Università degli Studi di Firenze

Dipartimento di Statistica "G. Parenti"

Control of the false discovery rate with frequentist p-values in Microarray data analysis

Stefano Cabras

Dottorato di Statistica Applicata XVI Ciclo

April 2004

Supervisor: Prof. L. Fattorini
Assessor: Prof. F. M. Stefanini

Abstract

Microarrays are emerging as a powerful and cost-effective tool for large scale analysis of gene expression (Brown and Botstein, 1999). These experiments are typically done in a *case-control study framework* where thousand of genes are simultaneously compared in order to assess which, among them, are differentially expressed (*Discoveries*) switching from case to control target samples. This approach typically involves *Multiple Hypothesis Testing* (MHT) procedures rather than *classical Multiple Comparisons* (MCPs) procedures because, in an exploratory data analysis on thousands of genes, the researcher is mainly interested in controlling the *False Discovery Rate* (FDR) rather than the probability of making one or more false discoveries ($FWER$). The control of FDR was initially introduced in the pioneer work of Benjamini and Hochberg (1995), further developed by Storey (2002) and generalized by Genovese and Wasserman (2002) that introduced the control of False Non-rejections Rate (FNR). Available literature makes use of *frequentist p-values* as a measure of evidence from each single hypothesis. By using the *Frequentist Principle* (Neymann, 1977), we have that a p-value is frequentist if it is uniformly distributed under the null hypothesis that the gene is not differentially expressed. This is not always the case, because the hypotheses under test are typically *composite null hypothesis* rather than simple null hypothesis. Moreover composite null hypotheses often involve *nuisance parameters* to be eliminated in order to calculate the p-value. This problem has been recently addressed by Bayarri and Berger (2000) in an objective Bayesian framework that makes use of non-informative priors, as the case in an exploratory data analysis. They introduced the *conditional predictive p-value* (p_{cpred}) and the *partial posterior predictive p-value* (p_{ppost}). Under fairly general conditions, the p_{cpred} is uniformly distributed under the null hypothesis (no matter the number of experimental replications), while the p_{ppost} only asymptotically, but with better approximation to the uniform distribution than other alternative p-values, such as, Plug-in p-values (p_{plug}) and Posterior Predictive p-values (p_{post}). The aim of this work is to extend the use of FDR controlling procedures to models that involves nuisance parameters and when no sufficient statistics are available. We do this by using the p_{cpred} and p_{ppost}. We found that they allow to control the FDR, by using the recent available techniques, as if we were dealing with simple null hypothesis. In the end they allow more power, in detecting differentially expressed genes, than other p-values for composite null models. Two models are considered here: *i*) the *Gamma model* with the shape parameter that represent the common variation coefficient in spotted cDNA microarrays and *ii*) the *Normal model* in order to match the theoretical results coming from p_{cpred} with the well known t-tests. The gamma model has been applied to three public data sets in order to make comparisons among different notions of p-values. The methodology here proposed is general and the results can be extended to more complicates models than those showed in this thesis. The methodology also can be applied to every other experimental situations where the control of FDR is needed.

Contents

List of Figures

List of Tables

Notation and Definitions

Symbol	Definition

	Not-Reject	*Reject*	Total
Null True	V'	V	m_0
Alt. True	O	L	m_1
Total	W	R	m

Outcome from m hypotheses (genes) under test

Γ_α	Rejection region with Type I error α
$FWER = \Pr(V \geq 1)$	Family Wise Error Rate
$FDR = \mathbf{E}(V/R \vert R > 0)\Pr(R > 0)$	False Discovery Rate
$pFDR = \mathbf{E}(V/R \vert R > 0)$	*Positive* False Discovery Rate
q-value$(\cdot) = \inf_{t \in \Gamma_\alpha} pFDR(\Gamma_\alpha)$ and $\widehat{q}(\cdot)$	The q-value and its estimate
$\widehat{pFDR}_\lambda(\cdot)$	Estimates of $pFDR$ for a fixed rejection region
$\mathbf{X}_i$ and $\mathbf{Y}_i, i = 1, \ldots, m$	m gene abundances in *case* ($\mathbf{X}$) and *control* ($\mathbf{Y}$)
$\mathbf{x}_i$ and $\mathbf{y}_i, i = 1, \ldots, m$	Samples of gene abundance of sizes n_X and n_Y
$T(\mathbf{X}, \mathbf{Y})$ and $t_{obs} = t(\mathbf{x}, \mathbf{y})$	Test statistic and its observed value
$T_1, T_2, \ldots, T_m$	Test statistics for m tests
$t_1 = t(\mathbf{x}_1, \mathbf{y}_1), \ldots, t_m = t(\mathbf{x}_m, \mathbf{y}_m)$	Observed statistics for $T_1, T_2, ..., T_m$
$P_1, P_2, ..., P_m$ and $p_1, p_2, ..., p_m$	Random and observed p-values for m tests
$N(\mu, \sigma^2)$	Normal distribution (mean μ and variance σ^2)
$\Phi(x)$	Standard Normal distribution
$N^{-1}(1-\alpha, \mu, \sigma^2)$	$1-\alpha$ quantile of a Normal distribution
$\Gamma(x)$ and $\psi^{(k)}(x) = \frac{\partial^k}{\partial a^k} \log \Gamma(x)$	Gamma and Polygamma functions of order k
$Gamma(a, \theta)$	Gamma distribution with shape a and scale θ
$Ga^{-1}(a, \theta)$	Inverse Gamma distribution with mean $a\theta$
$\boldsymbol{\theta} \in \boldsymbol{\Theta}$	Vector of parameters and its prior distribution
$\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_0$	Restriction of $\boldsymbol{\Theta}$ induced by the null model
$f(\mathbf{x}; \boldsymbol{\theta}_0)$	Density for the *null model*
$h(t)$	Marginal density of $T(\mathbf{X}, \mathbf{Y})$
$h\left(t; \widehat{\boldsymbol{\theta}}\right)$	*Plug-in* distribution
$\pi(\boldsymbol{\theta})$	*Prior* distribution of $\boldsymbol{\theta} \in \boldsymbol{\Theta}$
$\pi(\boldsymbol{\theta} \vert \mathbf{x}) \propto f(\mathbf{x} \vert \boldsymbol{\theta})\pi(\boldsymbol{\theta})$	*Posterior* distribution
$\pi(\boldsymbol{\theta} \vert \mathbf{x} \backslash t_{obs}) \propto f(\mathbf{x} \vert t_{obs}; \boldsymbol{\theta})\pi(\boldsymbol{\theta})$	*Partial posterior* distribution
$X \nsim F(x)$	X is not distributed with $F(x)$

Chapter 1

Introduction

This chapter contains the introduction to this thesis's work. Here it will be introduced the terminology and part of the notation used in the rest of the thesis and it will be provided the motivations for this work. We first start by stating the goal of the microarray experiment and the contribution of this work in helping the researcher while analyzing the experimental outcome. Then, we provide a brief description of microarray experiment with the two commons techniques: cDNA arrays and oligonucleotide arrays. Finally we underline the sources of experimental variations and statistical issues in data analysis.

The literature on microarray as recently experienced a great explosion. This can be monitored, for example, from microarray journal watch web sites such as the "Y. F. Leung's Functional Genomic page" at `http://genomicshome.com` or the "Microarray and Data Analysis" at `http://www.nslij-genetics.org/microarray/`.

1.1 Goal of a microarray experiment and thesis contribution

Microarray experiments are mainly designed to characterize the genetic profile of cells under different experimental conditions. This typically involves statistical testing on thousand of genes in order to asses which are differentially expressed in two or more experimental conditions. Multiple Hypothesis Testing (MHT) techniques address this issue, but they require calibrated measures of evidence for each hypothesis. This can be achieved by using frequentist p-values.

1.1.1 Microarrays experiments are performed to understand the genome through the messenger Ribonucleic Acid

Proteins are structural components of the cells and tissues and perform many key functions in biological systems. The production of proteins is controlled by genes, which are coded in Deoxyribonucleic Acid (DNA). DNA molecules are common to all cells of an organism. Protein production from genes involves two principal stages, known as transcription and translation, as illustrated schematic on Figure 1.1. During transcription, a single strand of messenger Ribonucleic Acid (mRNA) is used as a

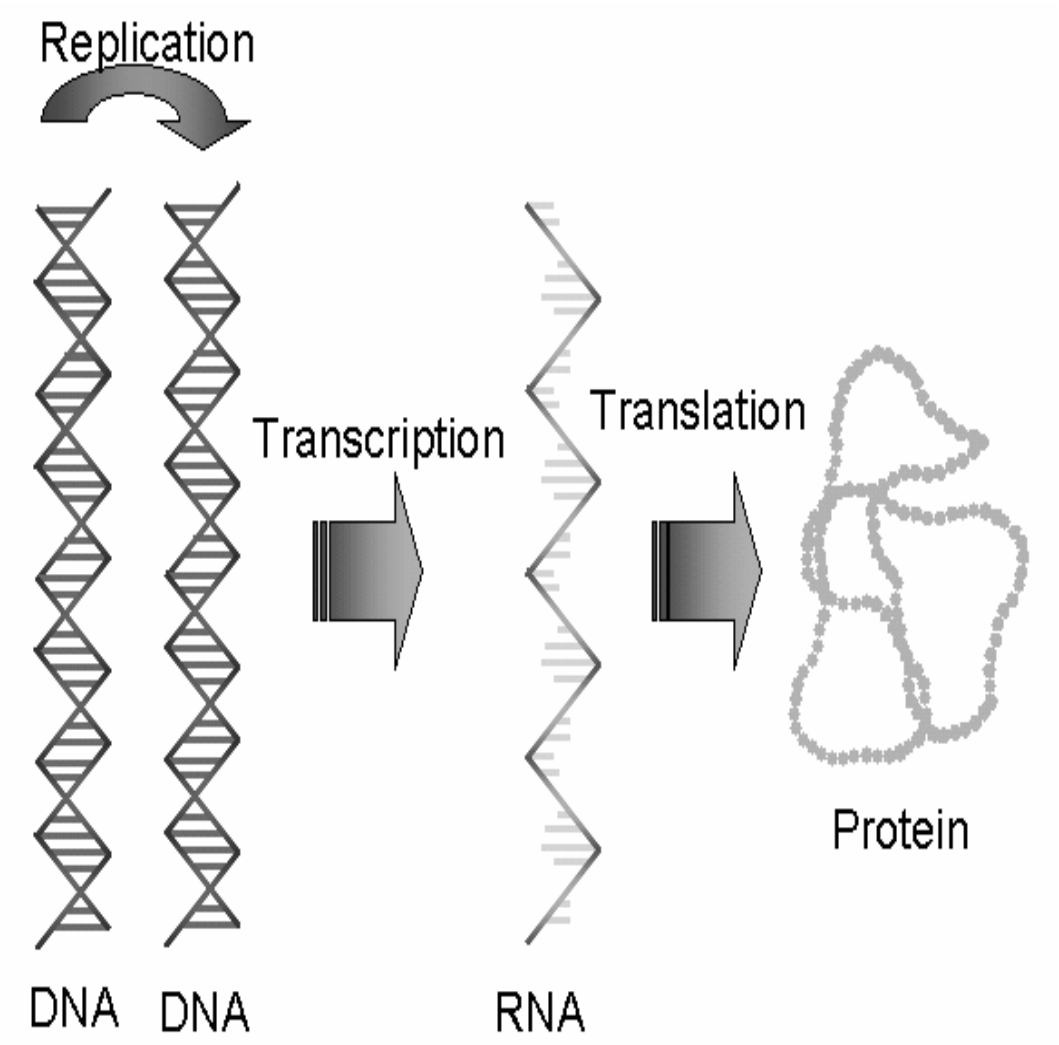

Figure 1.1: Central dogma of molecular biology (source: `www.swbic.org`).

template to assemble a chain of amino acids to form the protein. Gene expression investigations study the amount of transcribed mRNA in biological system. Although most proteins undergo modification after translation and before becoming functional, most changes in the state of a cell are related to changes in mRNA levels for some genes, making the transcription worthy of systematic measurement. Basic biochemistry and molecular biology textbooks, such as those of Bolsover *et al.* (1997) and Garret and Grisham (2002), provide background on gene expression and its biological significance.

1.1.2 Many genes are compared simultaneously and this involve Multiple Hypothesis Testing (MHT)

The outcome of a microarray experiment are quantities which are supposed to represent the genes abundances in mRNA from a tissue. In the simplest experiment we compare the gene expression of a tissue under two experimental conditions, regarded sometimes as biological populations. Biologists expect that many genes are differentially expressed at different levels. Therefore, if we measure the abundance of m genes we have to perform m individual tests in order to *discover* which of them are differently expressed in the two conditions. This involves the simultaneous test of m hypotheses, where the *null hypothesis* is "the gene i is not differentially expressed in the two biological populations" ($H_i = 0$) against the alternative "It is differentially expressed" ($H_i = 1$): either overexpressed or underexpressed. The hypothesis testing is conducted under the limitation that the available number of replications, n, is very small. The order of magnitude is $m \simeq 10000$ genes with $n \simeq 3$ in a small scale study and $n \simeq 100$ in largest studies. Comparing gene expression across two conditions for a single gene is an instance of the most classical statistical questions: the *two-sample comparison*. Estimating and testing in this case are very well developed. In genomic applications, however, there is an increasing consensus on the inefficiency due to a gene-by-gene analysis, and consistency is gained by considering the ensemble of gene expression measures at once. This occurs for at least two reasons: first, genes measured on the same array type in the same laboratory are all affected by a number of common sources of noise; second, changes in expression are all part of the same biological mechanism, and their magnitudes, although different, are not completely unrelated. This requires to take into account this multiplicity of tests by considering the dependency of the hypotheses under test.

1.1.3 MHT procedures need frequentist p-values

The proportion of false discoveries among all discoveries, the *False Discovery Rate* (FDR) is the main quantity considered in this thesis in order to decide whether or not the reject a subset of null hypotheses. The control of this quantity was originally introduced by Benjamini and Hochberg (1995) and further developed by Storey (2002, 2003 and 2004). Early ideas on the FDR were introduced by Seeger in 1968, but the popularity of FDR was mainly due to Benjamini and Hochberg (1995). It may seems somewhat arbitrary to focus the control only on this quantity, because in general we

should control all the errors produced when we wrongly reject H_i or we fail to reject H_i. The *False Non-rejection Rate* (FNR), and the joint control of the FDR and FNR, was recently introduced by Genovese and Wasserman (2002). Although the joint control of FDR and FNR seems quite appealing, most of the techniques developed until now, concern the control of the FDR and so we will mainly concentrated on them.

The control of FDR (and also of FNR) relies on our ability to construct a rejection region for each hypothesis $H_i, i = 1, 2, ..., m$ and then can calculate their Type I Error. Let T be the a test statistic. In this thesis we show that to control the FDR, we must be able to find a suitable set of nested rejection regions Γ_α. such that

$$\Pr\left(T_i \in \Gamma_\alpha | H_i = 0\right) = \alpha$$

The nested property means that $\alpha' \leq \alpha$ implies $\Gamma_{\alpha'} \subseteq \Gamma_\alpha$. Each hypothesis may have a different Γ_α, so it is convenient to consider a set which is equal for all hypotheses. In particular we consider the set of nested rejection regions defined by the p-values on each H_i. The p-value for an observed value t_i is defined as (Lehmann, 1986):

$$p\text{-value}\left(t\right) = \inf_{\{\Gamma_\alpha : t_i \in \Gamma_\alpha\}} \Pr\left(T_i \in \Gamma_\alpha | H_i = 0\right). \tag{1.1}$$

Therefore, since the set of rejected regions is nested, it can be seen that

$$t_i \in \Gamma_\alpha \Leftrightarrow p\text{-value}\left(t_i\right) = p \leq \alpha.$$

When this is true, a p-values is said to be a *frequentist p-value*, that is

$$\Pr\left\{P \leq \alpha\right\} = \alpha, \tag{1.2}$$

where P is the p-value and we regarded it as a random variable uniformly distributed on $(0, 1)$. The (1.2) simply restate the frequentist principle of Neyman (1977): "the long run error is less or equals than the declared one.". Only frequentist p-value can also be thought of as the level of the test at which the hypothesis H_i would just be rejected. Using frequentist p-values we obtain the following rejection region:

$$\Gamma_\alpha \equiv \left\{\left(p, 1\right), \alpha = p\right\}. \tag{1.3}$$

Unfortunately, in a parametric framework, the (1.2) and the (1.3) are always true if H_i is a *precise null hypothesis* (or *simple* null hypothesis), and on the other hands (1.2) and (1.3) are generally false when H_i is a *composite null hypothesis*. Hence when the null hypothesis is composite and we want to control the FDR, we found that it is necessary to make use of p-values uniformly distributed for a fixed sample size n.

In an objective Bayesian framework, that makes use of *non-informative priors* (often *improper priors*), Bayarri and Berger (2000) provide p-values that are uniformly distributed under the null hypothesis. The object of this thesis is to investigate the control of FDR using different notions of p-values. We show that only frequentist p-values are legitimate to control the FDR and, in regard to this task, the p-values proposed by Bayarri and Berger (2000) outperform other commonly used p-values. This thesis work allows to properly extend the procedure that control the FDR in

parametric composite null hypothesis (or models). Until now, the very recent literature on FDR does not seem to be interested on this extension.

Using p-values we are mainly focused on *model criticism* rather than classical hypothesis testing, because when considering only p-values in hypothesis testing we do not consider any other alternative model. The use of p-values in model criticism is very questionable from a *pure Bayesian* point of view. Here we agree with the usual criticisms that: *i*) models can only be compared and cannot be assessed singularly; *ii*) the integration over the sample space after knowing the data introduce additional noise, and so forth. These criticism are due to the violation of the likelihood principle in the inferential process, because we are considering values greater than t_i which have been never observed. However, here we are interested only in constructing a rejection region for each single hypothesis and assign its Type I error. Obviously, this has to be done beforehand with respect to the experiment, for any arbitrary choice of a rejection region. This goal can be conveniently achieved by using frequentist p-values that suggest a rejection region and provide the corresponding Type I error. The choice of using rejection regions based on p-values becomes then a merely fact of convenience, because recent literature on p-values makes possible to build a one-to-one relationship between the rejection region $\Gamma \equiv (p, 1)$ and the Type I error $\alpha = p$ under fairly general conditions. Other measures of evidence useful to this purpose are welcome.

1.2 The microarray experiment

Microarrays quantify gene expression by measuring the hybridization, or matching, of DNA immobilized on a small glass plastic, or nylon matrix to DNA from the sample under study. As a crude approximation, we can think to a separate experiment taking place in each of many individual spots, arranged in a regular lattice pattern on a matrix, whence the name array. Arrays may have hundred of thousands of spots. We call the spotted sequences "genes", whether or not they are actual genes, ESTs (expression sequence tags) or cDNA sequences from other sources.

Such ability to measure simultaneously a large fraction of an expression (the expressed part of the genome) opens the door to the investigation of large scale interactions among the genes, the discovery of the role of a vast number of genes whose function is not adequately understood, and the characterization of how metabolic pathways are changed under various conditions. Duggan *et al.* (1999) review the use of microarrays in genomic investigations and the impressive spectrum of biological applications.

This work is focused on microarrays viewed as a tool for screening genes before further investigations take place. These second investigations are usually performed with other techniques then microarray such as: serial analysis of gene expression, cDNA library sequencing, differential display, cDNA subtraction and multiplex quantile RT-PCR. For more details and history on these techniques see for example Zweiger (2001).

1.2.1 The two most important techniques: cDNA arrays and oligonucleotide arrays

There are several microarray technologies, but we briefly describe two prevalent approaches: cDNA arrays and oligonucleotide arrays. Although they both exploit hybridization, they differ in how DNA sequences are laid on the array and in the length of these sequences. For a brief overview of current microarray technologies see Southern (2001) or Hardiman (2002).

The cDNA arrays

In spotted DNA arrays, mRNA from two different biological samples is reverse-transcribed into complementary DNA, (whence the acronym cDNA), labelled with dyes of two different colors (Cy3 and Cy5), and then hybridized to DNA sequences. Each of these sequences is spotted on a small region, or spot, on a glass slide. After hybridization, a laser scanner measures the dye fluorescence at the two wavelength on a fine grid of pixels. A high florescence indicates high amounts of hybridized cDNA, which in turn indicates high gene expression in the sample. A spot typically consists of a number of pixels. Image analysis algorithms either assign pixels to a spot (*foreground*) and produce summaries of fluorescence in the surrounding unspotted areas (*background*). More technical details on cDNA arrays are beyond the scope of this thesis and can be found in Schena (2000) and at microarrays web sites such as `http://www.microarrays.org`.

For each location on the array, a typical output consists of at least four quantities: two pairs foreground-background one for each dye. Sometimes these are accompanied by measures of the quality of the spot, to flag technical problems, or by measures of variability at pixel level. It is conventional to refer to the two colors Cy3 and Cy5 as red and green color, denoted respectively with R and G. The use of two channels allows for measurement of relative gene expression across two sources of cDNA, controlling the amount of spotted DNA, which can be variable, as well as other experimental variation. This led to emphasize the ratio R/G at each spot.

The high density oligonucleotide arrays

The second common approach involves the use of high-density oligonucleotide arrays. This is an area of active technological development. As we write, the most widely used oligonucleotide array type is the Affymetrix GeneChip (henceforth abbreviated by Affy). In Affy arrays, expression of each gene is measured by computing hybridization of the sample mRNA to a set of probes, composed of 11-20 pairs of oligonucleotide, each of length 25 base pairs. The first type of probe in each pair is known as perfect match (PM) and is taken from the gene sequence. The second type is known as mismatch (MM) because it is created by changing the middle (13th) base of the PM sequence to reduce the rate of specific binding of mRNA for that gene. The goal of MM is controlling for the experimental variation which is related to non specific binding of mRNA from other part of the genome.

An mRNA sample is prepared, labelled with one fluorescent dye, and hybridized to

an array. Unlike the two-channel array, a single sample is hybridized on a given array. Arrays are then scanned, and images are produced and analyzed to obtain a fluorescence intensity value for each probe, measuring hybridization for the corresponding oligonucleotide. For each gene, or probe set, the typical output consists of two vectors of intensity readings, one for PMs and one for MMs. Oligonucleotide arrays are discussed by Lockhart *et al.* (1996). Details of Affy-arrays can be found in Affymetrix (1999).

1.2.2 Raw data from the experiment consist in digitalized images

Using microarray technologies, it is essential to visual inspect the array in order to diagnose the presence of possible artifacts. For most visualizations, logarithmic transformation of the data are recommended because of the marked differences of expression data values and because the scanning of arrays results in optical or background noise affecting pixel intensities.

In cDNA arrays, image processing will produce an absolute expression measure and a background measurement for each spot or cell. On the contrary in high density oligonucleotide arrays have minimal space between the segments of the array where probes are attached, therefore background information is difficult to obtain and it is not commonly used. High-density oligonucleotide arrays pose the challenge of summarizing data from a probe set into a single measure, which estimates the level of expression of the gene of interest. Affy software provides a default approach for this step by returning the *AD* quantity, that is the difference between the PM fluorescence intensity and the MM florescence intensity. Two important reasons suggest that both probe-level data, PM and MM, should be considered as an integral part in Affy data analysis: the first reason is that visualization of probe-level data can help to identify artifacts on Affy chips. The second reason is that there is evidence that alternative summarizations, to the defaults currently implemented by Affy, may improve the ability to detect biological signal.

1.2.3 Images provide quantities on genes abundance which have to be efficiently stored in relational databases

In microarray experiments, the quantified values are contained in the image files produced by the scanner. The pixel intensity, stored in these files, can be thought as raw data. Image analysis tools are then used to provide numerical quantification for the quantities of interest: foreground-background intensities in cDNA arrays and PM-MM in Affy chips. The process of draw numerical quantities from scanned microarray images can be separated into three tasks that we list here in order to convey the complexity of the experiment:

i) *Addressing.* The basic structure of the microarray image is determined by the arrayer and is therefore known. That is, it is known that there are some number of rows and column of spots. The addressing process consists in matching an idealized model of the array with the scanned image. A number of parameters

need to be estimated. These parameters include: separation between rows and columns of grids, individual translation of grids, etc... It is desirable, for the addressing procedure, to be as reliable as possible to ensure accuracy of the whole measurement process. Reliability of the procedure can be increased allowing user intervention, but this makes the process very slow.

ii) *Segmentation* is the process of partitioning the image into different regions: *foreground* and *background*. The foreground region is the region of the spot, where the genetic materials is supposed to lay. The background region is the region where it is supposed to be no genetic materials. Any segmentation method produces a *spot mask*, which consists of the set of foreground pixels for a given spot. The procedures are usually classified in two categories: adaptive and non adaptive. The former perform generally better as shown in Yang *et al.* (2002). Adaptive procedures starts from a point in the image and try to learn the shape of the spots.

iii) *Information extraction*. After detecting the location, size and shape of each spot using one of the previous methods, foreground is calculated together with background intensities and, possibly, spot quality measures. Most microarray analysis packages define the foreground intensity as the mean or median of pixel values within the segmented spot mask. More possibilities exist in the choice of background calculation method. Common approaches include taking the median of values in selected regions surrounding the spot mask.

Accounting for the details of these techniques constitutes an additional insight which is useful in the model selection process, but the comparison between different techniques is behind the scope of the thesis. Recent reviews of image analysis techniques and software can be found in Brown *et al.* (2001), Yang *et al.* (2002) and Jain *et al.* (2002).

The information associated with microarray experiment has four important components:

i) a table of numbers representing absolute or relative expression values at gene or spot level. The emerging standard is to have rows represent genes and columns represent samples/arrays;

ii) a table of covariates associated with the samples, which may include information on the samples phenotypes (e.g., cancer type) as well as design variable for controlled experiments (e.g., drug treatment);

iii) detailed description of genes represented and of phenotypic variables;

iv) information about the experiment itself, which generally consists in an identification number in a public database, the experimental protocols, normalization information, and so forth.

This resulting complexity creates substantial computational challenges, therefore for researchers it is crucial to store the flood of information efficiently. Relational databases make possible to efficiently store and access to complex data sets and facilitate combining information from multiple microarray experiments. Gardiner-Garden (2001) provide a survey and comparative analysis of microarray databases.

Various open source database servers are freely available for custom database development. However, proliferation of microarray databases, the growing appreciation for both the importance of analysis across experiments and the need for well-documented repositories, have stimulated work toward the development of standards. The Microarray Gene Expression Database (MGED) group (`http://www.mged.org`) is a movement to promote the adoption of standards in microarray experiments and data. MGED developed requirements for Minimum Information about a Microarray Experiment (MIAME) required to interpret and verify results. Scientific journals are beginning to require compliance with MIAME standards for data made available as supplementary information in microarray-oriented papers. All data sets analyzed in this thesis are MIAME complaint.

Proper storage and access to data is critical but it is not the only relevant aspect in microarray analysis, because a database needs also to interact efficiently with statistical analysis languages and environments. With regard to this computational aspect, microarray analysis tools have been developed following three approaches: *i*) as parts of comprehensive databases, *ii*) as stand-alone packages, and *iii*) as libraries within well-established programming and analysis languages/environments, such as S-PLUS, *R*, SAS, Excel and Matlab. The first approach addresses the storage/analysis interface but often requires new development for statistical analysis software. The second approach leaves the interface to the user. The one joined in this thesis is the third, because it allows the exploitation of standard tools for data manipulation and statistical analysis. With this approach tailored analysis and methods for special needs are easily obtained.

Extensive support for microarray analysis is available as part of the *Bioconductor Project* (`http://www.bioconductor.org`) whose relies on the *R* language. The *R project* provides the main support to almost all calculations in this thesis. The original code, both in *R* and *C*, which have been used to produce all the numerical results in this work is described in the Appendix and also available at the web site `http://www.stat.cmu.edu/~scabras/tesi/`.

1.3 The experimental variability

The most important step in a microarray experiment, like any other experiment, is the identification of biological questions of interest. The degree of specificity in the question can range from a precisely defined hypothesis about two groups. For example, we can study a specific question such as the effect on a particular organ of a toxic compound in a population of genetically identical laboratory animals, or we can study a much more broad questions, such as a novel hypothesis about yet-unidentified

subtypes of lung cancer. The chosen question gives rise to a design of the experiment, which in turn leads to data. In view of the multiple source of errors described below, a substantial effort is necessary to extract from the data a reliable signal, (that is, a reliable representation of the gene expression under the various experimental conditions). Gene expression measures are subsequently used to address the biological questions. Because of the high potential for false positive findings, there is a wide agrement that results should be validated using alternative assays, such as RNA blotting, or RT-PCR.

A broad spectrum of biological investigations is made possible by microarray technologies. On the one end of this spectrum, we have highly specific comparisons, for example, between treated and control groups of genetically identical mice. In such applications, the signal-to-noise ratios are relatively favorable and statistical questions, albeit hard to addresses, are generally better defined. At the opposite end, we have what we could describe as "*genome biometry*", that is, the description of the genetic variability in different biological populations. In this latter, the signal-to-noise ratios are less favorable and the more exploratory is the nature of biological investigations the more the statistical questions are less well defined. It is widely accepted that, the statistical tools currently used in microarray data analysis are much more useful to support data exploration rather than completely automate it.

Ideally the three different stages of the analysis: signal extraction, data analysis and validation stages should be integrated and uncertainties propagated across stages. Because of the complexity and the novelty of the tasks corresponding to the three stages, only preliminary progress have been made toward this integration.

We mainly concern on data analysis stage. The typical approach is to first perform what is called *normalization*, that is trying to model and then remove sources of noise due to experimental artifacts. These sources of noise, if not appropriately removed, translate into a BIAS in the data analysis which is performed on normalized gene expression measures. There is a gross and prevalent strategy for dealing with uncertainty in individual spot measurements, which is to exclude spots for which the uncertainty is considered too high to be acceptable. This happen, for example, when the amount of background intensity from a spot is higher than the amount of foreground intensity.

Gene expression microarrays are powerful, but *technical variability* arising throughout the measurement process can obscure the biological signals of interest. It is useful to classify source of error into five phases of data acquisition: *i*) microarray manufacturing, *ii*) preparation of mRNA from biological samples, *iii*) hybridization, *iv*) scanning and *v*) imaging. Each of these phases can introduce an amount of artifactually variation and/or bias that makes problematic the estimation of expression levels as well as the comparison of expression changes between arrays. We list some examples to convey a sense for the multiplicity of source of errors and the importance of quality control.

1.3.1 The relevant sources of variability across different experimental phases

Phase *i*). Manufacturing errors are specific to the technology. In cDNA microarrays they arise: in the amplification, purification and concentration of DNA clones for spotting; in the amount of material spotted; in the ability of spotted material to bind to the array, and in the shape of the deposited spot. Systematic variation can be determined by microscopic defects in the print tip of the robotic equipment used for spotting.

Phase *ii*). During the preparation of the samples, sources of variability depend on the protocol and the platform used. Important examples include labelling procedure, RNA extraction and amplification. In cDNA arrays, dye biases can arise from different physical properties of the dyes or from differential ability of the dyes to incorporate into the samples.

Phase *iii*). During hybridization, variability arises from ambient conditions such as temperature and humidity, from edge effects (that is, effects seen only at the gene spotted near the edges of the array), from slight inhomogeneity of the hybridization solution, from extraneous molecules or dust binding to the array, from cross-hybridization of molecules with high sequence identity, and from washing of nonhybridized materials from the array.

Phase *iv*). During scanning, natural florescence and binding of genetic material to the array in unspotted regions can introduce a nontrivial, spatially varying background noise. Scanning requires separating the fluorescent label from the biological material and capturing it with sensors; both phases involve randomness, and rescanned slides usually give slight different results. Scanning intensity is an important factor, as higher intensity improves the quality of the signal but increase the risk of saturation caused by ceiling occurring when a channel reaches maximum intensity, (which is 2^{16} on a 16 bit channel).

Phase *v*). In the imaging step, some technologies require human intervention for initialization of the image algorithms or the alignment of the image to a grid. Different imaging algorithms and options within these algorithms also typically lead to varying fluorescence quantifications.

Although many of these errors are relative small, the compound of their effects can be significant. As a result, we can generally expect variation in the expression of a given gene across different hybridizations using the same RNA sample. Also, when a sequence is spotted in multiple locations on an array, there is usually variation in the amount of hybridization measured across location. In cDNA arrays, many sources of noise can be quantified in the aggregate by a self-versus-self hybridization in which two subsamples, from the same pool of RNA, are labelled with different dyes and then hybridized on the same array.

The sources of variation described so far arise from limitations of the current techniques and will be referred to as technological. In most microarray experiments we will also need to consider the usual source of variation that arise in sampling from biological populations and treated individuals. For example, in comparing the gene

expression of tumor tissue to normal tissue in a patient, we need to consider that the overall expression profile of the tissue could be different if another patient would be analyzed. It might differ if a different portion of the tissue from the patient under consideration had been sampled. In the end, microarray experiments resemble to other investigation of biological variation.

1.3.2 Experimental replications and sample size considerations

It is now becoming widely accepted that microarray experiments need to be replicated due to the sources of variation.

We can distinguish two broad types of replicates experiments: *biological replicates*, which refer broadly to analysis of RNA of the same type from different subjects (for example, muscle tissue treated with the same drug in different mice); *technical replicates* refer to multiple-array analysis performed using the same RNA (for example, multiple samples from the same tissue). Depending on the experimental setting, one or both of these types of replicates need to be considered. In controlled experiments, replicates are generally used to increase the reliability of conclusions. In more complex or more exploratory experiments, where biological variability is likely to exceed technical errors, it is more critical to obtain biological replicates. Simon *et al.* (2002) provide a discussion of the relevant trade-offs.

In controlled experiments comparing gene expression in two or a small number of conditions, the goal of microarray study can be often described as identifying as many genes that are differentially expressed across conditions as possible while keeping the probability of making false declarations of differential expression acceptably low. If this is the goal, then we can address the question of how many replicates are required using well-developed hypothesis testing ideas. In general the answer depends on several factors: the signal-to-noise ratio, the desired sensitivity in detecting changes, and the tolerance for false findings (Robert *et al.*, 2003). In the context of microarray experiments Pan *et al.* (2002) discuss how to calculate the number of replicates given a normal-mixture model to detect changes in gene expression. In order to allow proper inference, we consider experiments with at least two replicates.

1.4 Some issues on microarray data analysis

The goal of many controlled microarray experiments is to identify genes that are regulated by modifying conditions of interest. For example, one may wish to compare wild type to knockout laboratory animals or alternative drug treatments. The objective of these experiments is to identify as many genes as possible that are differentially expressed across the compared conditions. There are two broad categories of situations which have to be kept in mind when choosing a statistical approach. In the first situation, we are comparing samples which the majority of the genes are expected to show some differences, albeit in varying degrees. An example would be cells at different stages of development. This problem is best approached by estimating the differences or ratios of expression across conditions for each gene. Considering the

variability of these ratios we can make a screening of the genes by using only "the most significant" ones and report the part of the genome that could be altered across the conditions under comparison. In a second approach we compare samples where a relative small fraction of genes are differentially expressed. For example, we may be studying alterations in expression caused by loss of a gene active in a specific pathway. In this type of applications, microarrays are often used to screen genes for further analysis by more reliable assays, and the data analysis is best approached by ranking genes and/or by selecting of genes for further validation. The methodology proposed here apply to both situations, because it is essentially focused on finding differentially expressed genes by formalizing the expression "the most significant genes".

1.4.1 Brief description of the Normalization process

As stated before the first step in microarray data analysis is the normalization process. This is briefly described here.

The notion that the normalization process is difficult to automate is widely accepted (Tseng *et al.*, 2001; Yang *et al.*, 2002). Normalization is best understood as an interactive process of visualization, identification of likely artifacts and their removal, when feasible. Examples of artifacts that can be at least partially removed include differentially non linear response of the two channels to hybridization intensity, biases in DNA spotting from defective print tips and the fainting of the signal in large regions of the array.

Evaluation of spatially varying bias is also critical. One, for example, can look at the original images or, more conveniently, at images of the processed absolute or relative expression values arranged by their location, or at "gradient plots" graphing intensity versus one of the spatial coordinates (Sellers *et al.*, 2003).

Investigation on normalization process is behind the scope of this thesis, in the sense that our aim is to capture the variability left from a normalization process by using a parametric models where the presence of nuisance parameters account for this variability. Nonetheless, we briefly discuss here the choices about normalization that have been made on the analyzed data sets. We generally distinguish the normalization in cDNA array from those in oligonucleotide array.

Although they are different experiments, they both have in common the problem of *background subtraction*, where the background here, B, has the meaning of either the background measured in cDNA arrays either the MM intensity measured in Affy chips. Let X represent the foreground in cDNA array and the PM intensity in Affy chips. It is widely accepted that X is likely to be the result of signal and additional background noise, then it is biased estimate of the true hybridization that we intend to measure (that is, it is likely to be systematically too high). To obtain an unbiased measure of expression, conventional wisdom is to subtract the background considering $X-B$ (the AD quantity in Affy chips). If both X and B are unbiased and background adds to the signal, then $X-B$ is unbiased. Even in these circumstances, however, there are important trade-offs to be evaluated in deciding whether and how to subtract background noise. Because both X and B are estimates, the variability of $X-B$ is

larger than that of X alone; thus, subtracting background adds variance. This is especially problematic in the low intensity range, where the variance of B can be of the same magnitude as X. Generally, the assumptions of unbiasedness and additivity are far too optimistic. Also, some researchers have found that the background estimates produced by popular image-processing algorithms are not sufficiently reliable (Yang *et al.*, 2002). One alternative is to avoid background subtraction altogether and only use X to estimate the expression level. This avoid to introduce the additional variance from inaccurately estimate background and is generally conservative in making declaration of differential expression in practical. To illustrate this, say that the true expressions in two sample being compared are e_1 and e_2. We observe $X_1 = e_1 + B_1 + \varepsilon_1$ and $X_2 = e_2 + B_2 + \varepsilon_2$ where ε_1 and ε_2 are measurement errors of the true signal. Because B's are positive, the log ratio of the non-background-corrected raw expression values X_1/X_2 is likely to be closer to 1 than the true ratio e_1/e_2. This bias toward one is stronger for low intensity genes. In summary, not subtracting background can be an attractive alternative, as it does not rely in potentially problematic background estimates and loses sensitivity mostly for low intensity genes; the exceptions are experiments with major spatial artifacts affecting only one channel. In practice, decisions about background subtraction need to be made based on careful visualization of the data and we decided to subtract it and not include in the analysis those spots where the $X < B$.

The background subtraction is a problem related to the normalization of spots in the array, but we also considered experiments that requires comparison of the gene expression measures across arrays. Variation across arrays reflects the genetic, experimental and environmental differences under study but will also include variation introduced during the sample preparation, during manufacture of the arrays, and during the processing of the arrays (labelling, hybridization and scanning). These are typical sources of variability that require technical replicates. The needed of technical replicates and the sources of noise generate by them require a normalization across arrays. This normalization process will be mainly centered on removing variability due to experimental artifacts.

Normalization in cDNA arrays

In two-channel cDNA arrays, the two signals allow for internal correction of a number of commonly occurring artifacts. In order to compare two sets of expression values (say, the two channels of a cDNA array or the expression levels across two arrays in Affy chips), it is useful to look at scatter plots of the two intensities or, as is more commonly done, to examine plots of differential expression versus overall intensity. This plot is called the *MA-plot* (also *RI-plot*) where the dimensions are $M = \log_2 R/G$ and $A = \log_2 \sqrt{RG}$. An *MA-plot* amounts to a 45° counterclockwise rotation of the $(\log_2 G, \log_2 R)$-coordinate system followed by a scaling of the coordinates. When the effect of artifacts are weak or they have been removed by normalization, we expect to see the points along the 45° ($\pi/2$) line from the origin of the plan spanned by MA (the bisectrix or angle bisector). In this case isolate points far from the bisectrix are

supposed to represent genes differentially expressed.

Normalization in oligonucleotide arrays

In Affy arrays and other platforms providing a single reading per spot, one can then construct MA-plots for each array versus the reference. Initial approaches for normalization across arrays are focused on standardizing overall intensity. This is useful, but often inadequate, as one commonly encounters systematic nonlinear distortions. Vertical residuals from robust regression of MA-plots against a common reference, as described above, can provide normalized values that account for nonlinear effects. Other popular alternatives, like quantile normalization (Bolstad *et al.*, 2002), won't be considered here.

1.4.2 The "Analyze and Then Summarize" approach

We mainly consider observational studies which describe the variation of genomic information in biological populations. These studies have broadly ranging goals including refinement of current taxonomies, identification of genome-phenotype relationships, classification and annotation of genes, and exploration of unknown pathways. The statistical tools brought to bear in these investigations cover the full range of traditional multivariate analysis, cluster analysis and classification. A challenge for the application of these analysis tools to genome-wide studies of gene expression comes from the large number of genes that are studied simultaneously and the high gene-to-sample ratio. Having many more genes than biological replicates makes possible a number of strategies for analysis. These can be categorized mainly into two broad categories: *Summarize Then Analyze* (STA) and *Analyze Then Summarize* (ATS). In the first category, STA, a multivariate procedure such as cluster analysis or multidimensional scaling is used to reduce the large number of genes to a smaller number of summary variables or profiles. These profiles are further then taken as outcome variables in, (say), a regression analysis of expression on experimental conditions, or as predictor variables in a model with a health outcome (typically in survival analysis). In this thesis we will consider the ATS approach, where the modelling is conducted for each gene, producing an estimate of a statistic of scientific interest (for instance: a difference or ratio among means, a regression coefficient, etc...) and its standard error. These gene-specific coefficients are then summarized, for example, by identifying those supposed to be differentially expressed according some statistical parametric model.

This approach lead to the data analysis steps adopted here: a first step is isolating those genes that are supposed to be significant or most significant differentially expressed (*gene shaving*), then we use these to draw inference on the underlying biological process or to make class prediction of the samples into known categories as shown in the application of a case study in Chapter 4.

1.4.3 Microarrays are useful to diagnose diseases

The complexity of gene expression analysis is stimulating the development of novel and specific statistical modelling tools to perform classification of class prediction. However, the existing body of pattern recognition and prediction algorithms developed in computer science and statistics can provide an excellent starting point for class prediction. Dudoit *et al.* (2002a) offer a practical comparison of discriminant methods for the classification of tumors using gene expression data. The one used here is the k-Nearest-Neighbor Classifiers (k-NNC) which is a simple and powerful class of algorithms for classification (Cover and Hart, 1967). Consider a sample of gene expression values (or *expression profile*), then the k-NNC classifies such sample of unknown phenotype by comparing the given gene expression profile to those of a sample of known phenotype. Essentially a k-NNC works in this way: suppose we want to classify the gene expression of sample into two phenotype class, say A and B. Let a_0 represent the sample point in the space of gene expression profile, then we will classify to A, if the majority of the k nearest points belongs to A otherwise B if the majority belongs to B. On the contrary we may decide of not classify the sample because there is not a majority. The choice of a metric is critical in order to measure distances among points in the expression profile. This discussion is behind the scope of this thesis and, as usual, we consider the Euclidean distance between points in the space of gene expression profile.

Modelling of gene expression data, by capturing the residual variability left from normalization, aims at usable classification. This requires the validation of a constructed classifier. The approach considered here is the most satisfactory to validation. It is based on the use of independent data, which can often be achieved by setting aside samples for validation purposes, as illustrated by Dudoit *et al.* (2002a). Statistical validation of probabilistic models (DeGroot and Fienberg, 1983) has two goals: assessing calibration (that is, the correspondence of the fraction predicted and the fraction observed in validation sample) and measuring refinement (that is, the ability of the model to discriminate between classes). As an alternative to setting aside samples for validation, one can use cross-validation, that is, splitting the data in portions and training the classifier a number of times equal to the number of portions and then setting aside each portion for validation. The resulting average classification rates is an unbiased estimated of the correct classification rate. Applying cross-validation techniques here results in an exponential increase of the computational efforts and it will not be considered here. It is also not needed, because of the availability of a relative large data set to be used for validation purpose.

The evaluation of classifiers on the same data that were used for training is a potentially serious mistake. When the number of predictors is very large, a relative large number of predictors will appear to be correlated with the phenotype of interest as a result of random variation present in the data. This spurious correlation has no biological foundation and does not generally reproduce outside of the sample studied. As a result, evaluation of classifiers on training data tends to give overly optimistic assessments of the validity of a classifier. It is more likely that a classifiers may

even have a near perfect classification ability in the training set without having any biological relation with phenotype.

1.4.4 We are mainly interested in gene screening using the False Discovery Rate approach

Perhaps, the simplest screening approach is to select genes based on average change in expression (say, difference in mean log expression across group). One problem with this, is that it ignores variation in how reliably each gene is measured. The problem is partially mitigated by careful normalization. Even after normalization, however, within the considered experimental conditions the variation of expression will be highly gene-dependent, therefore we have to account both for the uncertainty on these measures and also for the multiplicity.

Joint estimation of many related quantities is a time-honored problem in statistics, dating back at least to the pioneering work of Stein and colleagues (James and Stein, 1961) and continuing with empirical Bayes approaches (Efron and Morris, 1973) and hierarchical Bayesian multilevel models (Lindley and Smith, 1972). The idea behind the multilevel models and the associated empirical Bayes and hierarchical Bayes estimation techniques is to proceed in two stages. The first defines some useful summaries at the gene level, for example, a test statistics, or parameter estimates of the fold change and noise in parametric models. These describe the variability of samples for each gene. The second stage posits a distribution of these gene-level summaries. This approach has benefits in both estimation and selection. In the case of microarrays, examples of estimation are provided by Efron *et al.* (2001), Lönnstedt and Speed (2002) for selection and Ibrahim et el. (2002) and Newton *et al.* (2002) for both selection and estimation.

