Multiple testing using the posterior probability of half-space: application to gene expression data

by

Aurélie Labbe

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2005

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

We consider the problem of testing the equality of two sample means, when the number of tests performed is large. Applying this problem to the context of gene expression data, our goal is to detect a set of genes differentially expressed under two treatments or two biological conditions. A null hypothesis of no difference in the gene expression under the two conditions is constructed. Since such a hypothesis is tested for each gene, it follows that thousands of tests are performed simultaneously, and multiple testing issues then arise. The aim of our research is to make a connection between Bayesian analysis and frequentist theory in the context of multiple comparisons by deriving some properties shared by both p-values and posterior probabilities. The ultimate goal of this work is to use the posterior probability of the one-sided alternative hypothesis (or equivalently, posterior probability of the half-space) in the same spirit as a p-value. We show for instance that such a Bayesian probability can be used as an input in some standard multiple testing procedures controlling for the False Discovery rate.

The first chapter of this thesis presents an introduction to the problem of cDNA microarray data. The underlying biological principles of this type of data, as well as the associated statistical issues are discussed. In the second chapter, we follow the work of Dudley & Haughton (2002) regarding the asymptotic normality of posterior probabilities of half-spaces. We show that such a probability shares with the frequentist p-value the property of uniformity under the null hypothesis. This result holds asymptotically, when the number of observations available for each test is large enough. Our approach is based on the observation that uniformity under the null hypothesis (as p-values are assumed to be) is the main property used in the multiple testing procedure developed by Benjamini and Hochberg (1995). We are then able to use the posterior probability, defined as an input to this procedure, in the same spirit as a p-value. We note that such a probability can also be seen as a test statistic from which the distribution under the null hypothesis is known. As a result, it can also be used in any extension of the Benjamini-Hochberg procedure, providing a control of the False Discovery or False Negative Rate.

Motivated by the case of microarray data, where the number of observations per gene is small, we show in the third chapter that the uniform property holds in a non-asymptotic manner, under a non-informative or a conjugate gamma model. A goodness of fit study on several microarray datasets is performed as well as an extended simulation study. This gamma model is extended in the fourth chapter to a multiplicative random effect ANOVA model, taking the arrays and dyes effects into consideration. Other models, such as inverse Gaussian models are also considered in the fifth chapter. In such cases, the uniform property of the posterior probability considered can be observed empirically when the sample size is small. Results using these models are very encouraging. A case study is presented in Chapter 6 using three microarray datasets resulting from a collaborative study between the Universities of McMaster and Waterloo. The methods developed in this thesis are then applied and results are compared. Our future work is described in the last chapter, and a brief discussion of the work proposed in this thesis is finally included.

Acknowledgements

Working on this thesis represented five years of my life. Along the way, I met many people without whom I would not be where I am today. Without any doubts, the most important one has been my research supervisor Mary Thompson, for whom I have a very deep respect. Mary, I thank you for your support, technical, financial or even mental, your encouragements and your positivism. You were always there in the critical points of my research, at the same time giving me enough freedom so that I feel confident to continue now by myself. When my turn will come to supervise students, I will definitively remember the way you guided me through.

I also want to thank my committee members: Hugh Chipman, Robert Gentleman, Jerry Lawless and Barbara Moffatt, for the time they spent reading my thesis and for their very helpful comments and feedbacks.

A PhD represents a lot of time spent at the university...I want to thank the Statistics and Actuarial Sciences Department for the financial support I received as well as all the administrative staff, especially Mary-Lou, Lucy and Nandanee, who helped me with so many technical difficulties. Thanks also to those with whom I shared these 5 years: Marc, Chantal, Philippe, Cody, Zeny and all my Latino friends.

To finish, a special word for three persons that I met here, and who are very special for me: Leilei, Hanna and Luis. For you, Waterloo was just the beginning...

Finally, I wish to thank the Ontario Graduate Scholarship for supporting this research during two years.

Et non, je ne vous oublie pas...merci à ma famille, et tout spécialement à mes parents et à Adèle, pour savoir rester si proche de moi à 6000km de distance. Je termine avec une pensée toute spéciale pour mon grand-père, qui je crois aurait été bien fier de moi.

Contents

1 Some background on microarray data 1

1.1 Some terminology and motivation 2

1.2 Microarray technology 6

1.3 Statistical issues in a microarray experiment 8

1.3.1 Variation and design of experiment 9

1.3.2 Normalization for cDNA microarray data 10

1.3.3 Some statistical problems 12

2 Multiple testing using posterior probabilities of half spaces 17

2.1 The multiple testing issue 18

2.2 The Bayesian framework 23

2.2.1 An overview of Bayesian models 23

2.2.2 Definition of the model 25

2.3 Uniform property of the posterior probability of half-space, under H_0 . . 28

2.4 Using the posterior probability of the one-sided alternative hypothesis in a multiple testing framework 31

3 A particularly favorable case for the half-space approach 35

3.1 The non-informative case 36

3.2 An extension of the uniform property to the informative case 39

3.3 Application: Goodness of fit of the conjugate gamma model for microarray datasets 44

3.4 Simulation study 49

3.4.1 Simulation design 49

3.4.2 Influence of the controlled FDR level δ 54

3.4.3 Influence of p 57

3.4.4 Influence of m 59

3.4.5 Influence of the estimated proportion of differentially expressed genes, $\hat{p}$ 61

3.4.6 Summary and discussion 63

4 A multiplicative random effect ANOVA model for microarray data 67

4.1 An additive ANOVA model for microarray data 68

4.1.1 A fixed effect model 71

4.1.2 A mixed ANOVA model 73

4.2 A multiplicative random effects ANOVA model 74

4.2.1 Definition of the model 75

4.2.2 Inference about the parameters via Gibbs sampling algorithm . . 79

4.2.3 Estimation of the hyperparameters 82

4.2.4 Normalization of the data and detection of differentially expressed genes 83

4.3 Application: Study of the DBLFLIP dataset 86

4.3.1 Goodness of fit study 89

4.3.2 Testing for differentially expressed genes 92

4.4 Study of a simulated dataset 92

4.5 Discussion 98

5 Probability of half space using inverse Gaussian models 99

5.1 Hierarchical models involving the inverse Gaussian distribution 100

5.1.1 The inverse Gaussian distribution : an overview 101

5.1.2 The natural conjugate model 102

5.1.3 The non-informative model 106

5.1.4 Another hierarchical model involving the inverse Gaussian distribution 107

5.1.5 Estimation of the hyper-parameters and tests for differentially expressed genes 112

5.2 Goodness of fit study 113

5.2.1 Application to four microarray datasets 113

5.2.2 A simulation study: Goodness of fit and robustness 121

5.3 Multiple testing using inverse Gaussian models: a simulation study . . . 122

5.3.1 Error rates and robustness 122

5.3.2 Asymptotic study 124

5.4 Discussion 125

6 Case-study: Using functional genomics to discover genes important in abiotic stress tolerance 127

6.1 Overview of the project 128

6.2 Experiment pre-processing 128

6.3 Design of the microarray 130

6.4 Normalization of the data 132

6.5 Results 134

6.6 Discussion 141

7 Future research 143

7.1 A Multiplicative ANOVA model for inverse Gaussian data 144

7.2 Generalization to more than 2 treatments 144

7.3 Detection of regulatory motifs in DNA sequences 148

7.3.1 Statistical models 149

7.3.2 Linking gene expression data with motif finding analysis 151

8 Discussion 155

A Raw data plots 159

B COLD dataset: Plots of the gene detected 166

C SALINITY dataset: Plots of the gene detected 173

D DROUGHT dataset: Plots of the gene detected 180

List of Tables

2.1 Outcomes when testing n hypotheses for a t-dependent test procedure . . 18

3.1 Design of the experiment 45

3.2 Estimated hyperparameters of the gamma conjugate model for the three datasets 49

3.3 Error rates for the simulation study with respect to the level of FDR control δ. Data were simulated with $p = 0.1$ and $m = 10$. 55

3.4 Error rates for the simulation study with respect to the proportion of differentially expressed genes p. Data were simulated with $m = 10$. A level of FDR control of $\delta = 5\%$ was used. 58

3.5 Error rates for the simulation study with respect to the number of observations available for each gene, m. Data were simulated with $p = 0.1$. A level of FDR control of $\delta = 5\%$ was used. 60

3.6 Robustness of the p_i and the v_i method to the estimation of the parameter p. Data were simulated with $m = 10$ and $p = 0.1$. A level of FDR control of $\delta = 5\%$ was used. 62

4.1 A dye-swap experiment involving two treatments T_1 and T_2. 70

4.2 Confounding structure for the dye-swap experiment 70

4.3 A loop design involving K treatments $T_1, \ldots, T_K$. 71

4.4 Estimated parameters γ's for the DBLFLIP dataset 86

4.5 Estimated parameters a's and b's for the DBLFLIP dataset. Note that assumptions of Theorem 4.1 require the parameters (a, b) to be the same for both treatments. 86

4.6 Quantiles of the posterior expectations of the gene-specific effects for the DBLFLIP dataset 88

4.7 Real and estimated hyperparameters a's and b's 94

4.8 Real and estimated hyperparameters γ's 94

4.9 Error rates 96

5.1 Hyperparameters for the 2 informative inverse Gaussian models 114

5.2 Goodness of fit distances for the Gamma model and for the IG/IG model, for the four datasets 120

5.3 Robustness to the misspecification of the true distribution of the data, for the non conjugate inverse Gaussian model and for the gamma conjugate model 121

5.4 Error rates for the three models, for each type of dataset. FDR was controlled at a level of 5%. 123

6.1 The four experiments 130

6.2 experimental design 130

6.3 Size of each datasets. Initial datasets contain 4896 genes. 131

6.4 Number of genes detected. Numbers in bold represent the number of genes detected by the most appropriate method, according to the goodness of fit study. 139

7.1 illustration of primary and secondary ES target genes 146

7.2 Situation that may occur when genes are ES targets. Situations 1-4 represent primary targets and situations 5-8 represent secondary targets . . 147

List of Figures

1.1 Double helix DNA structure 3

1.2 Figure (a) represents DNA before transcription. Figure (b) represents the transcription step: the DNA should unwind so that one of its strands can be used as a template to synthesize a complementary strand. 4

1.3 The central dogma of molecular biology 5

2.1 Illustration of the hyperplane and half-space in the case $\Theta = \mathbb{R}^{2+}$ 29

3.1 Histograms versus predictive density function for the four datasets using the gamma conjugate model (log-scale) 48

3.2 Histograms of the p-values from the likelihood ratio test (left) and of the p_i's (right) under H_0: Simulated data, $m = 4$. 61

4.1 DBLFLIP dataset: convergence rate of the Gibbs sampling posterior expectations for the first gene 87

4.2 Box-plots of the residuals for the multiplicative model (on the right), and for the model considering only the treatment-gene effects (on the left) . . 90

4.3 Quantile-quantile plots for the DBLFLIP dataset, using the multiplicative model 91

4.4 Quantile-quantile plots for the DBLFLIP dataset, using MAANOVA . . . 93

4.5 Quantile-quantile plots for the simulated dataset, using the multiplicative model 95

4.6 Histogram of the posterior probabilities p_g under the null hypothesis . . . 97

5.1 Inverse Gaussian distribution function 101

5.2 Fit for the 3 informative models: TCDD dataset 115

5.3 Fit for the 3 informative models: DBLFLIP dataset 116

5.4 Fit for the 3 informative models: COLD dataset 117

5.5 Fit for the 3 informative models: FIELD dataset 118

5.6 Fit for the GAM and IG/IG models: Datasets are simulated according the right model 119

5.7 Posterior probability p_i under H_0 for $m = 5$, $m = 10$ and $m = 20$. Results based on a single simulated dataset 124

6.1 RI plot: Array2, Cold dataset 134

6.2 RI plot: Array5, Salinity dataset 135

6.3 QQplots for the multiplicative gamma model: Cold dataset 136

6.4 QQplots for the multiplicative gamma model: Drought dataset 137

6.5 QQplots for the multiplicative gamma model: Salinity dataset 138

6.6 Predictive distribution for the MUL model, for the three datasets, corrected for the gene-specific array and dye effects 139

6.7 Predictive distribution for the GAM model, for the three datasets 140

6.8 Predictive distribution for the IG model, for the three datasets 141

A.1 Raw data: TCDD dataset 160

A.2 Raw data: DBLFLIP dataset 161

A.3 Raw data: COLD dataset 162

A.4 Raw data: SALINITY dataset 163

A.5 Raw data: DROUGHT dataset 164

A.6 Raw data: FIELD dataset 165

B.1 Genes detected: COLD dataset. The symbol "1" represents the genes detected by the inverse-Gaussian model. Symbols "1" and "2" represent the genes detected by the GAM model (including those detected by the inverse-Gaussian model). Symbols "1", "2" and "3" represent the genes detected by the multiplicative gamma model 167

B.2 Genes detected: COLD dataset. The symbol "1" represents the genes detected by the inverse-Gaussian model. Symbols "1" and "2" represent the genes detected by the GAM model (including those detected by the inverse-Gaussian model). Symbols "1", "2" and "3" represent the genes detected by the multiplicative gamma model 168

B.3 Genes detected: COLD dataset. The symbol "1" represents the genes detected by the inverse-Gaussian model. Symbols "1" and "2" represent the genes detected by the GAM model (including those detected by the inverse-Gaussian model). Symbols "1", "2" and "3" represent the genes detected by the multiplicative gamma model 169

B.4 Genes detected: COLD dataset. The symbol "1" represents the genes detected by the inverse-Gaussian model. Symbols "1" and "2" represent the genes detected by the GAM model (including those detected by the inverse-Gaussian model). Symbols "1", "2" and "3" represent the genes detected by the multiplicative gamma model 170

B.5 Genes detected: COLD dataset. The symbol "1" represents the genes detected by the inverse-Gaussian model. Symbols "1" and "2" represent the genes detected by the GAM model (including those detected by the inverse-Gaussian model). Symbols "1", "2" and "3" represent the genes detected by the multiplicative gamma model 171

B.6 Genes detected: COLD dataset. The symbol "1" represents the genes detected by the inverse-Gaussian model. Symbols "1" and "2" represent the genes detected by the GAM model (including those detected by the inverse-Gaussian model). Symbols "1", "2" and "3" represent the genes detected by the multiplicative gamma model 172

C.1 Genes detected: SALINITY dataset. The symbol "1" represents the genes detected by the inverse-Gaussian model. Symbols "1" and "2" represent the genes detected by the GAM model (including those detected by the inverse-Gaussian model). Symbols "1", "2" and "3" represent the genes detected by the multiplicative gamma model 174

C.2 Genes detected: SALINITY dataset. The symbol ”1” represents the genes detected by the inverse-Gaussian model. Symbols ”1” and ”2” represent the genes detected by the GAM model (including those detected by the inverse-Gaussian model). Symbols ”1”, ”2” and ”3” represent the genes detected by the multiplicative gamma model 175

C.3 Genes detected: SALINITY dataset. The symbol ”1” represents the genes detected by the inverse-Gaussian model. Symbols ”1” and ”2” represent the genes detected by the GAM model (including those detected by the inverse-Gaussian model). Symbols ”1”, ”2” and ”3” represent the genes detected by the multiplicative gamma model 176

C.4 Genes detected: SALINITY dataset. The symbol ”1” represents the genes detected by the inverse-Gaussian model. Symbols ”1” and ”2” represent the genes detected by the GAM model (including those detected by the inverse-Gaussian model). Symbols ”1”, ”2” and ”3” represent the genes detected by the multiplicative gamma model 177

C.5 Genes detected: SALINITY dataset. The symbol ”1” represents the genes detected by the inverse-Gaussian model. Symbols ”1” and ”2” represent the genes detected by the GAM model (including those detected by the inverse-Gaussian model). Symbols ”1”, ”2” and ”3” represent the genes detected by the multiplicative gamma model 178

C.6 Genes detected: SALINITY dataset. The symbol "1" represents the genes detected by the inverse-Gaussian model. Symbols "1" and "2" represent the genes detected by the GAM model (including those detected by the inverse-Gaussian model). Symbols "1", "2" and "3" represent the genes detected by the multiplicative gamma model 179

D.1 Genes detected: DROUGHT dataset. The symbol "1" represents the genes detected by the inverse-Gaussian model. Symbols "1" and "2" represent the genes detected by the GAM model (including those detected by the inverse-Gaussian model). Symbols "1", "2" and "3" represent the genes detected by the multiplicative gamma model 181

D.2 Genes detected: DROUGHT dataset. The symbol "1" represents the genes detected by the inverse-Gaussian model. Symbols "1" and "2" represent the genes detected by the GAM model (including those detected by the inverse-Gaussian model). Symbols "1", "2" and "3" represent the genes detected by the multiplicative gamma model 182

D.3 Genes detected: DROUGHT dataset. The symbol "1" represents the genes detected by the inverse-Gaussian model. Symbols "1" and "2" represent the genes detected by the GAM model (including those detected by the inverse-Gaussian model). Symbols "1", "2" and "3" represent the genes detected by the multiplicative gamma model 183

D.4 Genes detected: DROUGHT dataset. The symbol "1" represents the genes detected by the inverse-Gaussian model. Symbols "1" and "2" represent the genes detected by the GAM model (including those detected by the inverse-Gaussian model). Symbols "1", "2" and "3" represent the genes detected by the multiplicative gamma model 184

D.5 Genes detected: DROUGHT dataset. The symbol "1" represents the genes detected by the inverse-Gaussian model. Symbols "1" and "2" represent the genes detected by the GAM model (including those detected by the inverse-Gaussian model). Symbols "1", "2" and "3" represent the genes detected by the multiplicative gamma model 185

D.6 Genes detected: DROUGHT dataset. The symbol "1" represents the genes detected by the inverse-Gaussian model. Symbols "1" and "2" represent the genes detected by the GAM model (including those detected by the inverse-Gaussian model). Symbols "1", "2" and "3" represent the genes detected by the multiplicative gamma model 186

Chapter 1

Some background on microarray data

Genetics, as practiced for most of its history, has been a science that deals with the transmission of traits from parental organisms to their offsprings. In 1944, molecular biology was born, when Oswald Avery (Oswald, Avery Colin & McCarty (1944)) discovered that chromosomes are composed of a polymer called deoxyribonucleic acid (DNA). Since then, molecular geneticists are able to approach the subject from its foundation: the molecule. Instead of examining phenotypic characters, molecular geneticists examine genes, studying their chemical structure, their activity and regulation. In the past decades, molecular approaches to genetics have become more and more sophisticated and recent advances in genome sequencing are opening whole new approaches to understanding the regulation and function of genetic material. Today, with the availability of complete genome sequences of numerous organisms, thousands of genes have been identified. The new task of molecular biologists is to understand the roles of each gene in the genome. Microarray technology can provide a partial answer to this question by measuring and comparing the expression of thousands of genes simultaneously, under two conditions. This new technology has also drawn the attention of statisticians: the amount of data available, the experimental design, the high variability, the significance of the results are all issues that need to be treated carefully.

In this chapter, we first review some basic terminology used in molecular biology. For a more detailed introduction, refer to the book Molecular Biology (Weaver (2002)). The second section provides a brief overview of DNA microarray technology and finally, the statistical issues associated with gene expression data are presented in the last section.

1.1 Some terminology and motivation

At a microscopic level, every living organism is composed of cells. Humans, for example, are composed of about 10 trillion cells, divided into about 200 types, each specialized for such distinctive functions as memory, sight, movement or digestion. Despite the diversity of their functions, all cells in an organism share a common fundamental property: the storage of the genetic information inherited from the gametes. This genetic material is maintained in a three dimensional structure: DNA. How genetic information is replicated and transmitted from cell to cell and organism to organism is a question that is central to all of biology. First, each trait of an individual is determined by a pair of inherited factors called genes with one gene copy (an allele) being inherited from each parent. The role of the genes in the cell is crucial: genes replicate faithfully, they direct the production of proteins and they accumulate mutations allowing evolution. In order to understand how genes work, one should examine their molecular composition. Most genes are made of DNA, composed of a double helical structure (two DNA strands wound around each other). Each DNA strand is composed of linked nucleotides, made of three elements: a phosphate, a deoxyribose sugar and a nucleic acid base. There are four distinct bases: adenine (A), guanine (G), cytosine (C) and thymine (T). The two DNA strands are complementary, in the sense that wherever we find an A in one strand, a T is present in the opposite one. Similarly, the base G is complementary to the base C. This complementarity allows DNA to be replicated faithfully. Finally, the two DNA strands are held together by hydrogen bonds: two between A-T basepair and three between G-C basepair.

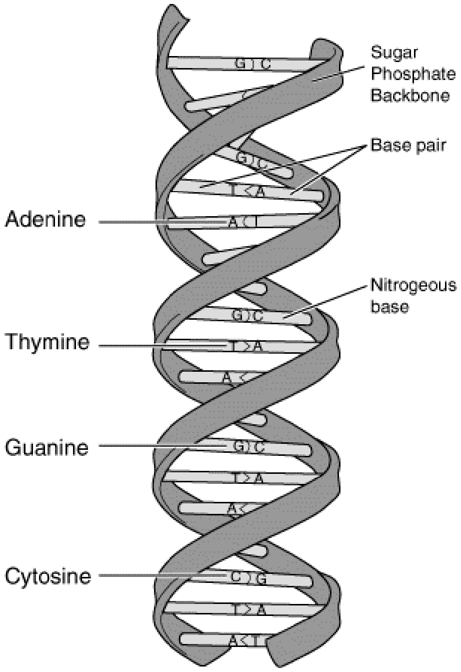


Figure 1.1: Double helix DNA structure

The double helical structure of DNA is illustrated in Figure 1.1.

As we saw earlier, genes direct the production of polypeptides, or proteins. This is a really crucial step in the cell development since protein activity directly determines the characteristic features (phenotype and functions) of cells (eg: cancer and normal cells). The process by which a gene product (RNA or polypeptide) is made is called gene expression, and involves two steps, called transcription and translation. These two steps are required to make a protein from the information carried in genes. In the transcription process, an enzyme called RNA polymerase makes a copy of one of the DNA strands, leading to the formation of messenger RNA (mRNA).

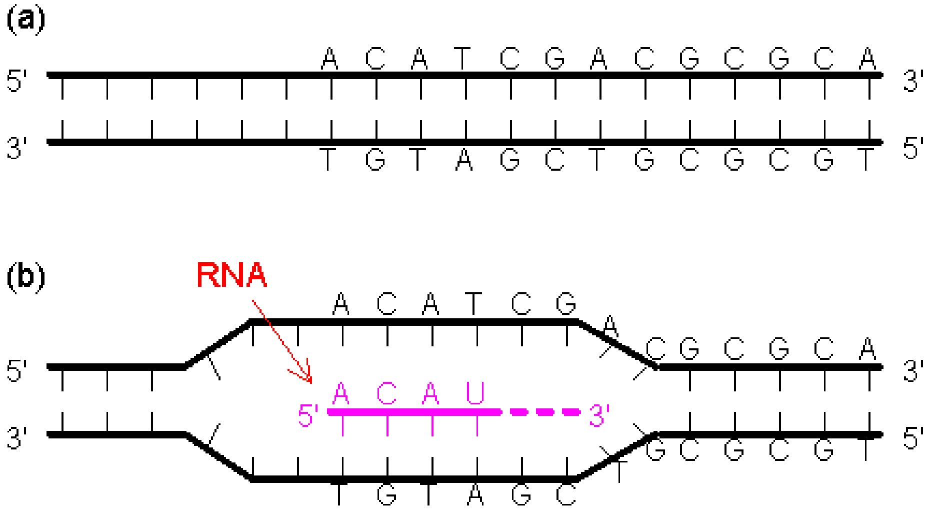


Figure 1.2: Figure (a) represents DNA before transcription. Figure (b) represents the transcription step: the DNA should unwind so that one of its strands can be used as a template to synthesize a complementary strand.

This step is illustrated in Figure 1.2. In the translation step, the mRNA carries the genetic instructions to the cell's protein factories, called ribosomes. The transcription

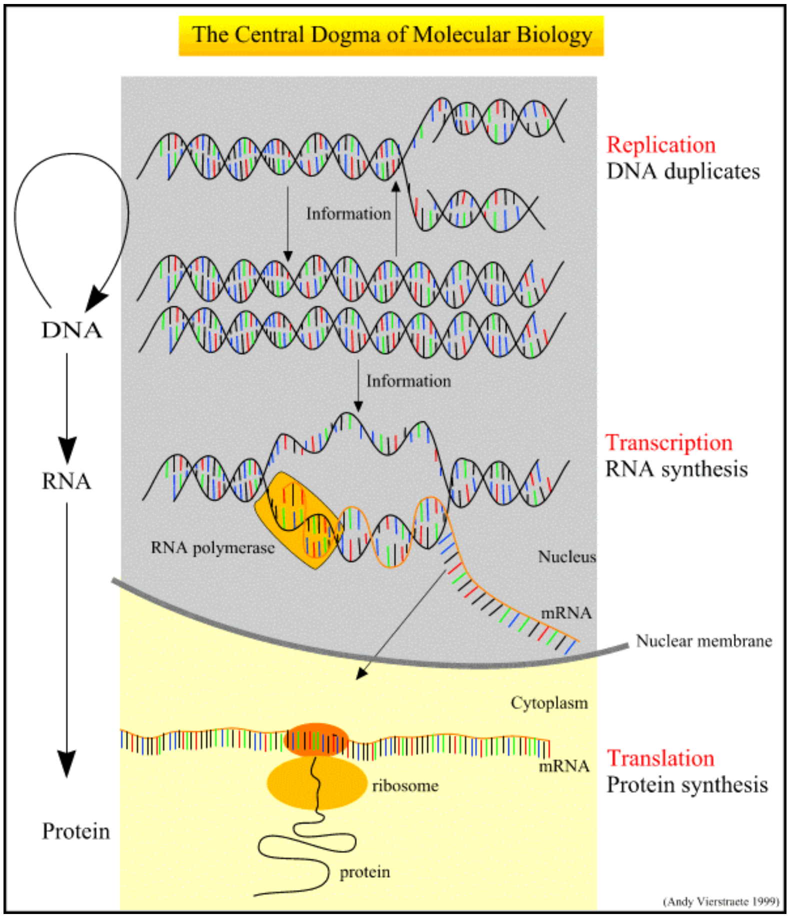


Figure 1.3: The central dogma of molecular biology

and translation steps are often referred as the central dogma of biology and they are illustrated in Figure 1.3 (picture: Andy Vierstraete, 1999).

Since most differences in a cell state (eg: cancer cell versus normal cell) are correlated with changes in mRNA levels of genes, and since the pattern of genes expressed in a cell is characteristic of its present state, it is of high interest to quantify the amount of expression of a set of expressed genes under two or more cell conditions.

1.2 Microarray technology

Gene expression can be measured during the transcription or translation process of a protein synthesis. To date, attention has focused primarily on monitoring changes at the transcriptional level, although protein arrays have recently been developed (Haab, Dunham & Brown (2001)). Although several types of microarray systems exist, we focus our attention on cDNA arrays, developed in the Brown & Botstein labs of Stanford. Using this type of array, gene expression can be followed using fluorescent cDNA probes and then measured using a scanning confocal laser microscope.

A microarray experiment can be divided into several stages, briefly described here. For a more detailed description of each of the following steps, we refer to Nguyen, Arpart, Wang & Carroll (2002).

1. Preparation of the biological samples: mRNA is extracted from cells grown under the two conditions studied.

2. cDNA synthesis: Complementary DNA (called cDNA) is prepared from each of the RNA samples. This is done using an enzyme called transcriptase which is able to synthesize a strand of DNA complementary to the mRNA sequence (A replaced by T, T by A, G is replaced by C and C by G).

3. Labeling of the cDNA samples: Fluorescently tagged nucleotides are used for the cDNA synthesis. Each cDNA sample is labeled with a different fluorescent dye (a red-fluorescent dye Cy5 for one sample and a green-fluorescent dye Cy3 for the other). Then, the two fluorescent samples are mixed together. This is called the "probe".

4. Array fabrication: after choosing the set of genes that will be investigated, the DNA sequences corresponding to the genes are printed onto a glass slide (array). Thus, each spot of the array contains a particular DNA gene sequence that, when denatured, can base-pair with a complementary cDNA, in the hybridization step.

5. Hybridization: this term refers to the binding of two complementary DNA strands by base pairing. Both the DNA on the array and the cDNA are denatured. The labeled cDNA is spread on the array and by using the complementarity between the labeled cDNA and the DNA spotted on the chip, some of the labeled cDNA binds to specific complementary sequences. As an example, consider a particular spot on the array. This spot contains the DNA sequence coding for a particular gene, say gene A. If gene A is expressed under at least one of the two conditions, cDNA coding for that particular gene will be present in the labeled cDNA solution and the cDNA sequence (labeled in green or red) for gene A will bind to the DNA sequence contained in the spot studied. When the hybridization process is over, unbound cDNA is washed off the array.

6. Construction of the raw data: The microscope slide containing the microarray is placed inside a darkened box. Inside the box, it is scanned with a green and a red laser to detect the bound labeled cDNA. Then, for cDNA arrays, the raw data consists of two images, one obtained from the Cy3 dye and one obtained from the Cy5 dye.

7. Image processing: this step allows us to obtain numerical data from the raw data obtained in the previous step. First, for each raw image, the location of each spot on the array is obtained using some image analysis techniques. Then, for each spot,

an estimate of the spot and background intensity is produced, based on the mean or median intensity values of the pixels in the spot area and in the background region. At the end of this step, we obtain background and spot intensities for each spot on the array and for each of the two dyes.

8. Background correction: most of the statistical analysis are based on the background corrected intensities. For each spot and for each dye, we subtract the background intensity from the spot intensity to obtain the final dataset. For an experiment involving m arrays and n genes (spots), we obtain two matrices: $\mathbf{R}$=(r_{ij}) and $\mathbf{G}$=(g_{ij}) corresponding to the red and green corrected intensities for gene i and array j.

1.3 Statistical issues in a microarray experiment

As we have seen, a microarray experiment involves a large number of complex procedures. The raw data are then generally influenced by systematic experimental variation, leading to noisy and biased measurements. To the extent the sources of variation are understood, they should be carefully controlled in a well designed experiment. However, in order to remove extraneous sources of variation that affect the measured gene expression levels, a normalization procedure must be applied to the data. Once the data are normalized, some statistical problems, like the discovery of a pattern of expression over several conditions, or the identification of differentially expressed genes, can be explored.

In addition to the high variability of the data, the cost of the technology allows the scientist to use only a limited number of arrays in an experiment, which greatly complicates the variance estimation. Furthermore, how to model the dependence structure between genes remains an open issue in the microarray literature. Certain types of well-designed experiments, such as microarray time-series data for example, can provide a possible mean for identification of transcriptional regulation relationship among genes. However, identifying correlation structure between genes is a difficult task to achieve in most of the datasets.

1.3.1 Variation and design of experiment

The exact factors contributing to the high variability of gene expression data are not completely understood yet. They can come from the hybridization conditions, the preparation of the mRNA samples, the type of arrays used or the labeling procedure. Kerr & Churchill (2001a), Kerr, Churchill & Martin (2001b), Kerr, Afshari, Bennett, Bushel, Martinez, Walker & Churchill (2002) studied in a series of three papers the potential sources of variation that can lead to misleading and biased results. To the extent these sources are known, they should be incorporated in the statistical model. Several factors were identified such as the array, dye, treatment and gene effects. The authors also pointed out a significant dye/gene interaction that seems to be present in most of the datasets studied: it was observed in multiple datasets that some genes exhibit higher expression when they are labeled with one particular dye, compared to the other one. This phenomenon was identified regardless of the sample. If such an effect occurs, the estimates of relative expression will be biased, leading to a wrong interpretation of the results. The contribution of Kerr et al to microarray data analysis was really significant, since they emphasized the importance of the design of the experiment in this particular context. They also proposed several designs, like the loop design when several factors of interest are studied, as well as ANOVA models that enable controlling for the systematic sources of variation. The most significant aspect of the proposed designs is the dye-reversal strategy: every experiment (array) should be replicated with the Cy5 and Cy3 labeling reversed. This strategy is proposed in order to control the possibly high gene-dye interaction effect and is commonly called a dye-swap experiment. Furthermore, it allows the design to be balanced with respect to the dyes.

The second major tool allowing a better understanding of the basic variability of gene expression data is replication. Replication, in a microarray experiment, is divided into three classes: the subsampling strategy, the array replication and the biological replication. Subsampling refers to the replication of the spots (genes) on the array. This practice is commonly used since every gene can easily be spotted at least two times on the array without any supplementary cost. The array replication refers to the replication of the hybridization step, on another array, using the same RNA source. This type of replication allows the estimation of the between-arrays variation. Finally, the biological replication allows the estimation of the biological variation since this type of replication samples multiple individuals, through different RNA sources.

1.3.2 Normalization for cDNA microarray data

When the different sources of variation cannot be controlled or estimated through a well designed experiment, a normalization procedure is required, in order to remove systematic effects that may bias the results. Normalization of gene expression data is a very challenging issue and, although more research is needed, more and more sophisticated methods are available. However, for any method, the set of genes that should be used for the normalization must be determined. In the case where a small proportion of genes are expected to vary in expression between the two mRNA samples, all the genes on the array should be used (Yang, Dudoit, Luu, Lin, Peng, Ngai & Speed (2002)). Another approach is to use a set of housekeeping genes for the normalization. These genes are expected to be expressed in a constant manner over the two mRNA samples and can provide a baseline for normalization. The third possibility involves the use of spiked control sequences. These DNA sequences (in general not genes) are chosen such that we expect to have equal red and green intensities across the range of intensities. Usually, these DNA sequences come from an organism different from the one studied. Unfortunately, this method is very challenging, from a technical point of view, and is not commonly used in practice.

Depending on the experimental set-up, several methods of normalization can be used. Yang et al. (2002) proposed three types of normalization procedure: the within-slide normalization, the paired-slides normalization (for dye-swap experiments) and the multiple-slides normalization. In the first case, the normalization is done separately for each slide (array). Several approaches can be used, but the most popular one is the global normalization, which assumes that the green and red intensities are related by a constant factor k. Several methods of estimating the constant k are available (Ideker, Thorsson, Siegel & Hood (2000)). The second type of normalization applies to the dye-swap experiments and is similar in spirit to the global normalization. Finally, the last type of normalization adjusts for multiple array experiments.

We note that most of the normalization methods are applied to the ratio Red/Green of the expression measurements. Also, many statistical methods are based on the logarithm of this ratio. Although the logarithm scale has the advantage of reducing the variability of the data, we believe that some useful information is lost by using such a statistic. For example, a ratio of 2 can be obtained from very low or very high expressions under both treatments. Intrinsic differences exist in the behavior of the gene's expression between these two types of situations. Modeling the expression's intensity allows one to consider such differences, which the ratio does not. In the next chapters, all the methods developed will be based on the raw background corrected intensity measurements. We note however the following point: if it is implicit in the use of microarrays that the measured intensity is a measure of the abundance of the mRNA in the sample, there may be some situations where it is not the case. For instance, it is possible that some genes are sequence-related, like those that codify for a family of proteins, for example. In such a case, a microarray spot could show mixed mRNA populations (it is also known as cross-hybridization). This result is not specific, because the intensity is not related with the abundance of a specific mRNA. As we mentioned, the labeling reaction could also present some bias: some mRNAs are more efficiently labeled than others with a particular dye. It follows that low intensity spots should not necessarily be treated differently.

1.3.3 Some statistical problems

Depending on the set-up of the experiment, a wide range of interesting statistical problems are associated with microarray data. As we have seen earlier, normalization procedures, the planning of the experiment (design, sample size) as well as the analysis of factorial experiments are of high interest.

Tumor classification (Dudoit, Fridlyand & Speed (2002)) is an example of a topic that can be investigated through microarrays: new/unknown tumor classes can be identified using some statistical methods like cluster analysis or other unsupervised learning methods. Discriminant analysis or other supervised learning methods can also be used to classify the tumors into known classes. Finally, some variable selection methods can provide a way to identify a “super gene” (or marker gene) that could characterize the different tumor classes. In a typical tumor classification problem, the experiment usually involves a large number of arrays, each of them comparing a specific tumor cell to a normal cell (Control). For such a problem, the very large number of variables (genes) relative to the number of observations (tumors) provides an interesting statistical challenge.

Another task is to find a set of genes that behave similarly in various conditions. In such a case, some unsupervised learning methods, like cluster analysis applied on the genes, have proven efficiency and many new cluster algorithms are now available.

Finally, in the next chapters, our attention will focus on a question that brings several interesting statistical problems: the discovery of differentially expressed genes. Basically, we consider replicated experiments comparing two cell types: a treated one and a control one. We are interested in finding a set of genes that are expressed significantly different in one cell type as compared to the other one. If the expression level of the treated cell is higher than that of the control one, for a specific gene, we say that the gene is over-expressed. On the other hand, if the expression level of the treated cell is lower than that of the control one, we say that the gene is under-expressed.

The first problem related to the open question: “which genes are up/down”, is the multiple testing issue (*cf* Dudoit, Shaffer & Boldrick (2003)). Since we want to make inference at a gene level, we need to test the null hypothesis that the mean expression level is equal for the treated and control cell for gene i, for each gene on the array. In practice, thousands of genes are studied simultaneously and then, thousands of statistical tests are performed. Since the probability that at least one Type I error is committed increases a lot with the number of hypotheses, numerous methods have been proposed to control the Family Wise Type I Rate (FWER), by adjusting the p-values of the tests. However, even if these methods allow several hypotheses to be tested simultaneously, they hardly handle a very high number of tests. Some approaches that minimize the False Discovery Rate (FDR) have recently been proposed. Instead of fixing the error rate to estimate the rejection region, these methods propose the opposite approach: fixing the rejection region to estimate the corresponding error rate. The optimal region is then chosen such that the error rate is controlled at a specified level. These types of methods have proven efficiency, increase the power of the tests and can be easily applied in a context where the number of hypotheses tested is very high.

The second statistical issue with the type of problem studied is the distribution assumption. Most of the work published so far assumes a log-normal distribution or a gamma distribution (*cf* Ibrahim, Chen & Gray (2002), Newton, Kendziorski, Richmond, Blattner & Tsui (2001) and Kerr et al. (2002)) of the expression levels. With the high variability of the data, there is need for models that allow a greater flexibility and where the parametric assumptions are lightened. A large number of models have been suggested to identify differentially expressed genes. Some methods propose a frequentist approach (Kerr et al. (2002), while some others use a Bayesian framework (Ibrahim et al. (2002), Newton et al. (2001)) or a mixture of models (Kendziorski, Newton, Lan & Gould (2002)). We can also find some parametric approaches (Kerr et al. (2002)) and others that are non-parametric (Efron, Storey & Tibshirani (2001)) or semi-parametric (Newton, Noueiry, Sarkar & Ahlquist (2003)). We believe that the frequentist approaches are too restrictive and that a Bayesian framework provides a more flexible alternative. The use of a mixture of distributions for the two hypotheses tested can also allow us to obtain a good approximation of the expected false discovery rate. Our goal for the next chapters is to construct a test statistic, as robust as possible to the variability contained in the data, based on an underlying Bayesian informative model, which is as flexible as possible. A hypothesis test will be constructed, from the test statistics, minimizing a criterion related to the False Discovery Rate in a Bayesian context.

Chapter 2

Multiple testing using posterior probabilities of half spaces

The main goal of this chapter is the construction of a formal hypothesis test to identify differentially expressed genes under a Bayesian framework. A null hypothesis of no difference in the gene expression under two conditions is constructed. Since a test is performed for each gene, it follows that thousands of tests are performed simultaneously, and multiple testing issues then arise. The aim of our research is to make a connection between Bayesian analysis and frequentist theory in the context of multiple comparisons by deriving some properties shared by both p-values and posterior probabilities. Following the work of Dudley & Haughton (2002) regarding the asymptotic normality of posterior probabilities of half- spaces, we show that the posterior probability of the one-sided alternative hypothesis, defined as a half-space in our case, shares with the frequentist p-value the property of uniformity under the null hypothesis. This result holds asymptotically, when the number of observations available for each test is large enough. Uniformity under the null hypothesis (as p-values are assumed to be) is the main property used in most of the multiple testing procedures developed recently controlling for the False Discovery Rate and the False Negative Rate. We are then able to use this posterior probability as an input to these procedures, in the same spirit as a p-value.

The first section of this chapter reviews the issues brought by multiple tests. In the second section, we present the Bayesian framework under which multiple comparisons are performed. The uniform property of the posterior probability studied is presented in the third section, as well as a review of the work done by Dudoit et al. (2002). Finally, the procedure allowing the integration of such a Bayesian quantity into the methods developed recently is presented in the last section.

2.1 The multiple testing issue

Consider the problem of testing simultaneously n null hypotheses H_{0i} $(i = 1, \ldots, n)$. For each test i, a statistic is constructed from which a p-value, P_i, is derived. We consider here the test procedure that rejects H_{0i} if $P_i \leq t$, for all $i = 1, \ldots, n$ and a specified $t \in [0, 1]$. The various outcomes for the n tests can be summarized in Table 2.1, where the numbers n_0 and n_1 of true and false null hypotheses are unknown parameters. The quantities R(t), S(t), T(t), U(t) and V(t) are all random empirical processes, for a t-dependent rejection region, from which only R can be observed. The variable V is often referred to as the

	# not rejected	# rejected	
# true null hypotheses	U(t)	V(t)	n_0
# non true null hypotheses	T(t)	S(t)	n_1
	n-R(t)	R(t)	n

Table 2.1: Outcomes when testing n hypotheses for a t-dependent test procedure

number of false positives, or Type I errors, whereas the variable T represents the number of false negatives, or Type II errors. The standard methods seek tests that minimize the Type II error rate (or maximize the power) within the class of tests where the Type I error rate is fixed at a reasonable level δ. The main concern raised by the multiple number of tests is that the probability of at least one Type I error (also called Family-Wise Error Rate or $FWER$) increases dramatically with the number of hypotheses tested. As a result, there is the need for a redefinition of the Type I error rate, in the case of multiple tests, allowing a global control over the n tests of the proportion of errors.

