy-number profile of a tumor sample. Top: total copy numbers (c): the same panel is displayed twice. Bottom: allelic ratios (b) on the left, and decrease in heterozygosity (d) on the right. Red vertical bars indicate the presence of change points. Gray and black dots correspond to SNP that were called homozygous and heterozygous, respectively, in a paired normal sample.

Importantly, Figure 9.3 also illustrates the fact that change points occur at the same position in both signal dimensions. This is the case because changes in one of the parental copy numbers induce changes in both c and b (or d). In order to maximize the power of a change point detection method, it makes sense to require that a method can be applied jointly (rather than independently) to the two dimensions of the signals.

The problem of inferring the location of DNA copy number changes (also called copy-number change points) from locus-level estimates is an instance of the widely-studied change-point detection problem. Two main statistical models have been used to model DNA copy

number changes: change-point models and Hidden Markov Models (HMM). Before focusing on change point models, we give a brief overview of HMM. HMM assume that observed copy numbers at the locus level are emitted by an underlying Markov chain according to a small number of hidden true copy number states. HMM naturally incorporate and take advantage of the fact that different segments can have the same true copy number. Several HMM-based methods have been proposed for segmenting total copy numbers [148, 105, 106]. These methods mainly differ in the assumptions that are made for the dynamics of the underlying Markov chain, and the approaches used for the estimation of the hidden states. Extensions of HMM to two-dimensional signals have also been proposed [79].

We refer to [23] for a more comprehensive review on the general problem of change-point detection, which also covers the known statistical guarantees for the different methods. In the rest of this section we review change point models in the light of their application to copy number data. This section is adapted and updated from [B1]. In particular, we are interested in two key aspects:

- trade-offs between statistical accuracy and computational efficiency: sub-quadratic complexity in time or space is necessary for an algorithm to be applicable to copy number data;
- applicability to the joint segmentation of two-dimensional piecewise constant signals (c, d) which was not addressed in [B1].

9.3.2 A change-point model

A simple model for the observed DNA copy numbers $\mathbf{c}$ is:

$$c_j = \gamma_j + \varepsilon_j, \quad 1 \leq j \leq p \tag{9.1}$$

where the errors $(\varepsilon_j)_{1 \leq j \leq p}$ are iid and the vector γ of true copy numbers is are assumed to be piecewise constant. For simplicity of exposition, (9.1) is formulated in terms of total copy numbers, but extending this model to two dimensions ((c, d) or $\underline{c}, \overline{c}$) is straightforward. The main difficulty with the two-dimensional version is of computational nature, as explained below.

We consider here the problem of minimizing the ℓ^2 loss, both for simplicity and for its relevance to DNA copy number segmentation. More general loss functions can be considered, see e.g. [53] which addresses the problem of minimizing the cost of a segmentation under the assumption that the cost is segment-additive. Under the ℓ^2 loss, our goal is to solve the following optimization problem:

$$\min_{\boldsymbol{\gamma}} \sum_{j=1}^{p} (c_j - \gamma_j)^2 \quad \text{s.t.} \quad \|D\boldsymbol{\gamma}\|_0 \leq K \tag{9.2}$$

where $D\boldsymbol{\gamma} = (\gamma_{j+1} - \gamma_j)_{1 \leq j \leq p}$ is the vector of first order differences associated to the true copy numbers $\boldsymbol{\gamma}$, and $\|\cdot\|_0$ is the ℓ_0 norm, that is, the number of non-null entries. The constraint $\|D\boldsymbol{\gamma}\|_0 \leq K$ encodes the fact that γ is piecewise constant, with at most K change points. Note that if we assume that the errors are Gaussian and homoscedastic, i.e. $\varepsilon_j \sim \mathcal{N}(0, \sigma^2)$, then maximizing the likelihood of the model (9.1) for a given number of change points K is equivalent to solving the optimization problem (9.2). Also, under this assumption, given change point locations, the optimal value of the true copy number between two change points is the average observed copy number in this region. Therefore, for a given K, solving (9.2) reduces to the combinatorial problem of finding the best possible change point locations according to (9.2). In practice, we are thus faced with two problems:

Combinatorial problem: (9.2) is non-convex, and the number of admissible change point locations is $\binom{K}{p-1}$, that is, $O(p^K)$, so an exhaustive search of the best partition is practically infeasible in the case where p is large. This problem is discussed in Section 9.3.3;

Model selection problem: K is usually not known. As the segmentation models are nested with respect to K, the objective function in (9.2) is non-increasing in K. This model fit term has to be compensated by a penalty term that is increasing with the model size K:

$$\min_{\boldsymbol{\gamma}} \sum_{j=1}^{p} (c_j - \gamma_j)^2 + \mathrm{pen}(K)\,. \tag{9.3}$$

A variety of penalties have been proposed for the change point problem, see e.g. [23] for a general review and [68] for a focus on model section for DNA copy number change point detection. Formally, when the penalty is linear in K, 9.3 can be seen as the (Lagrangian) dual problem of (9.2). This is often the case, e.g. for the classical Akaike Information Criterion (AIC, [190]) or Bayesian Information Criterion (BIC, [184]).

9.3.3 Estimation of change point locations

The main difficulty to address the computational problem of estimating the change point locations is to find an appropriate balance between computational complexity, and statistical accuracy. We are interested in applications where K is of the order of 10 to 100, and p of the order of 10^5 to 10^6. In such situations, a quadratic time or space complexity is already prohibitive. We review three types of approaches that we find both relevant to the biological problem, and statistically sound, and then discuss how these approaches may be combined.

Exact solutions. By taking advantage of the additivity of the objective function in the segments, it is possible to use a dynamic programming strategy to reduce the complexity of an exhaustive search from $O(p^K)$ to $O(K{\cdot}p^2)$ [139], at the price of increasing the space complexity from $O(1)$ to $O(p^2)$. This algorithm recovers the entire sequence of optimal segmentations in k segments for all $k \leq K$, and it can be extended two multi-dimensional signals. However, a quadratic time and space complexity may be too high for recent microarray or sequencing copy number data. A pruned dynamic programming algorithm with linear space complexity has been proposed, that recovers the set of solutions faster [53]. In particular, although the worst case time complexity of the pDPA algorithm is still $O(K{\cdot}p^2)$, in practical situations it is almost linear in p. A related algorithm called PELT [73] provides the solution to the dual problem (9.3) with linear penalty $\mathrm{pen}(K) = \lambda K$ in a linear time for a fixed value of λ. Although it does not provide the entire path of solutions (like pDPA does), it can be extended to multi-dimensional signals (unlike pDPA).

Binary segmentation. A widely used approach in various types of applications is binary segmentation [188], which recursively looks for the best partition of the data into two segments. This greedy algorithm has a very low (linear) time complexity of $O(p \log K)$. It has been adapted to the problem of copy-number segmentation by looking for the best partition in three segments at each step, with the constraint that the two extreme segments have the same true copy number [150]. This method is called Circular Binary Segmentation (CBS). The depth of the recursion is determined by the estimated significance of the change points, which implicitly determines K. The original algorithm is quadratic in time due to the fast that two segment boundaries are looked for at each step. However, a pruned version which is also almost linear in time has been proposed [121]. We have proposed an extension of this method to the problem of segmenting both c and d [J19, s1]. It consists in a conditional segmentation method, where (i) each chromosome is segmented by CBS using the c statistic, and (ii) each segment is segmented by CBS using the d statistic. This extension inherits the quasi-linear complexity or CBS.

Convex relaxations. A classical approach in statistical machine learning is to replace a non-convex optimization problem with an approximate, but convex version of the problem,

which can therefore be solved efficiently. Problem (9.2) can be convexified by replacing the ℓ_0 constraint on the number of change points by a ℓ_1 constraint. An adaptation of the fused lasso [144] has been proposed [112], which solves the constrained optimization problem

$$\min_{\boldsymbol{\gamma}} \sum_{j=1}^{p} (c_j - \gamma_j)^2 \quad \text{s.t.} \quad \|D\boldsymbol{\gamma}\|_1 \leq v \text{ and } \|\boldsymbol{\gamma} - 2\|_1 \leq u\,. \tag{9.4}$$

The optimization problem (9.4) incorporates both a total variation constraint, that is, a constraint on the ℓ_1 norm of the jumps in $\boldsymbol{\gamma}$, and sparsity constraint on $\boldsymbol{\gamma} - 2$, enforcing that most loci correspond to the normal copy number state. The complexity of the algorithm proposed in [112] is (at best) $O(p^2)$, which as already discussed can be prohibitive for recent data sets. The second method [94] only keeps the total variation constraint, i.e. it simply convexifies $\|D\boldsymbol{\gamma}\|_0$ in (9.2) into $\|D\boldsymbol{\gamma}\|_1$. This optimization problem can be written as a Lasso-type regression problem via the change of variable $\boldsymbol{\delta} = D\boldsymbol{\gamma}$. Therefore be solved in $O(p{\cdot}K^2)$ using a Least Angle Regression (LARS) algorithm [147] to select the first K change points. An extension of this method to multi-dimensional signals into group-fused LARS has been proposed by [95, 77], which inherits the linear complexity of the original method.

Combining approximate and exact methods. We also mention an interesting two-step strategy that consists in

1. performing a fast (yet approximate) search of K candidate change points, e.g. using binary segmentation or total variation penalty;

2. running dynamic programming only on this reduced set of candidates.

This two-step strategy can improve the exploration of all possible segmentations, while maintaining a low computational complexity. Indeed, when $K \ll p$, the computational cost of the second step ($O(K^3)$) can be negligible compared to the initial search. Such strategies have been proposed by [104] for binary segmentation, and by [94] for the total variation penalty. Importantly, this two-step approach can be used for the joint segmentation of multi-dimensional signals. These methods are implemented in the **R** `jointseg` [S2] which is available from CRAN[2]: in particular, the group-fused LARS method of [95, 77] has been ported from Matlab to **R**, the CART/binary segmentation method of [104] has been implemented. The fast univariate implementation of pDPA [53] is also available in `jointseg`.

9.4 Performance evaluation

It is often the case that in absence of rich enough gold standard data sets, the numerical performance of statistical methods is assessed using unrealistic simulation studies, or based on small real data analyses with limited ground truth available. We have designed and implemented a framework to generate realistic DNA copy number profiles of cancer samples with known truth [J11, S4], that aims at combining the advantages and bypass some of the limitations of previous approaches [145, 75, 135, 67]. In this section we describe this framework, and demonstrate how it can be applied to compare copy number segmentation methods.

9.4.1 Genomic regions with known copy-number state

Figure 9.4 displays microarray copy-number profiles (c, b) for two chromosomes (in columns) in two samples (in rows) from the tumor cell line H1395, whose karyotype is displayed in Figure 1.1. These samples are part of a dilution series [85] available from the Gene

[2] `http://cran.r-project.org/package=jointseg`

Expression Omnibus (GEO, [157]). In this series, the cell line H1395 is mixed with a matched blood sample with several mixture proportions: 0, 30, 50 70 and 100% tumor cells. Only the last two mixture proportions are shown in Figure 9.4. The regions highlighted in purple were manually labelled as corresponding to two different copy number states (gain and loss of one DNA copy). Importantly, the labelling was performed on the 100% tumor sample where the signal to noise ratio is higher, but it is valid as well for the 70% tumor sample (and for the lower ones), where the signal to noise ratio is weaker. The regions highlighted in purple are then extracted and serve as a basis for a resampling-based data generation framework. These annotated data sets are available in the R package `acnr` [S4] which is available from CRAN[3].

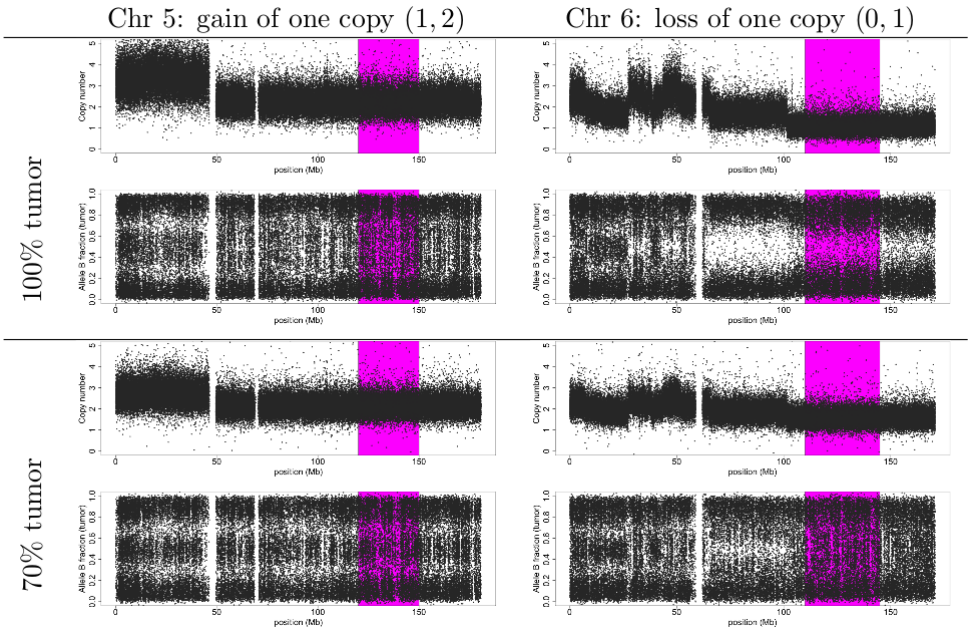


Figure 9.4: (c, b) signals for two annotated regions from tumor cell line NCI-H1395-4W. Left: gain on chromosome 5; right: loss on chromosome 6. Top: sample with pure tumor; Bottom: sample with a mixture of 70% tumor cells and 30% normal cells.

9.4.2 Generating copy number profiles with known truth

A synthetic copy-number profile of length n with K change points can be generated in two steps from the annotated data described above:

generation of truth: K change point positions (drawn uniformly out of the $n-1$ possible intervals between two successive loci), and $K+1$ copy-number state labels for all $K+1$ regions between two consecutive change points, drawn from those of the existing annotated regions;

generation of signal: a $n \times 3$ matrix of total copy numbers (c), allelic ratios (b), and genotypes. For each region of size n_R between two change points, we sample n_R data points from real copy-number data (such at the one displayed in Figure 9.1) corresponding to this type of region.