In analysis aimed at selecting differentially expressed genes, there are several approaches for reporting the degree of reliability of results. Conventional approaches based on gene-specific p-values are generally criticized on the grounds of the multiplicity of comparisons involved. Several proposals exist for adjusting p-values (see, for example, Dudoit *et al.*, 2002a and references therein) to account for multiplicities. A second approach, not considered here, is to compute the posteriori probability that a single gene is differentially expressed. For a discussion of the differences between the two approaches, see Berger and Delampady (1987).

This thesis is focused on the third and, perhaps, currently the most popular approach, that is, to estimate the *False Discovery Rate* (FDR) for a group of genes or for a specific cutoff value of a statistic. Assuming that the population of genes were truly divided into two groups, the altered and unaltered genes, and that a statistical approach selects a set of significant genes, the FDR is an estimate of the fraction of truly altered genes among the genes declared significant. This approach often reflects appropriately the fact that array experiments are performed to guide future validation work on individual genes, which is usually expensive and time-consuming. Another version of the FDR, namely $pFDR$, is also directly interpretable as the posterior probability that a gene is not differentially expressed in a list of genes declared as

differentially expressed.

An alternative to estimate the FDR could be to consider using classical Multiple Comparison procedures (MCPs) such as the well known Bonferroni correction. These procedures are too conservative when the amount of hypothesis under testing is high, as is the case in microarray application. In fact even though MCPs have been in use since early 1950s, and in spite of the advocacy for their use (e.g. it is strongly recommended by some journals, as well as in some institution such as the Food and Drug Administration in USA), researchers have not yet widely adopted these procedures. In medical research, for example, Godfrey (1985) and Pocock *et al.* (1987) examined sample of reports of comparative studies from major medical journals. They found that researchers overlook various kinds of multiplicity, and as a result reporting tends to exaggerate treatment differences (Pocock *et al.*, 1987). This underutilization in applied analysis is also due to other two difficulties with classical MCPs:

i) much of the methodology concerns comparisons of multiple treatments and families whose test statistics are multivariate normal (or t-student). In practice, many of the problems encountered are not of the multiple-treatments type, and test statistics are not multivariate normal (or t-student). In fact, families of hypothesis are often combined with statistics of different types;

ii) classical MCPs are focused on controlling the probability of making a false rejection and often this is not quite needed in the analysis. The control of this probability is important when a conclusion from the various individual inferences is likely to be erroneous when at least one of them is. This may be the case, for example, when several new treatments are competing against a standard, and a single treatment is chosen from the set of treatments which are declared significantly better than the standard. However, a treatment group and a control group are often compared by testing various aspects of the effect (different end points in clinical trials). The overall conclusions that the treatment is superior need not be erroneous even if some of the null hypotheses are falsely rejected.

The first difficulty has been partially addressed by the recent line of research advancing Bonferroni-type procedures, which use the observed individual p-values, while remaining faithful to control the probability of making a false rejection (see Simes (1986), Hommel (1988) and Hochberg (1988)). The other difficulties are still present and they constitute a serious problem. This is probably why other procedures which amounts to ignoring the multiplicity problem altogether, were recommended by some (e.g. Saville, 1990).

Several papers (Efron *et al.*, (2001); Efron and Tibshirani, (2002)) connect empirical Bayes methods with false discovery rates, but it is not clear whether these procedures allow to control the FDR as suggested in Dudoit *et al.* (2002b).

Chapter 2

Methodology

In this chapter we show the methodology adopted to perform *genes screening* (or *genes shaving*), of the genes under study in a microarray experiment. We want to remark the need of frequentist p-values in order to control the FDR and we want to show that the p-values proposed in Bayarri and Berger (2002) are useful in order to control the FDR. In this way we are allowed to extent the control of FDR to models where the p-values have to be calculated accounting the uncertainty on nuisance parameters. We do this by numerically investigating the control of FDR using the p-values proposed by Bayarri and Berger (2002). Furthermore we match the results about the frequentist property of the Bayarri and Berger p-values with those that allow the control FDR.

We start by setting up the single hypothesis test on each gene by introducing the use of p-values. We then introduce the p-values of Bayarri and Berger (2002), bringing some results from model criticism to hypothesis testing, being aware that these are two separate approaches to the inference. We then concentrate on Multiple Hypothesis testing MHT by illustrating the use and control of two main quantities: the FDR and the $pFDR$ introduced, respectively, by Benjamini and Hochberg (1995) and Storey (2002). We show the lack of efficiency of these methods when we use p-values that are not uniformly distributed under the null hypothesis. We will consider two situations when all genes are supposed independent and when there exists a kind of dependency called *clumpy dependency*, meaning that genes expression are dependent in small groups.

We apply the methodology to two relevant models for gene expression: the normal model and the gamma model. The aim of using the former is to show the relation between the Conditional Predictive p-value (p_{cpred}) with the classical t-test for two populations with unknown equal variance. In this way we justify the popular analysis on t-test in terms of Conditional Predictive p-value. We show how the Partial Posterior Predictive p-value (p_{ppost}) improves the inference using the Gamma model, which is a flexible and popular model in microarrays analysis (Newton *et. al*, 2002).

2.1 Single Hypothesis Testing

We introduce here the notation of single hypothesis testing. Lehmann (1986) covers hypothesis testing in detail, so the reader should refer there for a more thorough discussion of it. Suppose we are given a set of data, generally noted with $\mathbf{X}$, and we suppose that the data follow some distribution $F_{\boldsymbol{\theta}}$, where $F_{\boldsymbol{\theta}}$ comes from a family of distributions indexed by $\boldsymbol{\theta} \in \Theta$. Some subset of Θ, say, Θ_0 represent the *null hypothesis*, which is usually the state of $\boldsymbol{\theta}$ that one hopes to find evidence against. Some other subset Θ_1 represent the *alternative hypothesis*. For example, we are interesting in assessing the differential expression of a gene in two tissues, the null hypothesis Θ_0 would tend to be the set of $\boldsymbol{\theta}$ that indicate the gene is not differentially expressed. The alternative hypothesis Θ_1 contains the set of $\boldsymbol{\theta}$ values representing differential expression for the considered gene: either overexpressed or underexpressed. It is always the case that $\Theta_0 \cap \Theta_1 = \emptyset$, and sometimes $\Theta_0 \cup \Theta_1 = \Theta$. If Θ_i consists of a single value, then this hypothesis is said to be *simple*, otherwise, Θ_i is a *composite* hypothesis.

A test statistic $T = t(\mathbf{X})$ is chosen to investigate compatibility of the model $F_{\boldsymbol{\theta} \in \Theta_0}$ with the observed data, $\mathbf{x}_{obs}$, and/or decide whether $\boldsymbol{\theta} \in \Theta_0$ or $\boldsymbol{\theta} \in \Theta_1$. We think that the choice of the test statistic is driven by the data analysis problem and we do not discuss choices of T, which can be anything from the unchanged set of data to an univariate quantity. However, we follow the convention that large values of T indicate less compatibility or evidence against de decision $\boldsymbol{\theta} \in \Theta_0$. The decision is based on a rejection region, which we will denote by Γ. If $T \in \Gamma$, then we decide that the evidence is not in favors of $\boldsymbol{\theta} \in \Theta_0$; if $T \notin \Gamma$, then we decide $\boldsymbol{\theta} \in \Theta_0$. There are two kinds of errors that can be committed when testing a hypothesis. The first is a *Type I error* (false positive or false discovery), which occurs when $T \in \Gamma$ yet $\boldsymbol{\theta} \in \Theta_0$. Therefore, the Type I error rate for a specific $\boldsymbol{\theta}_0 \in \Theta_0$ is $\int_{\{T \in \Gamma\}} dF_{\boldsymbol{\theta}_0}$, which we will denote by $\Pr(T \in \Gamma | \boldsymbol{\theta}_0)$. The second class is *Type II error* (false negative or false non-rejection), and this occurs when $T \notin \Gamma$ yet $\boldsymbol{\theta} \in \Theta_1$. The Type II error rate for a specific $\boldsymbol{\theta}_1 \in \Theta_1$ is $\Pr(T \notin \Gamma | \boldsymbol{\theta}_1)$. The *power* of the test (T, Γ) at $\boldsymbol{\theta}_1$ is $\Pr(T \in \Gamma | \boldsymbol{\theta}_1)$, and this is equal to $1 - \Pr(T \notin \Gamma | \boldsymbol{\theta}_1)$; in words, the power is the probability of rejecting, given the alternative hypothesis is true.

The optimality criterion defined for hypothesis testing is to find the *most powerful test*. Let Y be a test statistic and Δ a rejection region for Y. For a given $\boldsymbol{\theta}_1 \in \Theta_1$, (T, Γ) represents the most powerful test of size $\sup_{\boldsymbol{\theta} \in \Omega_0} \Pr(T \in \Gamma | \boldsymbol{\theta})$ if for all (Y, Δ) such that

$$\sup_{\boldsymbol{\theta} \in \Theta_0} \Pr(T \in \Gamma | \boldsymbol{\theta}) \leq \sup_{\boldsymbol{\theta} \in \Theta_0} \Pr(Y \in \Delta | \boldsymbol{\theta}),$$

we have

$$\Pr(T \in \Gamma | \boldsymbol{\theta}_1) \geq \Pr(Y \in \Delta | \boldsymbol{\theta}_1).$$

If (T, Γ) is most powerful $\forall \boldsymbol{\theta}_1 \in \Theta_1$, then it is said to be *uniformly most powerful*.

Given the optimality criterion for testing a single hypothesis, it is convenient to consider only the nested set of rejection regions, such as those defined by p-values.

For the remainder of this thesis, we will use the symbol H to denote the state of the hypothesis: we pose $H = 0$ when $\boldsymbol{\theta} \in \Theta_0$ and $H = 1$ when $\boldsymbol{\theta} \in \Theta_1$. Whether the

hypothesis is simple or composite will be explicitly stated.

2.1.1 Hypothesis testing versus model criticism

Let $f(\mathbf{x};\boldsymbol{\theta})$ be the corresponding density of $F_{\boldsymbol{\theta}}$. In this thesis we assume $f(\mathbf{x};\boldsymbol{\theta})$ is a continuous density (with respect to Lebesgue measure), this is not strictly necessary in single hypothesis testing, but it is generally required in MHT.

Consider now the previous setup in terms of densities

$$H = 0 : \mathbf{X} \sim f(\mathbf{x};\boldsymbol{\theta}_0) \text{ against } H = 1 : \mathbf{X} \nsim f(\mathbf{x};\boldsymbol{\theta}_0)$$

The null hypothesis $H = 0$ means that one of the distributions $F_{\boldsymbol{\theta}\in\Theta_0}$ governs the system with reasonable approximation, or that the model $f(\mathbf{x};\boldsymbol{\theta}_0)$ is compatible with the observed data. It is worthy here to stress the difference between hypothesis testing and model criticism (or model validation or model checking) following the recent paper of (O'Hagan, 2003). What makes the difference between hypothesis testing and model criticism is how we specify the alternative hypothesis. More formally, we are in a hypothesis testing framework if the alternative is

$$\mathbf{X} \nsim f(\mathbf{x};\boldsymbol{\theta}_0) \equiv \mathbf{X} \sim f(\mathbf{x};\boldsymbol{\theta}_1 \in \Theta_1),$$

while we are doing model criticism (or model checking) if

$$\mathbf{X} \nsim f(\mathbf{x};\boldsymbol{\theta}_0) \equiv \mathbf{X} \sim f_1(\mathbf{x};\boldsymbol{\omega}),$$

where $f_1(\mathbf{x};\boldsymbol{\omega})$ belongs to the set of all models but $f(\mathbf{x};\boldsymbol{\theta}_0)$. The tools that we are going to present, that relay on p-values as measure of evidence against $H = 0$, are explicitly designed for model criticism and not for hypothesis testing. However, here we are only interested in constructing a nested set of rejection regions for $H = 0$ and assign to that $\Pr(T \in \Gamma|\boldsymbol{\theta}_0)$. This can be done without specifying any alternative hypothesis. In this way we can produce proper inference using MHT, because in MHT the assumptions on the distribution of the p-values under the alternative hypothesis are very weak and operative procedures need only to know the distribution of the p-values under the null hypothesis.

We recognize that this point of view regarding p-values is not shared in part of statistical literature as clearly shown in Hubbard and Bayarri (2003). This is originate by the debate on the differences between the Fisher's ideas on significance testing (and inductive inference) and Neyman-Pearson's views on hypothesis testing (and inductive behavior). According to Hubbard and Bayarri (referring to Berger, 2002) the agrement between the two approaches can be reached if the p-values are calibrated with respect to some quantity of interest such as the posterior probability of $H = 0$. By calibration we mean that the p-values are values in some known scale. It is worthy here to remark that:

a) we calculate the p-values approximating a probability, but we interpret them as test statistics instead of probabilities. That, is P is a test statistic in the range $(0, 1)$ whose distribution under $H = 0$ belongs to the Beta family. In particular we

want this distribution to be fixed under the null model. This is necessary because of convergency theorems on MHT which will be show later. Moreover, in order to readily obtain the $\Pr(P \geq p)$ we also want $P \sim \mathcal{U}(0,1)$ (at least approximately). In the same spirit of Berger (2002) we remark that main point here is whether or not the P are calibrated with respect to some probability law regarding P, that is, whether or not we can make the statement $\Pr(P \leq p | H = 0) = p$;

b) we define the rejection region for $H = 0$, Γ_α as in (1.3) and therefore we consider a gene not differentially expressed if the random rejection region obtained from a p-value leads to α greater than some values. In particular the threshold level is decided after (and not prior to) the experiment and it is explicitly suggested by the MHT procedures that leads to Err, which is a compound error measures for all test.

With respect to the application on Microarrays it is very important to note that $H = 0$ means that the gene in two biological target samples under comparison is not differentially expressed according to the parametric model $f(\mathbf{x}; \boldsymbol{\theta}_0)$ but $f_1(\mathbf{x}; \boldsymbol{\omega})$ could be a model were the gene is still not differentially expressed. This is the price we are paying to combine parametric model criticism with hypothesis testing, it is the price we are paying to have assumed $f(\mathbf{x}; \boldsymbol{\theta}_0)$ as a reference model for the null hypothesis. This may seems to be a critical issue. The solution could be to use non parametric methods, such as the permutation tests, that Westfall and Young (1993) introduced in MHT data analysis based on original ideas of Tukey and further developed by Ge, Dudoit and Speed (2003). However, errors in approximating the null hypothesis with non-parametric methods is very large when the sample size is small, as it is the case for small scale microarray experiments. These experiments are the most popular due the prohibitive cost of a large scale study. For this reason we do not consider non-parametric null hypothesis, therefore we will not consider in detail MHT procedures based on permutation tests (Pesarin, 2001), such as those implemented in popular software `SAM` (Significance Analysis of Microarrays) which is a package in the `Bioconductor` project.

2.1.2 The p-value for composite null model

We simplify the notation by considering $f(\mathbf{x}; \boldsymbol{\theta})$ to be the model under $H = 0$, because we do not specify any alternative model. We consider composite null hypothesis where $\boldsymbol{\theta} \in \Theta$ and further we partition $\boldsymbol{\theta} = (\theta, \theta')$ and we regard θ as a nuisance parameter. This setup is a particular case of those in Berti, Fattorini and Rigo (2000) where they provide a characterization to the problem of eliminating nuisance parameters. Here we will mainly consider eliminating θ by maximum likelihood estimators and integrated likelihood and we do not aim explicitly towards the *mixture model* in Berti, Fattorini and Rigo (2000), which has been derived by eliminating the nuisance parameter θ.

To assess the compatibility of a model with the data it is necessary to chose a test statistic T. We do not discuss the choice of T and we will consider natural choices of T, in particular we will favor choice of statistics whose distribution under the null

hypothesis is known at least in their kernel. Let $f(t;\theta)$ be the density of T under the null hypothesis we will require that

$$f(t;\theta) \propto f^*(t;\theta)$$

where $f^*(t;\theta)$ is the kernel density t with parameter θ.

There are several measures of compatibility (or *measures of surprise*) of the data with $f(\mathbf{x};\boldsymbol{\theta})$, Bayarri and Castellanos (2001) investigate the behavior of some, for different choice of T in a goodness-of-fit framework. The most commonly used one is the p-value defined in terms of test statistics

$$\begin{aligned} p &= \Pr(T(\mathbf{X}) \geq t(\mathbf{x}_{obs})) \\ &= \Pr(T \geq t_{obs}) \\ &= \int_{T \geq t_{obs}} f(t;\theta)\, dt \end{aligned} \tag{2.1}$$

where $\mathbf{x}_{obs}$ are the observed data.

When $\theta = \theta_0$ the null model consists of a single distribution, the p-value is readily obtained, and it is uniformly distributed under $H = 0$. On the on the other hands, when the null model depends on unknown nuisance parameter θ, one must somehow eliminate θ. Various proposals have been suggested to remove θ, each yielding a different candidate p-value. Before analyze some of the most popular candidate p-values, we further motivate the need for p-values to be uniformly distributed on $[0,1]$.

2.1.3 The desirable finite and asymptotic sample property of candidate p-values

We call the random variable $p(\mathbf{X})$ a candidate p-value if it range in $[0,1]$; if it is also uniform under $H = 0$ ($p(\mathbf{X}) \sim \mathrm{U}(0,1)$), we say that $p(\mathbf{X})$ is a *frequentist* p-value. When a candidate p-value is not uniform, we say that it is conservative (anti-conservative) at θ if $\Pr\{p(\mathbf{X}) < \alpha\}$ is less (greater) than α for all $\alpha < 1/2$. Finally, a candidate p-value is globally conservative (anti-conservative) if it is conservative (anti-conservative) for all $\theta \in \Theta$.

This terminology was motivated by the following considerations. All considered candidate p-values have range $[0,1]$, but, because $H = 0$ is composite, they may not be uniform distributed, even when the null is true, the only exception is when T is an *ancillary* statistic. Yet in practice we use small values of $p(\mathbf{x}_{obs})$ to denote surprise or incompatibility, in analogy with the non-composite case, we act as if $p(\mathbf{X})$ were $\mathrm{U}(0,1)$. Seriously anti-conservative candidate p-values lead to discard the null model even when it is quite compatible with the data, while seriously conservative candidates may cause to fail to discard models that are grossly incompatible with the data.

The point here is that a p-value is useful to assess compatibility of the null model with the data only if its distribution under the null model, $G_0(p)$, is known to the analyst; otherwise there is no way of assessing whether or not observing, for example, $p = 0.1$ is surprising. The fact that we require the distribution to be uniform is largely a matter of convention, but it made readily usable all the results in MHT. We could

admit other distribution on $[0,1]$, because we can always find a transformation of $p(\mathbf{X})$ such that the resulting random variable is uniform in $[0,1]$.

Hence, for frequentist testing purposes, we should require that candidate p-values to be frequentist p-values. This requirements is generally unfulfillable for all sample sizes, with the exception of special models, many of which are discussed by Bayarri and Berger (1999), but it can often be approximately satisfied in large samples. We may argue that $f(t;\theta)$ can be chosen so that $p(\mathbf{X})$ is *asymptotic frequentist p-value*, i.e. one whose distribution converges in law to $\mathrm{U}(0,1)$ distribution under $H=0$ as $n \to \infty$ (Robins *et al.*, 2000). In this sense Bayesian statisticians, who use p-values to assess the compatibility of a model with the data, should require to use asymptotic frequentist p-values. In fact if the goal is to check the model rather than the prior, any procedure should perform adequately whatever the prior, and this would imply that p-values should be required to be frequentist p-values. However as the sample size increases, the data dominate any prior with support on all Θ, therefore Bayesians should both expect and require that any model checking procedure should perform adequately in the limit as $n \to \infty$.

2.1.4 Some candidate p-value

Let $h(t)$ be the marginal density of $T(\mathbf{X})$, that is the distribution obtained eliminating the nuisance parameter. Each way of eliminating θ lead to a different $h(t)$ and then to different p-values. We examine, here, the various solutions proposed to eliminate the nuisance parameter θ. We start with the classical solution and then we illustrate the Conditional predictive p-value and the Partial posterior predictive p-value.

Classical p-values

The similar p-value. The first and most promising solution is to *condition on an ancillary statistic for* θ. Let $T'(\mathbf{X}) = t'$ be an ancillary statistic for θ then $f(\mathbf{x}|t'_{obs};\theta) = f(\mathbf{x}|t'_{obs})$ which is a completely specified function and $h(\cdot|t'_{obs})$ can be obtained by applying the random variable transformation, $T' = T'(\mathbf{X})$. In this case we define the *similar p-value* as follow:

$$p_{sim} = \Pr{}^{h\left(\cdot|t'_{obs}\right)}\left(T(X) \geq t(\mathbf{x}_{obs})\right)$$

Weakness of this approach are:

i) computation can be burdensome;

ii) a suitable sufficient and ancillary statistic T' often does not exist;

iii) after conditioning upon T', the choice of compatibility statistic T may be forced on the user, and this may not be a desirable choice.

The plug-in p-value. Another alternative is to replace θ by same estimate, say the Maximum Likelihood Estimate (MLE), $\widehat{\theta}$, this lead to the *plug-in* p-value, where

$h(t) = h\left(t; \widehat{\theta}\right)$ (*plug-in distribution*) and

$$p_{plug} = \Pr^{h\left(\cdot;\widehat{\theta}\right)}\left(T(X) \geq t(\mathbf{x}_{obs})\right) \tag{2.2}$$

The main weakness is that p_{plug} does not account for the uncertainty on θ and it also over estimates the evidence against the null hypothesis.

The prior predictive p-value. A first Bayesian approach for the problem is to use the *predictive distribution*: $h(t)$ can be obtained by a transformation rule on $m(\mathbf{x}) = \int f(x;\theta)\,\pi(\theta)\,d\theta$. The *prior predictive p-value* is

$$p_{prior} = \Pr^{m(\cdot)}\left(T(X) \geq t(\mathbf{x}_{obs})\right) \tag{2.3}$$

Here $m(\mathbf{x})$ measures the likelihood of the data relative to both the model and the prior, therefore an excellent model could come under suspicion if a poor prior distribution were used. For this reason, in an exploratory data analysis, a non subjective analysis might be desired. This makes attractive to use non informative priors, but, unfortunately, these are typically improper priors, in which case the prior predictive $m(\mathbf{x})$ would also be improper and this precludes computation of (2.3). For this reason we no further consider the prior predictive p-value that was originally popularized by Box (1980).

The posterior predictive p-value. From now on we will always assume $\pi(\theta)$ to be an improper prior of $\theta \in \Theta$. This concerns leads many Bayesians, beginning with Guttman (1967) and Rubin (1984), to consider the *posterior predictive p-value*: θ is integrated out with its marginal posterior distribution and calculating the posterior predictive distribution $m_{post}(\mathbf{x}|\mathbf{x}_{obs}) = \int f(\mathbf{x};\theta)\,\pi(\theta|\mathbf{x}_{obs})$,

$$p_{post} = \Pr^{m_{post}(\cdot|\mathbf{x}_{obs})}\left(t(X) \geq t(\mathbf{x}_{obs})\right) \tag{2.4}$$

The strengths for this approach are:

i) improper non-informative priors can readily be used;

ii) $m_{post}(\mathbf{x}|\mathbf{x}_{obs})$ will be much more influenced by the model than by the prior and the posterior distribution will essentially concentrate at $\widehat{\theta}$ so that $p_{post} \to p_{plug}$ as $n \to \infty$.

The weakness of this approach is that we are making a double use of the data: first to train the improper prior into a proper distribution and then computing the tail area corresponding to $t(\mathbf{x}_{obs})$. This approach is very conservative with respect to the rejection of $H = 0$ in particular when the sample size is small as in the case of microarray data analysis.

The Partial Posterior Predictive p-value and the Conditional Predictive p-value.

On the spirit of avoid double use of the data Bayarri and Berger (2000) propose the *partial posterior predictive p-value*:

$$p_{ppost} = \Pr{}^{m(\cdot|\mathbf{x}_{obs}\backslash t_{obs})}\left(t\left(\mathbf{X}\right) \geq t\left(\mathbf{x}_{obs}\right)\right) \tag{2.5}$$

here $m\left(\cdot|\mathbf{x}_{obs}\backslash t_{obs}\right)$ and the *partial posterior* $\pi\left(\cdot|\mathbf{x}_{obs}\backslash t_{obs}\right)$ are given by

$$\begin{aligned} m\left(\cdot|\mathbf{x}_{obs}\backslash t_{obs}\right) &= \int f\left(t|\theta\right)\pi\left(\theta|\mathbf{x}_{obs}\backslash t_{obs}\right)d\theta, \\ \pi\left(\theta|\mathbf{x}_{obs}\backslash t_{obs}\right) &\propto f\left(\mathbf{x}_{obs}|t_{obs};\theta\right)\pi\left(\theta\right) \propto \frac{f\left(\mathbf{x}_{obs};\theta\right)\pi\left(\theta\right)}{f\left(t_{obs};\theta\right)}, \end{aligned} \tag{2.6}$$

where $\pi\left(\theta\right)$ is an improper prior and $\pi\left(\cdot|\mathbf{x}_{obs}\backslash t_{obs}\right)$ is assumed proper. Conditioning the posterior to the density of the observed statistic avoids the double use of the data that occurs in the posterior predictive p-value because the contribution of t_{obs} to the posterior is removed before eliminate θ by integration. The notation $\mathbf{x}_{obs}\backslash t_{obs}$ indicates this.

The second p-value they propose is a specific case of what they termed a *U-conditional predictive p-value*, defined, for some conditioning statistic $U = u\left(\mathbf{x}\right)$ as

$$p_{cpred(u)} = \Pr{}^{m(\cdot|u_{obs})}\left(t\left(\mathbf{X}\right) \geq t\left(\mathbf{x}_{obs}\right)\right); \tag{2.7}$$

here $u_{obs} = u\left(\mathbf{x}_{obs}\right)$, and the conditional distribution $T|U$ is

$$m\left(t|u\right) = \int f\left(t|u;\theta\right)\pi\left(\theta|u\right)d\theta,$$

assuming that

$$\pi\left(\theta|u\right) = \frac{f\left(u;\theta\right)\pi\left(\theta\right)}{\int f\left(u;\theta\right)\pi\left(\theta\right)d\theta}$$

is proper. $f\left(t|u;\theta\right)$ and $f\left(u;\theta\right)$ are defined as conditional and marginal densities of T and U under the null hypothesis.

The specific proposal that they recommend, for $m\left(t|u\right)$ to be invariant under parameter transformation, is to choose U in (2.7) to be the conditional MLE of θ, $\widehat{\theta}_{cMLE}\left(\mathbf{x}\right)$, given by

$$\widehat{\theta}_{cMLE}\left(\mathbf{x}\right) = \arg\max_{\theta\in\mathbf{\Theta}} f\left(\mathbf{x}|t,\theta\right) = \arg\max_{\theta\in\mathbf{\Theta}} \frac{f\left(\mathbf{x};\theta\right)}{f\left(t;\theta\right)} \tag{2.8}$$

and the resulting p-value is called the *conditional predictive p-value* denoted by $p_{cpred} = p_{cpred\left(\widehat{\theta}_{cMLE}\right)}$. Given the invariant property of MLE under one-to-one transformations of $u\left(\mathbf{x}\right)$, then also $m\left(t|u\right)$ is invariant to any one-to-one transformation of (2.8).

When $f\left(\mathbf{x};\theta\right)$ belongs to the natural scale exponential family the following Theorem 1 apply.

Theorem 1 *(Bayarri and Berger, 2000). Suppose $f\left(\mathbf{x};\theta\right)$ is a continuous density from the natural scale exponential family, and that statistics $T > 0$ and $U > 0$ exists such that $S = T + U$ is sufficient and the joint density*

$$f\left(t;u;\theta\right) = k\theta^{\alpha}t^{\gamma}u^{\alpha-\gamma-2}\exp\left\{-\theta\left(t+u\right)\right\},$$

for some constants k, $\alpha > -1$, *and* $\gamma < \alpha - 1$. *Under the prior* $\pi(\theta) \propto 1/\theta$ *then* $p_{cpred} = p_{post} = p_{sim}$.

Theorem 1 can be particularized by noting that when T is conditionally independent of $\widehat{\theta}_{cMLE}$ and $\left(T, \widehat{\theta}_{cMLE}\right)$ are jointly sufficient for $\boldsymbol{\theta}$, then $p_{ppost} = p_{cpred}$. But when this condition does not hold then we can consider p_{ppost} to approximate p_{cpred}, in fact Robins *et al.* (2000) showed that p_{cpred} and p_{ppost} are asymptotic frequentist p-values: their asymptotic distribution is $\mathcal{U}(0,1)$ for all θ when $n \to \infty$. Successively Bayarri and Berger (2000) showed the following theorem.

Theorem 2 *(Bayarri and Berger, 2000). Let* $p(\mathbf{X})$ *be any* U*-conditional* p*-value for a proper* $\pi(\theta)$*, and consider it as a random variable with respect to the distribution* $f(\mathbf{x};\theta)$. *If the distribution of* $p(\mathbf{X})$ *does not depend on* θ*, then* $p(\mathbf{X})$ *is a frequentist* p*-value for all* θ. *The conclusion also holds for* $\pi(\theta)$ *improper under this condition: suppose a sequence of increasing compact sets* $\Theta_k \subset \Theta$ *such that* $\cup_{k\geq 1}\Theta_k = \Theta$, $0 < m_k = \int_{\Theta_k} \pi(\theta)\, d\theta$, $m(u) = \int_{\Theta} f(u;\theta)\,\pi(\theta)\, d\theta < \infty$, *and* $\lim_{k\to\infty} m_k \int \frac{[m_k(u)]^2}{m(u)} du = 1$, *where* $m_k(u) = \left(\int_{\Theta_k} f(u;\theta)\,\pi(\theta)\, d\theta\right)/m_k$. *This condition is satisfied if* U *has location or scale parameter distribution and* $\pi(\theta)$ *is the reference prior (Berger and Bernardo, 1992abc).*

For the p_{cpred} this result may help to remove any concerns about asymptotic arguments, while, for the p_{ppost} the result of Robins *et al.* (2000) may help to alleviate them. However calculate p_{ppost} is much easier than calculate p_{cpred}, in fact it is only necessary to know the density $f(t;\theta)$, or at least its kernel $f^*(t;\theta)$. However in the case we don't know the density even on its kernel $f^*(t;\theta)$, we can approximate it by generating $\theta \sim \pi(\theta|\mathbf{x}_{obs})$ and for each θ generate $\mathbf{x}$, calculate t and approximate $f(t;\theta)$ with a kernel density. This is a very computational intense procedure that has not been used here because we considered application where we know at least $f^*(t;\theta)$. However this computationally intense procedure shows the potential of the methodology that can be applied to almost every practical situations.

2.2 Microarray data analysis with two models: the Normal model and the Gamma model

We apply the above methodology to two models: the Normal model and the Gamma model. In the former it is possible to calculate p_{cpred} and, using Theorem 2 we show that it is $\mathcal{U}(0,1)$ for all sample size. The p_{cpred} in the Normal model is the p-value obtained from the classical t-test for the difference of two means when the variance is unknown but equal in the two populations. t-tests are widely used in microarray analysis to perform gene screening and here we characterize them in terms of the used methodology.

The Gamma model has been also used in microarray data analysis because of its flexibility (Newton *et al.*, 2002). It allows equal variation coefficient on the measures of mRNA abundance in two samples (say, R and G signals in cDNA array experiments).

This feature has been found relevant in many experiments as noted in Chen *et al.* (1997). For this model we didn't found p_{cpred} and we only calculate p_{ppost} and we show that it outperforms the p_{plug} and p_{post} in terms of asymptotic approximation to the uniform distribution. In the end this leads to more efficient MHT procedures.

The sketches of the calculations are reported here while the details are showed in the Appendices A and B.

2.2.1 The Normal model for single gene expression

Let $\mathbf{X}$ and $\mathbf{Y}$ be two independent measures of mRNA abundance for a single gene in two biological target samples under comparison. Suppose that

$$\begin{aligned} \mathbf{X} &= X_1, \ldots, X_{n_X} \sim N\left(\mu_X, \sigma^2\right), \quad i.i.d. \\ \mathbf{Y} &= Y_1, \ldots, Y_{n_Y} \sim N\left(\mu_Y, \sigma^2\right), \quad i.i.d. \end{aligned}$$

where all the parameters are unknown. We want to investigate the compatibility of the data with the model that support the hypothesis of no differential expression. Parametrically, this requires the two Normal distributions to have the same mean. Formally,

$$H = 0 : \mu_X = \mu_Y = \mu, \forall \sigma^2 > 0$$

against a non specified alternative. We found convenient to use the following statistic:

$$T\left(\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}\right) = t\left(\mathbf{x}, \mathbf{y}\right) = \overline{x} - \overline{y}$$

where $\mathbf{x}$ and $\mathbf{y}$ represent the observations of sizes respectively n_X and n_Y, and $\overline{x}$, $\overline{y}$ their respectively means. The null distribution of the test statistic is

$$\begin{aligned} T\left(\mathbf{X}, \mathbf{Y}\right) | H_0 &\sim f\left(t|\sigma^2\right) = N\left(0, \left(\frac{1}{n_X} + \frac{1}{n_Y}\right)\sigma^2\right) \\ &= \frac{1}{\sqrt{2\pi \frac{n_X + n_Y}{n_X n_Y}\sigma^2}} \exp\left(-\frac{t^2}{2\frac{n_X + n_Y}{n_X n_Y}\sigma^2}\right) \end{aligned}$$

and σ^2 is the nuisance parameter. We denote with $t\left(\mathbf{x}, \mathbf{y}\right)$ the observed value of T and with t a point in the space defined by T, the real line in this case.

The Plug-in p-value for the Normal model

The MLE of μ and σ^2 are

$$\begin{aligned} \widehat{\mu} &= \frac{n_X}{n_X + n_Y}\overline{x} + \frac{n_Y}{n_X + n_Y}\overline{y} \\ \widehat{\sigma}^2 &= \frac{\sum_{i=1}^{n_X} x_i^2 + \sum_{i=1}^{n_Y} y_i^2}{n_X + n_Y} - \widehat{\mu}^2 \end{aligned}$$

Therefore the Plug-in p-value is:

$$p_{plug} = 2\left(1 - \Phi\left(\frac{|t\left(\mathbf{x}, \mathbf{y}\right)|}{\sqrt{\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)\widehat{\sigma}^2}}\right)\right)$$

where $\Phi(\cdot)$ is the cumulative distribution function (cdf) of a standard normal distribution.

The posterior predictive p-value for the Normal model

Using the reference prior for μ and σ^2:

$$\pi\left(\mu,\sigma^2\right) \propto 1/\sigma^2, \sigma \in \mathcal{R}^+$$

the marginal posterior distribution for σ^2 is

$$\pi\left(\sigma^2|\mathbf{x},\mathbf{y}\right) = Ga^{-1}\left(\frac{n_X+n_Y}{2}-1,\widehat{\sigma}^2\right)$$

where $Ga^{-1}\left(\frac{n_X+n_Y}{2}-1,\widehat{\sigma}^2\right)$ denotes the inverse gamma distribution with scale parameter $\widehat{\sigma}^2$. The marginal posterior distribution for T is

$$m_{post}\left(t|\mathbf{x},\mathbf{y}\right) = \zeta_{n_X+n_Y-2}\left(0,\widehat{\sigma}^2\frac{n_X+n_Y}{n_Xn_Y}\frac{2}{n_X+n_Y-2}\right)$$

where $\zeta_{n_X+n_Y}\left(0,\widehat{\sigma}^2\frac{n_X+n_Y}{n_Xn_Y}\frac{2}{n_X+n_Y-2}\right)$ represents the density of a centered t-student distribution with n_X+n_Y degrees of freedom and scale parameter equals to $\widehat{\sigma}^2\frac{n_X+n_Y}{n_Xn_Y}\frac{2}{n_X+n_Y-2}$. Therefore the Posterior p-value is

$$p_{post} = 2\left(1-\Upsilon_{n_X+n_Y}\left(\frac{|t\left(\mathbf{x},\mathbf{y}\right)|}{\sqrt{\widehat{\sigma}^2\frac{n_X+n_Y}{n_Xn_Y}\frac{2}{n_X+n_Y-2}}}\right)\right)$$

where $\Upsilon_{n_X+n_Y}(\cdot)$ is the c.d.f. of a standard t-student distribution with n_X+n_Y degrees of freedom.

The conditional predictive p-value for the Normal model

The U statistic is a vector of two elements $\widehat{\mu}_{cMLE}$ and $\widehat{\sigma}^2_{cMLE}$

$$u\left(\mathbf{x},\mathbf{y}\right) = \arg\max_{\mu,\sigma^2}\frac{L\left(\mu,\sigma^2\right)}{f\left(t|\sigma^2\right)} = \left(\widehat{\mu}_{cMLE}=\widehat{\mu},\widehat{\sigma}^2_{cMLE}=\frac{n_XS_x^2+n_YS_y^2}{n_X+n_Y-1}\right),$$

where $S_x^2=\overline{x^2}-\overline{x}^2$, $S_y^2=\overline{y^2}-\overline{y}^2$ and $\overline{x^2}=\sum_{i=1}^n x_i^2/n_X$, $\overline{y^2}=\sum_{i=1}^n y_i^2/n_Y$ are the second sampling moments. Note that only $\widehat{\sigma}^2_{cMLE}$ is relevant in order to eliminate σ^2. Further we note that $\widehat{\sigma}^2_{cMLE}$ is independent from T and they are jointly sufficient for σ^2 and μ; so by Theorem 1 the Partial Posterior Predictive p-value equals Conditional predictive p-value.

Let $f\left(u\left(\mathbf{x},\mathbf{y}\right),\mu,\sigma^2\right)$ represent the joint density of U and the parameters, then the U-conditional distribution for the parameter is given by:

$$\pi\left(\mu,\sigma^2|u\left(\mathbf{x},\mathbf{y}\right)\right) = \pi\left(\mu|\sigma^2,\frac{n_X\overline{x}+n_Y\overline{y}}{n_X+n_Y}\right)\pi\left(\sigma^2|\widehat{\sigma}^2_{cMLE}\right)$$

where:

$$\begin{aligned}\pi\left(\mu|\sigma^2,\frac{n_X\overline{x}+n_Y\overline{y}}{n_X+n_Y}\right) &= N\left(\frac{n_X\overline{x}+n_Y\overline{y}}{n_X+n_Y},\frac{\sigma^2}{n}\right)\\ \pi\left(\sigma^2|\widehat{\sigma}^2_{cMLE}\right) &= Ga^{-1}\left(\frac{n_X+n_Y-2}{2},\frac{n_X+n_Y-1}{2}\widehat{\sigma}^2_{cMLE}\right)\end{aligned}$$

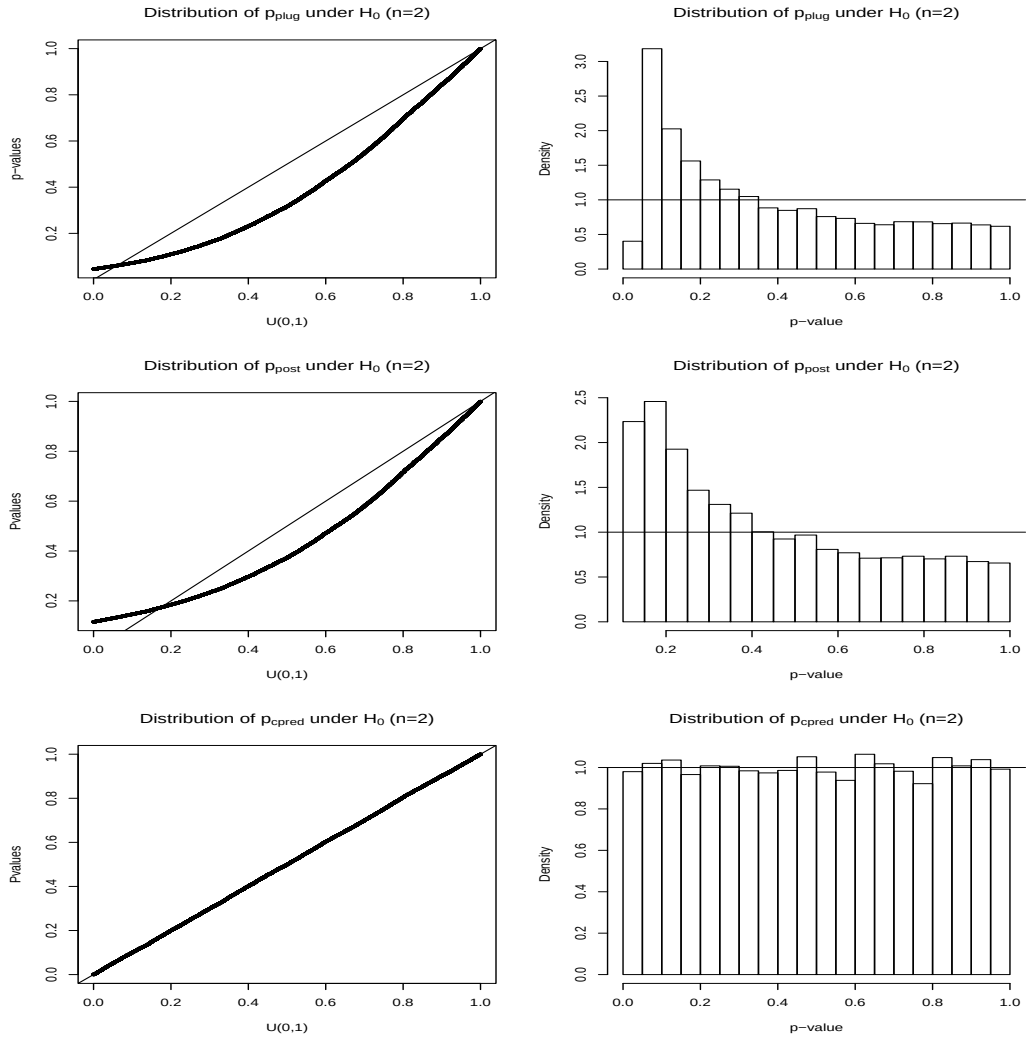

Figure 2.1: Distribution of the p-values under the null model: $\mathbf{X}, \mathbf{Y} \sim \Phi(\cdot)$ and $n_X = n_Y = 2$. We can see that only the p_{cpred} is uniformly distributed. In the left columns are showed the quantiles of the empirical distribution of the p-values against the quantile of the $\mathcal{U}(0,1)$ distribution (QQ-plot). The right column show the histograms of the same distribution.

The marginal distribution of $t|u(\mathbf{x},\mathbf{y})$ is given by:

$$m\left(t|u\left(\mathbf{x},\mathbf{y}\right)\right) = \zeta_{n_X+n_Y-2}\left(0, \frac{\left(n_X+n_Y\right)\left(n_X+n_Y-1\right)}{\left(n_X+n_Y-2\right)n_X n_Y}\widehat{\sigma}^2_{cMLE}\right)$$

Then the Conditional Predictive p-value is equal to:

$$p_{cpred} = 2\left(1 - \Upsilon_{n_X+n_Y-2}\left(\frac{\left|t\left(\mathbf{x},\mathbf{y}\right)\right|}{\sqrt{S_p^2}}\right)\right)$$

where $S_p^2 = \left(\frac{1}{n_X}+\frac{1}{n_Y}\right)\left(\frac{n_X S_x^2 + n_Y S_y^2}{n_X+n_Y-2}\right)$ is the pooled sample variance in the classical t-test for the difference of two means under normal assumption with equal variance.

Figure 2.1, shows the differences in the null distribution in the three p-values for two experimental replications. We simulated 1000 samples of $\mathbf{X}$ and $\mathbf{Y}$ of size $n_X = n_Y = 2$ from a standard normal distribution. We can see the only p_{cpred} is

$\mathcal{U}(0,1)$ under the null hypothesis. Obviously it is not necessary to show this by using simulations because $|t(\mathbf{x},\mathbf{y})| / \sqrt{S^2_{pooled}}$ is a pivotal quantity and we know that the p-value from a t-test are frequentist. However, here we want to illustrate the differences among the three p-values using the p_{ppost} (which equals the p_{cpred}) that will be used in the for the Gamma model where we cannot calculate the p_{cpred}. We can further note that p_{plug} is less conservative than the p_{post}. This is due to the double use of the data in the p_{post}, in fact it becomes quite hard to criticize $H = 0$ when we first used the data to calculate the posterior and then to calculated the p-value using the data again in t_{obs}. The effect the double use of the data becomes stronger as the sample size get smaller, in fact the differences between the three p-values are negligible for a sample size larger than ≈ 50. However, in the case of microarray analysis this sample size is practically unreachable in small studies and therefore we have to be careful in choosing the p-values that provides evidence against the null hypothesis.