Typical solutions to the problem described can be divided into two approaches. Frequentist approaches usually compute a p-value for each test, that can be integrated into three main types of multiple testing framework: the single-step procedures, the step-down and step-up procedures (see Dudoit et al. (2003)). In single-step procedures, the rejection region of each test is constant and does not depend on the results of tests of other hypotheses. A well known example of such procedure is the Bonferroni procedure. In step-down procedures, p-values are ordered from the most significant to the least significant, and corresponding hypotheses are considered successively. When one fails to reject a null hypothesis, no further hypotheses are rejected. An example of such procedure is the Benjamini-Hochberg procedure (see Benjamini & Hochberg (1995)). Finally, step-up procedures work in the opposite direction as the step-down procedures, and p-values are sorted from the least significant to the most significant one.

A review of multiple testing procedures in a frequentist setting, as well as the different error rates developed can be found in Dudoit, van der Laan & Pollard (2004). It is common practice now to seek procedures providing a control for the False Discovery Rate (FDR): the expected proportion of Type I errors among rejected null hypotheses, and the

False Negative Rate (FNR): the expected proportion of Type II errors among rejected alternative hypotheses. Using the notation defined in Table 2.1, the False Discovery rate is defined as an empirical process (rejecting all null hypotheses with p-values $P_i \leq t$) such that

$$FDR(t) = E\left[\frac{V(t)}{R(t) \vee 1}\right] = E\left[\frac{V(t)}{R(t)}|R(t) > 0\right] P(R(t) > 0), \tag{2.1}$$

where $R(t) \vee 1$ =max$(R(t), 1)$. A recent and very powerful approach, proposed by Storey (2002), is to fix the rejection region of the tests (ie, fix the threshold t) in order to provide an estimate of the FDR, whose expectation was shown to be greater than or equal to the true FDR. Using Bayes' theorem, a conservative point estimate of $FDR(t)$ can be derived such that

$$F\hat{D}R_\lambda(t) = \frac{\hat{p_0}(\lambda)t}{[R(t) \vee 1]/n} \tag{2.2}$$

where $\hat{p_0}$ is an estimate of $p_0 \equiv n_0/n$, the proportion of true null hypotheses. This estimate depends on a tuning parameter λ such that

$$\hat{p_0}(\lambda) = \frac{n - R(\lambda)}{(1-\lambda)n}. \tag{2.3}$$

Storey (2002) showed how to pick λ optimally in order to minimize the mean square error of the estimates. We also note here that the estimated FDR presented can be adjusted for the finite sample case. The FDR-controlling procedure proposed by Storey, Taylor & Siegmund (2004) can be summarized as follows:

1 : Let δ be the pre-chosen level at which to control the FDR

2 : For any significance region $[0, t]$, estimate $FDR(t)$ by $F\hat{D}R_{\lambda}(t)$ given in (2.2)

3 : Let $t^* = sup\{0 \leq t \leq 1 : F\hat{D}R_{\lambda}(t) \leq \delta\}$ and reject all null hypotheses corresponding to $P_i \leq t^*$.

It can be shown that, under convergence assumptions that are easily verified when n is large enough, the limsup of $FDR(t^*)$ can be controlled at a level δ. Several other asymptotic results have been proved and we refer to Storey et al. (2004), Genovese & Wasserman (2002*a*) and Genovese & Wasserman (2002*b*) for more details. The well-known Benjamini-Hochberg procedure (*cf* Benjamini & Hochberg (1995)) appears to be a particular case of the procedure defined, with $\lambda = 0$. For $\lambda \neq 0$, the procedure defined is equivalent to the Benjamini-Hochberg procedure with n replaced by $\hat{p_0}n$. Storey (2003) also developed a procedure controlling for the pFDR, referred as the positive FDR (conditioning on the fact that we reject at least one hypothesis) and developed a Bayesian version of the p-value, referred as the q-value.
The major issue regarding these procedures deals with the estimation of the distribution of the test statistic under the null hypothesis (in the case of a p-value, assumed uniform on the interval $[0, 1]$). In some cases, this distribution may be based on some asymptotic properties that may not be verified when the number of observations per test is small. Some methods to estimate the distribution of these statistics under the null hypothesis have been developed, based on resampling strategies, and we refer for example to Van Der Laan, Dudoit & Pollard (2004). However, these methods have in general the main disadvantage of being computationally intensive.

The second type of approach commonly used in practice is defined in a context of Bayesian, empirical Bayes or random effects models, where a second level of randomness is assumed on the parameter of interest. Consider for example the situation where n subjects are studied, with m observations available for each subject. A hypothesis test is then performed independently for each subject, based on the set of m observations. Empirical Bayes methods have proven to be very efficient, particularly in the situation where the number of observations available for each subject is small or when the number of subjects is very large. Such methods in the context of microarray experiments (where the subjects represent the genes) have been used and described in Efron et al. (2001), Efron (2003), Ibrahim et al. (2002), Newton et al. (2003) and Smyth (2004) for example. The main advantage of this type of approach is that it allows a subject-specific inference, through the use of posterior probabilities, without the need of estimating a set of parameters for each subject. The parameters of the model are typically estimated using the data for all the subjects (in an empirical Bayes manner), allowing for a certain sort of dependence between subjects. Furthermore, the use of a mixture of distributions under the null and alternative hypothesis respectively accounts for a within-subject dependence. Recent work, accounting for multiple testing issues, can be found in Muller, Parmigiani, Robert & Rousseau (2004), who developed a test procedure controlling for the posterior expected FDR and FNR. In this context, a typical Bayesian measure of significance for each test is the posterior probability of the alternative hypothesis, closely related to the well known odds ratio. This probability can usually be computed easily, using the predictive distribution of the data under the null and alternative hypothesis. This type of approach has two main drawbacks. First, non-informative models cannot be considered, since the predictive distribution is typically improper in such a context (however, note that the use of non-informative prior distributions leads in general to proper posterior distributions). The second disadvantage is that one-sided alternative hypotheses cannot be considered since it is not possible to distinguish one side or the other when the measure of significance is above a threshold. Some other Bayesian approaches have been developed, where posterior probabilities are adjusted to account for multiple hypotheses (see Westfall, Johnson & Utts (1997)). This can be accomplished by calibrating the prior distribution for the joint parameter space to provide strong control of the family-wise error rate (FWER). However, this approach has been argued to be overly conservative. Other methods have been proposed, which provide a weak control of the FWER (see Shaffer (1999)) or of the probability of a sign error (see Gelman & Tuerlinckx (2000)) but it remains an open question how multiplicity adjustments should be performed under a Bayesian framework.

2.2 The Bayesian framework

As we mentioned, Bayesian or random effects models have been extensively used in the literature to model gene expression data. These types of models are found to be very good candidates to capture complex patterns of variation in the data, as well as the relationship existing between their first two moments. Also, even with a small number of observations per gene, the hierarchical framework allows a gene-specific inference, while keeping the number of parameters reasonable.

2.2.1 An overview of Bayesian models

Let $\underset{\sim}{X}$ be a random sample in $\mathcal{R}^n$, with probability density function $f(\underset{\sim}{x}|\theta)$, where $\theta \in \Theta \subset \mathcal{R}^d$ represents the parameter of the distribution. Assuming a distribution with density π on Θ enables us to incorporate the prior knowledge we have on the behavior of the data. Bayesian analysis combines this prior information with the information brought by the sample into the posterior distribution $\pi(\theta|\underset{\sim}{x})$. Therefore, the posterior distribution allows us to update our beliefs about θ after observing the sampled data. The posterior distribution is a direct consequence of Bayes' theorem and has density

$$\pi(\theta|\underset{\sim}{x}) = \frac{f(\underset{\sim}{x}|\theta)\pi(\theta)}{m(\underset{\sim}{x})}, \tag{2.4}$$

where $m(\underset{\sim}{x}) = \int_{\Theta} f(\mathbf{x}|\theta)\pi(\theta)d\nu(\theta)$ is the predictive (unconditional) density function of $\underset{\sim}{X}$ and $f(\underset{\sim}{x}|\theta)$ is the likelihood of the data. In this case, note that ν represents a sigma-finite measure on Θ's space, not involving hyperparameters. In Bayesian analysis, the choice of the prior is an important step of the inference process. It can be determined in a subjective way, using approaches such as the histogram approach, the relative likelihood approach or the CDF determination (Berger (1980)). In situations where little or no *a priori* information is available, a *non-informative* prior is generally used. This type of prior contains no information about the parameter θ and in practice, the choice of an improper distribution (one having an infinite mass) is not unusual. However, improper priors can lead to proper posterior distributions, as we will see in the next section, and inference about the parameter of interest may not be affected.

One of the main disadvantages of Bayesian models, compared to frequentist ones, is the computational complexity of the posterior distribution. The use of conjugate priors requires little computational effort, and their simple parametrization is often a nice alternative to more complex models. In such a situation, the prior-likelihood pair forms a conjugate family, where the prior and posterior distribution have the same form.

2.2.2 Definition of the model

Consider a replicated cDNA microarray experiment involving two treatments (T1 and T2). Let X_{ij} be the intensity (expression) recorded for gene i on array (replication) j under T1 and Y_{ij} be the intensity for gene i on array j under T2, where $i = 1, \ldots, n$ and $j = 1, \ldots, m$. We consider here that the data have been normalized between and within arrays such that systematic effects have been removed. We assume, for now, that all the variables are independent. We denote by $\underset{\sim}{X}_i$ the random sample $(X_{i1}, \ldots, X_{im})$ corresponding to gene i, and by $\underset{\sim}{X}$ the random sample $(\underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_n)$. Similarly, we denote by $\underset{\sim}{Y}_i$ the expression of gene i under T2 and let $\underset{\sim}{Y}$ represent the complete set of expression measurements under the second treatment. In the following, we suppose the m replications in $\underset{\sim}{X}_i$ and $\underset{\sim}{Y}_i$ of a given gene to be identically distributed, with mean θ_{xi} and θ_{yi} respectively. First, consider the two sets of hypothesis, for $i = 1, \ldots, n$:

$$\begin{array}{lcl} H_{0i} & : & \text{Gene } i \text{ is equally expressed under } T_1 \text{ and } T_2, \\ H_{1i} & : & \text{Gene } i \text{ is differentially expressed under } T_1 \text{ and } T_2, \end{array}$$

and define for each gene the variable H_i such that $H_i = 1$ if H_{1i} is true, and 0 otherwise. These variables are assumed to be distributed according to the Bernoulli distribution with parameter p. In other words, we assume that for each gene independently, H_{1i} occurs with probability p and H_{0i} occurs with probability $(1 - p)$. We also assume a specific distribution on the intensities recorded for each gene i and for each treatment, by defining

$$\begin{aligned} X_{ij}|\theta_{xi} &\sim f_x(.|\theta_{xi}), \\ Y_{ij}|\theta_{yi} &\sim f_y(.|\theta_{yi}), \end{aligned}$$

where $\underset{\sim}{X}$ and $\underset{\sim}{Y}$ are independent, conditionally on their mean θ. Furthermore, we consider the addition of a mixture of prior distributions on $(\theta_{xi}, \theta_{yi})$ to this model, such that, conditionally on the indicator variables H_i's,

$$\begin{aligned}
\theta_{xi} &\sim \pi(.), \\
\theta_{yi}|\theta_{xi}; H_i = 1 &\sim \pi(.), \\
\theta_{yi}|\theta_{xi}; H_i = 0 &\sim \delta_{\theta_{xi}}(\theta_{yi}),
\end{aligned}$$

where $\delta()$ is a Dirac δ-function. Under the alternative hypothesis that the gene is differentially expressed under the two treatments, the marginal means are considered to be independent of each other. Under the null hypothesis, they are assumed to be identical. This type of model is equivalent to the following one :

$$\text{With probability (1-p),} \begin{cases} X_{ij}|\theta_i \sim f_x(.|\theta_i), \\ Y_{ij}|\theta_i \sim f_y(.|\theta_i), \\ \theta_i \sim \pi_\theta(.). \end{cases}$$

$$\text{With probability p,} \begin{cases} X_{ij}|\theta_{xi} \sim f_x(.|\theta_{xi}), \\ Y_{ij}|\theta_{yi} \sim f_y(.|\theta_{yi}), \\ \theta_{xi} \sim \pi_\theta(.), \\ \theta_{yi} \sim \pi_\theta(.), \end{cases}$$

where θ_{xi} and θ_{yi} are assumed to be independent. We note that the dependence of the gene expression between the two treatments is addressed by assuming a mixture of models under the null and alternative hypothesis. However, the independence assumption between genes may not be completely realistic in practice. This concern is partially addressed by the fact that empirical Bayes techniques are used to estimate the hyperparameters of the model. In this case, these parameters are estimated using the whole set of data. However, how dependence between genes should be modeled remains an open question in the literature.

Under such a model, we can rewrite the hypothesis H_{0i} and H_{1i} as

$$\begin{aligned} H_{0i} &: \quad \theta_{xi} = \theta_{yi}, \\ H_{1i} &: \quad \theta_{xi} \neq \theta_{yi}. \end{aligned}$$

The predictive density function of $\underset{\sim}{X_i}$ and $\underset{\sim}{Y_i}$ is then a mixture of distributions such that $m(\underset{\sim}{x_i}, \underset{\sim}{y_i}) = (1-p)\ m_0(\underset{\sim}{x_i}, \underset{\sim}{y_i}) + p\ m_1(\underset{\sim}{x_i}, \underset{\sim}{y_i})$, where

$$\begin{aligned} m_0(\underset{\sim}{x_i}, \underset{\sim}{y_i}) &= \int_\theta f_x(\underset{\sim}{x_i}|\theta) f_y(\underset{\sim}{y_i}|\theta)\pi_\theta(\theta)d\theta, \\ m_1(\underset{\sim}{x_i}, \underset{\sim}{y_i}) &= \int_\theta f_x(\underset{\sim}{x_i}|\theta)\pi_\theta(\theta)d\theta \times \int_\theta f_y(\underset{\sim}{y_i}|\theta)\pi_\theta(\theta)d\theta, \\ f_x(\underset{\sim}{x_i}|\theta) &= \prod_{j=1}^{m} f_x(\underset{\sim}{x_{ij}}|\theta), \\ f_y(\underset{\sim}{y_i}|\theta) &= \prod_{j=1}^{m} f_y(\underset{\sim}{y_{ij}}|\theta). \end{aligned} \tag{2.5}$$

Finally, we can write the predictive density of the samples $\underset{\sim}{X}$ and $\underset{\sim}{Y}$ as $m(\underset{\sim}{x}, \underset{\sim}{y}) = \prod_{i=1}^{n} m(\underset{\sim}{x_i}, \underset{\sim}{y_i})$.

2.3 Uniform property of the posterior probability of half-space, under H_0

We present here an interesting result regarding the posterior probability of half-spaces, obtained by Dudley & Haughton (2002). Note that this result is first presented in its general form, and we do not consider the Bayesian framework defined in the previous section yet. Connection with this model will be made in the next section. Consider Θ as an open subset of an Euclidean space $\mathbf{R}^d$. A half-space of Θ is defined as

$$A = \{\theta \in \Theta : \theta.\upsilon_A \geq M\}, \tag{2.6}$$

where $\upsilon_A \in \mathbf{R}^d$, $|\upsilon_A| = 1$, $\theta.\upsilon_A$ represents the dot product of θ and υ_A and $M \in \mathbf{R}$. The boundary hyperplane of A is defined as

$$\partial A = \{\theta \in \Theta : \theta.\upsilon_A = M\}. \tag{2.7}$$

Now, to introduce the next theorem, we consider the set $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ of family of laws dominated by a σ-finite measure μ on a sample space $(\mathcal{X}, \mathcal{B})$. Let $f(x,\theta)$ be the density $(dP_\theta/d\mu)(x)$, $x \in \mathcal{X}$, $\theta \in \Theta$. We take $0 \leq f(x,\theta) \leq \infty$. We consider a set of *iid* random variables, $X_1, X_2, \ldots$ with values in $\mathcal{X}$ and with some law $P \in \mathcal{P}$. We denote $\underset{\sim}{x_m} = (x_1, \ldots, x_m)$ and $\underset{\sim}{X_m} = (X_1, \ldots, X_m)$. Finally, we let $\hat{\theta}_m(\underset{\sim}{x}) = \hat{\theta}_m(\underset{\sim}{x_m})$ be the maximum likelihood estimate of $\theta \in \Theta$.

Let us first consider a hyperplane ∂A of the form (2.7) and consider A to be one of its two possible half-spaces.

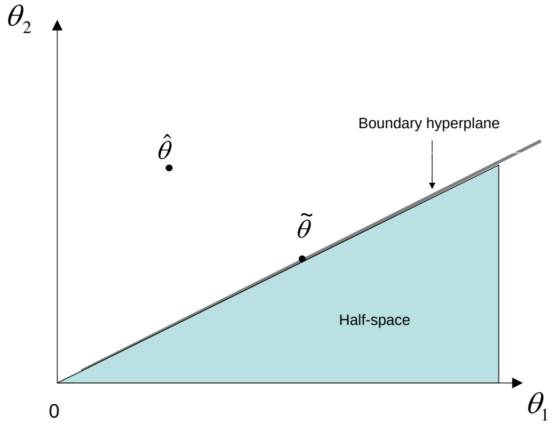


Figure 2.1: Illustration of the hyperplane and half-space in the case $\Theta = \mathbb{R}^{2+}$

An illustration of this setting, when $\Theta = \mathbb{R}^{2+}$, is provided in Figure 2.1. We also denote by $\tilde{\theta}_m(\underset{\sim}{x_m}) = \tilde{\theta}_m(\underset{\sim}{x})$ the maximum likelihood estimate of θ in ∂A. If we consider the null hypothesis $H_0 : \theta \in \partial A$, the likelihood ratio statistic is then defined as

$$\Delta_m = \Delta_m(\underset{\sim}{x}) = 2\left[\sum_{j=1}^{m} log(f(x_j, \hat{\theta}_m)) - \sum_{j=1}^{m} log(f(x_j, \tilde{\theta}_m))\right].$$

We consider also that a prior probability is given on Θ with continuous density $\pi_0(\theta) > 0$ for all θ and $\int_\Theta \pi_0(\theta)d\theta = 1$. The posterior distribution of θ on Θ is then defined as $\pi_{x,m}$. Under such a framework, the following theorem can be stated

Theorem 2.1. *Under some regularity conditions, for all $\epsilon > 0$,*

$$\frac{\Phi(Q_m)}{1+\epsilon} \leq \pi_{x,m}(A) \leq (1+\epsilon)\Phi(Q_m), \textit{ for } m \textit{ large enough, almost surely,}$$

where Φ is the standard normal distribution function and where $\pi_{x,m}(A) = P(\theta \in A|x_1, \ldots, x_m)$.

The variable Q_m *is defined as*

$$-Q_m = \begin{cases} \sqrt{\Delta_m} \textit{ if } \hat{\theta}_m \notin A, \\ -\sqrt{\Delta_m} \textit{ if } \hat{\theta}_m \in A. \end{cases}$$

It follows that if $\theta \in \partial A$,

$$\pi_{x,m}(A) \textit{ converges in distribution to } U, \textit{ as } m \to \infty,$$

where $U \sim \mathcal{U}(0,1)$.

The regularity conditions of this theorem are stated in the following:

(A_1) Define $LL_m(\theta)$ as being the log likelihood function. There is a $\theta_0 \in \Theta$, called the pseudo-true value of θ, such that for every neighborhood $\mathcal{N}$ of θ_0, there is a $\kappa > 0$ such that almost surely for m large enough, $\sup_{\theta \notin \mathcal{N}} LL_m(\theta) < \sup_{\theta \in \mathcal{N}} LL_m(\theta) - m\kappa$.

(A_2) For θ in a small enough neighborhood W of θ_0, the function $f(x, .)$ is strictly positive and C^2 in θ for P-almost all x, and the P-Fisher information matrix $E(\theta) := \{E_{ij}(\theta)\}_{i,j=1}^{d} := \{E_P(-\partial^2 \text{log} f(., \theta)/\partial\theta_i\theta_j)\}_{i,j=1}^{d}$ exists and is finite, strictly positive definite and continuous in θ.

(A_3) Let $P_m := (\delta_{X_1} + \ldots + \delta_{X_m})/m$, with $\delta_x(A) := \mathbb{1}_A(x)$. For some neighborhood W of θ_0, the class $\mathcal{F}_W := \{-\partial^2 \text{log} f(., \theta)/\partial\theta_i\theta_j\}$ of functions for $\theta \in W$, $i, j = 1, \ldots, d$, is a Glivenko-Cantelli class for P; this implies that $\sup_{g \in \mathcal{F}_W} |\int g d(P_m - P)| \to 0$ almost surely as $m \to \infty$.

Note that condition A_1 can be found from sufficient conditions for consistency of approximate MLE's. Furthermore, conditions A_2 and A_3 hold for exponential families. These conditions ensure the existence of $\hat{\theta}$, for m large enough.

The proof of the first part of the theorem is given in Dudley & Haughton (2002). The second part of the theorem is a direct corollary and the proof is straightforward. The theorem is based on the fact that if $\theta \in \partial A$, the likelihood ratio statistic, $\Delta_m = \Delta(\underset{\sim}{X}_m)$, is distributed asymptotically according to a Chi-square distribution with 1 degree of freedom (since the dimension of the boundary hyperplane is equal to $d-1$). This implies that $\sqrt{\Delta_m}$ is distributed as $|Z|$, where Z has a standard normal distribution. Adding a nearly random sign to Δ_m, through the variable Q_m, brings us to normality, by symmetry of the normal distribution. It follows that the random variable $\Phi(Q_m)$ is asymptotically uniform on the interval $[0,1]$. For the corollary, by letting ϵ tend to 0, we obtain that the posterior probability $\pi_{x,m}(A)$ converges in distribution to a Uniform variable, on the interval $[0,1]$. Note that the variable Q_m can be viewed as a signed-root likelihood ratio statistic, defined in the case of testing the boundary of a half space.

2.4 Using the posterior probability of the one-sided alternative hypothesis in a multiple testing framework

The aim of this section is to link the results presented in the last section with our problem of interest. We are presenting here a test statistic, in the Bayesian framework presented in Section 2.2, from which we can derive, or estimate, the distribution under the null hypothesis. Doing so, it is then possible to integrate such a statistic into the frequentist multiple testing procedures described earlier, based usually on frequentist p-values or statistics.

Considering a set of null hypotheses $H_{0i} : \theta_{xi} = \theta_{yi}$ for each gene i, we propose to use the posterior probability of the one-sided alternative as a measure of significance for each test, defined as

$$p_i = P\left(\theta_{xi} > \theta_{yi} \mid \underset{\sim}{x_i}, \underset{\sim}{y_i}, H_i = 1\right), \tag{2.8}$$

where H_i has been defined in Section 2.2.2. Making the link with Theorem 2.1, we consider for each gene i the set of *iid* variables $Z_{i1}, \ldots, Z_{im}$, such that $Z_{ij} = (X_{ij}, Y_{ij})$ for $j = 1, \ldots, m$. Gene expression being a positive measure, the spaces are defined as $\mathbf{R}^d = \Theta = \mathbb{R}^{2+}$, the set of positive real numbers on a space of dimension 2. Note that during the background correction step, genes with negative expression are usually flagged and removed from the analysis, or set to a very low expression value. Without loss of generality, we take $A_i = \{(\theta_{xi}, \theta_{yi}) \in \Theta : \theta_{xi} \geq \theta_{yi}\}$, with boundary hyperplane $\partial A_i = \{(\theta_{xi}, \theta_{yi}) \in \Theta : \theta_{xi} = \theta_{yi}\}$. In this case, the half-space A_i represents the space of the one-sided alternative hypothesis for gene i, whereas the boundary hyperplane ∂A_i represents the space covered by the null hypothesis. Applying Theorem 2.1, we then obtain the following Corollary:

Corollary 2.1. *Under the regularity conditions of Theorem 2.1 and under the null hypothesis (given $H_i = 0$),*

$$p_i \text{ converges in distribution to } U, \text{ as } m \to \infty,$$

where $U \sim \mathcal{U}(0, 1)$.

Defining the probability p_i given $H_i = 1$, whereas the result of the corollary is given conditionally on $H_i = 0$ may seem a little bit confusing at first. To make our idea clear, we note that the model being used here is the model defined under $H_i = 1$, where it may happen that θ_{xi} and θ_{yi} coincide, according to the framework under which Theorem 2.1 has been stated. Conditioning on $H_i = 1$ in the definition of the probability p_i emphasizes the fact that under the conditions of Theorem 2.1, we do not have a model which has an atom at the boundary. However, the result of the Corollary can also be applied to the model in which we do have an atom at the boundary. Since we are looking at the distribution under $H_i = 0$, the proportion $(1-p) = P(H_{0i})$ is not of interest in this case. However, this quantity will be used in the multiple testing procedure, as we will see next.

Two main points motivated our work. First, models using p-values are restrictive in many aspects, but offer in the other hand the advantage of providing a control of frequentist error rates, ie. that do not depend on the data. Inversely, Bayesian theory brings an attractive flexibility to the models, but the control of posterior error rates can be criticized (see the Discussion provided in Chapter 8). Motivated by the result obtained in Corollary 2.1, we can use the probability p_i in the same spirit as a p-value, in the FDR-control procedure described in Section 2.1. For example, if one wants to use the Benjamini-Hochberg procedure, the following steps are required:

1. Compute the probability p_i for each gene i. This probability can be easily computed numerically, using procedures such as Monte Carlo, as long as the posterior joint density function of the parameters θ_{xi} and θ_{yi} is known under the alternative hypothesis.

2. Adjust the probability p_i by computing $p_i^* = min(p_i, 1 - p_i)$. This step allows us

to obtain a probability p_i^* having the same interpretation as a p-value, in the sense that we reject the null hypothesis for small values of p_i^*. Using the original p_i would require to reject the null hypothesis for small or high values of p_i. Note that now, the probability p_i^* is asymptotically distributed according to a Uniform variable, on the interval $[0, 1/2]$, under the null hypothesis.

3. Order the n adjusted probabilities such that $p^*_{(1)} \leq \ldots \leq p^*_{(n)}$

4. Reject all null hypotheses corresponding to $p^*_{(1)}, \ldots, p^*_{(i_\delta)}$, where

$$i_\delta = max\{1 \leq i \leq n : p^*_{(i)} \leq \frac{\delta i}{2n(1-\hat{p})}\}, \tag{2.9}$$

   δ represents the control level of the FDR and $\hat{p}$ is the estimated value of the proportion of true alternative hypotheses in the data.

Chapter 3

A particularly favorable case for the half-space approach

In this chapter, we study two cases where the half-space approach described in Chapter 2 is particularly favorable. Under a general scale model associated with a non-informative prior distribution, or under a gamma conjugate model, we show that the uniform property of the posterior probability p_i holds not only asymptotically, but for any values of m. In the first case, the uniformity holds exactly, since we show that the probability p_i can be actually written as a conditional p-value. In the second case, the uniformity holds only approximately, under some conditions easily satisfied in microarray experiments. The approach described in Section 2.4 is then particularly favorable in the context of microarray data, this case representing the perfect illustration of a large number of tests (large n) with a small number of observations per test (small m). As we show in this chapter, the use of the posterior probability p_i provides advantages over traditional p-values (parametric or non-parametric), especially when m is small.

The first section of this chapter deals with the special case $m = 1$, in a generalized scale parameter model. The case of the gamma conjugate model is treated in the second section. Section 3 provides an application of the methodology proposed in section 2.4, for the gamma model, to microarray gene expression data. A simulation study is performed in the last section, presenting a comparison of the performance of our proposed method with some parametric and non parametric p-values, as well as with another Bayesian approach.

3.1 The non-informative case

In this section, we show that under some conditions on a generalized-scale parameter model associated with a Jeffrey's prior, the posterior probability considered can be seen as a conditional p-value, given one of the two observed samples. We consider here the case where $m = 1$. Note that in the case of microarray data, it is standard to spot each gene several times (usually two to three times) on the cDNA arrays, in a consecutive way. In many statistical methods available to analyze this kind of data, it is advised to take the mean of the expression for each gene and each treatment. The dataset obtained has effectively only one observation per gene, if no biological replication is available, for each condition, as stated in the next theorem. Following the notation defined in Section 2.2, we note here that the vector $\underset{\sim}{X_i}$ is replaced by the variable X_i, and similarly that the vector $\underset{\sim}{Y_i}$ is replaced by Y_i.

Theorem 3.1. *If the density $f(.)$ and the prior $\pi(.)$ satisfy the following:*

$$i) \quad f(z|\theta) = \tfrac{1}{\theta} h(\tfrac{z}{\theta})$$

$$ii) \qquad \pi(\theta) \propto 1/\theta$$

then, the quantity p_i, defined as $P(\theta_{xi} > \theta_{yi} | \underset{\sim}{x_i}, \underset{\sim}{y_i}, H_i = 1)$, can be written as

$$p_i = P(X_i \leq x_i \mid Y_i, H_i = 0).$$

Proof. First, we note that the posterior distribution of θ_{xi} under H_{1i} is

$$\pi(\theta_{xi}|x_i) = \frac{f(x_i|\theta_{xi})\pi(\theta_{xi})}{m(x_i)},$$

where $m(x_i) = \int f(x_i|\theta_{xi})\pi(\theta_{xi})d\theta_{xi}$ is the marginal density of X_i. Then, we can write

$$\begin{aligned} p_i &= \int_0^{+\infty} \frac{f(y_i|\theta_{yi})\pi(\theta_{yi})}{m(y_i)} \left[\int_{\theta_{yi}}^{+\infty} \frac{f(x_i|\theta_{xi})\pi(\theta_{xi})}{m(x_i)} d\theta_{xi} \right] d\theta_{yi}, \\ &= \int_0^{+\infty} \frac{f(y_i|\theta_{yi})\pi(\theta_{yi})}{m(y_i)} A(\theta_{yi}) d\theta_{yi}, \end{aligned}$$

where

$$\begin{aligned} A(\theta_{yi}) &= \int_{\theta_{yi}}^{+\infty} \frac{f(x_i|\theta_{xi})\pi(\theta_{xi})}{m(x_i)} d\theta_{xi}, \\ &= \left[\int_{\theta_{yi}}^{+\infty} \frac{1}{\theta_{xi}} h\left(\frac{x_i}{\theta_{xi}}\right) \frac{1}{\theta_{xi}} d\theta_{xi} \right] / m(x_i). \end{aligned}$$

By doing to successive changes of variable: $z = x_i/\theta_{xi}$ and then $x = \theta_{yi}z$, we obtain

$$\begin{aligned} A(\theta_{yi}) &= \frac{1}{m(x_i)} \int_0^{x_i} h(\frac{x}{\theta_{yi}}) \frac{1}{\theta_{yi}} \frac{1}{x_i} dx, \\ &= \frac{1}{m(x_i)} \int_0^{x_i} f(x|\theta_{yi}) \frac{1}{x_i} dx \end{aligned}$$

By making the same change of variables in the integral $m(x_i)$, we have

$$m(x_i) = \int_0^{+\infty} f(x|\theta_{yi}) \frac{1}{x_i} dx$$

Then,

$$\begin{aligned} A(\theta_{yi}) &= \frac{\displaystyle\int_0^{x_i} f(x|\theta_{yi}) dx}{\displaystyle\int_0^{+\infty} f(x|\theta_{xi}) dx}, \\ &= P(X_i \leq x_i \mid \theta_{yi} = \theta_{xi}). \end{aligned}$$

Then, the probability p_i can be rewritten as

$$\begin{aligned} p_i &= \int_0^{+\infty} \pi(\theta_{yi}|y_i) P(X_i \leq x_i \mid \theta_{yi}, H_i = 0) d\theta_{yi}, \\ &= P(X_i \leq x_i \mid Y_i, H_i = 0). \end{aligned}$$

□

First, note the symmetry of the result presented here: we showed that $p_i = p_i(x_i, y_i) = P(X_i \leq x_i \mid Y_i, H_i = 0)$, which is equivalent to $p_i(y_i, x_i) = P(Y_i \leq y_i \mid X_i, H_i = 0)$. This theorem implies that under the conditions *i)* and *ii)* stated above, the posterior probability of the alternative hypothesis can be written as a p-value, conditionally on the observed value of Y_i. A conditional p-value being uniformly distributed under the null hypothesis, as a p-value is, it is then relevant to integrate this probability into the multiple testing procedures presented in the last chapter. In this case, note that the uniform property holds not only asymptotically, but also for $m \geq 1$. Furthermore, the uniform distribution for $m = 1$ is not an approximation of the distribution of the probability p_i under H_0, but represents its exact distribution.

3.2 An extension of the uniform property to the informative case

In this section, we consider $2n$ independent random samples $\underset{\sim}{X}_i$ and $\underset{\sim}{Y}_i$ $(i = 1, \ldots, n)$ from $\mathbb{R}^m$. We suppose here that the hierarchical framework describing the data is the well known conjugate gamma model, with $m \geq 1$, where:

$$\text{With probability (1-p),} \begin{cases} X_{ij}|\theta_i \sim G\left(\alpha, \alpha/\theta_i\right), \\ Y_{ij}|\theta_i \sim G\left(\alpha, \alpha/\theta_i\right), \\ 1/\theta_i \sim G\left(\alpha_0, \alpha\nu\right). \end{cases} \tag{3.1}$$

$$\text{With probability p,} \begin{cases} X_{ij}|\theta_{xi} \sim G\left(\alpha, \alpha/\theta_{xi}\right), \\ Y_{ij}|\theta_{yi} \sim G\left(\alpha, \alpha/\theta_{yi}\right), \\ 1/\theta_{xi} \sim G\left(\alpha_0, \alpha\nu\right), \\ 1/\theta_{yi} \sim G\left(\alpha_0, \alpha\nu\right), \end{cases} \tag{3.2}$$

where θ_{xi} and θ_{yi} are assumed to be independent. Note that here, the gamma distribution is parameterized such that $E[X_{ij}|\theta_{xi}] = \theta_{xi}$ (and similarly for $E[Y_{ij}]$). In such a model, a set of hyperparameters, denoted (α, α_0, ν), is present. We consider here the hyperparameters of the distribution to be the same for X and for Y. In practice, we observed that such an assumption does not have any serious impact on the fit provided by the model. However, note that a generalization of this model, where the shape of the data is allowed to vary across each treatment, array and dye is presented in the next chapter. In the absence of prior knowledge that would allow us to fix the hyperparameters to some specified values, they need to be estimated. An empirical Bayes approach can be used, and the hyperparameters are set to maximize the predictive distribution of the data defined in (2.5), in a spirit similar to a maximum likelihood approach. The probabilities p_i actually used are then calculated using estimates of the hyperparameters from all the data, and we will assume the data set to be large enough (ie n large) that the hyperparameters are essentially known. Note that despite the fact that data are used twice, to estimate the parameters and to compute the posterior probability, the type of model considered here is more flexible than any model considering θ as being fixed. Furthermore, the set of parameters to be estimated is small.

Theorem 3.2. *Under the model (3.1), (3.2) above, the probability p_i defined in (2.8) can be written as*

$$p_i = P(\alpha m\bar{X}_i + \xi_i \leq \alpha m\bar{x}_i + \alpha\nu | \underset{\sim}{Y_i}, H_i = 0),$$

where ξ_i is a gamma variable with shape parameter α_0 and rate parameter $1/\theta_{yi}$ and where $\bar{x}_i = \sum_{j=1}^{m} x_{ij}/m$.

Furthermore, suppose the following conditions are satisfied:

i) α_0 *is a positive integer,*

ii) for all $\beta > 0.5$*, the* β*th quantile of* $\alpha m\bar{X}_i + \xi_i$*, given* $\bar{Y}_i$ *and* $H_i = 0$*, exceeds* M *with probability at least* $1 - \epsilon$ *(with respect to the distribution of* $\bar{Y}_i$*); the constant* M *is a specified positive number much larger than* $\alpha\nu$*, and* ϵ *is a specified small positive number.*

Then, p_i *is approximately uniformly distributed under* H_{0i} *on the interval* $[0, 1]$.

Proof. Using the same notations as in Theorem 3.1, we have

$$p_i = \int_0^{+\infty} \frac{f(\underset{\sim}{y_i}|\theta_{yi})\pi(\theta_{yi})}{m(\underset{\sim}{y_i})} A(\theta_{yi}) d\theta_{yi}.$$

By computing the integral $A(\theta_{yi})$, we obtain

$$\begin{aligned} A(\theta_{yi}) = \quad & \frac{1}{\Gamma(m\alpha + \alpha_0)} \frac{1}{(\alpha m\bar{x}_i + \alpha\nu)} \\ & \times \int_{\theta_{yi}}^{+\infty} \exp[-(\alpha m\bar{x}_i + \alpha\nu)/\theta_{xi}] \left(\frac{\alpha m\bar{x}_i + \alpha\nu}{\theta_{xi}}\right)^{m\alpha+\alpha_0+1} d\theta_{xi}. \end{aligned}$$

By letting $w = (\alpha m\bar{x}_i + \alpha\nu)/\theta_{xi}$ and then letting $x_0 = \theta_{yi}w$, we obtain

$$A(\theta_{yi}) = \int_0^{\alpha m\bar{x}_i + \alpha\nu} \frac{1}{\Gamma(m\alpha + \alpha_0)} \left(\frac{1}{\theta_{yi}}\right)^{m\alpha+\alpha_0} x_0^{m\alpha+\alpha_0-1} \exp\left(-x_0/\theta_{yi}\right) dx_0.$$

By noting that under the null hypothesis, $m\bar{X}_i|\theta_{yi} \sim G(m\alpha, \alpha/\theta_{yi})$, and by defining a new variable ξ_i, independent of $\bar{X}_i$, such that $\xi_i|\theta_{yi} \sim G(\alpha_0, 1/\theta_{yi})$, we finally get

$$A(\theta_{yi}) = P(\alpha m\bar{X}_i + \xi_i \leq \alpha m\bar{x}_i + \alpha\nu \mid \theta_{yi}, H_i = 0).$$

Then, the probability p_i can be written as

$$\begin{aligned} p_i &= \int_0^{+\infty} \pi(\theta_{yi}|\underset{\sim}{y_i}) P(\alpha m\bar{X}_i + \xi_i \leq \alpha m\bar{x}_i + \alpha\nu \mid \theta_{yi}, H_i = 0) d\theta_{yi}, \\ &= P(\alpha m\bar{X}_i + \xi_i \leq \alpha m\bar{x}_i + \alpha\nu \mid \underset{\sim}{Y_i}, H_i = 0). \qquad (3.3) \end{aligned}$$

To prove the second part of the theorem, we need to show that for all $\beta \in [0,1]$, $P(p_i \leq \beta|H_i = 0) \simeq \beta$. Since $p_i = p_i(\underset{\sim}{x_i}, \underset{\sim}{y_i}) = 1 - p_i(\underset{\sim}{y_i}, \underset{\sim}{x_i})$, it suffices to prove this result for $\beta > 0.5$. Here, we actually show that $P(p_i \leq \beta|\bar{Y}_i, H_i = 0) \simeq \beta$. First, we note that $g(w|\bar{y}_i)$ and $h(w|\bar{y}_i)$, the probability density functions under the null hypothesis of $\alpha m\bar{X}_i|(\bar{Y}_i = \bar{y}_i)$ and $\alpha m\bar{X}_i + \xi_i|(\bar{Y}_i = \bar{y}_i)$, respectively, are

$$\begin{aligned} g(w|\bar{y}_i) &= \frac{\Gamma(2m\alpha + \alpha_0)}{\Gamma(m\alpha)\Gamma(m\alpha + \alpha_0)} \frac{(\alpha m\bar{y}_i + \alpha\nu)^{m\alpha + \alpha_0}}{(w + \alpha m\bar{y}_i + \alpha\nu)^{2m\alpha + \alpha_0}} w^{m\alpha - 1}, \\ h(w|\bar{y}_i) &= \frac{\Gamma(2m\alpha + 2\alpha_0)}{\Gamma(m\alpha + \alpha_0)^2} \frac{(\alpha m\bar{y}_i + \alpha\nu)^{m\alpha + \alpha_0}}{(w + \alpha m\bar{y}_i + \alpha\nu)^{2m\alpha + 2\alpha_0}} w^{m\alpha + \alpha_0 - 1}. \end{aligned}$$

Furthermore, by defining $q_\beta(\bar{y}_i)$ as the βth-quantile of the variable $\alpha m\bar{X}_i + \xi_i$ under H_{0i} and given $\bar{Y}_i$, and by defining the following function

$$F(T, a, b, c) = \int_0^T \frac{t^a}{(t + c)^b} dt,$$

we can write

$$\begin{aligned} \beta &= \frac{\Gamma(2m\alpha + 2\alpha_0)}{\Gamma(m\alpha + \alpha_0)^2} (\alpha m\bar{y}_i + \alpha\nu)^{m\alpha + \alpha_0} \\ &\quad \times F\left(q_\beta(\bar{y}_i), m\alpha + \alpha_0 - 1, 2m\alpha + 2\alpha_0, \alpha m\bar{y}_i + \alpha\nu\right), \qquad (3.4) \end{aligned}$$

$$\begin{aligned} P\left[p_i \leq \beta | \bar{y}_i, H_i = 0\right] &= \frac{\Gamma(2m\alpha + \alpha_0)}{\Gamma(m\alpha)\Gamma(m\alpha + \alpha_0)}(\alpha m\bar{y}_i + \alpha\nu)^{m\alpha+\alpha_0} \\ &\quad \times F(q_\beta(\bar{y}_i) - \alpha\nu, m\alpha - 1, 2m\alpha + \alpha_0, \alpha m\bar{y}_i + \alpha\nu). \end{aligned}$$

By using the fact that, for all k,

$$F(T, a, b, c) = \left(\frac{T}{T+k}\right)^{a-b+1} F\left(T+k, a, b, c\frac{T+k}{T}\right), \ \forall\epsilon$$

and that under condition ii), $(q_\beta(\bar{y}_i) - \alpha\nu)/q_\beta(\bar{y}_i)$ is approximately 1 with high probability, we can rewrite (3.4) as

$$\begin{aligned} \beta &\simeq \frac{\Gamma(2m\alpha + 2\alpha_0)}{\Gamma(m\alpha + \alpha_0)^2}(\alpha m\bar{y}_i + \alpha\nu)^{m\alpha+\alpha_0} \\ &\quad \times F\left(q_\beta(\bar{y}_i) - \alpha\nu, m\alpha + \alpha_0 - 1, 2m\alpha + 2\alpha_0, \alpha m\bar{y}_i + \alpha\nu\right). \end{aligned} \tag{3.5}$$

Also, we note that for any α_0 integer, and for $a < b$, we have the following relation

$$F(T, a, b, c) = -\sum_{k=1}^{\alpha_0} \frac{a^{(k-1)}}{(b-1)^{(k)}} \frac{T^{a-k+1}}{(T+c)^{b-k}} + \frac{a^{(\alpha_0)}}{(b-1)^{(\alpha_0)}} F(T, a - \alpha_0, b - \alpha_0, c),$$

where $a^{(k)} = a(a-1)\ldots(a-k+1)$. By using this relation in (3.5) and using condition *ii*), we finally obtain

$$\beta \simeq P\left[p_i \leq \beta | \bar{Y}_i, H_i = 0\right].$$

□

The fact that under the conditions of the above theorem, p_i can be written in the form (3.3) does not imply that it is an exact p-value. This is partly due to the fact that the quantity $\alpha\nu$ cannot be seen as an observed value of the variable ξ_i, since it does not depend on the observed data. However, the quantity $\alpha\nu$ represents the expected value of ξ_i. The second part of the theorem shows that the probability p_i can be approximated in probability under the null hypothesis by a uniform distribution. Again, note that this result was not proven asymptotically, but for any value of m. As in the previous section, this statement allows us to use the quantity p_i in the spirit of the p-value, in some multiple testing procedures that require the uniformity under the null hypothesis. We note that these theorems can find very good applications in any area where multiple testing is an issue, and in particular, microarray data. In this case, we also note that assumption *ii*) of Theorem 2 is easily verified in practice, as illustrated in Section 3.4.1. This assumption is related to the magnitude of the gene expression and in general, the accuracy of the approximation increases with the level of expression of the gene. Assuming α_0 an integer does not have a great impact on the model and on the posterior probabilities, as we will see in the next sections. We note that the case $\alpha_0 = 0$ is treated in Theorem 3.1 and in this case, the posterior probability is exactly uniformly distributed under H_{0i}.