Using this framework, a variety of synthetic copy-number profiles can be generated. Figure 9.5 displays examples of four such profiles, which were generated from the same

[3] `http://cran.r-project.org/package=acnr`

"truth" (change point position and region labels) as in Figure 9.3 with different tumor cell fractions (in rows), and from different annotated data sets (in column; the first column uses the data set described above, and displayed in Figure 9.4). Two attractive features of this

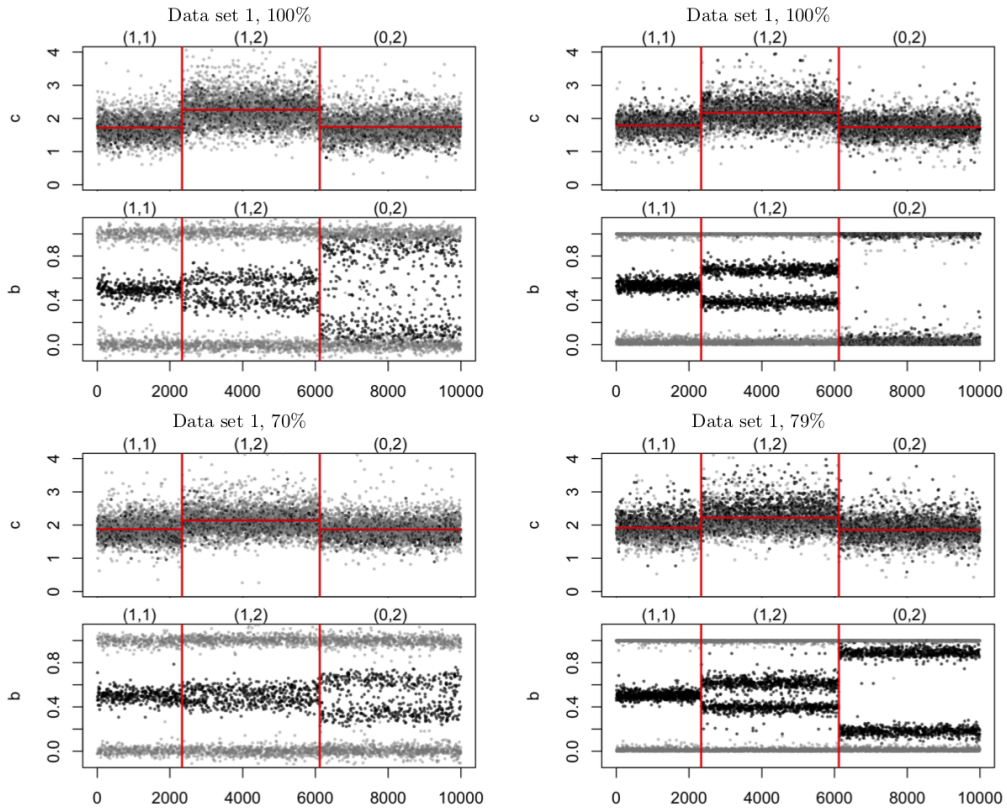


Figure 9.5: Synthetic copy-number profiles that generated from the same "truth" as in Figure 9.3. Each block of two plots corresponds to total copy numbers (c) and allelic ratios (b) for one particular combination of fraction of tumor cells (in rows) and data set (in columns). Red vertical lines represent change points. Heterozygous SNPs are colored in black, and homozygous SNPs in gray.

framework are the following:

- the distribution of copy number data in a given region matches a distribution actually observed in real data; therefore it does not rely on any probabilistic model assumption;
- the signal to noise ratio of these profiles can be tuned via the fraction of tumor cells, which has a clear biological interpretation.

9.4.3 Application to the joint segmentation of copy number profiles

In Section 9.3 we reviewed methods to segment copy number profiles with a focus on three main criteria: whether an algorithm solves the optimization problem (9.2) exactly or approximately, whether its time and space complexity is linear, quadratic, or worse, and whether it can be extended to two-dimensional signals (such as the ones displayed in Figure 9.3). To complement this theoretical review, we empirically compare the performance of some of these algorithms on synthetic data generated as described in the preceding subsection.

Several measures can be defined to assess the performance of segmentation methods, see e.g. [23, Section 3]. Here we focus on how well true change points are recovered, by casting the problem of change point detection as a binary classification problem. We

define the number of true positives (TP) as the number of true change points for which at least one change point is detected closer than a given tolerance parameter. The number of false positives FP is FP = P-TP, where P is the total number of detected change points. With these definitions we are able for each method to plot a ROC curve parametrized by the number of change point found by the method. Figure 9.6 (left panel) summarizes the results obtained for a scenario with the HCC1395 breast cancer cell line, $n = 5000$, $K = 5$, and a tolerance of one data point on each side of the change points. One sample from this scenario is represented in the right part of Figure 9.6. The relatively small number of

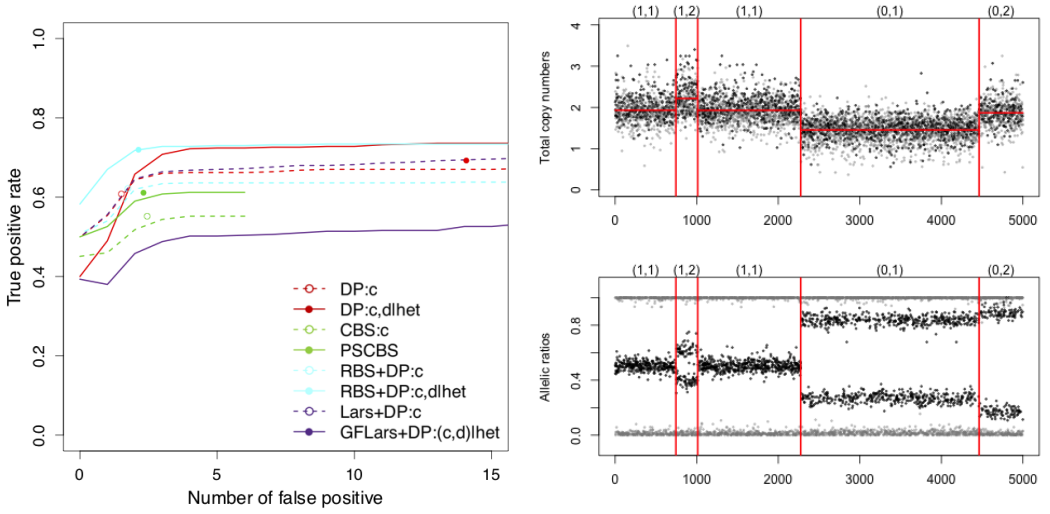


Figure 9.6: Comparing the accuracy of change point detection of different change point methods. Left: ROC curves; right: an example of sample illustrating the simulation scenario. $2d$ methods are in solid lines, 1d methods are in dashed lines.

data points ($n = 5,000$) was chosen in order for the dynamic programming (DP) approach to be applicable on the two-dimensional data (c, d), where it is quadratic as explained in Section 9.3.3. The methods compared are dynamic programming (DP, [139, 53]), (recursive) binary segmentation followed by dynamic programming (RBS + DP, [104]), circular binary segmentation (CBS [121] and PSCBS [J19]), and total variation penalty (LARS [94] and group-fused LARS [77]). For all these methods, we compared the 1d-segmentation of total copy numbers (c) alone to the $2d$ segmentation of (c, d). The main observations that may be drawn from these experiments are the following:

- The true positive rate seems to be bounded away from 1, meaning that on average, only 4 of the 5 segments could be identified.
- Most 2d methods (DP, RBS, CBS) generally perform better than their 1d counterpart, although 2d-DP is outperformed by 1d-DP at a high specificity (ie for a low number of false positives);
- The fact that the 2d total variation (GFLars) performs worse than is 1d counterpart (Lars) is an artifact of the implementation: in the implementation the data points for which one of the dimensions of the signal is missing are not considered, resulting in a substantial loss of power. Indeed, in this application only SNPs that were heterozygous in the germline have non-missing signals for d, and these SNP typically represent less than one third of the total number of SNPs.
- Although DP is the only method that exactly minimizes the objective function, it is not necessary the best performer. In the example of Figure 9.6 the 2d-RBS+DP method clearly outperforms DP at high specificities.

The above results illustrate the type of evaluation that can be performed using this framework. A general conclusion of this performance assessment is that no single method performs always better than the other ones. In particular, some methods are more robust than others to a decreasing tumor fraction, and the performance of the methods may also differ across types of change points (i.e. between a normal region and a region of gain).

An important point from a statistician's perspective is that most of the above observations would most probably not have been made if the performance of the methods had been assessed by classical simulations based on a probabilistic model for the noise. In particular, one would expect DP to have better statistical performance than its competitors, because it solves the original optimization problem (9.2) and not a relaxation. However, the statistical justification of considering this optimization problem relies on an assumption of homoskedastic Gaussian noise, which may not be appropriate for these data.

Chapter 10

Adjacency-constrained clustering

In the context of Genome Wide Association Studies (GWAS), region-scale approaches taking haplotype blocks into account can result in substantial statistical gains [158]. Hi-C studies [72] have demonstrated the existence of topological domains, which are megabase-sized local chromatin interaction domains correlating with regions of the genome that constrain the spread of heterochromatin. Both contexts raise the natural question of segmenting similarity matrices, either in the form of linkage disequilibrium (LD) matrices or of Hi-C contact maps.

In this chapter we study a particular segmentation method, hierarchical agglomerative clustering (HAC) with adjacency constraints. After defining the method and studying its conditions of applicability, we describe a quasi-linear algorithm to perform adjacency-constrained HAC under an additional assumption derived from the biological context. Finally we describe how this method can be applied to GWAS studies, and combined with group lasso regression to detect LD blocks associated to a given phenotype.

References:

[J1] N. Randriamihamison, N. Vialaneix, and P. Neuvial. *Applicability and Interpretability of Ward Hierarchical Agglomerative Clustering With or Without Contiguity Constraints*. Journal of Classification. 2020

[J5] C. Ambroise, A. Dehman, P. Neuvial, G. Rigaill, and N. Vialaneix. "Adjacency-constrained hierarchical clustering of a band similarity matrix with application to Genomics". *Algorithms for molecular biology* 14.22 (Nov. 2019)

[J9] A. Dehman, C. Ambroise, and P. Neuvial. "Performance of a Blockwise Approach in Variable Selection using Linkage Disequilibrium Information". *BMC Bioinformatics* (2015)

[S3] C. Ambroise, S. Chaturvedi, A. Dehman, P. Neuvial, G. Rigaill, and N. Vialaneix. *adjclust: Adjacency-Constrained Clustering of a Block-Diagonal Similarity Matrix*. R package version 0.5.8. 2018

Contents

10.1 HAC and adjacency-constrained HAC

HAC was initially described by Ward [193] for data in $\mathbb{R}^d$. Let $\Omega := \{x_1, \cdots, x_p\}$ be a set of p objects to be clustered, with $x_i \in \mathbb{R}^d$ for $i = 1, \ldots, d$. A cluster is a subset of Ω. The degree of inhomogeneity of a cluster $G \subset \Omega$ is quantified by the inertia (also known as *Error Sum of Squares*, ESS):

$$I(G) = \sum_{i \in G} \|x_i - \bar{x}_G\|^2, \tag{10.1}$$

where $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^d$, $\bar{x}_G = |G|^{-1} \sum_{i \in G} x_i$ is the center of gravity of G and $|G|$ denotes the cardinal of G. The loss of information when merging two disjoint clusters G and G' into $G \cup G'$ is quantified by :

$$\delta(G, G') := I(G \cup G') - I(G) - I(G'). \tag{10.2}$$

The quantity δ is known as Ward's linkage and it is equal to the variation of within-cluster inertia (also called *within-cluster sum of squares*) after merging two clusters. It also corresponds to a scaled version of the squared distance between centers of gravity:

$$\delta(G, G') = \frac{|G||G'|}{|G| + |G'|} \|\bar{x}_G - \bar{x}_{G'}\|^2. \tag{10.3}$$

The HAC algorithm is described in Algorithm 2. Starting from the trivial partition $\mathcal{P}_1 = \{\{x_1\}, \{x_2\}, \cdots, \{x_p\}\}$ with p singletons, the HAC algorithm creates a sequence of partitions by successively merging the two clusters whose linkage δ is the smallest, until all objects have been merged into a single cluster.

Algorithm 2 (Contiguity-constrained) Hierarchical Agglomerative Clustering (HAC)

1: **Initialization:** $\mathcal{P}_1 = \{\{x_1\}, \{x_2\}, \cdots, \{x_p\}\}$
2: **for** $t = 1$ to $p - 1$ **do**
3: Compute all pairwise linkage values between (contiguous) clusters of the current partition $\mathcal{P}_t$
4: Merge the two (contiguous) clusters with minimal linkage value to obtain the next partition $\mathcal{P}_{t+1}$
5: **end for**
6: **return** $\{\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_p\}$

Contiguity-constrained HAC is a simple modification of Algorithm 2, where only clusters that are adjacent are candidate mergers, as indicated by the word "contiguous" in line 3 and 4. The idea of incorporating such constraints was previously mentioned by Lebart [183] to incorporate geographical (two-dimensional) constraints to cluster socio-economic data, and by [74] to cluster functional Magnetic Resonance Imaging (fMRI) data into contiguous (three-dimensional) brain regions.

While contiguity-constrained HAC can be defined for any symmetric relation indicating which pairs of objects are considered as "contiguous", we focus in this chapter on the specific case of adjacency-constrained clustering. That is, we assume that the objects to cluster are ordered along a line, and only allow for adjacent objects or clusters of objects to be merged. This particular case has been studied by Grimm [178], and an **R** package implementing this algorithm, `rioja` [20], has been developed[1].

Formulation using pairwise Euclidean distances. The inertia of a cluster may be expressed only in function of sums of the entries of the pairwise distances ($\|x_i - x_j\|, 1 \leq$

[1] available on CRAN at `https://cran.r-project.org/package=rioja`.