2.2.2 The Gamma model for single gene expression

As noted above, measured intensity levels (either R and G in cDNA array, or two AD measures in oligonucleotide arrays) depends on signal strength (Chen *et al.*, 1997). On other hands it is convenient to express the measurement error in terms of relative variance (*variability*) rather than in terms of variance. Therefore a quantity worthy of estimations is the common Coefficient of Variation (CV) of the measurements. We take into account this by modelling the two measurements with distinct distributions with the same CV. We found convenient to work with Gamma distributions where the common shape parameter represents the common CV. Formally:

$$\begin{aligned} \mathbf{X} &= X_1, \ldots, X_{n_X} \sim Gamma(a, \theta_X) \quad i.i.d. \\ \mathbf{Y} &= Y_1, \ldots, Y_{n_Y} \sim Gamma(a, \theta_Y) \quad i.i.d. \end{aligned}$$

where $n_X = n_Y = n$. Extension to unbalanced sample cases are also possible, but they have not considered here for seek of simplicity in the notation. Note that in this parametrization we have that $CV_{\mathbf{X}} = CV_{\mathbf{Y}} = 1/\sqrt{a}$. The mean mRNA abundance for measurements $\mathbf{X}$ and $\mathbf{Y}$ is given respectively by $\mathbf{E}(\mathbf{X}) = a\theta_X$ and $\mathbf{E}(\mathbf{Y}) = a\theta_Y$, therefore the null model is

$$H = 0 : \theta_X = \theta_Y = \theta, \forall a > 0$$

where θ represents the common unknown mean. We found useful to work with the ratio of sample means, this leads us to define the following test statistic,

$$T(\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}) = t(\mathbf{x}, \mathbf{y}) = \begin{cases} \frac{\overline{x}}{\overline{y}}, \overline{x} \geq \overline{y} \\ \frac{\overline{y}}{\overline{x}}, \overline{x} \leq \overline{y} \end{cases} .$$

Note that the dual role of $\mathbf{X}$ and $\mathbf{Y}$ in using $T(\mathbf{X}, \mathbf{Y})$, in fact we loose the information of whether the gene is overexpressed in $\mathbf{X}$ or in $\mathbf{Y}$. This information is not probabilistically relevant because we are interested only in measuring the compatibility of

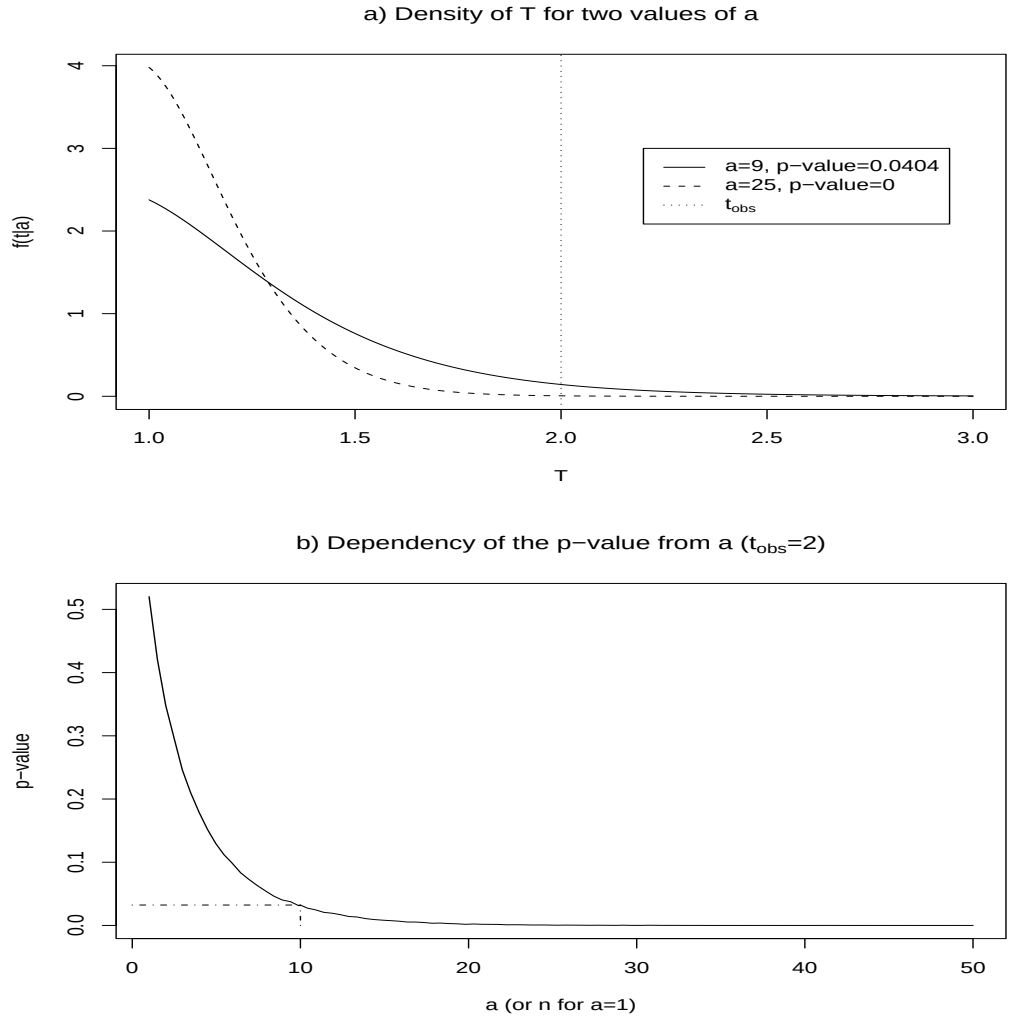

Figure 2.2: The effect of the nuisance parameter a on the distribution of $T = \overline{x}/\overline{y}$ for the gamma model. (a) Distribution of $T > 1$ for two different values of the nuisance parameter a. (b) Dependency of the p-values from a and n (for $a = 1$).

$H = 0$. The null distribution of T is a *Multiple Scaled Beta distribution of II kind* (Kendall and Stuart, 1969)

$$T(\mathbf{X}, \mathbf{Y}) | H = 0 \sim f(t; a) = \frac{2^{2na}\Gamma\left(\frac{1}{2} + na\right)}{\Gamma(na)\sqrt{\pi}} \frac{t^{na-1}}{(1+t)^{2na}}, t \geq 1 \tag{2.9}$$

and a is the nuisance parameter.

Figure 2.2 provides an idea of the dependency of (2.9) from the common CV. In the top of Figure 2.2 we plotted the (2.9) for $a = 9$ and $a = 25$ that correspond respectively to $CV = 1/3$ and $CV = 1/5$. For $T = 2$ and $n = 2$ we have that in the former situation the p-value is ≈ 0.05 while in the latter is ≈ 0.00. The interpretation of this is quite straight forward: it says that by observing a double fold change of the means does provide evidence against the null hypothesis when the variability in the system is sufficiently small. On the contrary, if we observe a double fold change then it does not provide enough evidence against $H = 0$ if the relative variability of the experiment is large. In many microarray data analysis, practitioners used to adopt the

practical rule of consider, as differentially expressed, only those genes that exhibits a double fold change in the means. Therefore, if the considered Gamma model fits the data, this role is not supported by the data.

We can reinforce this conclusion by considering the bottom of Figure 2.2 where we plotted the p-values for different values of a. The gradient of this curve is larger for $a < 10$, moreover $a \in (1, 10)$ is plausible in microarrays analysis, therefore it becomes a critical task estimate the nuisance parameter a. In (2.9) we can see the dual role of a and n, the horizontal axe in the bottom of Figure 2.2 represents also the sample size for $a = 1$. In fact when $a = 1$, the amount of signal equal the amount of noise and we need more than 10 replications to detect that a gene is differentially expressed. It is not surprising that playing with the sample size we may make the signal-to-noise ratio in our favor, but here we provide a numerical quantification. In fact the bottom of Figure 2.2 suggests that when the sample size is small, $n \leq 10$, proper inference on a is critical. For this reason is useful to consider the p_{ppost} as an alternative way to the p_{plug} and p_{post}.

The Plug-in p-value for the Gamma model

Let and $\widehat{a}$ represent the MLE of and a from the Likelihood function under the null model. The p_{plug} is obtained generating T from (2.9) having $a = \widehat{a}$. We can see that under the null hypothesis the (2.9) does not depend on θ and the MLE of θ will not be used. Details are in Algorithm 16 in Appendix.

Posterior p-value for the Gamma model

Two non-informative priors were considered: the Jeffreys's prior and the Reference prior. The inference under the Reference prior provides better results than with the Jeffreys's prior as showed in Liseo (1993). The reference prior is improper, it is given in the following formula

$$\pi(\theta, a) \propto \frac{1}{\theta}\sqrt{\frac{a\psi^{(1)}(a) - 1}{a}} \tag{2.10}$$

where $\psi^{(1)}(a)$ is the trigamma function. The posterior predictive distribution for T is given by the integral

$$\begin{aligned} m_{post}(t|\mathbf{x}, \mathbf{y}) &= \int_{\mathcal{R}^+} \frac{2^{2na}\Gamma\left(\frac{1}{2} + na\right)}{\Gamma(na)\sqrt{\pi}} \frac{t^{na-1}}{(1+t)^{2na}} \frac{\Gamma^2(2na)}{\Gamma^2(na)\Gamma^{2n}(a)} \frac{\prod_{i=1}^{n}(x_i y_i)^{a-1}}{\left[n(\overline{x} + \overline{y})\right]^{2na}} \times \\ &\quad \times \sqrt{\frac{a\psi^{(1)}(a) - 1}{a}} da. \end{aligned}$$

The p_{post} has been approximated with a Monte Carlo sum, by first approximating $\pi(a|\mathbf{x}, \mathbf{y})$ using a Metropolis Hastings algorithm (MH) and then using Algorithm 16 (see Appendix) where the values of a are obtained from the marginal posterior distribution of a, $\pi(a|\mathbf{x}, \mathbf{y})$.

The Partial Posterior Predictive p-value for the Gamma model

In this paragraph we derive the marginal partial posterior distribution for parameter a, $\pi(a|\mathbf{x},\mathbf{y}\backslash t(\mathbf{x},\mathbf{y}))$. As far as we know this distribution has been never derived before. We provide a simulation algorithm and we demonstrate that this is a proper probability distribution. $\pi(a|\mathbf{x},\mathbf{y}\backslash t(\mathbf{x},\mathbf{y}))$ is proportional to

$$\pi(a|\mathbf{x},\mathbf{y}\backslash t(\mathbf{x},\mathbf{y})) \propto \frac{\Gamma^2(2na)}{\Gamma^2(na)\Gamma^{2n}(a)} \frac{\prod_{i=1}^{n}(x_i y_i)^{a-1}}{[n(\overline{x}+\overline{y})]^{2na}} \sqrt{\frac{a\psi^{(1)}(a)-1}{a}} \times \tag{2.11}$$
$$\times \left(\frac{2^{2na}\Gamma\left(\frac{1}{2}+na\right)}{\Gamma(na)\sqrt{\pi}} \frac{t(\mathbf{x},\mathbf{y})^{na-1}}{(1+t(\mathbf{x},\mathbf{y}))^{2na}} \right)^{-1}$$

and its approximation has been obtained in the same way as for the posterior, except that the proposal distribution has been set up according to the kernel of the partial distribution $\pi(a|\mathbf{x},\mathbf{y}\backslash t(\mathbf{x},\mathbf{y}))$ instead of the kernel of the posterior $\pi(a|\mathbf{x},\mathbf{y})$. We remark that because we have to run thousand of Markov chains, it is not feasible to check the behavior of each chain and it is necessary to have an automatic choice of the proposal distribution. The one adopted here, produces a chain that behaves acceptably as shown in the Appendix.

We showed the following proposition (3) that assures a proper inference using the marginal partial posterior distribution for a.

Proposition 3 *For $n \geq 2$ the (2.11) is a proper distribution.*

Proof. Detailed proof is showed in Appendix. The proof is based on the analysis of $\pi(a|\mathbf{x},\mathbf{y}\backslash t(\mathbf{x},\mathbf{y}))$ for $a \to \infty$ and $a \to 0^+$. For $a \to \infty$ the proof is based on some considerations about the geometric and arithmetic means of $\mathbf{x}$ and $\mathbf{y}$. ■

Figure 2.3 shows a typical inference using the three p-values under the alternative model $\mathbf{X} \sim N(3, \sigma = 3/2)$, $\mathbf{Y} \sim N(1, \sigma = 1/2)$ with $n = 3$ replications. This model simply reproduce the fact the two sets of measurements for $\mathbf{X}$ and $\mathbf{Y}$ have the same CV but the two genes are differentially expressed (the gene in population X is overexpressed). We considered a set of simulated data where no negative numbers appeared (of course it is not worthy to check the compatibility of negative observations with a gamma model), and the three p-values are: $p_{plug} = 0.021$, $p_{post} = 0.085$ and $p_{ppost} = 0.009$. As expected, the partial posterior predictive p-value provides more evidence against the null hypothesis and it may allows us to detect that the gene is actually differentially expressed.

Considering the top of Figure 2.3 we see why the p_{ppost} provides more evidence against the null model. We know that the true value of a is 4 and only the partial posterior predictive distribution, $\pi(a|\mathbf{x},\mathbf{y}\backslash t(\mathbf{x},\mathbf{y}))$, has the mean near to this value so providing larger density around 4 than the posterior distribution $\pi(a|\mathbf{x},\mathbf{y})$. The variance of $\pi(a|\mathbf{x},\mathbf{y}\backslash t(\mathbf{x},\mathbf{y}))$ is larger than the variance of $\pi(a|\mathbf{x},\mathbf{y})$ because the latter takes into account only the variability of the parameter while the former also the variability of T. The partial posterior distributions $\pi(a|\mathbf{x},\mathbf{y}\backslash t(\mathbf{x},\mathbf{y}))$ learns more about a then both the posterior distribution and the MLE. The result of this is to

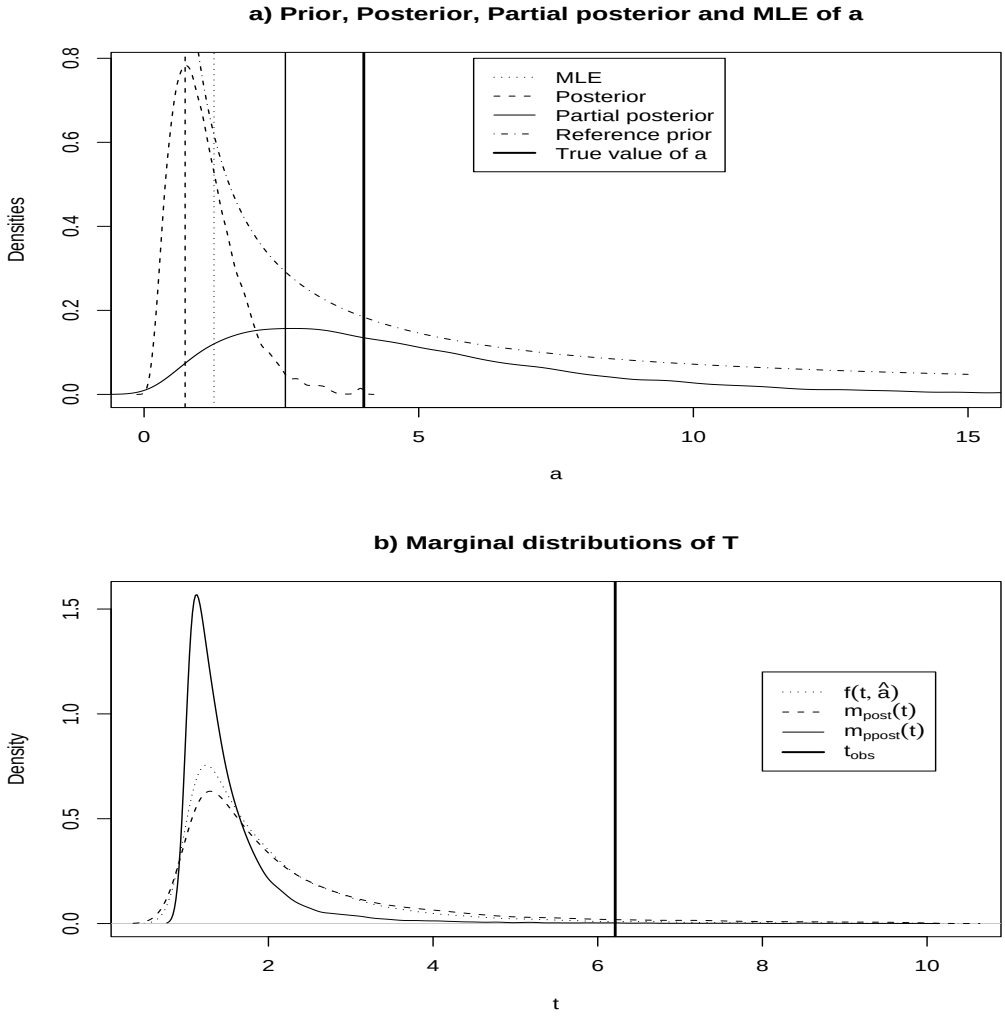

Figure 2.3: Gamma model: typical inference with the three p-values under an alternative model: $\mathbf{X} \sim N(3, \sigma = 3/2)$,$\mathbf{Y} \sim N(1, \sigma = 1/2)$ and $n = 3$, so that the common variation coefficient $a = 2$. (a) The reference prior, the posterior and partial posterior are showed together with their modes (Vertical lines). The bold vertical line is the true value of a. (b) The marginal distributions of T according to the different ways of eliminating a. The bold vertical line is the observed value of T for the simulated data set.

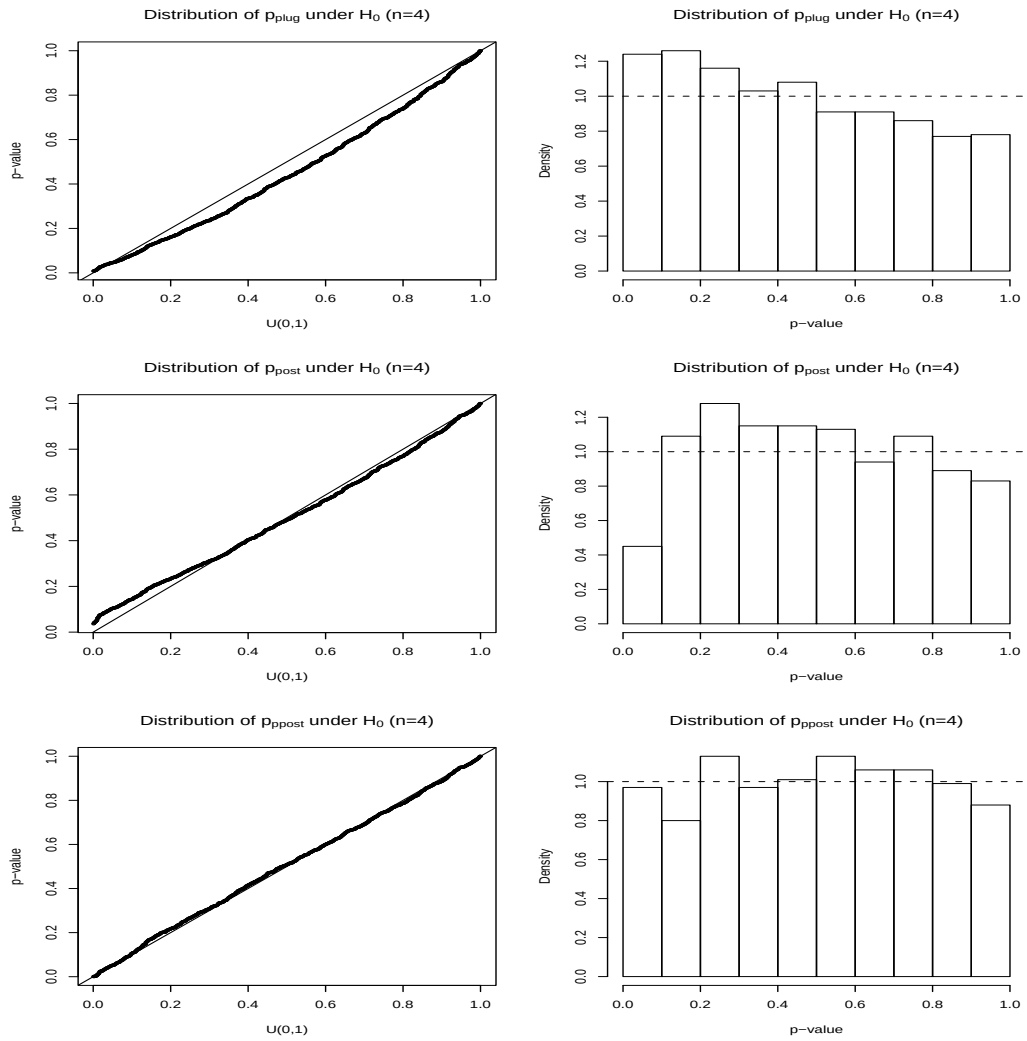

Figure 2.4: Distribution of the p-values under the null model: $\mathbf{X}, \mathbf{Y} \sim Gamma\,(2,1)$ and $n = 4$. We can see that only the p_{ppost} is approximately distributed $\mathcal{U}\,(0,1)$.

consider the observed value of T very rare under $H = 0$ as shown in the bottom of Figure 2.3. As we noted above, other measures of surprise can be considered in order to carry out the test. Bayarri and Castellanos (2001) recommended that these should be based on the partial posterior distribution or the conditional predictive distribution when available.

In the end the p_{ppost} is a p-value asymptotically uniform distributed under the null hypothesis. Also the other p-values are asymptotically uniform distributed under the null hypothesis, but for a fixed n the error in approximating the null distribution with $\mathcal{U}\,(0,1)$ is much less for p_{ppost} than for the other p-values as shown in Figure 2.4. We produced these quantile-quantile plots (in the sequel QQ-plots) by simulating 1000 p-values under the null hypothesis with $n = 4$ replicates of $\mathbf{X}, \mathbf{Y} \sim Gamma\,(2,1)$. We can see that with only 4 replicates p_{ppost} behaves like a $\mathcal{U}\,(0,1)$ random variables while the p_{post} is a conservative p-value and the p_{plug} is anticonservative. Considering the sample size that we can use to make proper inference, that is $n = 2$, we can see that the p_{ppost} tends to behave in a conservative way (see top of Figure 2.5), while with

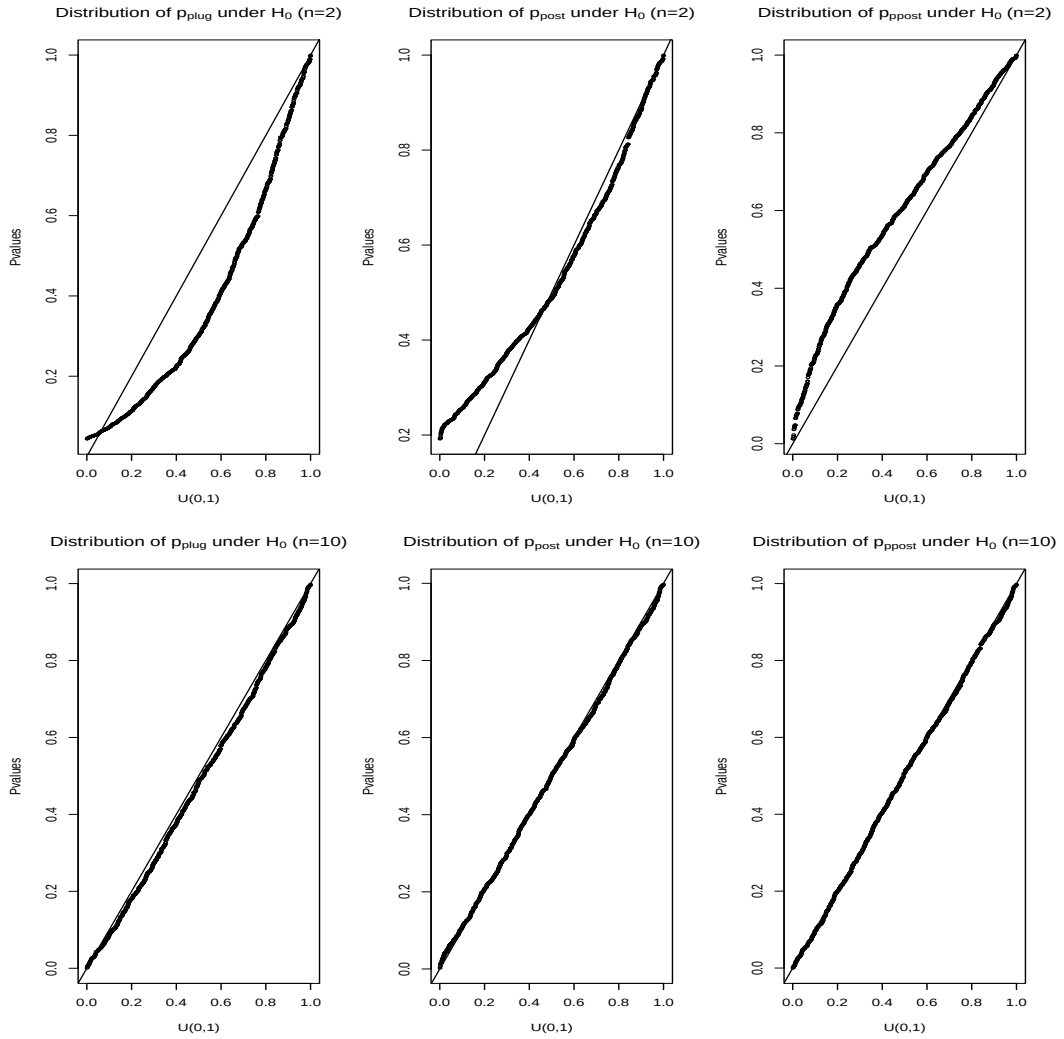

Figure 2.5: Distribution of the p-values under the null model: $\mathbf{X}, \mathbf{Y} \sim Gamma\,(2,1)$ and $n = 2$ (first row) and $n = 10$ (second row). We can see that with a large sample size all p-values are approximately distributed $\mathcal{U}\,(0,1)$. With very small sample sizes the p_{plug} is anticonservative, the p_{post} is very conservative and the p_{ppost} tend to be conservative.

$n = 10$ replicates we find that all three p-values are approximately $\mathcal{U}\,(0,1)$ as shown in the bottom of Figure 2.5. The differences between the three p-values may not seem to be very large, but when we use the p-values in MHT they become relevant. Intuitively this is due to the fact that we are drawing inference from thousand of p-values and the compound error arising from each test determines the results of the analysis.

2.3 Multiple Hypothesis Testing (MHT)

The literature on MHT has experienced an increase in the recent years. Here we try to summarize those topics that are relevant for this thesis. We briefly discuss the most important issues indicating the relevant bibliography. For the reasons indicated above the MHT techniques based on non-parametric methods will be marginally considered in this work.

	Not-Reject	Reject	Total
Null True	V'	V	m_0
Alternative True	O	L	m_1
Total	W	R	m

Table 2.1: Outcomes in testing m hypotheses, based on Table 1 of Benjamini and Hochberg (1995)

We remember here the notation: with H_i we indicate the state of the hypothesis i on gene i, $H_i = 0$ if the null hypothesis is true, otherwise $H_i = 1$. We will always assume H_i a random variable and its probability will explicitly specified. In this way we are legitimate to write conditional probabilities on H_i.

2.3.1 Compounds error measures for MHT

When testing multiple hypotheses (MHT) the statistical setup becomes much more complicated. Each test has is own Type I and Type II errors, and it becomes unclear how one should measure the overall error rate. Specifically, let's consider Table 2.1 which lists the possible outcomes when testing m hypothesis simultaneously. We suppose there exists an unknown subset of true null hypothesis, $\mathcal{M}_0 = \{i : H_i = 0\}$ the cardinality of $\mathcal{M}_0$ is m_0, with possibly $m_0 = 0$. The m hypotheses are assumed to be known in advance. R is an observable random variable, while V', V, O and L are unobservable.

For example, V is the number of Type I errors (false positives), O is the number of Type II errors (false negatives), and $R = V + L$ is the total number of significant hypothesis. In parallel with single hypothesis testing the notion of power is approximately given by the ratio L/R, again here we look for the most powerful testing procedure, that is a procedure that maximize L/R while bounding V.

In order to measure the errors occurred in MHT, it is convenient to define a compound error measure Err. We consider here six Err that are common to the MHT literature:

i) Per Comparison Error Rate: $PCER = \mathbf{E}(V)/m$;

ii) Family Wise (also called experiment-wise) Error Rate: $FWER = \Pr(V \geq 1)$;

iii) False Discovery Rate: $FDR = \mathbf{E}(V/R|R > 0)\Pr(R > 0)$;

iv) False Non-rejection Rate: $FNR = \mathbf{E}(O/W|W > 0)\Pr(W > 0)$;

v) Positive False Discovery Rate: $pFDR = \mathbf{E}(V/R|R > 0)$ and Positive False Non-rejection Rate $\mathbf{E}(O/W|W > 0)$;

vi) False Discovery Proportion: $FDP = \Pr(V/R > \gamma)$.

Note that all expectations and probabilities are conditioned on $\mathcal{M}_0$.

The $PCER$ and the $FWER$ have been used for many years, but the FDR was relative recently introduced by Benjamini and Hochberg (1995). The FNR was introduced by Genovese and Wasserman (2001). The *positive* versions of these compound error measures were introduced by Storey (2002, 2003).

The FDP was introduced by Korn *et al.* (2001).

Even if it may have no meaning to compare this error measure because each one has a different interpretation, we numerically expect that the following inequality hold in general (Ge *et al.*, 2003):

$$PCER \leq FDR \leq pFDR \leq FWER. \tag{2.12}$$

We reported (2.12) because it is very common in literature to fix a significance level and then compared the result of multiple testing across all Err definitions.

For each of these measures it has been studied a decision procedure on the m hypothesis under test that is focused on controlling a particular error measure. The main goal of each procedure is to gain in power while controlling the error rate at some fixed level α: $Err \leq \alpha$.

There exists different types of controls on Err. We say that an algorithm *strongly controls* the error rate if it is controlled for all values of m_0 simultaneously. In other words it is not necessary to include m_0 as an argument in the algorithm. An algorithm *weakly controls* the error rate when it only holds for $m_0 = m$; that is, when all null hypothesis are assumed true. Given the two types of control, Shaffer (1995) pointed out that in MHT literature strong control is usually preferred over weak control, simply because it implies weak control and it is adaptive over all possible values of m_0. We further note that the expectations and probabilities for the error rates $i)$ - $v)$ are calculated with respect to the cardinality of $\mathcal{M}_0$. We say that a procedure has an *exact control* of Err if it is conditioned on $\mathcal{M}_0$ (and not m_0). In general strong control implies exact control, but neither of weak control and exact control implies the other. In the microarray setting, where it is often unlikely that none of the genes is differentially expressed, it seems that weak control without any other safeguards is unsatisfactory, and it is important to have exact or strong control. The advantage of exact control is higher power, and we will show that we are more likely to obtain exact control by using the p_{ppost} and p_{cpred}.

We can consider MHT algorithms to be carried out on the ordered p-values for a given set of rejection regions Γ_α. The MHT literature has mostly been concerned with deriving algorithms based on the order statistics. As stated at the very beginning, instead of using p-values, we could use other measures of evidence if the choice leads to a nested set of rejection regions Γ_α and allows one-to-one mapping between Γ and α. The investigation of the behavior of other measures of evidence (or surprise) is behind the scope of this thesis, and the reader is referred to the originals works of Good (1953, 1956) recently commented by Bayarri and Berger (1997, 1999). The problem with using other measure of evidence is that most of them are not calibrated, i.e. they are not in the error scale such as the frequentist p-values. However, the relevant goal is to classify the hypotheses $H_1, H_2, \ldots, H_m$ based on respectively p-values $p_1, p_2, \ldots, p_m$

and another more general way to achieve this goal is to consider the MHT problem as missing label problems (Genovese and Wasserman, 2001) where we have to estimate the label (0 or 1) on each observed p-values. This more general point of view may lead to consider other MHT procedures which could be not based on ordered p-values.

We will refer to the algorithms based on ordered p-values as *sequential p-values methods*. This is how a sequential p-value method works: using the observed data, it estimates the rejection regions for all hypothesis, so that on average $Err \leq \alpha$ for some pre-chosen α. The result of a sequential p-value method is an estimate $\widehat{k}$ that indicates to reject the null hypotheses corresponding to $p_{(1)}, p_{(2)}, \ldots, p_{\left(\widehat{k}\right)}$, where $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$ are the ordered observed p-values.

To further illustrate the general goals in MHT and the most relevant results obtained from using different controlling procedure, we use Figure 2.6. Here test $m = 50$ null hypotheses $H_i = 0 : X \sim N(1, 1)$ for $i = 1, \ldots, m$. We simulated 25 observations from the null model and 25 from an alternative model $N(3, 1)$. The corresponding frequentist p-value from H_i is $p_i = 1 - \Pr^{N(1,1)}(X > x_i)$. It is reasonable to suspect that small p-values may come from the alternative hypothesis (`P` and `b` points). With respect to this intuition, it is convenient to consider the ordered sequence of p-values as showed in Figure 2.6, where are all p-values are plotted against their rank order. In this work we limit to consider six procedures that control $Err = \alpha$, three are aimed to control the $FWER$, and the others control the FDR and $pFDR$.

In order to simplify the notation, we use α to indicate the amount of Err that we are bearing with while controlling some of the previous compound error measure. However, each error measure has its own interpretation and it is not fair to compare all the error measures by using the same level α. We will do this having in mind that we are dealing with different procedures that are aimed to control different definitions of Err. Moreover in MHT there not exists any kind of "conventional level" α like the 0.05 (or 0.01) as in single hypothesis testing. In the end, the choice of some reference level α, should be done in a more broadly decision theory framework, for example, by assigning an utility function to α and to the power of the test. This is what is done in Genovese and Wasserman (2001) where they considered to control both the FDR and FNR.

It is important to note that if we choose a set of hypotheses to be rejected, $\widehat{\mathcal{M}}_0$, then the corresponding Err refers to the whole set and cannot be pointed to any of the single hypothesis belonging to $\widehat{\mathcal{M}}_0$. This is due to the fact that we are taking into account the multiplicity and considering the outcome from Table 2.1 we are, in some sense, loosing information on single tests. This point was suggested by Glonek and Solomon in the discussion of the paper of Ge *et al.* (2003). They used a numerical example on the $pFDR$ in order to show that if we reject a set of hypotheses for $pFDR = \alpha$ then the $pFDR$ we are bearing with is larger. The claim there was made only for the $pFDR$ but it applies in general to the p-values. As stated above, a p-value defines a rejection region of the form $P > p$, but once we observed p, then we can only decide to reject $H = 0$ based on the equality $P = p$ which is not the same rejection region as $P > p$. The result of these point, applied to MHT, has been showed by Glonek and

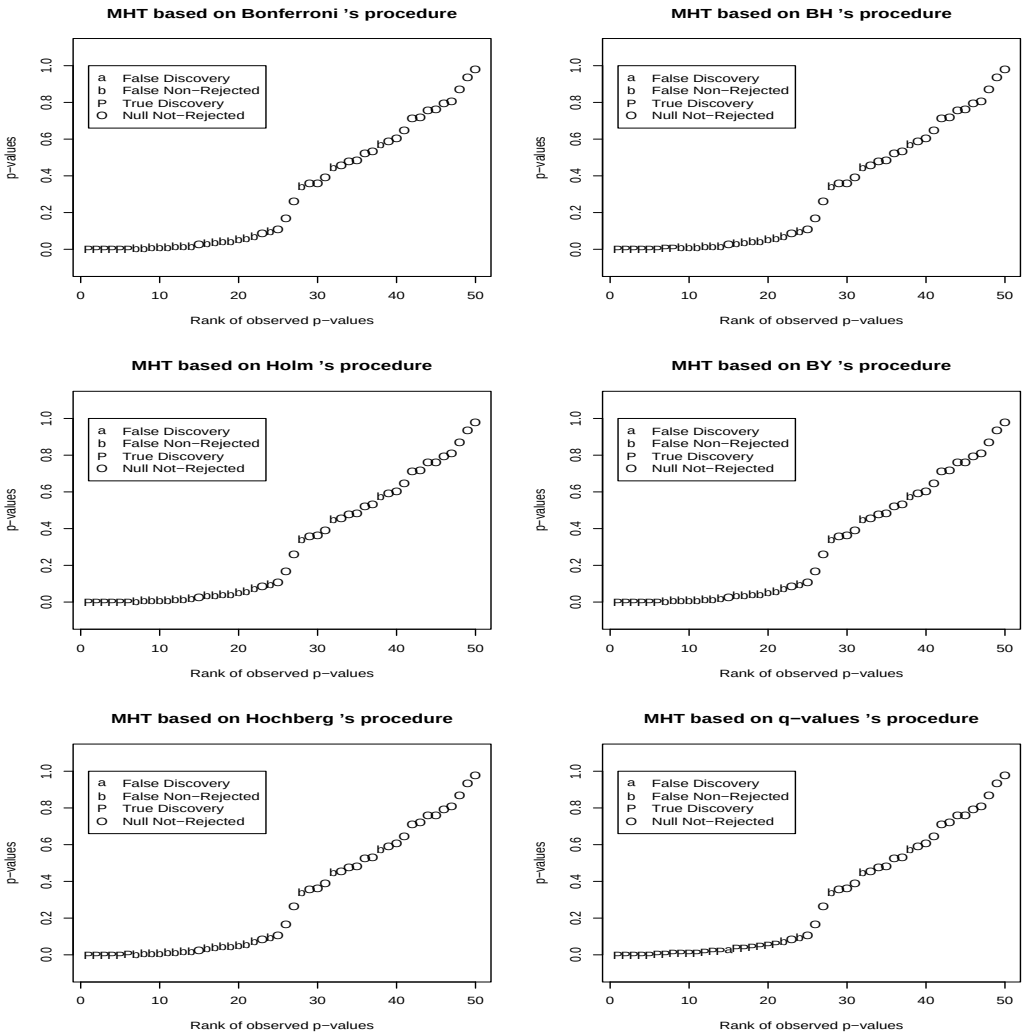

Figure 2.6: Hypothesis testing with ordered p-values. Lower case letters represent error: (a) $H = 0$ and null hypothesis rejected, (b) $H = 1$ and null hypothesis not rejected; Capital case represent success: (P) $H = 1$ and null hypothesis rejected and (O) $H = 0$ and null hypothesis not rejected. Therefore (P) represents the power of the procedure, (a) the False Discovery Rate and (b) the False Non-Rejection Rate.

Solomon, but it was yet known since the work in Selke, Bayarri and Berger (2001) and it is curious that Glonek and Solomon used almost the same numerical investigation as in Selke, Bayarri and Berger (2001).

Summing up, in MHT is relevant only the frequentist interpretation of the p-values.

2.3.2 Procedures that control compound error measures

In literature we encounter two different classes of MHT procedures:

a) *single-step procedures*: equivalent multiplicity adjustments are performed for all hypotheses, regardless of the ordering of the test statistics or raw p-values;

b) *stepwise procedures* in which the rejection of particular hypothesis is based not only on the total number of hypotheses, but also on the outcome of the tests of other hypotheses. Step-down procedures order the raw p-values starting with the most significant, while step-up procedures start from the least significant.

We consider here six procedures that control the Err. The first and the last are single-step procedures while the others are step-up procedures.

We consider three procedures that control the $FWER$:

i) Bonferroni method: $\widehat{k} = \max\left\{k : p_{(k)} \leq \alpha/m\right\}$,

ii) Holm procedure (1979): $\widehat{k} = \max\left\{k : p_{(j)} \leq \alpha/\left(m-j+1\right) \text{ for } j \leq k\right\}$,

iii) Hochberg procedure (1988): $\widehat{k} = \max\left\{k : p_{(k)} \leq \alpha/\left(m-k+1\right)\right\}$.

Simes (1986) developed a procedure that weakly control the $FWER$. This is important in this thesis and we summarize it in Algorithm 4.

Algorithm 4 *(Simes, 1986 - BH procedure).*
1. Let $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$ *be the ordered observed p-values.*
2. Calculate

$$\widehat{k} = \max\left\{k : p_{(k)} \leq \alpha \cdot k/m\right\}. \tag{2.13}$$

3. Reject the null hypothesis corresponding to $p_{(1)}, p_{(2)}, \ldots, p_{\left(\widehat{k}\right)}$.

We consider two procedures that control the FDR:

iv) the Benjamini Hochberg (BH) procedure (1995): throughout this work, we will refer to Algorithm 4 as the BH procedure, mainly because we are interested in the fact that it strongly controls the FDR as was originally proved in Benjamini and Hochberg (1995). This algorithm was initially studied by Seeger (1968), who conjectured that, when $m_0 = m$, Algorithm 4 provides weak control of the $FWER$. The formal proof of the conjecture was given in Simes (1986). The proof that Algorithm 4 strongly control the FDR was given by Benjamini and Hochberg using an induction argument, while the latest proof using martingales was given in Storey (2003);

v) the Benjamini and Yuketieli (BY) procedure (2001) substitutes the (2.13) with

$$\widehat{k} = \max\left\{k : p_{(k)} \leq \alpha \cdot k / \sum_{i=1}^{m} 1/i\right\}. \tag{2.14}$$

Sometimes it is convenient to consider these five procedures in terms of *adjusted p-values*, because raw p-values obtained from m tests are adjusted (or calibrated) to control a particular Err. More formally, let $\widetilde{p}_{(i)}$ be the adjusted p-value of a single hypothesis H_i, for a particular adjusting procedure that control the Err at level α we have

$$\widetilde{p}_i = \inf\{\alpha : H_i \text{ is rejected at level } Err = \alpha\}.$$

For example, consider the Bonferroni procedure, the adjusted p-values are the raw p-values divided by the number of tests: $\widetilde{p}_{(i)} = p_{(i)}/m$. Assuming that the raw p-values are $\mathcal{U}(0,1)$ under the null hypothesis, that is $p_{(i)} = \alpha$, then rejecting all hypotheses with $\widetilde{p}_{(i)} \leq \alpha$ assures that we are bearing with a $FWER \leq \alpha$. If the raw p-values were conservative than we are bearing with $FWER < \alpha$ and if they were anticonservative then we have no control of the $FWER$. These considerations apply to all other Err controlling procedures, so it is necessary to have frequentist p-values for the MHT.

Finally, we consider the control of the $pFDR$ by using the q-values that have been introduced by Storey (2003). Let q_i be the corresponding q-values for the p-value p_i. Storey (2003) pointed out that, differently from procedures, the q-values are not adjusted p-values, but this point is not shared by other authors (Ge *et al.*, 2003). q_i provides an estimation of $pFDR$ that we are bearing in rejecting all hypotheses with q-values less than q_i. This is our sixth procedure which also need frequentist p-values as we will show later.

We mention here a procedure that strongly controls the $PCER$ at level α. It simply compare each p-value with α: $\widehat{k} = \max\{k : p_{(k)} \leq \alpha\}$. Because we are not interested in controlling the $PCER$ in microarray analysis, we will no longer consider this procedure.

2.3.3 The comparison of several control procedures

We numerically compare the performance for the six MHT procedures introduced above by considering Figure 2.6. In the simulated data set we know whether each p-value came from the null or alternative model, therefore for a level $\alpha = 0.05$, we can state whether we made a False Discovery (`a` points), a True Discovery (`P` points), a False Non-rejection (`b` points) or when the null has been not rejected (`O` points). So `P` indicates the power, `a` the amount of FDR and `b` the amount of FNR. The three procedures, which control the $FWER$ are in the first column of plots. We can note that procedures *i*) trough *iii*) are very conservative: it is unlikely to have False Discoveries, but the power is smaller than with the other procedures and the FNR is higher. The practical difference between controlling $FWER$ or FDR is neither trivial nor small and the larger m the more dramatic the difference is.

Figure 2.6 shows that in controlling the FDR or $pFDR$ it is likely to have false discoveries, but the power is significantly higher for BH procedure and q-values, while

the BY procedure is quite conservative because it deals with a form of dependent hypotheses which is not needed in this simulated data. Moreover, we can note that controlling $pFDR$ with q-values provides more power than with the BH procedure, in fact we can see that with q-values and $\alpha = 0.05$ we make only 1 false rejection out of 20 rejections.

Comparing the behavior of all procedures by using different notions of p-value is the goal of the thesis, but before doing this, it is important to formally compare the procedures and to provide their operating characteristics when using frequentist p-values. In particular we will concentrate on describing the operational characteristics of the BH procedure and q-values. We will then introduce the problem of dependence tests and then characterize the BY procedure and the q-values when the tests are dependent.

The Bonferroni's procedure

Before to provide details about the relative most recent procedures it is worthy to make some comments about the Bonferroni procedure which is the oldest among the considered procedures. The Bonferroni procedure is based on the Bonferroni inequality again based on the Boole's inequality (2.15). Let Err represents the control level in the Bonferroni procedure and let $\mathcal{M}_0 = \mathcal{M}$ where $\mathcal{M}$ is the set of all hypothesis we have that the weak control of the $FWER$ is obtained by the following development

$$\begin{aligned} FWER &= \Pr\left(V > 0\right) \\ &= \Pr\left(\overline{\bigcap_{i=1}^{m} H_i = 0} \middle| \mathcal{M}\right) \\ &= \Pr\left(\bigcup_{i=1}^{m} H_i = 1 \middle| \mathcal{M}\right) \\ &= \sum_{i=1}^{m} \Pr\left(H_i = 1 | \mathcal{M}\right) - \Pr\left(\bigcap_{i=1}^{m} H_i = 1 \middle| \mathcal{M}\right) \\ &\leq \sum_{i=1}^{m} \frac{Err}{m} \end{aligned} \tag{2.15}$$

Using the Bonferroni procedure we are not making any assumption on the dependency structure among the hypotheses, because we do not compute the term $\Pr\left(\bigcap_{i=1}^{m} H_i = 0 | \mathcal{M}\right)$. The Bonferroni procedure is the most conservative procedures among all MHT procedures.