3.3 Application: Goodness of fit of the conjugate gamma model for microarray datasets

In this section, four microarray datasets are studied, generated from three different sources. The goal here is not to apply the method proposed to these datasets, but to evaluate the goodness of fit of the conjugate gamma model presented in the last section. The motivation behind this goodness of fit study is to show that the method we proposed can really be applied in practice. The first dataset, denoted TCDD, has been previously analyzed in Kerr et al. (2002). This experiment studies a compound (TCDD) known to induce a wide range of biological and biochemical responses, including gene induction. Two types of cells are studied in the experiment: a control one and a TCDD-treated cell line. The experiment was conducted over 6 arrays, with the same set of $n = 1907$ genes probed on each array, in a "triple dye-swap" experimental design. For now, we ignore the design component of the experiment and consider the 6 arrays as being $m = 6$ replications of the experiment. The design of this experiment is described in table 3.1. The second dataset, denoted DBLFLIP, is a sample data available with the R library MAANOVA, developed by the Jackson Laboratory (*cf* Kerr et al. (2002), Kerr & Churchill (2001a) and Kerr et al. (2001b)). This experiment is a four-array double dye swap experiment and the design is very similar to the one of the TCDD data. In this experiment, 14593 genes are studied and each gene is spotted twice on each array. The third and fourth datasets, denoted COLD and FIELD respectively, come from an experiment conducted at the University of Waterloo. This experiment is described in detail in Chapter 6 and contains 4896 genes, printed 3 times on each of six arrays, according to the design presented in Chapter 6, Table 6.2. Note that the median intensity of the three adjacent spots were taken. These two datasets represent two different treatments assigned to the plant organ-

	Dye1	Dye2
Array 1	Treatment	Control
Array 2	Treatment	Control
Array 3	Control	Treatment
Array 4	Treatment	Control
Array 5	Control	Treatment
Array 6	Control	Treatment

Table 3.1: Design of the experiment

ism *Thellungiella Salsuginea.* Throughout this section, the intensities corresponding to the treatment and control cases will be denoted $\underset{\sim}{X}$ and $\underset{\sim}{Y}$ respectively. Plots of the raw data $log(X)$ versus $log(Y)$ for these four datasets are presented in Figures A.1, A.2, A.3 and A.6.

We note that the datasets have been first normalized independently for each array using the LOWESS procedure (*cf* Yang et al. (2002)), described in Chapter 6. Furthermore, another data normalization process needs to be performed, in order to calibrate the signals from different channels and arrays to a comparable scale. Several methods are available, like the use of ANOVA models for example. A review of the normalization procedure used in the case study is provided in Chapter 6. For the current chapter, a simple transformation was applied to the data, so that the mean intensity for each array and for each dye is the same. Furthermore, in order to avoid any bias due to a strong treatment-dye effect, we transform the data such that the ratios of the means (Treatment/Control) are equal for each of the two dye combinations (Dye1/Dye2 and Dye2/Dye1). In particular, we denote by $\underset{\sim}{U}$ and $\underset{\sim}{V}$ the original data, and by $\underset{\sim}{X}$ and $\underset{\sim}{Y}$ the normalized dataset. For each gene i, we denote by $\bar{y_1}^{(i)}$ and $\bar{y_2}^{(i)}$ the control means for Dye1 and Dye2 respectively. Similarly, we denote by $\bar{x_1}^{(i)}$ and $\bar{x_2}^{(i)}$ the treatment means for Dye1 and Dye2. The notations $\bar{u_1}^{(i)}$, $\bar{u_2}^{(i)}$, $\bar{v_1}^{(i)}$ and $\bar{v_2}^{(i)}$ are similar and refer to the original dataset. The dataset is normalized such that

$$\frac{\bar{y_1}^{(i)}}{\bar{x_2}^{(i)}} = \frac{\bar{y_2}^{(i)}}{\bar{x_1}^{(i)}}, \tag{3.6}$$

for all $i = 1, \ldots, n$. By letting $\Delta^{(i)} = log(\bar{v_1}^{(i)}) - log(\bar{u_2}^{(i)}) - log(\bar{v_2}^{(i)}) + log(\bar{u_1}^{(i)})$, we can easily verify that the variables

$$x_{ij} = \begin{cases} \dfrac{u_{ij}}{\exp\left(\Delta^{(i)}/4\right)} \text{ for Dye 1} \\ u_{ij} \exp\left(\Delta^{(i)}/4\right) \text{ for Dye 2} \end{cases}$$

and

$$y_{ij} = \begin{cases} v_{ij} \exp\left(\Delta^{(i)}/4\right) \text{ for Dye 2} \\ \dfrac{v_{ij}}{\exp\left(\Delta^{(i)}/4\right)} \text{ for Dye 1} \end{cases}$$

satisfy (3.6). To see how well the gamma conjugate model fits the data, we chose to plot the histogram of the data (on the log-scale) versus the predictive density function (on the log scale) $m(\underset{\sim}{x})$, as we can see in Figure 3.1. This was done separately for the treatment and control data and only the results regarding the treatment are presented here (similar results were obtained regarding the control cases). We can see that the shape of the data varies a lot from one dataset to another, which confirms the fact that microarray data strongly depend on many factors that may be specific to laboratories or to the organism studied. As we can observe, the histograms of the TCDD and COLD datasets suggest that data may be adequately described by a mixture of models. However, the difference between the transformed and untransformed data (plots not shown here) does not suggest that this mixture can be explained by any effects due to the dye, array or treatment. We can see that the fit provided by the conjugate gamma model is not bad, considering the shape of the data, and considering that the predictive distribution we obtain is unimodal. We note however that the fit provided by our model to the DBLFLIP dataset is excellent,

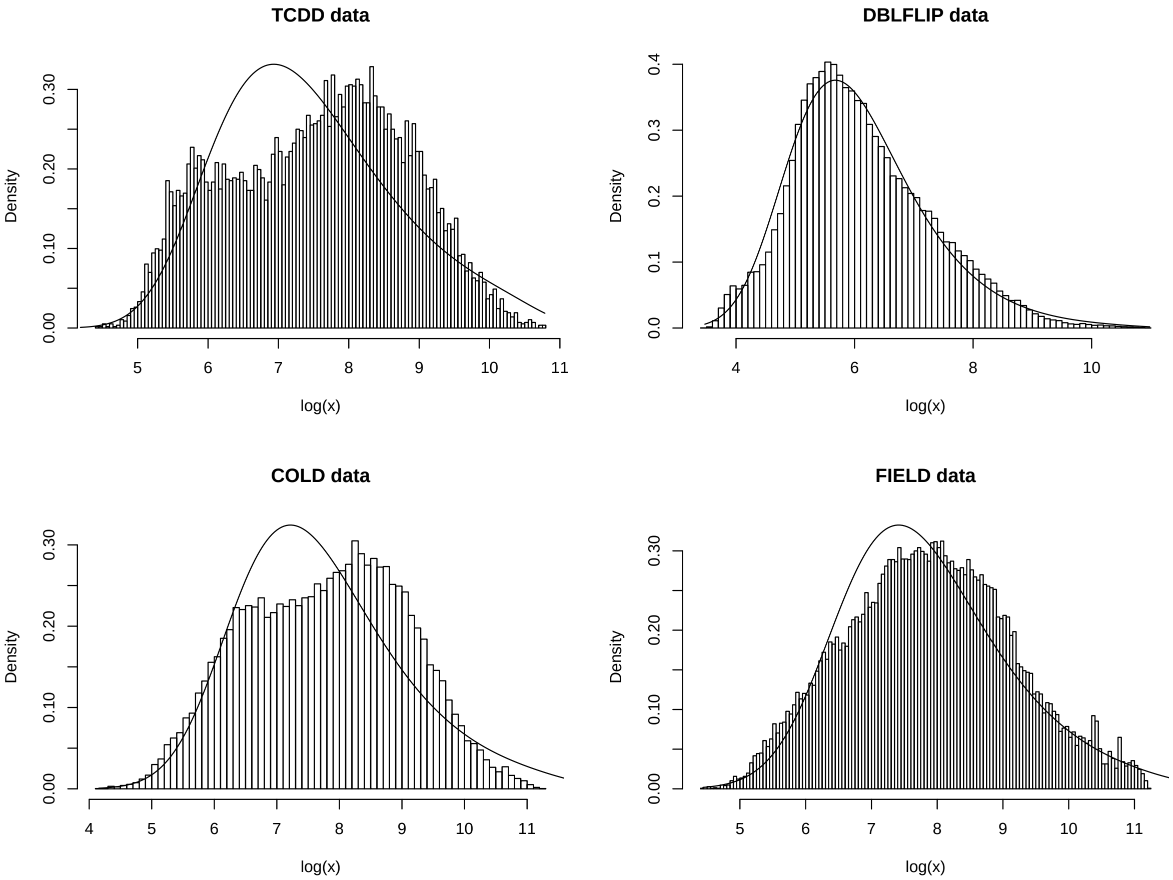


Figure 3.1: Histograms versus predictive density function for the four datasets using the gamma conjugate model (log-scale)

and we will see in Chapter5 that the fit to the FIELD dataset can be greatly improved by considering another type of model. Finally, note that the estimated hyperparameters of the model for each of the data are presented in Table 3.2. These parameters were obtained numerically, using a Newton-Raphson algorithm. Note that in the case of the data studied, this algorithm appeared to be very stable and seemed to converge to a global optimum (as observed in the simulations, where the estimated parameters were closed to the anticipated values).

Dataset	α	α_0	ν	p
TCDD	10.714	0.831	79.58	0.018
DBLFLIP	9.796	1.138	33.78	0.00143
COLD	5.814	0.891	210.1	0.1487
FIELD	5.309	0.9579	325.83	0.2193

Table 3.2: Estimated hyperparameters of the gamma conjugate model for the three datasets

3.4 Simulation study

The advantage of working with simulated datasets in the context of a mixture of models is that we know the true state of each observation. Then, for simulated gene expression data, we know if each gene is truly differentially expressed or not. This allows us to compute several types of error rates, and provides us with interesting tools to evaluate the performance of each of the methods and models tested.

In our case, several parameters may influence the performance of the methodology proposed. In particular, the aim of the simulation study is to provide insights regarding the influence of the parameters δ (level at which one wishes to control the FDR), p (proportion of differentially expressed genes in the dataset), m (number of replications available for each gene) and $\hat{p}$ (estimated proportion of differentially expressed genes).

3.4.1 Simulation design

The performance of the method is measured with respect to five criteria: the observed False Discovery Rate (FDR), False Negative Rate (FNR), Sensitivity (SENS), Specificity

(SPEC) and Risk (RISK). Using the notations of Table 2.1, these criteria are defined as

$$\begin{aligned} FDR &= V/R, \\ FNR &= T/(n-R), \\ SENS &= U/n_1, \\ SPEC &= S/n_0, \\ RISK &= (V+T)/n. \end{aligned} \tag{3.7}$$

In other words, we define the FDR as the proportion of Type I errors among the null hypotheses that are rejected. Similarly, the FNR represents the proportion of Type II errors among the null hypotheses that are accepted. Of course, very small values of FDR and FNR are expected. The sensitivity represents the proportion of alternative hypotheses that have been correctly rejected. It is then a measure of the method's ability to detect the differentially expressed genes among the alternative hypotheses. Note that this criterion should be the one under control in multiple hypotheses testing. Unfortunately, the number of true null and alternative hypotheses being difficult to estimate accurately, this criterion is extremely hard to control in practice, and we will see in our simulation study that values greater than 0.8 are rarely obtained. Similarly to the sensitivity, the specificity represents the proportion of null hypotheses that have been correctly accepted. It can be seen as a measure of the method's ability to judge correctly the equally expressed genes among the null hypotheses. Since the number of genes detected as differentially expressed is typically very small, it is not rare to obtain values of specificity that are very close to 1. Finally, the RISK represents the total number of errors (of Type I and II) among the n hypotheses tested and small values are expected.

We chose to apply our method using the multiple testing framework described in (2.9) and we refer to it as the p_i method. The probabilities p_i are computed using a Monte-Carlo approach. For instance, we generate $(\theta^*_{x,1}, \ldots, \theta^*_{x,K})$ and $(\theta^*_{y1}, \ldots, \theta^*_{yK})$ such that for all $i = 1, \ldots, K$

$$\begin{aligned} \theta^*_{xi} &\sim \pi(\theta_{xi} | \underset{\sim}{x_i}, H_{1i}), \\ \theta^*_{yi} &\sim \pi(\theta_{yi} | \underset{\sim}{y_i}, H_{1i}). \end{aligned}$$

In the case of the conjugate gamma model, note that the posterior distribution of the parameter θ is given under the alternative hypothesis by

$$1/\theta_{xi} | \underset{\sim}{x_i}, H_{1i} \sim G\left(m\alpha + \alpha_0, \alpha(m\bar{x}_i + \nu)\right),$$

and similarly for θ_{yi}. The probabilities p_i are then estimated such that

$$\hat{p}_i = \#\{i \in 1, \ldots, K : \theta^*_{xi} > \theta^*_{yi}\}/K.$$

In this document, note that for readability reasons p_i refers to the actual estimated value $\hat{p}_i$.

In the simulation study, the performance of our proposed methodology is compared with three other methods: two frequentist methods using p-values (one parametric and one non-parametric method) and one Bayesian method. The first method involves a likelihood ratio test for each gene (noted as LR), under the gamma model described in (3.1) and (3.2), ignoring the part regarding the prior distribution on θ. In order to avoid any over-parametrization problem, the parameters θ were estimated independently for each gene, using a maximum likelihood approach. The mixture parameter p was set to be the true value under which data were generated. Of course, in reality, this parameter would be unknown and would need to be estimated. Several problems can be encountered by doing so since it implies that the parameters θ of the model cannot be estimated independently from gene to gene. However, the focus of this simulation study being to compare the performance of our proposed method with some other ones, we will not comment more on that point. The second method is a non-parametric Wilcoxon test (unpaired, noted WILC). In both methods, the p-values obtained are used in the multiple testing framework described in Section 2.4, similar to the one used for our method.

The third method we use is the one described by Muller et al. (2004). In this paper, several loss functions were studied. In order to get comparable results across the different methods, we chose to use the decision rule minimizing a bivariate loss function, representing the posterior expected FDR and the posterior expected FNR. Using the Lagrangian method, it was shown that the decision rule minimizing this loss function is equivalent to the decision rule providing a control of the posterior expected FDR at a level δ. In this sense, this approach can be seen as the Bayesian version of the methods controlling for the FDR, and it is then relevant to compare it with our approach. Another interesting reason for comparing these two methods is that the test statistic used by Muller et al. is $v_i = P(H_{1i}|\underset{\sim}{x_i}, \underset{\sim}{y_i})$, which is the unconditional two-sided version of our probability p_i. Note that this method is referred as the v_i method throughout this document. Similarly to the probabilities p_i, the probabilities v_i are computed under the conjugate gamma model described in (3.1) and (3.2), and we obtain

$$v_i = \frac{p\ m_1(\underset{\sim}{x_i}, \underset{\sim}{y_i})}{pm_1(\underset{\sim}{x_i}, \underset{\sim}{y_i}) + (1-p)m_0(\underset{\sim}{x_i}, \underset{\sim}{y_i})},$$

where $m_1(\underset{\sim}{x_i}, \underset{\sim}{y_i})$ and $m_0(\underset{\sim}{x_i}, \underset{\sim}{y_i})$ have been defined in (2.5). The estimated posterior expected FDR is computed for each threshold t (rejecting H_{0i} if $v_i > t$) such that

$$F\bar{D}R(t) = \frac{\sum_{i=1}^{n} d_i(1-v_i)}{D+\epsilon},$$

where $(d_1, \dots, d_n)$ represents the vector of decisions, such that $d_i = 1$ if the null hypothesis H_{0i} is rejected, and 0 otherwise, and where ϵ is a small positive number avoiding a division by zero. The quantity D represents the total number of rejected null hypotheses ($D = \sum_{i=1}^{n} d_i$). The optimal decision rule rejects all hypotheses with $v_i > t^*$, where

$$t^* = min\{s : F\bar{D}R(s) \leq \delta\},$$

and where δ is the level of control of the posterior FDR.

In the following sections, datasets were simulated according to the conjugate model described in (3.1) and (3.2). The parameters chosen are identical to those chosen by Kendziorski et al. (2002), with $\alpha = 10$, $\alpha_0 = 0.8$ and $\nu = 91$. For each dataset, $n = 1000$ genes were simulated, with a proportion p of them being under H_1 and with m replications each (values of p and m may change according to the type of simulation and will be specified later). Each result presented in the next sections is based on 500 simulated datasets, and values of the observed error rates FDR, FNR, SENS, SPEC and RISK are computed as an average over the 500 simulations. Note that standard errors of the estimated rates are not presented here since for each estimated value x, it is of the order of $(x(1-x)/500)^{1/2}$. In most of the case, standard errors are of the order 10^{-3}. We also note that here, the condition $i)$ of Theorem 3.2 does not hold since α_0 is not an integer, but we will see that the violation of this condition does not have a serious impact on the results. The condition $ii)$ of this theorem holds in this example and in microarray datasets typically. As an illustration, we mention that the minimum average of $\alpha m\bar{x}_i + \xi_i$ was found to be 4570, over 500 simulated datasets with $m = 10$ and $p = 0.1$.

3.4.2 Influence of the controlled FDR level δ

In this section, we study the influence of the controlled FDR level δ on the performance of the methods. This section also allows us to have a global appreciation of the performance of the methods with respect to each others in terms of the different error rates. We simulate here datasets with $m = 10$ and $p = 0.1$. Values of δ varied as $\delta = (1\%, 2\%, 3\%, 4\%, 5\%)$. Results are presented in Table 3.3. Regarding the False Discovery Rate, note that we expect values as close to δ as possible for the four methods. The first thing that we observe is that our proposed method is able to provide an excellent control of the FDR with values not exceeding the level of control. The performance of the v_i method is also very good, but values of FDR are slightly over the level of control. The Wilcoxon test provides the lowest values of FDR, but does not control it as well as the two Bayesian methods. Finally, the likelihood ratio test does not perform well, with values of FDR systematically over the level of control by 1% to 2%. Regarding the False Negative Rate, the performance of the two Bayesian methods is identical and

Error Rate	Method	$\delta = 1\%$	$\delta = 2\%$	$\delta = 3\%$	$\delta = 4\%$	$\delta = 5\%$
FDR	p_i	0.010	0.020	0.030	0.039	0.048
	v_i	0.012	0.023	0.033	0.043	0.053
	LR	0.017	0.032	0.048	0.063	0.078
	WILC	0.008	0.016	0.024	0.034	0.043
FNR	p_i	0.023	0.022	0.021	0.020	0.020
	v_i	0.023	0.021	0.021	0.020	0.020
	LR	0.026	0.023	0.022	0.021	0.021
	WILC	0.029	0.027	0.025	0.024	0.024
SENS	p_i	0.787	0.800	0.807	0.814	0.819
	v_i	0.791	0.803	0.811	0.818	0.822
	LR	0.764	0.785	0.797	0.805	0.811
	WILC	0.732	0.754	0.767	0.776	0.784
SPEC	p_i	0.999	0.998	0.997	0.996	0.995
	v_i	0.999	0.998	0.997	0.996	0.995
	LR	0.999	0.997	0.995	0.994	0.992
	WILC	0.999	0.999	0.998	0.997	0.996
RISK	p_i	0.022	0.022	0.022	0.022	0.022
	v_i	0.022	0.022	0.022	0.022	0.022
	LR	0.025	0.024	0.024	0.025	0.026
	WILC	0.027	0.026	0.025	0.025	0.025

Table 3.3: Error rates for the simulation study with respect to the level of FDR control δ. Data were simulated with $p = 0.1$ and $m = 10$.

both provide the lowest rate compared to the other two methods. The likelihood ratio test performs also reasonably well. However, the Wilcoxon test gives the highest FNR, which is the price to pay for the low rate of false discoveries. Regarding the sensitivity, we can see again that the two Bayesian methods give the best performance, with slightly better rates for the v_i method. Again, the likelihood ratio test performs well and the Wilcoxon test provides the lowest rate. Regarding the specificity, all methods perform very well, with slightly lower rates for the likelihood ratio test. Finally, in terms of RISK, the performance of the two Bayesian methods is very good, with an average of 2.2% of errors among the 1000 tests. The performance of the likelihood ratio and Wilcoxon tests is also good, but the RISK is higher than for the Bayesian methods.

Regarding the behavior of the methods with respect to δ, we observe results that could be predicted. As we could expect, the False Negative Rate acts in the opposite way to the FDR, and values of FNR decrease with δ. Regarding the sensitivity, we note that whenever δ decreases, the number of detected genes should also decrease. Since the number of positive genes remains constant for any value of δ, we observe an increasing trend for this criterion when δ increases. The same argument can be applied to the specificity: since the number of correctly judged negative genes should increase with δ, we observe an increasing trend of this criterion with δ. The behavior of the RISK with respect to δ is harder to predict since the number of false negative and the number of false positive move in opposite directions with respect to δ. For the Bayesian methods, we observe a constant RISK over the values of δ, meaning that the loss due to the false discoveries is compensated with an equal weight by a gain in false negative. The likelihood ratio test and the Wilcoxon test do not behave similarly: it seems that globally, the RISK increases with δ for the LR method, whereas it decreases for the WILC method. In this last case, it may suggest that the gain in terms of true false negative is more important than the loss in terms of false positives, whenever δ increases.

3.4.3 Influence of p

In this section, we compare the influence of the parameter p on the different error rates for the four methods. Data were generated with $m = 10$ and the FDR was controlled at a level of $\delta = 5\%$. Values of p were taken to be $p = (0.01, 0.05, 0.1, 0.3, 0.5, 0.7, 0.9)$. Note that in microarray data, p rarely exceeds 0.5, but it is interesting to see the behavior of the methods when p is large. Results are presented in Table 3.4. Note that the same patterns regarding the performance of the methods with respect to each other occur, and we will not comment on that since it was done in the previous section.

By looking at the results, it seems that the FDR is difficult to control when p is small. For all the methods, we note that as p increases, the level of the FDR gets closer to the level of control $\delta = 5\%$. However, our proposed method is the one that seems to be the least influenced by the small value of p, with values of FDR extremely close to 5% even when p is as small as 0.05. As expected, the False Negative Rate increases with p, since the number of rejected hypotheses decreases with p. When p is below 0.7, it seems that all the methods perform similarly (except the Wilcoxon test, that tends to have a higher FNR, as we saw in the previous section). However, when p is very high, the v_i method performs significantly better (with an FNR of 38%) than the other ones who all have FNR values around 44%. Regarding the sensitivity, specificity and RISK, the same comments can be made as in the last section. However, it seems that all the methods are influenced the same way by the increase of p. Specifically, sensitivity and RISK increase

Error Rate	Method	$p = 0.01$	$p = 0.05$	$p = 0.1$	$p = 0.3$	$p = 0.5$	$p = 0.7$	$p = 0.9$
FDR	p_i	0.064	0.048	0.050	0.048	0.049	0.049	0.050
	v_i	0.101	0.057	0.053	0.051	0.050	0.050	0.051
	LR	0.088	0.082	0.079	0.071	0.066	0.062	0.053
	WILC	0.044	0.043	0.044	0.045	0.047	0.048	0.049
FNR	p_i	0.002	0.010	0.020	0.062	0.119	0.212	0.441
	v_i	0.002	0.010	0.020	0.061	0.116	0.204	0.387
	LR	0.003	0.011	0.021	0.063	0.118	0.021	0.440
	WILC	0.003	0.012	0.024	0.071	0.131	0.228	0.457
SENS	p_i	0.776	0.804	0.817	0.848	0.870	0.897	0.952
	v_i	0.780	0.808	0.820	0.853	0.875	0.902	0.961
	LR	0.742	0.791	0.809	0.848	0.874	0.902	0.955
	WILC	0.704	0.761	0.780	0.825	0.855	0.886	0.947
SPEC	p_i	0.999	0.998	0.995	0.981	0.955	0.891	0.538
	v_i	0.999	0.997	0.995	0.980	0.953	0.888	0.529
	LR	0.999	0.996	0.992	0.972	0.938	0.862	0.516
	WILC	1.000	0.998	0.996	0.983	0.958	0.895	0.563
RISK	p_i	0.003	0.012	0.023	0.058	0.087	0.105	0.089
	v_i	0.003	0.012	0.022	0.058	0.086	0.102	0.081
	LR	0.003	0.014	0.026	0.065	0.094	0.110	0.089
	WILC	0.003	0.014	0.025	0.064	0.093	0.111	0.091

Table 3.4: Error rates for the simulation study with respect to the proportion of differentially expressed genes p. Data were simulated with $m = 10$. A level of FDR control of $\delta = 5\%$ was used.

with p, whereas specificity decreases, especially when p is very large.

3.4.4 Influence of m

A very interesting aspect of the methodology proposed is its validity for any value of m. As we mentioned previously, no asymptotic assumptions are required for Theorems 3.1 and for 3.2. It is then very interesting to compare the performance of our method with respect to the others when m is small. In this section, data were generated with $p = 0.1$ and the FDR was controlled at a level of 5%. Values of m were taken to be $m = (4, 10, 20)$. Results are presented in Table 3.5.

We can see that the two Bayesian methods are not affected by the values of m, as far as the control of the FDR is concerned. However, both performances increase with respect to the other rates when m increases, and we can say that the two methods are equally affected by the increase or decrease of m. As expected, the two methods using p-values are the most affected by the value of m. First, when m is very small ($m = 4$), the Wilcoxon test is not able to detect any gene. Furthermore, as we could expect, the performance of the likelihood ratio test increases significantly with m, since p-values of this test are based on the asymptotic properties of the likelihood ratio statistic. As an illustration, we show in Figure 3.2 the histograms, under the null hypothesis, of the p-values based on the likelihood ratio statistic, as well as the histogram of the probability p_i, for a very small value $m = 4$. Both probabilities are assumed to be uniform under the null hypothesis, and the multiple testing procedure used is based on this assumption. We can clearly see that our approach, as we showed in Theorems 3.1 and 3.2, is not affected by the small value of m, whereas the p-value from the LR test is clearly not uniform. Again, this is an important advantage of the method we propose over the frequentist methods using

Error Rate	Method	$m = 4$	$m = 10$	$m = 20$
FDR	p_i	0.046	0.050	0.051
	v_i	0.051	0.055	0.055
	LR	0.113	0.078	0.065
	WILC	–	0.044	0.048
FNR	p_i	0.030	0.020	0.014
	v_i	0.030	0.019	0.014
	LR	0.037	0.020	0.014
	WILC	–	0.023	0.015
SENS	p_i	0.698	0.821	0.871
	v_i	0.705	0.825	0.873
	LR	0.631	0.813	0.870
	WILC	–	0.785	0.858
SPEC	p_i	0.974	0.995	0.993
	v_i	0.974	0.995	0.992
	LR	0.969	0.992	0.991
	WILC	–	0.996	0.993
RISK	p_i	0.032	0.022	0.017
	v_i	0.031	0.022	0.018
	LR	0.043	0.026	0.019
	WILC	–	0.025	0.018

Table 3.5: Error rates for the simulation study with respect to the number of observations available for each gene, m. Data were simulated with $p = 0.1$. A level of FDR control of $\delta = 5\%$ was used.

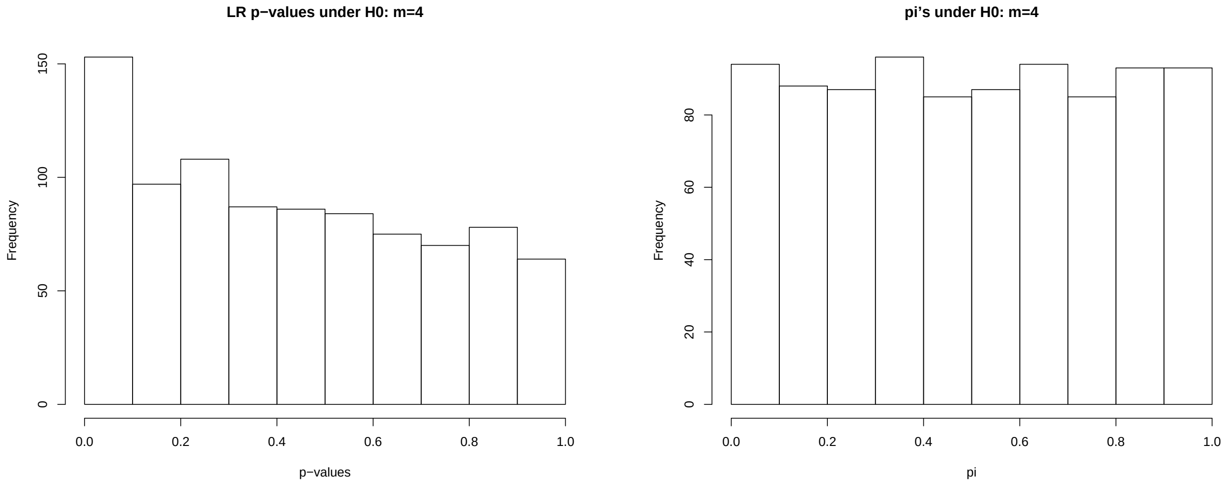


Figure 3.2: Histograms of the p-values from the likelihood ratio test (left) and of the p_i's (right) under H_0: Simulated data, $m = 4$.

p-values.

3.4.5 Influence of the estimated proportion of differentially expressed genes, $\hat{p}$

In this last section, we finally study the influence of the estimated proportion of differentially expressed genes, $\hat{p}$. As we mentioned before, regarding the two Bayesian methods (p_i and v_i), this proportion is estimated using an empirical Bayes approach, maximizing the predictive distribution of the data. The same value of $\hat{p}$ was taken for the Wilcoxon approach (in the multiple testing procedure) since this method is obviously not model-dependent. Finally, for the likelihood ratio test, the estimation of p by maximum likelihood leading to an over-parametrization of the model, we chose to use the real value of p.

If the model is correctly specified (which is the case in this simulation study), values of p

Error Rate	Method	$\hat{p} = 0.01$	$\hat{p} = 0.05$	$\hat{p} = 0.1$	$\hat{p} = 0.3$	$\hat{p} = 0.5$	$\hat{p} = 0.7$	$\hat{p} = 0.9$
FDR	p_i	0.046	0.047	0.050	0.065	0.090	0.148	0.443
	v_i	0.016	0.037	0.054	0.102	0.152	0.224	0.400
FNR	p_i	0.020	0.020	0.020	0.020	0.019	0.017	0.014
	v_i	0.023	0.021	0.020	0.018	0.017	0.016	0.014
SENS	p_i	0.815	0.815	0.816	0.822	0.830	0.843	0.885
	v_i	0.793	0.812	0.820	0.836	0.847	0.859	0.884
SPEC	p_i	0.996	0.995	0.995	0.994	0.991	0.983	0.920
	v_i	0.999	0.996	0.995	0.989	0.983	0.972	0.935
RISK	p_i	0.023	0.023	0.023	0.024	0.025	0.031	0.083
	v_i	0.022	0.022	0.023	0.026	0.030	0.039	0.071

Table 3.6: Robustness of the p_i and the v_i method to the estimation of the parameter p. Data were simulated with $m = 10$ and $p = 0.1$. A level of FDR control of $\delta = 5\%$ was used.

are usually correctly estimated. However, when p is very small and when n is very large, maximum likelihood methods or empirical Bayes methods may not provide accurate estimates. Storey (2002) proposed a method to estimate the parameter p, as we mentioned in (2.3), but again, accurate estimates are difficult to obtain. For these reasons, it is of interest to study the behavior of the methods when the parameter p is not well estimated. In this section, data were simulated with $m = 10$ and the FDR was controlled at a level of 5%. A proportion $p = 10\%$ of the genes were generated under the alternative hypothesis. The parameter estimate $\hat{p}$ was artificially taken to be $\hat{p} = (0.01, 0.05, 0.1, 0.3, 0.5, 0.8, 0.9)$. Since this parameter is used exactly in the same way for the p_i , the WILC and the LR method, only the results regarding the p_i method are presented here. These results can be compared with those obtained by the v_i method. Note that this method requires the value of $\hat{p}$ to compute the probabilities v_i, which is not the case for the p_i. As we can see in Table 3.6, the v_i method is the least robust to the lack of knowledge of the true value of p. Of course, both methods are influenced by the value of $\hat{p}$, but we can see for instance that with the v_i method, the FDR goes further away to the level of control, $\delta = 5\%$ as $\hat{p}$ goes away from the true value $p = 0.1$. The FNR is also more stable using the p_i method, as well as the RISK. Regarding the sensitivity and the specificity, both methods seem to be influenced the same way by an over or under estimation of p.

3.4.6 Summary and discussion

We presented in this chapter a particularly favorable case of the half-space approach described in Chapter 2. Through the proof of two theorems, we showed that the results presented in Chapter 2 regarding the posterior probability of half-space hold in a non-asymptotic manner under a non-informative gamma model, or under a conjugate gamma model. Note that this has been proven under some conditions that are easily satisfied in microarray experiments. Using four different microarray datasets, we could illustrate the fact that the gamma conjugate model may be a good candidate to describe this type of data. Using an extensive simulation study, we also illustrated several properties and advantages of our proposed approach (using the posterior probability of the one-sided alternative hypothesis), over other approaches using traditional p-values, parametric or non-parametric. We could also compare our method with the one proposed by Muller et al. (2004), that uses the two-sided version of the posterior probability p_i.

In general, we have seen that the two Bayesian approaches perform better than the frequentist ones in terms of the error rates FDR, FNR, SENS, SPEC and RISK. In terms of the control of the FDR, these Bayesian methods have the main advantage of being insensitive to small values of m, whereas the Wilcoxon test or the likelihood ratio tests are especially inefficient in this particular situation, typical of microarray experiments. Comparing the two Bayesian approaches with respect to each other, we note several advantages of our approach: it is able to provide a strict control of the FDR below the specified level without a loss in terms of the False Negative Rate, it is significantly less influenced by a poor estimation of p, which is a great advantage since this quantity can be hard to estimate accurately in practice, and finally, it performs better when p is small, which is typically the case in microarray experiments.

Working with the posterior probabilities defined here provides some great advantages. First, they can be used as p-values in a situation where p-values cannot be computed accurately. In microarray experiments, the number of replications, and hence the number of data points available for each gene, is not large enough to make inference on each gene in a frequentist setting. Bayesian hierarchical models, in the other hand, can provide a gene-specific inference keeping the number of parameters estimated reasonably small. The second advantage of the posterior probability used here is that it does not require the computation of a likelihood, or a marginal distribution. This allows us to work also with non-informative models, that may have improper priors, and consequently improper marginals. Some popular Bayesian hypothesis testing techniques require the use of the odds ratio of the marginals under the null and alternative hypothesis, as we have seen with the approach proposed by Muller et al. (2004). In such cases, the statistic of the test directly depends on the value of p, the proportion of alternative hypotheses in the dataset. This makes these types of methods less robust to a misspecification of this parameter. Furthermore, these methods are unfortunately unable to deal with non informative models, that have been proven to be very useful. Finally, we note that our approach allows us to differentiate between over expressed genes and under expressed genes, or allow us to test one-sided hypotheses, which is not the case of what we have called the v_i method. Looking at the direction of the fold change in the data, once a gene is detected as differentially expressed might be misleading: this is especially the case of genes for which a very strong array or dye effect is present.

Chapter 4

A multiplicative random effect ANOVA model for microarray data

In the previous two chapters, we have developed a multiple testing strategy, in a Bayesian framework, that can be used through the computation of the posterior probability of the one-sided alternative hypotheses. In such procedures, the design structure of the microarray experiment is not taken into account and there is the possibility of a potential confounding between the effect of the treatment and some sources of variation due to the arrays, the dyes or some genes themselves. It is clear that, to the extent these sources of variations are understood, they should be incorporated in the model. Kerr & Churchill (2001a), Kerr et al. (2001b) and Kerr et al. (2002) studied in a series of three papers the potential sources of variation in a microarray experiment. They were the first to incorporate the potential sources into an ANOVA model for gene expression data. Such a model enables the normalization of the data from array to array, but can also be used to detect differentially expressed genes through the estimation of the effect of the treatment for each genes.

In the first section of this chapter, we review the ANOVA model developed by Kerr & Churchill (2001a) as well as its extension to the case where some effects are treated as random. In the second section, we present a new multiplicative ANOVA model, that will be shown to be a generalization of the gamma hierarchical model introduced in the third chapter. Furthermore, we see that the posterior probabilities introduced in the third chapter can be estimated using a Gibbs sampling strategy. Under some specific constraints on the covariates of the ANOVA model, we also show that the distribution of these posterior probabilities of differential expression can be approximated under the null hypothesis by a uniform distribution. This result is similar to the one presented in Chapter 3. Finally, the different ANOVA models presented are applied to the DBLFLIP dataset and to a simulated dataset in the last two sections of this chapter.

4.1 An additive ANOVA model for microarray data

Several factors have been identified so far as contributing significantly to the variability of the data. Either these factors need to be incorporated in the model describing the data, or the data need to be normalized to correct for these effects, prior to the analysis. In both approaches, an ANOVA model seems to be a good choice to control this variability. Kerr & Churchill (2001a), Kerr et al. (2001b) and Kerr et al. (2002) were among the first statisticians to really emphasize the importance of a controlled design in a microarray experiment. In particular, they identified that the variation of the data arises from four main sources: the arrays (A), the dyes (D), the treatments (T) and the genes (G). Here, we note that we consider an experiment involving two dyes and K treatments. In addition to the main effects, some interactions contribute to the variation in the data. In particular, the array-dye interaction effect (AD) reflects the variability of the hybridization procedure (that allows the cDNA to bind to the DNA on each spot of the array) from array to array. Any interaction involving the gene effect is called a gene-specific effect. The gene-treatment interaction effect (TG) is the effect of interest and allows us to measure, for each gene, the effect of the treatment. The gene-dye interaction (DG) effect arises when, for example, the efficiency of incorporation of a dye varies for some specific genes. Finally, the array-gene interaction effect (AG) is due to the difference, in terms of amount of cDNA available for hybridization, from one spot (gene) to another, on each array.

Due to the structure of a microarray experiment, a careful design is necessary to avoid confounding between the effect of interest and some ancillary effects. To have a better understanding of the issues brought by such experiments in terms of the design, we first consider one gene per array. Since the same set of genes is printed on each array, the generalization to n genes is trivial. As we mentioned in the first chapter, each array is probed with two differently labeled cDNA samples (treatments). Then, we can informally say that two dyes and two treatments are applied on each array, for a total of two measurements per gene. The arrays can then be considered as being experimental blocks factors of size two. When there are more than two treatments to compare, not every treatment appears on every array and therefore, the experimental design is referred as an incomplete block design. Now, let us consider the case of two treatments. If each treatment is constantly labeled with the same dye, the issue of non-balance with respect to the dyes is then raised. In this case, the dye and treatment effects are completely confounded. In order to estimate the effect of the different fluorescent labels and then

	Green	Red
Array 1	T_1	T_2
Array 2	T_2	T_1

Table 4.1: A dye-swap experiment involving two treatments T_1 and T_2.

mean	ADT
A	DT
D	AT
T	AD
G	ADTG
TG	ADG
AG	DTG
DG	ATG

Table 4.2: Confounding structure for the dye-swap experiment

to minimize potential biases, the concept of dye-swapping has been introduced, which balances the dye and treatment factors. In dye swap experiments, each of the two treatments is labeled with both the red and the green dyes, over two replicated arrays (using the same mRNA samples). Although this technique is more costly in that it doubles the total number of arrays used, it is worthwhile since it allows the estimation of the dye effect in the model as well as a the reduction of the estimated variance of the data,

The design of a dye-swap experiment is presented in Table 4.1. We note that this design has the structure of a Latin Square and its confounding structure is presented in Table 4.2.