$i, j \leq p$):

$$I(G) = \frac{\Delta(G, G)}{2|G|}, \tag{10.4}$$

where Δ is defined by $\Delta(G, G') = \sum_{x_i \in G, x_{i'} \in G'} \|x_i - x_j\|^2$ for any clusters G and G'. Moreover, Ward's linkage between any two clusters G and G' may be itself be written in function of these pairwise distances:

$$\delta(G, G') = \frac{1}{2}\left(\frac{\Delta(G \cup G', G \cup G')}{|G \cup G'|} - \frac{\Delta(G, G)}{|G|} - \frac{\Delta(G', G')}{|G'|}\right). \tag{10.5}$$

10.2 Extensions to possibly non-Euclidean settings

The HAC algorithm of Ward [193] has been designed to cluster elements of $\mathbb{R}^d$. In this chapter, we focus on the situation where the objects to be clustered are indirectly described by a matrix of pairwise similarities $S = (s_{ij})_{i,j=1,\ldots,p}$. The motivation of this section is an important difference between GWAS and Hi-C from a statistical perspective. While the LD (r^2) similarity is a kernel (as explained below in Section 10.4), it is not necessarily the case for Hi-C similarity matrices, possibly leading to reversals in dendrograms obtained from constrained HAC as studied in detail in [J1].

Kernels. If S is positive definite, the theory of Reproducing Kernel Hilbert Spaces [197] implies that the data can be embedded in an implicit Hilbert space. This allows to define the inertia of a cluster, and consequently to formulate Ward's linkage between any two clusters in terms of the similarity using the so-called "kernel trick": $\forall C,\ C' \subset \{1, \ldots, p\}$,

$$\delta(C, C') = \frac{S(C)}{|C|} + \frac{S(C')}{|C'|} - \frac{S(C \cup C')}{|C \cup C'|}, \tag{10.6}$$

where $S(C) = \sum_{(i,j) \in C^2} s_{ij}$ only depends on S and not on the embedding. This expression shows that Ward's Linkage also has a natural interpretation as the decrease in average intra-cluster similarity after merging two clusters. Equation (10.6) is derived from (10.3) using the formalism of kernels in Section S1.1 of the Supplementary material of [J5]. Another way to prove it is to use Equation (10.5) with the distance associated to S by the kernel mapping, $d_{ij}^2 = s_{ii} + s_{jj} - 2s_{ij}$, and noticing that the diagonal terms cancel out.

Similarities. An extension of this approach to the case of a general (that is, possibly non-positive definite) similarity matrix has been studied by Miyamoto, Abe, Endo, and Takeshita [51]. Noting that (i) for a large enough λ, the matrix $S_\lambda = S + \lambda I_p$ is positive definite and that (ii) Ward's linkage associated to S_λ is simply given by $\delta_\lambda(C, C') = \delta(C, C') + \lambda$, applying Ward's HAC to S and S_λ yields the exact same hierarchy, only shifting the linkage values by $+\lambda$ [51, Theorem 1]. This result, which a fortiori holds for adjacency-constrained Ward's HAC, justifies the use of Equation (10.6) in the case of a general similarity matrix.

Dissimilarities. For completeness, we also describe a generic construction of HAC for arbitrary dissimilarity data, proposed by [17]. This construction is based on an analogy between distance and dissimilarity. If the p objects are described by a dissimilarity matrix D, that is, a matrix satisfying: for all $i, j \in \{1, \ldots, p\}, d_{ij} \geq 0; \quad d_{ii} = 0; \quad d_{ij} = d_{ji}$, then one can simply define the inertia of a cluster by (10.4), with (sums of squared) distances replaced by (sums of squared) dissimilarities. The corresponding HAC is then formally obtained as the output of Algorithm 2. This construction generalizes the (less intrinsic) one proposed by [143] and, later, [28], for specific dissimilarities: in those papers, instead of starting from Equation (10.4), the linkage between two clusters is *defined* by a reformulation of (10.5) using the fact that $\Delta(G \cup G', G \cup G') = \Delta(G, G) + \Delta(G', G') + 2\Delta(G, G')$.

10.3 Fast segmentation of a band similarity matrix

When the p objects to be clustered belong to $\mathbb{R}^d$, with $d < p$, the computation of Ward's linkage between two clusters can be done in $O(d)$ by exploiting its explicit alternative formulation as the distance between centers of gravity (10.3). In such cases, it is possible to obtain unconstrained HAC in $O(p^2 \log^2 p)$ in time [164], and lower complexities could possibly be achieved for adjacency-constrained HAC. However, we focus in this chapter in the situation where the input objects are represented by pairwise similarities. In such cases there is generally no explicit or finite-dimensional representation of the centers of gravity, and the time complexity of adjacency-constrained HAC is intrinsically quadratic in p because all of the p^2 similarities are used to compute all of the required linkage values required by Algorithm 2. This is the case for the implementation of adjacency-constrained HAD provided in the CRAN/R package `rioja` [20].

Band similarity assumption. In applications where adjacency-constrained clustering is relevant, such as Hi-C and GWAS data analysis, this quadratic time complexity is a major practical bottleneck because p is typically of the order of 10^4 to 10^5 for each chromosome. Fortunately, in such applications it also makes sense to assume that the similarity between physically distant objects is small. Specifically, we assume that S is a band matrix of bandwidth $h+1$, where $h \in \{1 \dots p\}$: $s_{ij} = 0$ for $|i-j| \geq h$. This assumption is not restrictive, as it is always fulfilled for $h = p$. However, we will be mostly interested in the case where $h \ll p$.

We now describe an algorithm proposed in [J5] to perform adjacency-constrained HAC under this assumption. This algorithm relies on (*i*) constant-time calculation of each of the Ward linkages involved in Algorithm 2 using Equation (10.6), and (*ii*) storage of the candidate fusions in a min-heap.

Ward's linkage as a function of pre-calculated sums. We have shown in [J5] that the sum of all similarities in any cluster $C = \{i, \dots, j-1\}$ of size $k = j - i$ can easily be obtained from sums of elements in the first $\min(h,k)$ sub-diagonals of S:

Lemma 10.1. *For $1 \leq r, l \leq p$, let $P(r,l) = \sum\{s_{ij} : 1 \leq i,j \leq r, |i-j| < l\}$ and $\bar{P}(r,l) = P(p+1-r,l)$. Letting $h_k := \min(h,k)$, we have*

$$P(j,h_k) + \bar{P}(i,h_k) = S(C) + P(p,h_k)\,. \tag{10.7}$$

$P(p,h_k)$ is the "full" pencil of bandwidth h_k (which also corresponds to $\bar{P}(1,h_k)$). Lemma 10.1 is illustrated in Figure 10.1, with $r \in \{i,j\}$, for $l = k \leq h$ in the left panel, and $l = h \leq k$ in the right panel. In both panels, $P(j,\min(h,k))$ is the sum of elements in the yellow and green regions, while $\bar{P}(i,\min(h,k))$ is the sum of elements in the green and blue regions. Because P and $\bar{P}$ are sums of elements in pencil-shaped areas, we call $P(r,l)$ a *forward pencil* and $\bar{P}(r,l)$ a *backward pencil*.

Lemma 10.1 makes it possible to compute $\delta(C,C')$ in constant time from the pencil sums using Equation (10.6). By construction, all the bandwidths of the pencils involved are less than h. Therefore, only pencils $P(r,l)$ and $\bar{P}(r,l)$ with $1 \leq r \leq p$ and $1 \leq l \leq h$ have to be pre-computed, so that the total number of pencils to compute and store is less than $2ph$. These computations can be performed recursively in a $O(ph)$ time complexity.

Storing candidate fusions in a min-heap. Iteration t of Algorithm 2 consists in finding the minimum of $p-t$ elements, corresponding to the candidate fusions between the $p-t+1$ clusters in $\mathcal{C}^{t-1}$, and merging the corresponding clusters. Storing the candidate fusions in an *unordered array* and calculating the minimum at each step would mean a quadratic time complexity. One intuitive strategy would be to make use of the fact that all but 2 to 3 candidate fusions at step t are still candidate fusions at step $t-1$. However, maintaining a *totally-ordered* list of candidate fusions is not efficient because the cost of deleting and

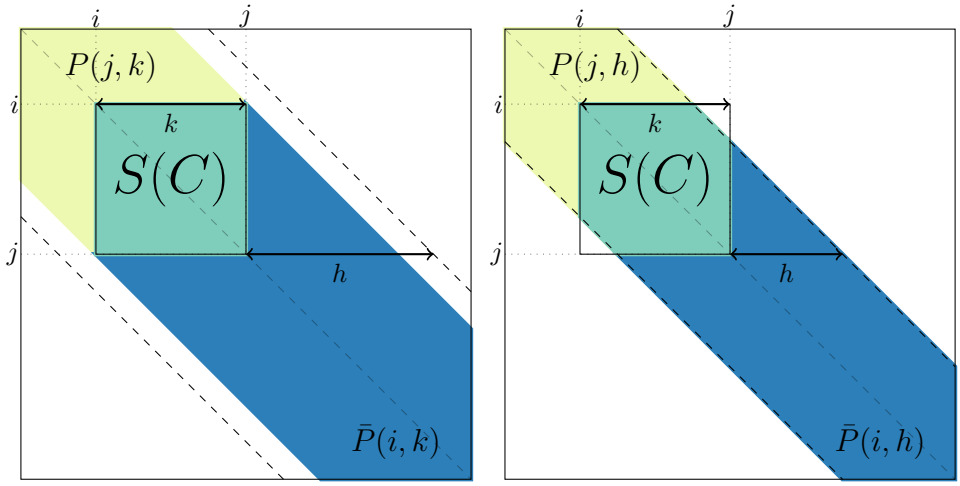


Figure 10.1: Example of forward pencils (in yellow and green) and backward pencils (in green and blue), and illustration of Equation (10.7) for cluster $C = \{i, \dots, j-1\}$. Left: cluster smaller than bandwidth ($k \leq h$); right: cluster larger than bandwidth $k \geq h$.

inserting an element in an ordered list is linear in p, again leading to a quadratic time complexity. Instead, we propose to store the candidate fusions in a *partially-ordered* data structure called a *min heap* [192]. This type of structure achieves an appropriate trade-off between the cost of maintaining the structure and the cost of finding the minimum element at each iteration.

A min heap is a binary tree such that the value of each node is smaller than the value of its two children. The advantage of this structure is that all the operations required in Algorithm 2 to create and maintain the list of candidate fusions can be done in $O(\log p)$. A detailed description of the algorithm, which is implemented in the `adjclust` package [S3], is given in [J5].

10.4 Application to Genome-Wide Association Studies

As explained in Chapter 1, genome-Wide Association Studies (GWAS) aim at identifying markers associated with a phenotype of interest. Given that an individual's genotype is characterized by millions of markers (known as SNPs) this approach yields a large multiple testing problem. Due to recombination phenomena, the hypotheses corresponding to SNPs that are close to each other along the genome are statistically dependent. This dependence is usually quantified by the linkage disequilibrium (LD), as displayed in the right panel of Figure 1.2. A widely used measure of LD in the context of GWAS is the r^2 coefficient, which can be estimated directly from genotypes measured by genotyping array or sequencing data using standard methods [43]. The similarity $S = (r^2_{ij})_{i,j}$ induced by LD can be shown to be a kernel [J5, Supplementary Materials][2].

In this section, we show how the above-described algorithm for adjacency-constrained clustering may be used in this context. In Subsection 10.4.1 we illustrate the relevance of the band similarity assumption introduced above, and the computational gains offered by the algorithm. In Subsection 10.4.2 we show how this algorithm can be used to perform tests of association at the level of LD blocks instead of SNPs. The numerical experiments below have been performed on a SNP dataset coming from a GWA study on HIV [103] mentioned in Chapter 1.

[2]The proof is a consequence of the fact that r^2 can be formulated as the squared correlation between (latent) variables that depend on haplotypes.

10.4.1 Linkage disequilibrium block inference in GWAS

We used genotype data corresponding to five chromosomes that span the typical number of SNPs per chromosome observed on the 317k Illumina genotyping microarray ($5,000 \leq p \leq 25,000$).

Quality of the band approximation. We compared the dendrogram obtained with $h < p$ to the reference dendrogram obtained with the full bandwidth ($h = p$) by recording the index t of the last clustering step (among $p-1$) for which all the preceding fusions in the two dendrograms are identical. The quantity $t/(p-1)$ can then be interpreted as a measure of similarity between dendrograms, ranging from 0 (the first fusions are different) to 1 (the dendrograms are identical). Figure 10.2 (left) displays the evolution of $t/(p-1)$ for different values of h for the five chromosomes considered here. For example, for all five chromosomes,

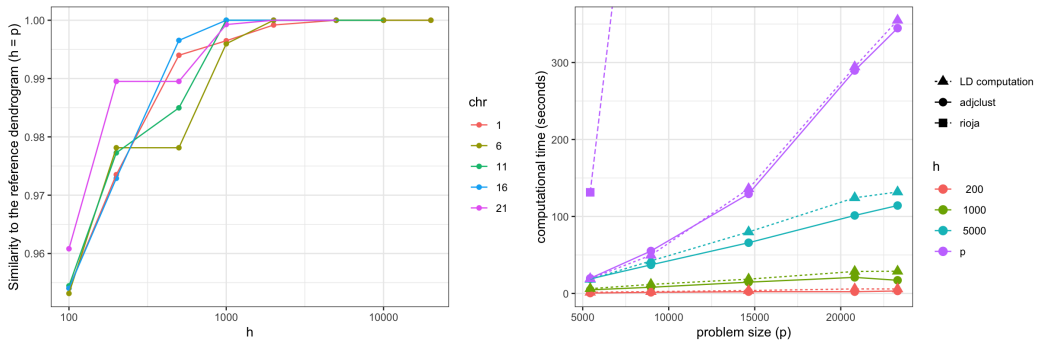


Figure 10.2: Left: quality of the band approximation as a function of the bandwidth h for five different chromosomes. Right: computation times versus p for LD matrices, and clustering with `rioja` and `adjclust`.

at $h = 1000$, the dendrograms differ from the reference dendrogram only in the last 0.5% of the clustering step. We obtained similar results when using other criteria for evaluating the quality of the band approximation, including Baker's Gamma correlation coefficient [189], which corresponds to the Spearman correlation between the ranks of fusion between all pairs of objects. Importantly, the influence of the bandwidth parameter is the same across chromosomes, that is, across values of p. Therefore, it makes sense to assume that h does not depend on p and that the time and space complexity of our proposed algorithm, which depends on h, is indeed quasi-linear in p.