The Benjamini and Hochberg procedure

The FDR criterion, and the BH procedure that controls it, has been used successfully in some problems: tresholding of wavelets coefficients (Abramovich and Benjamini, 1996), studying weather maps (Yuketieli and Benjamini, 1999), and multiple trait location in genetics (Weller *et al.*, 1998). Their approach calls for controlling FDR at a desired level α, while maximizing $\mathbf{E}\left(R\right)$. Two properties of this error rate are easy to be shown (Benjamini and Hochberg, 1995):

i) if all null hypotheses are true, then FDR is equivalent to the $FWER$. So, when $m_0 = m$ the control of the FDR implies the control of the $FWER$ in the weak sense;

ii) when $m_0 < m$, the FDR is smaller than or equal to the $FWER$, therefore any procedure that control the $FWER$ also control the FDR. However, if a procedure control the FDR only, it can be less stringent, and a gain in power maybe expected. In particular, the larger the number of non-true null hypotheses is, (the larger L tend to be), so is the differences between the $FWER$ and FDR. As a result, the potential for increasing the power is larger when more of the hypotheses are non-true. This is the case in microarray analysis;

iii) another attractive feature of the FDR criterion is that if it is controlled separately in several families of hypotheses at some level, then it is also controlled at the same level in the whole family (as long as the families are large enough, and do not consist only of true null hypotheses).

It is worthy to examine the relationship between Hochberg's procedure and the BH procedure. Both Hochberg and BH procedure are step-up procedures, which start by comparing $p_{(m)}$ with α, and if $p_{(m)} < \alpha$ then all hypotheses are rejected (as if a $PCER$ approach had been taken), otherwise if $p_{(m)} > \alpha$ proceed to smaller p-values until one satisfies the condition (2.13). The procedure ends, if not terminate earlier, by comparing $p_{(1)}$ with α/m, as a pure Bonferroni comparison. At the two ends the Hochberg and BH are similar, but in between the sequence of $p_{(i)}$is compared with $\alpha\left(1-\left(i-1\right)/m\right)$ in the BH procedure, rather then with $\alpha/\left(m+1-i\right)$ in the Hochberg's procedure. The series of linearly decreasing constants of the BH method is always larger than the hyperbolically decreasing constant of Hochberg's procedure and the extreme ratio is as large as $4m/\left(m+1\right)^2$ at $i = \left(m+1\right)/2$. This imply that the BH procedure rejects samplewise at least as many hypotheses as the Hochberg's method and therefore the BH has also greater power than other $FWER$ controlling methods such as the Holm's procedure. The Hochberg (1988) procedure was suggested as a different way to use Simes's procedure so that it does control the $FWER$ in the strong sense.

Genovese and Wasserman (2001) investigated the operating characteristics of the BH procedure. They achieved the following conclusion: asymptotically, the BH procedure correspond to reject $H_i = 0$ when the raw p-value $p_i \leq u^*$, where u^* is the solution to the equation

$$G_1\left(u^*\right) = \frac{\left(\frac{1}{\alpha} - \frac{m_0}{m}\right)}{\left(1 - \frac{m_0}{m}\right)} u^*,$$

where G_1 is the common distribution of the p-value under the alternative model. Furthermore Genovese and Wasserman (2001) showed that

$$\alpha/m \leq u^* \leq \alpha,$$

so BH procedure is intermediate between Bonferroni (corresponding to α/m) and uncorrected testing (corresponding to α).

The control of FDR can be problematic when $R = 0$ (when $m = m_0$), on this purpose Benjamini and Hochberg (1995) defined the ratio $Q = V/(V + L) = 0$ when $R = 0$, as no error of false rejection can be committed. Q is an unobserved (unknown) random variable, because we do not observe the realizations v and l, and thus $q = v/(v + l)$, even after experimentation and data analysis. Undoubtedly, controlling the random variable Q at each realization is most desirable. This is impossible, for example, if $m = m_0$ and even if a single hypothesis is rejected $v/r = 1$ and Q cannot be controlled. Controlling $(V/R|R > 0)$ has the same problem - it is identically 1 in the above configuration. For this problem, let's consider another formulation given by Soriç (1989): the proportion of false discoveries among the discoveries $Q' = \mathbf{E}(V)/r$. This quotient is neither the random variable Q nor its expectation but a mixture of expectation and realizations. It is not the conditional expectation of Q, namely $\mathbf{E}(Q|R = r) = \mathbf{E}(V|R = r)/r$, which have again the problem of control for $m_0 = m$. Third, consider $Q'' = \mathbf{E}(V)/\mathbf{E}(R)$ and when all hypotheses are true $Q'' = 1$ and again it is impossible to control. A remedy may be given by either adding 1 to the denominator, but this is a somewhat artificial solution, or another solution could be to change the denominator to $\mathbf{E}(R|R > 0)$. Modifying both numerator and denominator in the same way will again run into problems of control when $m_0 = m$. The FDR, instead, is $\Pr(R > 0)\,\mathbf{E}(V/R|R > 0)$ and this is possible to control, but when $\Pr(R > 0)$ is very small and we write the controlling equation as

$$\frac{FDR}{\Pr(R > 0)} = \mathbf{E}(V/R|R > 0)$$

then the quantity in which we are mainly interested, $\mathbf{E}(V/R|R > 0)$, is controlled at higher level than FDR. The additional term $\Pr(R > 0) \to 1$ as $m \to \infty$, therefore when m is small then $\Pr(R > 0)$ could also be too small and the control of FDR can be problematic and obfuscate its interpretation. This motivated Storey (2003) to introduce the notion of $pFDR$ and the q-values. In order to take into account $\Pr(R > 0)$ Benjamini and Hochberg (2000) introduced a new procedure based on Algorithm 4, but instead of computing (2.13) compute

$$\widehat{k} = \max\left\{k : p_{(k)} \leq \alpha/\eta_0 \cdot k/m\right\}$$

where η_0 is an estimator of the fraction m_0/m. In the Benjamini and Hochberg (2000) η_0 is estimated adaptively. In this thesis we will consider the original work of Benjamini and Hochberg (1995) where $\eta_0 = 1$.

Controlling the positive false discovery rate with the q-values

The $pFDR$ has a Bayesian interpretation. It can be written as a posterior probability of making a false discovery once we observed the data. Suppose we wish to perform m identical tests of a null hypothesis based on statistics $T_1, T_2, \ldots, T_m$. For a given rejection region Γ, define the positive False Discovery Rate as previously done:

$$pFDR(\Gamma) = \mathbf{E}\left(\frac{V(\Gamma)}{R(\Gamma)}\middle| R(\Gamma) > 0\right)$$

where $V(\Gamma) = \#\{T_i \text{ under } H_i = 0 : T_i \in \Gamma\}$ and $R(\Gamma) = \#\{T_i : T_i \in \Gamma\}$. Suppose H_i is a random variable $H_i = 0$ when the i^{th} null hypothesis is true and $H_i = 1$ when the alternative is true, $i = 1, \ldots, m$. Let π_0 be the priori probability that the null hypothesis is true. That is, we assume that H_i are *i.i.d.* Bernoulli random variables with $\Pr(H_i = 0) = \pi_0$ and $\Pr(H_i = 1) = 1 - \pi_0 = \pi_1$. Storey (2003) showed the following Theorem 5.

Theorem 5 *(Storey, 2003). Suppose m identical hypothesis tests are performed with the statistics $T_1, \ldots, T_m$ and rejection region Γ. Assume that $T_i|H_i \overset{i.i.d.}{\sim} (1 - H_i) \cdot F_0 + H_i \cdot F_1$ for null distribution F_0 and alternative distribution F_1 and assume $H_i \overset{i.i.d.}{\sim} Bernoulli(\pi_1)$ for $i = 1, \ldots, m$. Then*

$$pFDR(\Gamma) = \Pr(H = 0|T \in \Gamma) \tag{2.16}$$

where $\pi_0 = 1 - \pi_1$ is the implicit prior probability used in the above posterior probability.

Note the following facts (Storey, 2003):

i) posterior probability (2.16) does not depend on m;

ii) $\Pr(H_i = 0|T_i \in \Gamma)$ is the same for each i and for this reason we left out the index in the statement of the theorem;

iii) we can explicitly write

$$\begin{aligned} pFDR(\Gamma) &= \Pr(H = 0|T \in \Gamma) \\ &= \frac{\pi_0 \Pr(T \in \Gamma|H = 0)}{\pi_0 \Pr(T \in \Gamma|H = 0) + \pi_1 \Pr(T \in \Gamma|H = 1)} \\ &= \frac{\pi_0 \{\text{Type I error of } \Gamma\}}{\pi_0 \{\text{Type I error of } \Gamma\} + \pi_1 \{\text{Power of } \Gamma\}} \end{aligned}$$

this shows that the $pFDR$ increases with increasing Type I error and decreases with increasing power.

Note that if the H_i were not random, then this theorem no longer holds since there is the deterministic constraint that $\sum_{i=1}^{m} H_i = m_1$. However, the mixture distribution assumption, $T_i|H_i \overset{i.i.d.}{\sim} (1 - H_i) \cdot F_0 + H_i \cdot F_1$ avoid this problem, and this assumption generally holds for large m (Storey, 2002).

Two corollaries easily follow from Theorem 5:

Corollary 6 *(Storey, 2003). Under the assumptions of Theorem 5, for $k > 0$ we have*

$$\mathbf{E}\left(\frac{V(\Gamma)}{R(\Gamma)}\middle| R(\Gamma) = k\right) = \Pr(H = 0|T \in \Gamma)$$

and the Bayesian interpretation of $pFDR$ holds for all $k > 0$.

Corollary 7 *(Storey, 2003). Under the assumptions of Theorem 5,*

$$\mathbf{E}\left(\frac{V(\Gamma)}{R(\Gamma)}\middle| R(\Gamma) > 0\right) = \frac{\mathbf{E}[V(\Gamma)]}{\mathbf{E}[R(\Gamma)]}.$$

$pFDR(\Gamma) = \Pr(H = 0|T \in \Gamma)$ gives a global measure in which it doesn't provide specific information about the value of each statistic: only whether it falls in Γ or not.

The measure of significance with respect to the $pFDR$, for each test statistic, is called the q-value. This continues to have a Bayesian interpretation, allowing one to make simultaneous inferences.

Definition 8 *(Storey, 2003). For an observed statistic $T = t$ define the q-value of t to be:*

$$\begin{aligned} q\text{-}value(t) &= \inf_{\{\Gamma_a : t \in \Gamma_\alpha\}} pFDR(\Gamma_\alpha) \\ &= \inf_{\{\Gamma_a : t \in \Gamma_\alpha\}} \Pr(H = 0|T \in \Gamma_\alpha) \\ &= \frac{\pi_0 \Pr(T \geq t|H = 0)}{\pi_0 \Pr(T \geq t|H = 0) + (1 - \pi_0) \Pr(T \geq t|H = 1)} \end{aligned}$$

The q-value is a measure of strength of an observed statistic with respect to the $pFDR$. It is the minimum $pFDR$ which can occur when rejecting a single hypothesis H once observed t. The q-values are thus Bayesian versions of p-values, analogous to the "Bayesian posterior p-values" of Morton (1955).

According to Storey (2003) the q-value it is not an adjusted p-value in order to control the $pFDR$ in the sense that it does not satisfy the definition of adjusted p-value given by Shaffer (1995): "Given any test procedure, the adjusted p-value corresponding to a test of a single hypothesis H_i can be defined as the level of entire test procedure at which H_i would be rejected, given the values of all test statistics involved". Therefore, since the $pFDR$ cannot be controlled by a test procedure, then it cannot be used to define adjusted p-values. An argument against to this original view of Storey (2003) was given in Ge *et al.* (2003) where the q-values are viewed as the product of a step wise procedure.

We can calculate the q-value by finding the minimizer of $\Pr(H = 0|T \in \Gamma_\alpha)$, that is

$$\begin{aligned} & \arg \min_{\{\Gamma_a : t \in \Gamma_\alpha\}} \Pr(H = 0|T \in \Gamma_\alpha) \\ = \ & \arg \min_{\{\Gamma_a : t \in \Gamma_\alpha\}} \frac{\pi_0 \Pr(T \in \Gamma_\alpha|H = 0)}{\pi_0 \Pr(T \in \Gamma_\alpha|H = 0) + (1 - \pi_0) \Pr(T \in \Gamma_\alpha|H = 1)} \\ = \ & \arg \min_{\{\Gamma_a : t \in \Gamma_\alpha\}} \frac{\Pr(T \in \Gamma_\alpha|H = 0)}{\Pr(T \in \Gamma_\alpha|H = 1)} \end{aligned} \tag{2.17}$$

Therefore the rejection region which determines the q-value minimizes the ratio of the Type I error to the power over all rejection regions that contain the observed test statistic. This interpretation agrees with the fact that the $pFDR$ is concerned with measuring how frequent the false positives occur in relation to true positives.

Let $G_1(\alpha)$ and $G_0(\alpha)$ be respectively the cdf of the null and alternative p-values:

$$\begin{aligned} G_1(\alpha) &= \Pr(T \in \Gamma_\alpha|H = 1) = \int_{\Gamma_\alpha} dF_1 \\ G_0(\alpha) &= \Pr(T \in \Gamma_\alpha|H = 0) = \int_{\Gamma_\alpha} dF_0 \end{aligned}$$

then the following theorem hold.

Theorem 9 *(Storey, 2002). For m identical hypothesis tests,*

$$pFDR\left(\Gamma_\alpha\right) = pFDR\left(p : p \leq \alpha\right),$$

which implies that the q-value can be calculated from either the original statistics or their p-values. Also, when the statistics are independent and we assume $G_0\left(\alpha\right) = \alpha$ then

$$q\text{-}value\left(t\right) = pFDR\left(p : p \leq p\text{-}value\left(t\right)\right)$$

if and only if $G_1\left(\alpha\right)/a$ is decreasing in α.

This theorem justifies the definition of a q-value in terms of a p-value rather than the original test statistic, because the q-value is the same as if it were calculated on the original test statistics.

For p-values based rejections, all rejection regions are of the form $[0, p]$ for some $0 \leq p \leq 1$. Instead of denoting rejection regions by the more abstract Γ we denote them by p which refers to the interval $[0, p]$. In terms of p-values we can write the result of Theorem 5 as

$$pFDR\left(p\right) = \frac{\pi_0 \Pr\left(P \leq p | H = 0\right)}{\Pr\left(P \leq p\right)} = \frac{\pi_0 \cdot p}{\Pr\left(P \leq p\right)},$$

where P is the random p-value variable resulting from any test.

In Storey (2002) is showed the following non parametric method which have been used in this work to estimate π_0 and $\Pr\left(P \leq p\right)$ from the data. However the estimation of $\Pr\left(P \leq p\right)$ can be performed using the parametric model available for the single hypothesis testing. This is done in the relevant example in the next Chapter.

Since $\pi_0 \cdot m$ of the p-values are expected to be null, then the largest p-values are most likely to come from the null. Moreover $\pi_0 \cdot (1 - \lambda)$ of the null are expected to fall in the region $(\lambda, 1]$, and a small proportion of alternative p-values will fall outside $(\lambda, 1]$. Hence a good estimate of π_0 (Storey, 2002) is

$$\widehat{\pi_0}\left(\lambda\right) = \frac{\#\left\{p_i > \lambda\right\}}{\left(1 - \lambda\right) m} = \frac{W\left(\lambda\right)}{\left(1 - \lambda\right) m} \tag{2.18}$$

for some well chosen λ, where $W\left(\lambda\right) = \#\left\{p_i > \lambda\right\}$. From now on we assume λ to be fixed, the estimation of λ we will be provided in Algorithm 11 later showed. Storey (2002) showed that the estimation of $pFDR$ with $\lambda = 0$ leads to the BH procedure, generalizing in this way the BH procedure and building a family of estimation procedures, $\widehat{pFDR}_\lambda\left(p\right)$, indexed by λ. The construction of this family follows by considering a natural estimate of $\Pr\left(P \leq p\right)$

$$\widehat{\Pr}\left(P \leq p\right) = \frac{\#\left\{p_i \leq p\right\}}{m} = \frac{R\left(p\right)}{m},$$

where $R\left(p\right) = \#\left\{p_i \leq p\right\}$. Therefore, a good estimate of $pFDR\left(p\right)$ for a fixed λ is

$$\widehat{Q}_\lambda\left(p\right) = \frac{\widehat{\pi_0}\left(\lambda\right) p}{\widehat{\Pr}\left(P \leq p\right)} = \frac{W\left(\lambda\right) p}{\left(1 - \lambda\right) R\left(p\right)} \tag{2.19}$$

It is also possible to shows that $\widehat{Q}_\lambda\left(p\right)$ is the MLE of $pFDR\left(p\right)$ for a fixed λ and it has good consistency properties for $m \to \infty$ (Storey, 2002). However, due to finite sample

considerations, we have to make two slight adjustments in order to estimate the $pFDR$ (Storey, 2003). When $R(p) = 0$, the estimate would be undefined, which is undesirable for finite samples. Therefore we replace $R(p)$ with $R(p) \vee 1 = \max\{R(p), 1\}$. This is equivalent to take a linear interpolation between the estimate of $pFDR$ at $\left[0, p_{(1)}\right]$ and the origin. Note also that $1-(1-p)^m$ is the lower bound for $\Pr(R(p) > 0)$ when the tests are independent, therefore since the $pFDR$ is conditioned on $R(p) > 0$, we divide (2.19) by $1-(1-p)^m$.Therefore we estimate $pFDR$ as

$$\widehat{pFDR}_\lambda(p) = \frac{\widehat{\pi_0}(\lambda)\, p}{\widehat{\Pr}(P \leq p)} = \frac{W(\lambda)\, p}{(1-\lambda)(R(p) \vee 1)(1-(1-p)^m)} \tag{2.20}$$

If $\widehat{pFDR}_\lambda(p) > 1$ then we set $\widehat{pFDR}_\lambda(p) = 1$ since it is obviously $pFDR(p) \leq 1$.

Assuming independence of p-values we have that $\widehat{pFDR}_\lambda(p)$ is a conservative estimates of $pFDR(p)$ and that truncating $\widehat{pFDR}(p)$ at 1 provides a decrease in the mean square error.

Since q-value$(p) = \inf_{s \geq p} pFDR(s)$, we can estimate it by

$$\widehat{q}_\lambda(p) = \inf_{s \geq p} \widehat{pFDR}_\lambda(s)$$

where $\widehat{pFDR}_\lambda(s)$ has been defined in (2.20). The following Algorithm 10 provides estimate of the q-value(p).

Algorithm 10 *(Storey, 2002).*

1. *For the m hypotheses tests, calculate the p-values $p_1, \ldots, p_m$.*
2. *Let $p_{(1)} \leq \ldots \leq p_{(m)}$ be the ordered p-values.*
3. *Set $\widehat{q}_\lambda\left(p_{(m)}\right) = \widehat{pFDR}_\lambda\left(p_{(m)}\right)$.*
4. *Set $\widehat{q}_\lambda\left(p_{(i)}\right) = \min\left\{\widehat{pFDR}_\lambda\left(p_{(i)}\right), \widehat{q}_\lambda\left(p_{(i+1)}\right)\right\}$ for $i = m-1, m-2, \ldots, 1$.*

The q-values can be used in practice in the following way: they give us the minimum $pFDR$ we can achieve for rejection regions containing $\left[0, p_{(i)}\right]$ for $i = 1, \ldots, m$. In other words, for each p-value there is a rejection region with $pFDR$ equal to q-value$\left(p_{(i)}\right)$ so that at least $p_{(1)}, \ldots, p_{(i)}$ are rejected.

Estimator $\widehat{pFDR}_\lambda(p)$ involves the estimator $\widehat{\pi_0}(\lambda)$ and, in the end, the tuning parameter λ. We consider choosing λ adaptively to simultaneously minimize the bias and the variance of the procedures.

Recall that

$$\widehat{\pi_0}(\lambda) = \frac{W(\lambda)}{m(1-\lambda)} = \frac{\#\{p_i : p_i > \lambda\}}{m(1-\lambda)}.$$

Suppose $G_1(\lambda)$ is the power over $[0, \lambda]$ averaged over all alternative hypotheses. Then (Storey, 2003)

$$\mathbf{E}\left[\widehat{\pi_0}(\lambda)\right] = \pi_0 + \frac{1-G_1(\lambda)}{1-\lambda}(1-\pi_0) \geq \pi_0,$$

so clearly $\widehat{\pi_0}(\lambda)$ is conservatively biased estimate of π_0. Moreover, if the p-values are independent, and each alternative p-value follows the distribution G_1, then (Storey, 2003)

$$\mathbf{Var}\left[\widehat{\pi_0}(\lambda)\right] = \frac{\pi_0 \lambda}{m(1-\lambda)} + \frac{(1-\pi_0)\left[1-G_1(\lambda)\right] G_1(\lambda)}{m(1-\lambda)^2}.$$

Now it is easy to see that $\lim_{\lambda\to 0}\mathbf{Var}\left[\widehat{\pi_0}\left(\lambda\right)\right]=0$ and $\lim_{\lambda\to 1}\mathbf{Var}\left[\widehat{\pi_0}\left(\lambda\right)\right]=\infty$. The bias of $\widehat{\pi_0}\left(\lambda\right)$ is greatest at $\lambda=0$ and then it decreases as λ increase. Therefore, there is clearly a bias-variance trade-off in the choice of λ and we choose λ that minimizes the mean-squared error, $MSE\left(\lambda\right)$, which is a common measure to balance a bias-variance trade-off situation. If one is interested in estimating the rejection region in order to control the $pFDR$ or the q-value then two functions to minimize in λ would be

$$\mathbf{E}\left[\left(\widehat{pFDR}_{\lambda}\left(p\right)-pFDR\left(p\right)\right)^{2}\right] \tag{2.21}$$

and

$$\sum_{i=1}^{m}\mathbf{E}\left\{\left[\widehat{q}_{\lambda}\left(p_i\right)-q\left(p_i\right)\right]^{2}\right\} \tag{2.22}$$

But since both estimators (2.21), (2.22) involve estimating $\widehat{\pi_0}\left(\lambda\right)$ then we choose λ that minimize

$$\mathbf{E}\left[\left(\widehat{\pi_0}\left(\lambda\right)-\pi_0\right)^{2}\right]$$

However, we do not know π_0, so we have to form a plug-in estimate of this quantity. Notice that for any λ we have:

$$\mathbf{E}\left[\widehat{\pi_0}\left(\lambda\right)\right]\geq\min_{\lambda'}\mathbf{E}\left[\widehat{\pi_0}\left(\lambda\right)\right]\geq\pi_0$$

therefore the considered plug-in estimate is $\min_{\lambda'\in\Lambda}\left[\widehat{\pi_0}\left(\lambda'\right)\right]$ where Λ is a prefixed grid of values of λ. According to Storey (2003) when the p-values are independent, we estimate $MSE\left(\lambda\right)$ for all $\lambda\in\Lambda$ with the bootstrap estimator

$$\widehat{MSE}\left(\lambda\right)=\frac{1}{B}\sum_{b=1}^{B}\left(\widehat{\pi_0}^{*b}\left(\lambda\right)-\min_{\lambda'\in\Lambda}\left[\widehat{\pi_0}\left(\lambda'\right)\right]\right)^{2} \tag{2.23}$$

where $\widehat{\pi_0}^{*b}\left(\lambda\right)$ is the bootstrap estimate of $\widehat{\pi_0}\left(\lambda\right)$. We choose $B=100$. The whole estimate procedure is listed in the following Algorithm 11.

Algorithm 11 *(Storey et al., 2004).*
1. For some range of λ*, say* $\Lambda=\{0,0.01,0.02,\ldots,0.95\}$*, calculate* $\widehat{\pi_0}\left(\lambda\right)$ *using (2.18).*
2. For each $\lambda\in\Lambda$*, form* $B=100$ *bootstrap version* $\widehat{\pi_0}^{*b}\left(\lambda\right)$ *of the estimate,* $b=1,\ldots,B$*.*
3. For each $\lambda\in\Lambda$*, estimate its respective mean square error using (2.23).*
4. Set $\widehat{\lambda}=\arg\min_{\lambda\in\Lambda}\widehat{MSE}\left(\lambda\right)$ *and the overall estimate of* π_0*:* $\widehat{\pi_0}=\widehat{\pi_0}\left(\widehat{\lambda}\right)$*.*

We note that, no procedures can give strong or weak control for $pFDR$, as $pFDR=1$ when $m_0=m$, but this is very unlikely with microarray data, because we expect to have genes differentially expressed and so $m_0<m$. Therefore the $pFDR$ can be conservatively estimated with the q-values under the unknown set of null true hypotheses $\mathcal{M}_0$. The q-values provide an exact control of $pFDR$. As showed in Storey *et al.* (2004) the $pFDR$ is less conservative than the FDR, however the two procedures are asymptotically equivalent.

2.3.4 MHT under dependency

It is possible to show (Storey, 2002) that $pFDR$ calculated for a fixed number of tests, $pFDR_m(\Gamma_\alpha)$ converge for $m \to \infty$ to $pFDR(\Gamma_\alpha)$ if $\sum_{i=1}^{m}(1-H_i)/m \to \pi_0$ and if the distribution of the p-values under the null $G_{0_m}(\alpha)$ and alternative hypothesis $G_{1_m}(\alpha)$ converge respectively to some functions $G_0(\alpha)$ and $G_1(\alpha)$. This is true even if the tests are weakly dependent. However for fixed m we could consider some correction in order to take into account the dependency.

Storey (2002) suggested that a correction for dependency in (2.20) could be to consider $\Pr(R(p)>0)$ instead of the lower bound $1-(1-p)^m$. This probability is computed by simulating $B = 100$ p-values under the null hypothesis and then using the Monte Carlo approximation:

$$\Pr(\widehat{R(p)>0}) = \frac{1}{B}\sum_{b=0}^{B}\mathbf{1}_{(p^b>\alpha)}. \tag{2.24}$$

The Benjamini and Yekutieli procedure

Benjamini and Yekutieli (2001) proposed the BY procedure that account for multiple hypothesis testing under dependency. The (2.14) compare the observed p-values with $\alpha \cdot k/\sum_{i=1}^{m} 1/i$ instead of $\alpha \cdot k/m$. The adjustment by $\sum_{i=1}^{m} 1/i \approx \log(m) + \frac{1}{2}$ is quite often not needed, and yields a too conservative procedure. Still, even if only a small proportion ($\approx \log(m)/m$) of the tested hypotheses are detected as not true, the BY procedure is more powerful than the comparable $FWER$ controlling procedures. However, what Benjamini and Yekutieli (2001) demonstrated is that the BH procedure control the FDR under the dependency form that they called *Positive Regression Dependency on each one from a Subset* (PRDS). The PRDS is formally defined on an increasing set of statistics. The PRDS is defined an increasing set $\mathcal{T}$ and the set of test statistics $\mathbf{T} = \{T_1, T_2, \ldots, T_m\}$.

Definition 12 *(Benjamini and Yekutieli, 2001). For any increasing set $\mathcal{T}$, and for each subset of null hypotheses I_0, $\Pr(\mathbf{T} \in \mathcal{T} | T_i = t)$ is non-decreasing in t for $i \in I_0$.*

Recall that $\mathcal{T}$ is increasing if $t \in \mathcal{T}$ and $t' \geq t$ implies that also $t' \in \mathcal{T}$.

The PRDS property is a relaxed form of the positive regression dependency property that characterizes the multivariate normal distributions, often used in modelling microarray genes expression. The following Theorem 13 makes this point more formal.

Theorem 13 *(Benjamini and Yekutieli, 2001). Consider $\mathbf{T} \sim N(\mu, \mathbf{\Sigma})$ a vector of test statistics each testing hypothesis $\mu_i = 0$ against the alternative $\mu_i > 0$, for $i = 1, \ldots, m$. For $i \in I_0$, the set of true null hypotheses, $\mu_i = 0$. Otherwise $\mu_i > 0$. Assume that for each $i \in I_0$, and for each $j \neq i$, $\mathbf{\Sigma}_{ij} \geq 0$, then the distribution of $\mathbf{T}$ is PRDS over I_0.*

Benjamini and Yekutieli (2001) demonstrated that the BY procedures always control the FDR, no matter what kind of dependency is assumed.

Finally Storey *et al.* (2004) demonstrated that the $\widehat{pFDR}_\lambda(p)$ control the $pFDR$ for m fixed when the PRDS holds and the p-values are frequentist.

The minP adjustment procedure

We do not further investigate the operating characteristics of other MHT procedures under dependency, however we mention here other two procedure that takes into account the dependency structure among the tests. The first was developed by Westfall and Young (1993). It is a single-step procedure that calculate the *minP adjusted p-values*, which are defined by

$$\widetilde{p}_{(i)} = \Pr\left(\min_{1 \leq k \leq m} P_k \leq p_{(i)} \middle| I_0 \right). \tag{2.25}$$

The (2.25) is usually estimated non parametrically using a permutation test procedure (Westfall and Young, 1993). Another way to incorporate parametrically the dependency structure of the m tests is the single-step *maxT adjusted p-values* which are defined in terms of the test statistics T_i,

$$\widetilde{p}_{(i)} = \Pr\left(\max_{1 \leq k \leq m} |T_k| \geq |t_i| \middle| I_0 \right). \tag{2.26}$$

By assuming a parametric law for the joint distribution on $(T_1, T_2, ..., T_m)$ we can calculate the (2.26). Ge *et al.* (2003) showed that when adjusted p-values are estimated by permutation and m is large, procedures based on the $minP$ adjusted p-values tend to be more sensitive to the number of permutations and more conservative than those based on the $maxT$ adjusted p-values. If we are using frequentist p-values than the adjusted p-values in (2.25) and (2.26) are the same as those obtained with the Holm procedures *ii*). For more on comparison of multiple hypothesis test procedures under dependency see Ge *et al.* (2003) the discussion therein.

Chapter 3

Results

So far we have showed the literature concerning p-values and MHT separately. In this chapter we use simulations to support the following conjecture.

Conjecture 14 *We are more likely to find differentially expressed genes, while maintaining a control in the FDR (and pFDR) using the Partial posterior predictive p-value and the Conditional Predictive p-value.*

However the goal of this chapter is broader, in fact we will characterize the inference using the above six MHT procedures jointly with the above four definitions of p-values: p_{plug}, p_{post}, p_{cpred} and p_{ppost}. We will do this before using simulations and than considering an application of the Gamma model to three public data sets. The way we implemented the inference for the Gamma model is really new and we show that it provides useful results.

Before look at these results, it is useful to have an idea of the error encountered in using p-values that are not frequentist, or which is the same, when we don't know their null distribution $G_0(\alpha)$. This idea is provided in the following relevant example.

3.1 A relevant example

We illustrate the numerical convergency of $pFDR_m(p)$ to $pFDR(p)$ for $m \to \infty$ either when we know the distribution of the p-values under the null hypothesis and when we don't know it because we are not able to make appropriate inference on the nuisance parameters involved in hypothesis testing.

Suppose this state of nature

$$T_i|H_i = 0 \sim N(0,1)$$

and

$$T_i|H_i = 1 \sim N(2,1).$$

with the following covariance structure

$$\mathbf{Cov}(T_i, T_{i+k}) = \begin{cases} \rho, & \text{for } k = 1, 2, \ldots, 9 \text{ and } i = 1, 11, 21, \ldots, m \\ 0, & \text{otherwise} \end{cases}$$

where $0 \leq \rho \leq 1$.

In other words the statistics have correlation ρ in groups of 10. We are simulating a kind of dependency called *clumpy dependency* among genes (Storey, 2002), which is often encountered when measuring genes expression. This kind of dependency is realistic and it is mainly due to the fact that genes interact in a pathway and there also tends to be cross hybridization because of molecular similarity at the sequence level.

Let $N^{-1}\left(1-\alpha, \mu, \sigma^2\right)$ indicates the quantile of $N\left(\mu, \sigma^2\right)$ at point $1-\alpha$ and suppose to be under two inferential situations on the test statistic T.

a) We want to perform this hypothesis testing of each gene

$$H_i = 0 : T_i \sim N\left(\mu = 0, 1\right)$$

against

$$H_i = 1 : T_i \sim N\left(\mu \neq 0, 1\right).$$

In this case the null and alternative distribution of T_i are completely specified and the rejection region for all tests is

$$\Gamma_\alpha = \left[N^{-1}\left(1-\alpha, 0, 1\right), \infty\right)$$

and so $\Pr\left(T_i \in \Gamma_\alpha | H = 0\right) = \alpha$.

b) We don't know the variance σ^2 and we have to test

$$H_i = 0 : T_i \sim N\left(\mu = 0, \sigma^2\right), \forall \sigma^2 > 0$$

against

$$H_i = 1 : T_i \sim N\left(\mu \neq 0, \sigma^2\right), \forall \sigma^2 > 0$$

considering some estimator $\widehat{\sigma}^2$ where $\Pr\left(\widehat{\sigma}^2 \neq 1\right) > 0$. The rejection region for all tests is

$$\Gamma'_a = \left[N^{-1}\left(1-\alpha, 0, \widehat{\sigma}^2\right), \infty\right], \forall \widehat{\sigma}^2 > 0.$$

Suppose, now, that we ignore that

$$\Pr\left(T_i \in \Gamma'_\alpha | H = 0\right) = \alpha' \neq \alpha$$

but we behaves as $\Pr\left(T_i \in \Gamma'_\alpha | H = 0\right) = \alpha$.

Situation a) refers to the use of frequentist p-values, while in situation b) we are not using frequentist p-values, but we treat them as if they were. In particular if the p-values were conservative (anticonservative) then $\alpha' < \alpha$ ($\alpha' > \alpha$) which is the case if $\widehat{\sigma}^2 > 1$ ($\widehat{\sigma}^2 < 1$).

Let again assume $H_i \overset{i.i.d.}{\sim} Bernoulli\left(1 - \pi_0 = 0.1\right)$ and $\alpha = 0.005$ and $\widehat{\sigma}^2 = 2$. Storey (2003) provides the following result

$$\lim_{m \to \infty} \sup_{\alpha \geq 0} \left| pFDR_m\left(\Gamma_a\right) - \Pr_\infty\left(H = 0 | T \in \Gamma_\alpha\right)\right| = 0.$$

Therefore

m	ρ=0	ρ=0.2	ρ=0.6	ρ=0.8	ρ=1
100	0.136 (0.007)	0.131 (0.007)	0.105 (0.006)	0.088 (0.006)	0.049 (0.006)
500	0.130 (0.003)	0.135 (0.003)	0.128 (0.003)	0.114 (0.004)	0.089 (0.006)
1000	0.138 (0.002)	0.135 (0.002)	0.129 (0.002)	0.128 (0.003)	0.110 (0.005)
3000	0.139 (0.001)	0.137 (0.001)	0.137 (0.001)	0.134 (0.002)	0.134 (0.003)
5000	0.138 (0.001)	0.137 (0.001)	0.137 (0.001)	0.136 (0.001)	0.130 (0.002)
10000	0.136 (0.001)	0.137 (0.001)	0.137 (0.001)	0.138 (0.001)	0.136 (0.002)

Table 3.1: Simulation results: $pFDRm\left(\Gamma_a\right) \rightarrow \Pr_\infty\left(H=0|T\in\Gamma_\alpha\right)=0.137$

m	ρ=0	ρ=0.2	ρ=0.6	ρ=0.8	ρ=1
100	0.020 (0.004)	0.019 (0.004)	0.023 (0.005)	0.014 (0.003)	0.008 (0.003)
500	0.020 (0.003)	0.021 (0.003)	0.021 (0.003)	0.015 (0.003)	0.005 (0.002)
1000	0.027 (0.002)	0.026 (0.002)	0.021 (0.002)	0.017 (0.002)	0.006 (0.002)
3000	0.025 (0.001)	0.023 (0.001)	0.023 (0.001)	0.022 (0.002)	0.019 (0.003)
5000	0.026 (0.001)	0.023 (0.001)	0.023 (0.001)	0.023 (0.001)	0.018 (0.002)
10000	0.024 (0.001)	0.024 (0.001)	0.022 (0.001)	0.024 (0.001)	0.022 (0.002)

Table 3.2: Simulation results: $pFDRm\left(\Gamma_a\right) \rightarrow \Pr_\infty\left(H=0|T\in\Gamma'_{\alpha'}\right)=0.024$

- in situation a) we have

$$\Pr_\infty\left(H=0|T\in\Gamma_\alpha\right)=\frac{\pi_0\alpha}{\pi_0\alpha+\left(1-\pi_0\right)\Pr\left(N\left(2,1\right)\geq N^{-1}\left(1-\alpha,0,1\right)\right)}=0.137;$$

- in situation b) where we erroneously believe $\alpha'=\alpha$ we do not have any convergence to the quantity which we are supposing to control

$$\Pr_\infty\left(H=0|T\in\Gamma'_\alpha\right)=\frac{\pi_0\alpha}{\pi_0\alpha+\left(1-\pi_0\right)\Pr\left(N\left(2,1\right)\geq N^{-1}\left(1-\alpha,0,\widehat{\sigma}^2\right)\right)}=0.47,$$

but to the following quantity

$$\Pr_\infty\left(H=0|T\in\Gamma'_{\alpha'}\right)=\frac{\pi_0\alpha'}{\pi_0\alpha'+\left(1-\pi_0\right)\Pr\left(N\left(2,1\right)\geq N^{-1}\left(1-\alpha,0,\widehat{\sigma}^2\right)\right)}=0.024,$$

where

$$\alpha'=\Pr\left(N\left(0,1\right)\geq N^{-1}\left(1-\alpha,0,\widehat{\sigma}^2\right)\right)=0.00013$$

In fact in situation b) the following two results apply

$$\lim_{m\rightarrow\infty}\sup_{\alpha\geq0}\left|pFDR_m\left(\Gamma'_a\right)-\Pr_\infty\left(H=0|T\in\Gamma'_\alpha\right)\right|>0$$

and

$$\lim_{m\rightarrow\infty}\sup_{\alpha\geq0}\left|pFDR_m\left(\Gamma'_{a'}\right)-\Pr_\infty\left(H=0|T\in\Gamma'_{\alpha'}\right)\right|=0.$$

Table 3.1 shows, for different values of ρ, the convergency of $pFDR_m\left(\Gamma_a\right)$ to $\Pr_\infty\left(H=0|T\in\Gamma_\alpha\right)$ and Table 3.2 to $\Pr_\infty\left(H=0|T\in\Gamma'_{\alpha'}\right)$. For high values of m the differences are within the Monte Carlo standard errors (in parenthesis).

In this numerical example the difference between $pFDR_m\left(\Gamma'_{a'}\right)=0.024$ and $pFDR_m\left(\Gamma'_a\right)=0.47$ is remarkable. We believe to control the $pFDR$ at level $pFDR_m\left(\Gamma'_a\right)$, but in the

end we are controlling at the smaller and more conservative level $pFDR_m\left(\Gamma'_{a'}\right)$. In this way we are likely to miss some genes truly altered in the two experimental conditions. Moreover using non frequentist p-value we may also loose all the advantages, in terms of power, that the recent MHT procedure has provided to the researcher.

3.2 Controlling FDR and $pFDR$ using different p-values.

We further investigate the problem of choosing the appropriate p-value in MHT by examining the behavior of the six MHT procedures across the three p-values. We start with the normal model and then we consider the gamma model.

We simulated B replications of microarray experiments under different setups. Each experiment is composed by a panel of m genes, but only five genes are simulated as differentially expressed. After each simulations we calculated the adjusted p-values and the q-values according to procedures illustrated Chapter 2. We finally ranked all genes according to the adjusted p-values and save the B ranks assigned to the five overexpressed genes. The empirical distributions of this ranks was showed (in box plots) in order to compare the results among the across different notions of p-values and MHT procedures. We will see that p_{cpred} in the normal case and the p_{ppost} in the gamma case outperform the others p-values uniformly, that is over all the considered MHT procedures.

3.2.1 Results for the Normal model using different controlling procedures across different notions of p-value

We simulated $B = 1000$ microarray experiments of $m = 20000$ genes each. Each independent experiment has been replicated three times: $n_X = n_Y = 3$. We considered two situations: when all genes are independent and when there exists strong clumpy dependency. We draw the first five genes in the case ($\mathbf{X}$) according to

$$\mathbf{X} \sim N_{20000}\left(\boldsymbol{\mu}_X = (10.0, 8.5, 7.0, 5.5, 4.0, 2.0, \ldots, 2.0), \boldsymbol{\Sigma} = \mathbf{I}_{20000}\right)$$

and while for the control ($\mathbf{Y}$) according to

$$\mathbf{Y} \sim N_{20000}\left(\boldsymbol{\mu}_Y = (2, \ldots, 2), \boldsymbol{\Sigma} = \mathbf{I}_{20000}\right),$$

where $N_{20000}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ is a multivariate normal of dimension 20000 and $\mathbf{I}_{20000}$ is the identity matrix of the same dimension. We repeat the same experiment for the dependency case by simulating $\mathbf{X}$ and $\mathbf{Y}$ according to

$$\mathbf{X} \sim N_{20000}\left(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}\right)$$

and

$$\mathbf{Y} \sim N_{20000}\left(\boldsymbol{\mu}_Y, \boldsymbol{\Sigma}\right),$$

where $\mathbf{\Sigma}$ is the following covariance matrix of dimension $m \times m$

$$\mathbf{\Sigma}_{m\times m}= \begin{bmatrix} \rho_{11} & \cdots & \rho_{1k} & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & 0 & 0 & 0 & \cdots & 0 \\ \rho_{k1} & \cdots & \rho_{kk} & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \rho_{11} & \cdots & \rho_{1k} & \cdots & 0 \\ 0 & 0 & 0 & \vdots & \ddots & \vdots & \cdots & 0 \\ 0 & 0 & 0 & \rho_{k1} & \cdots & \rho_{kk} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \cdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & \rho_{kk} \end{bmatrix}$$

therefore for $\rho_{ij} = 0.9$, $i = 1, \ldots, k$, $j = 1, \ldots, k$ we have that all genes are dependent in groups of $k = 10$.

Figures 3.1-3.6 show the results in the situation of independents genes expression, while Figures 3.7-3.12 under the clumpy dependency described above.

The box plots represent the distributions of the ranks for the five genes: the more the distribution is centered on lower values, the more is the likely to recognize the five altered genes as differentially expressed. We note that this happened if we use the q-values and the BH procedure combined with the p_{cpred}. Using the other MHT methods it is in general very unlikely to find these genes in the top list of the differentially expressed genes, but still using the p_{cpred} we are more likely to find them. In fact we can see that the isolated observations of very small rank's values happened using the p_{cpred} rather then p_{plug} or p_{post}.

In the case of dependency we can see that the conclusions do not change, in fact we can note only an inflation in the variance of the rank distributions, but still, the location of the medians, for q-values and the BH procedure with the p_{cpred}, corroborate our conjecture.

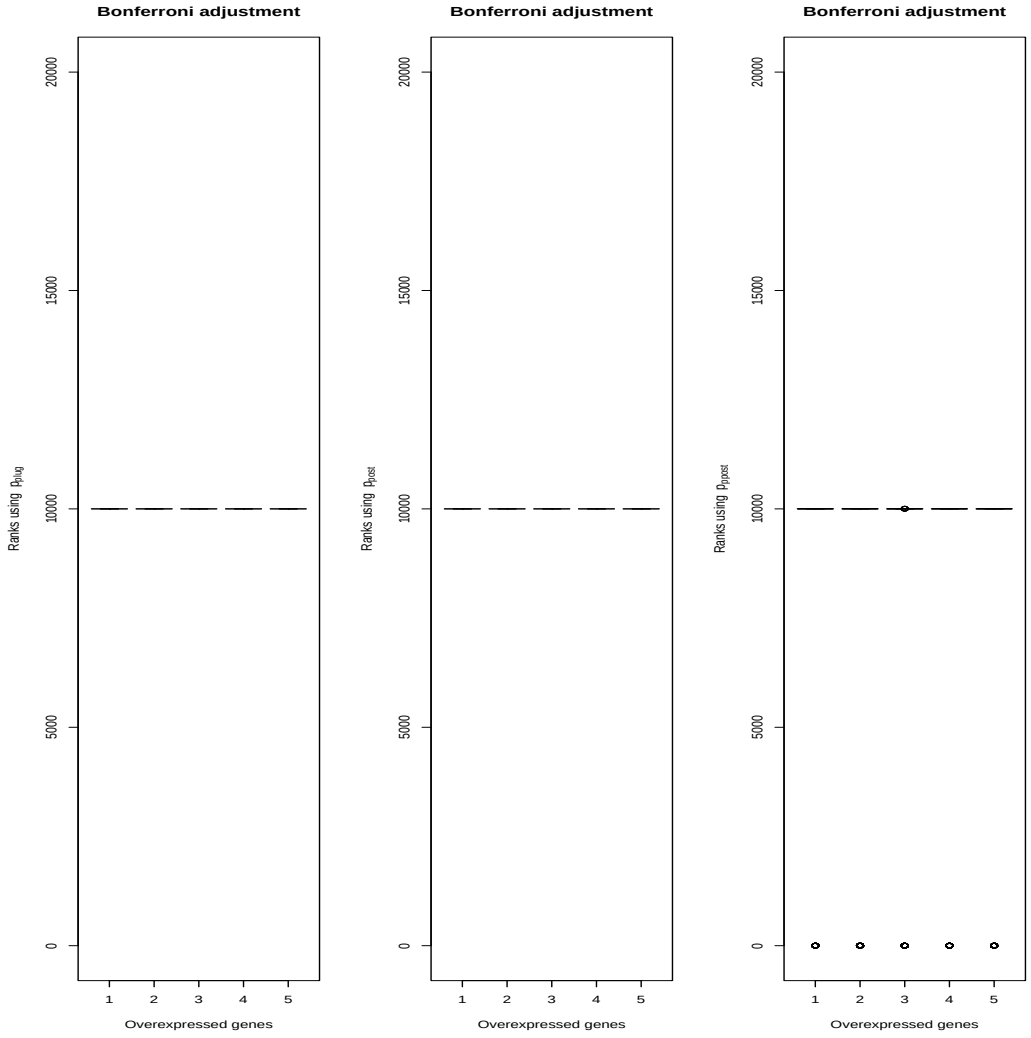

Figure 3.1: Normal model (Independence case): Distribution of the ranks of the 5 overexpressed genes with the p-values adjusted according to the Bonferroni's procedure. We can see that the boxplots are concentrated even if some spikes appear for p_{ppost}. In order to recognize the 5 overexpressed genes the boxplots should be concentrated around 0.