In the case of more than two treatments, a loop design can be efficient if the number of treatments is not too large. This design is presented in Table 4.3 and we can see that similarly to the dye swap, it is balanced with respect to the dyes since each treatment is labeled twice. This design loses in efficiency (average precision) when the number of

	Green	Red
Array 1	T_1	T_2
Array 2	T_2	T_3
...	...	...
Array (K-1)	T_{K-1}	T_K
Array K	T_K	T_1

Table 4.3: A loop design involving K treatments $T_1, \ldots, T_K$.

treatments compared increases, or when one wants to make every pairwise comparison. Furthermore, this design, like the dye-swap one, is not robust in the sense that the loss of a single array greatly affects the efficiency. For alternative designs, we refer to Kerr & Churchill (2001a).

Finally, as we mentioned in the first chapter, replication is a crucial step in microarray experiments. Genes are usually spotted two or three times on each array, using the same RNA source (subsampling). But inference about the biological population can only be made through the use of several arrays, where the mRNA extracted for each array comes from different biological samples. In practice, the designs described should be replicated at least two or three times using different biological samples of mRNA, to account for the biological variability in the model. These replicated experiments are referred as multiple dye-swap or multiple loop designs.

4.1.1 A fixed effect model

The ANOVA model used in the types of experiments described above is the following:

$$log(X_{ijkgr}) = \mu + A_i + D_j + (AD)_{ij} + G_g + (TG)_{jg} + (AG)_{ig} + S_{r(ig)} + \epsilon_{ijkgr},$$

where X_{ijkgr} is the intensity for array i, dye j, treatment k, gene g and spot r (in the case where genes are spotted several times on the arrays). In many cases, the intensity X has been previously transformed using a normalization procedure for each array separately. The term μ of the model refers to the overall average intensity and the following terms refers to the variation due to the array, dye, treatment, gene, as well as some specific interaction involving each gene. The term S captures the difference among the duplicated spots within an array. Finally, the terms ϵ's represent the error of the model and the only random quantity, if we consider all the effects in the model to be fixed. They are assumed to be independent and identically distributed with mean 0. We note that a normality assumption on the residuals is not required here. Note that treatment main effect was omitted on purpose in the model, since it is aliased with the array-dye interaction. This aliased structure does not have an impact on the analysis, since we are only interested in treatment differences that are specific to one gene, meaning that we care about the gene-treatment interaction.

The estimated effects of the model are the least squares estimates, minimizing the quantity $\sum_{ijkgr}\left[log(X_{ijkgr}) - \mu - A_i - D_j - (AD)_{ij} - G_g - (TG)_{jg} - (AG)_{ig} - S_{r(ig)}\right]^2$, under the constraints $\sum A_i = \sum D_j = \sum T_k = \sum G_g = \sum_g (AG)_{ig} = \sum_i (AG)_{ig} = \sum_g (TG)_{kg} = \sum_k (TG)_{kg} = \sum_r S_{r(ig)} = 0$. The effect of interest is the treatment-gene interaction, $(TG)_{kg}$, for each gene g and treatment k, with least squares estimate $(\hat{TG})_{kg} = t_{..kg.} - t_{..k..} - t_{...g.} + t_{.....}$. Here, t represents the logarithm of the intensity and a "." as an index means to average over that index. Under such a model, detecting the set of genes differentially expressed under at least one of the K treatments requires the test of the null hypothesis $H_{0g} : (TG)_{1g} = \ldots = (TG)_{Kg} = 0$ for each gene g. For the case of two treatments, under the latin square design, this is equivalent to testing the hypothesis of no differential expression between two treatments, where the null hypothesis becomes $H_{0g} : (TG)_{1g} = (TG)_{2g}$.

Three types of F-test statistics are proposed by Cui (2004), that differ in their denominator. The first statistic (F1) is a gene-specific F-statistic that compares the variation among replicated samples within and between treatment conditions. Its denominator is then based on the residual sum of squares of the model. The second statistic (F3) is identical to F1, but assumes a common error variance for all genes. The third statistic's denominator (F2) is a weighted combination of global and gene-specific variance estimates. We note that F2 and F3 do not follow the tabulated F distributions. Furthermore, the assumption of normality of the residuals does not seem to be appropriate (*cf* Kerr et al. (2001b)). A bootstrap analysis of the residuals is often required to assess the significance of the test statistics (see Efron & Tibshirani (1986)).

4.1.2 A mixed ANOVA model

A natural variation of the ANOVA model described earlier is to treat some of the effects as being random. Cui (2004) propose to treat the effects (AG) and S as being normally distributed with mean 0. The three types of F-tests can still be computed and we refer to Littel, Milliken, Stroup & Wolfinger (1996) for more details. Another type of mixed model was also introduced by Wolfinger, Gibson, Wolfinger, Bennett, Hamadeh, Bushel, Afshari & Paules (2001). This model is a two-stage model, the first stage being referred as a normalization model, where

$$log_2(X_{ijkgr}) = \mu + A_i + D_j + T_k + (AT)_{ik} + \epsilon_{ijkgr}$$

The terms A_i, $(AT)_{ik}$ and ϵ_{ijkgr} are assumed to be normally distributed with mean 0. The second stage of the model is referred to as a gene model, where the residuals of the first model r_{ijkgr} are fitted such that

$$r_{ijkgr} = G_g + (TG)_{kg} + (AG)_{ig} + \gamma_{ijkgr}$$

Here, the terms $(AG)_{ig}$ and γ_{ijkgr} are also assumed to be normally distributed with mean 0. All the random effects of the two-stage model are assumed to be independent both across their indices and with each other. Furthermore, the variance of the effects $(AG)_{ig}$ and γ_{ijkgr} is different across the genes, accounting for heterogeneity between the genes.

4.2 A multiplicative random effects ANOVA model

Traditionally, microarray data (and in general, design of experiment data) are analyzed using an additive regression model, either applied on the raw data, or on a suitable transformation of the data. In the case of microarray data, models are traditionally applied on the logarithm, in base 2 of the raw intensities. As we mentioned in the introduction of this chapter, our goal is to extend the multiple testing procedure developed in Chapter 2 and 3, in order to include important factors in microarray experiments, such as arrays and dyes. In order to do so, the model must satisfy the following criteria: the intensities are assumed to be gamma distributed, and the gene-specific effects considered are randomly distributed according to an inverse gamma distribution (to match the conjugate gamma model used before). Many computational challenges arise with this type of model, mostly due to the fact that the gamma distribution is not as tractable as the normal distribution, for instance. A solution to these computational issues is to keep the data on its natural scale, and to consider the model as being multiplicative, instead of additive. Using such a framework, the random effect model can be accurately defined, and effects can be estimated using a Gibbs sampling approach. Note that the multiplicative model described in the following subsection is not equivalent to a log-linear model, using the log-gamma distribution.

4.2.1 Definition of the model

The factors we consider in the model are the same as the factors presented in the last section: Array, Dye and Treatment. We first assume that the data have been previously corrected such that no global Array, Dye or Treatment effects are present. A simple, efficient and easy way to do so is, as we did in Section 3.3, to normalize the data such that the mean is the same across the arrays, dyes and treatments. For instance, we may assume the following model:

$$log(X_{ijkgr}) = \mu + A_i + D_j + T_k + \epsilon_{ijkgr},$$

where μ represents the global mean, A_i is the ith array effect, D_j is the jth Dye effect and T_k is the kth treatment effect. Note that we consider an experiment with I arrays ($i = 1, \ldots, I$), 2 dyes ($j = 1, 2$), K treatments ($k = 1, \ldots, K$), n genes ($g = 1, \ldots, n$) and m technical replications on the same array ($r = 1, \ldots, m$). The error term is noted ϵ_{ijkgr} and is assumed to have a mean of zero. The least square estimates for such a model can be easily obtained, and we find

$$\begin{aligned}\hat{\mu} &= \bar{X},\\ \hat{A}_i &= \bar{X}_{A_i} - \bar{X},\\ \hat{D}_j &= \bar{X}_{D_j} - \bar{X},\\ \hat{T}_k &= \bar{X}_{T_k} - \bar{X},\end{aligned}$$

where $\bar{X}$ is the global average of the data, $\bar{X}_{A_i}$ is the average for the ith array, $\bar{X}_{D_j}$ is the average for the jth dye and $\bar{X}_{T_k}$ represents the average over the kth treatment. Under such a simple model, data can be normalized so that

$$(log(X_{ijkgr}))^{New} = log(X_{ijkgr}) - \hat{A}_i - \hat{D}_j - \hat{T}_k.$$

Of course, this model is over-simple, but has the advantage of pre-processing the data in a very efficient way, for further gene-specific analysis. Note that in the following, the notation X will be used instead of X^{New} and then, data are assumed to be pre-processed. The model we propose is a gene-specific model, where all the effects are multiplicative, and where the data remain in their original scale. For instance, we assume that

$$E\left[X_{ijkgr}|(AG)_{ig},(TG)_{kg},(DG)_{jg}\right] = (AG)_{ig} \times (TG)_{kg} \times (DG)_{jg},$$

where (AG), (TG) and (DG) are the gene-specific Array, Treatment and Dye effects. In other words, we assume here that the gene-specific effects are additive for $log(E[X_{ijkgr}])$, instead of assuming additivity for $E[log(X_{ijkgr})]$ in the usual additive models. In order to allow more flexibility, and in order to minimize the impact of the small number of observations available for each gene, all the effects are assumed to be random, making the model a multiplicative random effects ANOVA model. The distributional assumptions are a generalization of the model described in the previous chapter, and we have

$$\begin{array}{rcl} X_{ijkgr}|(AG)_{ig},(TG)_{kg},(DG)_{jg} & \sim & Gamma\left(\gamma_{ij}, \dfrac{\gamma_{ij}}{(AG)_{ig} \times (DG)_{jg} \times (TG)_{kg}}\right), \\ 1/(AG)_{ig} & \sim & Gamma(a_i^{(A)}, b_i^{(A)}), \\ 1/(DG)_{jg} & \sim & Gamma(a_j^{(D)}, b_j^{(D)}), \\ 1/(TG)_{kg} & \sim & Gamma(a_k^{(T)}, b_k^{(T)}), \end{array} \tag{4.1}$$

Note that this choice of distribution allows us to work with a conjugate framework, and seems to be the only choice leading to a computation of posterior distributions that are mathematically tractable. The scale of the data is specified for each gene, dye, array and treatment. The shape is assumed to be constant across the genes, but specific to each array/dye (i, j) combination. We note that, due to the design of microarray experiments, assuming the shape to be different for each array i and dye j is equivalent to assuming a specific shape for array i, dye j and treatment k. Then, writing the terms γ's as γ_{ij} or γ_{ijk} is equivalent.

When we define an ANOVA model, we need to impose some constraints in order to provide an adequate estimation of the effects of interest (in this case, the gene-specific effects). In additive models, the choice of the type of constraints determines the estimation of the effects, but does not have a real impact on the model itself. In our case, this step is particularly important, since it has a direct influence on the tractability of the calculations leading to the estimation of the posterior distributions of the gene-specific effects. For instance, in the models described in the previous section, the constraints $\sum A_i = \sum D_j =$

$\sum T_k = \sum G_g = \sum_g (AG)_{ig} = \sum_i (AG)_{ig} = \sum_g (TG)_{kg} = \sum_k (TG)_{kg} = 0$ are used. In our context, this type of constraints (adapted to the multiplicative setting) cannot be used, for computational reasons that would prevent us from finding the posterior distribution of the gene-specific effects. As a results, the following constraints are used, for each $g = 1, \dots, n$:

$$\begin{aligned} (AG)_{Ig} &= 1, \\ (DG)_{2g} &= 1. \end{aligned} \tag{4.2}$$

We note here that no constraints were imposed on the gene-treatment effects $(TG)_{kg}$. The choice of constraints (4.2) has some consequences on the interpretation of the effects of interest. For example, the effect of array i is computed with respect to the effect of the last array I, for each $i = 1, \dots, I-1$. In other words, testing the hypothesis $(AG)_{ig} = 1$ (no effect of array i in gene g) is equivalent to testing the hypothesis $(AG)_{ig} = (AG)_{Ig}$. A similar interpretation holds for the gene-dye effects. However, since no constraints were imposed on the treatment effects, the hypothesis $(TG)_{kg} = (TG)_{k'g}$ really tests the hypothesis of no difference in expression between treatments k and k', for the gene g, which is the focus of our analysis.

The constraints defined lead to the definition of the covariates of the model. Each covariate is denoted by z_{ij} (noting that z_{ij} is equivalent to z_{ijk}) and is an indicator variable. For the gene-array effects, $(I-1)$ covariates, $z^{(A_1)}, ..., z^{(A_{I-1})}$, are defined such that $z^{(A_i)} = 1$ if the observation corresponds to array i, and 0 otherwise. Similarly, we define a single covariate, $z^{(D)}$, for the dye effect. Regarding the treatment effects, K indicator variables (one for each treatment) are defined and are noted $z^{(T_1)}, \dots, z^{(T_K)}$. From now on, for better readability, the set of $(I-1)+1+K$ covariates corresponding to the array, dye and treatment effects is denoted $z^{(1)}, \ldots, z^{(p)}$. The model, defined in (4.1), can then be written as

$$\begin{aligned} X_{ijgr}|\theta_{1g}, \ldots, \theta_{pg} &\sim Gamma\left(\gamma_{ij}, \frac{\gamma_{ij}}{\theta_{1g}^{z_{ij}^{(1)}} \times \ldots \times \theta_{pg}^{z_{ij}^{(p)}}}\right), \\ 1/\theta_{tg} &\sim Gamma\left(a_t, b_t\right), \text{ for } t = 1, \ldots, p, \end{aligned}$$

where the parameters $(\theta_{1g}, \ldots, \theta_{pg})$ are the random coefficients arising from the multiplicative structure of the model. Furthermore, the observations are assumed to be independent, conditionally on θ. The variables θ are also assumed to be independent.

4.2.2 Inference about the parameters via Gibbs sampling algorithm

The advantage of working with the model defined above is that inference about $\theta_{tg}|\underset{\sim}{x_g}$ is possible, for $t = 1, \ldots, p$ ($\underset{\sim}{x_g}$ representing the set of measurements available for gene g). This allows us to work with the posterior probability p_g, defined similarly to the probability p_i in the last chapter, such that

$$p_g = P(\theta_{tg} > \theta_{t'g}|\underset{\sim}{x_g}),$$

where, for example, θ_{tg} and $\theta_{t'g}$ represent the mean expression for treatment t and t' respectively.

We are going to see now that the computation of such statistics can be done via the Gibbs sampling algorithm. Suppose that one wants to make inference about θ_{tg} using the quantities $E[\theta_{tg}|\underset{\sim}{x_g}]$ or p_g, for some $t \neq t'$. The Gibbs sampling algorithm strategy can be used as followed:

0.) Generate $\theta_{1g}^{(0)}, \ldots, \theta_{pg}^{(0)}$ from $\pi(\theta_{1g}), \ldots, \pi(\theta_{pg})$.

1.1) Generate $\theta_{1g}^{(1)}$ from $\pi(\theta_{1g}|\theta_{2g}^{(0)}, \ldots, \theta_{pg}^{(0)}, \underset{\sim}{x_g})$

1.2) Generate $\theta_{2g}^{(1)}$ from $\pi(\theta_{2g}|\theta_{1g}^{(1)}, \theta_{3g}^{(0)}, \ldots, \theta_{pg}^{(0)}, \underset{\sim}{x_g})$

...

1.p) Generate $\theta_{pg}^{(1)}$ from $\pi(\theta_{pg}|\theta_{1g}^{(1)}, \ldots, \theta_{(p-1)g}^{(1)}, \underset{\sim}{x_g})$

2.) Repeat steps (1.1)-(1.p) K times to obtain $\underset{\sim}{\theta_g}{}^{(1)}, \ldots, \underset{\sim}{\theta_g}{}^{(K)}$.

It can be shown that for all $t = 1, \ldots, p$

$$\theta_{tg}^{(k)} \to \theta \sim \pi(\theta_{tg}|\underset{\sim}{x_g}) \text{ in distribution, when } k \to \infty.$$

Furthermore, we have, for all $t = 1, \ldots, p$, when $k \to \infty$

$$\begin{aligned} \frac{1}{k}\sum_{q=1}^{k} \theta_{tg}^{(q)} &\to E\left[\theta_{tg}|\underset{\sim}{x_g}\right], \\ \frac{1}{k} \times \#\{q : \theta_{tg}^{(q)} > \theta_{t'g}^{(q)}, q = 1, \ldots, k\} &\to P(\theta_{tg} > \theta_{t'g}|\underset{\sim}{x_g}). \end{aligned}$$

In the context of our model, the Gibbs sampling algorithm can be easily applied and then, numerical computation of $E[\theta_{tg}|\underset{\sim}{x_g}]$ and $P(\theta_{tg} > \theta_{t'g}|\underset{\sim}{x_g})$ is possible for each gene $g = 1, \ldots, n$ and each $(t, t') \in \{1, \ldots, p\}$. Of course, it requires the knowledge of the conditional probability density function of $\pi(\theta_{tg}|\theta_{[-t]g}, \underset{\sim}{x_g^*})$, where $\theta_{[-t]g}$ represents the set of variables $(\theta_{1g}, \ldots, \theta_{(t-1)g}, \theta_{(t+1)g}, \ldots, \theta_{pg})$. This density can be easily obtained using the relation

$$\pi(\theta_{tg}|\theta_{[-t]g}, \underset{\sim}{x}_g) \propto f(\underset{\sim}{x}_g|\underset{\sim}{\theta}_g)\pi(\theta_{tg})$$

and by noting that the exponential part of $f(\underset{\sim}{x}_g|\underset{\sim}{\theta}_g)$ can be decomposed into two exponentials, where the second part does not depend on θ_{tg}. In particular, we can write, for any $t \in \{1, \ldots, p\}$,

$$\begin{aligned} f(\underset{\sim}{x}_g|\underset{\sim}{\theta}_g) \;\; \propto \;\; & \left(\frac{1}{\theta_{1g}}\right)^{m\sum_{ij}\gamma_{ij}z_{ij}^{(1)}} \cdots \left(\frac{1}{\theta_{pg}}\right)^{m\sum_{ij}\gamma_{ij}z_{ij}^{(p)}} \\ & \times \exp\left[-\frac{1}{\theta_{tg}} m \sum_{ij} \frac{\gamma_{ij} z_{ij}^{(t)} \bar{x}_{ijg.}}{\displaystyle\prod_{u \neq t} \theta_{ug}^{z_{ij}^{(u)}}}\right] \exp\left[-m \sum_{ij} \frac{\gamma_{ij}(1 - z_{ij}^{(t)})\bar{x}_{ijg.}}{\displaystyle\prod_{u \neq t} \theta_{ug}^{z_{ij}^{(u)}}}\right], \end{aligned}$$

where $\bar{x}_{ijg.}$ represents the average of $(x_{ijg1}, \ldots, x_{ijgm})$ over the m replicated spots. We finally obtain

$$\frac{1}{\theta_{tg}}| \; \theta_{[-t]g}, \underset{\sim}{x}_g \sim Gamma\left(a_t + m\sum_{ij} z_{ij}^{(t)}\gamma_{ij} \;;\; b_t + m\sum_{ij}\frac{\gamma_{ij}z_{ij}^{(t)}\bar{x}^*_{ijg.}}{\displaystyle\prod_{u \neq t}\theta_{ug}^{z_{ij}^{(u)}}}\right),$$

for $t = 1, \ldots, p$. We can then see that if we condition on all θ's except one, we retrieve the gamma conjugate model defined in Chapter 2. Again, note that the computationability of this problem is a direct consequence of the choice of distribution for the random effects, as well as the choice of the type of constraints on the effects.

4.2.3 Estimation of the hyperparameters

Before fitting the model to the data, several hyperparameters need to be estimated. In total, $(2I)$ γ's parameters (for the conditional gamma distribution) and $2p$ parameters for the inverse-gamma prior (parameters a's and b's) are present in the model. Despite the fact that the conditional posterior distribution of each parameter θ_{tg} conditionally on $\theta_{[-t]g}$ can be obtained, the unconditional posterior distribution of θ_{tg} cannot be found. As a result, the marginal distribution of the data $\underset{\sim}{x_g}$ cannot be obtained and an empirical Bayes approach maximizing the unconditional likelihood is not possible. We then propose the following approach: first, naive estimates of the parameters θ's are obtained using a traditional least-square method, ignoring their random component. In particular, for the replicated dye-swap design (with $K = 2$ treatments) presented in Table 4.1 we get, for each gene g

$$\begin{aligned}
\tilde{\theta}_{\text{Array } i,\, g} &= \left[\prod x_{\text{Array } i,\, g} \Big/ \prod x_{\text{Array } I,\, g}\right]^{\frac{1}{2m}}, \text{ for } i = 1, \ldots, I-1, \\
\tilde{\theta}_{\text{Dye},\, g} &= \left[\prod x_{\text{cy3},\, g} \Big/ \prod x_{\text{cy5},\, g}\right]^{\frac{1}{Im}}, \\
\tilde{\theta}_{\text{Treat 1},\, g} &= \frac{\left[\prod x_{\text{Treat 1},g}\right]^{\frac{1}{mI}}}{(\tilde{\theta}_{A_1,\, g})^{1/I} \times \ldots \times (\tilde{\theta}_{A_{I-1},\, g})^{1/I} \times (\tilde{\theta}_{\text{Dye},\, g})^{1/2}}, \\
\tilde{\theta}_{\text{Treat 2},\, g} &= \tilde{\theta}_{\text{Treat 1},\, g} \times \left[\frac{\prod x_{\text{Treat 1},g}}{\prod x_{\text{Treat 2},g}}\right]^{\frac{1}{mI}},
\end{aligned}$$

where $x_{S,\, g}$ represents the set of all the observations belonging to the set S, for gene g. Once these naive estimates are obtained, we can fit the following model:

$$1/\tilde{\theta}_{tg} \sim G(a_t, b_t),$$

for each $t = 1, \ldots, p$, in order to obtain estimates of the coefficients a_t and b_t. Regarding the coefficients γ's, we first normalize the data for the gene-specific array and dye effects, using the naive estimates, such that

$$X^*_{ijkgr} = \frac{X_{ijkgr}}{\tilde{\theta}_{\text{Array } i,\ g} \times \tilde{\theta}_{\text{Dye, } g}}$$

Then, we can fit the following model

$$\begin{aligned} X^*_{ijkgr} &\sim G(\gamma_{ij}, \frac{\gamma_{ij}}{\theta_{\text{Treat k, } g}}), \\ 1/\theta_{\text{Treat k,}} &\sim G(\hat{a}_{\text{Treat k,}}, \hat{b}_{\text{Treat k,}}), \end{aligned}$$

where $\hat{a}_{\text{Treat k,}}$ and $\hat{b}_{\text{Treat k,}}$ are the estimated parameters obtained previously. In this case, the model is identical to the one fitted in Chapter 3, and estimates of the parameters γ's can be obtained by maximizing the predictive density function, as we did before. This way of estimating the hyperparameters of the model is very naive, and in practice, the naive parameters $\tilde{\theta}$'s tend to be less variable than the real ones. However, as we will see in our simulation study in Section 4.4, estimates appear to be very accurate if the model assumed is true.

4.2.4 Normalization of the data and detection of differentially expressed genes

The model we presented can be used for two different purposes: normalization or detection of differentially expressed genes. If one wants to normalize the data to correct for the gene-specific effects, we suggest to normalize such that

$$X_{ijkgr}^{new} = \frac{X_{ijkgr}}{E[\theta_{(K+1)g}|\underset{\sim}{x_g}]^{z_{ij}^{(K+1)}} \times \ldots \times E[\theta_{pg}|\underset{\sim}{x_g}]^{z_{ij}^{(p)}}},$$

where we suppose that the first K θ's represent the K treatments. We note here that we divide the raw data by the posterior expectation of the gene-specific effects, except for the gene-treatment interaction effects. Usually, the first interest of the biologists is to detect differentially expressed genes under the treatments. Normalizing the data by the gene-treatment effects would then remove the effect of interest.

By making inference about the parameters $\theta_{\text{treat k, }g}$ and $\theta_{\text{treat k', }g}$ corresponding to the treatments k and k', for example, we are able to detect the genes that are differentially expressed under these two treatments. This can be done by computing the posterior probabilities p_g defined as $p_g = P(\theta_{\text{treat k, }g} > \theta_{\text{treat k', }g}|\underset{\sim}{x_g})$. If we assume the other gene-specific effects as being fixed and known, the model (4.1) can be written exactly like the hierarchical gamma model described in the third chapter, involving data from two treatments k and k'. For instance, assuming $a_k = a_{k'}$ and $b_k = b_{k'}$, we can define

$$\begin{aligned} X_{ijgr}^* &= X_{ijkgr} \Big/ \prod_{t \neq (k,k')} \theta_{tg}^{z_{ij}^{(t)}}, \\ Y_{ijgr}^* &= X_{ijk'gr} \Big/ \prod_{t \neq (k,k')} \theta_{tg}^{z_{ij}^{(t)}}. \end{aligned}$$

In this case, the variables X^* and Y^* represent normalized data for treatment k and k' respectively. We can also denote θ_{kg} and $\theta_{k'g}$ as being θ_{xg} and θ_{yg}. Then, under the hypothesis $H_{1g} : \theta_{xg} \neq \theta_{yg}$, the model (4.3) can be rewritten as

$$\begin{array}{rclcrcl} X_{igr}|\theta_{xg} & \sim & Gamma\left(\gamma_{xi}, \dfrac{\gamma_{xi}}{\theta_{xg}}\right), \ (4.3) & \qquad & Y_{igr}|\theta_{yg} & \sim & Gamma\left(\gamma_{yi}, \dfrac{\gamma_{yi}}{\theta_{yg}}\right), \\ 1/\theta_{xg} & \sim & Gamma(a,b), & & 1/\theta_{yg} & \sim & Gamma(a,b). \end{array}$$

Then, the Theorem 3.2 proven in the third chapter can be generalized, and we obtain the following theorem.

Theorem 4.1. *Under the model (4.3) above, the probability p_g can be written as*

$$p_g = P\left(\sum_i m\gamma_{xi}\bar{X}_{ig} + \xi_g \leq \sum_i m\gamma_{xi}\bar{x}_{ig} + b|\underset{\sim}{Y}_g, H_g = 0\right),$$

where ξ_g is a gamma variable with shape parameter a and rate parameter $1/\theta_{yg}$ and where the variable H_g is defined as the variable H_i in the previous chapter. Furthermore, if the following conditions are satisfied

i) a is an integer

ii) the βth quantile of $m\sum_i \bar{X}_{ig}\gamma_{xi} + \xi_g$ under H_{0g}, given $m\sum_i \bar{Y}_{ig}\gamma_{yi}$ is large enough, for $\beta > 0.5$

then, p_g is approximately uniformly distributed under H_{0g} on the interval $[0,1]$.

The proof of this theorem is very similar to the proof derived for Theorem 3.2. Furthermore, since this theorem can be proven conditionally on all the gene-specific effects, by integrating over the distribution of these effects, the theorem is still valid unconditionally. As a result, we can use the posterior probability p_g as a p-value, for detecting differentially expressed genes, and controlling the FDR at the same time.

Design (Dye,Array,Treat)	2,1,1	1,1,1	2,2,1	1,2,1	1,3,2	2,3,2	1,3,2	2,4,2
Estimated γ	98.00	58.54	62.70	87.21	91.73	60.69	66.69	73.33

Table 4.4: Estimated parameters γ's for the DBLFLIP dataset

Parameters a	a_{Dye}	a_{Array1}	a_{Array2}	a_{Array3}	a_{Treat}
Value	63.78	1.88	4.91	3.30	0.94
Parameters b	b_{Dye}	b_{Array1}	b_{Array2}	b_{Array3}	b_{Treat}
Value	63.28	1.41	4.42	2.81	230.63
Expected value of	Dye	Array 1	Array 2	Array 3	Treatment
the effects: $b/(a-1)$	1.007	1.60	1.13	1.22	Not defined

Table 4.5: Estimated parameters a's and b's for the DBLFLIP dataset. Note that assumptions of Theorem 4.1 require the parameters (a, b) to be the same for both treatments.

4.3 Application: Study of the DBLFLIP dataset

In this section, we present the results of the multiplicative ANOVA model proposed, when applied to the DBLFLIP dataset, first described in Section 3.3. Three other datasets will be studied in details in Chapter 6. We recall that this experiment is constituted of four arrays ($I = 4$), two dyes and two treatments ($K = 2$). The estimated coefficients (a, b)'s and γ are presented in Tables 4.4 and 4.5. It is hard to compare these values with those obtained in the same dataset with gamma model developed in the last chapter (see Table 3.2), since the expected value of the gene expression is now decomposed into three different effects. However, as a reference, we note that we obtained $\gamma = 9.79$, $a = 1.138$ and $b = 795.29$. Furthermore, the expected value of the effects is also given in Table 4.5. We can see that even if the data were pre-corrected for the global effects, a strong gene interaction remains present regarding the first array.

In order to compute the posterior expectations of the gene-specific effects as well as the posterior probabilities p_g, the Gibbs sampling algorithm was applied, as described in

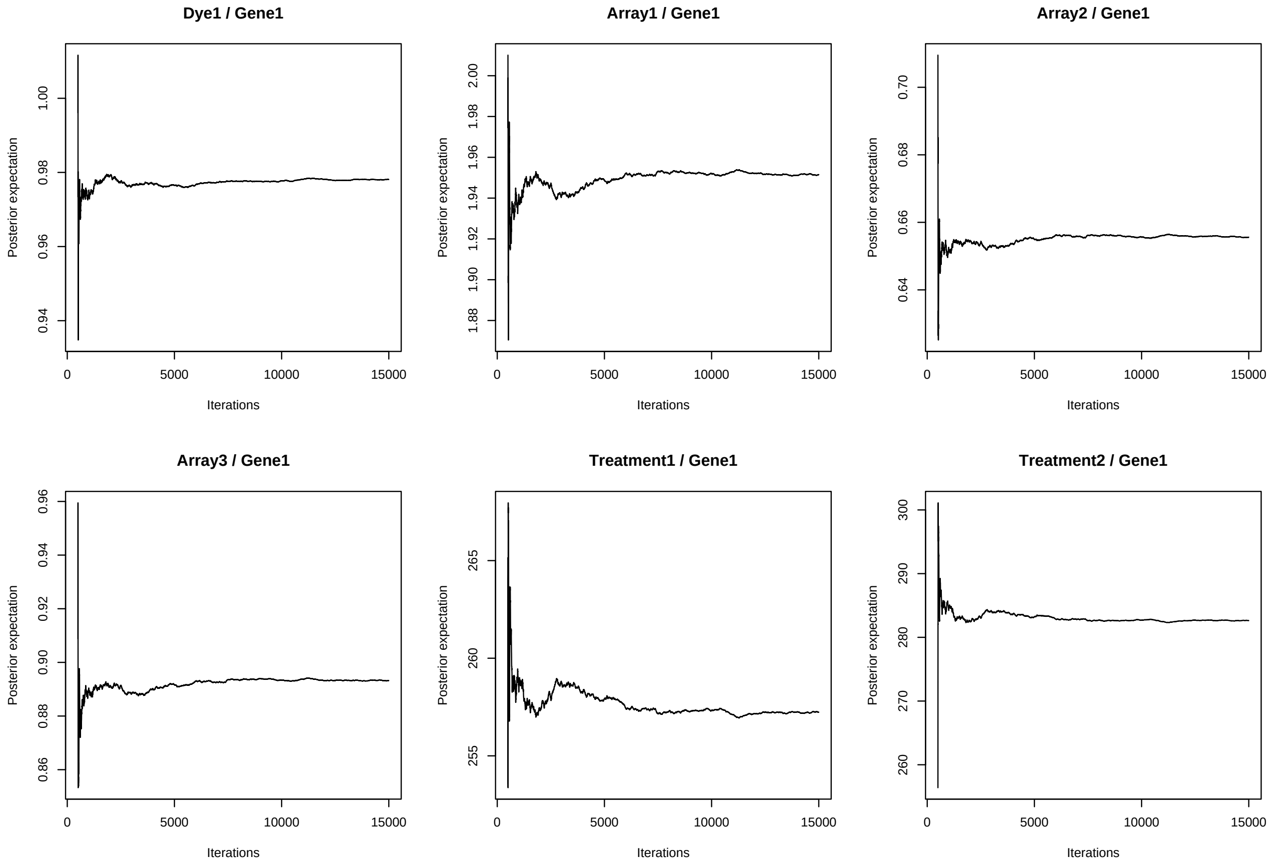


Figure 4.1: DBLFLIP dataset: convergence rate of the Gibbs sampling posterior expectations for the first gene

Section 4.2.2. The number of iterations used was 2500 (for each gene), where the first 500 iterations were ignored in the computation of the estimates. Different starting values were tried, generated form the prior distribution of the effects, or arbitrary fixed. In all the cases, values of the six estimates were exactly identical after the first 500 iterations (ignored in the computation).

Quantile	Dye	Array 1	Array 2	Array 3	Treatment 1	Treatment 2
0%	0.677	0.013	0.100	0.090	42.455	41.561
10%	0.889	0.429	0.605	0.513	103.014	102.690
20%	0.924	0.622	0.739	0.670	156.803	157.411
30%	0.949	0.788	0.834	0.803	218.767	218.811
40%	0.971	0.946	0.917	0.918	297.453	297.634
50%	0.992	1.103	1.006	1.039	401.925	404.052
60%	1.014	1.272	1.108	1.176	558.812	558.391
70%	1.042	1.444	1.233	1.336	823.850	819.416
80%	1.078	1.671	1.399	1.551	1298.590	1282.088
90%	1.142	2.027	1.668	1.888	2468.606	2465.324
100%	2.052	62.977	65.432	78.722	45857.500	47673.000

Table 4.6: Quantiles of the posterior expectations of the gene-specific effects for the DBLFLIP dataset

An illustration of the convergence rate of the posterior expectations, for the first gene of the dataset is presented in Figure 4.1. The posterior expectations of the gene-specific effects are presented in Table 4.6. Note that in this dataset, a very strong array effect is present. This is true especially for the first array, whose expression average was half that of the other arrays, before the correction of the data. This dataset illustrates perfectly a situation where ANOVA models are very useful and where there is the need to correct for these effects. The remaining analysis of this dataset is going to be divided into two parts: a goodness of fit study, where we compare the fit provided to the data under different conditions, and a study of the set of genes detected, including a comparison with the mixed model developed by Wu, Kerr, Cui & Churchill (2003) and implemented in the software package MAANOVA.

4.3.1 Goodness of fit study

Goodness of fit can be evaluated by looking at the residuals from the model we assumed. They can be defined as

$$\mathrm{R}_{ijkgr} \;=\; \frac{X_{ijkgr}}{E[\theta_{dye}|\underset{\sim}{x}_g^*]^{z_{ij}^{(D)}}\; E[\theta_{A_1}|\underset{\sim}{x}_g^*]^{z_{ij}^{(A_1)}} \dots E[\theta_{A_{I-1}}|\underset{\sim}{x}_g^*]^{z_{ij}^{(A_{I-1})}}\; E[\theta_{T_1}|\underset{\sim}{x}_g^*]^{z_{ij}^{(T_1)}} \dots E[\theta_{T_K}|\underset{\sim}{x}_g^*]^{z_{ij}^{(T_K)}}}.$$

The first question of interest is to know if fitting a multiplicative model (including all the gene-specific effects) is efficient, compared to the model defined in the previous chapter, where only the treatment-gene specific effects were considered.

Figure 4.2 presents the boxplots of the logarithm of the residuals (for the two treatments) when we fit a model with only the treatments effects, and when we fit a multiplicative model including all the array and dye effects. Note that in both cases, data have been corrected for the main Array, Dye and Treatment effects. We can see that there is a very strong improvement in the residuals by considering the multiplicative model. Not only their mean is closer to 1, but the variability is considerably reduced, illustrating the advantage of considering a more complex model.

Using these residuals, we can also construct some diagnostic plots, such as some quantile-quantile plots. By using the posterior expectation of the effects as the estimated coefficients of the regression model, and following model (4.1), we should have approximatively

$$\mathrm{Residual}_{ijkgr} \sim G(\gamma_{ij}, \gamma_{ij}),$$

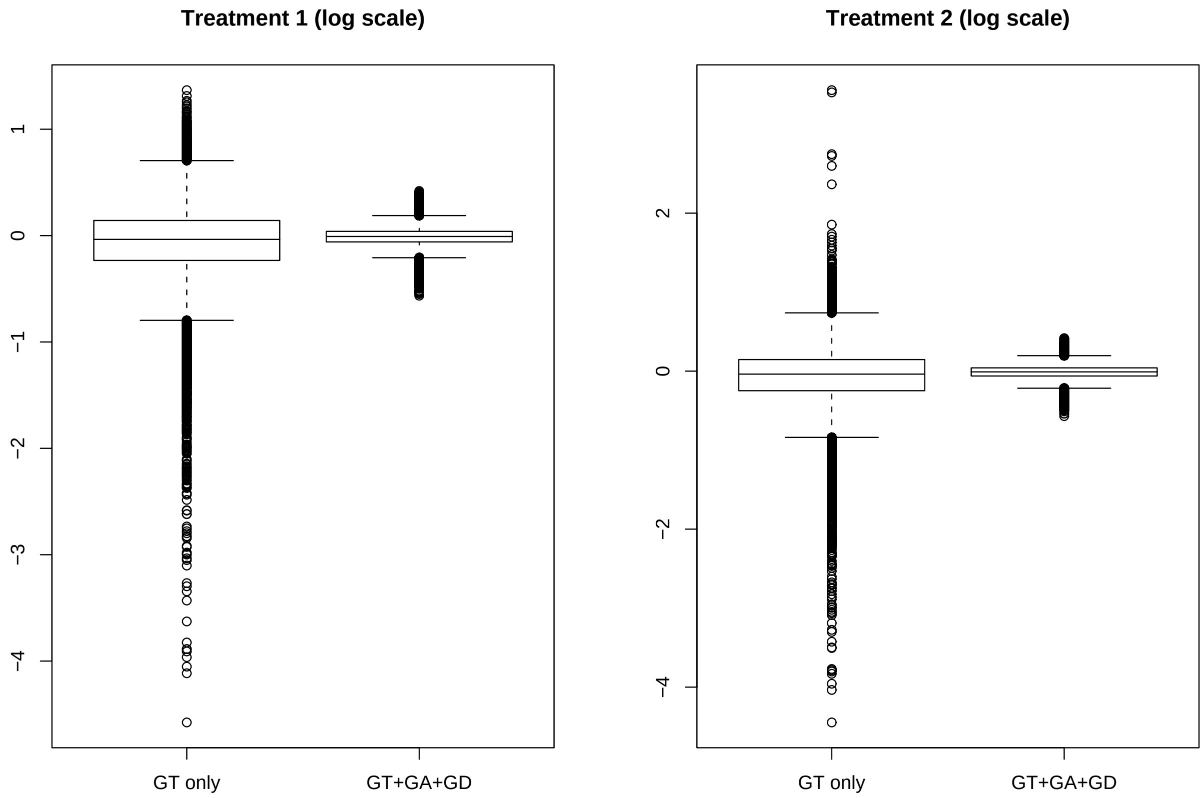


Figure 4.2: Box-plots of the residuals for the multiplicative model (on the right), and for the model considering only the treatment-gene effects (on the left)

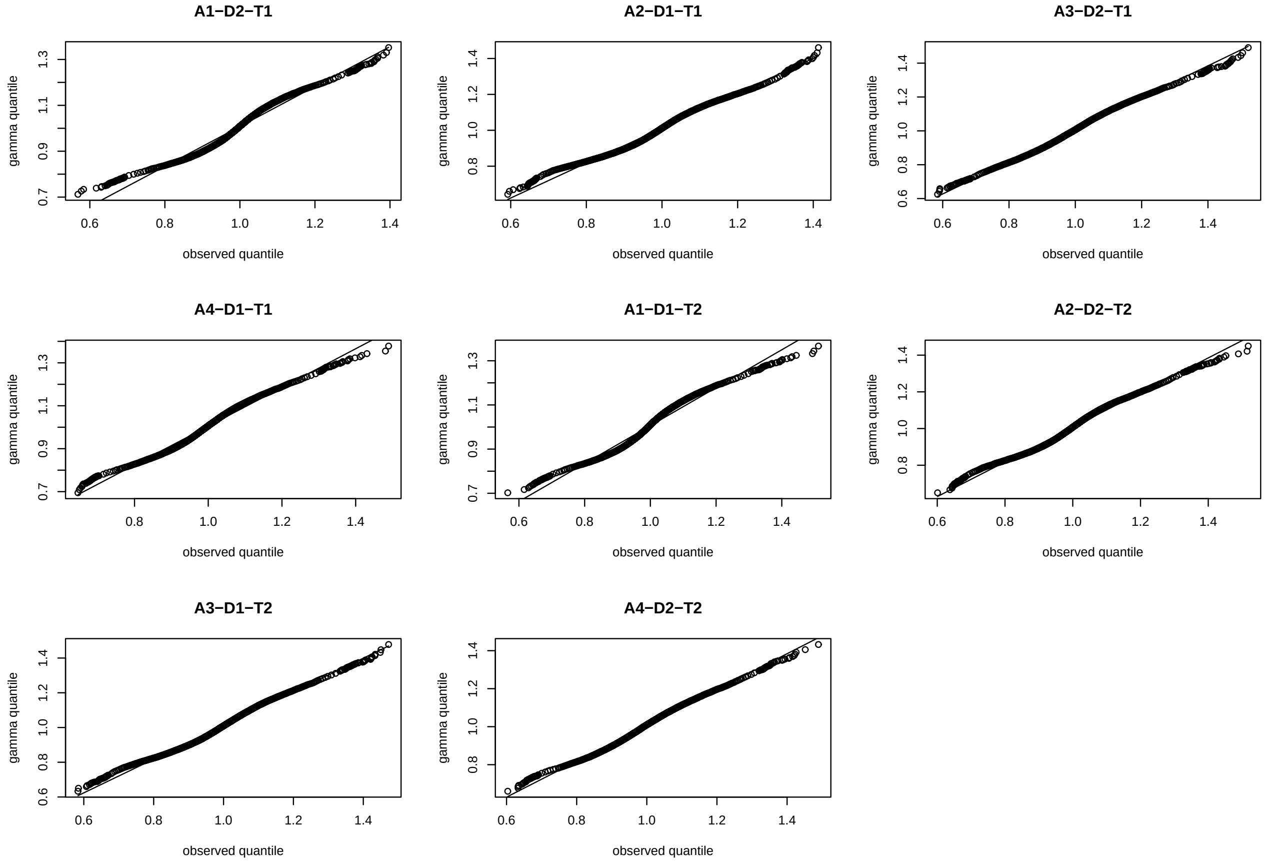


Figure 4.3: Quantile-quantile plots for the DBLFLIP dataset, using the multiplicative model

for each set (i, j). We can then plot the observed quantiles of the residuals versus the quantiles of a gamma distribution.