Scalability and computation times. Figure 10.2 displays the computation time for the LD matrix (dotted lines) and for the adjacency-constrained HAC with respect to the size of the chromosome (x axis), both for `rioja` (dashed line) and `adjclust` (solid lines). As expected, the computation time for `rioja` did not depend on the bandwidth h, so we only represented $h = p$. For `adjclust`, the results for varying bandwidths are represented by different colors.

First, the computation times of `rioja` are several orders of magnitude larger than those of `adjclust`, even when $h = p$ where both methods implement the exact same algorithm. As expected, the complexity of `adjclust` with $h = p$ is quadratic in p, while it is essentially linear in p for fixed values of $h < p$. For large values of p the gain of the band approximation is substantial. We also note that regardless of the value of h, the total time needed for the clustering is of the order of (and generally lower than) the time needed for the computation of the LD. This implies that runnning the `adjclust` algorithm in this case comes essentially for free from a computational point of view.

10.4.2 Block-wise GWAS

To overcome the intrinsic limitations of classical Genome-wide association studies (GWAS) involving univariate tests between statistically dependent markers, we have proposed in [J9] the following approach: (i) infer LD blocks using `adjclust`, (ii) estimate the number of LD blocks using a model selection criterion based on the Gap statistic [163], and (iii) perform Group Lasso regression to identify which of the inferred LD blocks are associated with the response.

We have investigated numerically the efficiency of this approach compared to state-of-the art regression methods: haplotype association tests, single marker analyses, and Lasso and Elastic-Net regressions. Our numerical experiments show that the proposed method outperforms state-of-the-art approaches (the haplotype association module of the PLINK genome association analysis tool [119], single marker analyses and Lasso and Elastic-Net regressions) as soon as the number of causal SNPs within a LD block exceeds 2. This method has also been applied to the analysis of the HIV data set introduced in Chapter 1, where one goal is to identify LD blocks associated to the viral load (that is, the abundance of the virus in blood cells). This is illustrated by Figure 10.3, where a small part of the LD similarity matrix is displayed together with the results of the proposed method (left panel) and PLINK (right panel). The region displayed corresponds to 68 contiguous markers located in the major histocompatibility complex (MHC), of which 8 (marked with a red star *) had previously been identified as (marginally) associated to the viral load [103]. The proposed method (Figure 10.3, left plot) is able to identify LD blocks (delimited with a red contour line) larger than PLINK.

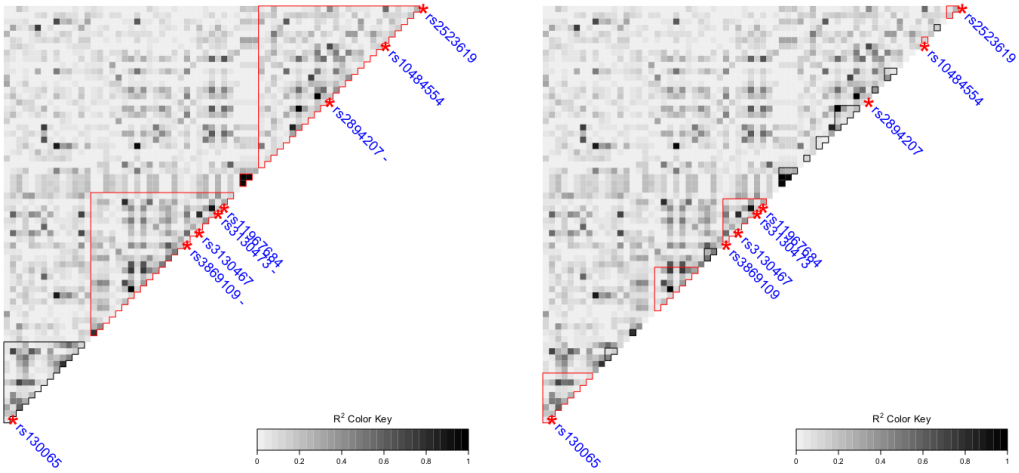


Figure 10.3: A linkage disequilibrium (r^2) plot with the inferred block structures (black and red contour lines) for a set of 68 contiguous SNPs located on the MHC region. Left: within the structure inferred by the proposed method, the blocks selected by the Group Lasso are delimited with a red contour line. The SNPs selected by single marker analysis are plotted with a red star (*), and the SNPs missed by Lasso with a blue dash (-). Right: within the structure inferred by the haplotype association method, the blocks selected by the competing method are delimited with a red contour line.

Chapter 11

Directions for future research

The previous chapters summarize some attempts to address the complexity of genomic data by considering their sparsity, heterogeneity and structure as constraints that alleviate the curse of dimensionality, and guide statistical methods toward solutions that are more plausible from a biological standpoint. My recent research contributions, particularly in the context of the SCALES project (2017-2019) funded by the CNRS MITI and the SansSouci project (2016-2020) funded by ANR have also raised a number of new research questions, that cover a broad spectrum from theory to applications. In this chapter we describe some research directions toward exploiting the multiscale nature of biological problems, and providing inferential guarantees for exploratory findings. We will pay particular attention to the evaluation of the statistical and computational "price to pay" for accounting for these important features.

Some open questions in the field of post hoc inference are discussed in Section 11.1. In Section 11.2 specific methodological challenges for inference for signals that are structured along the genome or in three dimensions are described. Section 11.3 is an opening to broader scientific challenges which I consider to be of primary importance, especially in the context of interdisciplinary research.

Contents

11.1 Emerging challenges in post hoc inference

11.1.1 Improved post hoc bounds

New JER controlling procedures. The generic framework based on JER control introduced in Section 5 makes it possible to construct several post hoc bounds based on different reference families. One natural extension is to look for new probabilistic inequalities that yield JER control, or to extend the validity of existing one. In particular, it is conjectured in the multiple testing field that the Simes inequality holds for Gaussian tests statistics with arbitrary covariance matrix. If this conjecture was proved, then the validity of the Simes post hoc bound would be automatically extended as well. Another interesting research direction is to obtain JER control for dependent tests by building reference families based on harmonic mean p-values [13].

Aggregation of reference families. Bounds based on distinct reference families may perform differently on different subsets of selected hypotheses. For example, Figure 5.6 in Chapter 5, suggests (consistently with our theoretical results) that the Linear bound performs better than the Beta bound for smaller subsets, and worse for larger subsets. Similar observations have been made in the case of locally-structured hypotheses [J3]. An interesting perspective is to use *aggregation techniques* to build bounds performing almost as well as the best possible bound for any size or shape of subset.

11.1.2 Toward Post Selection Inference and online multiple testing

Connection to post selection inference. The post hoc bounds originally proposed by [81] and our JER-based bounds [J2] intrinsically rely on testing *marginal* hypotheses. Valid inference methods for *multiple regression* after arbitrary model selection have been proposed by [65, 2]. These methods involve inflating standard confidence intervals by a constant factor called "Post Selection Inference (PoSI) constant". However, this approach cannot be used in high-dimensional settings due to the exponential time complexity (in the number p of variables) required to calculate the PoSI constant for a given design matrix. Kuchibhotla, Brown, Buja, Cai, George, and Zhao [4] have recently introduced confidence intervals enjoying two remarkable properties: they have a much reduced computational cost of $O(p^2)$, and their volume scales with the size of the selected model and not with the largest possible model size considered. An important perspective is to derive post hoc bounds from [4], and to compare these bounds to our "marginal" bounds.

Post hoc bounds for online tests. Online multiple testing consists in performing sequence of tests, when the number of tests is unknown an possibly infinite. Motivated by applications to A/B testing, much progress has been recently made for online FDR and FDP control, see [19] and the works of Aaditya Ramdas on the subject, e.g. [3, 10]). A natural question is the construction of online post hoc bounds.

11.2 Inference for structured signals

The problem of differential analysis along the genome has been studied for DNA methylation data [63] and ChIP seq data [62] among other types of genomic data. Just like differential expression analyses, the goal is to identify genomic features whose "activity" is significantly different between two biological or clinical conditions, based on the observation of several activity profiles. However, one is generally more interested in detecting differentially active regions (differentially methylated regions, or differentially bound regions in the above two examples) than differentially active individual loci. The main statistical difficulty is that the candidate regions are not known a priori and thus have to be constructed from the data. We are interested in two specific practical instances of such structured differential

analysis problems. The first one is the case of differential DNA copy number analysis in cancers between cancer subtypes. The second one is differential analysis of Hi-C contact maps. For the latter, at least two "regional" scales should be considered, which are known as Topologically Associated Domains (TAD) and A/B compartments. These scales correspond to two levels of chromosomal organization that can be detected from Hi-C contact maps from a single DNA sample [72].

In both contexts, it is possible to first perform differential analysis at the most local level, using methods derived from differential expression studies proposed in [63, 62, 49], and then to aggregate this information into a regional signal. Using the strategy outlined in [J3] it is possible to derive post hoc bounds for any candidate region. Depending on the constraints on the shape of the regions of interest, finding efficient ways of scanning all relevant regions raises interesting computational challenges.

From a statistical perspective, following the argument developed in Chapter 8 to motivate the method proposed in [J16], aggregating marginal tests can be under-powered compared to performing multivariate tests. An additional difficulty in the above applications is that contrary to the case of differential expression of pathways, the regions of interest are generally not known in advance. Accounting for this feature will rely on testing strategies tailored to each problem. For differential copy number analysis, one direction to investigate consists in fitting a Hidden Markov Model in order to obtain locus-level probabilities of differential expression that account for spatial dependence, and to derive a tailored JER control. Another relevant direction consist in first performing a joint segmentation of the genome as in Chapter 9, and then design tailored tests of differential copy number analysis of such data-driven segments. For differential Hi-C analysis, one possibility is to elaborate on the TAD detection method presented in Chapter 10. We expect that these methodological developments will also raise interesting statistical and computational challenges.

11.3 Broader scientific challenges

In this section I discuss broader scientific challenges that I believe as a statistician to be fundamental for the research community to work in a sound and efficient manner. While I will certainly not solve these challenges by myself, I hope to be able to make some contributions to some of the points they raise in future years.

11.3.1 Interdisciplinarity: methods to connect theory and practice

An important context element for method development in genomic data analysis is the rapid evolution of genomic assays, which has several consequences for statisticians. First, the increasing complexity of these assays makes it difficult for non-specialists to gain enough understanding to propose relevant models for statistical and experimental variability. Then, the number of proposed bioinformatic methods makes it difficult to identify state-of-the-art methods, and the fact that these methods are typically pipelines made of several tools hampers the statistical understanding of the strengths and weaknesses of such methods. Finally, and as a result, the time needed to finely address the statistical issues raised by the analysis of a given technology can be much longer than the half-life of this technology. Thus, the breadth of the interface between biology/genomics applications and statistical theory is expanding. An important point for future years is the ability for the scientific community to strengthen the links between theory and applications.

My personal experience is that it is difficult to foster scientific interactions between biology and mathematics (or even biology and statistics) at a macroscopic scale. I have found it more effective and more stimulating to organize seminars, reading groups or workshops dedicated to the statistical, bioinformatic and biological challenges raised by the analysis of one particular type of data, or one particular biological question. In the context of the PhD thesis of Nathanaël Randriamihamison on Hi-C data analysis, we have created in 2018 in Toulouse a very active reading group called chrocogen (for chromatin Conformation and gene

expression), which gathers statisticians, biostatisticians, bioinformaticians and biologists. We have also organized a successful workshop on Hi-C data analysis in December 2019.

11.3.2 Reproducible research and performance evaluation

Reproducible research and the role of statisticians. The ability to reproduce or replicate scientific experiments is at the very core of the scientific method. Almost fifteen years ago, a paper raised awareness by claiming that an important proportion of published biomedical studies could not be replicated [134]. This issue is not specific to biomedical sciences[1]. While the actual proportion of "wrong" papers in the scientific literature can be discussed (see e.g. [48, 69]), there is consensus that this proportion is unacceptably large. Most results in biomedical studies are obtained from some data analysis. A first and obvious necessary condition for reproducing such computational results is the presence of all the required data and code. A number of journals now favor methods reproducibility by imposing to publish code and data allowing to reproduce the results of a paper; see also the special collection "Challenges in irreproducible research" from the Nature journals. The question of "reproducibility" of research findings has several important aspects from a statistician's perspective:

1. Part of non-reproducibility can simply be attributed to the fact that a number of studies still disregard or poorly address the issue of selection or multiple testing. Even when all of the hypotheses tested are explicitly reported in a study, it is not uncommon that only those whose p-value is less than 0.05 are retained and mentioned in the abstract. This issue can readily be addressed by "classical" multiple testing procedures, and it is worrisome that not all scientific journals seem to be able to catch these errors before publication.

2. Scientific results may be non-reproducible because of selection biases in the study, such as silently adding/removing some cases in a study, reporting only the results of one particular combination of method and tuning parameters whereas many such combinations have also been tried. This issue of *fishing for significance* can partly be addressed by pre-registration of the methods to be used: such practices are highly relevant to confirmatory analyses, clinical trials, but less so in exploratory contexts. We believe that developments in post hoc inference could help address this issue.

3. Similar issues can be found in the statistical methodology literature itself. In papers presenting new statistical methods, it is not uncommon that the performance of several methods are tested on several synthetic or experimental data sets and/or performance criteria, but only a subset including those favoring the method proposed by the authors of the paper are reported. This issue has been studied in depth in [41], and practical guidelines for benchmarking studies have been proposed in [11].

Importance of performance evaluation in scientific literature. The bias in the methodological literature discussed in the last item partly comes from the current publication system itself. Indeed, in order for a statistical method to be published, most journals require evidence of the *superiority* of this method on existing ones in simulations or data set analyses, which encourages the above-discussed selective reporting of numerical experiments. This lack of objectivity is problematic per se (leading for example to very contrasted performance assessments for a given method in different papers), but also because the literature on statistical methods often provides only limited insight into *when* and *why* a given method is appropriate or not, whereas these questions are actually those that contribute to a global understanding and improvement of the available methods.

While the need for neutral comparison studies (also known as benchmarking studies) is widely acknowledged by us as statisticians, it is not always (or even often not) applied in our

[1] "Is There a Reproducibility Crisis in Science?". Nature Video, Scientific American. 28 May 2016.

own methods publications, as noted in [26]. The question of designing such benchmarking studies in several fields has itself become a subject of research, see e.g. [15, 8, 9, 11] for recent contributions and the STRATOS Initiative[2] in clinical biostatistics. Some statistical journals including *Briefings in Bioinformatics* or *Genome Biology* explicitly welcome neutral comparison studies. These stimulating questions provide food for thought for building novel forms of scientific publications that could be specifically designed for these types of studies.