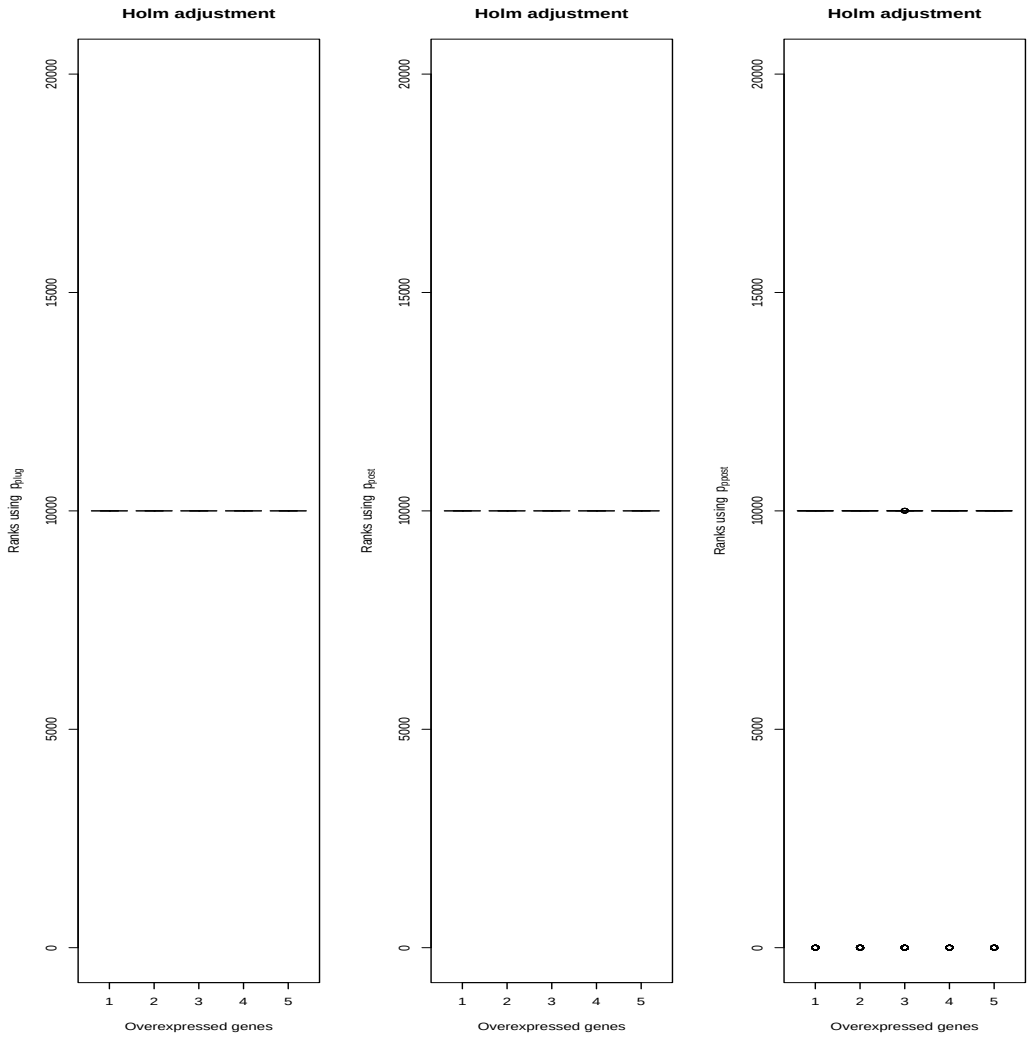

Figure 3.2: Normal model (Independence case): Distribution of the ranks of the 5 overexpressed genes with the p-values adjusted according to the Holm's procedure. We can see that the boxplots are concentrated even if some spikes appear for p_{ppost}. In order to recognize the 5 overexpressed genes the boxplots should be concentrated around 0.

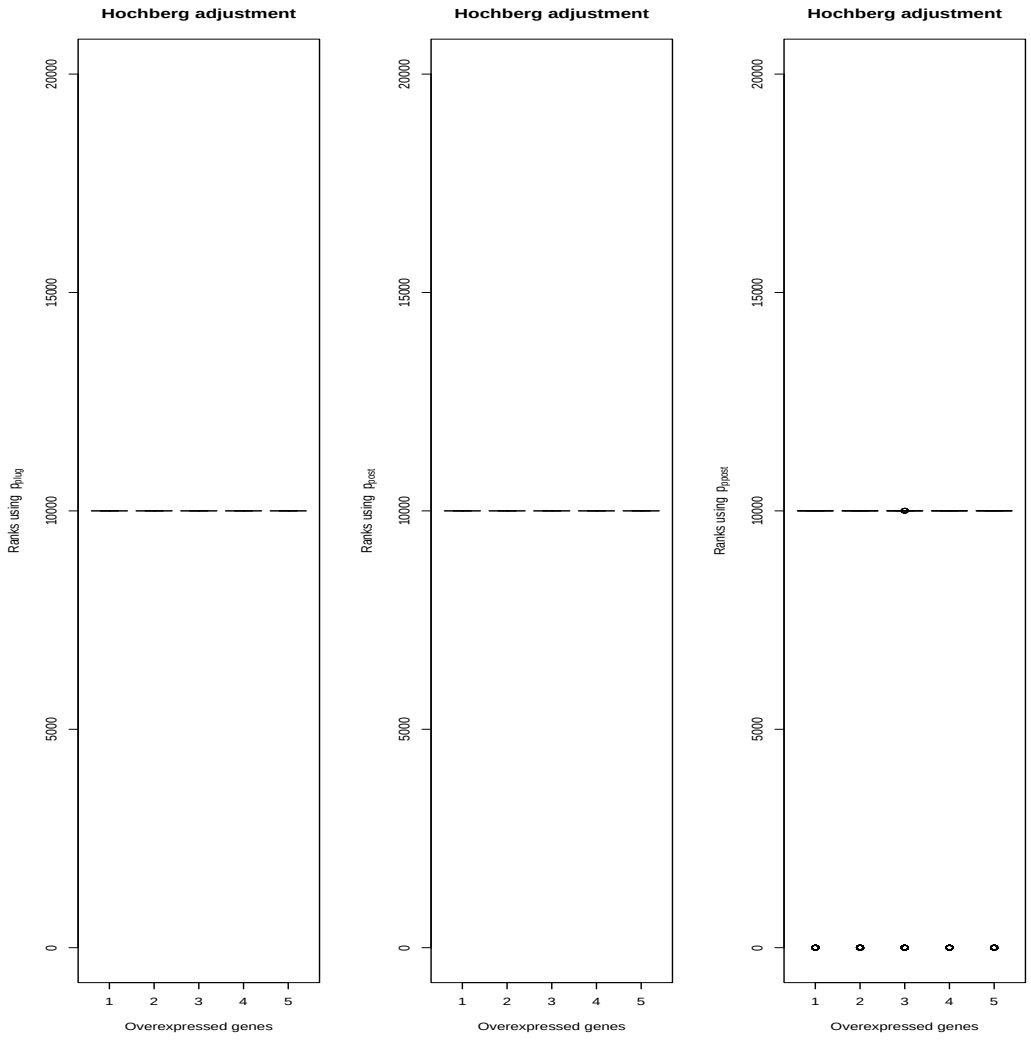

Figure 3.3: Normal model (Independence case): Distribution of the ranks of the 5 overexpressed genes with the p-values adjusted according to the Hochberg's procedure. We can see that the boxplots are concentrated even if some spikes appear for p_{ppost}. In order to recognize the 5 overexpressed genes the boxplots should be concentrated around 0.

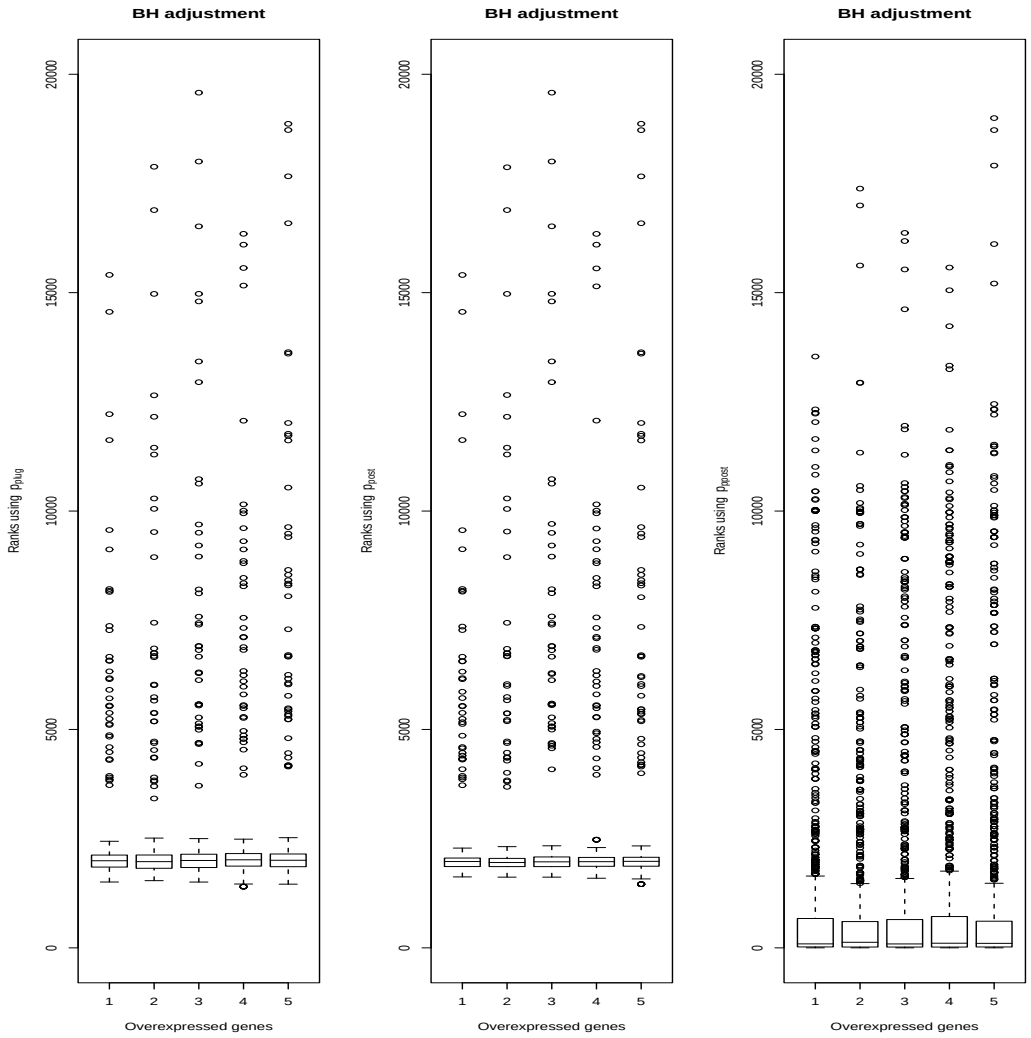

Figure 3.4: Normal model (Independence case): Distribution of the ranks of the 5 overexpressed genes with the p-values adjusted according to the BH's procedure. We can see that the boxplots referring to the p_{ppost} are more concentrated around zero than those for other p-values. In order to recognize the 5 overexpressed genes the boxplots should be concentrated around 0.

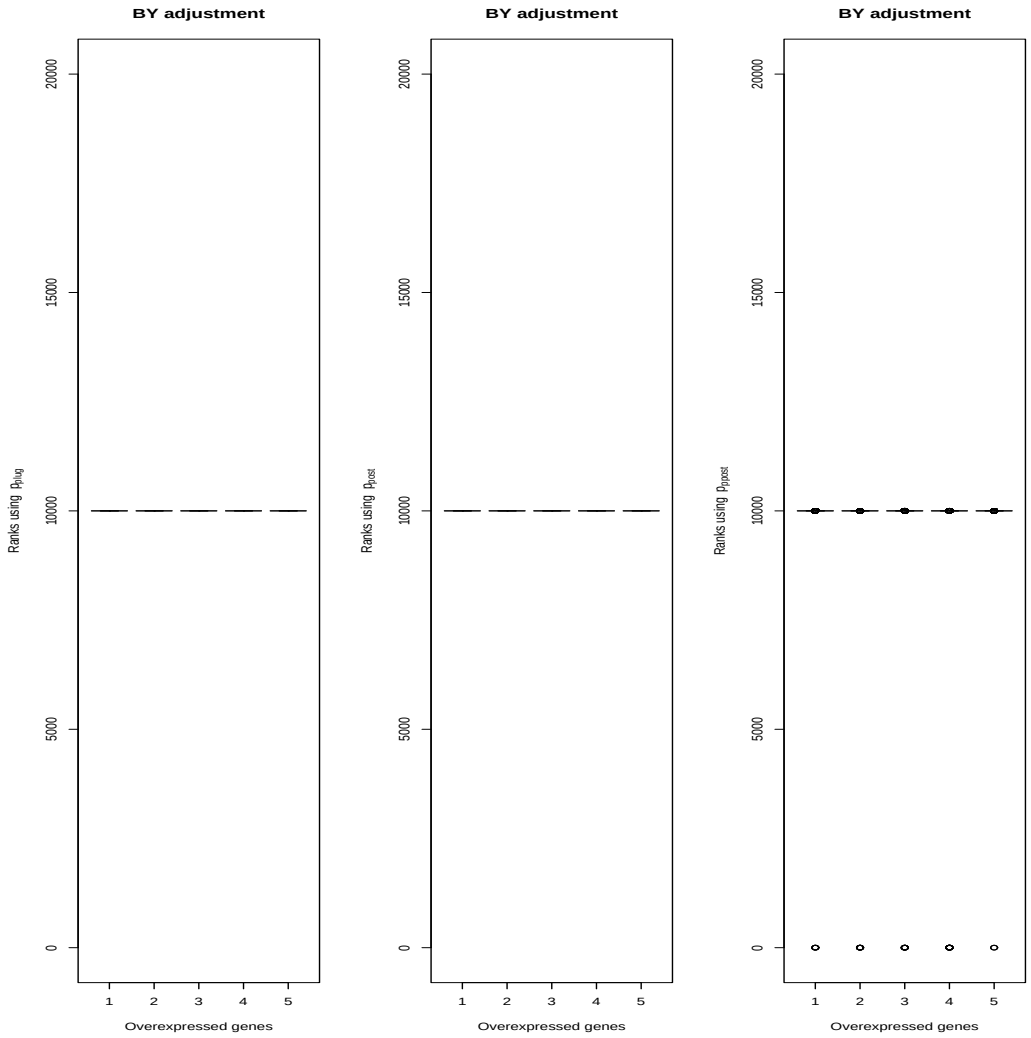

Figure 3.5: Normal model (Independence case): Distribution of the ranks of the 5 overexpressed genes with the p-values adjusted according to the BY's procedure. We can see that the boxplots are concentrated even if some spikes appear for p_{ppost}. In order to recognize the 5 overexpressed genes the boxplots should be concentrated around 0.

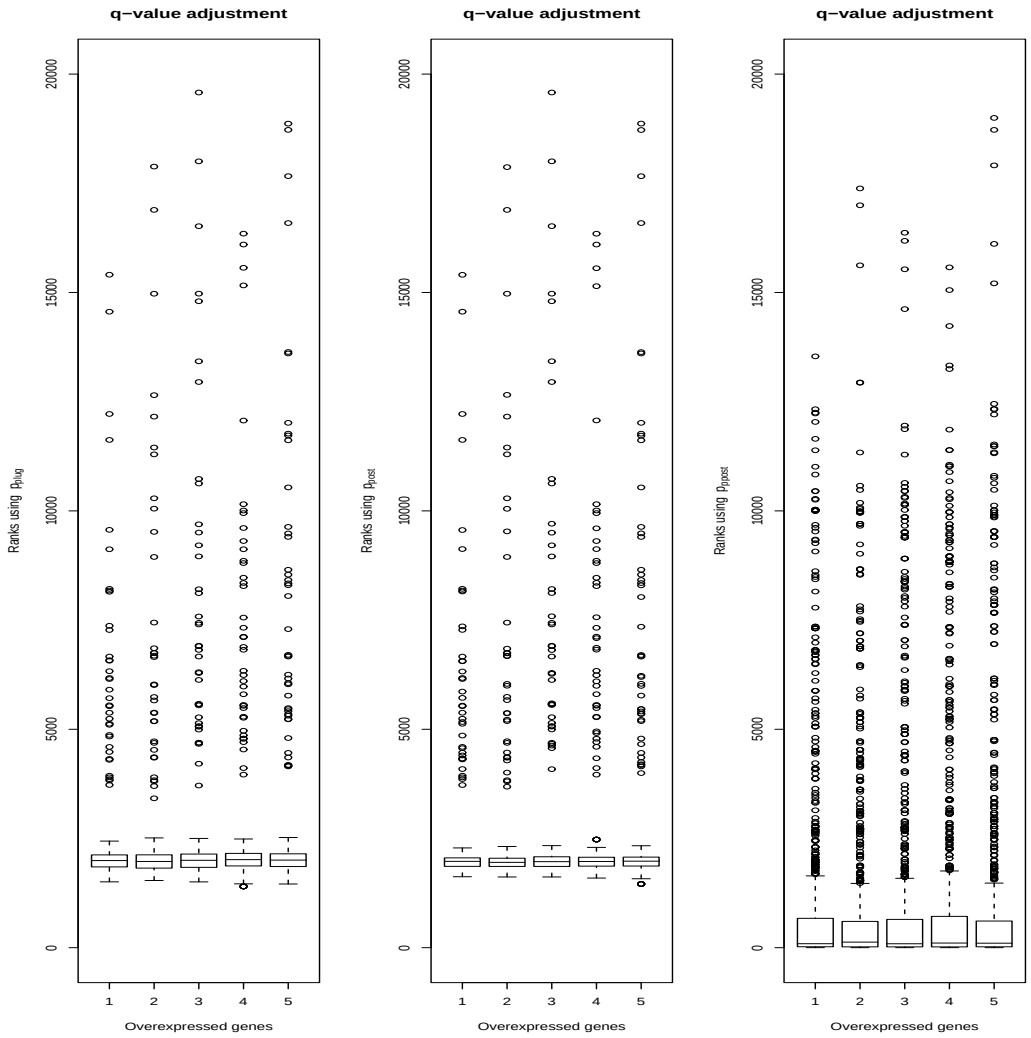

Figure 3.6: Normal model (Independence case): Distribution of the ranks of the 5 overexpressed genes with the p-values adjusted according to the q-values. We can see that the boxplots referring to the p_{ppost} are more concentrated around zero than those for other p-values. In order to recognize the 5 overexpressed genes the boxplots should be concentrated around 0.

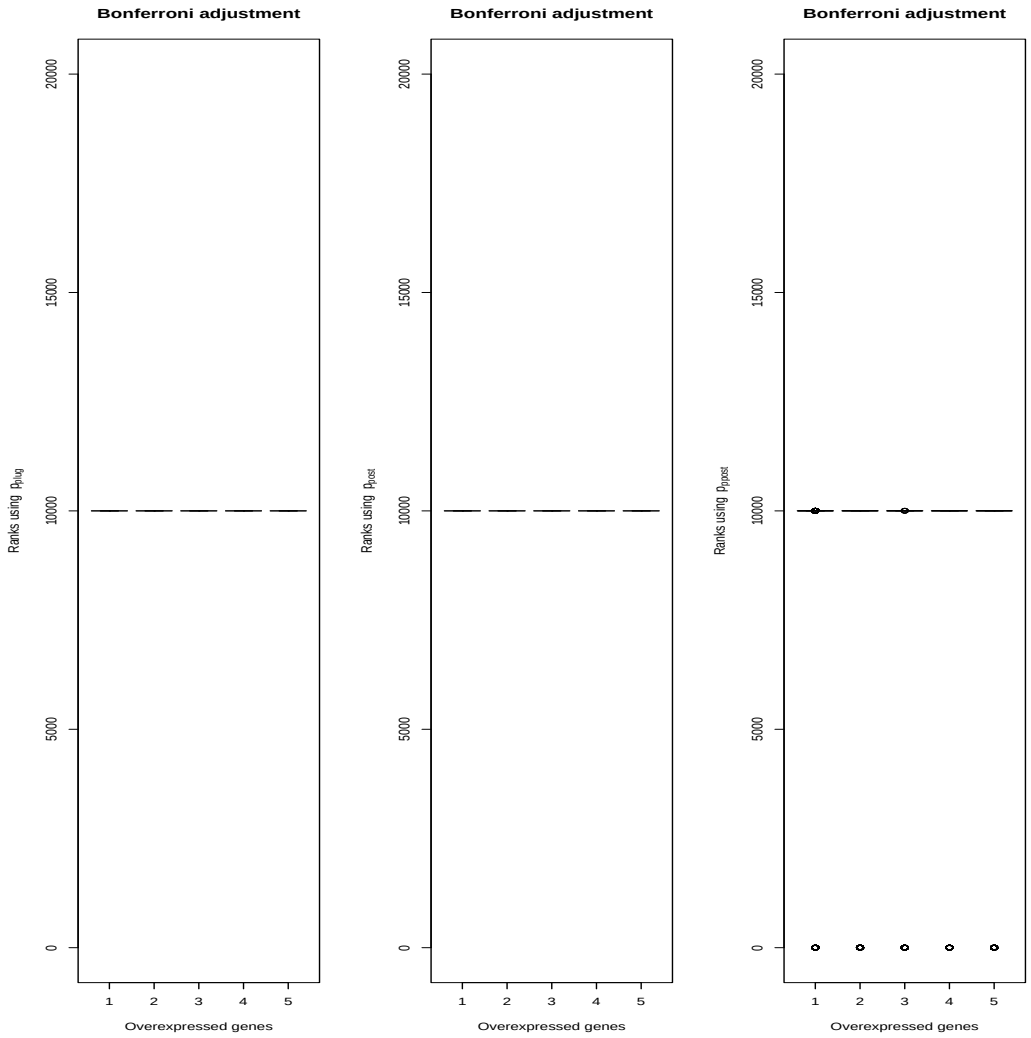

Figure 3.7: Normal model (Dependency case, $\rho = 0.9$): Distribution of the ranks of the 5 overexpressed genes with the p-values adjusted according to the Bonferroni's procedure. We can see that the boxplots are concentrated even if some spikes appear for p_{ppost}. In order to recognize the 5 overexpressed genes the boxplots should be concentrated around 0.

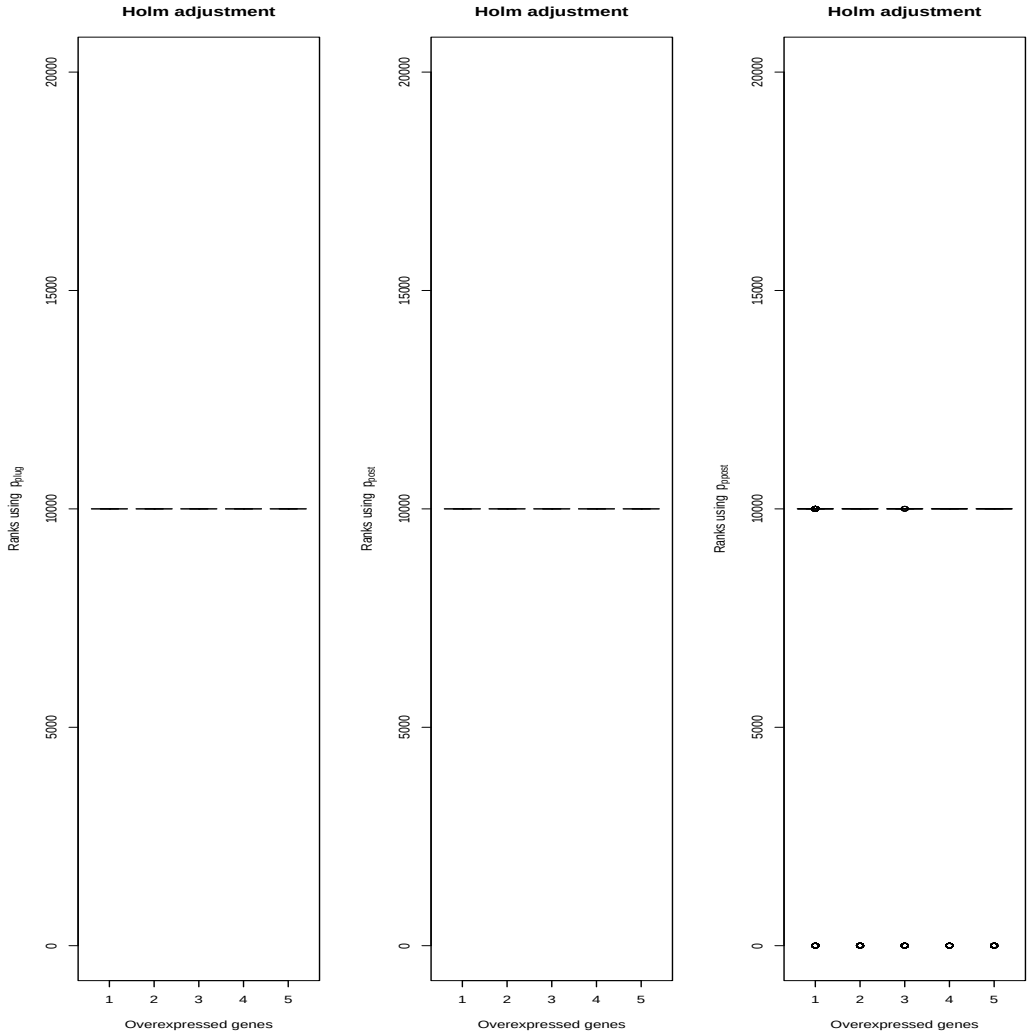

Figure 3.8: Normal model (Dependency case, $\rho = 0.9$): Distribution of the ranks of the 5 overexpressed genes with the p-values adjusted according to the Holm's procedure. We can see that the boxplots are concentrated even if some spikes appear for p_{ppost}. In order to recognize the 5 overexpressed genes the boxplots should be concentrated around 0.

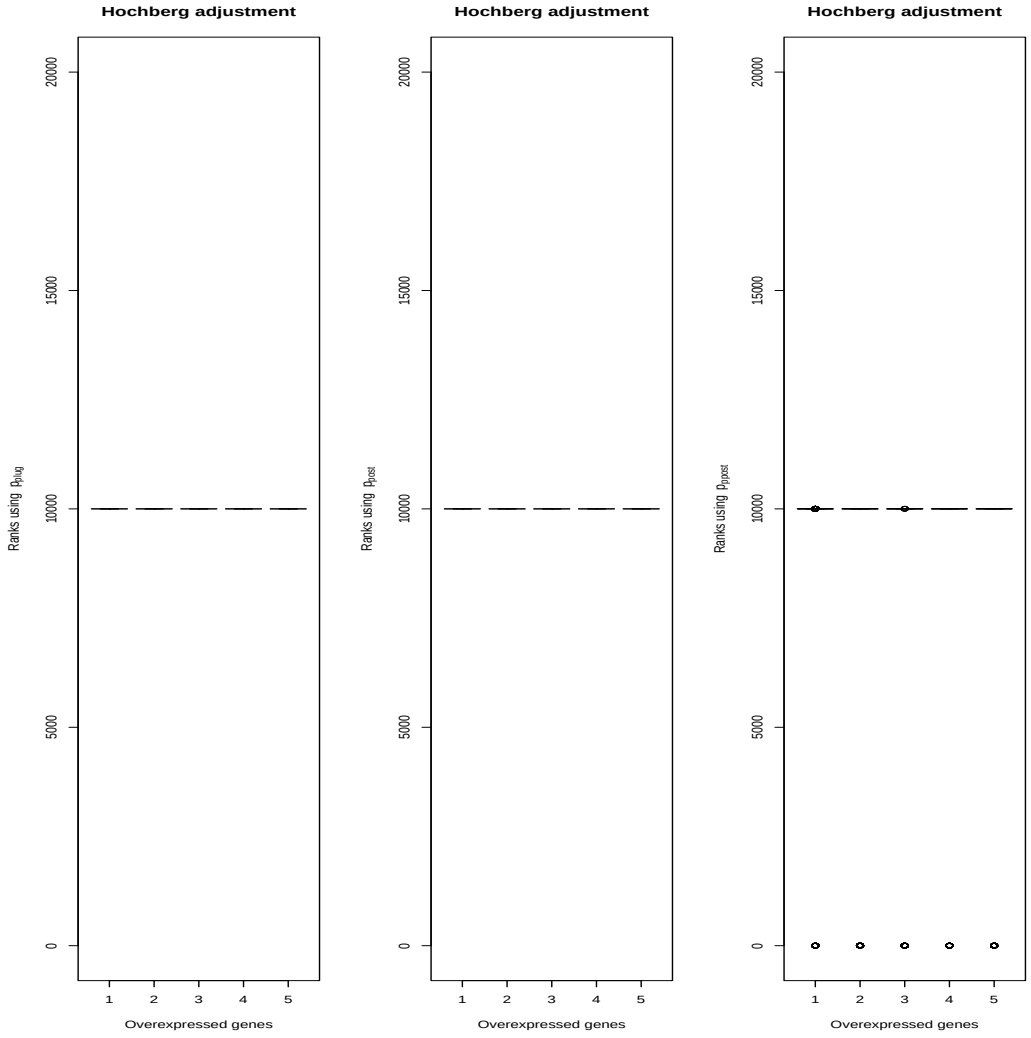

Figure 3.9: Normal model (Dependency case, $\rho = 0.9$): Distribution of the ranks of the 5 overexpressed genes with the p-values adjusted according to the Hochberg's procedure. We can see that the boxplots are concentrated even if some spikes appear for p_{ppost}. In order to recognize the 5 overexpressed genes the boxplots should be concentrated around 0.

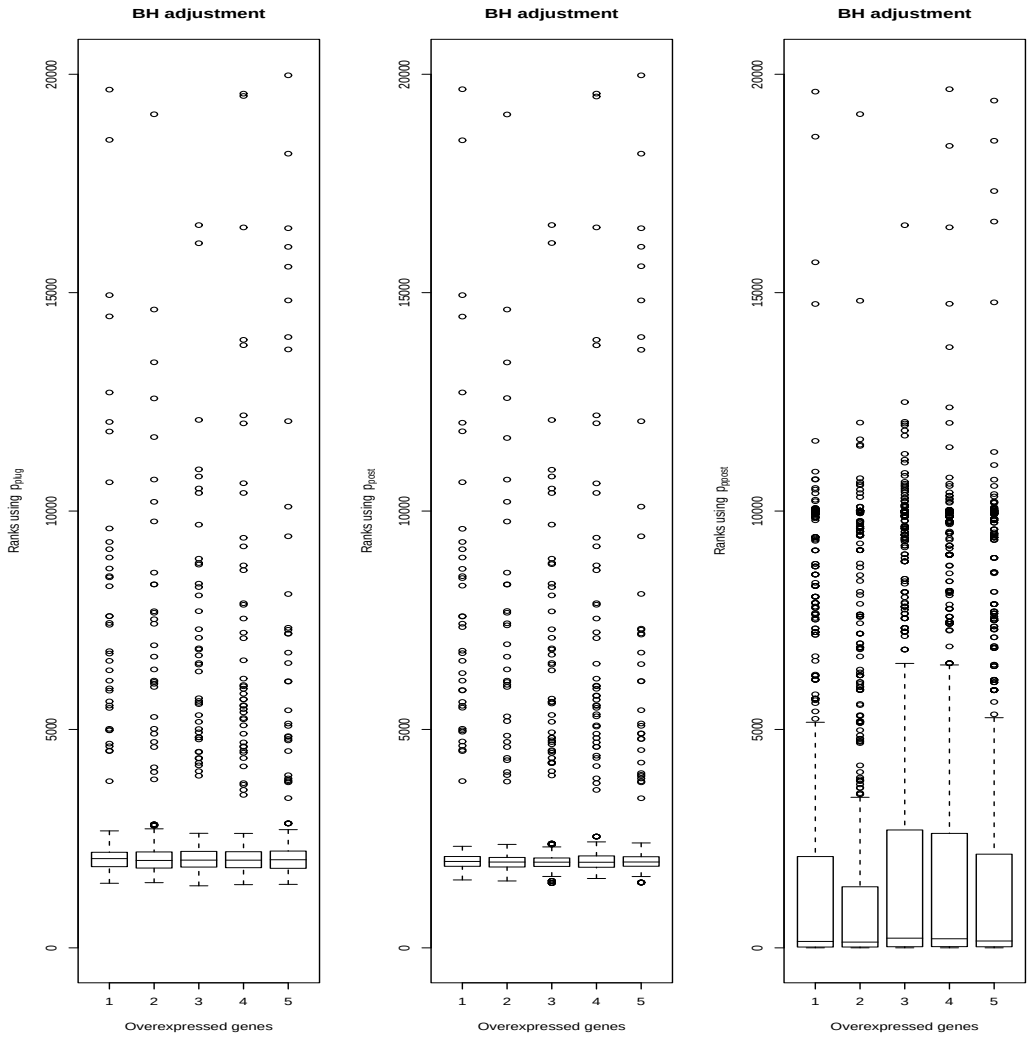

Figure 3.10: Normal model (Dependency case, $\rho = 0.9$): Distribution of the ranks of the 5 overexpressed genes with the p-values adjusted according to the BH's procedure. We can see that the boxplots referring to the p_{ppost} are more concentrated around zero than those for other p-values. In order to recognize the 5 overexpressed genes the boxplots should be concentrated around 0.

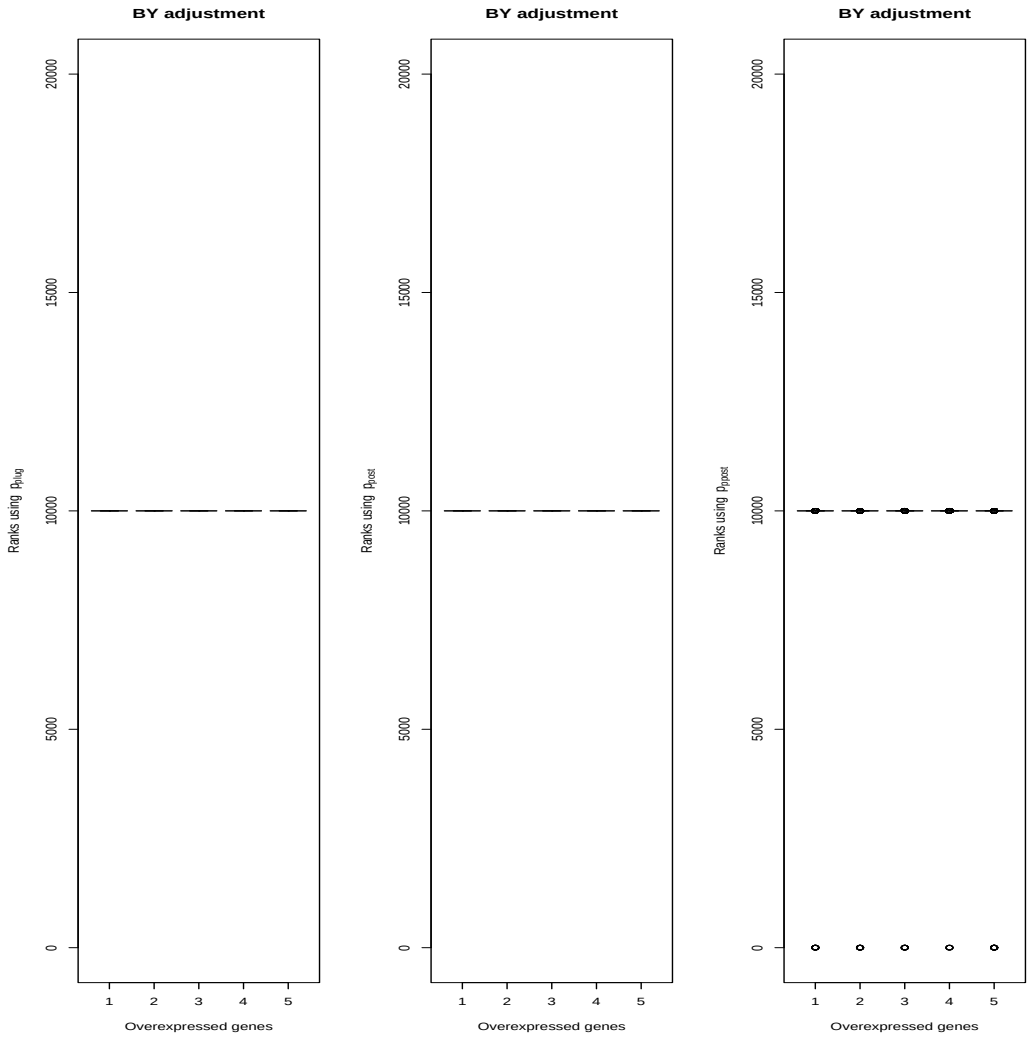

Figure 3.11: Normal model (Dependency case, $\rho = 0.9$): Distribution of the ranks of the 5 overexpressed genes with the p-values adjusted according to the BY's procedure. We can see that the boxplots are concentrated even if some spikes appear for p_{ppost}. In order to recognize the 5 overexpressed genes the boxplots should be concentrated around 0.

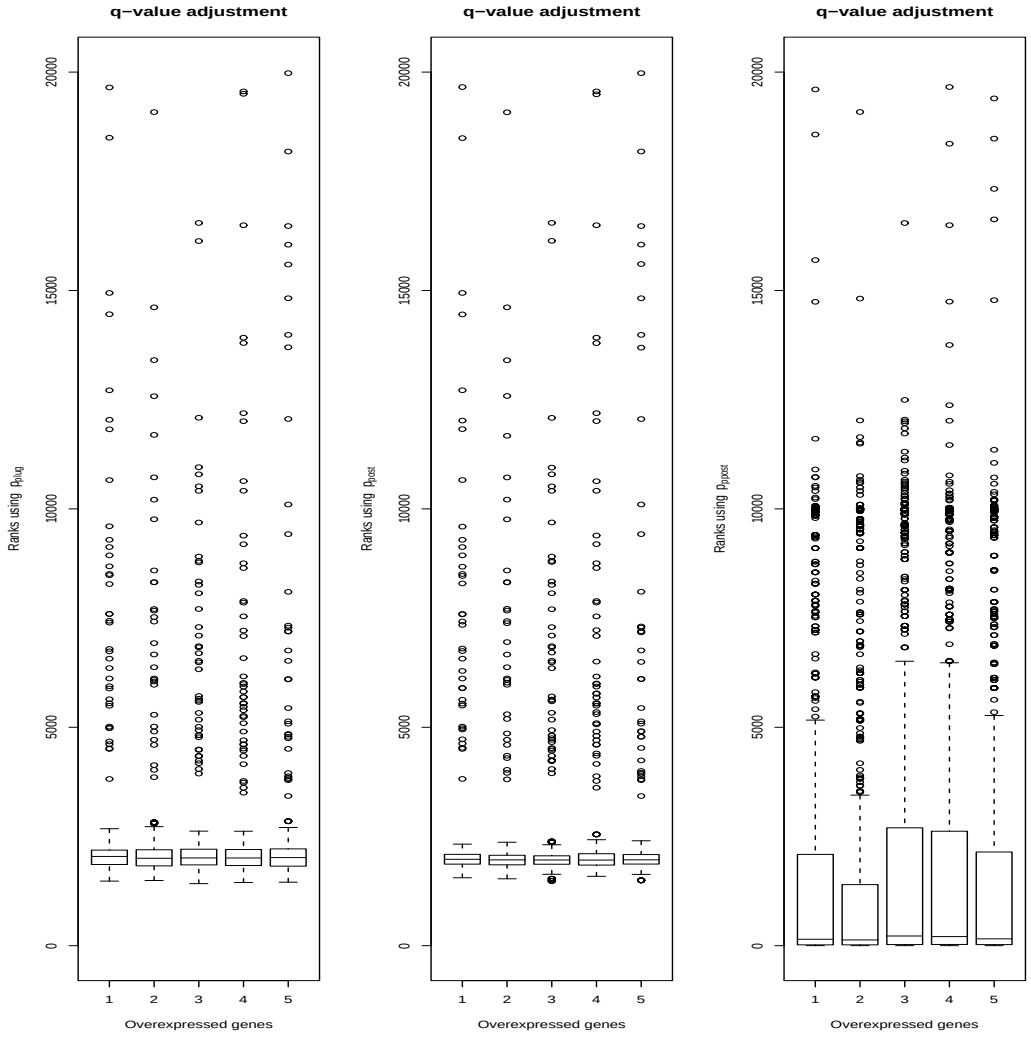

Figure 3.12: Normal model (Dependency case, $\rho = 0.9$): Distribution of the ranks of the 5 overexpressed genes with the p-values adjusted according to the q-values. We can see that the boxplots referring to the p_{ppost} are more concentrated around zero than those for other p-values. In order to recognize the 5 overexpressed genes the boxplots should be concentrated around 0.

3.2.2 Results for the Gamma model using different controlling procedures across different notions of p-value

We simulated $B = 100$ microarray experiments of $m = 1000$ genes each. Each independent experiment has been replicated five times: $n_X = n_Y = 5$. Each replication in the case has been drawn independently according to:

$$\mathbf{X}_i \sim Gamma\,(1, \theta_i)$$

where

$$\theta_1 = 20.00, \theta_2 = 16.25, \theta_3 = 12.50, \theta_4 = 8.75, \theta_5 = 5.00$$

and

$$\theta_{5<i\leq m} = 1.00.$$

Each replication in control has been drawn according to

$$\mathbf{Y}_i \sim Gamma\,(1, 1)\,, \quad i = 1, \ldots, m.$$

In this simulation setup $\mathbf{X}$ and $\mathbf{Y}$ exhibit the same relative error, in particular the choice of $CV = 1$ correspond to experiments where the amount of signal equals the amount of noise. This situation is commonly encountered in practical cases.

The number of simulated experiment, $(B = 100)$, maybe somewhat low, but we replicated another time the simulations and we did not find any significative differences from the results reported in Figures 3.13-3.18.

The results for the gamma model lead to the same conclusions as for the normal model: they corroborated the conjecture, because we recognize the five altered genes using the q-values and the BH combined with the p_{ppost}. Moreover, in this case, it is even more clear that using the p_{ppost} we are more likely to detect the five genes also with very conservative methods that control the $FWER$ rather than the FDR.