Such a qq-plot, for the multiplicative model, is presented in Figure 4.3, for the 8 combinations of Dye/Array/treatment. We can see that the fit provide by the model to this dataset is very good, with almost straight lines for each set of residuals.

4.3.2 Testing for differentially expressed genes

We applied the methodology proposed, using the posterior probabilities p_i, to this dataset. At a FDR control level of 5%, 287 genes were detected, which represents 1.96% of the genes printed in this experiment. The mixed model presented in Section 4.1.2 was also fitted, using the R package MAANOVA. In this model, the logarithm of the data is assumed to be distributed according to a normal distribution, and the gene-array interactions are the only random effects of the model. In that sense, note that the difference between our approach and this model is not only in the type of distribution chosen for the data since we assume all the gene-specific effects to be randomly distributed. FDR-corrected permuted p-values were computed, from four different F-tests, introduced briefly in Section 4.1.1.

Figure 4.4 presents the normal qq-plots of the residuals obtained. We can see that the fit provided by the mixed normal model is very poor, and residuals do not follow a normal distribution. As a result, none of the four F-tests are able to detect any genes, at a FDR-control level of 5% (the smallest FDR-corrected p-value being around 6%). Note that even if the truth regarding the state of expression of the genes is unknown, at least one gene, clearly over expressed in this dataset, should be detected by any method. In conclusion, this dataset illustrates very well a situation where log-normal models are not appropriate and where other distributions need to be considered.

4.4 Study of a simulated dataset

We study in this section the application of the multiplicative model described above to a single simulated dataset. The main purpose of this study is to evaluate the performance

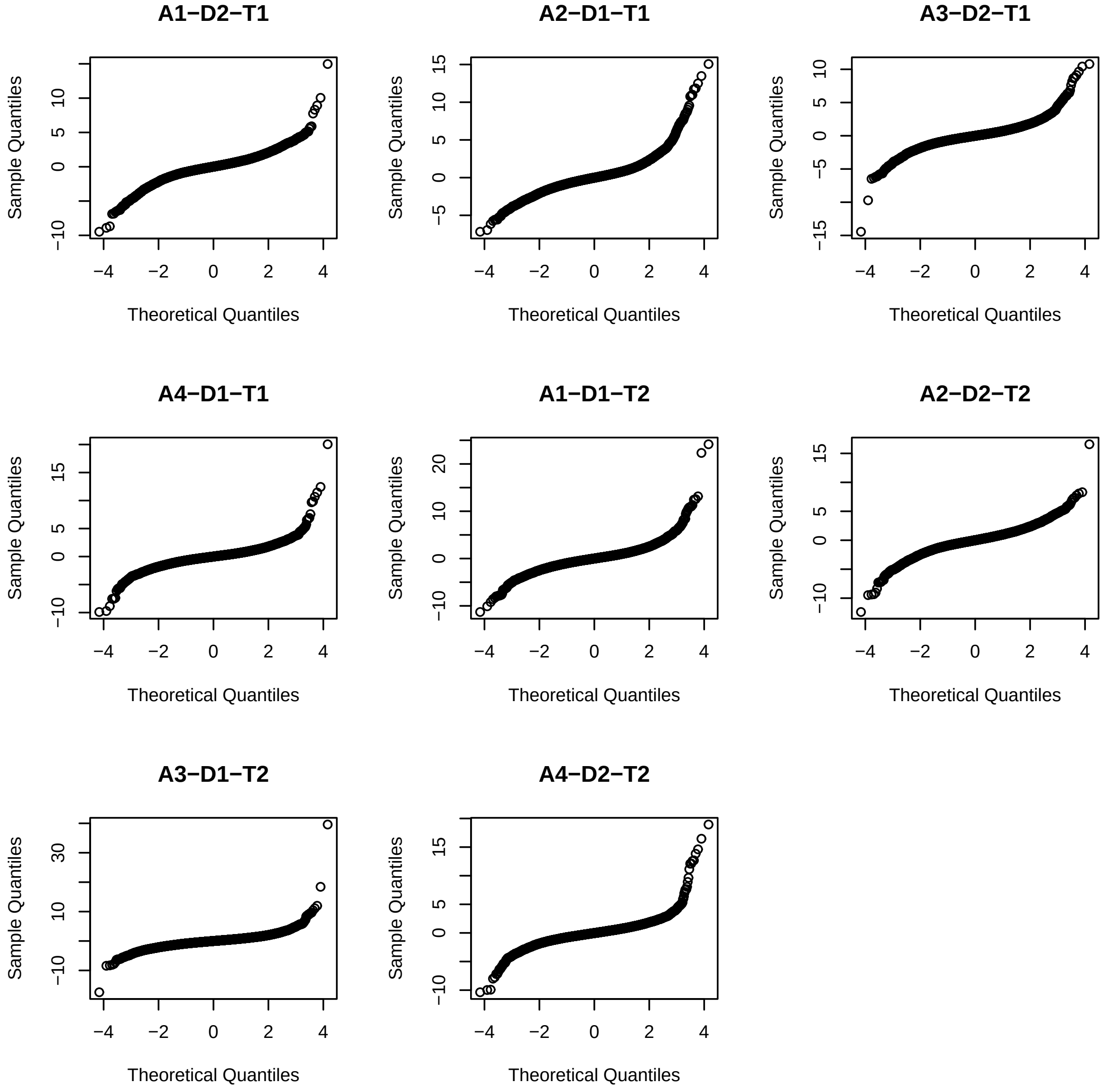


Figure 4.4: Quantile-quantile plots for the DBLFLIP dataset, using MAANOVA

Parameters	Dye	Array 1	Array 2	Array 3	Treatment
Real parameters a	63.78	1.88	4.91	3.30	0.94
Estimated parameters	52.28	1.87	4.72	3.22	0.93
Real parameters b	63.28	1.41	4.42	2.81	230.63
Estimated parameters	51.64	1.40	4.24	2.75	228.91

Table 4.7: Real and estimated hyperparameters a's and b's

Design (Dye,Array,Treat)	2,1,1	1,1,1	2,2,1	1,2,1	1,3,2	2,3,2	1,3,2	2,4,2
Real γ	98.00	58.54	62.70	87.21	91.73	60.69	66.69	73.33
Estimated γ's	97.97	58.79	61.61	88.21	90.83	60.25	65.80	74.10

Table 4.8: Real and estimated hyperparameters γ's

of our proposed method, in terms of error rates, when the parameters are estimated using the naive approach described in Section 4.2.3. It also allows us to confirm the results from Theorem 4.1 regarding the uniformity of the probabilities p_g under the null hypothesis. Data were simulated according to the multiplicative model described in this chapter. The parameters chosen correspond to the ones obtained from the DBLFLIP dataset (shown in the previous section, in Tables 4.4 and 4.5). Note also that data were simulated according to the same design as the DBLFLIP dataset, with 2 dyes, 2 treatments and 4 arrays. We simulated $n = 15000$ genes, from which 5% are differentially expressed. In practice, we obtained 726 truly differentially expressed genes.

Regarding the first stage model, the true and estimated parameters are presented in Tables 4.7 and 4.8. We can see that the estimated parameters are very well estimated for both (a, b)'s and γ's. The only important difference between the real and estimated parameters occurs when the parameters (a, b) are estimated for the dye effect. One reason may be the very small variance of the naive dye effect, since in this dataset, gene-dye effects are almost absent.

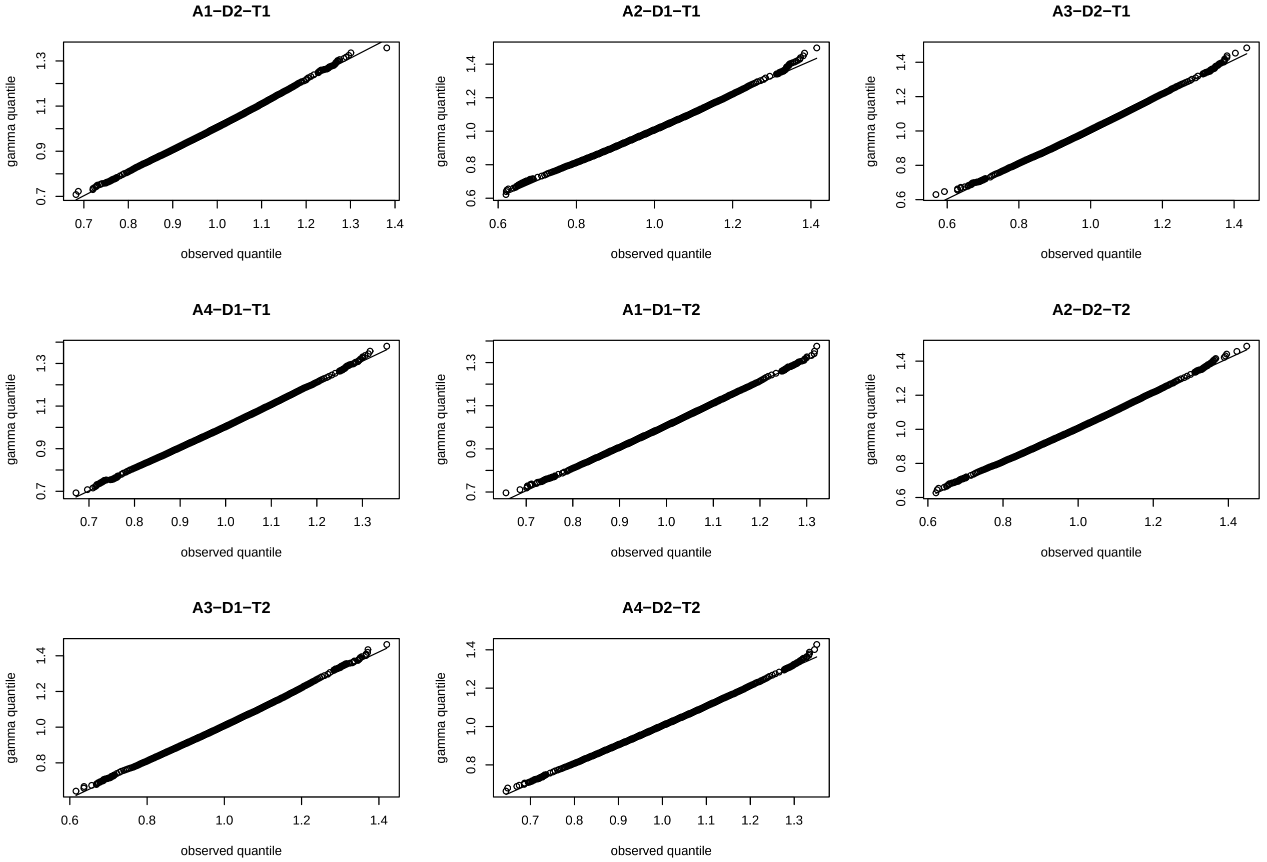


Figure 4.5: Quantile-quantile plots for the simulated dataset, using the multiplicative model

FDR level	# detected (true = 726)	FDR	FNR	SENS	SPEC	RISK
1%	683	0.033	0.0046	0.909	0.998	0.0059
5%	700	0.051	0.0043	0.914	0.997	0.0065

Table 4.9: Error rates

The quantile-quantile plots of the residuals are presented is Figure 4.5. Of course, as expected, we observe an almost perfect straight line, which indicates that those types of plots, using the posterior expectation of the gene-specific effects, are a good indicator of the goodness of fit.

Applying the Benjamini-Hochberg method, adapted to the posterior probabilities p_g, we can detect the differentially expressed genes. Finally the error rates observed (FDR, FNR, SENS and SPEC, defined in Section 3.4) are presented in Table 4.9. We can see that the results obtained are excellent. At a level of FDR control of 5%, the observed FDR is almost equal to the expected one. At a level of 1%, the observed FDR is greater than the observed one, but we expect that if several datasets were simulated, the average observed FDR would be very close to the expected one. The level of FNR and RISK is also very small, and the sensitivity is particularly high.

Finally, the histogram of the posterior probabilities p_g used, under the null hypothesis, is presented in Figure 4.6. As stated in Theorem 4.1, they are uniformly distributed on the interval $[0, 1]$.

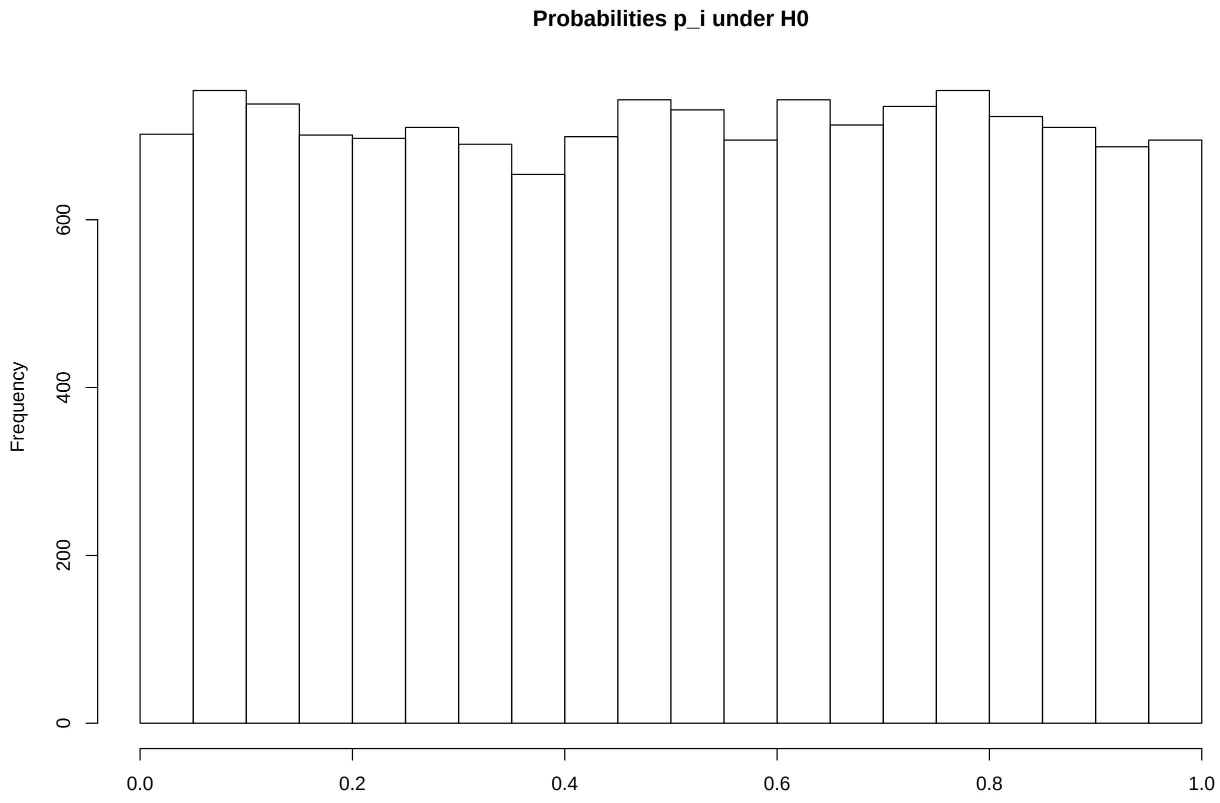


Figure 4.6: Histogram of the posterior probabilities p_g under the null hypothesis

4.5 Discussion

We have proposed in this chapter a new way of defining ANOVA models for microarray data. The first particularity of this approach is its multiplicative setting, which allows us to work with the data (normalized) on the raw scale, assuming a gamma distribution. In the traditional ANOVA models for microarray data, if the effects are all assumed to be fixed, no assumption about the distribution of the residuals is usually made. However, when one or more of the effects are assumed to be random (mixed models), the normal distribution is often the most appropriate choice (in an additive setting). Assuming that the effects of the microarray experiment act in a multiplicative way allows us to use another kind of distribution, and especially the gamma distribution, which has proven to be very efficient in modeling gene expression data. Furthermore, the study of the DBLFLIP dataset illustrated the fact that gamma models may be more appropriate than log-normal ones in some situations.

In the usual ANOVA mixed models for microarray experiments, the Array/Gene interactions are considered to be random. The model we propose considers all the gene interactions to be random. We actually showed that it is possible to construct a Bayesian conditional conjugate model to describe the behavior of the data, for each gene individually, in function of the effects of interest. This model can be seen as a generalization of the conjugate model presented in the second chapter, and the theorems presented in the third chapter can be generalized as well. We can then test for differentially expressed genes using some FDR-controlled procedures, in the context of this gamma Bayesian ANOVA model.

Chapter 5

Probability of half space using inverse Gaussian models

As we mentioned previously, the use of hierarchical Bayesian models, in the analysis of gene expression data, has proven their effectiveness, especially when one wants to make inference on a given gene. Up to now, the gamma/gamma and normal/normal conjugate models have been used successfully in the literature (*cf* Newton et al. (2001), Kendziorski et al. (2002), Newton et al. (2003), Smyth (2004) and Ibrahim et al. (2002)). We show here how the inverse Gaussian distribution can provide a good alternative to these models, and how the multiple testing framework developed in Chapter 2 (Theorem 2.2) can be applied successfully.

In the first section of this chapter, we develop three types of hierarchical inverse Gaussian models: a conjugate model, a non-informative model and a non-conjugate model. A goodness of fit study on four real microarray datasets is performed in the second section, as well as a simulation study of the robustness of these models to the true distribution of the data. The multiple testing procedure using the posterior probability of half-space is finally applied in Section 3, and error rates are compared between the inverse Gaussian and the gamma models.

5.1 Hierarchical models involving the inverse Gaussian distribution

The inverse Gaussian distribution has been used in many different areas of statistics and has proven its usefulness as a model to describe positively skewed data. However, most of the applications are based on the idea of a first passage time for an underlying process (see Chhikara (1986) for a standard source regarding the inverse Gaussian distribution). Because it tends to fit ratios of positive random variables well, a natural extension of the applications of this distribution would be to use it as a way to describe gene expression data. We believe that the wide variety of shapes generated by its probability density function makes it a good competitor to the gamma or log-normal distributions that are typically used with such data. In this section, we propose three models involving the inverse Gaussian distribution. The first one has the computational advantage of being a conjugate model, but we will see that it is actually quite restrictive with respect to the shape of the shape of the distribution of the means of $\underset{\sim}{X}$ and $\underset{\sim}{Y}$. The second model is non-informative and does not require the estimation of any hyper-parameters. Finally, the last model, first proposed by Betro and Rotondi (1991), is expected to provide a good fit to the data, but is numerically more complex than the other two.

5.1.1 The inverse Gaussian distribution : an overview

Let X be an inverse Gaussian random variable with parameters μ and λ, denoted by $X \sim IG(\mu, \lambda)$. The probability density function of X is given by

$$f_X(x; \mu, \lambda) = \sqrt{\frac{\lambda}{2\pi}} x^{-3/2} \exp\left[-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right], \; x > 0,$$

where μ and λ are both positive parameters. This density, which is seen to be a member of the exponential family, is unimodal and positively skewed. The parameter μ corresponds to the mean of the distribution and λ is a scale parameter. The shape of this distribution depends on both μ and λ and can conveniently be indexed by $\phi = \lambda/\mu$.

As we can see in Figure 5.1, the inverse Gaussian density accommodates a variety of shapes and allows the representation of a wide class of unimodal distributions. It can be shown that when $\phi \to 0$, the distribution is highly skewed, whereas high values of ϕ lead

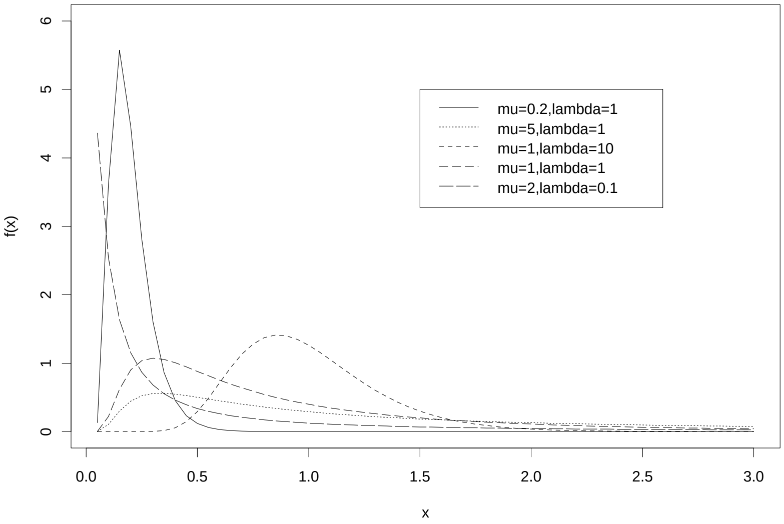


Figure 5.1: Inverse Gaussian densities for five values of (μ, λ)

to a more symmetrical distribution.

The moments of the distribution can be found using the characteristic function, and the mean and the variance are respectively μ and μ^3/λ.

The inverse Gaussian distribution has also many convenient properties for the associated sampling distributions and some are quite similar to those of the normal distribution. One useful property of the normal distribution is shared only partially: a linear combination of inverse Gaussian random variables is not inverse Gaussian, unless some specific conditions on the parameters are satisfied.

Proposition 5.1. *Let $X_1, ..., X_n$ be independent inverse Gaussian random variables with parameters (μ_i, λ_i), for $i = 1, \ldots, n$ and let us consider the random variable $X = \sum_{i=0}^{n} c_i X_i$. If $\lambda_i/(\mu_i^2 c_i) = \xi$ for all i, then*

$$X \sim IG\left(\sum_{i=1}^{n} c_i\mu_i, \xi(\sum_{i=1}^{n} c_i\mu_i)^2\right).$$

Therefore, if $X_1, \ldots, X_n$ is a random sample such that $X_i \sim IG(\mu, \lambda)$ for $i = 1, \ldots, n$, the sample mean $\bar{X}$ has also an inverse Gaussian distribution with parameters μ and $n\lambda$.

5.1.2 The natural conjugate model

The type of parametrization has a significant impact on the choice of the prior when the inverse Gaussian distribution is used under a Bayesian framework. With the parametrization (μ, λ) seen in Section 5.1.1, a natural conjugate prior does not exist, unless μ is known and λ is unknown (Palmer 1973). However, the parametrization $(1/\mu, \lambda)$ provides a natural conjugate model that is mathematically tractable.

Consider the following hierarchical framework, under the same mixture defined in Chapter 2, where the new parametrization implies $E[X_{ij}] = 1/\theta_{xi}$ and $E[Y_{ij}] = 1/\theta_{yi}$: Under the alternative hypothesis H_{1i}, which occurs with probability p

$$\begin{aligned} X_{ij}|\theta_{xi}, \lambda_{xi} &\sim IG(\frac{1}{\theta_{xi}}, \lambda_{xi}), & Y_{ij}|\theta_{yi}, \lambda_{yi} &\sim IG(\frac{1}{\theta_{yi}}, \lambda_{yi}), \\ \theta_{xi}|\lambda_{xi} &\sim \mathcal{N}_0\left(\frac{1}{\beta_x}, \frac{1}{r_x\beta_x\lambda_{xi}}\right), & \theta_{yi}|\lambda_{yi} &\sim \mathcal{N}_0\left(\frac{1}{\beta_y}, \frac{1}{r_y\beta_y\lambda_{yi}}\right), \\ \lambda_{xi} &\sim G^*\left(\frac{r_x\alpha_x}{2}, \frac{r_x - 1}{2}, \frac{r_x}{\beta_x}\right) & \lambda_{yi} &\sim G^*\left(\frac{r_y\alpha_y}{2}, \frac{r_y - 1}{2}, \frac{r_y}{\beta_y}\right), \end{aligned} \tag{5.1}$$

Under the null hypothesis H_{0i}, which occurs with probability $(1 - p)$,

$$\begin{aligned} X_{ij}|\theta_i, \lambda_i &\sim IG(\frac{1}{\theta_i}, \lambda_i), \\ Y_{ij}|\theta_i, \lambda_i &\sim IG(\frac{1}{\theta_i}, \lambda_i), \\ \theta_i|\lambda_i &\sim \mathcal{N}_0\left(\frac{1}{\beta}, \frac{1}{r\beta\lambda_i}\right), \\ \lambda_i &\sim G^*\left(\frac{r\alpha}{2}, \frac{r-1}{2}, \frac{r}{\beta}\right), \end{aligned}$$

for $i = 1, \dots, n$ and $j = 1, \dots, m$. $\mathcal{N}_0$ denotes the normal distribution truncated at 0 and G^* denotes the modified gamma distribution. We use the notation $Y \sim G^*(a, b, c)$ to represent a random variable with probability density function

$$g(y) = \frac{a^b \phi(\sqrt{cy})}{\Gamma(b) F_{2b}(\sqrt{a/bc})} y^{b-1} e^{-ay} \text{ , } y > 0,$$

where $\phi(.)$ stands for the standard normal distribution function , $\Gamma(.)$ is the gamma function, $F_k(.)$ is the distribution function of a Student random variable with k degrees of freedom, and the parameters a, b and c are positive.

This model being conjugate, the posterior and prior distributions have the same form. We can show that

$$\begin{aligned}
\theta_{xi}|\underset{\sim}{x_i}, H_{1i} &\sim \mathcal{T}_0\left(r'_x - 1, \frac{1}{\beta'_{xi}}, \frac{\alpha'_{xi}}{(r'_x - 1)\beta'_{xi}}\right), \\
\theta_{yi}|\underset{\sim}{y_i}, H_{1i} &\sim \mathcal{T}_0\left(r'_y - 1, \frac{1}{\beta'_{yi}}, \frac{\alpha'_{yi}}{(r'_y - 1)\beta'_{yi}}\right), \qquad (5.2)\\
\theta_i|\underset{\sim}{x_i}, \underset{\sim}{y_i}, H_{0i} &\sim \mathcal{T}_0\left(r' - 1, \frac{1}{\beta'_i}, \frac{\alpha'_i}{(r' - 1)\beta'_i}\right),
\end{aligned}$$

where $\mathcal{T}_0(a, b, c^2)$ represents a Student distribution truncated at 0, with a degrees of freedom, location parameter b and scale parameter c^2. In other words, if $Z \sim \mathcal{T}_0(a, b, c^2)$, the variable $(Z - b)/c$ follows a Student distribution with a degrees of freedom, truncated at $-b/c$. The parameters of the posterior distribution under the alternative hypothesis H_{1i} are

$$\begin{aligned}
r'_x &= m + r_x, \\
\beta'_{xi} &= \frac{(m\bar{x}_i + r_x\beta_x)}{r'_x}, \qquad (5.3)\\
\alpha'_{xi} &= \left[m u_{xi} + r_x\alpha_x + \frac{m}{\bar{x}_i} + \frac{r_x}{\beta_x} - \frac{r'_x}{\beta'_x}\right] / r'_x,
\end{aligned}$$

where

$$\begin{aligned}
\bar{x}_i &= \frac{1}{m}\sum_{j=1}^{m} x_{ij}, \\
u_{xi} &= \frac{1}{m}\sum_{j=1}^{m} \frac{1}{x_{ij}} - \frac{1}{\bar{x}_i},
\end{aligned}$$

for $i = 1, \ldots, n$. The parameters r'_y, β'_{yi}, α'_{yi} are defined in a similar way. Finally, the parameters of the posterior distribution under the alternative hypothesis H_{1i} are:

$$\begin{aligned} r' &= 2m + r, \\ \beta'_i &= \frac{(m\bar{x}_i + m\bar{y}_i + r\beta)}{r'}, \\ \alpha'_i &= \left[mu_{xi} + mu_{yi} + r\alpha + \frac{m}{\bar{x}_i} + \frac{m}{\bar{y}_i} + \frac{r}{\beta} - \frac{r'}{\beta'} \right] / r'. \end{aligned} \tag{5.4}$$

In order to compute the odds of differential expression for each gene, we need to compute the marginal densities under the null and alternative hypothesis. We find $m_1(\underset{\sim}{x_i}, \underset{\sim}{y_i}) = m_1(\underset{\sim}{x_i}) m_1(\underset{\sim}{y_i})$, where

$$\begin{aligned} m_1(\underset{\sim}{x_i}) &= \left(\frac{1}{2\pi}\right)^{m/2} \left(\prod_{j=1}^{m} x_{ij}\right)^{-3/2} \frac{(\beta_x/\alpha_x)^{1/2}}{(\beta'_{xi}/\alpha'_{xi})^{1/2}} \frac{(r_x\alpha_x/2)^{r_x/2}}{(r'_x\alpha'_{xi}/2)^{r'_x/2}} \frac{\Gamma((r'_x - 1)/2)}{\Gamma((r_x - 1)/2)} \\ &\quad \times \frac{F_{r'_x - 1}(\sqrt{\alpha'_{xi}\beta'_{xi}/(r'_x - 1)})}{F_{r_x - 1}(\sqrt{\alpha_x\beta_x/(r_x - 1)})}, \end{aligned}$$

for $i = 1, \ldots, n$. The function $m_1(\underset{\sim}{y_i})$ has a similar expression. Under the null hypothesis, we find

$$\begin{aligned} m_0(\underset{\sim}{x_i}, \underset{\sim}{y_i}) &= \left(\frac{1}{2\pi}\right)^{m} \left(\prod_{j=1}^{m} x_{ij} y_{ij}\right)^{-3/2} \frac{(\beta/\alpha)^{1/2}}{(\beta'_i/\alpha'_i)^{1/2}} \frac{(r\alpha/2)^{r/2}}{(r'\alpha'_i/2)^{r'/2}} \frac{\Gamma((r' - 1)/2)}{\Gamma((r - 1)/2)} \\ &\quad \times \frac{F_{r' - 1}(\sqrt{\alpha'_i\beta'_i/(r' - 1)})}{F_{r - 1}(\sqrt{\alpha\beta/(r - 1)})}. \end{aligned}$$

5.1.3 The non-informative model

As we have introduced in Section 5.1.1, when no prior information on the data is available, one may use a non-informative prior. A popular choice of non-informative prior is the Jeffreys' prior, based on the Fisher information matrix. As is often the case, with the inverse Gaussian distribution, the posterior distribution induced by Jeffreys' prior is not a proper distribution. By adapting the model developed by Banerjee & Bhattacharyya (1979), using the locally uniform reference prior, to the context of a mixture of models, we obtain:

Under H_{1i}, which occurs with probability p,

$$\begin{aligned} X_{ij}|\theta_{xi},\lambda_{xi} &\sim IG(\frac{1}{\theta_{xi}},\lambda_{xi}), & Y_{ij}|\theta_{yi},\lambda_{yi} &\sim IG(\frac{1}{\theta_{yi}},\lambda_{yi}), \\ \pi(\theta_{xi},\lambda_{xi}) &\propto \lambda_{xi}^{-1}. & \pi(\theta_{yi},\lambda_{yi}) &\propto \lambda_{yi}^{-1}. \end{aligned} \tag{5.5}$$

Under H_{0i}, which occurs with probability $(1-p)$,

$$\begin{aligned} X_{ij}|\theta_{i},\lambda_{i} &\sim IG(\frac{1}{\theta_{i}},\lambda_{i}), \\ Y_{ij}|\theta_{i},\lambda_{i} &\sim IG(\frac{1}{\theta_{i}},\lambda_{i}), \\ \pi(\theta_{i},\lambda_{i}) &\propto \lambda_{i}^{-1}. \end{aligned} \tag{5.6}$$

This type of prior is commonly used in Bayesian analysis for location-scale parameters. Although this prior is improper, it leads to the following proper posterior densities:

$$\begin{aligned}
\theta_{xi}|\underset{\sim}{x_i}, H_{1i} &\sim \mathcal{T}_0\left(m-1, \frac{1}{\bar{x}_i}, \frac{u_{xi}}{(m-1)\bar{x}_i}\right), \\
\theta_{yi}|\underset{\sim}{y_i}, H_{1i} &\sim \mathcal{T}_0\left(m-1, \frac{1}{\bar{y}_i}, \frac{u_{yi}}{(m-1)\bar{y}_i}\right), \qquad (5.7)\\
\theta_{i}|\underset{\sim}{x_i}, \underset{\sim}{y_i}, H_{0i} &\sim \mathcal{T}_0\left(r'-1, \frac{1}{\beta'_i}, \frac{\alpha'_i}{(r'-1)\beta'_i}\right),
\end{aligned}$$

where

$$\begin{aligned}
r' &= 2m, \\
\beta'_i &= \frac{(m\bar{x}_i + m\bar{y}_i)}{r'}, \qquad (5.8)\\
\alpha'_i &= \left[m u_{xi} + m u_{yi} + \frac{m}{\bar{x}_i} + \frac{m}{\bar{y}_i} - \frac{r'}{\beta'}\right]/r'.
\end{aligned}$$

We note that these posteriors have the same form as the natural conjugate model and are identical to the extreme case $r \to 0$, which corresponds to the case of an infinite variance of the gene expression under both treatments. Finally, although the use of an improper prior distribution on the parameter yields a proper posterior distribution, the marginal densities remain improper and cannot be computed.

5.1.4 Another hierarchical model involving the inverse Gaussian distribution

As we have seen earlier, the main disadvantage of the natural conjugate model is that it uses a symmetric distribution for the reciprocal of the mean of highly skewed data.

Betro & Rotondi (1991) introduced a model that does not impose such a constraint: the prior for the mean is taken to be inverse Gaussian as well, which brings more flexibility to the model. Note that this model uses the parametrization (μ, ϕ), where $\phi = \mu/\lambda$ is the shape parameter of the inverse Gaussian distribution. This allows the hierarchy of the model to focus directly on the scale and the shape parameter separately. In terms of this parametrization, the density of an inverse Gaussian random variable can be written as

$$f(x|\mu,\phi) = \left(\frac{\phi\mu}{2\pi}\right)^{1/2} x^{-3/2} \exp\left\{\phi - \frac{\phi}{2}\left(\frac{x}{2} + \frac{\mu}{2}\right)\right\} \ , x > 0.$$

Consider the following model under the alternative H_{1i} (with probability p):

$$\begin{aligned} X_{ij}|\theta_{xi},\phi_{xi} &\sim IG(\theta_{xi},\phi_{xi}), & Y_{ij}|\theta_{yi},\phi_{yi} &\sim IG(\theta_{yi},\phi_{yi}), \\ \theta_{xi}|\phi_{xi} &\sim IG(\eta_x, w_x\phi_{xi}), & \theta_{yi}|\phi_{yi} &\sim IG(\eta_y, w_y\phi_{yi}), \\ \phi_{xi} &\sim G(a_x,b_x). & \phi_{yi} &\sim G(a_y,b_y). \end{aligned} \tag{5.9}$$

Under the null hypothesis H_{0i} (with probability $(1-p)$), we consider

$$\begin{aligned} X_{ij}|\theta_i,\phi_i &\sim IG(\theta_i,\phi_i), \\ Y_{ij}|\theta_i,\phi_i &\sim IG(\theta_i,\phi_i), \\ \theta_i|\phi_i &\sim IG(\eta, w\phi_i), \\ \phi_i &\sim G(a,b). \end{aligned} \tag{5.10}$$

This model requiring the computation of the marginal density $m(\underset{\sim}{x},\underset{\sim}{y})$ is mathematically more complex. However, it is possible to find an expression for the posterior distribution of θ and we get:

$$\begin{aligned}
\pi(\theta_{xi}|\underset{\sim}{x_i}, H_{1i}) &= \frac{\theta_{xi}^{m+a_x-1}}{[v_{1xi}\theta_{xi}^2 - 2v_{2xi}\theta_{xi} + v_{3xi}]^{(m+2a_x+1)/2}}\ I_{xi}^{-1}, && (5.11)\\
\pi(\theta_{yi}|\underset{\sim}{y_i}, H_{1i}) &= \frac{\theta_{yi}^{m+a_y-1}}{\left[v_{1yi}\theta_{yi}^2 - 2v_{2yi}\theta_{yi} + v_{3yi}\right]^{(m+2a_y+1)/2}}\ I_{yi}^{-1}, && (5.12)\\
\pi(\theta_i|\underset{\sim}{x_i}, \underset{\sim}{y_i}, H_{0i}) &= \frac{\theta_i^{2m+a-1}}{[v_{1i}\theta_i^2 - 2v_{2i}\theta_i + v_{3i}]^{(m+a+1/2)}}\ I_i^{-1}, && (5.13)
\end{aligned}$$

where

$$\begin{aligned}
v_{1xi} &= m\bar{x}_{ri} + w_x\eta_x^{-1},\\
v_{2xi} &= m + w_x - b_x,\\
v_{3xi} &= m\bar{x}_i + w_x\eta_x,\\
v_{1i} &= m\bar{x}_{ri} + m\bar{y}_{ri} + w\eta^{-1},\\
v_{2i} &= 2m + w - b,\\
v_{3i} &= m\bar{x}_i + m\bar{y}_i + w\eta,\\
\bar{x}_{ri} &= \left(\sum_{j=1}^{m} \frac{1}{x_{ij}}\right)/m,
\end{aligned}$$

and

$$\begin{aligned}
I_{xi} &= \int_0^{+\infty} \frac{t^{m+a_x-1}}{(v_{1xi}t^2 - 2v_{2xi}t + v_{3xi})^{(m+2a_x+1)/2}} dt, && (5.14)\\
I_i &= \int_0^{+\infty} \frac{t^{2m+a-1}}{(v_{1i}t^2 - 2v_{2i}t + v_{3i})^{(m+a+1/2)}} dt. && (5.15)
\end{aligned}$$

The parameters v_{1yi}, v_{2yi}, v_{3yi}, I_{yi} are defined similarly as those for $\underset{\sim}{x_i}$. Betro & Rotondi (1991) showed that the integrals I_{xi}, I_{yi} and I_i always exist since a_x, a_y and a are all positive.

If we make the assumption that a_x, a_y and a are integers, the integral I_{xi} can be computed using a double recursion by writing $I_{xi} = I_{xi}(A_x, B_x)$, where $A_x = m + a_x - 1$, $B_x = m + 2a_x$ and (removing the subscript i)

$$I(A,B) = \int_0^{+\infty} \frac{t^A}{(v_1 t^2 - 2v_2 t + v_3)^{(B+1)/2}} dt,$$

where A, $B \in \mathcal{N}$. Clearly, if a_x is an integer, A_x and B_x are both integers as well. We can show that

$$I(A,B) = \frac{v_1}{v_2} \frac{(B - 2A + 1)}{(B - A)} I(A-1, B) + \frac{(A-1)}{(B-A)} \frac{v_3}{v_1} I(A-2, B).$$

The integral $I(A,B)$ can be computed for any integers $A > 2$ and any integers $B > 0$ if we can find an expression for $I(1,B)$ and $I(0,B)$. Actually, using a recursion on B, we can show that

$$\begin{aligned} I(1,B) &= \frac{1}{v_1(B-1)} \frac{1}{v_3^{(B-1)/2}} + \frac{v_2}{v_1} I(0,B), \\ I(0,B) &= \frac{1}{d(B-1)} \frac{v_2}{v_3^{(B-1)/2}} + \frac{v_1(B-2)}{d(B-1)} I(0, B-2), \end{aligned}$$

where $d = v_1 v_3 - v_2^2$.

To complete the recursive algorithm, we can also show

$$\begin{aligned} I(0,1) &= \begin{cases} \frac{1}{\sqrt{d}}\left(\frac{\pi}{2} + \arctan\left(\frac{v_2}{\sqrt{d}}\right)\right) & \text{if } d > 0 \\ \frac{1}{2\sqrt{|d|}}\log\left(1 - \frac{2\sqrt{|d|}}{v_2 + \sqrt{|d|}}\right) & \text{if } d < 0. \end{cases} \\ I(0,2) &= \frac{1}{d}\left(v_1^{1/2} + \frac{v_2}{v_3^{1/2}}\right). \end{aligned}$$

The integrals I_{yi} and I_i will be computed in a similar way. We will see in the last section of this chapter that the assumption that a_x, a_y and a are integers is not very strong and in practice, it does not affect significantly the results.