11.3.3 Toward truly applicable methods

Although the applied statistics community has made tremendous efforts to make newly developed methods available for the bioinformatics or biology community, I believe that there is still much room for improvement in this direction, and that such improvement is crucial to consolidate the links between theory and applications.

R packages. The past decades have undergone massive developments of open-source implementations of data analysis methods, especially in R and python. The current standard for the implementation of statistical methods is in the form of R packages. From a statistician's perspective, there is a clear tension between the *collective* interest of putting together and maintaining well-documented packages, and the corresponding *individual* time requirements – in particular, the more a package is used, the more costly it is to maintain it. As a result, the sweet spot for developers of statistical methods is currently to put together a (github) repository where their code is available, generally in the form of an R package[3]. However, an important caveat of this package-centered model is that several (and possibly dozens of) packages using different inputs and outputs can be devoted to similar statistical problems. This makes it difficult for end users to know which implementation to use for a particular task[4].

The computational statistics community should continue to foster the development of open-source implementations of new statistical methods. The scikit-learn Python library for machine learning is based on a different model, relying on a growing user and developer community to contribute to the code and documentation of a single module, which is a well-recognized entry point. This project has recently been awarded a prestigious prize from the Académie des Sciences, but it has required a huge time investment from the main contributors [5]. The time saved (collectively) by having a readily usable implementation of recent statistical methods is invaluable, not only for potential users, but also for developers of competing methods. I believe that it is important for the efficiency and quality of current and future scientific research to increase the (individual) support and recognition of this type of contributions.

Interactive visualization and inference. Interactive graphical user interfaces (GUI) are an important tool for the diffusion of statistical methods toward end users with little or no programming skills. Post hoc inference methods play particularly nicely with interactive visualization tools because these methods by construction allow the user to perform data

[2]STRATOS stands for STRengthening Analytical Thinking for Observational Studies, see `http://www.stratos-initiative.org/`.

[3]This is a clear improvement with respect to the standards of 10-15 years ago, where implementations were generally "available upon request" to the authors. This progress is a consequence of (i) the more and more frequent requirements of journals to make code truly available, (ii) the development of efficient packages for package development [12], testing [89], and documentation [25], and (iii) the recognition of R packages as a contribution in the applied statistic community, which can even lead to dedicated publications [J8].

[4]To address this difficulty, the Comprehensive R Archive Network (CRAN) offers Task Views, which give a list of packages (curated by specialists) pertaining to a given topic. For example, the recent task view on missing data lists more than 130 CRAN packages related to this particular topic. Also, the developers of Bioconductor packages, which are specifically dedicated to computational biology, are encouraged to use standard methods and classes.

[5]See the blog post Getting a big scientific prize for open-source software by Gaël Varoquaux.

snooping without compromising the statistical guarantees offered by the methods, thereby making statistical analysis truly interactive. The development of such tools has been made easier for R users with the advent of Shiny Applications. For example, we have quickly developed a proof of concept application as a supplementary material for the book chapter [P1], in order to illustrate the interest of post hoc bounds for localized signals[6]. I plan to make future contributions available in the form of such applications, in particular for the interactive visualization and analysis of DNA copy number profiles in cancers, and for the differential analysis of Hi-C data. Finally, such developments are a not only a way to bridge the gap between theory and practice: developing such applications together with end users can also yield a better understanding of which type of signal is relevant, therefore possibly opening the way for new statistical developments to specifically detect this particular type of signal.

[6]This app is available at `https://pneuvial.shinyapps.io/posthoc-bounds_ordered-hypotheses/`.

Chapter 12

Scientific production

Preprints

[P1] G. Blanchard, P. Neuvial, and E. Roquain. *On agnostic post hoc approaches to false positive control*. Book chapter in revision for Handbook of Multiple Comparisons; available from `https://hal.archives-ouvertes.fr/hal-02320543`. Oct. 2019 (cit. on pp. 25, 41, 102).

Journal papers

[J2] G. Blanchard, P. Neuvial, and E. Roquain. "Post Hoc Confidence Bounds on False Positives Using Reference Families". *Annals of Statistics* 48.3 (2020), pp. 1281–1303 (cit. on pp. 16, 25, 41, 42, 45–47, 49, 98).

[J3] G. Durand, G. Blanchard, P. Neuvial, and E. Roquain. "Post hoc false positive control for structured hypotheses". *Scandinavian Journal of Statistics* (2020) (cit. on pp. 25, 41, 52–54, 98, 99).

[J4] F. Pont, M. Tosolini, Q. Gao, M. Perrier, Madrid-Mencía, T. S. Huang, P. Neuvial, M. Ayyoub, K. Nazor, and J.-J. Fournié. "Single-Cell Virtual Cytometer allows user-friendly and versatile analysis and visualization of multimodal single cell RNAseq datasets". *NAR Genomics and Bioinformatics* (2020).

[J1] N. Randriamihamison, N. Vialaneix, and P. Neuvial. *Applicability and Interpretability of Ward Hierarchical Agglomerative Clustering With or Without Contiguity Constraints*. Journal of Classification. 2020 (cit. on pp. 60, 89, 91).

[J5] C. Ambroise, A. Dehman, P. Neuvial, G. Rigaill, and N. Vialaneix. "Adjacency-constrained hierarchical clustering of a band similarity matrix with application to Genomics". *Algorithms for molecular biology* 14.22 (Nov. 2019) (cit. on pp. 60, 89, 91–93).

[J6] F. Bachoc, G. Blanchard, and P. Neuvial. "On the post selection inference constant under restricted isometry properties". *Electron. J. Statist.* 12.2 (Nov. 2018), pp. 3736–3757. ISSN: 1935-7524 (cit. on p. 25).

[J7] B. Sadacca, A.-S. Hamy-Petit, C. Laurent, P. Gestraud, H. Bonsang-Kitzis, A. Pinheiro, J. Abecassis, P. Neuvial, and F. Reyal. "New insight for pharmacogenetics studies from the transcriptional analysis of two large-scale cancer cell line panels". *Scientific Reports* 7 (2017), p. 15126.

[J8] A. Chambaz and P. Neuvial. "tmle.npvi: targeted, integrative search of associations between DNA copy number and gene expression, accounting for DNA methylation". *Bioinformatics* 31.18 (2015), pp. 3054–6 (cit. on pp. 58, 61, 63, 101).

[J9] A. Dehman, C. Ambroise, and P. Neuvial. “Performance of a Blockwise Approach in Variable Selection using Linkage Disequilibrium Information”. *BMC Bioinformatics* (2015) (cit. on pp. 60, 89, 95).

[J10] T. Picchetti, J. Chiquet, M. Elati, P. Neuvial, R. Nicolle, and E. Birmelé. “A model for gene deregulation detection using expression data”. *BMC Systems Biology* (Dec. 2015) (cit. on pp. 58, 59).

[J11] M. Pierre-Jean, G. J. Rigaill, and P. Neuvial. “Performance evaluation of DNA copy number segmentation methods”. *Briefings in Bioinformatics* 4 (2015), pp. 600–615 (cit. on pp. 60, 77, 83).

[J12] I. Brito, P. Hupé, P. Neuvial, and E. Barillot. “Stability-based comparison of class discovery methods for array-CGH profiles”. *PLoS One* 8.12 (Dec. 2013), e81458 (cit. on p. 59).

[J13] P. Neuvial. “Asymptotic Results on Adaptive False Discovery Rate Controlling Procedures Based on Kernel Estimators”. *Journal of Machine Learning Research* 14 (2013), pp. 1423–1459 (cit. on pp. 24, 27, 28, 30–34).

[J14] A. Chambaz, P. Neuvial, and M. J. van der Laan. “Estimation of a Non-Parametric Variable Importance Measure of a Continuous Exposure”. *Electron. J. Statist.* 6 (2012), pp. 1059–1099 (cit. on pp. 58, 61–63, 65).

[J15] L. Heiser et al. “Subtype and pathway specific responses to anticancer compounds in breast cancer”. *Proceedings of the National Academy of Sciences* 109.8 (Feb. 2012), pp. 2724–2729 (cit. on p. 60).

[J16] L. Jacob, P. Neuvial, and S. Dudoit. “More Power via Graph-Structured Tests for Differential Expression of Gene Networks”. *Annals of Applied Statistics* 6.2 (2012), pp. 561–600 (cit. on pp. 58, 69, 74, 75, 99).

[J17] P. Neuvial and E. Roquain. “On false discovery rate thresholding for classification under sparsity”. *Annals of Statistics* 40.5 (2012), pp. 2572–2600 (cit. on pp. 25, 35, 37–39).

[J18] M. Ortiz-Estevez, A. Aramburu, H. Bengtsson, P. Neuvial, and A. Rubio. “CalMaTe: A Method and Software to Improve Allele-Specific Copy Number of SNP Arrays for Downstream Segmentation”. *Bioinformatics* 28.13 (July 2012), pp. 1793–1794 (cit. on pp. 59, 77, 80).

[J19] A. B. Olshen, H. Bengtsson, P. Neuvial, P. T. Spellman, R. A. Olshen, and V. E. Seshan. “Parent-specific copy number in paired tumor-normal studies using circular binary segmentation”. *Bioinformatics* 27.15 (Aug. 2011), pp. 2038–2046 (cit. on pp. 59, 77, 82, 86).

[J20] The Cancer Genome Atlas Research Network. “Integrated Genomic Analyses of Ovarian Carcinoma”. *Nature* 474.7353 (June 2011), pp. 609–615 (cit. on p. 60).

[J21] H. Bengtsson, P. Neuvial, and T. P. Speed. “TumorBoost: Normalization of allele-specific tumor copy numbers from a single pair of tumor-normal genotyping microarrays”. *BMC Bioinformatics* 11.1 (2010), p. 245 (cit. on pp. 59, 77, 79).

[J22] H. Noushmehr et al. “Identification of a CpG Island Methylator Phenotype that Defines a Distinct Subgroup of Glioma”. *Cancer Cell* 17.5 (Apr. 2010), pp. 510–522 (cit. on p. 60).

[J23] M. A. Bollet et al. “High-resolution mapping of DNA breakpoints to define true recurrences among ipsilateral breast cancers.” *J Natl Cancer Inst* 100.1 (2008), pp. 48–58 (cit. on p. 59).

[J24] P. Neuvial. “Asymptotic properties of false discovery rate controlling procedures under independence”. *Electron. J. Statist.* 2 (2008). With corrigendum in EJS 2009(3):1083, pp. 1065–1110 (cit. on pp. 24, 27–32, 34).

[J25] M. Elati, P. Neuvial, M. Bolotin-Fukuhara, E. Barillot, F. Radvanyi, and C. Rouveirol. "LICORN: LearnIng COoperative Regulation Networks". *Bioinformatics* 23.18 (2007), pp. 2407–2414 (cit. on p. 58).

[J26] P. La Rosa et al. "VAMP: visualization and analysis of array-CGH, transcriptome and other molecular profiles." *Bioinformatics* 22.17 (Sept. 2006), pp. 2066–2073 (cit. on p. 59).

[J27] S. Liva, P. Hupé, P. Neuvial, I. Brito, E. Viara, P. La Rosa, and E. Barillot. "CAPweb: a bioinformatics CGH array Analysis Platform." *Nucleic Acids Res* 34.Web Server issue (July 2006), pp. 477–481 (cit. on p. 59).

[J28] P. Neuvial, P. Hupé, I. Brito, S. Liva, E. Manié, C. Brennetot, F. Radvanyi, A. Aurias, and E. Barillot. "Spatial normalization of array-CGH data." *BMC Bioinformatics* 7.1 (May 2006), p. 264 (cit. on p. 59).

In proceedings

[C1] M. Champion, J. Chiquet, P. Neuvial, M. Elati, and E. Birmelé. "Identification of deregulated transcription factors involved in subtypes of cancers". *International Conference on Bioinformatics and Computational Biology*. 2020 (cit. on pp. 58, 59).

Book chapters

[B1] P. Neuvial, H. Bengtsson, and T. P. Speed. "Statistical analysis of Single Nucleotide Polymorphism microarrays in cancer studies". *Handbook of Statistical Bioinformatics*. Ed. by H. H.-S. Lu, B. Schölkopf, and H. Zhao. Springer Handbooks of Computational Statistics. Springer, 2011 (cit. on pp. 60, 77, 80, 81).

Popular science

[V1] P. Neuvial. "Vers une médecine personnalisée grâce à la recherche en génomique". *Variances* 48 (Oct. 2013), pp. 31–33.

[V2] P. Neuvial. "Tests multiples en génomique". *La gazette des mathématiciens* 130 (Oct. 2011), pp. 71–76.

[V3] P. Neuvial and P.-Y. Bourguignon. "Problématiques statistiques à l'heure de la post-génomique". *Variances* 35 (Feb. 2009), pp. 56–60.

PhD thesis

[T1] P. Neuvial. "Contributions à l'analyse statistique des données de puces à ADN". PhD thesis. Institut Curie et Université Paris VII (France), 2008 (cit. on p. 28).

Unpublished technical reports

[R1] E. Hauvuy, B. Lebrave, and P. Neuvial. "Analyse statistique du lien entre les plages homogènes de séquences d'ADN de différentes bactéries". MA thesis. ENSAE Paristech et Université Paris Diderot, 2003 (cit. on p. 2).

[R2] R. Elie, A. Frachot, P. Georges, and P. Neuvial. *A Model of Prepayment for the French Residential Loan Market*. Tech. rep. Groupe de Recherche Opérationnelle, Crédit Lyonnais, France, 2002 (cit. on p. 2).

Software

[S1] G. Blanchard, G. Durand, P. Neuvial, and E. Roquain. *sansSouci: Post Hoc Multiple Testing Inference*. R package version 0.8.0. 2019 (cit. on pp. 25, 41).

[S2] M. Pierre-Jean, G. Rigaill, and P. Neuvial. *jointseg: Joint segmentation of multivariate (copy number) signals*. R package version 1.0.2. 2019 (cit. on pp. 60, 77, 83).