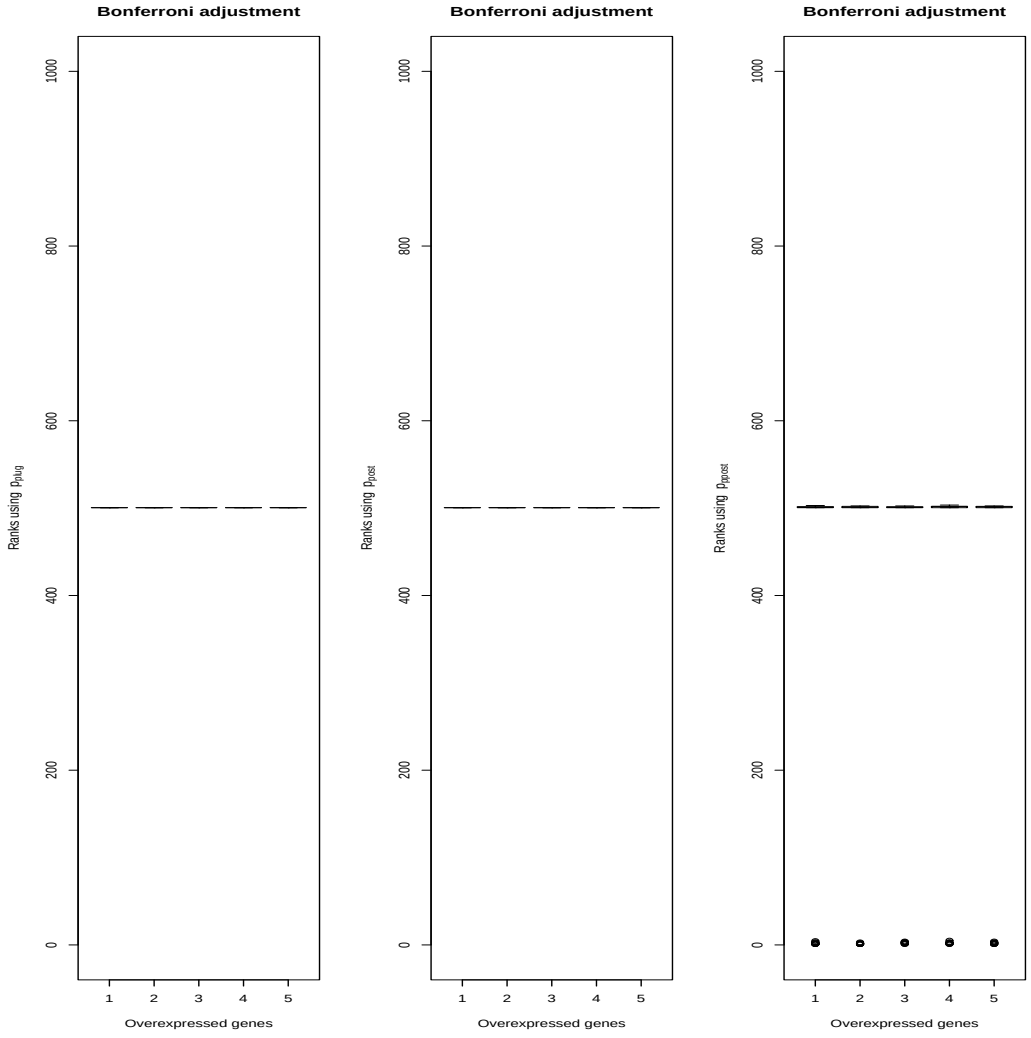

Figure 3.13: Gamma model: Distribution of the ranks of the 5 overexpressed genes with the p-values adjusted according to the Bonferroni's procedure. We can see that the boxplots are concentrated even if some spikes appear for p_{ppost}. In order to recognize the 5 overexpressed genes the boxplots should be concentrated around 0.

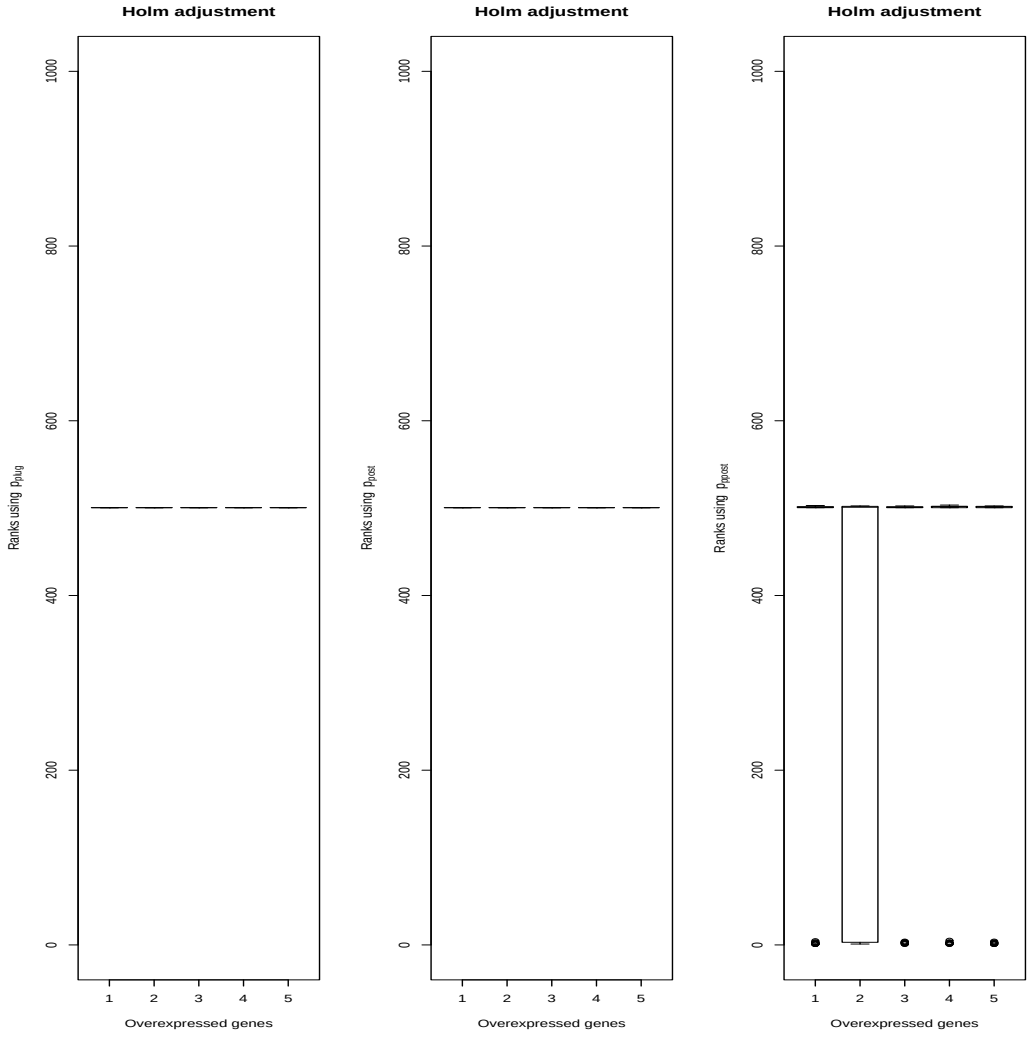

Figure 3.14: Gamma model: Distribution of the ranks of the 5 overexpressed genes with the p-values adjusted according to the Holm's procedure. We can see that the boxplots are concentrated even if some spikes appear for p_{ppost}. In order to recognize the 5 overexpressed genes the boxplots should be concentrated around 0.

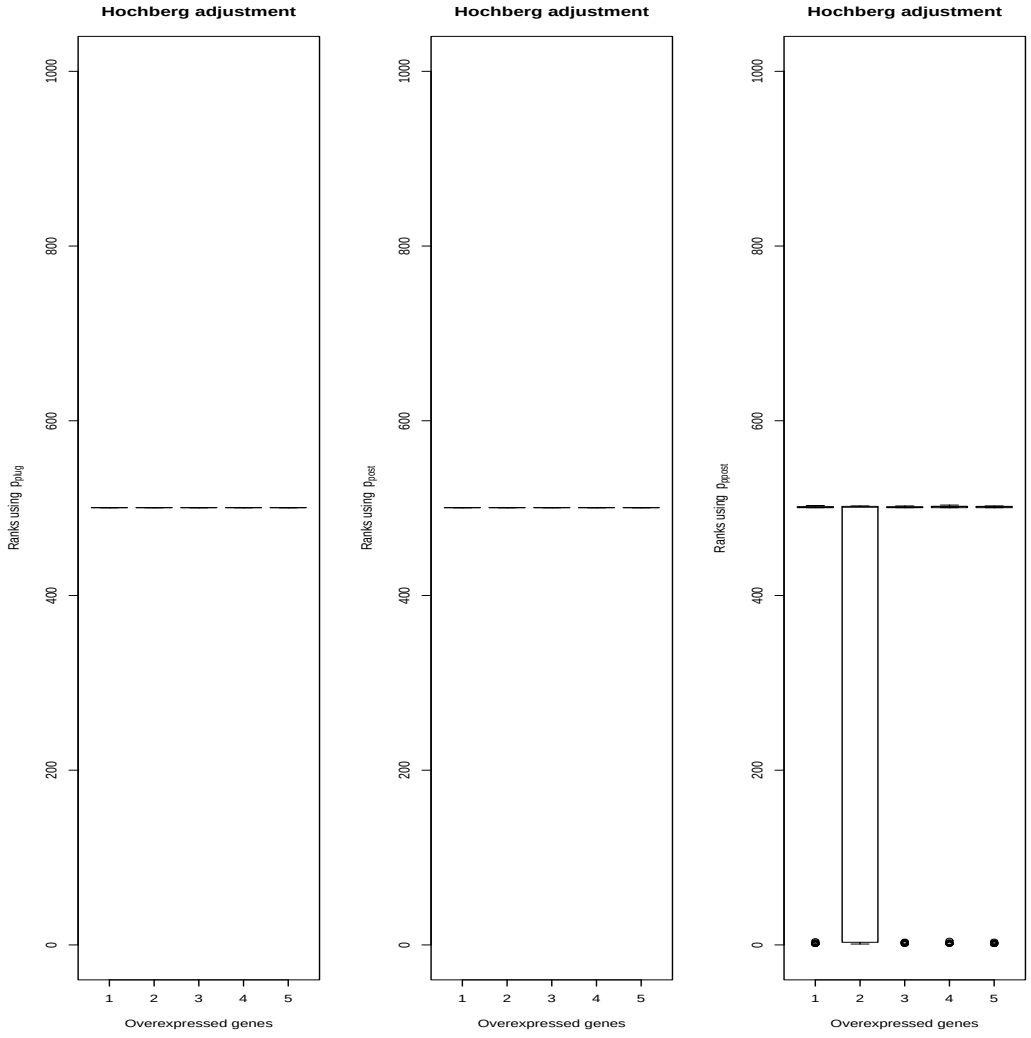

Figure 3.15: Gamma model: Distribution of the ranks of the 5 overexpressed genes with the p-values adjusted according to the Hochberg's procedure. We can see that the boxplots are concentrated even if some spikes appear for p_{ppost}. In order to recognize the 5 overexpressed genes the boxplots should be concentrated around 0.

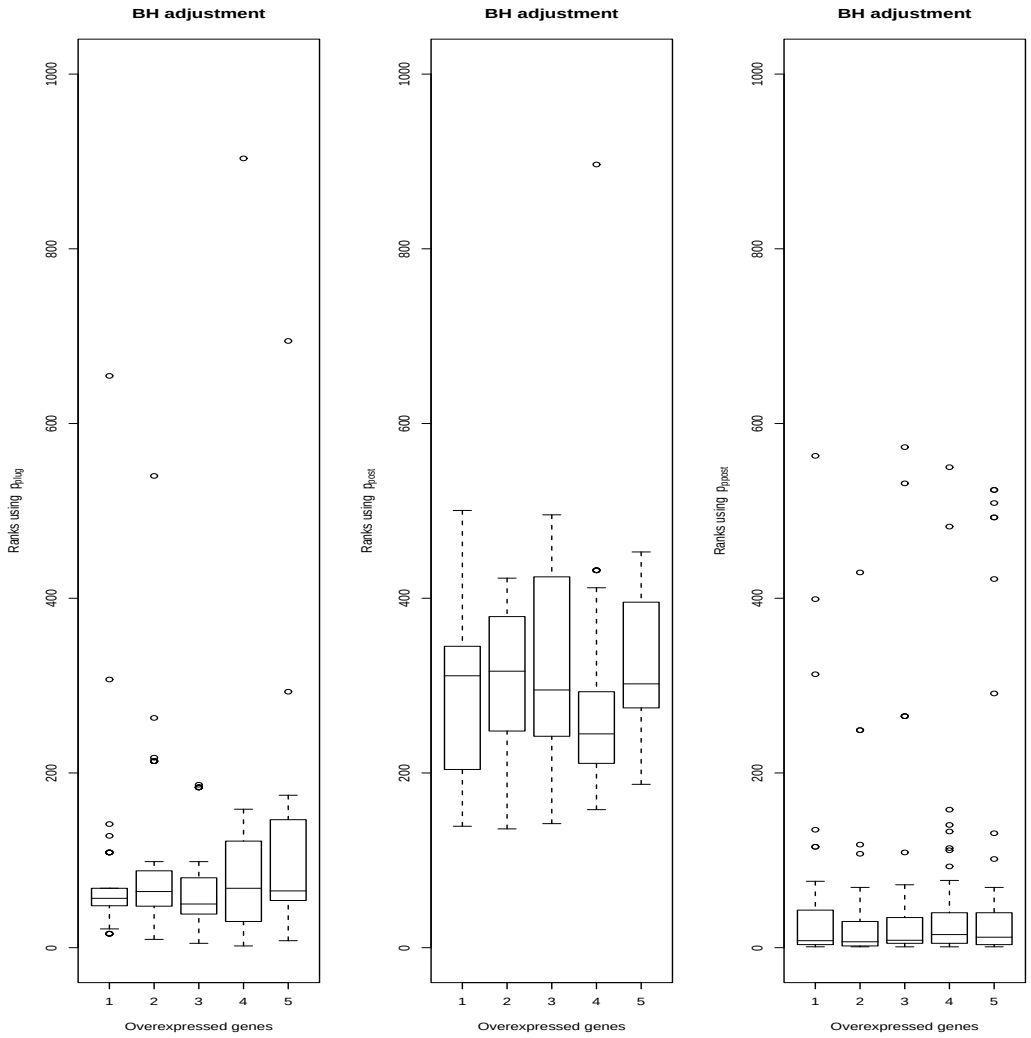

Figure 3.16: Gamma model: Distribution of the ranks of the 5 overexpressed genes with the p-values adjusted according to the BH's procedure. We can see that the boxplots referring to the p_{ppost} are more concentrated around zero than those for other p-values. In order to recognize the 5 overexpressed genes the boxplots should be concentrated around 0.

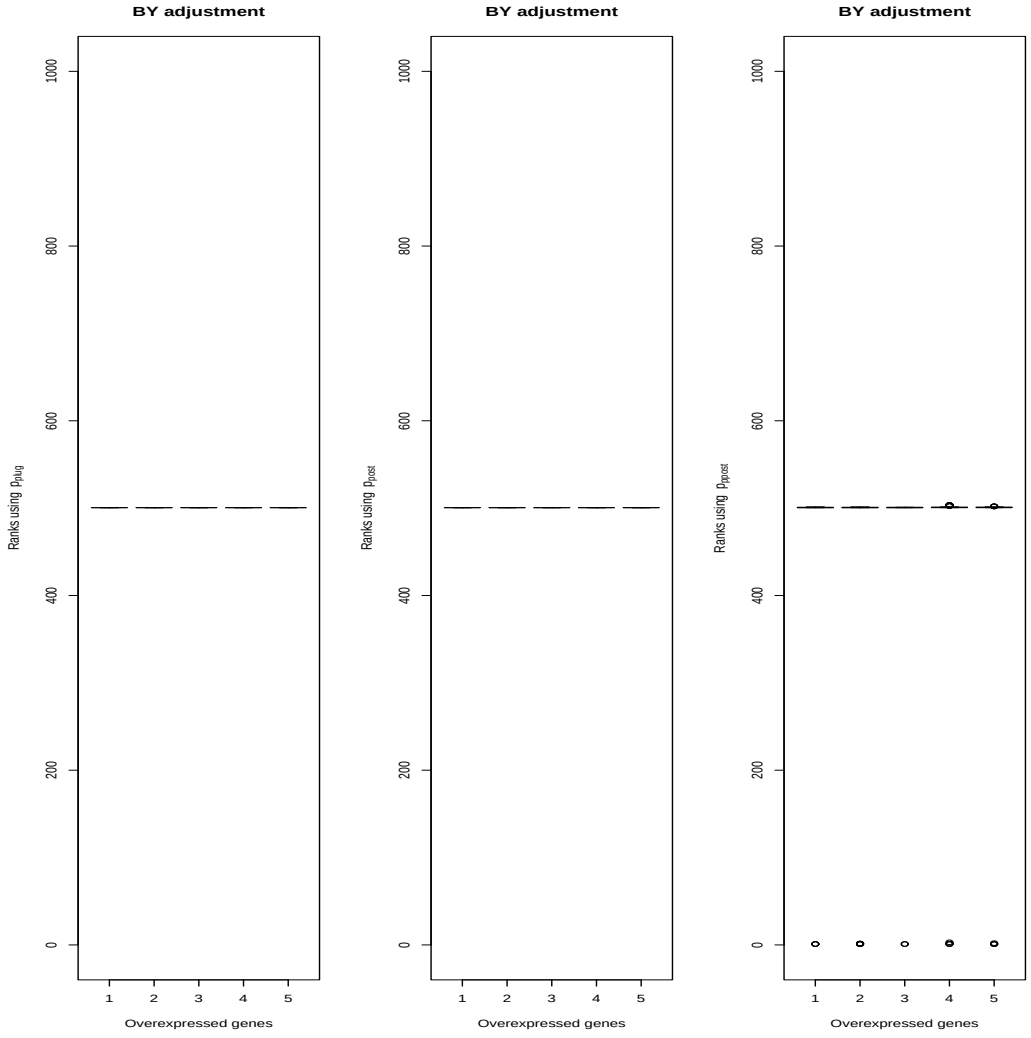

Figure 3.17: Gamma model: Distribution of the ranks of the 5 overexpressed genes with the p-values adjusted according to the BY's procedure. We can see that the boxplots are concentrated even if some spikes appear for p_{ppost}. In order to recognize the 5 overexpressed genes the boxplots should be concentrated around 0.

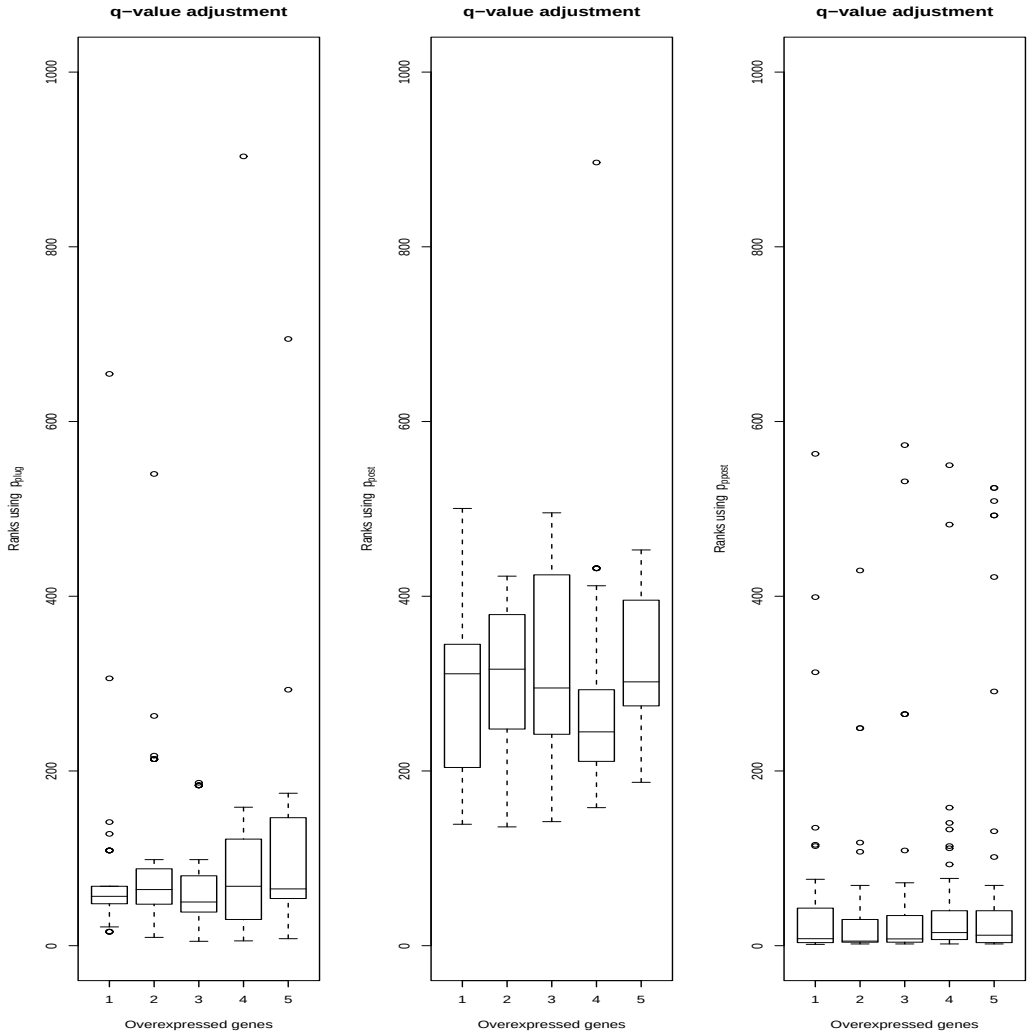

Figure 3.18: Gamma model: Distribution of the ranks of the 5 overexpressed genes with the p-values adjusted according to the q-value procedure. We can see that the boxplots referring to the p_{ppost} are more concentrated around zero than those for other p-values. In order to recognize the 5 overexpressed genes the boxplots should be concentrated around 0.

3.3 Applications to three public data sets

We consider here an application of the Gamma model to three public and well known data sets.

The Swirl data set is usually used as an example to show some issues in normalization across arrays and within the array, because we have to take into account the print-tips effects in cDNA experiments. The goal here in using this data set is to show that only by using the p_{ppost} and the q-values we are ably to detect those genes that we expect to be differentially expressed.

The second data set is the Golub data set which is an extended study on Leukemia. The experiment was designed to find a classification role for two different kinds of Leukemia and even discover other type of Leukemias. A classifier for the two kinds of Leukemia is constructed using the expression profile of a small number of genes. The goal, in this case, is to show that only the p_{ppost} allows us to find those genes that can be fruitfully used to discriminate among the two types of Leukemia.

The third data set is what is usually referred as a controlled experiment where sixteen genes have been spiked in at different known concentrations in different hybridizations and they are thus differentially expressed. Therefore among these sixteen genes some are very likely to be detected while other less. For this data set we applied the gamma and the normal model and we show that only the gamma model along with q-values with the p_{ppost} is able to detect those genes with altered expression.

3.3.1 The Swirl data set

This experiment was carried out with cDNA arrays using zebrafish as a model organism to study early development in vertebrates and to determine the gene that distinguish the vertebrate wild-type zebrafish (wt), from the invertebrate swirl mutant.

Swirl is a point mutant in the BMP2 gene. The BMP2 gene causes defects in the organization in the developing embryo along its dorsal-ventral axis. When BMP2 is not expressed then ventral fates such as blood are reduced, whereas dorsal structures such as somites and notochord are expanded.

The data come from four replicate slides: two sets of dye-swap pairs. For each of these slides the target cDNA from the swirl mutant was labelled using one of the Cy3 and Cy5 dyes and the same has been done for the cDNA from the wild-type zebrafish. Target cDNA was hybridized to microarrays containing 8,448 cDNA probes. The microarrays were printed using 4×4 print-tips and are thus partitioned into 4×4 matrix of subarrays. Each sub-array consists of 22×24 spot matrix that was printed with a single print-tip. The expression quantities from each probe were measured using the GenePix software (`http://www.axon.com`). These data were provided by Katrin Wuennenberg-Stapleton from the Ngai Lab at UC Berkeley (U.S.A.). The swirl embryos for this experiment were provided by David Kimelman and David Raible at the University of Washington (U.S.A.).

This well known experiment is the case study number one in Speed *et al.* (2003). It is often used to show the main issues in normalization of raw data from the experiment,

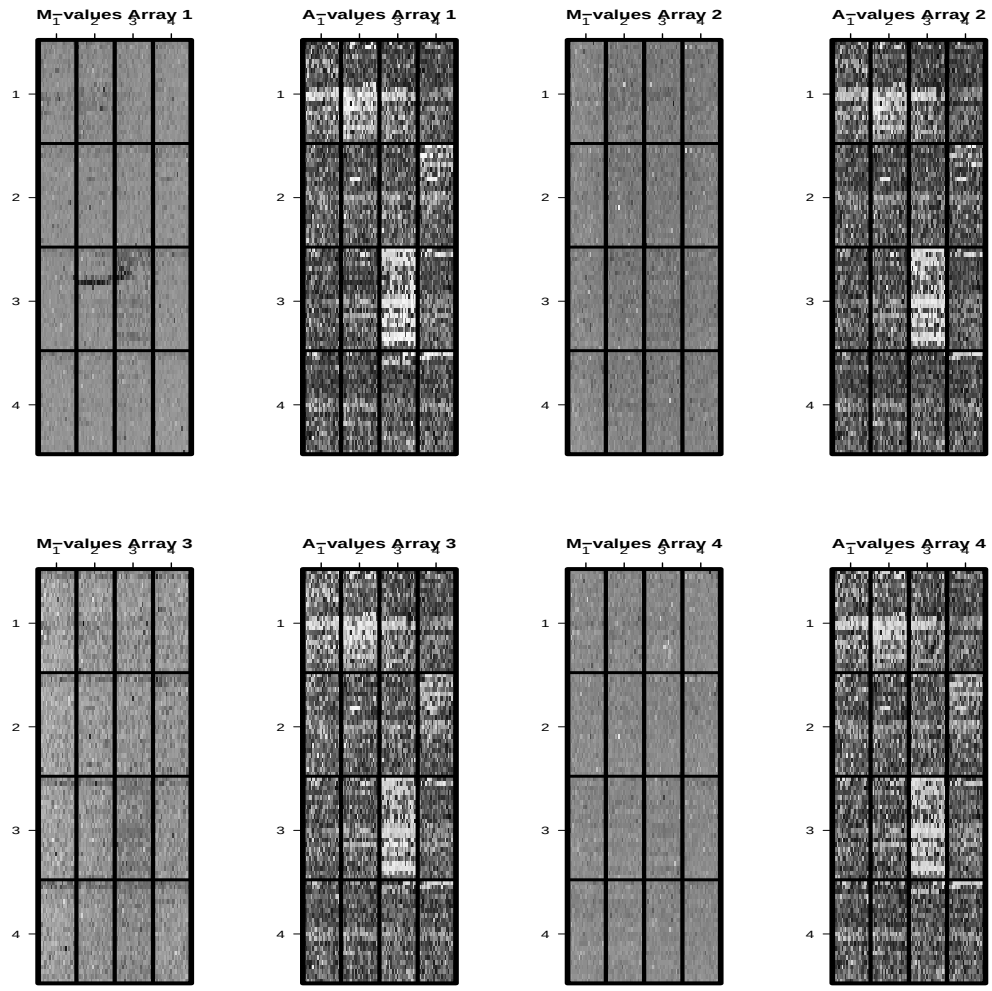

Figure 3.19: Analysis of Swirl data set: M and A values for the 6 array under analysis. The grids on the array represent different subarrays spotted with different print tips. It is evident a print tip effect, a spatial effect (on the edges of the array) and an array effect, in fact some arrays are systematically brighter than others.

because it is evident the print-tip effects and the array effect as shown in Figure 3.19. These artifacts can be modelled and then removed by using non-parametric regression techniques with splines. Here we will use the normalized data and we sent the reader to the review book of Speed *et al.* (2003) for more details on the normalization process for this data set.

We modelled the experiment outcome according to the Gamma model and for each of 8,448 genes we computed the p_{plug}, p_{post} and p_{ppost}. We then computed the adjusted p-values and the q-values using the (2.24) which takes into account dependency in the tests. We compare the results obtained from our ranks to the rank obtained using only the $M = \log_2 R/G$ values computed on the normalized data set. The goal here is to show that the q-value calculated on p_{ppost} can be useful in order to detect the BMP2 gene.

Table 3.3 shows the list of the top 10 genes ordered according to the observed M values. We can see that the BMP2 gene is in the list and it is located in position 27

Gene names	Rank of p_{plug}	p_{plug}	Rank of p_{post}	p_{post}	Rank of p_{ppost}	p_{ppost}
18-F10	843	1.000	843	1.000	8	0.045
BMP2	1608	1.000	1608	1.000	27	0.203
Dlx3	28	1.000	28	1.000	1	0.000
6-I1	61	1.000	61	1.000	29	0.203
11-L19	1300	1.000	1300	1.000	6	0.000
7-K10	699	1.000	699	1.000	20	0.174
7-K22	702	1.000	702	1.000	820	1.000
18-G18	845	1.000	845	1.000	30	0.203
vent	2150	1.000	2150	1.000	21	0.184
7-G6	698	1.000	698	1.000	20	0.174

Table 3.3: Analysis of Swirl data set: rank of the genes according to the M values. The q-values with the p_{ppost} provide the greatest evidence for the BMP2 gene to be differentially expressed in the two organisms.

using the q-values and p_{ppost}, while using the other p-values we have to consider the first 1608 genes in order to find this gene. The declared $pFDR$ of gene BMP2 is 0.2 which means that, given the data, this is the probability of making a false discovery by declaring as differentially expressed the BMP2 gene and other 26 genes. This probability may seems too high, but it can be accepted if we are in a merely exploratory data analysis where instead of bounding the probability of making a mistakes we are essentially interested in the list of the most differentially expressed genes. So usually a list from 10 to 50 genes is selected and then analyzed with other laboratory techniques.

Using the other MHT methods we are not able to find the BMP2 gene in the first positions. The behavior of these methods is shown in Figures 3.20, 3.21 and 3.22.

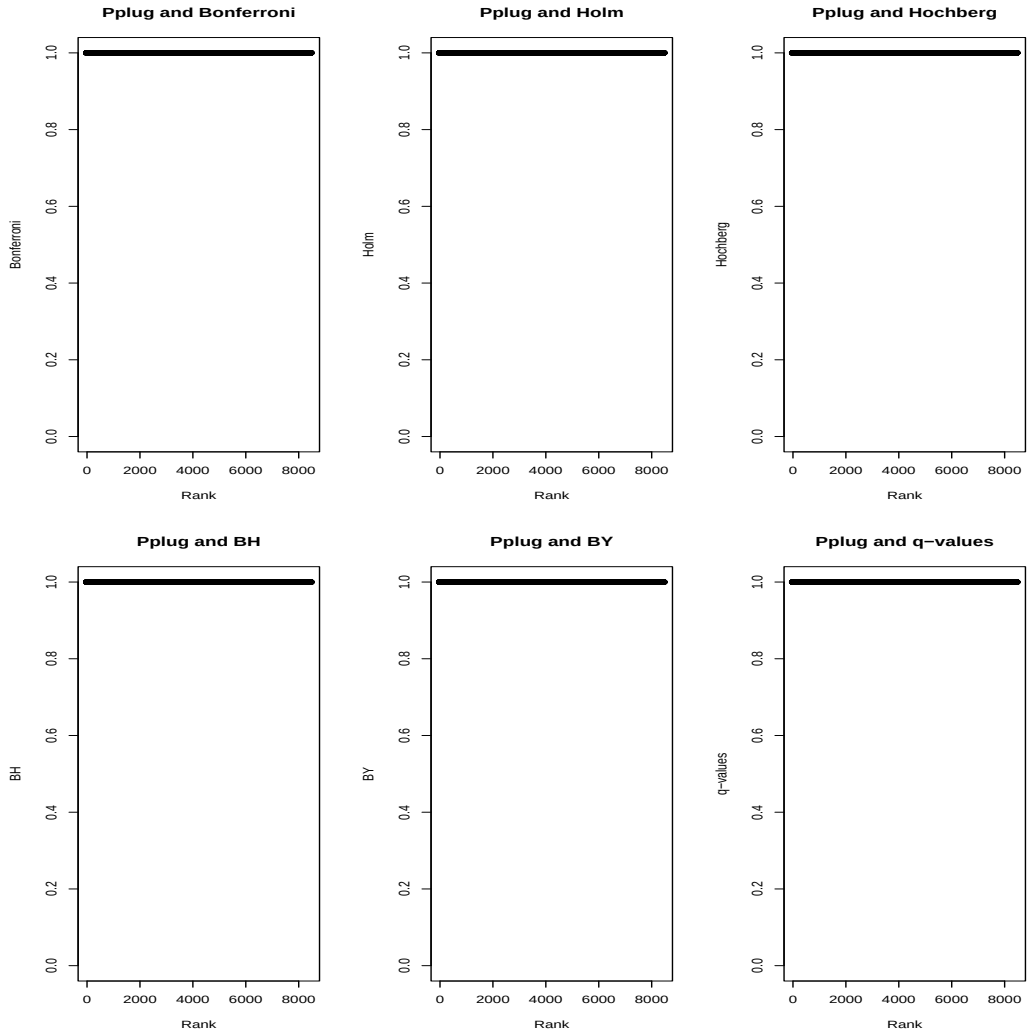

Figure 3.20: Analysis of Swirl data set: MHT with p_{plug}. We can see that after the p-value adjustment no gene is suspected to be differentially expressed.

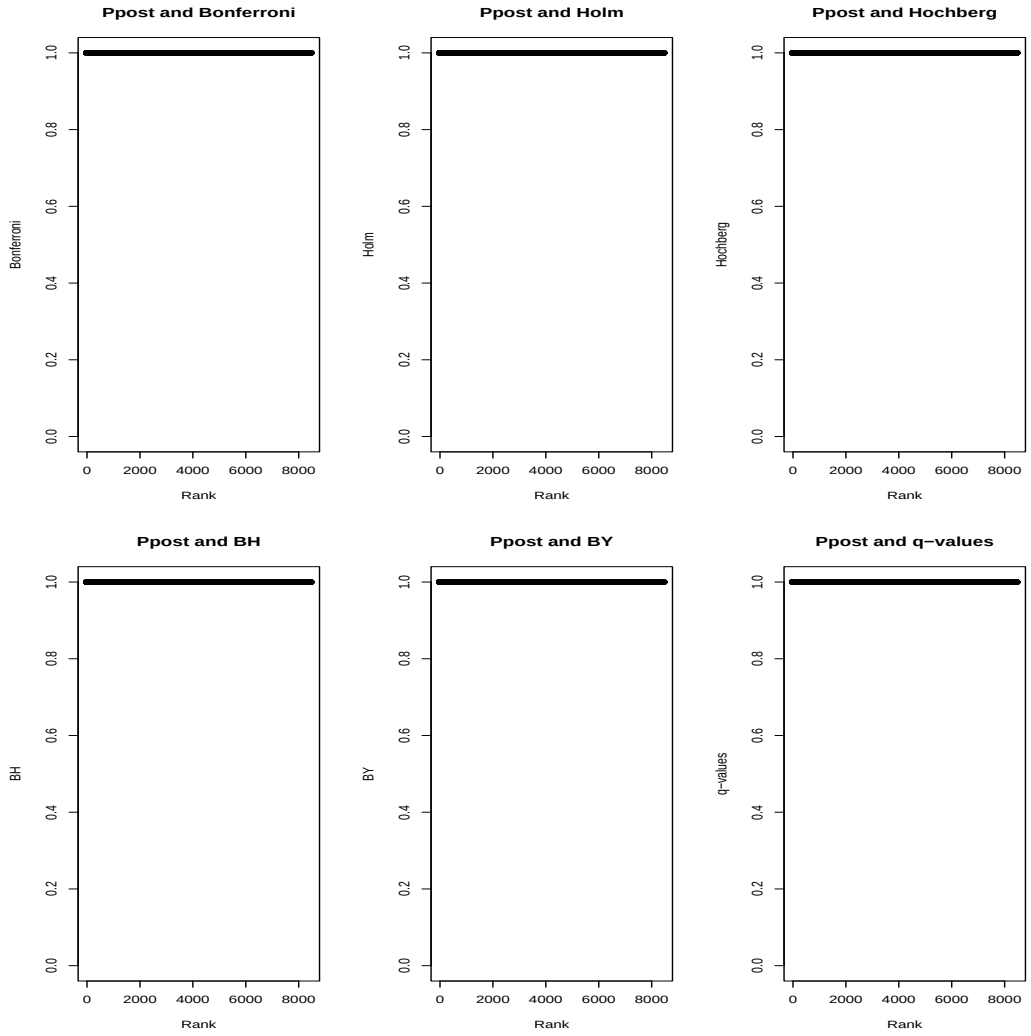

Figure 3.21: Analysis of Swirl data set: MHT with p_{post}. We can see that after the p-value adjustment no gene is suspected to be differentially expressed.

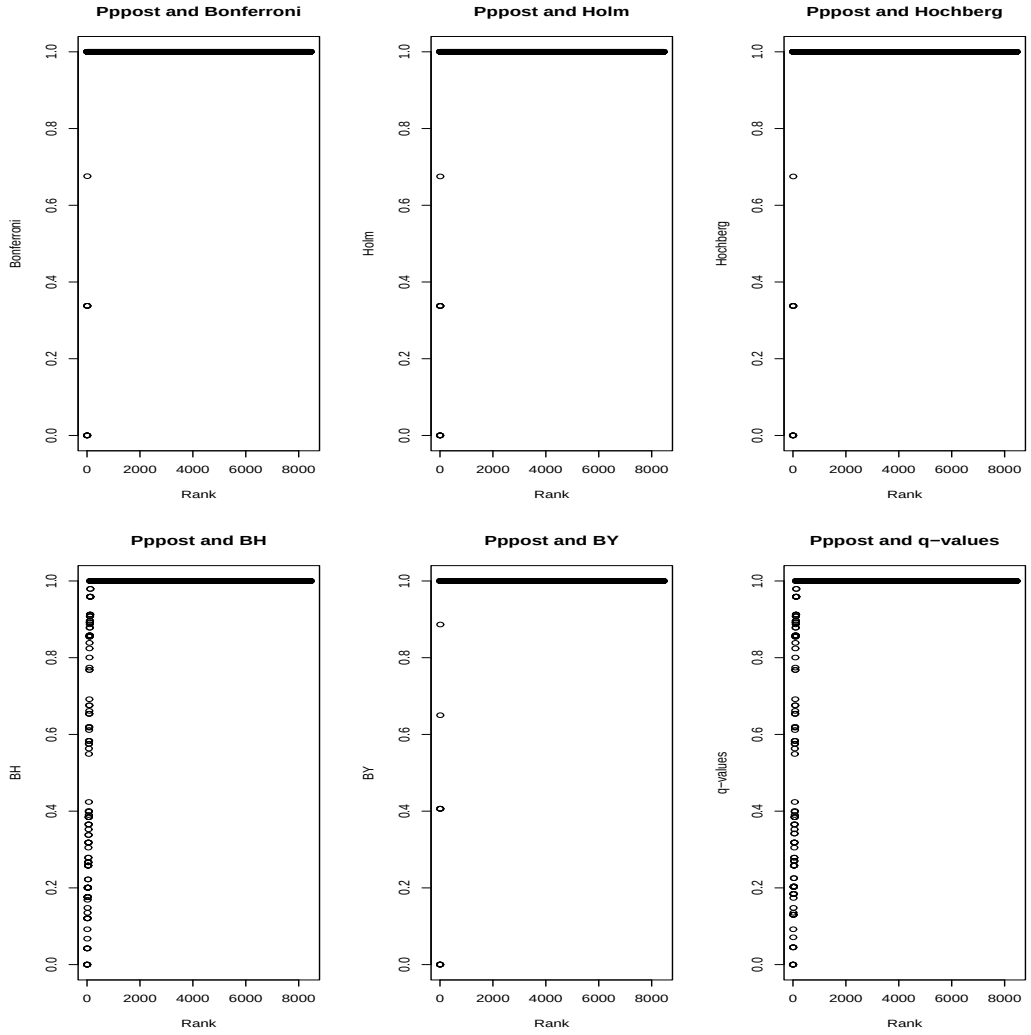

Figure 3.22: Analysis of Swirl data set: MHT with p_{ppost}. We can see that we using the BH's procedure and the q-value we can find many altered genes. Among this there is the BMP2 gene.

3.3.2 The Golub data set

The goal of this experiment is to find a suitable subset of genes that allows to make cancer classification. In particular we want to discover which genes are able to differentiate the Acute Lymphoblastic Leukemia (ALL) from the Acute Myeloid Leukemia (AML). It is important to analyze gene expression profile of human blood in order to predict which kind of Leukemia a patient is exhibiting, because they are treated differently. A correct classification give the opportunity to increase the probability of a patient to survive from Leukemia.

The experiment was performed and first analyzed by Golub *et al.* (1999). It was carried out by using 72 Affymetrix U68 gene chips. Each chip produce measurements for 7129 different expressed sequence tags (ESTs). Some of these ESTs map to the same genes and others are used for quality control purposes, this lead to 6817 different genes mapped on the array. The data are available at the following web site: `http://www.genome.wi.mit.edu/MPR`. We choose to use a subset of 56 chips divided into two subsets:

i) a training subset that consist of 22 samples: 11 from patient that exhibit AML and 11 from those that exhibit ALL;

ii) a test subset that consist of 34 samples: 14 AML and 20 AML.

We proceed in this way: we use the training subset to select those genes that are the most differentially expressed, then we use this list in order to make class prediction in the test subset. We will suppose that the gene expression follows a gamma distribution and exhibit the same variation coefficient. The normalized AD quantity from the experiment are distributed around 0 and in order to have all positives number we simply shifted all the observations by $C = 1 + \min(\mathbf{x}, \mathbf{y})$. This translation does not alter the features of $\mathbf{x}$ and $\mathbf{y}$ distribution with respect to a common variation coefficient for $\mathbf{X}$ and $\mathbf{Y}$. Where $\mathbf{x}$ and $\mathbf{y}$ are the observed samples of 7129 genes of 11 replications for both the case and the control. We calculate the p-values on each gene using 11 replications, then we calculate the q-values and pick up the first 10 genes that provide the lowest q-values. We use these genes in order to make class prediction using the k-Nearest Neighbor Classifier as described in Chapter 1. In particular, we used $k = 3$ neighborhoods to assign a test set to a class. We are interested in making class prediction using the smallest number of genes, because the larger the number of genes we use the more is the likely to make unreliable predictions due to gene's expression measurement uncertainty. Figure 3.23 shows the number of misclassified cases using up to the first 10 genes. It is possible to see that the genes picked up according to the q-values calculated on p_{ppost} are those that provide the lowest number of misclassified cases.

We underline that we are not saying that the q-values calculated by using p_{ppost} are tailed to make class prediction, but that they just select relevant genes for further statistical analysis, such as classification. We again remark that the p-values should be employed in an merely exploratory data analysis.

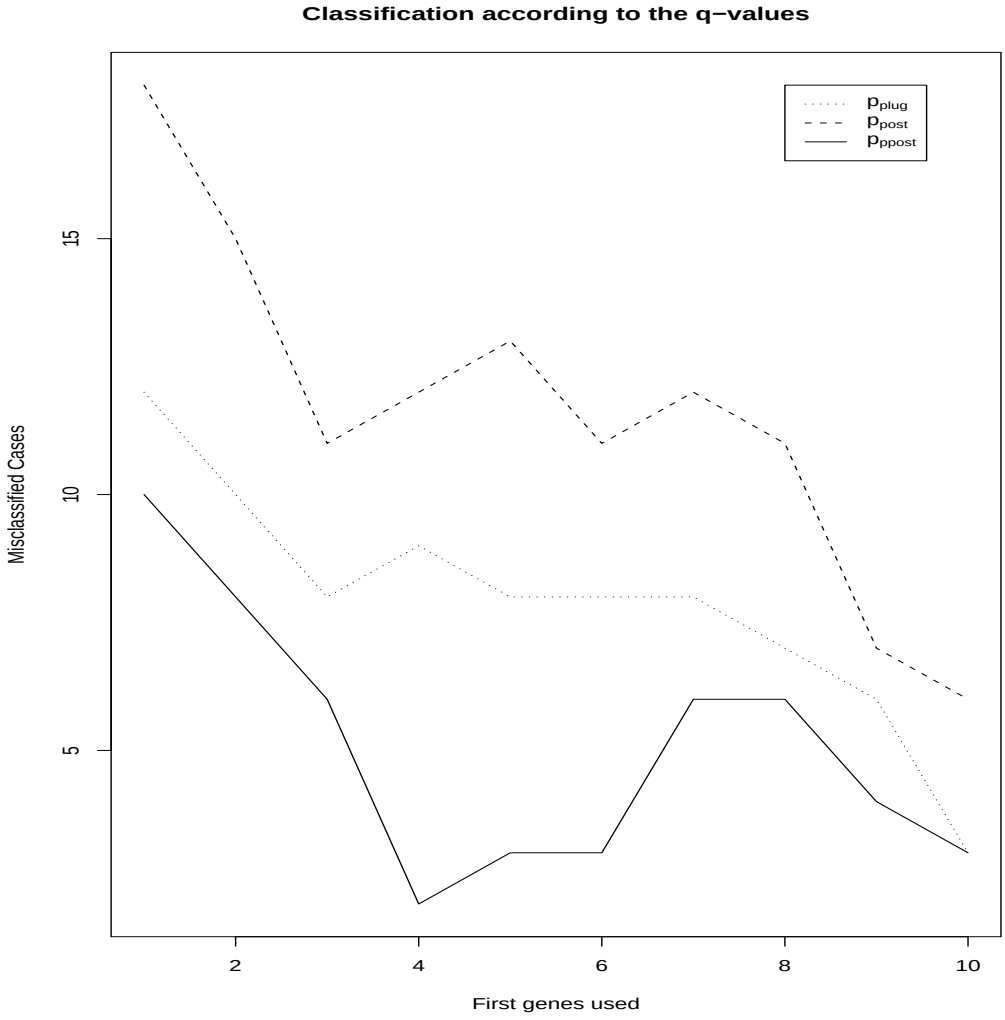

Figure 3.23: Analysis of Golub data set: number of misclassified cases according to the q-values and the three p-values. We can see that in the training set the genes detected with the p_{ppost} lead to a smaller number of misclassified cases.

Gene names	Conc. in Pop.1	Conc. in Pop.2
684-at	0.00	1024.00
38734-at	0.25	0.50
39058-at	0.50	1.00
36311-at	1.00	2.00
36889-at	2.00	4.00
1024-at	4.00	8.00
36202-at	8.00	16.00
36085-at	16.00	32.00
40322-at	32.00	64.00
1091-at	128.00	256.00
1708-at	256.00	512.00
33818-at	32.00	128.00
546-at	8.00	16.00
37777-at	512.00	1024.00
407-at	512.00	1024.00
1597-at	0.00	0.25

Table 3.4: Analysis of Eset3 data set: Concentrations of the sixty genes in the two reference populations.