Finally, the predictive densities for this model are given by

$$\begin{aligned} m_1(\underset{\sim}{x_i}) &= \frac{b^a}{(2\pi)^{m/2}}(\prod_{j=1}^{m} x_{ij})^{-3/2}\left(\frac{\eta w}{2\pi}\right)^{1/2} 2^{m/2+1/2+a}\frac{\Gamma(m/2+1/2+a)}{\Gamma(a)} I_{xi}, \\ m_0(\underset{\sim}{x_i}, \underset{\sim}{y_i}) &= \frac{b^a}{(2\pi)^{m}}(\prod_{j=1}^{m} x_{ij}y_{ij})^{-3/2}\left(\frac{\eta w}{2\pi}\right)^{1/2} 2^{m+1/2+a}\frac{\Gamma(m+1/2+a)}{\Gamma(a)} I_{i}. \end{aligned}$$

The marginal $m_1(\underset{\sim}{y_i})$ is computed in the same way as $m_1(\underset{\sim}{x_i})$.

As we have seen, this model seems to be much more flexible than the conjugate one. Even if it is mathematically more complex, by assuming a_x, a_y and a are integers, we are able to find simple expressions for the posterior distributions. Furthermore, we note that in practice, assuming a integer does not seem to have any important impact on the fit this model provides to the data, or on the list of genes that are detected as differentially expressed.

5.1.5 Estimation of the hyper-parameters and tests for differentially expressed genes

So far, three hierarchical models involving the inverse Gaussian distribution have been proposed. For the first and last models, the last level of the hierarchy involves some hyperparameters that need to be estimated. The second model, being non-informative, does not involve any hyperparameters. For the natural conjugate model, the hyperparameters are $(p, r_x, \alpha_x, \beta_x, r_y, \alpha_y, \beta_y, r, \alpha, \beta)$. In the case of the non-conjugate model, they are $(p, \eta_x, w_x, a_x, b_x, \eta_y, w_y, a_y, b_y, \eta, w, a, b)$. As we did for the gamma model in Chapter 3, these hyperparameters can be estimated using an empirical Bayes approach by maximizing the marginal function $m(\underset{\sim}{x}, \underset{\sim}{y})$. In practice, a numerical algorithm such as Newton-Raphson is required to obtain these parameters. Note that in the next two sections, we use

$$\begin{gathered} r_x = r_y = r, \alpha_x = \alpha_y = \alpha, \beta_x = \beta_y = \beta, \qquad (5.16) \\ \eta_x = \eta_y = \eta, w_x = w_y = w, a_x = a_y = a, b_x = b_y = b, \end{gathered}$$

in order to simplify the computation of the parameters. However, different parameters for X and Y were also tried, and no loss in terms of fit or in terms of error rates was observed by assuming (5.16).

As we did in Chapter 3, the posterior probability of the one-sided alternative hypothesis was computed for each gene. Note that due to the re-parametrization in the conjugate and non-informative models, the one-sided alternative hypothesis is written as $H'_{1i} : 1/\theta_{xi} > 1/\theta_{yi}$ in these two cases. From the results of Theorem 2.2, the posterior probability p_i obtained is asymptotically uniformly distributed on the interval $[0, 1]$ under the null hypothesis, when m is large enough. It follows that this posterior probability can also be applied in the Benjamini-Hochberg procedure, for example. Note that the exact procedure used in the next sections was described in Section 2.4.

5.2 Goodness of fit study

Before applying the multiple testing procedure developed in the second chapter to the inverse Gaussian models presented in this chapter, we believe that it is appropriate to first perform a goodness of fit study, and compare it with the conjugate gamma model used previously.

5.2.1 Application to four microarray datasets

In this section, we apply the two informative inverse Gaussian models described in Section 5.1.2 and 5.1.4 to four microarray datasets. These datasets have been described and used in Chapter 3, Section 3.3. As we did for the gamma model, the histograms of the data (on the log scale) were plotted against the predictive density functions (on the log scale) for each dataset. The predictive density function of the non-informative inverse Gaussian model being non-proper, this model is not used here.

For a better comparison between the gamma model and the two inverse Gaussian models, the fit for the three models is presented for each set of data in Figures 5.2, 5.3, 5.4, 5.5. Throughout this section, the models are labeled IG/Norm for the conjugate inverse Gaussian and IG/IG for the last inverse Gaussian model presented. The hyperparameters of the two inverse Gaussian models used here are also presented in Table 5.1. We note the good agreement between the gamma model presented in Chapter 3 (see Table 3.2) and

Dataset	IG/IG/G	IG/Norm/G
TCDD	$\eta = 4184.06$ $w = 0.027$ $a = 8,\ b = 0.62$ $p = 0.018$	$\alpha = 7.5 \times 10^{-6}$ $\beta = 3251.8$ $r = 1.397$ $p = 0.00108$
COLD	$\eta = 5912.18$ $w = 0.036$ $a = 2,\ b = 0.218$ $p = 0.159$	$\alpha = 4.5 \times 10^{-6}$ $\beta = 3296.1$ $r = 1.25$ $p = 0.082$
DBLFLIP	$\eta = 832.96$ $w = 0.029$ $a = 2,\ b = 0.1144$ $p = 3.5 \times 10^{-4}$	$\alpha = 3.6 \times 10^{-5}$ $\beta = 412.7$ $r = 1.34$ $p = 5.57 \times 10^{-8}$
FIELD	$\eta = 7274.8$ $w = 0.0435$ $a = 2,\ b = 0.222$ $p = 0.250$	$\alpha = 6.4 \times 10^{-6}$ $\beta = 3693.8$ $r = 1.35$ $p = 0.125$

Table 5.1: Hyperparameters for the 2 informative inverse Gaussian models

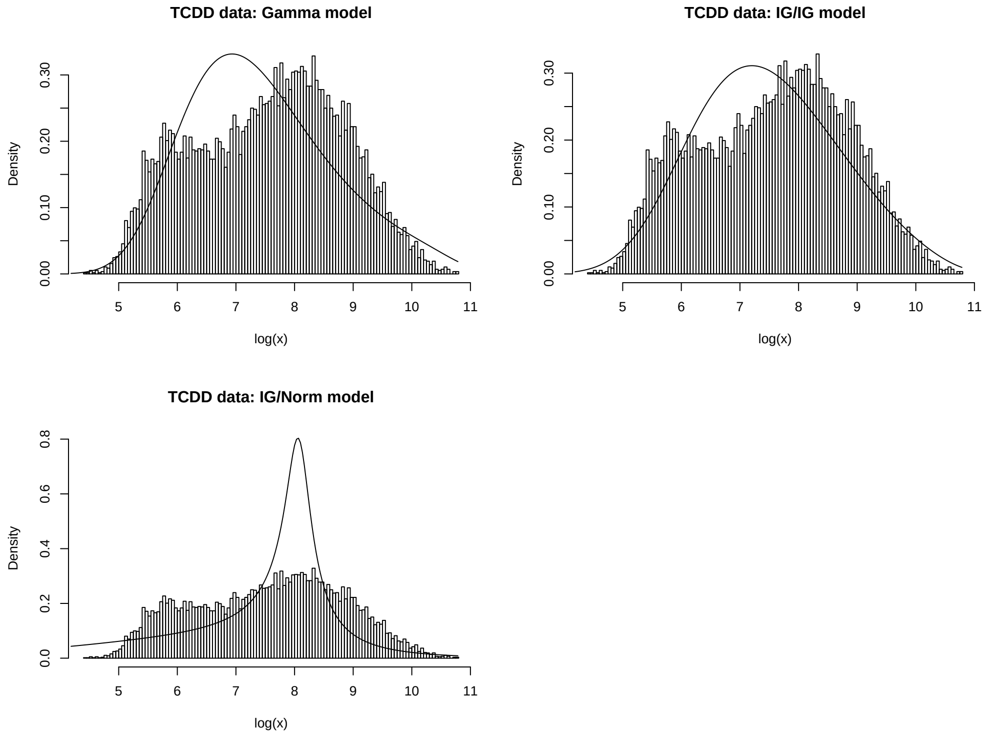


Figure 5.2: Fit for the 3 informative models: TCDD dataset

the IG/IG model, regarding the estimation of the parameter p. However, the fact that the IG/Norm model provides estimates of p that are very different from the other two models may be due to a lack of fit. By looking at the histograms, we note the very poor fit provided by the conjugate model, to the four datasets. This is probably due to the restriction imposed by this model, especially regarding the fact that the mean of the data is assumed to be symmetrically distributed (normal distribution). So, even if this model has the advantage of being conjugate and computationally easy to handle, it is not adequate to describe microarray data. Again, this illustrates the importance of the choice of the prior distribution in Bayesian or random effects models. However, we note that the fit provided by the gamma and the IG/IG model is in general not bad. Note that

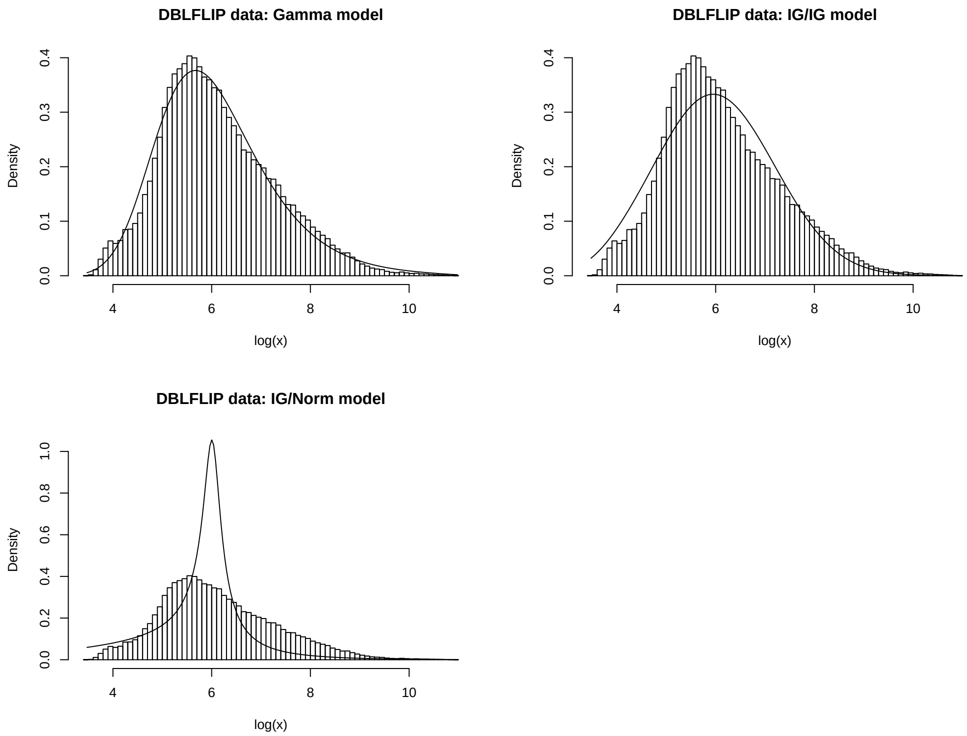


Figure 5.3: Fit for the 3 informative models: DBLFLIP dataset

the performance of these two models strongly depends on the characteristics of the data, and neither of the two models is better than the other for all four sets of data. However, regarding the DBLFLIP dataset, it is obvious that the gamma model provides a better fit. For the FIELD data, the opposite is true, with an excellent fit provided by the IG/IG model. In order to compare the fit provided by these models with the fit obtained if the model was true, we simulated two datasets according to the GAM and IG/IG models. The predictive distribution of each model was then plotted against the histogram of the data. This plot is presented in Figure 5.6.

In order to compare the fit provided by the models to the different datasets mathematically, three statistics were used, derived from three distribution-free tests of goodness of

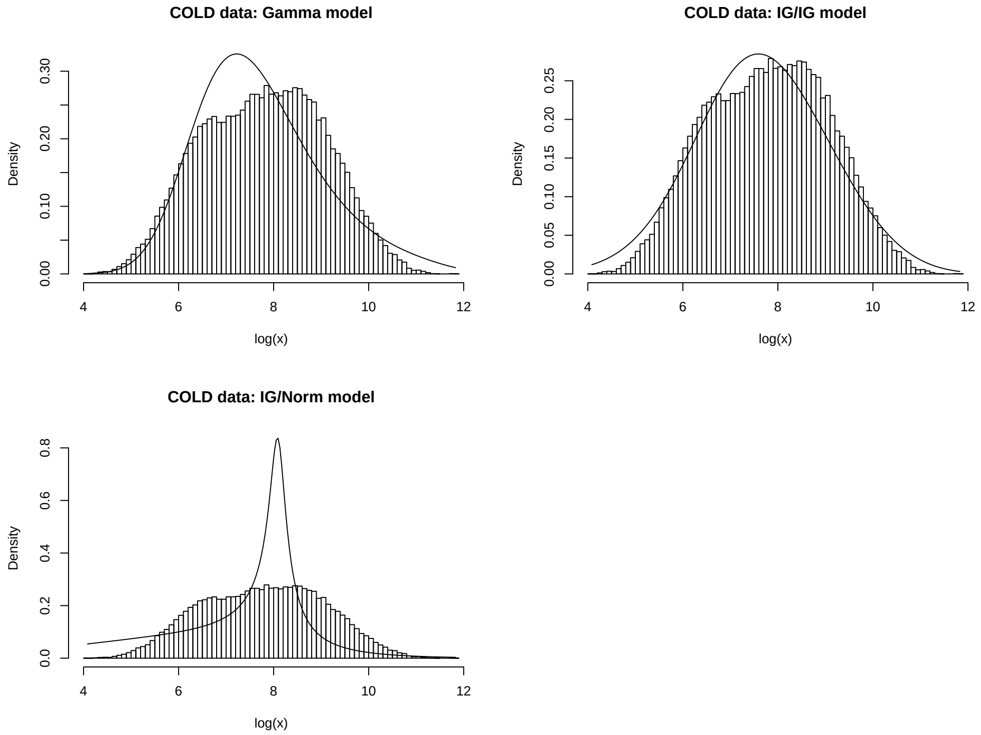


Figure 5.4: Fit for the 3 informative models: COLD dataset

fit. For instance, we consider a random variable X with unknown continuous distribution function $F_X(x)$ and we want to test the null hypothesis $H_0 : F_X = F_0$, where F_0 is a completely specified distribution function. In our case, the function F_0 is the distribution function induced by the marginal $m(\underset{\sim}{x})$ for the two models. The three statistics we use here are based on measure of the “distance” between F_0 and the empirical distribution function $F_n(x)$. We note that our interest focuses more on the values of the test statistics (in order to compare the models) rather than in the tests themselves. The three statistics used here are the Kolmogorov-Smirnov statistic, the Cramer-von Mises statistic and the Anderson and Darling statistic. Considering n random variables $X_1, \ldots, X_n$ with distribution $F_X(x)$ and order statistics $X_{(1)}, \ldots, X_{(n)}$, we define the Kolmogorov -Smirnov

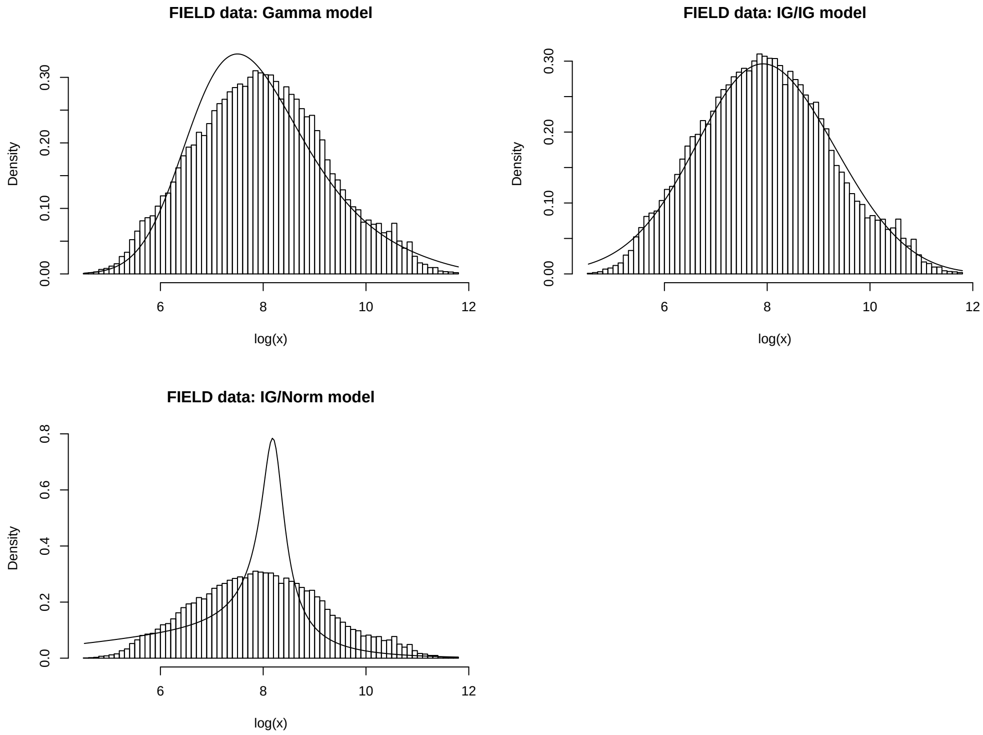


Figure 5.5: Fit for the 3 informative models: FIELD dataset

statistic as

$$
\begin{aligned}
D_n &= Sup_{-\infty<x<+\infty}|F_n(x) - F_0(x)| = max(D_n^+, D_n^-), \\
D_n^+ &= Sup_{-\infty<x<+\infty}F_n(x) - F_0(x) = max_{1\leq i\leq n}\{i/n - F_0(X_{(i)})\}, \\
D_n^- &= Sup_{-\infty<x<+\infty}F_0(x) - F_n(x) = max_{1\leq i\leq n}\{F_0(X_{(i)}) - (i-1)/n)\}.
\end{aligned}
$$

Then, this statistic represents the maximum distance between the F_0 and the empirical distribution function of X. A second type of distance measure, the Cramer-von Mises

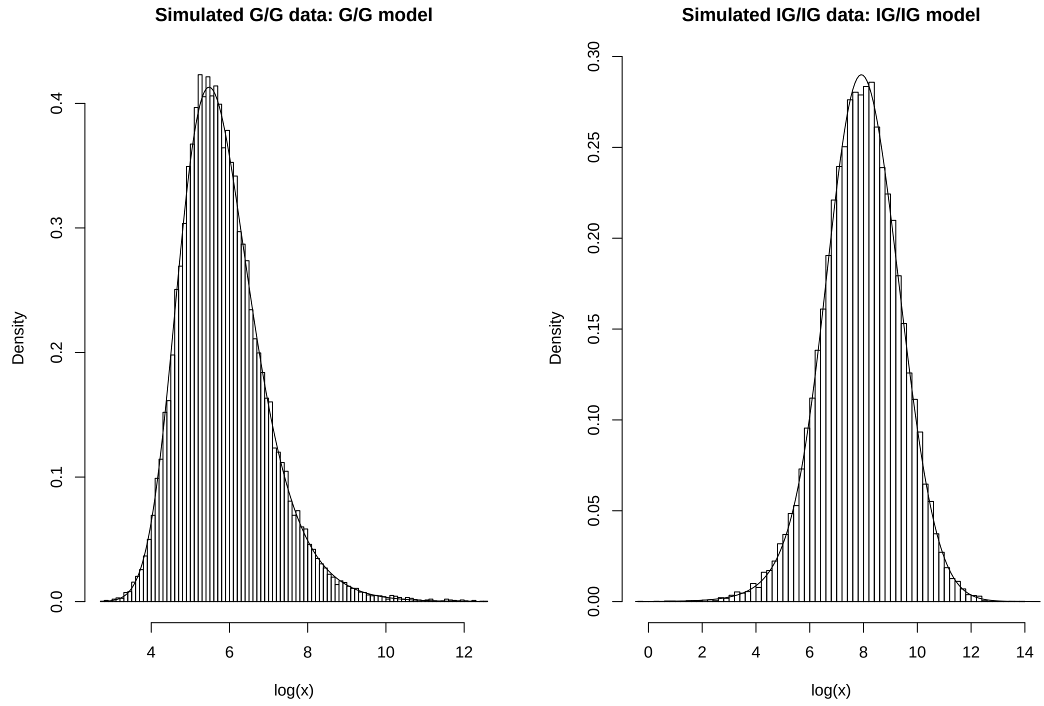


Figure 5.6: Fit for the GAM and IG/IG models: Datasets are simulated according the right model

statistic, is a type of mean square deviation between $F_n(x)$ and $F_0(x)$. It is defined as

$$\begin{aligned} w_n^2 &= \int_{-\infty}^{+\infty} \{F_n(x) - F_0(x)\}^2 dF_0(x), \\ &= \frac{1}{12n^2} + \frac{1}{n}\sum_{i=1}^{n}\left\{F_0(x_{(i)}) - \frac{(2i-1)}{2n}\right\}^2 . \end{aligned}$$

One of the features of the first two statistics is that they are more sensitive to departures in the middle of the range of $F_0(x)$. In our case, the main feature of microarray data is the long tail of the distribution. It would then be interesting to have a measure focusing also on the last part of the range of $F_0(x)$. The Anderson and Darling statistic is a weighted version of the Cramer-von Mises statistic by a non-negative weight function $\psi(u)$. In

		Kolmogorov	Cramer	Anderson
TCDD:	Gamma	0.115	0.004	243.058
	IG/IG	0.074	0.002	107.549
DBLFLIP	Gamma	0.017	7.12×10^{-5}	72.57
	IG/IG	0.050	0.000	539.70
COLD	Gamma	0.088	0.003	1125.78
	IG/IG	0.047	0.000	458.233
FIELD	Gamma	0.147	0.001	291.32
	IG/IG	0.052	4.77×10^{-5}	36.5

Table 5.2: Goodness of fit distances for the Gamma model and for the IG/IG model, for the four datasets

particular, $\psi(u) = n/[u(1-u)]$ is considered here. The statistic is then defined as

$$\begin{aligned} W_n^2 &= n\int_0^1 \left[\frac{F_n(u)-u}{\sqrt{u(1-u)}}\right]^2 du, \\ &= -n - \frac{1}{n}\sum_{i=1}^{n}\left[(2i-1)log(F_0(x_{(i)})) + (2n-2i+1)log(1-F_0(x_{(i)}))\right]. \end{aligned}$$

Regarding the four datasets studied, goodness of fit distances for the IG/IG models and the gamma model are presented in Table 5.2. These distances were not computed for the conjugate inverse Gaussian model, since by looking at the histograms of the fit, it is obvious that this model does not provide an adequate fit to the data. By analyzing the results from Table 5.2, we note that for all the datasets, except for the DBLFLIP, the inverse Gaussian model performs better than the gamma model, with respect to the distances defined, Kolmogorov, Cramer and Anderson. In the case of the three datasets TCDD, COLD and FIELD, we can conclude that the inverse Gaussian model is more

Dataset	Model	Kolmogorov	Cramer	Anderson
Gamma	Gamma	0.01913	0.00114	52.729
Gamma	Inverse Gaussian	0.084	0.00287	178.203
Inverse Gaussian	Gamma	0.1186	0.00548	268.83
Inverse Gaussian	Inverse Gaussian	0.01568	6.52×10^{-5}	3.82

Table 5.3: Robustness to the misspecification of the true distribution of the data, for the non conjugate inverse Gaussian model and for the gamma conjugate model

appropriate than the gamma model to describe the behavior of the gene expressions.

5.2.2 A simulation study: Goodness of fit and robustness

In this section, we study the robustness of the inverse Gaussian models, compared to the gamma model, in terms of goodness of fit. In order to do so, we simulated two sets of 100 datasets each. The first set of data were simulated according to the gamma model, with parameters $\alpha = 9$, $\alpha_0 = 0.8$, $\nu = 91$ and $p = 0.1$, as in Kendziorski et al. (2002). The second set of data was generated according to the non-conjugate inverse Gaussian model defined in Section 5.1.4, with parameters $p = 0.250$, $\eta = 7274.8$, $w = 0.0435$, $a = 2$ and $b = 0.222$ (as estimated in the FIELD dataset). Both models were fitted to the two types of datasets. Note that for the same reasons as before, the non-informative model was not applied here since its goodness of fit cannot be measured using the predictive distribution function. Furthermore, as we have seen in the previous section, the fit provided by the conjugate inverse Gaussian model was very poor, and for this reason, it was not considered here. Table 5.3 presents the average goodness of fit distances (as defined in the previous section) over the two types of 100 datasets. Of course, as expected, the fit provided to the simulated data is always better when the true model is applied, for the three types of distances chosen. However, if we look at the results when the model is misspecified

(Gamma data with IG/IG model, or inverse Gaussian data with the gamma model), we note that the difference between the good specification and the misspecification is much less important when the inverse Gaussian model is applied. These results may indicate that the inverse Gaussian model studied here is the most robust to the misspecification of the model. Note that this is a great advantage when one works with microarray data, since the true distribution of the data is hardly known and subject to many unknown factors.

5.3 Multiple testing using inverse Gaussian models: a simulation study

In this section, we apply the multiple testing procedure described in the second chapter, and reviewed in Section 5.1.5, using posterior probabilities based on the inverse Gaussian models.

5.3.1 Error rates and robustness

Since the results of the previous section showed a very poor fit when the conjugate model is used, only the non-conjugate and non-informative models are used here. In order to study the robustness of these models in terms of error rates, data were generated according to the non-conjugate inverse Gaussian model, and according to the gamma model presented in Chapter 3. Results were then compared with those using the posterior probability computed under the gamma model. Specifically, two sets of 500 datasets were generated. The first set is based on the non-conjugate inverse Gaussian model, with parameters $\eta = 7274$, $w = 0.0435$, $a = 2$, $b = 0.222$, $p = 0.23$, corresponding to those obtained

Dataset	Model	FDR	FNR	SENS	SPEC	RISK
GAMMA	Gamma	0.049	0.066	0.763	0.988	0.063
SIMULATED	IG/IG	0.017	0.079	0.710	0.996	0.069
DATA	IG/NI	0.016	0.094	0.649	0.996	0.082
IG	Gamma	0.1214	0.055	0.809	0.96	0.069
SIMULATED	IG/IG	0.038	0.056	0.802	0.990	0.052
DATA	IG/NI	0.031	0.073	0.73	0.99	0.065

Table 5.4: Error rates for the three models, for each type of dataset. FDR was controlled at a level of 5%.

with the FIELD dataset. The second set of data was also generated according to the parameters obtained by fitting a gamma model to the FIELD data, with $\alpha = 5.309$, $\alpha_0 = 0.9579$, $\nu = 325.83$, $p = 0.23$. In both cases, $n = 1000$ genes were simulated, with $m = 10$ replications each. For each set of data, the three models were fitted: IG/IG, IG/NI (for the non-informative model) and Gamma. For each of them, the average error rates were recorded. Results are presented in Table 5.4. Of course, we expect best results for the model from which data were generated. However, we observe that the error rates provided by the non-conjugate inverse-Gaussian model, when applied to gamma data, are very low, especially for the False Discovery rate. It seems that this model is robust enough to provide a low rate of False Discoveries, without a great loss in False Negative Rate, or in RISK. The opposite, regarding the gamma model, is not true. It seems that this model is not robust enough to provide an adequate control of the FDR (12%) when data are generated under the inverse Gaussian model. Finally, the performance of the non-informative model is not as good as expected: this model seems to be somehow robust to the true distribution of the data, but provides a very low FDR, compensated by a high FNR.

5.3.2 Asymptotic study

The multiple testing procedure we use is based on the asymptotic (when m is large) uniformity of the posterior probabilities p_i under the null hypothesis (see Theorem 2.2). Note that no result similar to the gamma model (Theorems 3.1 and 3.2), whose uniform property holds not only asymptotically, were presented in this section. However, Figure 5.7 shows histograms of the probabilities p_i under the null hypothesis, for $m = 5$, $m = 10$ and $m = 20$. These results are based on a single simulated dataset, generated under the non-conjugate inverse Gaussian model, with $\eta = 810.6$, $w = 0.068$, $a = 2$, $b = 0.266$, $p = 0.05$ and $n = 1000$. We can really see that the property of uniformity seems to be approximately true, even for small value of m, and non-asymptotic results

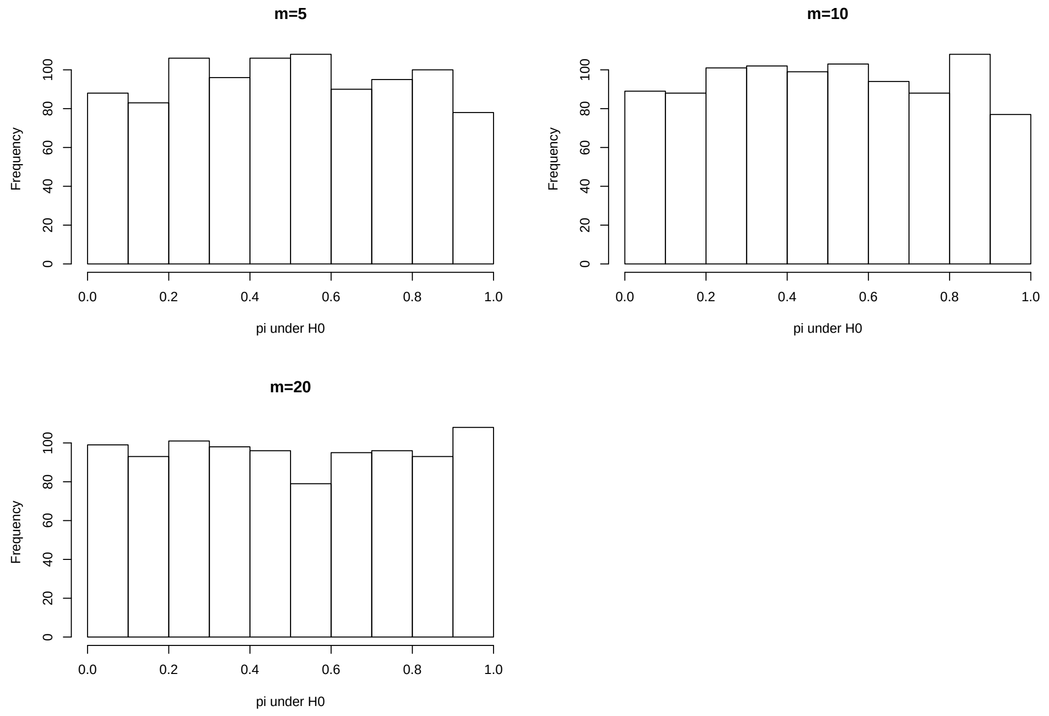


Figure 5.7: Posterior probability p_i under H_0 for $m = 5$, $m = 10$ and $m = 20$. Results based on a single simulated dataset

should hold in this case. However, our research in this direction is still preliminary, and we leave this aspect to future research.

5.4 Discussion

We have introduced in this chapter three hierarchical models involving the inverse Gaussian distribution, from which posterior probabilities can be derived similarly to the gamma model presented in the last two chapters. Using the results from Theorem 2.2, these probabilities are used in the same multiple testing procedures based on the control of the False Discovery Rate.

With the study of four datasets presenting different characteristics, it appears that the conjugate Inverse Gaussian model is not a good candidate to describe the behavior of gene expression data. However, the non-conjugate model provides a very good fit to these different types of data and may be a better alternative to the gamma model, in some situations. The flexibility of this model comes from the combination of two inverse Gaussian distributions used for the raw data and its mean, and it appears to be really suitable to describe long tail datasets, typical of microarray data. Furthermore, we showed through a simulation study that this model has some attractive robustness properties that the Gamma model does not show. We note that this could be explained by the fact that the inverse Gaussian model has one more level of randomness, compared to the gamma model, by assuming a prior distribution on the shape parameter. Such an assumption in the gamma model would greatly complicate the computation of the posterior distributions and we leave this aspect as a future work.

The use of the posterior probability of the half-space provides an adequate control of the

False Discovery level, as we showed in our simulation. Again, with respect to the error rates studied, the inverse Gaussian model tends to be more robust to the true distribution of the data than the gamma model. Finally, we note that the results we used, regarding the multiple testing procedure developed, are valid asymptotically, when m is large enough. However, it seems that this property is valid not only asymptotically, and further investigation is needed in order to obtain non-asymptotic results similar to the ones obtained with the gamma model.

Chapter 6

Case-study: Using functional genomics to discover genes important in abiotic stress tolerance

This is a four year project (April 2001-March 2005) supervised by Marilyn Griffith and Barbara Moffatt, from the Department of Biology of the University of Waterloo, and by Elizabeth Weretilnyk, from the Department of Biology of McMaster University. It also involves two bio-informaticians, Brian Golding and Paulo Nuin, both from the Biology Department at McMaster University. The main goal of the project is the identification of genes essential to the development of abiotic stress tolerance in the crucifer *Thellungiella salsuginea*. In a long term, this information will ideally be used to improve the stress tolerance of canola varieties.

6.1 Overview of the project

Thellungiella salsuginea is a plant from the wide family of crucifers, the member of which grow in many different types of habitats, ranging from alkaline mudflats to meadows, thickets, beaches and burnt-over woods. It has been found that this plant is extremely tolerant of abiotic stresses, and in particular shows a high freezing, salt and drought tolerance. *Thellungiella salsuginea* is also found in the region where the majority of Canada's canola crop is produced. The discovery of genes responsible for the different types of stress tolerance in *Thellungiella* would then be of high interest, in order to improve stress tolerance, and then productivity in canola. So far, among the crucifers, the genetic components of abiotic stress tolerance have been best characterized in *Arabidopsis thaliana* (see Motoaki & Mari (2002)), the model plant used internationally for studying plant molecular genetics. Furthermore, it has been shown that *Arabidopsis* and *Thellungiella* plants are closely related, in terms of their genetic information (see Bressan, Zhang, Zhang, Hasegawa, Bohnert & Zhu (2002)). The main strategy of the project is then to exploit the extensive genetic knowledge amassed for *Arabidopsis* to discover genes involved in stress tolerance in a plant that exhibits a significantly greater capacity to survive saline, drought and freezing conditions.

6.2 Experiment pre-processing

The microarray experimental procedure requires two essential steps. First, it requires the preparation of the biological samples of interest. In this particular case, *Thellungiella* seeds were collected from the Takhini Salt Flats, in the Yukon, and were grown under different controlled environmental conditions to determine the level of freezing, salt and drought tolerance of the plant. For each salinity/temperature/drought condition, total RNA was prepared from the leaves of the plant, from which mRNA was isolated. In addition, “control” plants were grown under identical light and temperature condition except for stress (ie. normal temperature with absence of salt or drought stress) and mRNA was isolated as well. During the microarray experiment, the mRNA obtained from plants grown under two different conditions is labeled and hybridized to the cDNA sequences printed on the array. Note that in this case, expression information will be obtained only from the genes printed on the array.

First, cDNA is synthesized from plants grown under conditions known to provoke the accumulation of transcripts associated with cold acclimation, freezing tolerance, drought and salt stress. This is done by using an experimental technique called reverse transcription, where mRNA strands are used to synthesize double strand complementary DNA (cDNAs). A cDNA clone is a section of cDNA that has been inserted into a vector molecule to form many copies.

For the project, five cDNA libraries were created, using RNA from different tissues acclimated to cold, salt and freezing conditions. At this point, each cDNA was sequenced. Using the close genetic connection between *Arabidopsis* and *Thellungiella*, this was done by conducting a BLAST search analysis against the *Arabidopsis* genomic sequence databank for identification of the genes and their map positions in the *Arabidopsis* genome. Finally, in order to have a sufficient quantity of each cDNA clone to print on the array, each clone was amplified using a technique called PCR (polymerase Chain Reaction). It is the PCR product that is actually printed on the array.

6.3 Design of the microarray

	Treatment 1	Treatment 2
Experiment 1	Cold	Control
Experiment 2	Drought	Control
Experiment 3	Salt	Control
Experiment 4	Drought	Rewater
Experiment 5	Field	Control

Table 6.1: The four experiments

The five different microarray experiments that were conducted for this project are presented in Table 6.1. For some technical reasons, the five experiments were conducted and analyzed separately. While all the treatments could be combined together in a loop design (see Kerr & Churchill (2001a)), we used instead a replicated latin square design for each of the experiments. For each experiment, six chips were used, using three biological replications and a dye-reversal strategy. The design is illustrated in Table 6.2. The biological replications A, B and C correspond to mRNA extracted from three pools of different plants, grown under the same conditions. We also note that the exact same set of genes (in the same order) was printed on the 6 arrays used for each experiment. Each gene was spotted three times (on adjacent spots) on each array. Thus, 36 expression

	Green dye	Red dye
Array 1	Treat.1-Sample A	Treat.2-Sample A
Array 2	Treat.2-Sample A	Treat.1-Sample A
Array 3	Treat.1-Sample B	Treat.2-Sample B
Array 4	Treat.2-Sample B	Treat.1-Sample B
Array 5	Treat.1-Sample C	Treat.2-Sample C
Array 6	Treat.2-Sample C	Treat.1-Sample C

Table 6.2: experimental design

measurements are available for each gene, for each experiment. We mention that it is rare, in the literature, to see cDNA microarray datasets with so many replications. This extra effort allows us to get good estimates of the different sources of variation and the data available are then more reliable than many datasets available to the public.

In total, 4896 genes (PCR products from cDNA clones) were spotted on each array, and of these 125 are control spots. These controls consist of 96 spiked genes that should be constantly equally expressed under any type of treatment, and of 29 buffer-only spots where no hybridization should occur. In this last case, no cDNA was printed on the spot and we should not obtain a positive value of the expression associated to these spots. Finally, we note that each array is divided in 48 sub-arrays, represented by a 12×4 matrix of sub-arrays, and each sub-array can be represented by a 17×18 matrix of spots. Then, on each sub-array, 102 different genes are spotted, 3 times each.

In this chapter, we focus our attention on the analysis of the Cold, Drought and Salinity experiments. The variability of each gene within the three spots being quite large, we decided to take the median of the three spots as a unique measure of expression, for each array-treatment combination. Thus, a total of 12 measurements for each gene were used. Finally, note that genes having intensities within the three replicated spots that differed by more than 2.5 fold were removed from the data. Plots of the raw data $log(X)$ versus $log(Y)$ are presented in Figures A.3, A.4 and A.5. New datasets sizes are presented in Table 6.3.

COLD	DROUGHT	SALINITY
4853	4891	4835

Table 6.3: Size of each datasets. Initial datasets contain 4896 genes.

6.4 Normalization of the data

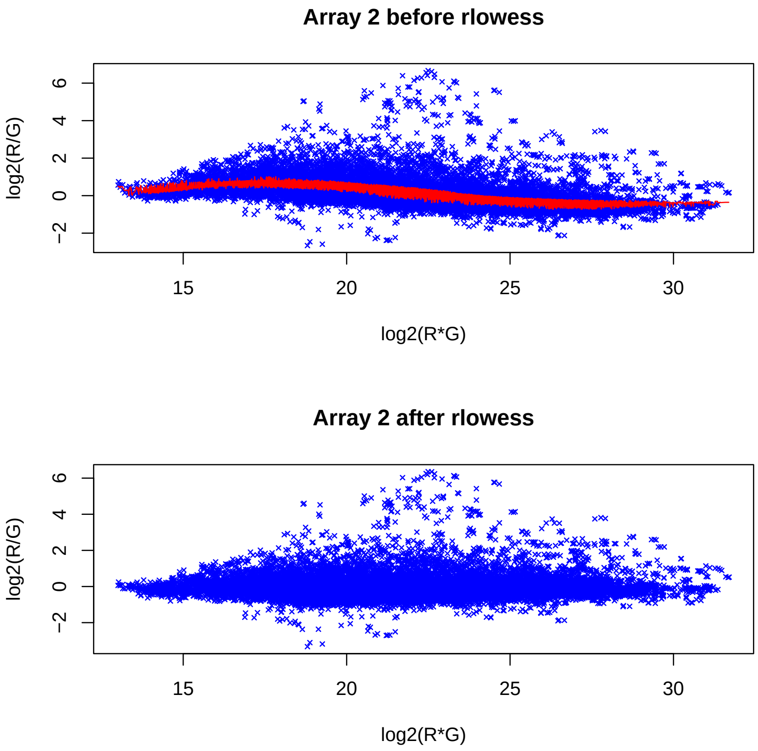
Due to the high variability of the expression data, a normalization procedure had to be performed on each array, for each experiment. This process of removing systematic effects is divided into three steps (see Cui, Kerr & Churchill (2002)). First, the background correction consists of subtracting the background intensity from the fluorescent signal at each spot. Then, the data transformation, applied to one microarray at a time, removes systematic effects from the log-ratios (Red/Green). Finally, the data normalization process calibrates the signals from different channels and arrays to a comparable scale. Regarding the data transformation (of each individual array), we observe two common features in microarray data that can be the result of a problem occurring during the hybridization step. The first one is the dependence of log-ratios on spot intensity and the second one is the spatial variation of the log-ratios, over the different sub-arrays. Since cDNA clones are typically spotted in a random fashion, one should not see any strong association between expression and the spatial region (sub-array) of a slide. The first type of dependence can be diagnosed by viewing a plot of the log-ratios versus the average of the log-intensities (over the two dyes). This type of representation is referred to as a RI plot (Ratio by Intensity), or MA plot and was first introduced by Yang et al. (2002) in the context of microarray data. Under the assumption that most of the genes are not differentially expressed (which should be the case in a large genomic study), most points of the RI plot should fall along an horizontal line, if no problem occurred in the hybridization. The second type of dependence (spatial heterogeneity) can be diagnosed by plotting, for each sub-array, the intensity of the red channel versus the green channel. If no problem in the hybridization occurred, the red and green intensities should be highly correlated and we should see a nearly linear curve in all plots. There are many data transformation methods available that correct for the former type of dependence. We chose to use the one developed recently by Cui et al. (2002), called the joint LOWESS method. It has the main advantage of combining two standard approaches, the intensity LOWESS correction and the spatial LOWESS correction (see Yang et al. (2002) for more details) to correct for intensity dependent bias and spatial bias simultaneously. Then, for each array, and for each gene g, the corrected log-ratio is

$$\begin{aligned} Z_g &= log\left(X_{red,g}/X_{green,g}\right) - C_g(I_g, row, col), \\ \text{where } I_g &= (log(X_{red,g}) + log(X_{green,g}))/2, \end{aligned}$$

and where C_g is a constant that depends on the location of the spot g on the array (col, row) as well as the average intensity I_g. This constant is determined by the common curve fitting LOWESS procedure. Here, X refers to the expression value for a particular spot and a particular dye on the array. We note that if a linear function is used for the local regression, the procedure is then referred to as the joint LOESS procedure.