[S3] C. Ambroise, S. Chaturvedi, A. Dehman, P. Neuvial, G. Rigaill, and N. Vialaneix. *adjclust: Adjacency-Constrained Clustering of a Block-Diagonal Similarity Matrix*. R package version 0.5.8. 2018 (cit. on pp. 60, 89, 93).

[S4] M. Pierre-Jean and P. Neuvial. *acnr: Annotated Copy-Number Regions*. R package version 1.0.0. 2017 (cit. on pp. 60, 77, 83, 84).

[S5] A. Chambaz and P. Neuvial. *Targeted Learning of a Non-Parametric Variable Importance Measure of a Continuous Exposure*. R package version 0.10.0. 2015 (cit. on pp. 58, 61, 63).

[S6] L. Jacob and P. Neuvial. *DEGraph: Two-sample tests on a graph*. Bioconductor R package version 1.37.0. 2012 (cit. on pp. 58, 69).

[S7] P. Neuvial and P. Hupé. *MANOR: Micro-Array data NORmalization*. Bioconductor R package version 1.58.0. 2006 (cit. on p. 59).

Contributions to other software

[s1] A. Olshen et al. *PSCBS: Analysis of Parent-Specific DNA Copy Numbers.* R package, CRAN. 2011 (cit. on pp. 59, 82).

[s2] M. Ortiz-Estevez et al. *CalMaTe: A post-calibration process to improve allele-specific copy number estimates from SNP microarrays.* R package, CRAN. 2011 (cit. on pp. 59, 80).

[s3] H. Bengtsson and P. Neuvial. *aroma.cn: Analysis of copy-number estimates obtained from various platforms.* R package, aroma-project. 2010 (cit. on pp. 59, 79).

[s4] H. Bengtsson and P. Neuvial. *aroma.cn.eval: Evaluating copy-number estimates.* R package, aroma-project. 2010.

[s5] P. La Rosa et al. *VAMP: Visualisation and Analysis of Molecular Profiles*. 2006 (cit. on p. 59).

[s6] S. Liva et al. *CAPweb: Copy Number Microarray Analysis Platform*. 2006 (cit. on p. 59).

Chapter 13

Bibliographic references

[1] F. Bachoc, H. Leeb, and B. M. Pötscher. "Valid confidence intervals for post-model-selection predictors". *Annals of Statistics* (to appear) (cit. on p. 24).

[2] F. Bachoc, D. Preinerstorfer, and L. Steinberger. "Uniformly valid confidence intervals post-model-selection". *Annals of Statistics* (to appear) (cit. on pp. 24, 98).

[3] E. Katsevich and A. Ramdas. "Simultaneous high-probability bounds on the false discovery proportion in structured, regression, and online settings". *Annals of Statistics* (to appear) (cit. on pp. 21, 22, 45, 98).

[4] A. K. Kuchibhotla, L. D. Brown, A. Buja, J. Cai, E. I. George, and L. Zhao. "Valid Post-selection Inference in Model-free Linear Regression". *the Annals to Statistics* (to appear) (cit. on pp. 24, 98).

[5] A. Rinaldo, L. Wasserman, M. G'Sell, J. Lei, and R. Tibshirani. "Bootstrapping and sample splitting for high-dimensional, assumption-free inference". *Annals of Statistics* (to appear) (cit. on p. 23).

[6] X. Tang, K. Li, and M. Ghosh. *Properties of multiple testing procedures for Student's t distributions*. Sumbitted to Statistica Sinica (2015) (cit. on p. 39).

[7] J. Hemerik, A. Solari, and J. J. Goeman. "Permutation-based simultaneous confidence bounds for the false discovery proportion". *Biometrika* 106.3 (July 2019), pp. 635–649 (cit. on p. 50).

[8] T. P. Morris, I. R. White, and M. J. Crowther. "Using simulation studies to evaluate statistical methods". *Statistics in medicine* 38.11 (2019), pp. 2074–2102 (cit. on p. 101).

[9] M. D. Robinson and O. Vitek. *Benchmarking comes of age*. 2019 (cit. on p. 101).

[10] J. Tian and A. Ramdas. "ADDIS: an adaptive discarding algorithm for online FDR control with conservative nulls". *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Curran Associates, Inc., 2019, pp. 9383–9391 (cit. on p. 98).

[11] L. M. Weber, W. Saelens, R. Cannoodt, C. Soneson, A. Hapfelmeier, P. P. Gardner, A.-L. Boulesteix, Y. Saeys, and M. D. Robinson. "Essential guidelines for computational method benchmarking". *Genome biology* 20.1 (2019), p. 125 (cit. on pp. 100, 101).

[12] H. Wickham, J. Hester, and W. Chang. *devtools: Tools to Make Developing R Packages Easier*. R package version 2.2.1. 2019 (cit. on p. 101).

[13] D. J. Wilson. "The harmonic mean p-value for combining dependent tests". *Proceedings of the National Academy of Sciences* 116.4 (2019), pp. 1195–1200 (cit. on p. 98).

[14] J.-M. Azaïs, Y. De Castro, S. Mourareau, et al. "Power of the spacing test for least-angle regression". *Bernoulli* 24.1 (2018), pp. 465–492 (cit. on p. 23).

[15] A.-L. Boulesteix, H. Binder, M. Abrahamowicz, W. Sauerbrei, and S. P. of the STRATOS Initiative. "On the necessity and design of studies comparing statistical methods". *Biometrical Journal* 60.1 (2018), pp. 216–218 (cit. on p. 101).

[16] A. Celisse, G. Marot, M. Pierre-Jean, and G. Rigaill. "New efficient algorithms for multiple change-point detection with reproducing kernels". *Computational Statistics & Data Analysis* 128 (2018), pp. 200–220 (cit. on p. 60).

[17] M. Chavent, V. Kuentz-Simonet, A. Labenne, and J. Saracco. "ClustGeo2: an R package for hierarchical clustering with spatial constraints". *Computational Statistics* 33.4 (2018), pp. 1799–1822 (cit. on p. 91).

[18] J. Hemerik and J. J. Goeman. "Exact testing with random permutations". *Test* 811–825 (27 2018) (cit. on p. 17).

[19] A. Javanmard and A. Montanari. "Online rules for control of false discovery rate and false discovery exceedance". *Ann. Statist.* 46.2 (Apr. 2018), pp. 526–554 (cit. on p. 98).

[20] S. Juggins. *rioja: Analysis of Quaternary Science Data*. R package version 0.9-15.1. 2018 (cit. on pp. 90, 92).

[21] D. Kivaranovic and H. Leeb. "Expected length of post-model-selection confidence intervals conditional on polyhedral constraints". *arXiv preprint arXiv:1803.01665* (2018) (cit. on p. 23).

[22] R. J. Meijer, T. J. Krebs, and J. J. Goeman. "Hommel's procedure in linear time". *Biometrical Journal* (2018) (cit. on p. 43).

[23] C. Truong, L. Oudre, and N. Vayatis. "A review of change point detection methods". *arXiv preprint arXiv:1801.00718* (2018) (cit. on pp. 81, 82, 85).

[24] M. J. van der Laan and S. Rose. *Targeted learning in data science: causal inference for complex longitudinal studies*. Springer Verlag, 2018 (cit. on p. 62).

[25] H. Wickham, P. Danenberg, and M. Eugster. *roxygen2: In-Line Documentation for R*. R package version 6.1.1. 2018 (cit. on p. 101).

[26] A.-L. Boulesteix, R. Wilson, and A. Hapfelmeier. "Towards evidence-based computational statistics: lessons from clinical research on the role and design of real-data benchmark studies". *BMC medical research methodology* 17.1 (2017), p. 138 (cit. on p. 101).

[27] A. Chatterjee, E. J. Rodger, I. M. Morison, M. R. Eccles, and P. A. Stockwell. "Tools and strategies for analysis of genome-wide and gene-specific DNA methylation patterns". *Methods in Molecular Biology*. Springer, 2017, pp. 249–277 (cit. on p. 4).

[28] T. Strauss and M. J. von Maltitz. "Generalising Ward's method for use with Manhattan distances". *PLoS ONE* 12 (2017), e0168288 (cit. on p. 91).

[29] S. Arlot, A. Celisse, and Z. Harchaoui. "A kernel multiple change-point algorithm via model selection". Preprint arXiv: 1202.3878. 2016 (cit. on p. 60).

[30] S. Delattre and E. Roquain. "On empirical distribution function of high-dimensional Gaussian vector components with an application to multiple testing". *Bernoulli* 22.1 (2016), pp. 302–324 (cit. on p. 34).

[31] G. Durand. "Multiple testing and post hoc bounds for heterogeneous data". PhD thesis. 2016 (cit. on p. 25).

[32] J. Gallier. "Spectral theory of unsigned and signed graphs. applications to graph clustering: a survey". *arXiv preprint arXiv:1601.04692* (2016) (cit. on p. 72).

[33] C. Gawad, W. Koh, and S. R. Quake. “Single-cell genome sequencing: current state of the science”. *Nature Reviews Genetics* 17.3 (2016), p. 175 (cit. on p. 4).

[34] P. Ghosh, X. Tang, M. Ghosh, and A. Chakrabarti. “Asymptotic Properties of Bayes Risk of a General Class of Shrinkage Priors in Multiple Hypothesis Testing Under Sparsity”. *Bayesian Anal.* 11.3 (Sept. 2016), pp. 753–796 (cit. on p. 39).

[35] S. Goodwin, J. D. McPherson, and W. R. McCombie. “Coming of age: ten years of next-generation sequencing technologies”. *Nature Reviews Genetics* 17.6 (2016), p. 333 (cit. on p. 2).

[36] J. D. Lee, D. L. Sun, Y. Sun, J. E. Taylor, et al. “Exact post-selection inference, with application to the lasso”. *The Annals of Statistics* 44.3 (2016), pp. 907–927 (cit. on p. 23).

[37] R. Nakato and K. Shirahige. “Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation”. *Briefings in bioinformatics* 18.2 (2016), pp. 279–290 (cit. on p. 4).

[38] M. Pierre-Jean. “Development of statistical methods for DNA copy number analysis in cancerology.” PhD thesis. 2016 (cit. on p. 60).

[39] J. E. Taylor, J. R. Loftus, and R. J. Tibshirani. “Inference in adaptive regression via the Kac-Rice formula”. *Ann. Statist.* 44.2 (Apr. 2016), pp. 743–770 (cit. on p. 23).

[40] R. J. Tibshirani, J. Taylor, R. Lockhart, and R. Tibshirani. “Exact Post-Selection Inference for Sequential Regression Procedures”. *Journal of the American Statistical Association* 111.514 (2016), pp. 600–620 (cit. on p. 23).

[41] A.-L. Boulesteix, V. Stierle, and A. Hapfelmeier. “Publication bias in methodological computational research”. *Cancer informatics* 14 (2015), CIN–S30747 (cit. on p. 100).

[42] J. Chiquet. “Contributions to Sparse Methods for Complex Data Analysis”. Habilitation thesis. Université d’Evry val d’Essonne, 2015 (cit. on p. 5).

[43] D. Clayton. *snpStats: SnpMatrix and XSnpMatrix classes and methods*. R package version 1.24.0. 2015 (cit. on p. 93).

[44] A. Dehman. “Spatial clustering of linkage disequilibrium blocks for genome-wide association studies”. PhD thesis. 2015 (cit. on p. 60).

[45] R. Dezeure, P. Bühlmann, L. Meier, N. Meinshausen, et al. “High-dimensional inference: Confidence intervals, p-values and R-software hdi”. *Statistical science* 30.4 (2015), pp. 533–558 (cit. on p. 23).

[46] C. Giraud. *Introduction to high-dimensional statistics*. Vol. 139. Monographs on Statistics and Applied Probability. CRC Press, Boca Raton, FL, 2015, pp. xvi+255. ISBN: 978-1-4822-3794-8 (cit. on p. 6).

[47] H. Leeb, B. M. Pötscher, K. Ewald, et al. “On various confidence intervals post-model-selection”. *Statistical Science* 30.2 (2015), pp. 216–227 (cit. on p. 23).

[48] J. T. Leek and R. D. Peng. “Statistics: P values are just the tip of the iceberg”. *Nature News* 520.7549 (2015), p. 612 (cit. on p. 100).

[49] A. T. Lun and G. K. Smyth. “diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data”. *BMC bioinformatics* 16.1 (2015), p. 258 (cit. on p. 99).

[50] R. J. Meijer, T. J. Krebs, and J. J. Goeman. “A region-based multiple testing method for hypotheses ordered in space or time”. *Statistical Applications in Genetics and Molecular Biology* 14.1 (2015), pp. 1–19 (cit. on p. 52).

[51] S. Miyamoto, R. Abe, Y. Endo, and J.-I. Takeshita. “Ward method of hierarchical clustering for non-Euclidean similarity measures”. *Proceedings of the VIIth International Conference of Soft Computing and Pattern Recognition (SoCPaR 2015)*. Fukuoka, Japan: IEEE, 2015 (cit. on p. 91).

[52] J. A. Reuter, D. V. Spacek, and M. P. Snyder. “High-throughput sequencing technologies”. *Molecular Cell* 58.4 (2015), pp. 586–597 (cit. on p. 59).

[53] G. Rigaill. “A pruned dynamic programming algorithm to recover the best segmentations with 1 to K_max change-points.” *Journal de la Société Française de Statistique* 156.4 (2015), pp. 180–205 (cit. on pp. 81–83, 86).

[54] E. Roquain. “Contributions to multiple testing theory for high-dimensional data”. Habilitation thesis. Université Pierre et Marie Curie, 2015 (cit. on pp. 11, 14).

[55] X. Tang, K. Li, and M. Ghosh. “Bayesian Multiple Testing Under Sparsity for Polynomial-Tailed Distributions”. *arXiv preprint 1509.08100* (2015) (cit. on p. 39).

[56] J. Taylor and R. J. Tibshirani. “Statistical learning and selective inference”. *Proceedings of the National Academy of Sciences* 112.25 (2015), pp. 7629–7634 (cit. on p. 21).

[57] Y. Benjamini and M. Bogomolov. “Selective inference on multiple families of hypotheses”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76.1 (2014), pp. 297–318 (cit. on p. 22).