3.3.3 The Eset3 data set

This data set consists of 6 HGU95a Affy chips, each containing 12626 genes. The data can be downloaded at the following `Bioconductor` web page `http://www.bioconductor.org/repository/Courses/bioclabs_0.1.zip`. Among these 12626 genes 16 have been altered in different concentrations as showed in Table 3.4. We have two reference populations and 3 replicates for each one. We may see that the gene `1597` is the most difficult to detect as differentially expressed, because the level of concentration is very low. However, this is in general true for all those genes that have extreme concentration levels either too low or too high. The MA-plot for all genes is showed in Figure 3.24. We can see that gene `684-at` is highly differentially expressed, because it lays out of the horizontal lines indicating the double fold change in the expression level.

In fact as pointed out above, their variability is related to their mean. Therefore they could be detected as differentially expressed if we take into account the variability instead of the variance. For this reason we compared the gamma model with the normal model where the nuisance parameter σ^2 represent the variance. The standard analysis on this data set is to consider the M and A values and model them as they were normally distributed. This lead to perform t-tests and adjust the p-values according to FDR and q-values. Here we show that the choice of the model is critical. In fact even if we consider the M and A as normally distributed we are not able to detect most of the 16 genes as shown in Figures 3.25-3.27. We have then to consider another model. By doing so we are more likely to face with the problem that we may not handle sufficient and ancillary statistics for the unknown parameter. This case is considered here by using the Gamma model. This can be solved by calculating the p_{ppost}. In fact

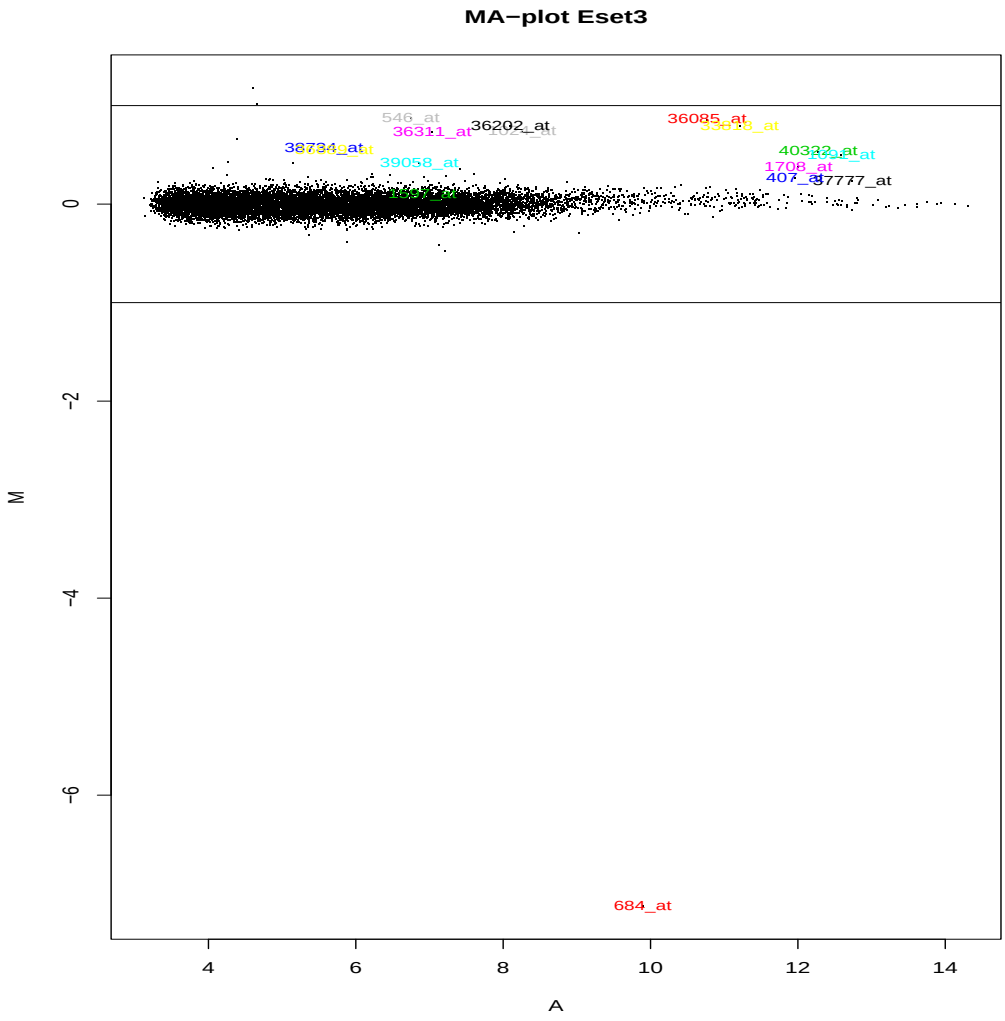

Figure 3.24: Analysis of Eset3 data set: MA-plot. We may quickly see that gene 684-at is differentially expressed.

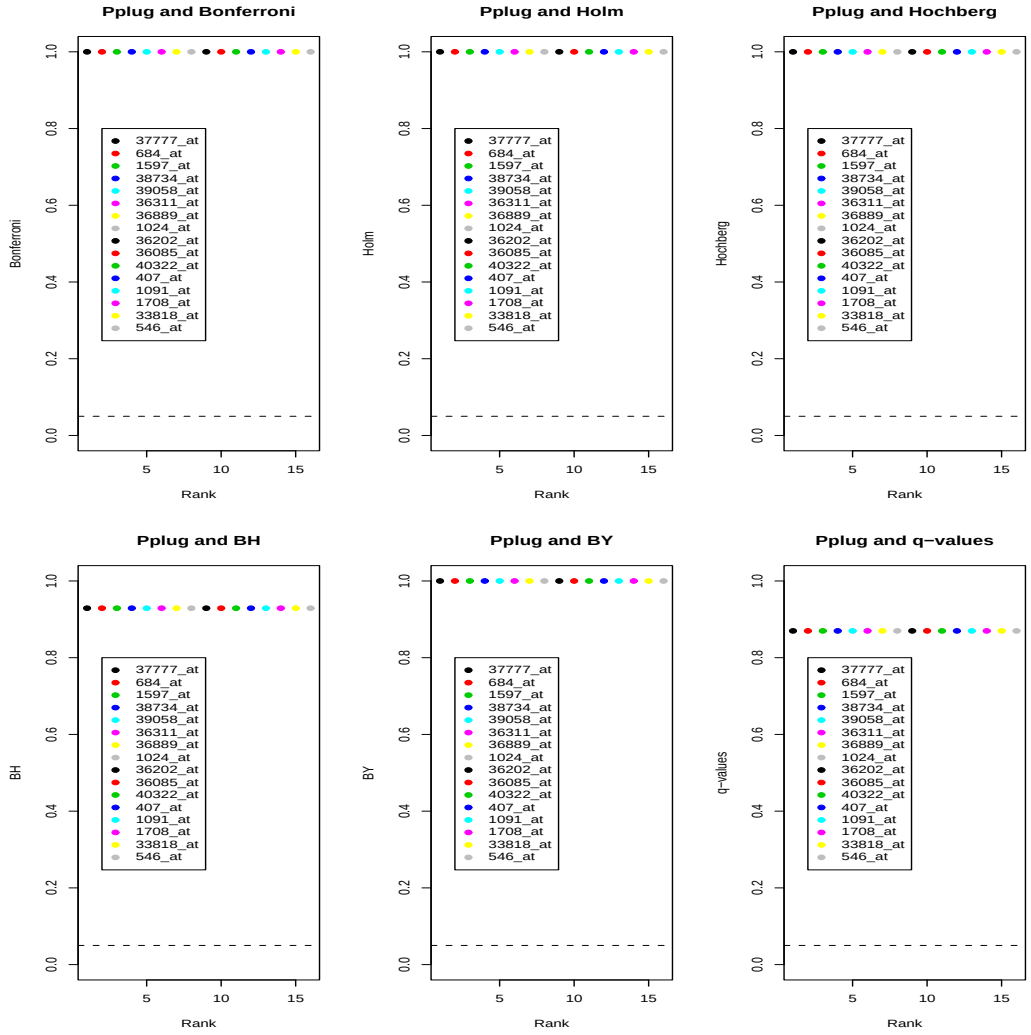

Figure 3.25: Analysis of Eset3 data set (Normal model): results of the MHT using the p_{plug}. The dotted line represent the reference error level 0.05. The legend reports the genes which are not significative differentially expressed.

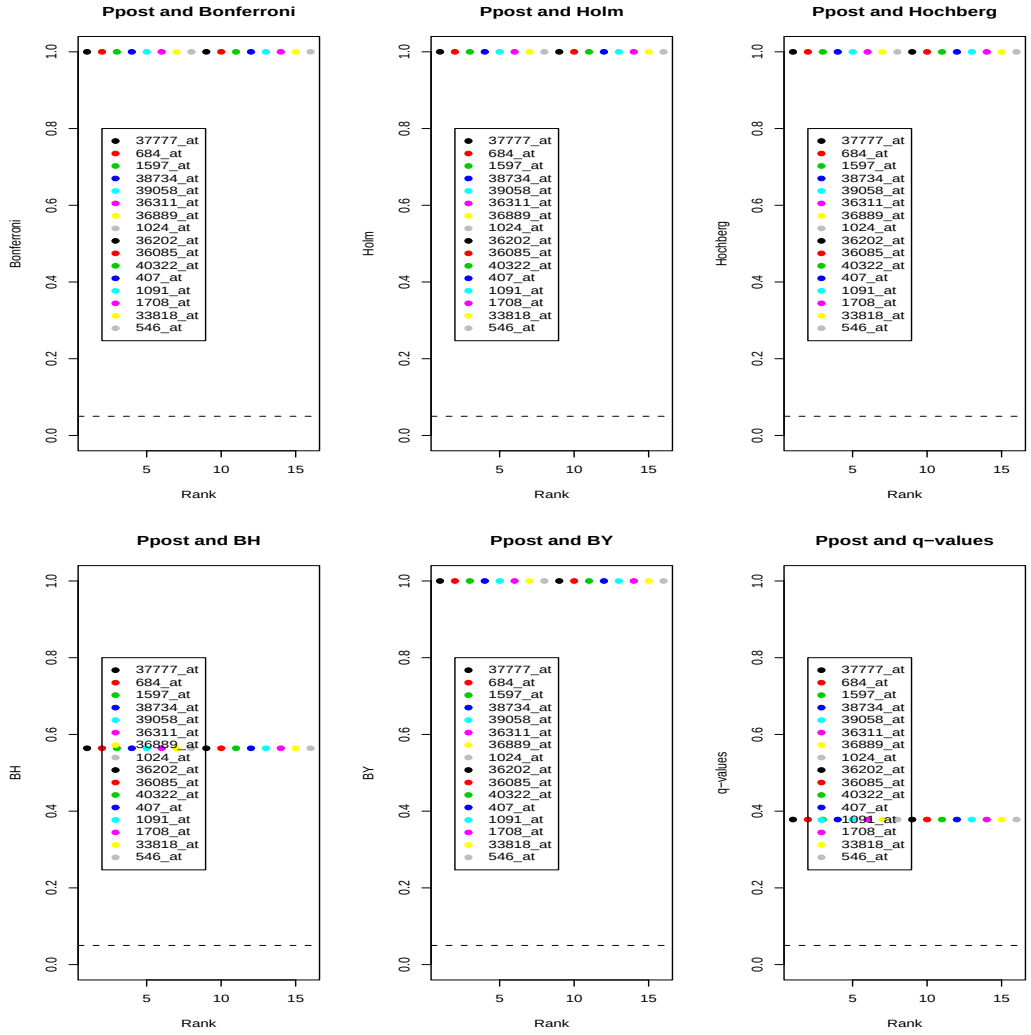

Figure 3.26: Analysis of Eset3 data set (Normal model): results of the MHT using the p_{post}. The dotted line represent the reference error level 0.05. The legend reports the genes which are not significative differentially expressed.

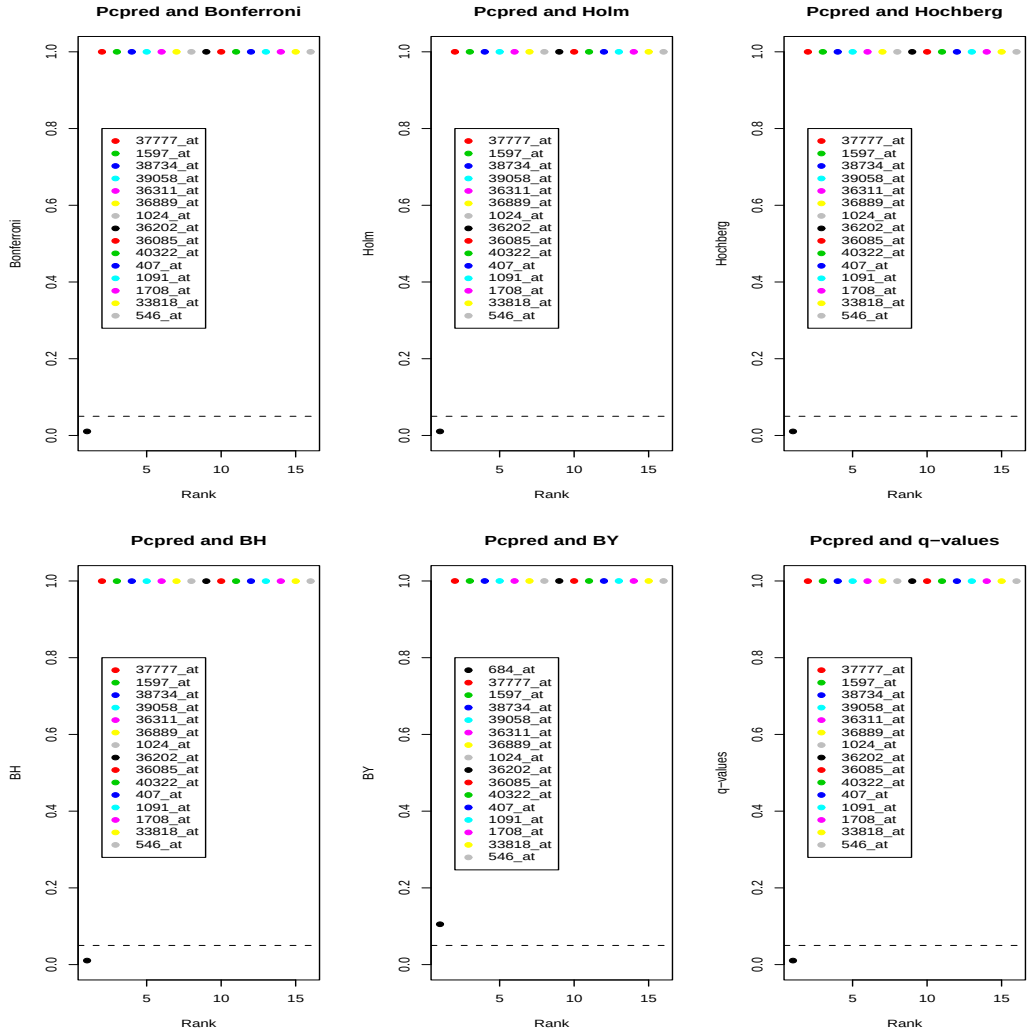

Figure 3.27: Analysis of Eset3 data set (Normal model): results of the MHT using the p_{cpred}. The dotted line represent the reference error level 0.05. The legend reports the genes which are not significative differentially expressed.

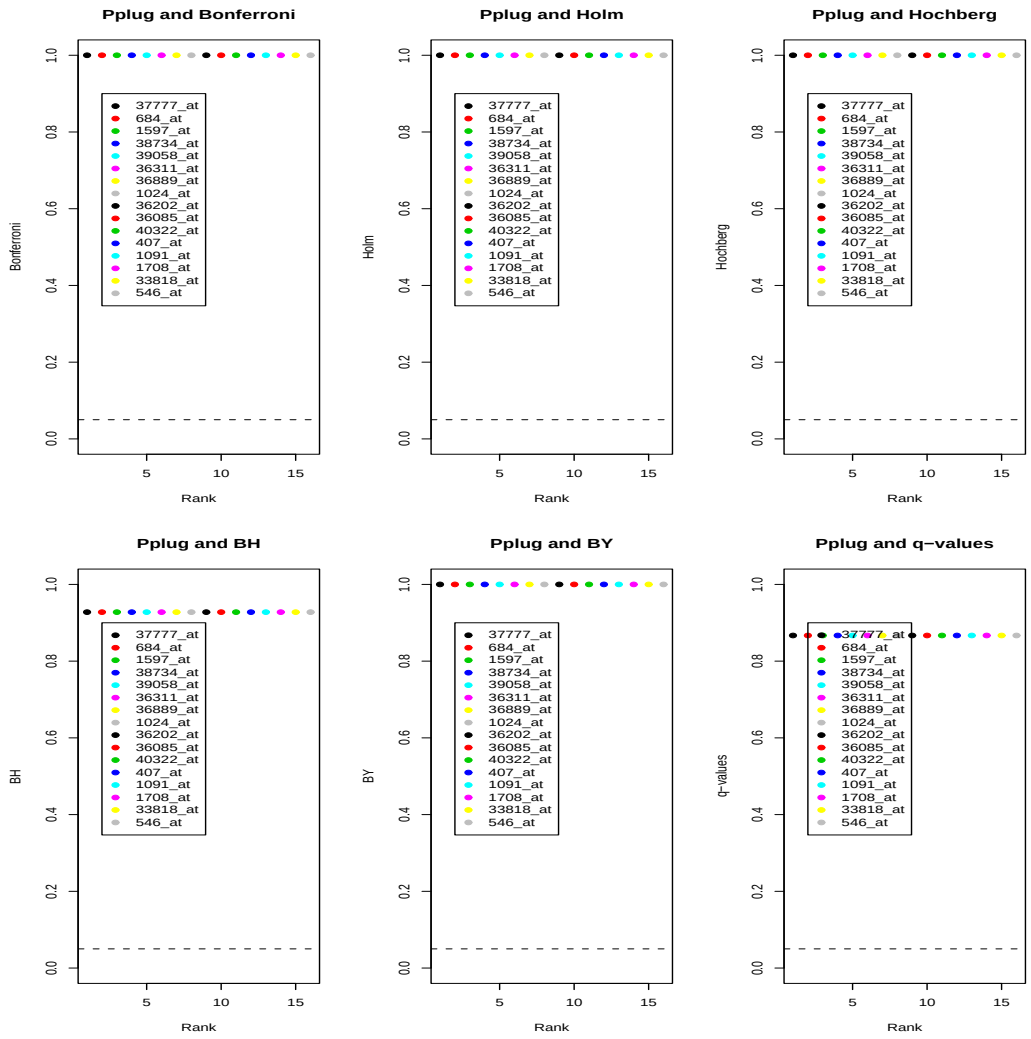

Figure 3.28: Analysis of Eset3 data set (Gamma model): results of the MHT using the p_{plug}. The dotted line represent the reference error level 0.05. The legend reports the genes which are not significative differentially expressed.

the behavior of the three p-values in the gamma model can be observed in Figures 3.28-3.30. We can see in Figure 3.30 that the combination of p_{ppost} and the q-values results in successfully detect almost all 16 genes. Comparing now the p_{ppost} with the other p-values we can see that they perform better because they do not make a double use of the data as the p_{post}, which is critical with 3 replications. Moreover using the p_{ppost} we are taking into account uncertainty on variability which is not considered in the p_{plug}.

In this data set the number of altered genes is very small with respect to the total number and then $\pi_0 = m_0/m = 16/12626 \approx 0$, so we find again that controlling the FDR or $pFDR$ correspond for this data set to weakly control the $FWER$. However we can see that the gain in power of the BH procedure and q-values can be achieved if we use an asymptotic frequentist p-value such as the Partial posterior Predictive p-value.

We also analyze a more extended version of this study which have 12 replications

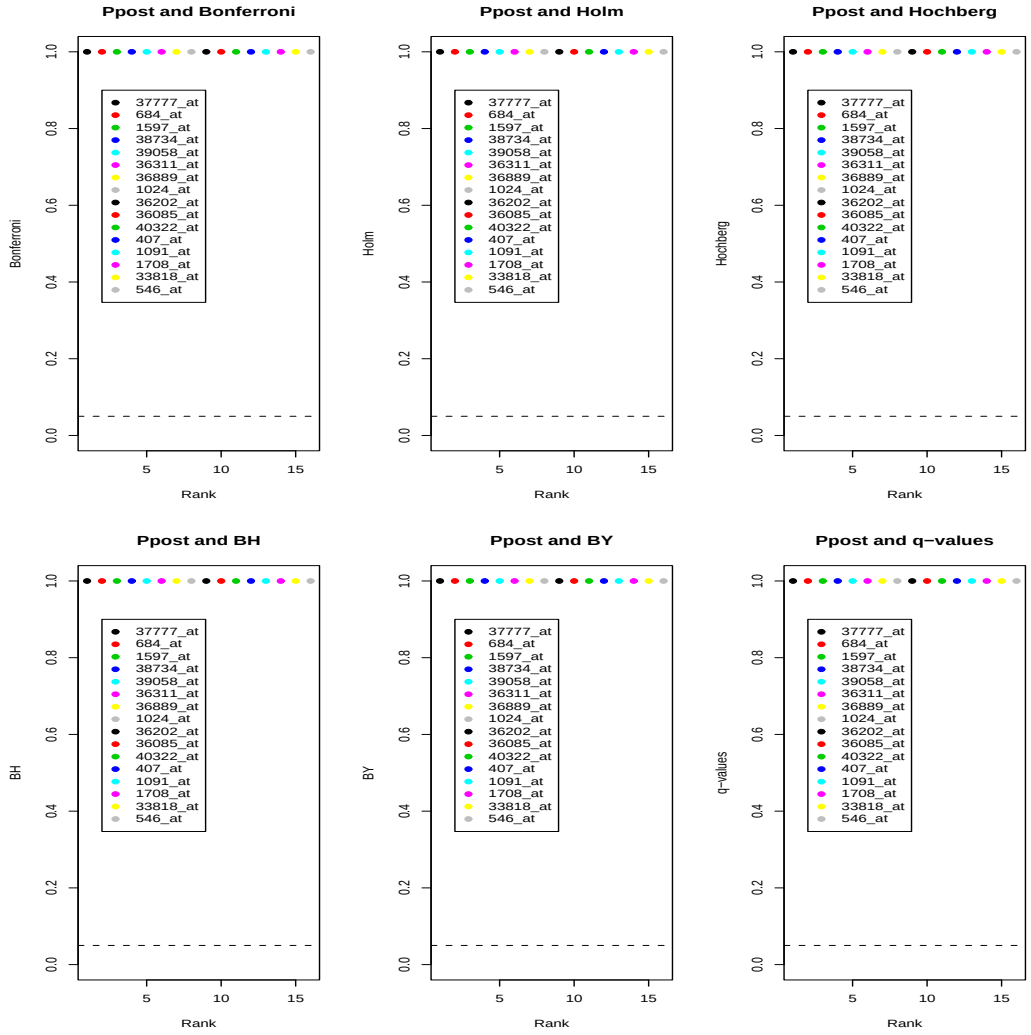

Figure 3.29: Analysis of Eset3 data set (Gamma model): results of the MHT using the p_{post}. The dotted line represent the reference error level 0.05. The legend reports the genes which are not significative differentially expressed.

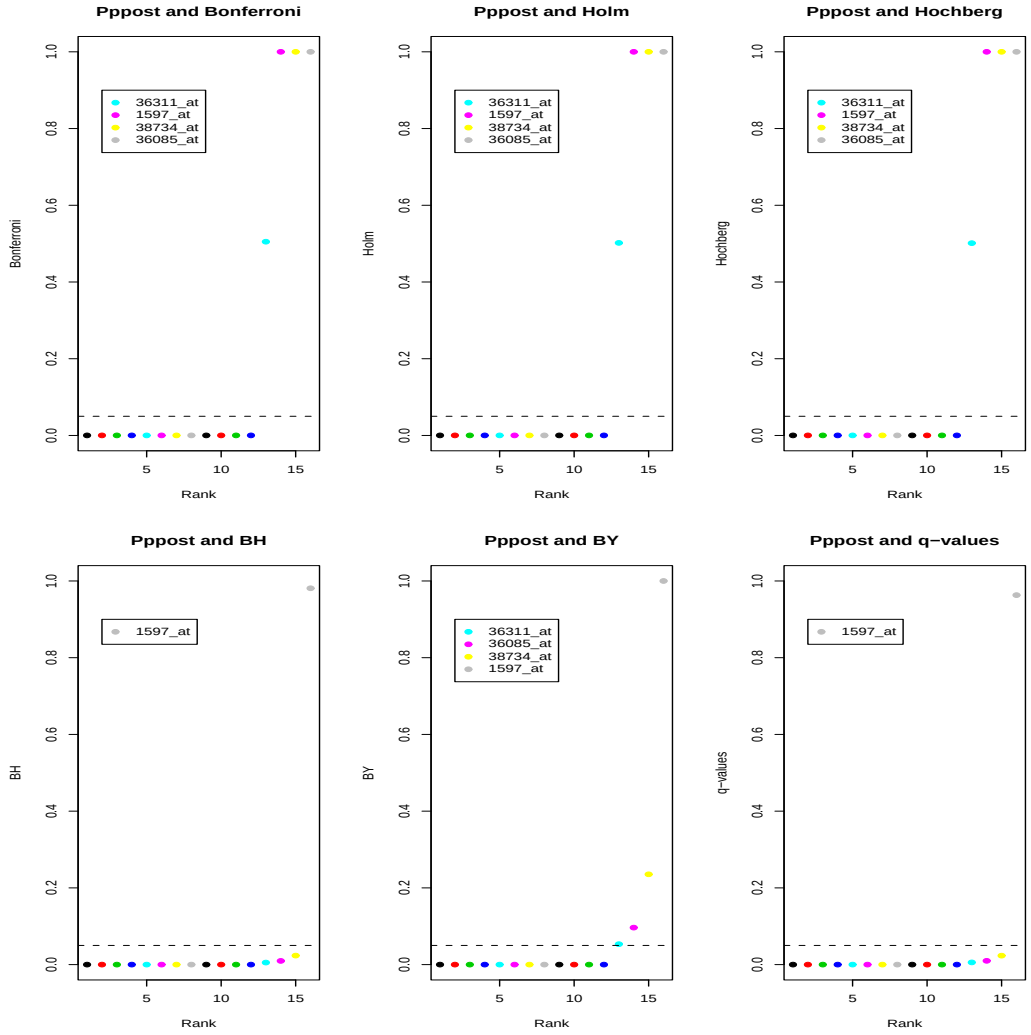

Figure 3.30: Analysis of Eset3 data set (Gamma model): results of the MHT using the p_{ppost}. The dotted line represent the reference error level 0.05. The legend reports the genes which are not significative differentially expressed.

per population. The `eset12` data set can be downloaded at the same location as `eset3`. As expected, and pointed out in Chapter 3, in this data set all three p-values provides the same results.

Chapter 4

Conclusions

Our main conclusions is that the MHT techniques investigated in this work need to use frequentist p-value. The case, where frequentist p-values are available, are limited to situations in which we have sufficient test statistics. Therefore in order to properly extend the MHT techniques to more general situations, (within a parametric framework), we showed that the p_{cpred} (when available) and p_{ppost} are useful.

We further illustrated that the p_{ppost} is approximately uniform distributed even with small sample sizes, such as 3-4 replications for each experimental condition. With larger sample size the p_{plug} and p_{post} are also approximately uniform distributed in $(0, 1)$ and the lack of uniformity is negligible for large sample size. However, we also showed that even if the null distribution of the p-values is not too far from the uniform the compound error in MHT is remarkable and convergency theorems for FDR procedures do not hold.

In this thesis we used a simple model, as the Gamma model, and we found that despite its simplicity it can be useful to draw inference if the evidence against $H = 0$ is obtained using appropriate p-values. We showed this with the analysis of the `eset3` data set. This reason is due to the fact the p_{ppost} avoids the double use of the data. This issue turns out to be relevant in MHT when the sample size is small.

4.1 Other approaches

We briefly mention other approaches either to MHT and to calibrated the p-values in the error scale. These approaches have not been considered here mainly because the are beyond the scope of the thesis or because they are computationally unfeasible.

4.1.1 The Empirical Bayes approach

Too simple models may be incompatible with the data and more complicated models should be used in order to capture the variance left from the normalization process. In this sense the Empirical Bayes models can be useful and they also have been applied to analyze the outcome of a microarray experiment (Efron *et al.*, 2001). However the relationship between the FDR and Empirical Bayes models is not clear, because the Empirical Bayes naturally account for the multiplicity of the tests. In fact if we

have m tests and each gene i came from the distribution $X_i \sim f(x_i, \theta_i)$ and we pose $\theta_i \sim g(\widehat{\gamma})$, we are naturally considering the multiple inference on the θ_i's because the MLE of γ, $\widehat{\gamma}$ has been calculated on the expression of all m genes. As pointed out by Kendiorski, in the discussion of Ge *et al.* (2003), the inference drawn from the posterior distribution of an Empirical Bayes model often does not require adjustments in order to control the FDR. Therefore the approach of Empirical Bayes models to MHT is different from the one considered here.

4.1.2 Calibrating p-values with the Bootstrap

Another different approach to derive frequentist p-values is the one illustrated by Davison and Hinkley (1997, pp. 175-176). They propose the following approach summarized in Algorithm 15.

Algorithm 15 *(Davison and Hinkley, 1997). Let P be the p-value regarded as a random variable and p the observed value. Let $\widehat{F}_0$ be the empirical distribution of the null model F_0 (corresponding to $H = 0$). We first simulate $P_1^*, P_2^*, \ldots, P_B^*$ by drawing B samples from $\widehat{F}_0$. The null distribution of the test statistic have to be approximated by in each of the B sample. We then approximate the adjusted p-value*

$$p_{adj} = \Pr{}^* \left(P^* \leq p | \widehat{F}_0 \right) \tag{4.1}$$

with a Monte Carlo sum.

Note that one must be careful to interpret the (4.1) because the outer probability relates to sampling from the empirical distribution $\widehat{F}_0$ rather than from F_0. This method has two non negligible issues in microarray data analysis:

a) the approximation of $\widehat{F}_0$ to F_0 can be problematic when the sample size is small;

b) the method can be potentially very computational intense, because we have to implement a bootstrap procedure for each replication $P_1^*, P_2^*, \ldots, P_B^*$ and we have to repeat it for every m genes under test.

For these reasons we did not implemented the method even if it provides useful results in some parametric models as shown in Davison and Hinkley (1997, pp. 176-177).

4.1.3 The Bayes Factors in MHT: a feasible approach

As pointed out in the first chapter we can construct Γ_a by using other test statistics more than the p-values. For example we could use Bayes Factors. Let $m_0(\mathbf{x}_i, \mathbf{y}_i)$ and $m_1(\mathbf{x}_i, \mathbf{y}_i)$ be the two predictive distribution respectively under $H_i = 0$ and $H_i = 1$. Here the alternative hypothesis can be not so vague as in the setup considered in this thesis. The alternative hypothesis must be embedded in a parametric model which is

supposed to hold when the gene is not differentially expressed. Then the Bayes Factor (BF) for a single gene under test is defined

$$BF_i = \frac{m_0\left(\mathbf{x}_i, \mathbf{y}_i\right)}{m_1\left(\mathbf{x}_i, \mathbf{y}_i\right)}.$$

The BFs are more in favor to Bayesian than the p-values because if the null hypothesis is true then $BF \to \infty$ as $n \to \infty$. However BFs present two relevant issues:

a) the distribution of BF is often unknown and can be very difficult to approximate when the marginal distributions of t are not known analytically;

b) often they are not defined when the prior are improper or the hypotheses are not nested.

Many techniques have been developed in order to solve issue (b) by choosing suitable noninformative priors on the unknown parameter involved in the hypotheses. See for example, the intrinsic procedure of Berger and Pericchi (1996), the fractional Bayes factor of O'Hagan (1995) and the intrinsic procedures in Moreno, Bertolino and Racugno (1998).

However, issue (a) still remains a problem especially in microarray data analysis where there are thousand of hypotheses under test and therefore thousand of BFs to approximate. This issue can be partially mitigated if we assume the same models and the same prior for all genes under testing. In this way we have only to approximate the unique null distribution for the BF and then calculate the Type I error concerning the decision of rejecting each $H_i = 0$ for $BF_i \leq \overline{BF}$. Once we have the rejection region $\Gamma_a = BF \leq \overline{BF}$ and we know α then we can estimate the FDR corresponding to the chosen threshold $\overline{BF}$.

Bibliography

[1] Abramovich, F. and Benjamini, Y. (1996). Adaptive thresholding of wavelet coefficients. *Computational Statistics and Data Analysis*, **22**, 351-361.

[2] Affymetrix (1999). *Affymetrix Microarray Suite User Guide*. Affymetrix, Santa Clara, CA.

[3] Bayarri, M. J. and Berger, J. O. (1997). Measures of surprise in Bayesian analysis. *ISDS Discussion Paper* 97-46, Duke University.

[4] Bayarri, M. J. and Berger, J. O. (1999). Quantifying Surprise in the Data and Model Verification. In *Bayesian Statistics* 6, J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (eds.), 53-82. Oxford University Press.

[5] Bayarri, M. J. and Berger, J. O. (2000). p-values for Composite Null Models. *Journal of the American Statistical Association*, **95**, 1127-1142.

[6] Bayarri, M. J. and Castellanos, M. E. (2001). A Comparison Between p-Values for Goodness-of-fit Checking. *Monographs of official Statistics, Bayesian Methods* (ed. : George I. Edward), 1-10.

[7] Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society B*, **57**, 289-300.

[8] Benjamini, Y. and Hochberg, Y. (1997). False Discovery Rate control in multiple testing using depentent test statistics. *Research Paper 97-1, Department of Statistics and O. R.* , Tel Aviv University.

[9] Benjamini, Y. and Liu, W. (1999). A ste$p$-down multiple hypothesis procedure that controls the false discovery rate under independence. *Journal of Statistical Planning and Inference*, **82**, 163-170.

[10] Benjamini, Y. and Hochberg, Y. (2000). On adaptive control of the false discovery rate in multiple testing with indepentent statistics. *Journal of Educational and Behavior Statistics*, **25**, 60-83.

[11] Benjamini Y. and Yekutieli, D. (2001). The control of the False Discovery Rate in multiple testing under dependency. *Annals of Statistics,* **29**, 1165-1188.

[12] Berger, J. O. and Sellke, T. (1987). Testing of a point null hypothesis: The irreconciability of p-value and evidence. *Journal of the American Statistical Association*, **82**, 112-139.

[13] Berger, J. O. and Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, **2**, 317-335.

[14] Berger, J. O. and Bernardo, J. M. (1992a). Ordered group reference priors with applications to multinomial problem. *Biometrika*, **79**, 25-37.

[15] Berger, J. O. and Bernardo, J. M. (1992b). Reference priors in a variance components problem. In *Bayesian Analysis in Statistics and Econometrics*, Ed. P. K. Goel and N. S. Iyengar, 177-94. New York: Springer-Verlag.

[16] Berger, J. O. and Bernardo, J. M. (1992c). On the development of the reference prior method (with discussion). In *Bayesian Statistics 4*, Ed. J. M. Bernardo, J. O. Berger, A. P. David and A. F. M. Smith, 35-60. Oxford University Press.

[17] Berger, J. O. and Pericchi, L. R. (1996). The intrinsic Bayes factor for linear models. In *Bayesian Statistics 5*, Ed. J. M. Bernardo, J. O. Berger, A. P. David and A. F. M. Smith, 25-44. Oxford University Press.

[18] Berger, J. O. (2002). Could Fisher, Jeffreys, and Neyman Have Agreed on Testing?. ISDS (Duke University) discussion paper 02-01.

[19] Bernardo, J. M. (1976). Algorithm AS 103: Psi (digamma) function. *Journal of Applied Statistics*, **25**, 315-317.

[20] Berti, P., Fattorini, L. and Rigo, P. (2000). Eliminating nuisance parameters: two characterizations. *Test*, **9**, 133-148.

[21] Bolsover, S. R., Hyams, J., Jones, S., Shepard, E. A. and White, H. A. (1997). *From Genes to Cells*. New York: Wiley

[22] Bolstad, B., Irizarry, R., Astrand, M., Speed, T. (2002). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Technical Report, UC Berkley.

[23] Box, G. E. P. (1980). Sampling and Bayes Inference in Scientific Modeling and Robustness. *Journal of the Royal Statistical Society A*, **143**, 383-430.

[24] Brown, P. and Botstein, D. (1999). Exploring the New World of the Genome with DNA Microarray. *Nature genetics supplement*, **21**, 33-37.

[25] Brown, C., Goodwin, P. and Sorger, P. (2001). Image metrics in the statistical analysis of DNA microarray data. *Proceedings of the National Academy of Science*, USA, **98**, 8944-8949.

[26] Chen, Y., Dougherty, E. R. and Bittner, M. L. (1997). Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Journal of Biomedical Optics*, **2**, 364-374.

[27] Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, IT-**13**, 21-27.

[28] Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap methods and their application*, New York: Cambridge University Press.

[29] DeGroot, M. H. and Fienberg, S. (1983). The comparison and evaluation of forecasters. *The Statistician*, **32**, 12-22.

[30] Dudoit, S., Yang, Y. H., Callow, M. J. and Speed, T. P. (2001). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, **12**, 111-139.

[31] Dudoit, S., Fridlyand, J. and Speed, T. P. (2002a). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, **97**, 77-87.

[32] Dudoit, S., Shaffer, J. P. and Boldrick, J. C. (2002b). Multiple hypothesis testing in microarray experiments. UC Berkeley, Division Biostatistics working paper series: 2002-110.

[33] Duggan, D., Bittner, M., Chen, Y., Meltzer, P. and Trent, J. (1999). Expression profiling using cDNA miroarrays. *Nature Genetics*, **21**, 10-14.

[34] Efron, B. and Morris, C. (1973). Combining possibly related estimation problems (with discussion). *Journal of the Royal Statistical Society, Series B,* **35**, 379-421.

[35] Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman and Hall.

[36] Efron, B. and Tibshirani, R. (1998). The problem of regression. *Annals of Statistics,* **26**, 1687-1718.

[37] Efron, B., Tibshirani, R., Storey, J. D. and Tusher V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, **96**, 1151-1160.

[38] Efron, B. and Tibshirani, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology*, **23**, 70-86.

[39] Garret, R. H. and Grisham, C. M. (2002). *Priciples of Byochemistry*. Pacific Grove, CA: Brooks/Cole.

[40] Ge, Y., Dudoit, S. and Speed, T. P. (2003). Resampling-based Multiple Testing for Microarray Data Analysis (with discussion). *Test*, **12**, 1-77.

[41] Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society B*, **64**, 499-518.

[42] Gibbons, J. and Pratt, J. (1975). P -values: interpretation and methodology. *American Statistician*, **29**, 20-25.

[43] Godfrey, K. (1985). Comparing the means of several groups. *New England Journal of Medicine*, **311**, 1450-1456.

[44] Good, I. J. (1953). The appropriate mathematical tools for describing and measuring uncertainty. In *Uncertainty and Business Decision*, eds. C. F. Carter, G. P. Meredith and G. L. S. Shackle, 19-34. Liverpool: University Press.

[45] Good, I. J. (1956). The surprise index for the multivariate normal distribution. *Annals of Mathematical Statistics*, **27**, 1130-1135.

[46] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. and Lander, E. S. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, **286(5439)**, 531-537.

[47] Guttman, I. (1967). The Use of the Concept of a Future Observation in Goodness-of-fit Problems. *Journal of the Royal Statistical Society B*, **29**, 83-100.

[48] Hardiman, G. (2002). Microarray technologies-an overview. *Pharmacogenomics* 3, 293-297.

[49] Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, **75**, 800-803.

[50] Hochberg, Y. and Tamhane, A. (1987). *Multiple Comparison Procedures*, Wiley and Sons, New York.

[51] Hochberg, Y. and Rom, Y. (1996). Extentions of multiple testing procedures based on Simes' test. *Journal of Statical Planning and Inference,* **48**, 141-152.

[52] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavia Journal of Statistics*, **6**, 65-70.

[53] Hommel, G. (1988). A stage-wise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, **75**, 383-386.

[54] Hubbard, R. and Bayarri, M. J. (2003). Confusion Over Measures of Evidence (p's) Versus Errors (α's) in Classical Statistical Testing. *The American Statistician*, **57**, 171-178.

[55] Ibrahim, J., Chen M. and Gray R. (2002). Bayesian models for gene expression with DNA microarray data. *Journal of the American Statistical Association,* **97**, 88-99.

[56] Jain, A., Tokuyasu, T., Snijders, A., Segraves, R., Albertson, D. and Pinkel, D. (2002). Fully automatic quantification of microarray image data. *Genome Research*, **12**, 325-332.

[57] James, W., Stein, C. (1961). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 361-380.

[58] Kendall, M. G. and Stuart, A. (1969). *The Advanced Theory of Statistics*, **1**, 3rd ed. Hafner, New York.

[59] Korn, E. L., Troendle, J. F., McShane, L. M. and Simon, R. (2001). Controlling the number of false discoveries: application to high dimentional genomic data. Technical Report 003, National Cancer Institute, Division of Cancer Treatment and Diagnosis.

[60] Lehmann, E. L. (1986). *Testing Statistical Hypotheses*, second edition. Springer-Berlag, New York.

[61] Lindley, D. and Smith, A. F. M. (1972). Bayes estimates for the linear model (with discussion). *Journal of the Royal Statistical Society, Series B,* **34**, 1-41.

[62] Lockhart, D., Dong, H., Byrne, M., Follettie, M., Gallo, M., Chee, M., Mittmann, M., Wang, C., Kobayashi, M., Horton, H. and Brown, E. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, **14**, 1675-1680.

[63] Liseo, B. (1993). Elimination of nuisance parameter with reference priors. *Biometrika*, **80**, 295-304.

[64] Lönnstedt, I. and Speed, T. P. (2002). Replicated microarray data. *Statistica Sinica* **12**, 31-46.

[65] Moreno, E., Bertolino, F. and Racugno, W. (1998). An intrinsic limiting procedure for model selection and hypothesis testing. *Journal of American Statistical Association*, **93**, 1451-1460.

[66] Morton, N. E. (1955). Sequential tests for detection of linkage. *American Journal of Human Genetics*, **7**, 277-318.

[67] Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattener, F. R. and Tsui, K. W. (2002). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology*, **8**, 37-52

[68] Neyman, J. (1977). Frequentist probability and frequentist statistics. *Synthèse*, **36**, 97-131.

[69] O'Hagan, A. (1995). Fractional Bayes factor for model comparison (with discussion). *Journal of the Royal Statistical Society B*, **57**, 99-138.

[70] O'Hagan, A. (2003). HSSS model criticism (with discussion). In *Highly Structured Stochastic Systems*, P. J. Green, N. L. Hjort and S. T. Richardson (eds), 423-453. Oxford University Press.

[71] Pan, W., Lin, J. and Le, C. (2002). How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach. *Genome Biology*, **3**, 22. 1-22. 10.

[72] Pesarin, F. (2001). *Multivariate permutation tests with applications in biostatistics*. John Wiley and Sons, Chichester.

[73] Pocock, S. J., Hughes, M. D. and Lee, R. J. (1987). Statistical problems in reporting of clinical trials. *Journal of the American Statistical Association*, **84**, 381-392.

[74] Robert, C., Müller, P., Parmigiani, G., Rousseau, J. (2003). Optimal Sample Size for Multiple Testing: The Case of Gene Expression Microarrays. Johns Hopkins University, Dept. of Biostatistics, technical report 3.

[75] Robins, J. M., van der Vaart, A. and Ventura, V. (2000). The asymptotic distribution of p-values in composite null models. *Journal of the American Statistical Association*, **95**, 1143-1156

[76] Rubin, D. B. (1984). Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician. *Annals of Statistics*, **12**, 1151-1172.

[77] Saville, D. J. (1990). Multiple Comparison Procedure: The Practical Solution. *American Statistician*, **44**, 174-180.

[78] Schneider, B. E. (1978). Algorithm AS 121: Trigamma function. *Journal of Applied Statistics*, **27**, 97-99.

[79] Schena, M., (2000). *Microarray Biochip Technology*. Sunnyvale, CA: Eaton.

[80] Seeger, P. (1968). A note on a method for the analysis of significance en masse. *Technometrics*, **10**, 586-593.

[81] Selke, T, Bayarri, M. J. and Berger, J. O. (2001). Calibration of p-values for testing precise null hypothesis. *The american stastisticians*, **55**, 62-71

[82] Sellers, K. F., Miecznikowski, J. and Eddy, W. F. (2003). Removal of Systematic Variation in Genetic Microarray Data. Technical report, Department of Statistics, Carnegie Mellon University.

[83] Shaffer, J. P. (1995). Multiple hypotheses-testing. *Annual Review of Psichology*, **46**, 561-584.

[84] Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance, *Biometrika*, **73**, 751-754.

[85] Simon, R., Radmacher, M. and Dobbin, K. (2002). Design of study using DNA microarrays. *Genetic Epidemiology* **23**, 21-26.