As an example of such correction, RI plots, where a strong trend is observed before correction, are presented in Figures 6.1 and 6.2. These plots correspond to the second and fifth arrays of the Cold and Salinity experiments, respectively. We can see that, for these two arrays, a trend was clearly noticeable before transformation, that was corrected by the joint LOWESS procedure.

Finally, in order to normalize the data between arrays and treatment to a comparable scale, the procedure described in Section 4.2.1 was applied.

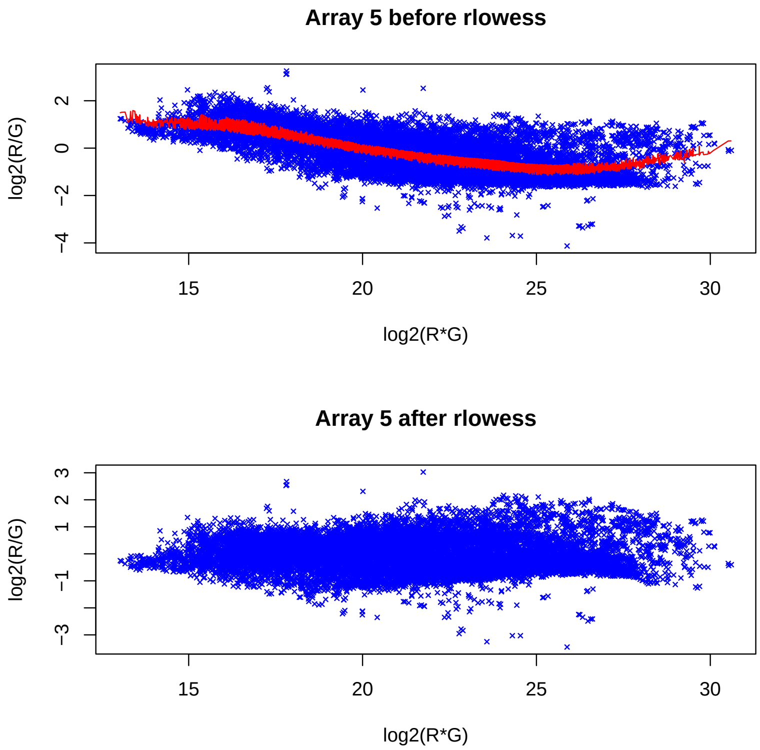

Array 2 before rlowess

Array 2 after rlowess

Figure 6.1: RI plot: Array2, Cold dataset

6.5 Results

We applied the three procedures developed in this thesis to the three datasets. The procedure involving the conjugate gamma model, described in Chapter 3, is denoted GAM. Its generalization to a multiplicative model accounting for the gene-array and gene-dye interactions, described in Chapter 4, denoted as GAM.MUL. As we did in Chapter 3, 2500 iterations were used in the Gibbs sampling algorithm, and starting values were generated according to the prior distribution of the gene-specific effects. Finally, the procedure based on a inverse Gaussian model, as described in Chapter 5, is denoted IG. Note that the multiplicative gamma model is the only method that takes into account


Array 5 before rlowess

Array 5 after rlowess

Figure 6.2: RI plot: Array5, Salinity dataset

the array-gene and the dye-gene interactions. However, we recall that data have been previously normalized to account for a global array, treatment and dye effects. We finally mention that for each model, genes have been detected at a FDR level of 1%.

First, let us evaluate the fit provided by the models, to the three datasets. The quantile-quantile plots, regarding the multiplicative gamma model as described in Section 4.3.1, are presented in Figures 6.3, 6.4, and 6.5 for the Cold, Drought and Salinity datasets, respectively. We note that these quantile plots can hardly be compared with those involving the histogram and predictive distribution of the data (as for the gamma and inverse-Gaussian models). The main issue in such plots is that in the case of the multiplicative model, the predictive distribution cannot be computed easily. A partial

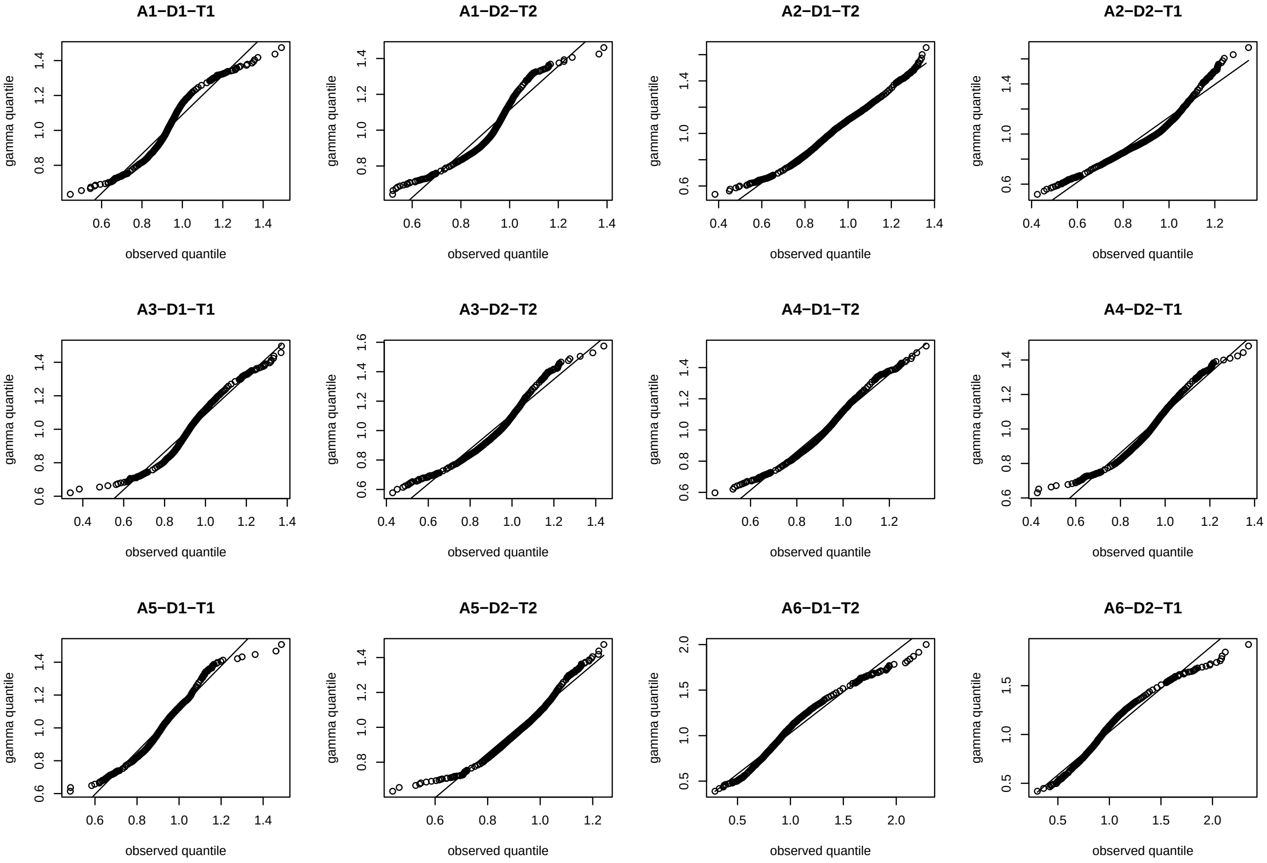


Figure 6.3: QQplots for the multiplicative gamma model: Cold dataset

solution to this problem is then to correct the data for the array and dye gene-specific effects, using the posterior expectations obtained from the multiplicative model. A simple gamma model can then be fitted on the data arising from the two types of treatments, as we did in Chapter 3. The histograms of such corrected data (on the logarithm scale), with the associated predictive distribution, are presented in Figure 6.6. The predictive distributions (on the logarithm scale) versus the histogram of the intensities are also presented in Figures 6.7 and 6.8 for the GAM and IG models respectively. Note that these histograms are not identical to those of Figure 6.6 since in this case, data have

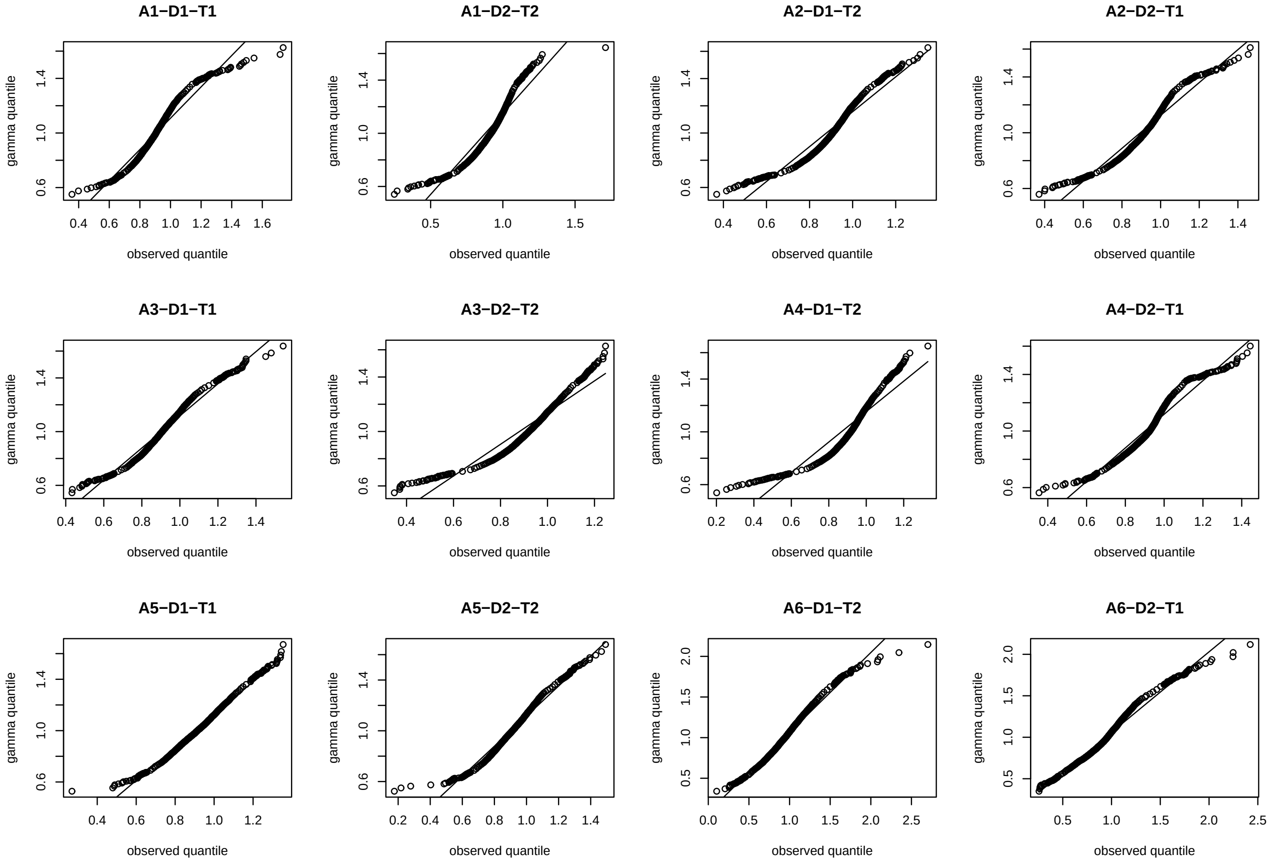


Figure 6.4: QQplots for the multiplicative gamma model: Drought dataset

been corrected for the gene-specific dye and array effects. First, these figures confirm the fact that there is no universal "best model" in microarray data, and that intrinsic features of the data determine which model is the most appropriate. Up to now, these features have not been discovered, and a goodness of fit study of several models seems an appropriate choice. Figures 6.5 and 6.6 both suggest that the multiplicative gamma model is the most appropriate for the SALINITY dataset. The quantile-quantile plots of the COLD and DROUGHT data clearly show a deviation from the gamma (especially for the DROUGHT) and this model is not appropriate in this case. We also note the small

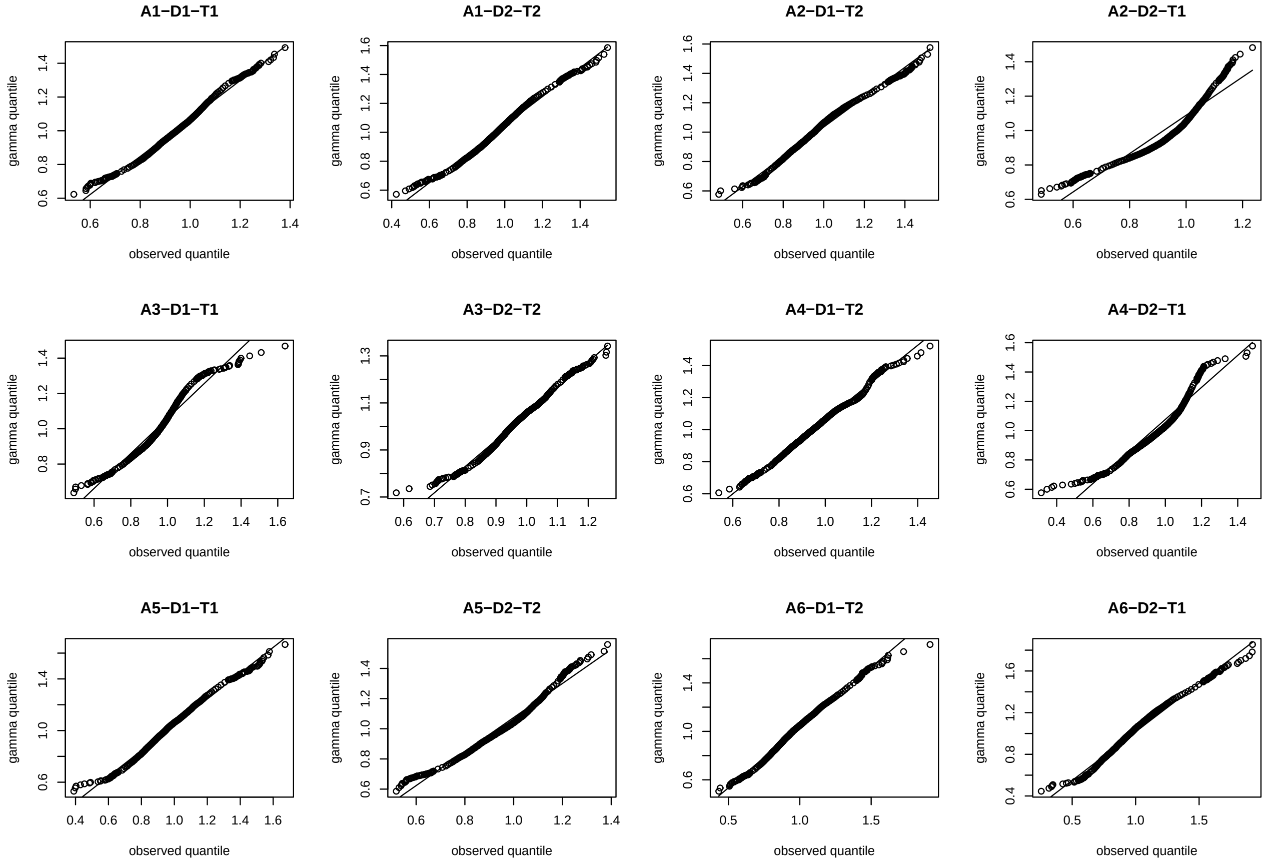


Figure 6.5: QQplots for the multiplicative gamma model: Salinity dataset

bias in the boxplots of the residuals for these two datasets, with a median and average of 0.95 instead of 1, as we could expect. Figures 6.7 and 6.8, which show the predictive distributions for the IG and GAM models seem to suggest that the IG model is more appropriate than the GAM model for both the DROUGHT and COLD data. Note that such a statement could be verified with the help of the goodness of fit distances described in Section 5.2.1 (results not shown here).

The number of genes detected by the three models is presented in Table 6.4. By looking at the results, we can observe the same pattern, regardless of the type of dataset

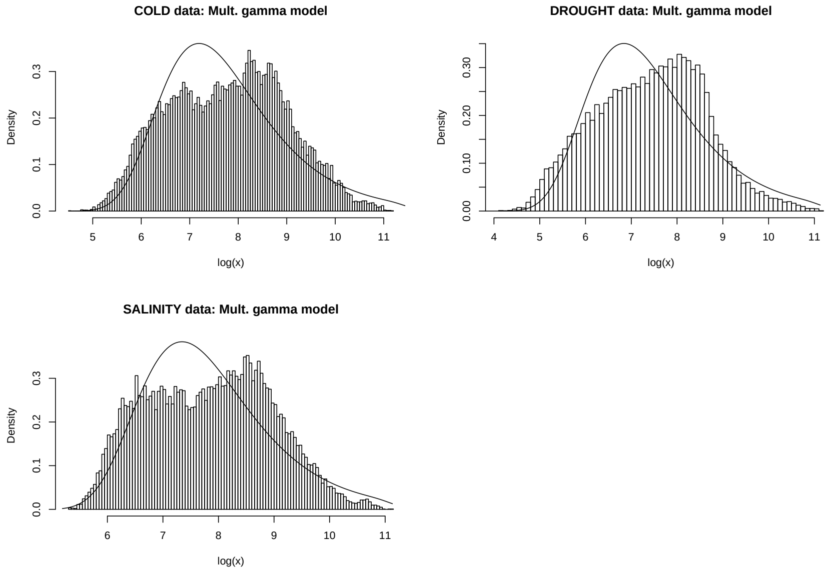


Figure 6.6: Predictive distribution for the MUL model, for the three datasets, corrected for the gene-specific array and dye effects

	GAM.MUL		GAM		IG
COLD	337	$< -$ 160 common $- >$	162	$< -$ 113 common $- >$	**116**
DROUGHT	295	$< -$ 195 common $- >$	202	$< -$ 108 common $- >$	**115**
SALINITY	**266**	$< -$ 60 common $- >$	62	$< -$ 25 common $- >$	26

Table 6.4: Number of genes detected. Numbers in bold represent the number of genes detected by the most appropriate method, according to the goodness of fit study.

analyzed. The gamma multiplicative model seems to be the least conservative approach, and detects more genes than the other models. On the contrary, the inverse Gaussian model is the most conservative model leading to the detection of fewer genes. An extreme illustration of this pattern is shown in the SALINITY dataset, with which the IG model detects only 26 genes, versus 266 genes for the multiplicative gamma model. Note that in this specific case, GAM.MUL seemed the most appropriate model, as we concluded from the goodness of fit study. However, it seems that the three models agree pretty well

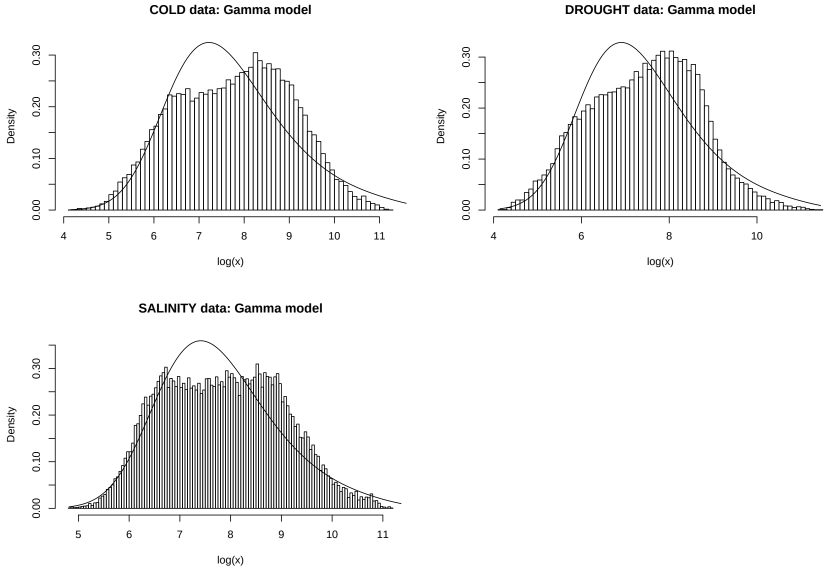


Figure 6.7: Predictive distribution for the GAM model, for the three datasets

in the list of genes detected since most of the genes from a short list also appear in the longer list. Furthermore, we note that none of the buffer-only genes (not expressed in any of the treatments) and the spiked genes (equally expressed controls) were detected by any of the models. A similar statement holds for 5 genes (ACTIN2, TUBULIN, EIF, UBQ10 and PIP) printed several times across the arrays, known in the literature to be equally expressed (constitutive genes) under the three types of treatments studied. For a visualization of which genes are detected by each method, we refer to the Figures B.1 to B.6 for the COLD dataset, to Figures C.1 to C.6 for the SALINITY dataset and finally, to Figures D.1 to D.6 for the DROUGHT dataset.

In addition to these constitutive genes, some positive genes have also been printed on the arrays. For instance, the gene COR15, known in the literature to be induced under the COLD treatment, was printed 3 times on each array. The three occurrences of this gene have been detected by the three models for the COLD dataset. Another gene, FL5-2D23,

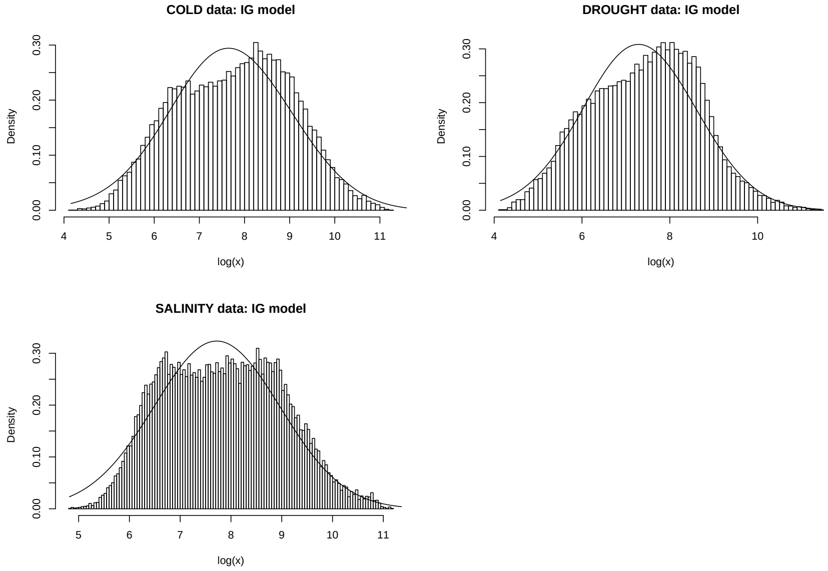


Figure 6.8: Predictive distribution for the IG model, for the three datasets

known to be induced in DROUGHT was also printed in three copies. The 3 occurrences were detected by the GAM.MUL and GAM model, whereas only 2 occurrences were detected using the IG model, for the DROUGHT dataset. Note that these same three occurrences were also detected by the all the models in the SALINITY dataset.

6.6 Discussion

From the analysis of the three datasets described, we can draw several conclusions. First, as we mentioned, it is very hard to find a model that can describe adequately microarray data, in general. We believe that several models should be applied to the datasets, and a goodness of fit study should be performed in order to know which model fits best. In our case, the quantile-quantile plots, or the histograms versus the predictive distribution plots are both good indicators of goodness of fit. In addition, the use of controls is very important (positive and negative) in order to ensure the quality of the results. In our case, the three models agree in this sense, since the known positive genes were detected by the three models and the spiked genes as well as the constitutive genes were not detected. Finally, we mention that the feature of the data may depend of several unknown factors, and we note that in our case, even if the three datasets come from the same laboratory, and were built in the same technical conditions, we observe a variation in the type of model that should be used. However, we still believe that Bayesian or random effects models should be preferred to frequentist one, since these models deals easily with a small number of observations and are generally more flexible.

Chapter 7

Future research

In this chapter, we outline some future research projects, that would allow us to generalize or link some of the results presented in the previous chapters to another area of statistical genetics. In particular, we first consider the possibility of developing a multiplicative model, similar to the one presented in Chapter 4, involving the inverse Gaussian model. Furthermore, motivated by a paper from Scholten, Miron, Merchant, Miller, Miron, Iglehart & Gentleman (2004), we consider extending the multiple testing procedure developed in this thesis to more than two treatments. Finally, we would like to explore the possibility of linking gene expression data with some specific DNA sequences corresponding to the genes, through a problem called DNA motif discovery.

7.1 A Multiplicative ANOVA model for inverse Gaussian data

As we have seen in Chapter 5, inverse Gaussian models can provide a very powerful alternative to the traditional gamma models. Specifically, these models demonstrate properties of robustness to the true distribution of the data that are very appealing, especially for microarray data whose distribution may vary a lot from one laboratory to another. Furthermore, results regarding the uniformity of the posterior probability p_g under the null hypothesis do not seem to be affected by a small number of observations. Regarding the non-informative model, we believe that computations of the posterior distribution of the gene-effects would be tractable, under the model

$$\begin{aligned} X_{ijkgr}|\theta_{1g},\ldots,\theta_{pg},\lambda_g &\sim IG\left(\frac{1}{\theta_{1g}^{z_{ij}^{(1)}} \times \theta_{pg}^{z_{ij}^{(p)}}},\lambda_g\right), \\ \pi(\theta_{1g},\ldots,\theta_{pg},\lambda_g) &\propto \frac{1}{\lambda_g}, \end{aligned}$$

using the same notations as in Chapter 4. However, computations would become much more challenging if one wanted to work with a model similar to the non-conjugate one, presented in Section 5.1.4. Clearly, it would be interesting to see if such a model can be integrated into the multiplicative framework presented in the fourth chapter.

7.2 Generalization to more than 2 treatments

In this thesis, we considered the particular case of two treatments. The multiplicative model presented in Chapter 4 can be applied to more than two treatments, but how the multiple testing issue should be considered in this case remains an open question. Suppose for example an experiment involves three treatments, with gene expressions noted $\underset{\sim}{X}$, $\underset{\sim}{Y}$ and $\underset{\sim}{Z}$ respectively. The three treatments means, for each gene g, are noted θ_{xg}, θ_{yg} and θ_{zg} respectively and we are interested in testing the null hypothesis $H_0 : \theta_{xg} = \theta_{yg} = \theta_{zg}$. Note that the parameter space is then $\mathcal{R}^3$. With such a space, a half-space would be defined as $\{\underset{\sim}{\theta} \in \mathcal{R}^3 : \; v_1\theta_x + v_2\theta_y + v_3\theta_z \geq M\}$ with $M \in \mathcal{R}$ and $|v_1 + v_2 + v_3| = 1$. Taking the posterior probability of such a space may not be very meaningful in many cases. We may then think about two alternative ways of testing the hypothesis H_0. The first solution would be to consider the alternative hypothesis H_1 : *At least one of the θ's is different*, and then consider the one-sided alternative $H_1^a : \theta_{xg} > \theta_{yg} > \theta_{zg}$. In such a case, note that H_1^a represents 1/6 of the total space and thus cannot really be seen as a "one-sided" alternative hypothesis. Such considerations are the object of current work and it can be shown that the computation of $P(H_1^a|data)$ involves the knowledge of the joint distribution of $Z|X$ and $Z|Y$, which we are currently trying to derive.

The second solution we propose is an iterative solution. First, the null hypothesis $H_0^{(1)}$: $\theta_{xg} = \theta_{yg}$ is tested using the posterior probability of the half-space $P(\theta_{xg} > \theta_{yg}|data)$. This can be done using the approach developed in Chapter 2. Taking only the subset of genes from which $H_0^{(1)}$ was not rejected, we can consider the parameter space $\{(\theta_x, \theta_y, \theta_z) \in \mathcal{R}^3 \text{ and } \theta_x = \theta_y\}$. Under such a parameter space, we consider the null hypothesis $H_0^{(2)} : \theta_{xg} = \theta_{yg} = \theta_{zg}$ and the one-sided alternative as $H_{1a} : \theta_{xg} = \theta_{yg}$ and $\theta_{yg} > \theta_{zg}$. Again, the probability $P(H_{1a}|data)$ represents the posterior probability of the half-space of the space considered and results presented in this thesis can be used. Such an iterative procedure can be generalized to any number of treatments and any type of contrasts tested. The main idea is then to take, in an iterative manner, the half-space of the

ES present	⇒	**Gene A** expressed. Production of a protein called Transcription Factor (TF)		
		⇓		
		TF binds to the promoter region of Gene B	⇒	Expression of **Gene B** is induced by the TF

Table 7.1: illustration of primary and secondary ES target genes

space under which a null hypothesis is tested.

A very interesting motivation and application regarding the generalization of our work to several treatments is the experiment described in Scholten et al. (2004). This paper emphasizes the fact that microarray experiments can be very efficiently designed by interpreting the biological questions in terms of the statistical parameters. In this specific case, 32 affymetrix arrays, from a 2^4 factorial design experiment with 2 replications each, were printed. Data come from an experiment on cells from an estrogen receptor positive human breast cancer cell line. Among the four 2-level factors studied, two are of particular interest: cyclohexamide (CX, present or absent) and estrogen (ES, present or absent). One of the goals of the study is to detect ES target genes (genes differentially expressed under the presence of estrogen) and among them, differentiate between the primary and secondary targets.

The definition of primary and secondary ES target genes is illustrated in Table 7.2. In this table, Gene A is a primary ES target since its expression is directly induced by the presence of ES. On the other hand, Gene B is called a secondary ES target since its expression is the results of the production of a transcription factor by the primary ES

Conditions	1	2	3	4	5	6	7	8
$\theta_0 \neq \theta_1$	No	Yes	Yes	Yes	No	Yes	Yes	Yes
$\theta_0 \neq \theta_2$	Yes	No	Yes	Yes	Yes	No	Yes	Yes
$\theta_0 \neq \theta_3$	Yes	No	Yes	No	No	Yes	Yes	Yes
$\theta_3 \neq \theta_1$	Yes	Yes	Yes	Yes	No	No	No	No
$\theta_3 \neq \theta_2$	No	No	Yes	Yes	Yes	Yes	Yes	Yes
$\theta_1 \neq \theta_2$	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes

Table 7.2: Situation that may occur when genes are ES targets. Situations 1-4 represent primary targets and situations 5-8 represent secondary targets

target gene A.

We consider here expression levels under four treatments, noted as

$$\begin{aligned} X_0 &: \text{ gene expression under both CX and ES absent,} \\ X_1 &: \text{ gene expression under CX only,} \\ X_2 &: \text{ gene expression under ES only,} \\ X_3 &: \text{ gene expression under both CX and ES.} \end{aligned}$$

Note that the index g for each gene is omitted here. Suppose that (X_0, X_1, X_2, X_3) have corresponding expression means $(\theta_0, \theta_1, \theta_2, \theta_3)$. Under such a framework, if a gene is an ES target, 8 types of biological situations may occur, that can be related to the parameters as described in Table 7.2. Situations are labeled from 1 to 8. Primary ES targets are represented by situations 1-4 and secondary ES targets by situations 5-8. Using this table, we can see that testing if a gene is an ES target can be done by testing the null hypothesis $H_{0a} : \theta_0 = \theta_1 = \theta_2$ (representing the hypothesis that a gene is not an ES target). Among the ES targets, secondary ES targets can be identified by testing $H_{0b} : \theta_3 = \theta_1$. Furthermore, genes could be clustered according to the type of biological situations they represent, by testing the hypothesis that a specific situation (shown in Table 7.2) occurs.

7.3 Detection of regulatory motifs in DNA sequences

As we have seen through this document, microarray technology provides insights about the level of expression of a set of thousands of genes in a cell, or differences in their expression between two cells. If we go deeper into the principles of molecular genetics, it is of interest to understand the biological mechanisms behind gene expression. It is known that the expression of specific genes can be controlled by some proteins called transcription factors. These proteins act by binding to short sequences of nucleotides, located in the upstream region of the gene. These corresponding families of sequences are distinctive and are referred as *cis elements* or *motifs*. Each occurrence of the motif in the sequence is called a *motif element*. The knowledge of these sites is of course of high interest for biologists.

One of the main issues with the discovery of new motifs is the variability of the motif elements. For example, consider the 12 binding sites for a transcription factor called the λ repressor, composed of 8 nucleotides each. It appears that only two of the eight nucleotides are conserved over all the sites, whereas the other nucleotides have a range of variability. We can then view the problem studied as detecting the occurrences of a word (motif), in a text (DNA sequence) composed of four letters, A, T, G and C, where many ”typos” can occur. The motif is then described as a stochastic word and it is natural to use probabilistic models to describe the motif pattern, as well as the associated statistical properties to lead us in the discovery of these patterns.

A motif of length w is represented by a Probability Weight Matrix (PWM) of size $4 \times w$, noted $\Theta = (\theta_1, \ldots, \theta_w)$. Each column j of the matrix, θ_j^T, represents a multinomial distribution with four cell probabilities, $\theta_{j1}, \ldots, \theta_{j4}$, representing the probability of occurrence of the four nucleotides A, T, G and C. Several types of models have been developed and, as a basis for our research, we consider the models developed by Keles, Van der Laan, Dudoit, Xing & Eisen (2003) and Liu, Gupta, Liu, Mayerhofere & Lawrence (2004). In both approaches, we consider the nucleotides of the DNA sequence arising from a two component multinomial mixture model. The first component is called the background model and assumes that the nucleotides observed at a particular site do not contribute to the motif, but are independent and identically distributed according to a multinomial distribution with parameter $\theta_0 = (\theta_{01}, \ldots, \theta_{04})$. The second component of the model is the motif model, described by the matrix Θ.

7.3.1 Statistical models

The first model we studied (Keles et al. (2003)) assumes zero or one occurrence of the motif per sequence. The occurrence (or not) of the motif as well as the starting position of the motif in the sequence are considered to be hidden variables. If the motif occurs once, its starting position is assumed to be uniformly distributed over the length of the sequence (minus the length of the motif). In this model, no prior distribution is assumed on the parameters Θ and θ_0, and the use of the E-M algorithm is required in order to estimate these parameters. The authors of the paper also showed that it is possible to impose some constraints on the PWM Θ, through its Information Content (IC) at a position l of the motif, defined as

$$IC(l) = log_2 J + \sum_{j=1}^{J} \theta_{lj} log_2 \theta_{lj},$$

where $J = 4$ represents the number of letters in the "alphabet". The constraints on the information content of the matrix arise as a function of the position l in the motif and examples of such constraints are the high-low-high or low-high-low information content profiles. In the case of the constrained Probability Weight Matrix, estimates of the parameters are also obtained using the E-M algorithm, but the computational complexity of the problem requires the use of some programming techniques such as SQP (Sequential Quadratic Programming). Inference about the motifs is made through the probability that the motif starts at a specific location, and through the probability that there is one motif occurrence in the sequence. In practice, we note that neither the length of the motif, nor the type of constraint for the PWM (eg: low-high-low) are known. There is the need here to search through a specified range of motif widths and the optimal width is chosen by optimizing a model selection criterion. In the case of the paper studied, the model selection method chosen is a likelihood-based cross-validation.

The second model (Liu et al. (2004)) deals with one or more motif occurrences per sequence. The main difference between this model and the one described in the previous section is the introduction of a Bayesian framework. Again, a hidden variable of the model is the starting position of the motif in the sequence, but a Dirichlet prior is assumed on the Probability Weight Matrix Θ and on the background probability vector θ_0. The parameters of the model are estimated using a Gibbs sampling technique. Furthermore, no constraints on the PWM are assumed, but a sequential model approach can be used, that allows for some "holes" in the motif.

7.3.2 Linking gene expression data with motif finding analysis

Very few papers have been published that link gene expression with motif discovery. We believe that the information available from gene expression (microarray data) could be of great use in the search for new motifs. For instance, if genes are clustered based on similarity in expression profile over a large number of different conditions, the upstream regions of the genes in the cluster can then be analyzed for the presence of shared sequence motifs (see Bussemaker, Li & Siggia (2001)). Among the papers available linking microarray data with motif discovery, we can cite Conlon, Liu, Lieb & Liu (2003) which uses linear regression models between a motif-matching score and gene expression. The algorithm they use can be summarized into 5 steps. First, the genes are ranked according to their expression, and their upstream sequence is obtained. We note here that the genes selected belong to a cluster of genes with high expression. Then, a computational method, called Motif Scan Discovery (MDscan) is applied, that searches for DNA sequence motifs. We refer to Liu, Brutlag & Liu (2002) for the details about this procedure, and we note that the underlying statistical details (Bayesian model) can be found in Liu, Neuwald & Lawrence (1995). This model is similar to the one described briefly in the previous section. The next step of the algorithm is to construct a score, for each sequence, for matches to each MDscan reported motif. For a motif m of length w and for a gene g, the score is defined as

$$S_{mg} = log_2 \left[\sum_{x \in X_{wg}} \frac{Pr(\text{ x from } \Theta_m)}{Pr(\text{ x from } \theta_0)} \right],$$

where Θ_m is the PWM of the motif m, θ_0 represents the background probability model (here, a third order Markov chain model for the background is assumed) and X_{wg} is the set of all w-mers (nucleotides sequences of size w) in the upstream sequence of gene g. In the fourth step of the algorithm, a simple linear regression is performed, for each motif m reported by MDScan, such that

$$Y_g = \alpha + \beta_m S_{mg} + \epsilon_g,$$

where Y_g is the log_2 expression value of gene g and where ϵ_g is the gene-specific error term. The motifs having a significant β coefficient are retained and the multiple regression model is then fitted:

$$Y_g = \alpha + \sum_{m=1}^{M} \beta_m S_{mg} + \epsilon_g$$

A stepwise regression procedure is finally applied, to detect the group of motifs acting together to affect gene expression.

As we can see, the problem of motif discovery can be divided into two different aspects: the construction of statistical models that allow us to obtain a list of potential motifs from the n DNA sequences, and the construction of a model linking gene expression with a score for each motif. Regarding the first aspect of the problem, we would like to consider a model with zero, one or more occurrences of the motif per sequence (we recall that the models presented in section 6.2 assume zero/one or one/more occurrences of the motif per sequence). As for the other models, the hidden variables considered would be the starting position of the motif in the sequence, as well as its number of occurrences. For instance, we could define the variable Y_k such that $Y_k = 1$ if there is at least one occurrence of the motif in the kth sequence, and $Y_k = 0$ otherwise. We could consider the introduction of a prior distribution on the variable Y_k, such that we could assume the variables Y_k as being distributed according to a Bernoulli distribution with parameter p_k, where p_k would be the posterior probability defined in the last two chapters, for the gene k. The motivation behind this is that if the gene is differentially expressed under the treatment, there is more chance to find a motif sequence that would induce the expression of this gene under this treatment.

Furthermore, introducing some constraints on the Information Content of the Probability Weight Matrix of the motif seems very appealing, and it would be interesting to find a more general function for this constraint, that would be a function of some parameters such that it could accommodate a variety of array shapes. The selection model procedure would then be defined on the length of the motif as well as on the parameters of the Information Content.

Regarding the second aspect of the problem of finding new motifs, we could use the work from the previous chapter by using the normalized gene expression ratio of the two treatments (using the posterior expectations) as the dependent variable. The posterior probability of differential expression could then be used as a weight for the motif score (a big part of the significant genes have an estimated posterior probability of 1) in the regression model. Some alternative to linear models (non-linear models) could also be considered.

Chapter 8

Discussion

The main contribution of this thesis is the attempt to connect Bayesian analysis and frequentist theory in a frequentist multiple testing framework. The half-space procedure described in this document provides an interesting tool, allowing a posterior probability to be considered in the same spirit as a p-value in such a context.

Two main points motivated our work. First, models using p-values are restrictive in many aspects, but offer in the other hand the advantage of providing a control of frequentist error rates, ie. that do not depend on the data. Inversely, Bayesian theory brings an attractive flexibility to the models, but the control of posterior error rates can be criticized. We can cite for example the weak repeated sampling principle, stating that "we should not follow procedures which for some possible parameter values would give, in hypothetical repetitions, misleading conclusions most of the time" (see Cox & Hinkley (1974)).

The linking of both approaches enables us to use the best of each simultaneously in a unique procedure. The posterior probability p_i we use is at the same time meaningful as a test statistic (we reject the null hypothesis for small values of p_i), computed under the posterior probability of the parameter of interest, and independent of the mixture parameter of the model (proportion of true null hypotheses). This makes it a statistic robust to a misspecification of this parameter, with the open possibility of using it in the Benjamni-Hochberg procedure, or any of its extension using frequentist p-values to control the False Discovery Rate. Note that the use of this type of probability was first introduced by Ibrahim et al. (2002) as a measure of significance for each test. In this paper, sub-models were created, for each threshold level δ such that H_{1i} is rejected when $p_i > 1 - \delta$ or $p_i < \delta$, and the optimal model was chosen using a Bayesian criterion.

The method we propose is described in the context of microarray experiments in this document. However, one should not limit its use to this type of datasets. If the case of two dimensions is treated here, the approach finds also very good application in the univariate case. Testing any hypothesis, in a Bayesian framework, of the form $H_0 : \theta = \theta_0$ can be done by considering the posterior probability of the half-space, $P(\theta > \theta_0|data)$, and similarly, by using it as an input in the BH procedure. In this case, exact computation of such probability only requires the knowledge of the posterior distribution of θ, which is often not difficult to obtain if the model is tractable enough. Note that in this thesis, only one or two level for the hierarchy of the model were considered, but highly hierarchical models could be used in a similar spirit.

The models we considered in the context of microarray data may be somehow questionable with respect to the independence assumption between genes. Modeling such a dependence is very challenging and this is still an open research question in the microarray literature. However, hierarchical models seem to have the potential of modeling certain dependence structure, as Ibrahim et al. (2002) showed, and more research in this area is certainly needed.