[58] P. Bühlmann and J. Mandozzi. “High-dimensional variable screening and bias in subsequent inference, with an empirical comparison”. *Computational Statistics* 29.3-4 (2014), pp. 407–430 (cit. on p. 23).

[59] T. Dickhaus. “Simultaneous statistical inference”. *AMC* 10 (2014), p. 12 (cit. on pp. 11, 17).

[60] J. J. Goeman and A. Solari. “Multiple hypothesis testing in genomics”. *Statistics in medicine* 33.11 (2014), pp. 1946–1978 (cit. on pp. 11, 14, 19).

[61] R. Lockhart, J. Taylor, R. J. Tibshirani, and R. Tibshirani. “A significance test for the lasso”. *Ann. Statist.* 42.2 (Apr. 2014), pp. 413–468 (cit. on p. 23).

[62] A. T. Lun and G. K. Smyth. “De novo detection of differentially bound regions for ChIP-seq data using peaks and windows: controlling error rates correctly”. *Nucleic acids research* 42.11 (2014), e95–e95 (cit. on pp. 98, 99).

[63] M. D. Robinson, A. Kahraman, C. W. Law, H. Lindsay, M. Nowicka, L. M. Weber, and X. Zhou. “Statistical methods for detecting differentially methylated loci and regions”. *Frontiers in genetics* 5 (2014), p. 324 (cit. on pp. 98, 99).

[64] C.-H. Zhang and S. S. Zhang. “Confidence intervals for low dimensional parameters in high dimensional linear models”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76.1 (2014), pp. 217–242 (cit. on p. 25).

[65] R. Berk, L. Brown, A. Buja, K. Zhang, L. Zhao, et al. “Valid post-selection inference”. *The Annals of Statistics* 41.2 (2013), pp. 802–837 (cit. on pp. 22–25, 98).

[66] F. Frommlet and M. Bogdan. “Some optimality properties of FDR controlling rules under sparsity”. *Electron. J. Statist.* 7 (2013), pp. 1328–1368 (cit. on p. 39).

[67] T. D. Hocking, G. Schleiermacher, I. Janoueix-Lerosey, V. Boeva, J. Cappo, O. Delattre, F. Bach, and J.-P. Vert. “Learning smoothing models of copy number profiles using breakpoint annotations”. *BMC Bioinformatics* 14.1 (2013), p. 164 (cit. on pp. 59, 83).

[68] T. Hocking, G. Rigaill, J.-P. Vert, and F. Bach. “Learning sparse penalties for change-point detection using max margin interval regression”. *International Conference on Machine Learning*. 2013, pp. 172–180 (cit. on p. 82).

[69] L. R. Jager and J. T. Leek. “An estimate of the science-wise false discovery rate and application to the top medical literature”. *Biostatistics* 15.1 (2013), pp. 1–12 (cit. on p. 100).

[70] K. Li. “Bayesian multiple testing under sparsity for exponential distributions”. PhD thesis. University of Florida, 2013 (cit. on p. 39).

[71] C. Xu, D. Tao, and C. Xu. “A survey on multi-view learning”. *arXiv preprint arXiv:1304.5634* (2013) (cit. on p. 6).

[72] J. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. Liu, and B. Ren. “Topological domains in mammalian genomes identified by analysis of chromatin interactions”. *Nature* 485 (2012), pp. 376–380 (cit. on pp. 4, 89, 99).

[73] R. Killick, P. Fearnhead, and I. A. Eckley. “Optimal detection of changepoints with a linear computational cost”. *Journal of the American Statistical Association* 107.500 (2012), pp. 1590–1598 (cit. on p. 82).

[74] V. Michel, A. Gramfort, G. Varoquaux, E. Eger, C. Keribin, and B. Thirion. “A supervised clustering approach for fMRI-based inference of brain states”. *Pattern Recognition* 45.6 (2012), pp. 2041–2049 (cit. on p. 90).

[75] D. Mosén-Ansorena, A. Aransay, and N. Rodríguez-Ezpeleta. “Comparison of methods to detect copy number alterations in cancer using simulated and real genotyping data”. *BMC bioinformatics* 13.1 (2012), p. 192 (cit. on p. 83).

[76] The Cancer Genome Atlas (TCGA) Network. “Comprehensive molecular portraits of human breast tumours”. *Nature* 490.7418 (2012), pp. 61–70 (cit. on p. 66).

[77] K. Bleakley and J.-P. Vert. *The group fused Lasso for multiple change-point detection.* Tech. rep. http://hal.archives-ouvertes.fr/hal-00602121/en, June 2011 (cit. on pp. 83, 86).

[78] M. Bogdan, A. Chakrabarti, F. Frommlet, and J. K. Ghosh. “Asymptotic Bayes optimality under sparsity of some multiple testing procedures”. *The Annals of Statistics* 39.3 (2011), pp. 1551–1579 (cit. on pp. 35–37).

[79] H. Chen, H. Xing, and N. R. Zhang. “Estimation of parent specific DNA copy number in tumors using high-density genotyping arrays”. *PLoS Computational Biology* 7.1 (2011), e1001060 (cit. on p. 81).

[80] S. Delattre and E. Roquain. “On the false discovery proportion convergence under Gaussian equi-correlation”. *Statistics & Probability Letters* 81.1 (Jan. 2011), pp. 111–115 (cit. on p. 34).

[81] J. Goeman and A. Solari. “Multiple testing for exploratory research”. *Statistical Science* 26.4 (2011), pp. 584–597 (cit. on pp. 15, 16, 21, 22, 24, 42, 43, 45–47, 98).

[82] D. Hanahan and R. A. Weinberg. “Hallmarks of cancer: the next generation”. *cell* 144.5 (2011), pp. 646–674 (cit. on p. 3).

[83] K. D. Hansen, Z. Wu, R. A. Irizarry, and J. T. Leek. “Sequencing technology does not eliminate biological variability”. *Nature biotechnology* 29.7 (2011), p. 572 (cit. on p. 6).

[84] R. Louhimo and S. Hautaniemi. “CNAmet: an R package for integrating copy number, methylation and expression data”. *Bioinformatics* 27.6 (2011), p. 887 (cit. on p. 61).

[85] M. Rasmussen, M. Sundström, H. Kultima, J. Botling, P. Micke, H. Birgisson, B. Glimelius, A. Isaksson, et al. “Allele-specific copy number analysis of tumor samples with aneuploidy and tumor heterogeneity”. *Genome Biology* 12.10 (2011), R108 (cit. on p. 83).

[86] E. Roquain. “Type I error rate control in multiple testing: a survey with proofs”. *Journal de la Société Française de Statistique* 152.2 (2011), pp. 3–38 (cit. on pp. 11, 18).

[87] Z. Sun et al. “Integrated analysis of gene expression, CpG island methylation, and gene copy number in breast cancer cells by deep sequencing”. *PLoS One* 6.2 (2011), e17490 (cit. on p. 61).

[88] M. J. van der Laan and S. Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer Verlag, 2011 (cit. on p. 62).

[89] H. Wickham. "testthat: Get Started with Testing". *The R Journal* 3 (2011), pp. 5–10 (cit. on p. 101).

[90] J. Andrews et al. "Multi-Platform Whole-Genome Microarray Analyses Refine the Epigenetic Signature of Breast Cancer Metastasis with Gene Expression and Copy Number". *PLoS ONE* 5.1 (Jan. 2010), e8665 (cit. on p. 61).

[91] R. Bourgon, R. Gentleman, and W. Huber. "Independent filtering increases detection power for high-throughput experiments". *PNAS* (2010) (cit. on p. 3).

[92] S. X. Chen and Y.-L. Qin. "A two-sample test for high-dimensional data with applications to gene-set testing". *Ann. Stat.* 38.2 (Feb. 2010), pp. 808–835 (cit. on p. 74).

[93] J. J. Goeman and A. Solari. "The sequential rejection principle of familywise error control". *The Annals of Statistics* 38.6 (2010), pp. 3782–3810 (cit. on p. 16).

[94] Z. Harchaoui and C. Lévy-Leduc. "Multiple change-point estimation with a total variation penalty". *Journal of the American Statistical Association* 105.492 (2010), pp. 1480–1493 (cit. on pp. 83, 86).

[95] J.-P. Vert and K. Bleakley. "Fast detection of multiple change-points shared by many signals using group LARS". *Advances in Neural Information Processing Systems* 23 (2010), pp. 2343–2351 (cit. on p. 83).

[96] H. Finner, T. Dickhaus, and M. Roters. "On the false discovery rate and an asymptotically optimal rejection curve". *The Annals of Statistics* 37.2 (Apr. 2009), pp. 596–618 (cit. on p. 29).

[97] Y. Gavrilov, Y. Benjamini, and S. K. Sarkar. "An adaptive step-down procedure with proven FDR control under independence". *The Annals of Statistics* (2009), pp. 619–629 (cit. on p. 19).

[98] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2009 (cit. on p. 6).

[99] N. Meinshausen, L. Meier, and P. Bühlmann. "P-values for high-dimensional regression". *Journal of the American Statistical Association* 104.488 (2009), pp. 1671–1681 (cit. on p. 23).

[100] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009 (cit. on pp. 33, 64).

[101] Z. Chi and Z. Tan. "Positive false discovery proportions: intrinsic bounds and adaptive control". *Statistica Sinica* 18.3 (2008), pp. 837–860 (cit. on pp. 30, 31).

[102] L. Chin and J. W. Gray. "Translating insights from the cancer genome into clinical practice". *Nature* 452.7187 (Apr. 2008), pp. 553–563. ISSN: 0028-0836 (cit. on p. 3).

[103] C. Dalmasso et al. "Distinct genetic loci control plasma HIV-RNA and cellular HIV-DNA levels in HIV-1 infection: the ANRS Genome Wide Association 01 study". *PLoS ONE* 3.12 (2008), e3907 (cit. on pp. 4, 5, 93, 95).

[104] S. Gey and E. Lebarbier. *Using CART to Detect Multiple Change Points in the Mean for Large Sample*. Tech. rep. Statistics for Systems Biology research group, 2008 (cit. on pp. 83, 86).

[105] S. Guha, Y. Li, and D. Neuberg. "Bayesian hidden Markov modeling of array CGH data". *Journal of the American Statistical Association* 103.482 (2008), pp. 485–497 (cit. on p. 81).

[106] T. L. Lai, H. Xing, and N. Zhang. "Stochastic segmentation models for array-based comparative genomic hybridization data analysis". *Biostat* 9.2 (Apr. 2008), pp. 290–307 (cit. on p. 81).

[107] H. Leeb and B. M. Pötscher. "Can one estimate the unconditional distribution of post-model-selection estimators?" *Econometric Theory* 24.2 (2008), pp. 338–376 (cit. on p. 23).

[108] S. K. Sarkar et al. "On the Simes inequality and its generalization". *Beyond parametrics in interdisciplinary research: Festschrift in honor of Professor Pranab K. Sen.* Institute of Mathematical Statistics, 2008, pp. 231–242 (cit. on p. 18).

[109] M. S. Srivastava and M. Du. "A test for the mean vector with fewer observations than the dimension". *Journal of Multivariate Analysis* 99 (3 Mar. 2008), pp. 386–402. ISSN: 0047-259X (cit. on p. 74).

[110] J. Staaf, D. Lindgren, J. Vallon-Christersson, A. Isaksson, H. Göransson, G. Juliusson, R. Rosenquist, M. Höglund, A. Borg, and M. Ringnér. "Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome SNP arrays." *Genome biology* 9.9 (Jan. 2008), R136. ISSN: 1465-6914 (cit. on p. 80).

[111] The Cancer Genome Atlas (TGCA) research Network. "Comprehensive genomic characterization defines human glioblastoma genes and core pathways". *Nature* 455 (2008), pp. 1061–1068 (cit. on p. 64).

[112] R. Tibshirani and P. Wang. "Spatial smoothing and hot spot detection for CGH data using the fused lasso". *Biostatistics* 9.1 (2008), pp. 18–29 (cit. on p. 83).

[113] W. N. van Wieringen and M. A. van de Wiel. "Nonparametric Testing for DNA Copy Number Induced Differential mRNA Gene Expression". *Biometrics* 5.1 (Mar. 2008), pp. 19–29 (cit. on p. 61).

[114] W. B. Wu. "On false discovery control under dependence". *Ann. Statist.* 36.1 (2008), pp. 364–380 (cit. on p. 34).

[115] Z. Chi. "On the performance of FDR control: constraints and a partial solution". *The Annals of Statistics* 35.4 (2007), pp. 1409–1431 (cit. on pp. 28, 30, 31).

[116] F. S. Collins and A. D. Barker. "Mapping the cancer genome". *Scientific American* 296.3 (Mar. 2007), pp. 50–57 (cit. on p. 64).

[117] J. J. Goeman and P. Bühlmann. "Analyzing gene expression data in terms of gene sets: methodological issues". *Bioinformatics* 23.8 (Apr. 2007), pp. 980–987 (cit. on p. 70).

[118] Z. Harchaoui and O. Cappé. "Retrospective mutiple change-point estimation with kernels". *Proceedings of the 14th Workshop on Statistical Signal Processing (SSP'07).* Madison, WI, USA: IEEE, 2007, pp. 768–772 (cit. on p. 60).

[119] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. De Bakker, M. J. Daly, et al. "PLINK: a tool set for whole-genome association and population-based linkage analyses". *The American Journal of Human Genetics* 81.3 (2007), pp. 559–575 (cit. on p. 95).

[120] M. J. van der Laan, E. C. Polley, and A. E. Hubbard. "Super Learner". *Stat. Appl. Genet. Mol. Biol.* 6 (2007), Article 25 (cit. on p. 63).

[121] E. S. Venkatraman and A. B. Olshen. "A faster circular binary segmentation algorithm for the analysis of array CGH data". *Bioinformatics* 23.6 (Mar. 2007), pp. 657–663. ISSN: 1460-2059 (Electronic) (cit. on pp. 82, 86).

[122] F. Abramovich, Y. Benjamini, D. L. Donoho, and I. M. Johnstone. "Adapting to Unknown Sparsity by controlling the False Discovery Rate". *Ann. Statist.* 34.2 (2006), pp. 584–653 (cit. on pp. 35–37).

[123] Y. Benjamini, A. M. Krieger, and D. Yekutieli. "Adaptive linear step-up procedures that control the false discovery rate". *Biometrika* 93.3 (2006), pp. 491–507 (cit. on pp. 19, 30).