[86] Soriç, B. (1989). Statistical "discoveries" and effect size estimation. *Journal of American Statistics Association*, **84**, 608-610.

[87] Southern, E. M. (2001). DNA microarrays. History and overview. *Methods in Molecular Biology*, **170**, 1-15.

[88] Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society B*, **64**, 479-498.

[89] Storey, J. D. (2003). The Positive False Discovery Rate: A Bayesian Interpretation and the q-value. *Annals of Statistics*, **31**, 2013-2035.

[90] Storey, J. D., Taylor J. E. and Siegmund D. (2004). Strong control, conservative point estimation and simultaneous conversative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society B*, **66**, 187-205.

[91] Tseng, G., Oh M., Rohlin, L., Liao, J. and Wong, W. (2001). Issues on cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Research*, **29**, 2549-2557.

[92] Weller, J. I., Song, J. Z., Heyen, D. W., Lewin, H. A. and Ron, M. (1998). A new approach to the Problem of Multiple Comparisons in the Genetic Dissection of Complex Traits. *Genetics*, **150**, 1699-1706.

[93] Westfall, P. H. and Young, S. S. (1993). *Resampling-based Multiple Testing: Examples and Methods for p-value Adjustment*, Wiley.

[94] Yang, Y. H., Buckley, M. J., Dudoit, S., and Speed, T. P. (2002). Comparison of methods for image analysis on cDNA microarray data. *Journal of Computational and Graphical Statistics*, in press.

[95] Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J. and Speed, T. P. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleid Acids Research*, **30**, e15.

[96] Yekutieli, D. and Benjamini, Y. (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference*, **82**, 171-196.

[97] Zweiger, G. (2001). *Transducing the genome: information, anarchy and revolution in the biomedical sciences*. McGraw-Hill.

Appendix A

P-values for the Normal model

In the Appendix we provide the details for the calculations of the p-values for the Normal model.

A.1 Plug-in p-value

The Likelihood under $H = 0$ is:

$$\begin{aligned} L\left(\mu, \sigma^2\right) &= \prod_{i=1}^{n_X} \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \prod_{i=1}^{n_Y} \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\sigma^2}} \exp\left(-\frac{(y_i-\mu)^2}{2\sigma^2}\right) \\ &\propto \left(\sigma^2\right)^{-\left(\frac{n_X+n_Y}{2}\right)} \exp\left(-\frac{1}{2\sigma^2} B\left(\mu\right)\right) \end{aligned}$$

where $B\left(\mu\right) = \sum_{i=1}^{n_X} x_i^2 + \sum_{i=1}^{n_Y} y_i^2 - 2\mu\left(\sum_{i=1}^{n_X} x_i + \sum_{i=1}^{n_Y} y_i\right) + \left(n_X + n_Y\right)\mu^2$. The MLE for μ and σ^2 are

$$\begin{aligned} \widehat{\mu} &= \frac{\sum_{i=1}^{n_X} x_i + \sum_{i=1}^{n_Y} y_i}{n_X + n_Y} = \frac{n_X}{n_X + n_Y}\overline{x} + \frac{n_Y}{n_X + n_Y}\overline{y} \\ \widehat{\sigma}^2 &= \frac{n_X^2 S_x^2 + n_Y^2 S_y^2 + n_X n_Y\left(\overline{x^2} + \overline{y^2} - 2\overline{xy}\right)}{\left(n_X + n_Y\right)^2} \\ &= \frac{\sum_{i=1}^{n_X} x_i^2 + \sum_{i=1}^{n_Y} y_i^2}{n_X + n_Y} - \widehat{\mu}^2 \end{aligned}$$

where $S_x^2 = \overline{x^2} - \overline{x}^2$, $S_y^2 = \overline{y^2} - \overline{y}^2$ and $\overline{x^2} = \sum_{i=1}^{n} x_i^2/n_X$, $\overline{y^2} = \sum_{i=1}^{n} y_i^2/n_Y$ are the second sampling moments. The Plug-in p-value is:

$$p_{plug} = 2\left(1 - \Phi\left(\frac{\left|t\left(\mathbf{x}, \mathbf{y}\right)\right|}{\sqrt{\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)\widehat{\sigma}^2}}\right)\right)$$

where $\Phi\left(\cdot\right)$ is the cumulative distribution function of a standard normal distribution.

A.2 Posterior p-value

Using the Reference prior for μ and σ^2:

$$\pi\left(\mu, \sigma^2\right) \propto 1/\sigma^2, \sigma \in \mathcal{R}^+$$

the marginal posterior distribution for σ^2 is

$$\begin{aligned}
\pi\left(\sigma^2|\mathbf{x},\mathbf{y}\right) &\propto \\
&\propto \int_{\mathcal{R}} \left(\sigma^2\right)^{-\left(\frac{n_X+n_Y+1}{2}\right)} \times \\
&\quad \times \exp\left\{-\frac{n_X+n_Y}{2\sigma^2}\left[\mu^2 - 2\mu\frac{\sum_{i=1}^{n_X} x_i + \sum_{i=1}^{n_Y} y_i}{n_X+n_Y} + \frac{\sum_{i=1}^{n_X} x_i^2 + \sum_{i=1}^{n_Y} y_i^2}{n_X+n_Y}\right]\right\} d\mu \\
&\propto \left(\sigma^2\right)^{-\left(\frac{n_X+n_Y+1}{2}\right)} \exp\left\{-\frac{1}{2\sigma^2}\left[\frac{\sum_{i=1}^{n_X} x_i^2 + \sum_{i=1}^{n_Y} y_i^2}{n_X+n_Y} - \widehat{\mu}^2\right]\right\} \times \\
&\quad \times \int_{\mathcal{R}} \underbrace{\exp\left\{-\frac{n_X+n_Y}{2\sigma^2}\left(\mu-\widehat{\mu}\right)^2\right\}}_{\text{Kernel of } N\left(\widehat{\mu},\frac{\sigma^2}{n_X+n_Y}\right)} d\mu \\
&\propto \left(\sigma^2\right)^{-\left(\frac{n_X+n_Y+1}{2}\right)} \exp\left\{-\frac{1}{\sigma^2}\widehat{\sigma}^2\right\}\sqrt{2\pi\frac{\sigma^2}{n_X+n_Y}} \\
&\propto \left(\sigma^2\right)^{-\left(\frac{n_X+n_Y}{2}-1+1\right)} \exp\left\{-\frac{\widehat{\sigma}^2}{\sigma^2}\right\} \\
&= Ga^{-1}\left(\frac{n_X+n_Y}{2}-1, \widehat{\sigma}^2\right)
\end{aligned}$$

where $Ga^{-1}\left(\frac{n_X+n_Y}{2}-1,\widehat{\sigma}^2\right)$ is an inverse gamma distribution with scale parameter $\widehat{\sigma}^2$. The marginal posterior distribution for $t\left(\mathbf{x},\mathbf{y}\right)$ is

$$\begin{aligned}
m_{post}\left(t|\mathbf{x},\mathbf{y}\right) &= \int_0^\infty f\left(t|\sigma^2\right)\pi\left(\sigma^2|\mathbf{x},\mathbf{y}\right)d\sigma^2 \\
&= \frac{1}{\sqrt{2\pi\frac{n_X+n_Y}{n_X n_Y}}}\frac{\left(\widehat{\sigma}^2\right)^{\frac{n_X+n_Y}{2}-1}}{\Gamma\left(\frac{n_X+n_Y}{2}-1\right)} \times \\
&\quad \times \int_0^\infty \underbrace{\left(\sigma^2\right)^{-\left(\frac{n_X+n_Y+1}{2}\right)}\exp\left\{-\frac{1}{\sigma^2}\left(\frac{n_X n_Y t^2}{2\left(n_X+n_Y\right)}+\widehat{\sigma}^2\right)\right\}}_{\text{kernel of } Ga^{-1}\left(\frac{n_X+n_Y+1}{2}-1,\frac{n_X n_Y t^2}{2\left(n_X+n_Y\right)}+\widehat{\sigma}^2\right)} d\sigma^2 \\
&= \frac{\left(\widehat{\sigma}^2\right)^{\frac{n_X+n_Y}{2}-1}}{\sqrt{2\pi\frac{n_X+n_Y}{n_X n_Y}}}\frac{\Gamma\left(\frac{n_X+n_Y+1}{2}-1\right)}{\Gamma\left(\frac{n_X+n_Y}{2}-1\right)}\left(\frac{n_X n_Y t^2}{2\left(n_X+n_Y\right)}+\widehat{\sigma}^2\right)^{-\left(\frac{n_X+n_Y+1}{2}-1\right)} \\
&= \frac{1}{\sqrt{2\widehat{\sigma}^2\frac{n_X+n_Y}{n_X n_Y}}}\frac{1}{\sqrt{\pi}}\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\nu/2\right)}\left(\frac{t^2}{2\widehat{\sigma}^2\frac{n_X+n_Y}{n_X n_Y}}+1\right)^{-\frac{\nu+1}{2}}, \nu = n_X+n_Y-2 \\
&= \frac{1}{\sqrt{2\widehat{\sigma}^2\frac{n_X+n_Y}{\nu n_X n_Y}}}\frac{1}{\sqrt{\nu\pi}}\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\nu/2\right)}\left[1+\frac{1}{\nu}\left(\frac{t}{\sqrt{2\widehat{\sigma}^2\frac{n_X+n_Y}{\nu n_X n_Y}}}\right)^2\right]^{-\frac{\nu+1}{2}} \\
&= \zeta_{n_X+n_Y-2}\left(0,\widehat{\sigma}^2\frac{n_X+n_Y}{n_X n_Y}\frac{2}{n_X+n_Y-2}\right)
\end{aligned}$$

where $\zeta_{n_X+n_Y}\left(0,\widehat{\sigma}^2\frac{n_X+n_Y}{n_X n_Y}\frac{2}{n_X+n_Y-2}\right)$ represents the density of a centered t-student

distribution with $n_X + n_Y$ degrees of freedom and $\widehat{\sigma}^2 \frac{n_X+n_Y}{n_X n_Y} \frac{2}{n_X+n_Y-2}$ is the scale parameter. Therefore the Posterior p-value is given by:

$$p_{post} = 2\left(1 - \Upsilon_{n_X+n_Y}\left(\frac{|t(\mathbf{x},\mathbf{y})|}{\sqrt{\widehat{\sigma}^2 \frac{n_X+n_Y}{n_X n_Y} \frac{2}{n_X+n_Y-2}}}\right)\right)$$

where $\Upsilon_{n_X+n_Y}(\cdot)$ is the c.d.f. of a standard t-student distribution with $n_X + n_Y$ degrees of freedom.

A.3 Conditional Predictive and Partial Posterior Predictive p-value

Using the same notation as in Bayarri-Berger

$$\begin{aligned}
u(\mathbf{x},\mathbf{y}) &= \text{the specific proposal for } U \\
&= \arg\max_{\mu,\sigma^2} \frac{L(\mu,\sigma^2)}{f(t|\sigma^2)} \\
&= \arg\max_{\sigma^2} \left(\sigma^2\right)^{-\left(\frac{n_X+n_Y}{2}\right)} \exp\left(-\frac{1}{2\sigma^2} B(\widehat{\mu})\right) \sqrt{\frac{n_X+n_Y}{n_X n_Y}\sigma^2} \exp\left(\frac{t(\mathbf{x},\mathbf{y})^2}{2\frac{n_X+n_Y}{n_X n_Y}\sigma^2}\right) \\
&= \arg\max_{\sigma^2} \left(\sigma^2\right)^{-\left(\frac{n_X+n_Y-1}{2}\right)} \exp\left(-\frac{1}{2\sigma^2}\left(B(\widehat{\mu}) - \frac{(\overline{x}-\overline{y})^2}{\frac{n_X+n_Y}{n_X n_Y}}\right)\right) \\
&= \arg\max_{\sigma^2} -\left(\frac{n_X+n_Y-1}{2}\right)\log\sigma^2 - \frac{1}{2\sigma^2}\left(B(\widehat{\mu}) - \frac{(\overline{x}-\overline{y})^2}{\frac{n_X+n_Y}{n_X n_Y}}\right) = l\left(\sigma^2\right) \\
&\Rightarrow\ l'\left(\sigma^2\right) = 0 \\
&\Rightarrow\ \widehat{\sigma}^2_{cMLE} = \frac{1}{(n_X+n_Y-1)}\left(B(\widehat{\mu}) - \frac{(\overline{x}-\overline{y})^2}{\frac{n_X+n_Y}{n_X n_Y}}\right) \\
&= \frac{n_X^2\overline{x^2} + n_Y^2\overline{y^2} + n_X n_Y\overline{x^2} + n_X n_Y\overline{y^2} - n_X^2\overline{x}^2 - n_Y^2\overline{y}^2 - n_X n_Y\overline{x}^2 - n_X n_Y\overline{y}^2}{(n_X+n_Y)(n_X+n_Y-1)} \\
&= \frac{n_X^2 S_x^2 + n_Y^2 S_y^2 + n_X n_Y\left(S_x^2+S_y^2\right)}{(n_X+n_Y)(n_X+n_Y-1)} \\
&= \frac{n_X}{n_X+n_Y-1}S_x^2 + \frac{n_Y}{n_X+n_Y-1}S_y^2 \\
&\Rightarrow\ u(\mathbf{x},\mathbf{y}) = \left(\widehat{\mu}_{cMLE} = \widehat{\mu}, \widehat{\sigma}^2_{cMLE} = \frac{n_X S_x^2 + n_Y S_y^2}{n_X+n_Y-1}\right)
\end{aligned}$$

Note that $\widehat{\sigma}^2_{cMLE}$ is stochastically proportional to two independent quantities: S_x^2 and S_y^2 (because of the independence between X and Y). S_x^2 and S_y^2 are independent from T (also from $\widehat{\mu}_{cMLE}$) and they are jointly sufficient for σ^2 and μ, therefore, the Partial Posterior Predictive p-value is equal to the Conditional predictive p-value.

Let $f\left(u(\mathbf{x},\mathbf{y}),\mu,\sigma^2\right)$ represent the joint density of U, then the U-conditional distribution for the parameter is given by:

$$\pi\left(\mu,\sigma^2|u(\mathbf{x},\mathbf{y})\right) \propto f\left(u(\mathbf{x},\mathbf{y}),\mu,\sigma^2\right)\pi\left(\mu,\sigma^2\right)$$

$$\begin{aligned}
&\propto \underbrace{f\left(\widehat{\mu},\mu,\sigma^2|H_0\right)}_{N\left(\mu,\left(\frac{1}{n_X}+\frac{1}{n_Y}\right)\sigma^2\right)} \underbrace{f\left(\widehat{\sigma}^2_{cMLE},\mu,\sigma^2|H_0\right)}_{\chi^2_{n_X+n_Y-1}} \frac{1}{\sigma} \text{ as } \widehat{\sigma}^2_{cMLE} \perp \widehat{\mu}_{cMLE} \\
&= \pi\left(\mu|\sigma^2, \frac{n_X\overline{x}+n_Y\overline{y}}{n_X+n_Y}\right)\pi\left(\sigma^2|\widehat{\sigma}^2_{cMLE}\right) \text{ marginal conditional distribution}
\end{aligned}$$

where:

$$\begin{aligned}
\pi\left(\mu|\sigma^2, \frac{n_X\overline{x}+n_Y\overline{y}}{n_X+n_Y}\right) &= N\left(\frac{n_X\overline{x}+n_Y\overline{y}}{n_X+n_Y}, \frac{\sigma^2}{n}\right) \\
\pi\left(\sigma^2|\widehat{\sigma}^2_{cMLE}\right) &= \pi\left(\sigma^2|\widehat{\sigma}^2_{cMLE}\right) \text{ does not depend on } \mu \\
&= Ga^{-1}\left(\frac{n_X+n_Y-2}{2}, \frac{n_X+n_Y-1}{2}\widehat{\sigma}^2_{cMLE}\right)
\end{aligned}$$

The marginal distribution of $t|u\left(\mathbf{x},\mathbf{y}\right)$ is given by:

$$\begin{aligned}
m\left(t|u\left(\mathbf{x},\mathbf{y}\right)\right) &= \int_{\mathcal{R}}\int_{\mathcal{R}^+} f\left(t|\sigma^2,u\left(\mathbf{x},\mathbf{y}\right)\right)\pi\left(\mu,\sigma^2|u\left(\mathbf{x},\mathbf{y}\right)\right)d\mu d\sigma^2 \\
&= \int_{\mathcal{R}^+} f\left(t|\sigma^2,\widehat{\sigma}^2_{cMLE}\right)\pi\left(\sigma^2|\widehat{\sigma}^2_{cMLE}\right)d\sigma^2 \int_{\mathcal{R}}\pi\left(\mu|\sigma^2,\frac{n_X\overline{x}+n_Y\overline{y}}{n_X+n_Y}\right)d\mu \\
&= \int_{\mathcal{R}^+} f\left(t|\sigma^2\right)\pi\left(\sigma^2|\widehat{\sigma}^2_{cMLE}\right)d\sigma^2 \text{ because } \widehat{\sigma}^2_{cMLE} \perp T \\
&= \int_{\mathcal{R}^+} \frac{1}{\sqrt{2\pi\frac{n_X+n_Y-1}{n_Xn_Y}\sigma^2}}\exp\left(-\frac{t^2}{2\frac{n_X+n_Y-1}{n_Xn_Y}\sigma^2}\right)\times \\
&\quad \times\frac{\left(\frac{n_X+n_Y-1}{2}\widehat{\sigma}^2_{cMLE}\right)^{\frac{n_X+n_Y-2}{2}}\left(\sigma^2\right)^{-\left(\frac{n_X+n_Y-2}{2}+1\right)}}{\Gamma\left(\frac{n_X+n_Y-2}{2}\right)}\times \\
&\quad \times\exp\left\{-\frac{\left(n_X+n_Y-1\right)\widehat{\sigma}^2_{cMLE}}{2\sigma^2}\right\}d\sigma^2 \\
&= \frac{1}{\sqrt{2\pi\frac{n_X+n_Y}{n_Xn_Y}}}\frac{\left(\frac{n_X+n_Y-1}{2}\widehat{\sigma}^2_{cMLE}\right)^{\frac{n_X+n_Y-2}{2}}}{\Gamma\left(\frac{n_X+n_Y-2}{2}\right)}\times \\
&\quad \times\int_{\mathcal{R}^+}\underbrace{\left(\sigma^2\right)^{-\left(\frac{n_X+n_Y-1}{2}+1\right)}\exp\left\{-\frac{1}{\sigma^2}\left(\frac{n_X+n_Y-1}{2}\widehat{\sigma}^2_{cMLE}+\frac{t^2}{2\frac{n_X+n_Y}{n_Xn_Y}}\right)\right\}}_{\text{Kernel of } Ga^{-1}\left(\frac{n_X+n_Y-1}{2},\frac{n_X+n_Y-1}{2}\widehat{\sigma}^2_{cMLE}+\frac{t^2}{2\frac{n_X+n_Y}{n_Xn_Y}}\right)}d\sigma^2 \\
&= \frac{1}{\sqrt{2\pi\frac{n_X+n_Y}{n_Xn_Y}}}\left(\frac{n_X+n_Y-1}{2}\widehat{\sigma}^2_{cMLE}\right)^{\frac{n_X+n_Y-2}{2}}\times \\
&\quad \times\frac{\Gamma\left(\frac{n_X+n_Y-1}{2}\right)}{\Gamma\left(\frac{n_X+n_Y-2}{2}\right)}\left(\frac{n_X+n_Y-1}{2}\widehat{\sigma}^2_{cMLE}+\frac{t^2}{2\frac{n_X+n_Y}{n_Xn_Y}}\right)^{-\frac{n_X+n_Y-1}{2}} \\
&= \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\frac{1}{\sqrt{\pi\frac{(n_X+n_Y)(n_X+n_Y-1)}{n_Xn_Y}\widehat{\sigma}^2_{cMLE}}}\times, \text{ where } \nu=n_X+n_Y-2 \\
&\quad \times\left(1+\frac{t^2}{\frac{(n_X+n_Y)(n_X+n_Y-1)}{n_Xn_Y}\widehat{\sigma}^2_{cMLE}}\right)^{-\frac{\nu+1}{2}}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\sqrt{\nu\pi}} \frac{1}{\sqrt{\frac{(n_X+n_Y)(n_X+n_Y-1)}{\nu n_X n_Y}\widehat{\sigma}^2_{cMLE}}} \times \\
&\quad \times \left[1+\frac{1}{\nu}\left(\frac{t}{\sqrt{\frac{(n_X+n_Y)(n_X+n_Y-1)}{\nu n_X n_Y}\widehat{\sigma}^2_{cMLE}}}\right)^2\right]^{-\frac{\nu+1}{2}} \\
&= \zeta_{n_X+n_Y-2}\left(0, \frac{(n_X+n_Y)(n_X+n_Y-1)}{(n_X+n_Y-2)n_X n_Y}\widehat{\sigma}^2_{cMLE}\right)
\end{aligned}
$$

Then the Conditional Predictive p-value is equal to:

$$
p_{cpred} = 2\left(1-\Upsilon_{n_X+n_Y-2}\left(\frac{|t(\mathbf{x},\mathbf{y})|}{\sqrt{S^2_{pooled}}}\right)\right)
$$

where $S^2_{pooled} = \frac{(n_X+n_Y)(n_X+n_Y-1)}{(n_X+n_Y-2)n_X n_Y}\widehat{\sigma}^2_{cMLE} = \left(\frac{1}{n_X}+\frac{1}{n_Y}\right)\left(\frac{n_X S^2_x + n_Y S^2_y}{n_X+n_Y-2}\right)$ is the pooled sample variance in the classical test for the difference of two means under normal assumption with equal variance.

Appendix B

P-values for the Gamma model

In this Appendix we provide the details of the calculations for the p-values in the Gamma model. We also show that the partial posterior distribution for the Gamma model is a proper probability distribution.

The density (2.9) has been obtained by constraining the random variable $D = \overline{X}/\overline{Y}$ to be greater than 1. For $\rho = \theta_Y/\theta_X$, the distribution of random variable D is a Multiple Scale Beta distribution of II kind (B.1) (see Kendall and Stuart 1969, p. 151):

$$f(d; a, \rho) = \frac{\Gamma(2na)}{\Gamma^2(na)} \left(\frac{1}{\rho}\right) \frac{(d/\rho)^{na-1}}{(1 + d/\rho)^{2na}} \tag{B.1}$$

Algorithm 16 *We estimate the c.d.f. of T for a given a using the following fact:*

$$b = \frac{d}{d+1} \sim Beta(na, na)$$

then we generate values of T in this way: first we generate b from $Beta(na, na)$ and then we apply the transformation $t^* = \frac{b}{1-b}$ and collect only the values $t^* \geq 1$. We generated I values of statistic T, $t_1^*, \ldots, t_I^*$, and approximate a p-value by the following Monte Carlo sum:

$$\frac{\sum_{i=1}^{I} \mathbf{1}\{t_i^* > t_{obs}\}}{I}$$

At the beginning we considered the statistic $t = \left|\frac{\overline{x}}{\overline{y}} - 1\right|$ with the following null density:

$$f\left(t = \left|\frac{\overline{x}}{\overline{y}} - 1\right| | a\right) = \frac{\Gamma(2na)}{\Gamma^2(na)} \left(\frac{(1+t)^{na-1}}{(2+t)^{2na}} + \mathbf{1}_{(0,1)}(t) \frac{(1-t)^{na-1}}{(2-t)^{2na}}\right)$$

This density leads to a numerical problem. In fact when $t < 1$ we have to compute the log of the sum $\left(\frac{(1+t)^{na-1}}{(2+t)^{2na}} + \mathbf{1}_{(0,1)}(t) \frac{(1-t)^{na-1}}{(2-t)^{2na}}\right)$ and is often the case that $\left(\frac{(1+t)^{na-1}}{(2+t)^{2na}} + \mathbf{1}_{(0,1)}(t) \frac{(1-t)^{na-1}}{(2-t)^{2na}}\right) \approx 0$. Therefore we cannot gain the numerical advantages of a log-scale. This does not happen when $t > 1$.

B.1 Plug-in p-value

The Likelihood under H_0 is:

$$L(\theta, a) = \theta^{2na}\Gamma^{-2n}(a)\left(\prod_{i=1}^{n} x_i y_i\right)^{a-1} \exp\{-\theta n(\overline{x}+\overline{y})\}$$

and the maximum likelihood estimators for θ and a are respectively $\widehat{\theta}$ and $\widehat{a}$ given by

$$\begin{aligned} \widehat{\theta} &= \widehat{a}\frac{2}{\overline{x}+\overline{y}} \\ \widehat{a} &: \psi^{(0)}(a) - \log(a) = \frac{1}{2}\left(\frac{\sum_{i=1}^{n}\log x_i + \sum_{i=1}^{n}\log y_i}{n}\right) - \log\left(\frac{\overline{x}+\overline{y}}{2}\right) \end{aligned} \quad \text{(B.2)}$$

where $\psi^{(k)}(a)$ is

$$\psi^{(k)}(a) = \frac{\partial^k}{\partial a^k}\log\Gamma(a)$$

therefore $\psi^{(0)}(a)$ and $\psi^{(1)}(a)$ are respectively the digamma and trigamma function. The equation (B.2) is implicit in a and it has been solved using the Bisection method (or Bolzano's method). The expression of $\widehat{\theta}$ has been reported here but has not been calculated. The algorithm used to calculate the trigamma is from Schneider, (1978) based on the following approximation

$$\psi^{(k)}(a) \approx (-1)^{k-1}\left[\frac{(k-1)!}{a^k} + \frac{k!}{2a^{k+1}} + \sum_{i=1}^{\infty} B_{2i}\frac{(2i+k-1)!}{(2i)!a^{2i+i}}\right],$$

where $B_{2i}, i = 1, \ldots, \infty$ are the Bernoulli Numbers. The algorithm used for the digamma function is based on Bernardo, (1976) where we can find the following remark.

Remark 17 *(Bernardo, 1976). Note that the trigamma function behaves in the limits according to*

$$\psi^{(1)}(a) \approx \frac{1}{a} + \frac{1}{2a^2} + \frac{1}{6a^3} - \frac{1}{30a^5} + \frac{1}{45a^7} - \frac{1}{30a^9} + O\left(\frac{1}{a^{11}}\right) \text{ for } a \to \infty \quad \text{(B.3)}$$

$$\psi^{(1)}(a) \approx \frac{1}{a^2} + o(1) \text{ for } a \to 0, \quad \text{(B.4)}$$

where

$$o(1) < \sum_{i=1}^{\infty}\frac{1}{i^2} = \frac{\pi^2}{6}$$

We approximate the p_{plug} using Algorithm 16 for $a = \widehat{a}$.

B.2 Posterior p-value

The prior

Two non-informative priors were considered: the Jeffreys's prior and the Reference prior (Berger and Bernardo, 1992a, 1992b, 1992c). The inference under the Reference prior provides better results than with the Jeffreys's prior as showed in Liseo (1993).

According to the algorithm provided by (Berger and Bernardo, 1992a) the reference prior is obtained by maximizing an asymptotic version of the expected Shannon information. Using Remarks 17 we have that this prior is improper because is unbounded for $a \to 0$. Applying the factorization theorem to the prior in (2.10) we have that θ and a are independent parameters

$$\pi(\theta, a) \propto \pi(\theta)\pi(a).$$

The kernel of the posterior of the joint distribution is given by:

$$\begin{aligned}\pi(\theta, a|\mathbf{x}, \mathbf{y}) &\propto L(\theta, a)\pi(\theta, a) \\ &\propto \Gamma^{-2n}(a)\prod_{i=1}^{n}(x_i y_i)^{a-1}\sqrt{\frac{a\psi^{(1)}(a)-1}{a}}\underbrace{\theta^{2na-1}\exp\{-\theta n(\overline{x}+\overline{y})\}}_{\pi(\theta|a,\mathbf{x},\mathbf{y})=Gamma(2na,n(\overline{x}+\overline{y}))}\end{aligned}$$

and the marginal posterior distribution for a is:

$$\pi(a|\mathbf{x}, \mathbf{y}) \propto \frac{\Gamma(2na)}{\Gamma^{2n}(a)}\frac{\prod_{i=1}^{n}(x_i y_i)^{a-1}}{[n(\overline{x}+\overline{y})]^{2na}}\sqrt{\frac{a\psi^{(1)}(a)-1}{a}}$$

the posterior predictive distribution for T is given by:

$$\begin{aligned}m_{post}(t|\mathbf{x}, \mathbf{y}) &= \int_{\mathcal{R}^+}\int_{\mathcal{R}^+} f(t|a)\pi(\theta, a|\mathbf{x}, \mathbf{y})\, d\theta da \\ &= \int_{\mathcal{R}^+} f(t|a)\pi(a|\mathbf{x}, \mathbf{y})\, da \\ &= \int_{\mathcal{R}^+}\frac{2^{2na}\Gamma\left(\frac{1}{2}+na\right)}{\Gamma(na)\sqrt{\pi}}\frac{t^{na-1}}{(1+t)^{2na}}\frac{\Gamma^2(2na)}{\Gamma^2(na)\Gamma^{2n}(a)}\frac{\prod_{i=1}^{n}(x_i y_i)^{a-1}}{[n(\overline{x}+\overline{y})]^{2na}}\times \\ &\quad \times\sqrt{\frac{a\psi^{(1)}(a)-1}{a}}da\end{aligned}$$

This integral has been approximated by first obtaining $\pi(a|\mathbf{x}, \mathbf{y})$ through a Metropolis Hastings Sampler (MH) with a gamma proposal distribution with mode on $\overline{a} = \arg\max_a \pi(a|\mathbf{x}, \mathbf{y})$ and variance $\left(-\frac{\partial^2}{\partial a^2}\pi(a|\mathbf{x}, \mathbf{y})|_{a=\overline{a}}\right)^{-1}$. The MH chain has been run for 51000 steps with a burn-in of 1000. Figure 2.3 (top) suggests that this number of replicates is enough because the modes of the simulated are located appropriately respect to the maximum estimate with a gaussian kernel. The initial point was set in the mode of the posterior and the percentage of acceptance rate of our MCMC sampler was around 30% - 50%.

We approximate the p_{post} using Algorithm 16 where the values of a are obtained from the posterior.

B.3 Partial Posterior Predictive p-value

The marginal partial posterior distribution for parameter a is given by:

$$\pi(a|\mathbf{x}, \mathbf{y}\backslash t(\mathbf{x}, \mathbf{y})) \propto \int_{\mathcal{R}^+}\frac{L(\theta, a)\pi(\theta, a)}{f(t(\mathbf{x}, \mathbf{y})|a)}d\theta$$

$$\begin{aligned}
&\propto \frac{\pi\left(a|\mathbf{x},\mathbf{y}\right)}{f\left(t\left(\mathbf{x},\mathbf{y}\right)|a\right)} \\
&\propto \frac{\Gamma^2\left(2na\right)}{\Gamma^2\left(na\right)\Gamma^{2n}\left(a\right)}\frac{\prod_{i=1}^{n}\left(x_i y_i\right)^{a-1}}{\left[n\left(\overline{x}+\overline{y}\right)\right]^{2na}}\sqrt{\frac{a\psi^{(1)}\left(a\right)-1}{a}}\times \\
&\quad\times\left(\frac{2^{2na}\Gamma\left(\frac{1}{2}+na\right)}{\Gamma\left(na\right)\sqrt{\pi}}\frac{t\left(\mathbf{x},\mathbf{y}\right)^{na-1}}{\left(1+t\left(\mathbf{x},\mathbf{y}\right)\right)^{2na}}\right)^{-1}.
\end{aligned}$$

The partial posterior distribution was approximated using the same MH chain as for the posterior, except that the initial point and the proposal setup was made according to $\pi\left(a|\mathbf{x},\mathbf{y}\backslash t\left(\mathbf{x},\mathbf{y}\right)\right)$ rather than $\pi\left(a|\mathbf{x},\mathbf{y}\right)$. We calculated the Partial Posterior predictive p-value, p_{ppost}, by approximating the integral

$$m\left(t|\mathbf{x},\mathbf{y}\backslash t\left(\mathbf{x},\mathbf{y}\right)\right)=\int_{\mathcal{R}^+} f\left(t|a\right)\pi\left(a|\mathbf{x},\mathbf{y}\backslash t\left(\mathbf{x},\mathbf{y}\right)\right)da$$

using Algorithm 16 where the values of a are those obtained from the partial posterior distribution $\pi\left(a|\mathbf{x},\mathbf{y}\backslash t\left(\mathbf{x},\mathbf{y}\right)\right)$.

B.3.1 The Metropolis Hasting algorithm for approximate the posterior and the partial posterior distribution of a.

We provide here more details on the approximation of $\pi\left(a|\mathbf{x},\mathbf{y}\backslash t\left(\mathbf{x},\mathbf{y}\right)\right)$ and $\pi\left(a|\mathbf{x},\mathbf{y}\right)$. The differences between the two algorithm stay only in the density to be approximated and not in the procedure, therefore we will illustrate it referring to the approximation of $\pi\left(a|\mathbf{x},\mathbf{y}\backslash t\left(\mathbf{x},\mathbf{y}\right)\right)$. We will refer to the numerical example of inference under the Gamma model provided in Chapter 3. In particular we will refer to the analysis reproduced in Figure 2.3 where we run a chain of length 51000 steps.

Figure B.1 compare the proposal distribution with the kernel of $\pi\left(a|\mathbf{x},\mathbf{y}\backslash t\left(\mathbf{x},\mathbf{y}\right)\right)$ and its approximation. For approximating the $\pi\left(a|\mathbf{x},\mathbf{y}\backslash t\left(\mathbf{x},\mathbf{y}\right)\right)$ we setup a gamma distribution with mode in $\overline{a}=\arg\max_a \pi\left(a|\mathbf{x},\mathbf{y}\backslash t\left(\mathbf{x},\mathbf{y}\right)\right)$ and variance

$$\left(-\frac{\partial^2}{\partial a^2}\pi\left(a|\mathbf{x},\mathbf{y}\backslash t\left(\mathbf{x},\mathbf{y}\right)\right)|_{a=\overline{a}}\right)^{-1}.$$

This choice of the proposal distribution was suggested because of the similarity of the shape of a Gamma density with the kernel density as shown in Figure B.1, therefore to mimic the $\pi\left(a|\mathbf{x},\mathbf{y}\backslash t\left(\mathbf{x},\mathbf{y}\right)\right)$ was just necessary to match the proposal mode with the mode of the $\pi\left(a|\mathbf{x},\mathbf{y}\backslash t\left(\mathbf{x},\mathbf{y}\right)\right)$. In this way we guarantee us to explore all the support where the kernel provide significant probability mass. The convergence of the Markov chain to the stationary distribution is guaranteed because the proposal distribution has the same support of the $\pi\left(a|\mathbf{x},\mathbf{y}\backslash t\left(\mathbf{x},\mathbf{y}\right)\right)$. In fact the Autocorrelation Function plotted in Figure B.2 (b) suggests that the chain behaves acceptably, because the mixture of the posterior and the proposal allows the chain to jump with a fairly high acceptance rate (about 50%). Moreover the burn in period of 1000 steps seems to be adequately because the chain seems to stationary as shown in Figure B.2 (a).

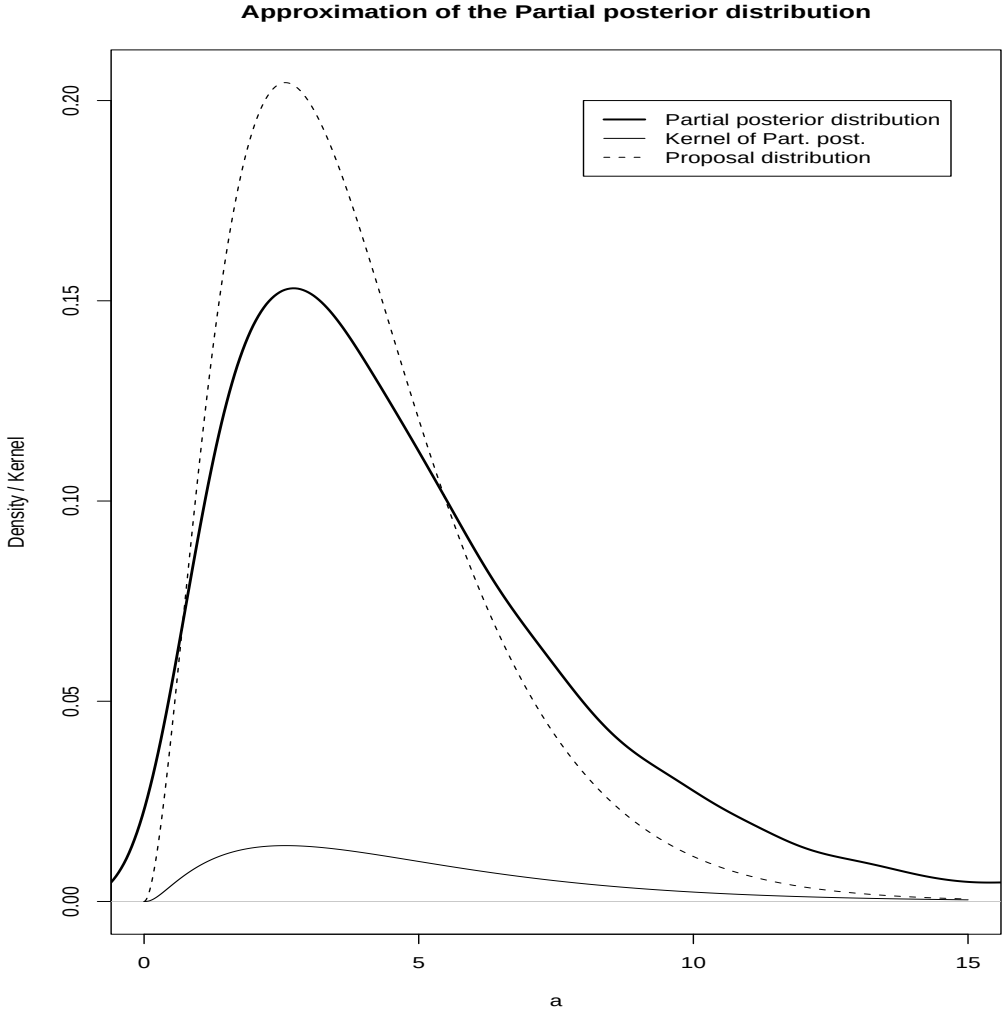

Figure B.1: Kernel of the partial posterior distribution, proposal and *approximated* partial posterior distribution.

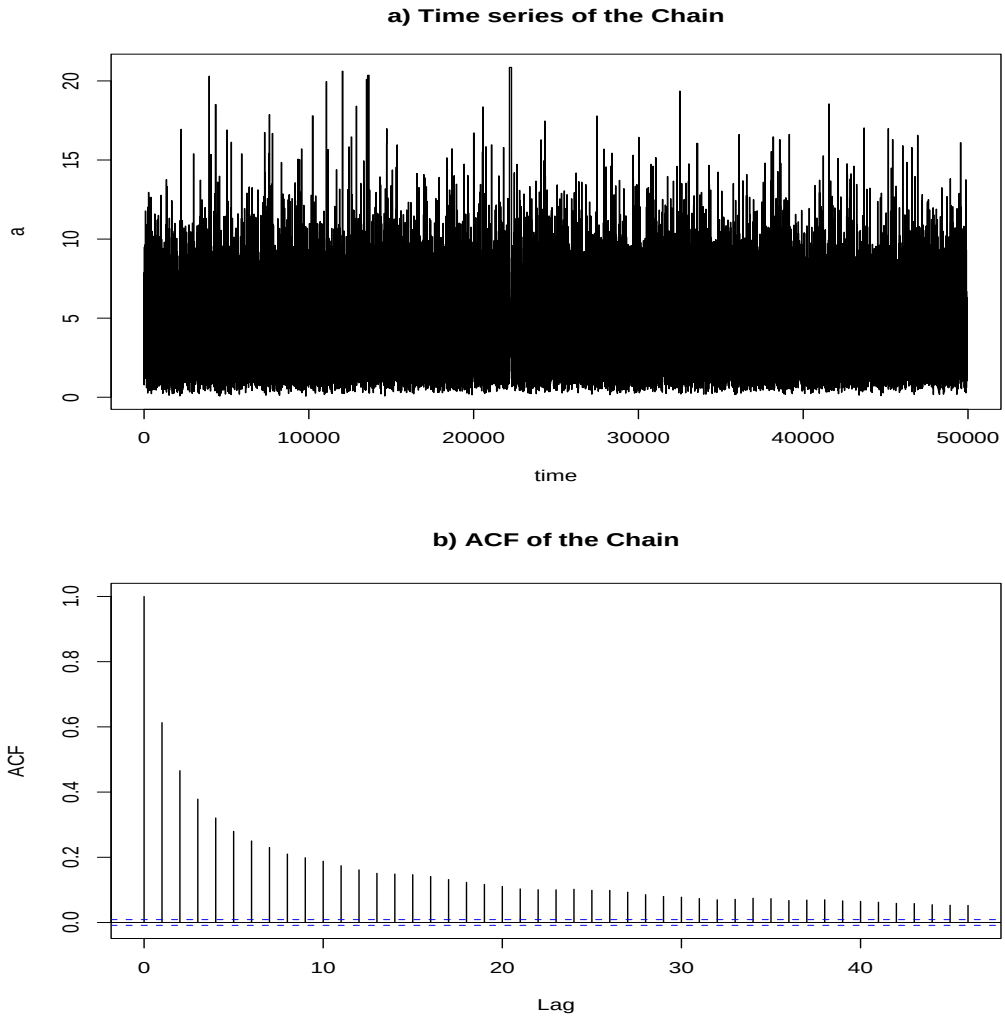

Figure B.2: (a) Time series of the Markov chain and (b) Auto Correlation Function.

B.3.2 The partial posterior distribution is proper

We provide here the proof of proposition (3), which states that $\pi\left(a|\mathbf{x},\mathbf{y}\backslash t\left(\mathbf{x},\mathbf{y}\right)\right)$ is a proper probability density for $n \geq 2$.

Proof. The proof is based on the integrability of $\pi\left(a|\mathbf{x},\mathbf{y}\backslash t\left(\mathbf{x},\mathbf{y}\right)\right)$ for $a \to \infty$ (a) and for $a \to 0$ (b).

a) $\pi\left(a|\mathbf{x},\mathbf{y}\backslash t\left(\mathbf{x},\mathbf{y}\right)\right)$ can be written in the following form

$$\pi\left(a|\mathbf{x},\mathbf{y}\backslash t\left(\mathbf{x},\mathbf{y}\right)\right) \propto K\left(a\right) = \frac{\Gamma^2\left(2na\right)}{\Gamma^2\left(na\right)\Gamma^{2n}\left(a\right)\Gamma\left(\frac{1}{2}+na\right)}\sqrt{\psi^{(1)}\left(a\right)-a^{-1}}C^a$$

where C is the following constant

$$\begin{aligned} C &= \frac{\prod_{i=1}^n x_i \prod_{i=1}^n y_i}{4\left(\sum_{i=1}^n x_i\right)^n\left(\sum_{i=1}^n y_i\right)^n} \\ &= \frac{1}{4n}\frac{G_x G_y}{\overline{x}\,\overline{y}}. \end{aligned}$$

and G_x, G_y represent respectively the geometric mean of $\mathbf{x}$ and $\mathbf{y}$. Using (B.3) we can approximate

$$\sqrt{\psi^{(1)}\left(a\right)-a^{-1}} \approx \frac{1}{2a^2}$$

and the asymptotic behavior of the gamma function are

$$\begin{aligned} \frac{\Gamma^2\left(2na\right)}{\Gamma^2\left(na\right)\Gamma^{2n}\left(a\right)\Gamma\left(\frac{1}{2}+na\right)} &\approx \frac{\Gamma^2\left(a\right)}{\Gamma^2\left(a\right)\Gamma^{2n}\left(a\right)\Gamma\left(a\right)} \\ &\approx \frac{1}{\Gamma^{2n+1}\left(a\right)}. \end{aligned}$$

Therefore for $a \to \infty$ we have

$$K\left(a\right) \to \frac{1}{2a^2}\frac{1}{\Gamma^{2n+1}\left(a\right)}C^a,$$

Note also that $C < 1$, which is always true because of the well known relation between geometric and arithmetic means:

$$G_x \leq \overline{x}, G_y \leq \overline{y}.$$

Therefore $\pi\left(a|\mathbf{x},\mathbf{y}\backslash t\left(\mathbf{x},\mathbf{y}\right)\right)$ is integrable for $a \to \infty$.

b) Using approximations (B.4) for $a \to 0$ we have

$$\begin{aligned} \lim_{a\to 0} K\left(a\right) &= \lim_{a\to 0}\frac{\sqrt{a^{-1}-1}}{\left(2na\right)^2}\left(na\right)a^{2n}\underbrace{\left(\frac{1}{2}+na\right)}_{\sim 1/2}\frac{1}{\sqrt{a}} \\ &= \lim_{a\to 0} Ca^{-1/2}a^{-2}aa^{2n}a^{-1/2}, C = \frac{1}{8n} \\ &= \lim_{a\to 0} Ca^{2n-2} \end{aligned}$$

and $\pi\left(a|\mathbf{x},\mathbf{y}\backslash t\left(\mathbf{x},\mathbf{y}\right)\right)$ is integrable if $n-1>0$ which is satisfied for $n \geq 2$. ■