[124] D. L. Donoho and J. Jin. "Asymptotic minimaxity of false discovery rate thresholding for sparse exponential data". *Ann. Statist.* 34.6 (2006), pp. 2980–3018 (cit. on pp. 35–37).

[125] C. R. Genovese and L. Wasserman. "Exceedance Control of the False Discovery Proportion". *J. Amer. Statist. Assoc.* 101.476 (2006), pp. 1408–1417 (cit. on pp. 14, 21, 22, 45, 46).

[126] P. Kabaila and H. Leeb. "On the large-sample minimal coverage probability of confidence intervals after model selection". *Journal of the American Statistical Association* 101.474 (2006), pp. 619–629 (cit. on p. 23).

[127] H. Leeb, B. M. Pötscher, et al. "Can one estimate the conditional distribution of post-model-selection estimators?" *The Annals of Statistics* 34.5 (2006), pp. 2554–2591 (cit. on p. 23).

[128] E. L. Lehmann and J. P. Romano. *Testing statistical hypotheses*. Springer Science & Business Media, 2006 (cit. on p. 11).

[129] N. Meinshausen. "False discovery control for multiple tests of association under general dependence". *Scandinavian Journal of Statistics* 33 (2006), pp. 227–237 (cit. on pp. 22, 45).

[130] N. Meinshausen and J. Rice. "Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses". *Ann. Statist.* 34.1 (2006), pp. 373–393 (cit. on p. 45).

[131] M. J. van der Laan and D. Rubin. "Targeted maximum likelihood learning". *Int. J. Biostat.* 2 (2006), Article 11 (cit. on p. 58).

[132] Y. Benjamini and D. Yekutieli. "False discovery rate-adjusted multiple confidence intervals for selected parameters". *Journal of the American Statistical Association* 100.469 (2005), pp. 71–81 (cit. on p. 22).

[133] S. Chiaretti, X. Li, R. Gentleman, A. Vitale, K. S. Wang, F. Mandelli, R. Foa, and J. Ritz. "Gene expression profiles of B-lineage adult acute lymphocytic leukemia reveal genetic patterns that identify lineage derivation and distinct mechanisms of transformation". *Clinical cancer research* 11.20 (2005), pp. 7209–7219 (cit. on p. 3).

[134] J. P. Ioannidis. "Why most published research findings are false". *PLoS medicine* 2.8 (2005), e124 (cit. on p. 100).

[135] W. R. Lai, M. D. Johnson, R. Kucherlapati, and P. J. Park. "Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data." *Bioinformatics (Oxford, England)* 21.19 (Oct. 2005), pp. 3763–70. ISSN: 1367-4803 (cit. on p. 83).

[136] H. Leeb and B. M. Pötscher. "Model selection and inference: Facts and fiction". *Econometric Theory* 21.1 (2005), pp. 21–59 (cit. on p. 23).

[137] E. L. Lehmann and J. P. Romano. "Generalizations of the familywise error rate". *Ann. Statist* 33.3 (2005), pp. 1138–1154 (cit. on pp. 18, 20).

[138] N. Meinshausen and P. Bühlmann. "Lower bounds for the number of false null hypotheses for multiple testing of associations". *Biometrika* 92 (2005), pp. 893–907 (cit. on p. 22).

[139] F. Picard, S. Robin, M. Lavielle, C. Vaisse, and J.-J. Daudin. "A statistical approach for array-CGH data analysis". *BMC Bioinformatics* 6.27 (2005), pp. 1471–2105 (cit. on pp. 59, 82, 86).

[140] J. P. Romano and M. Wolf. "Exact and approximate stepdown methods for multiple hypothesis testing". *J. Amer. Statist. Assoc.* 100.469 (2005), pp. 94–108. ISSN: 0162-1459 (cit. on p. 17).

[141] J. P. Romano and M. Wolf. "Exact and approximate stepdown methods for multiple hypothesis testing". *Journal of the American Statistical Association* 100.469 (2005), pp. 94–108 (cit. on p. 49).

[142] A. Subramanian et al. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." *Proc Natl Acad Sci U S A* 102.43 (Oct. 2005), pp. 15545–15550 (cit. on p. 70).

[143] G. J. Székely and M. L. Rizzo. "Hierarchical clustering via joint between-within distances: extending Ward's minimum variance method". *Journal of Classification* 22.2 (2005), pp. 151–183 (cit. on p. 91).

[144] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. "Sparsity and smoothness via the fused lasso". *Journal of the Royal Statistical Society. Series B, Statistical Methodology* (2005), pp. 91–108 (cit. on p. 83).

[145] H. Willenbrock and J. Fridlyand. "A comparison study: applying segmentation to array-CGH data for downstream analyses". *Bioinformatics* 21.22 (Nov. 2005), pp. 4084–91 (cit. on p. 83).

[146] T. Beissbarth and T. P. Speed. "GOstat: find statistically overrepresented Gene Ontologies within a group of genes." *Bioinformatics* 20.9 (June 2004), pp. 1464–1465 (cit. on p. 70).

[147] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. "Least angle regression". *Annals of statistics* 32.2 (2004), pp. 407–451 (cit. on p. 83).

[148] J. Fridlyand, A. Snijders, D. Pinkel, D. G. Albertson, and A. N. Jain. "Hidden Markov models approach to the analysis of array CGH data". *Journal of Multivariate Analysis* 90.1 (July 2004), pp. 132–153. ISSN: 0047259X (cit. on p. 81).

[149] C. R. Genovese and L. Wasserman. "A Stochastic Process Approach to False Discovery Control". *The Annals of Statistics* 32.3 (2004), pp. 1035–1061 (cit. on pp. 19, 21, 22, 28, 32, 45).

[150] A. B. Olshen, E. S. Venkatraman, R. Lucito, and M. Wigler. "Circular binary segmentation for the analysis of array-based DNA copy number data". *Biostatistics* 5.4 (2004), pp. 557–572 (cit. on p. 82).

[151] J. D. Storey, J. E. Taylor, and D. O. Siegmund. "Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 66.1 (2004), pp. 187–205 (cit. on p. 19).

[152] M. J. van der Laan, S. Dudoit, and K. S. Pollard. "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 (2004), Art. 15, 27 pp. (Cit. on pp. 21, 45).

[153] X. Cui and G. A. Churchill. "Statistical tests for differential expression in cDNA microarray experiments". *Genome Biol* 4.4 (2003), p. 210 (cit. on pp. 21, 42).

[154] J. D. Storey. "The Positive False Discovery Rate: A Bayesian Interpretation and the q-value". *The Annals of Statistics* 31.6 (2003), pp. 2013–2035 (cit. on pp. 14, 28).

[155] J. S. Yedidia, W. T. Freeman, and Y. Weiss. "Exploring Artificial Intelligence in the New Millennium". Ed. by G. Lakemeyer and B. Nebel. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2003. Chap. Understanding Belief Propagation and Its Generalizations, pp. 239–269. ISBN: 1-55860-811-7 (cit. on p. 59).

[156] G. Casella and R. L. Berger. *Statistical inference*. Vol. 2. Duxbury Pacific Grove, CA, 2002 (cit. on p. 12).

[157] R. Edgar, M. Domrachev, and A. Lash. "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository". *Nucleic acids research* 30.1 (2002), pp. 207–210 (cit. on p. 84).

[158] S. B. Gabriel et al. "The structure of haplotype blocks in the human genome". *Science* 296.5576 (2002), pp. 2225–2229 (cit. on p. 89).

[159] J. R. Pollack et al. "Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors." *Proc Natl Acad Sci U S A* 99.20 (Oct. 2002), pp. 12963–12968 (cit. on p. 61).

[160] J. D. Storey. "A direct approach to false discovery rates". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.3 (2002), pp. 479–498 (cit. on p. 19).

[161] Y. Benjamini and D. Yekutieli. "The control of the false discovery rate in multiple testing under dependency". *The Annals of Statistics* 29.4 (2001), pp. 1165–1188 (cit. on p. 18).

[162] B. Efron, R. J. Tibshirani, J. D. Storey, and V. G. Tusher. "Empirical Bayes Analysis of a Microarray Experiment". *Journal of the American Statistical Association* 96.456 (Dec. 2001), pp. 1151–1160 (cit. on pp. 28, 34).

[163] R. Tibshirani, G. Walther, and T. Hastie. "Estimating the number of clusters in a data set via the gap statistic". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.2 (2001), pp. 411–423 (cit. on p. 95).

[164] D. Eppstein. "Fast hierarchical clustering and other applications of dynamic closest pairs". *Journal of Experimental Algorithmics (JEA)* 5 (2000), p. 1 (cit. on p. 92).

[165] D. Hanahan and R. A. Weinberg. "The hallmarks of cancer." *Cell* 100.1 (Jan. 2000), pp. 57–70 (cit. on p. 3).

[166] L. C. Evans. *Partial Differential Equations (Graduate Studies in Mathematics, V. 19) GSM/19*. American Mathematical Society, June 1998. ISBN: 0821807722 (cit. on p. 72).

[167] J. Fan and S.-k. Lin. "Test of Significance when data are curves". *J. Am. Statist. Assoc* 93 (1998), pp. 1007–1021 (cit. on p. 74).

[168] D. Pinkel et al. "High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays". *Nat. Genet.* 20 (1998), pp. 207–211 (cit. on p. 3).

[169] A. W. van der Vaart. *Asymptotic Statistics*. Vol. 3. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press, 1998 (cit. on pp. 28, 31, 32).

[170] A. W. van der Vaart. *Asymptotic statistics*. Vol. 3. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press, 1998, pp. xvi+443 (cit. on p. 62).

[171] Z. Bai and H. Saranadasa. "Effect of high dimension : by an example of a two sample problem". *Statistica Sinica* 6 (1996), pp. 311, 329 (cit. on pp. 71, 74).

[172] R. Tibshirani. "Regression shrinkage and selection via the lasso". *J. R. Stat. Soc. Ser. B Stat. Methodol.* 58.1 (1996), pp. 267–288 (cit. on pp. 5, 23).

[173] Y. Benjamini and Y. Hochberg. "Controlling the false discovery rate: A practical and powerful approach to multiple testing". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 57.1 (1995), pp. 289–300 (cit. on pp. 5, 18, 31, 34).

[174] P. H. Westfall and S. S. Young. *Resampling-Based Multiple Testing*. NY John Wiley & Sons, 1993 (cit. on p. 17).

[175] P. Massart. "The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality". *The Annals of Probability* (1990), pp. 1269–1283 (cit. on p. 53).

[176] R. R. Picard and K. N. Berk. "Data splitting". *The American Statistician* 44.2 (1990), pp. 140–147 (cit. on p. 23).

[177] G. Hommel. "A stagewise rejective multiple test procedure based on a modified Bonferroni test". *Biometrika* 75.2 (1988), pp. 383–386 (cit. on p. 17).

[178] E. Grimm. "CONISS: a fortran 77 program for stratigraphically constrained analysis by the method of incremental sum of squares". *Computers & Geosciences* 13.1 (1987), pp. 13–35 (cit. on p. 90).

[179] R. J. Simes. "An improved Bonferroni procedure for multiple tests of significance". *Biometrika* 73.3 (1986), pp. 751–754 (cit. on pp. 15, 29).

[180] D. B. Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, 1984 (cit. on p. 33).

[181] G. Hommel. "Tests of the overall hypothesis for arbitrary dependence structures". *Biometrical J.* 25.5 (1983), pp. 423–430. ISSN: 0323-3847 (cit. on p. 15).

[182] S. Holm. "A simple sequentially rejective multiple test procedure". *Scandinavian Journal of Statistics* 6 (1979), pp. 65–70 (cit. on p. 15).

[183] L. Lebart. "Programme d'agrégation avec contraintes". *Les Cahiers de l'Analyse des Données* 3.3 (1978), pp. 275–287 (cit. on p. 90).

[184] G. Schwarz et al. "Estimating the dimension of a model". *The annals of statistics* 6.2 (1978), pp. 461–464 (cit. on p. 82).

[185] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm". *Journal of the royal statistical society. Series B (methodological)* (1977), pp. 1–38 (cit. on p. 58).

[186] R. Marcus, P. Eric, and K. R. Gabriel. "On closed testing procedures with special reference to ordered analysis of variance". *Biometrika* 63.3 (1976), pp. 655–660 (cit. on p. 16).

[187] D. R. Cox. "A note on data-splitting for the evaluation of significance levels". *Biometrika* 62.2 (1975), pp. 441–444 (cit. on p. 23).

[188] A. Sen and M. Srivastava. "On tests for detecting change in mean". *The Annals of Statistics* 3.1 (1975), pp. 98–108 (cit. on p. 82).

[189] F. B. Baker. "Stability of two hierarchical grouping techniques case I: sensitivity to data errors". *Journal of the American Statistical Association* 69.346 (1974), pp. 440–445 (cit. on p. 94).

[190] H. Akaike. "Information Theory and an Extension of the Maximum Likelihood Principle". *Selected Papers of Hirotugu Akaike*. New York, NY: Springer New York, 1973, pp. 199–213 (cit. on p. 82).

[191] Z. Šidák. "Rectangular confidence regions for the means of multivariate normal distributions". *Journal of the American Statistical Association* 62.318 (1967), pp. 626–633 (cit. on p. 16).

[192] J. W. J. Williams. "Algorithm 232 - heapsort". *Communications of the ACM* 7.6 (1964). Ed. by G. Forsythe, pp. 347–348 (cit. on p. 93).

[193] J. H. Ward. "Hierarchical grouping to optimize an objective function". *Journal of the American Statistical Association* 58.301 (1963), pp. 236–244 (cit. on pp. 90, 91).

[194] H. Scheffé. *The analysis of variance*. Wiley, New York, 1959 (cit. on p. 24).

[195] A. Dvoretzky, J. Kiefer, and J. Wolfowitz. "Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator". *The Annals of Mathematical Statistics* (1956), pp. 642–669 (cit. on p. 53).

[196] H. Scheffé. "A method for judging all contrasts in the analysis of variance". *Biometrika* 40.1-2 (1953), pp. 87–110 (cit. on pp. 21, 23, 24).

[197] N. Aronszajn. "Theory of reproducing kernels". *Transactions of the American Mathematical Society* 68.3 (1950), pp. 337–337 (cit. on p. 91).