Bimodal models could also be considered successfully, and some preliminary work in this direction shows that it could be the answer to the lack of fit provided by any models, especially regarding the COLD and SALINITY dataset presented here. In such a case, modeling low and high expressed models using a mixture of distribution seems to greatly improve the model fit.

Finally, we want to mention that the tools developed in this thesis are especially appropriate for the analysis of two color microarrays, where using ratios is the predominant approach. These tools would likely be less useful in the case of one colour arrays.

Appendix A

Raw data plots

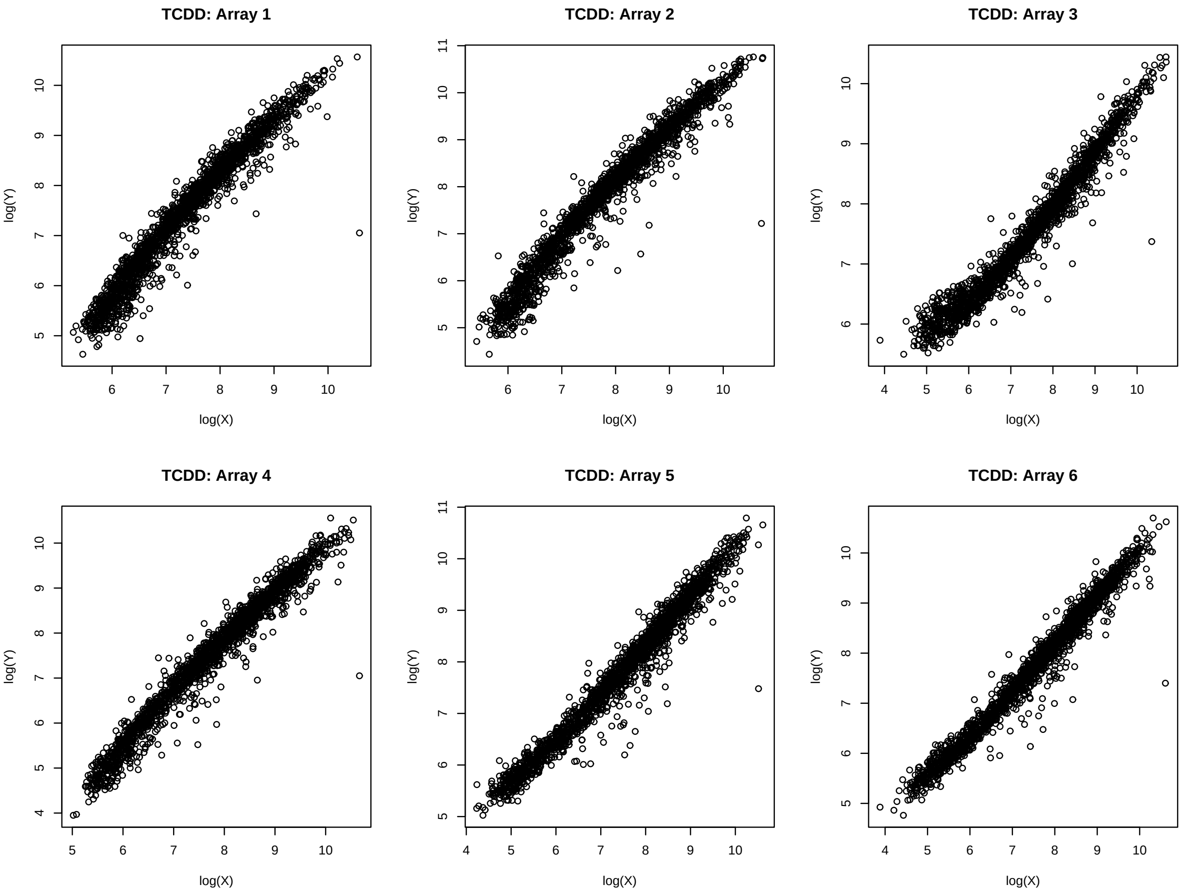


Figure A.1: Raw data: TCDD dataset

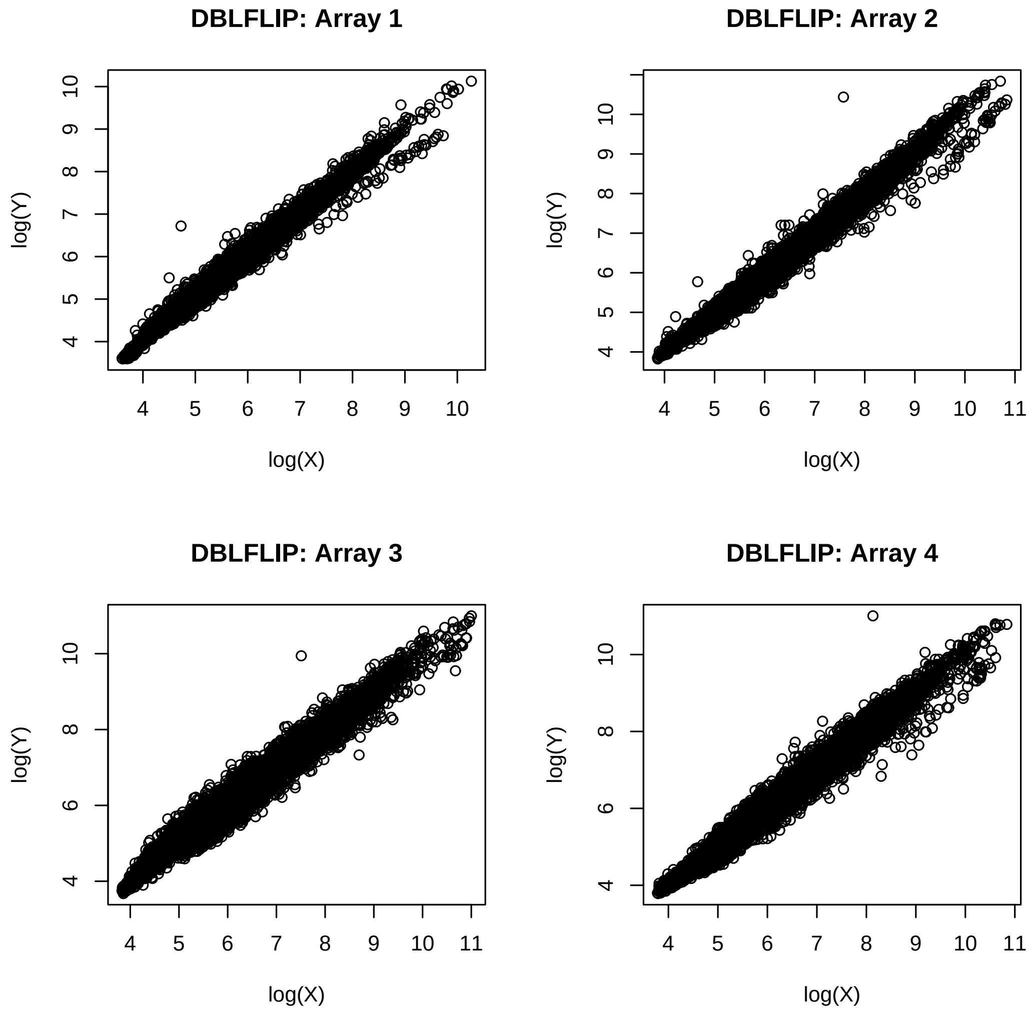


Figure A.2: Raw data: DBLFLIP dataset

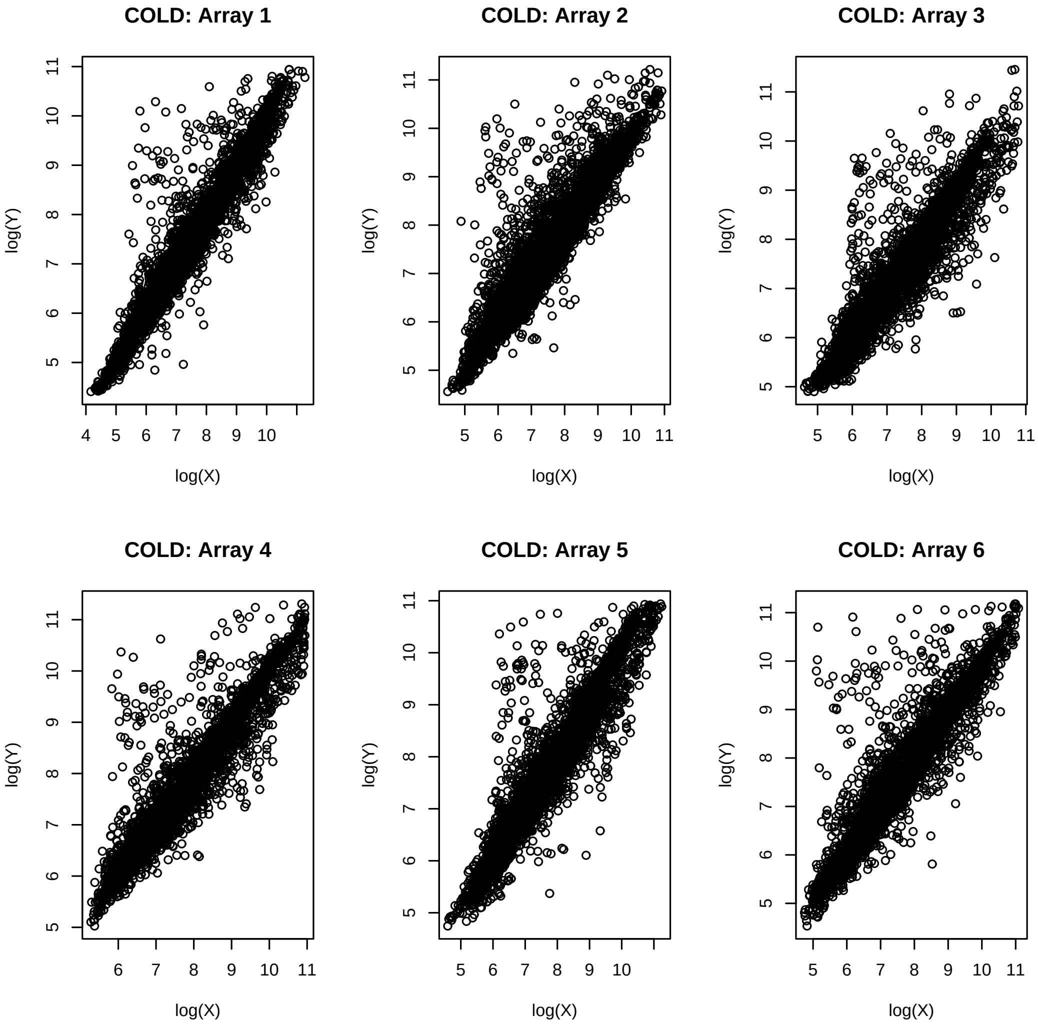


Figure A.3: Raw data: COLD dataset

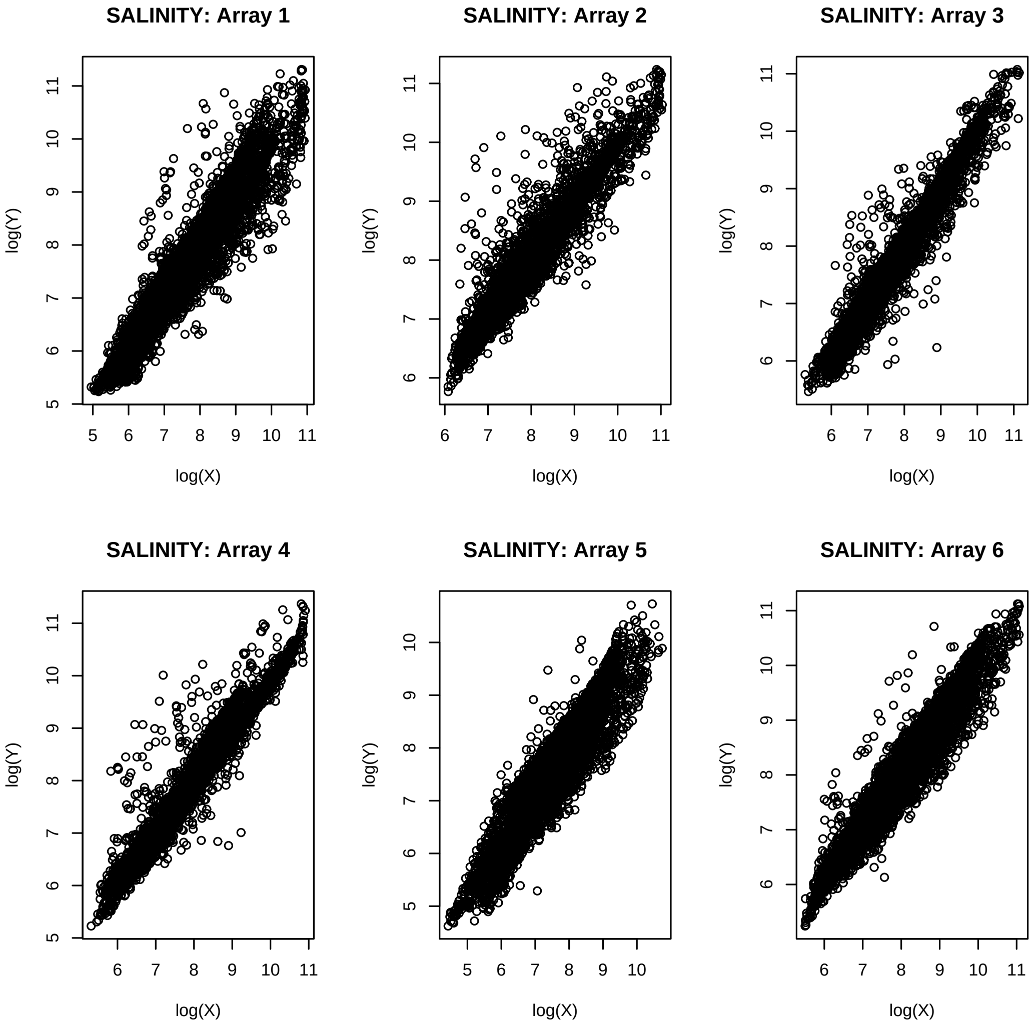


Figure A.4: Raw data: SALINITY dataset

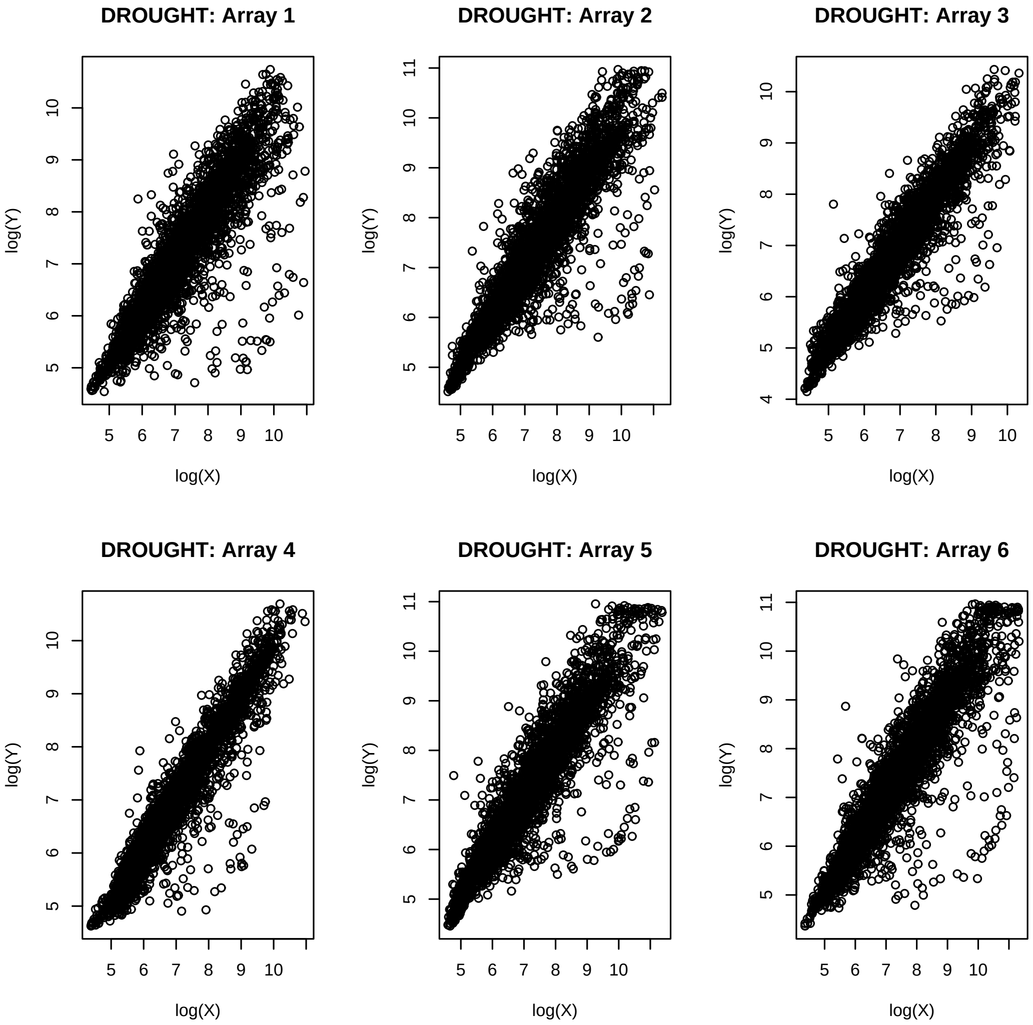


Figure A.5: Raw data: DROUGHT dataset

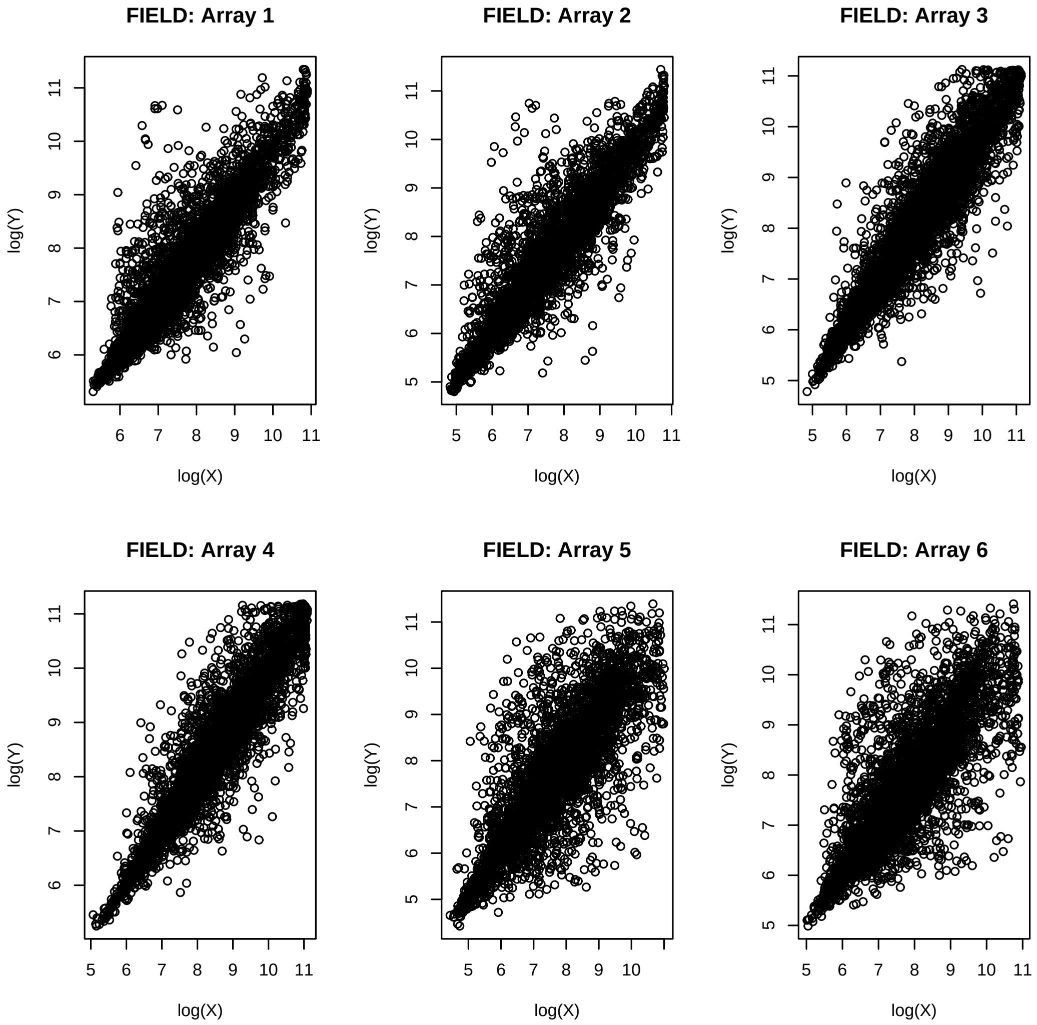


Figure A.6: Raw data: FIELD dataset

Appendix B

COLD dataset: Plots of the gene detected

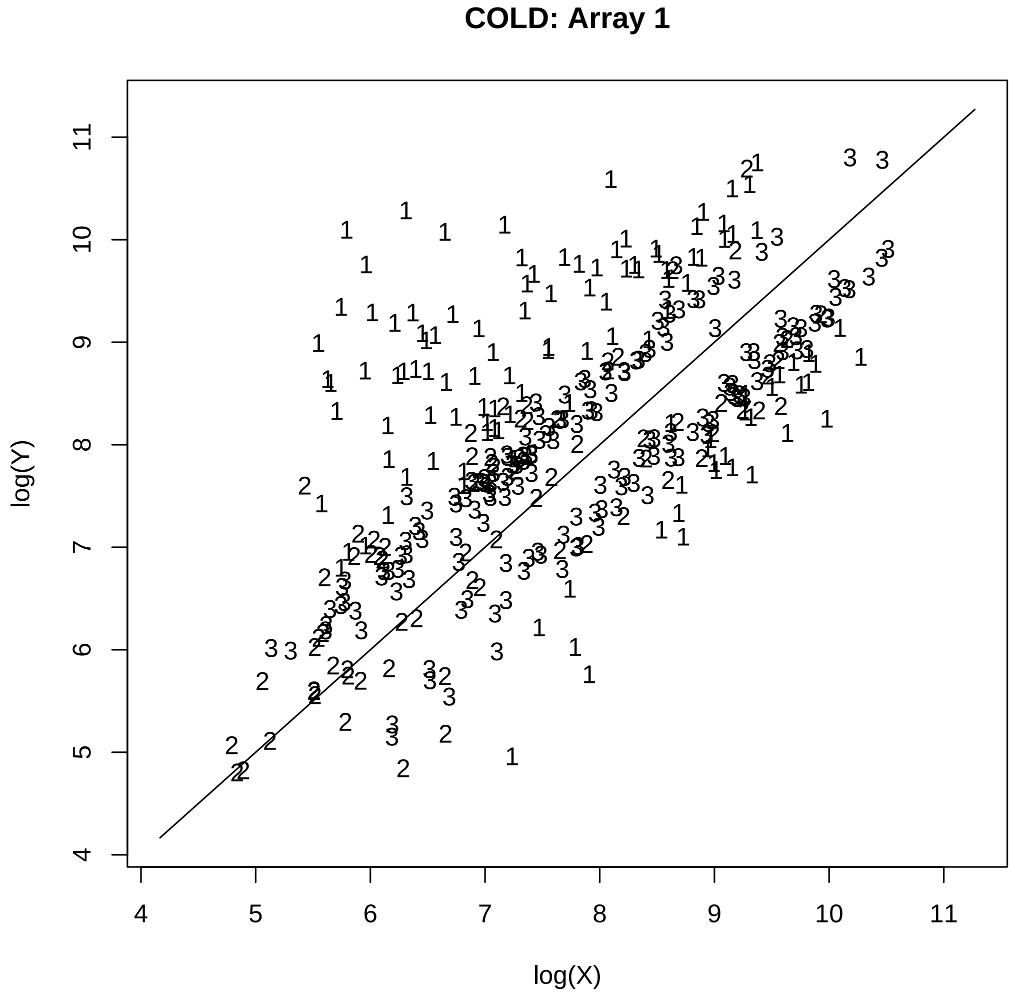


Figure B.1: Genes detected: COLD dataset. The symbol "1" represents the genes detected by the inverse-Gaussian model. Symbols "1" and "2" represent the genes detected by the GAM model (including those detected by the inverse-Gaussian model). Symbols "1", "2" and "3" represent the genes detected by the multiplicative gamma model

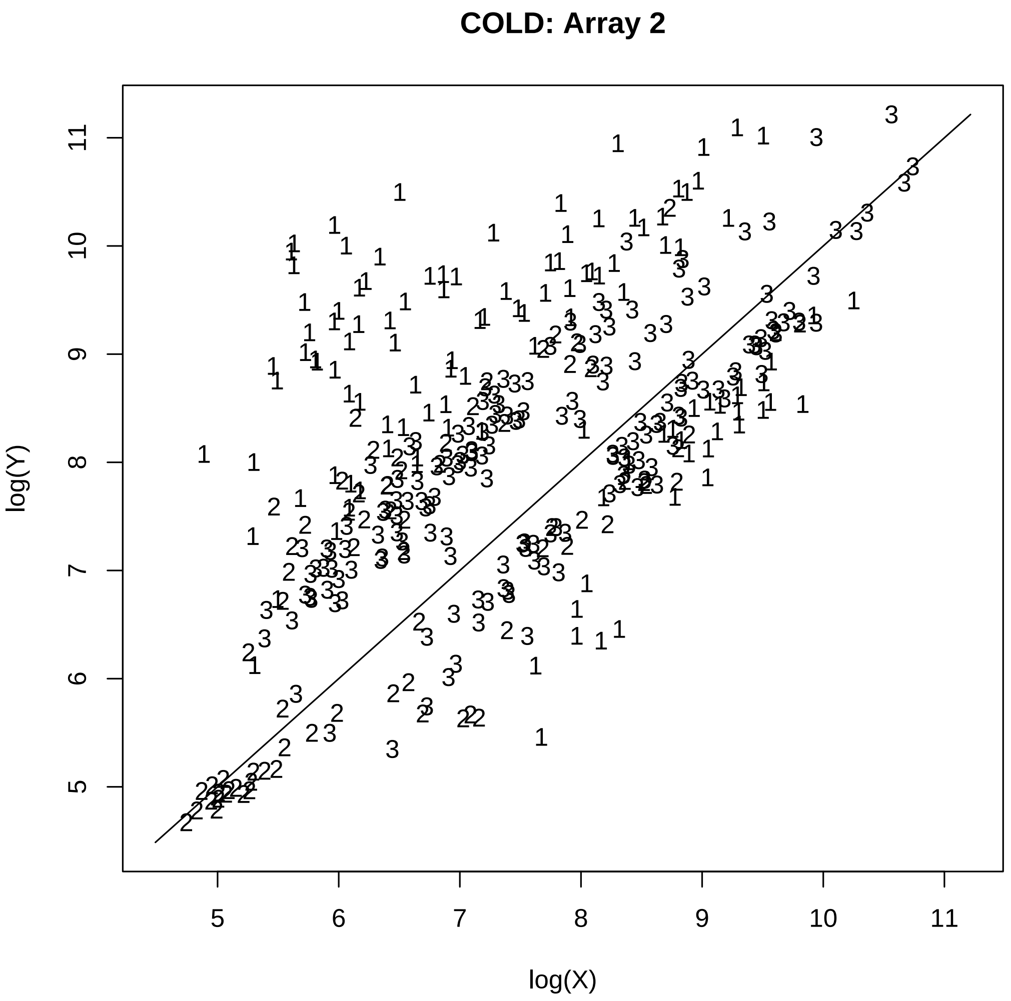


Figure B.2: Genes detected: COLD dataset. The symbol "1" represents the genes detected by the inverse-Gaussian model. Symbols "1" and "2" represent the genes detected by the GAM model (including those detected by the inverse-Gaussian model). Symbols "1", "2" and "3" represent the genes detected by the multiplicative gamma model

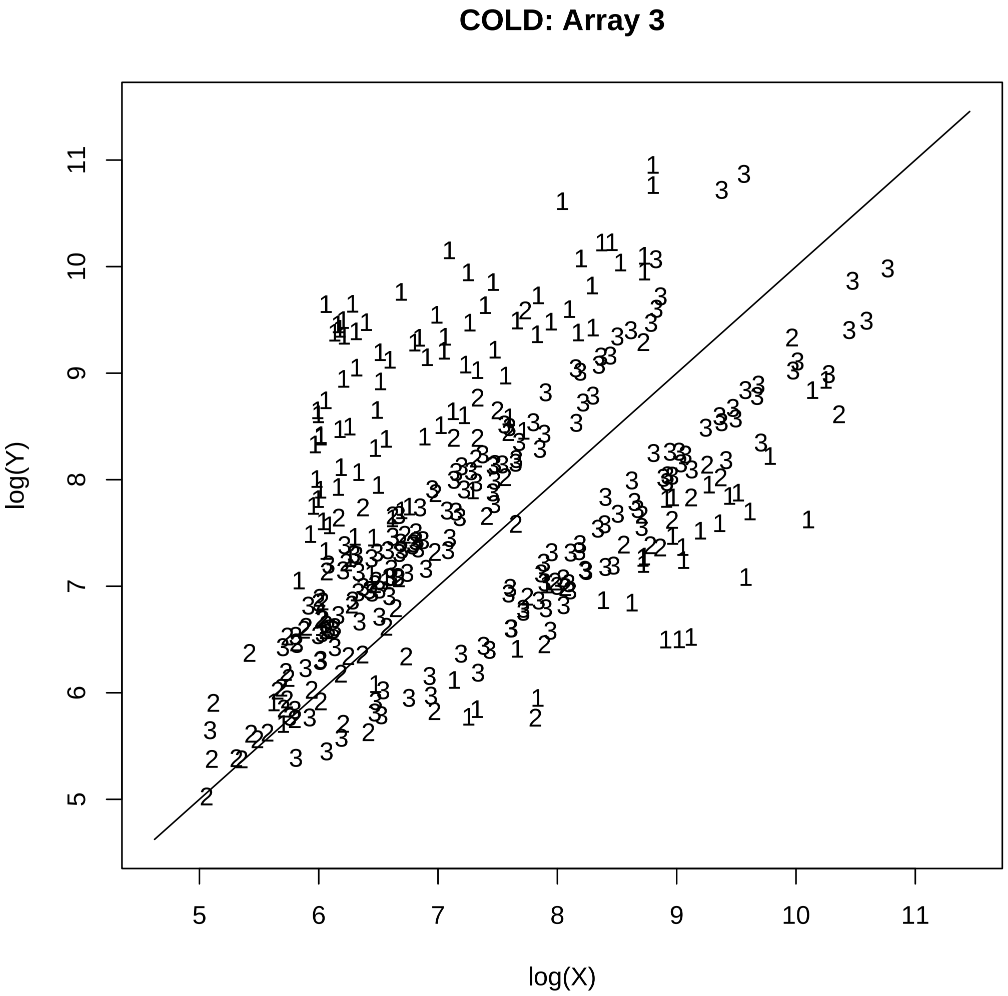


Figure B.3: Genes detected: COLD dataset. The symbol "1" represents the genes detected by the inverse-Gaussian model. Symbols "1" and "2" represent the genes detected by the GAM model (including those detected by the inverse-Gaussian model). Symbols "1", "2" and "3" represent the genes detected by the multiplicative gamma model

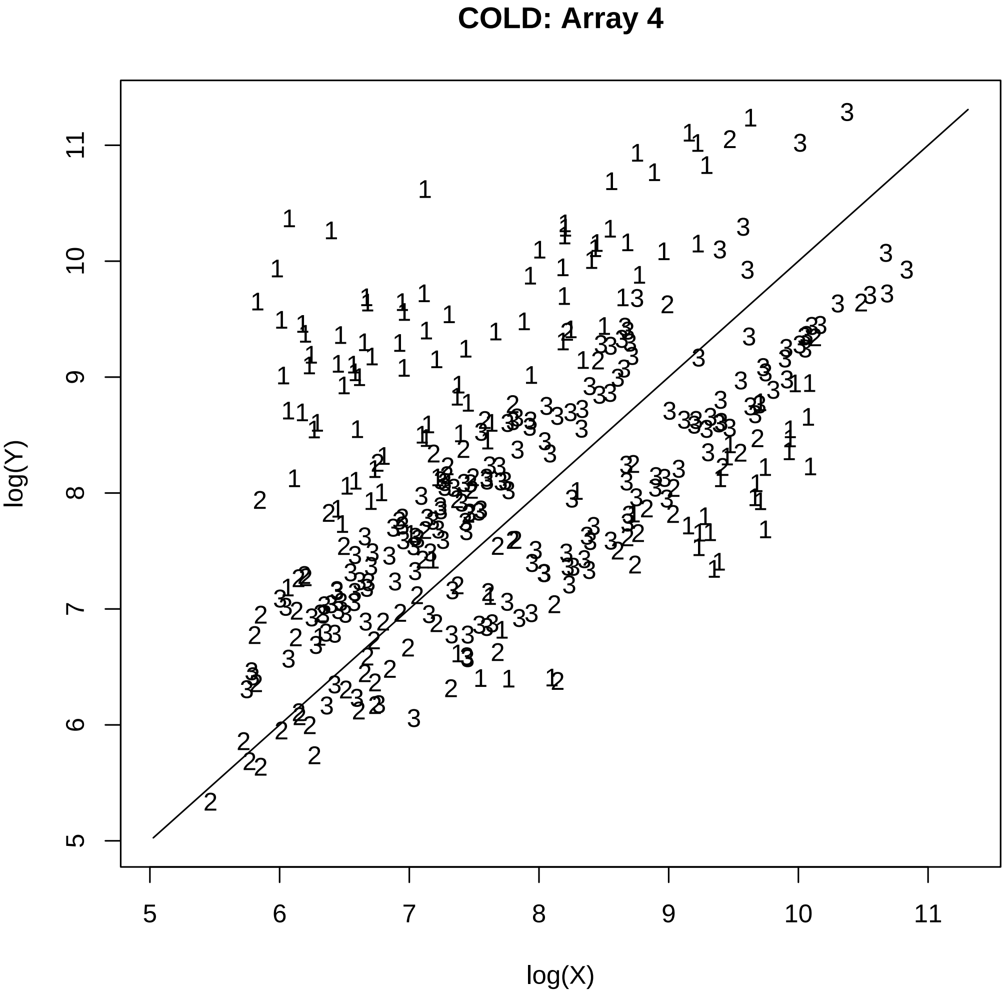


Figure B.4: Genes detected: COLD dataset. The symbol "1" represents the genes detected by the inverse-Gaussian model. Symbols "1" and "2" represent the genes detected by the GAM model (including those detected by the inverse-Gaussian model). Symbols "1", "2" and "3" represent the genes detected by the multiplicative gamma model

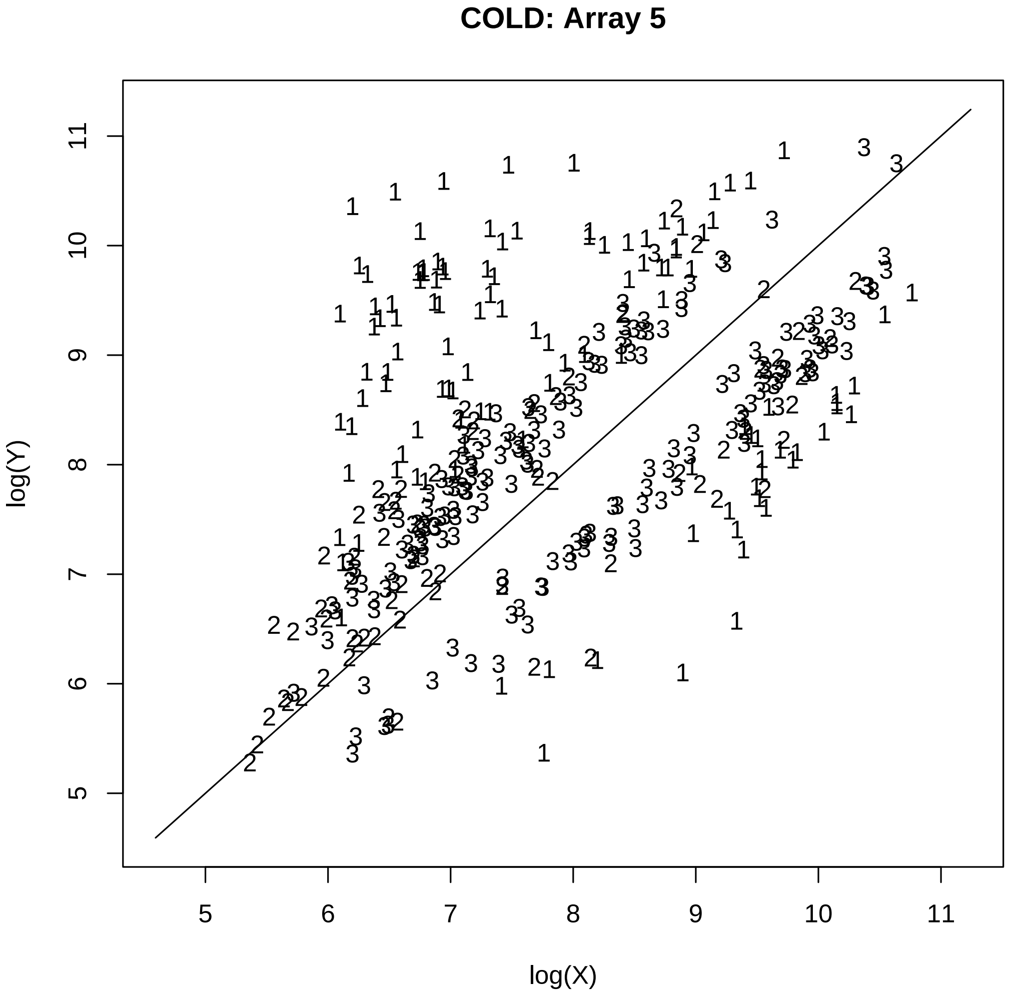


Figure B.5: Genes detected: COLD dataset. The symbol "1" represents the genes detected by the inverse-Gaussian model. Symbols "1" and "2" represent the genes detected by the GAM model (including those detected by the inverse-Gaussian model). Symbols "1", "2" and "3" represent the genes detected by the multiplicative gamma model

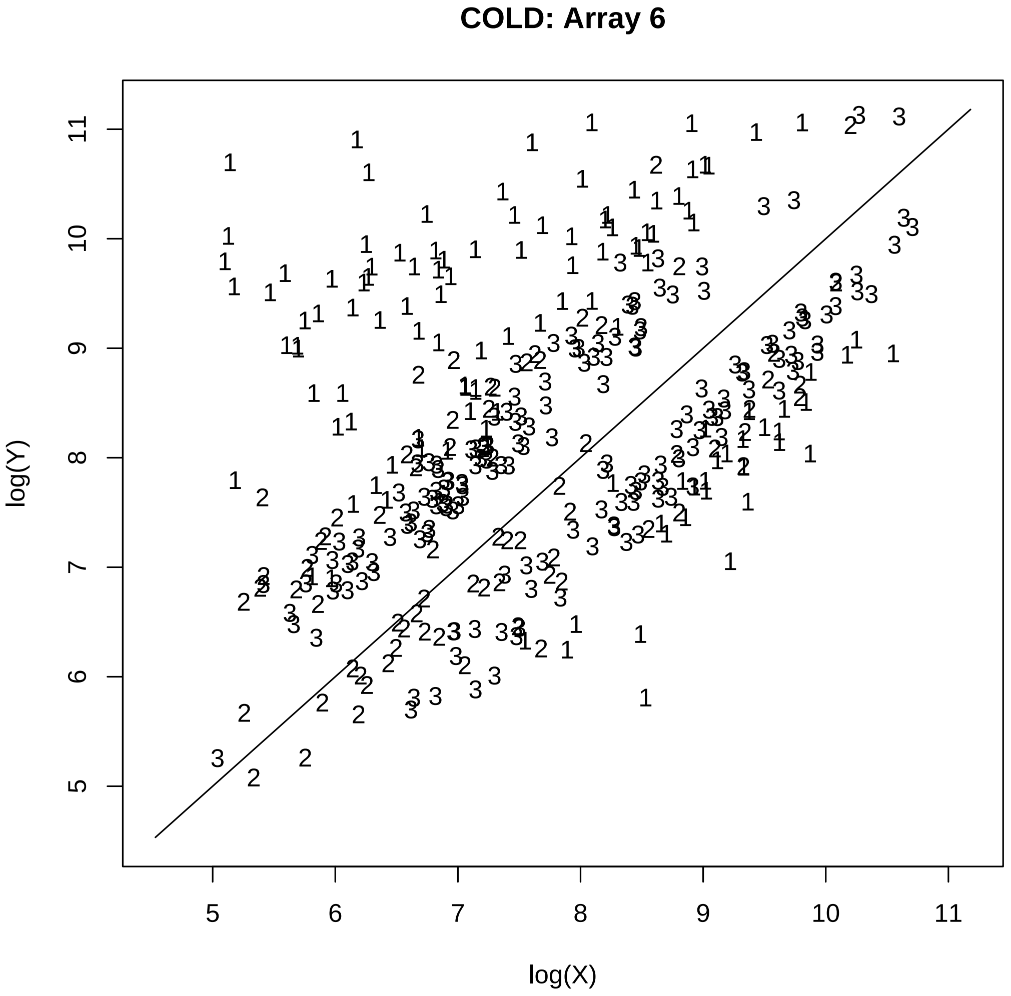


Figure B.6: Genes detected: COLD dataset. The symbol "1" represents the genes detected by the inverse-Gaussian model. Symbols "1" and "2" represent the genes detected by the GAM model (including those detected by the inverse-Gaussian model). Symbols "1", "2" and "3" represent the genes detected by the multiplicative gamma model

Appendix C

SALINITY dataset: Plots of the gene detected

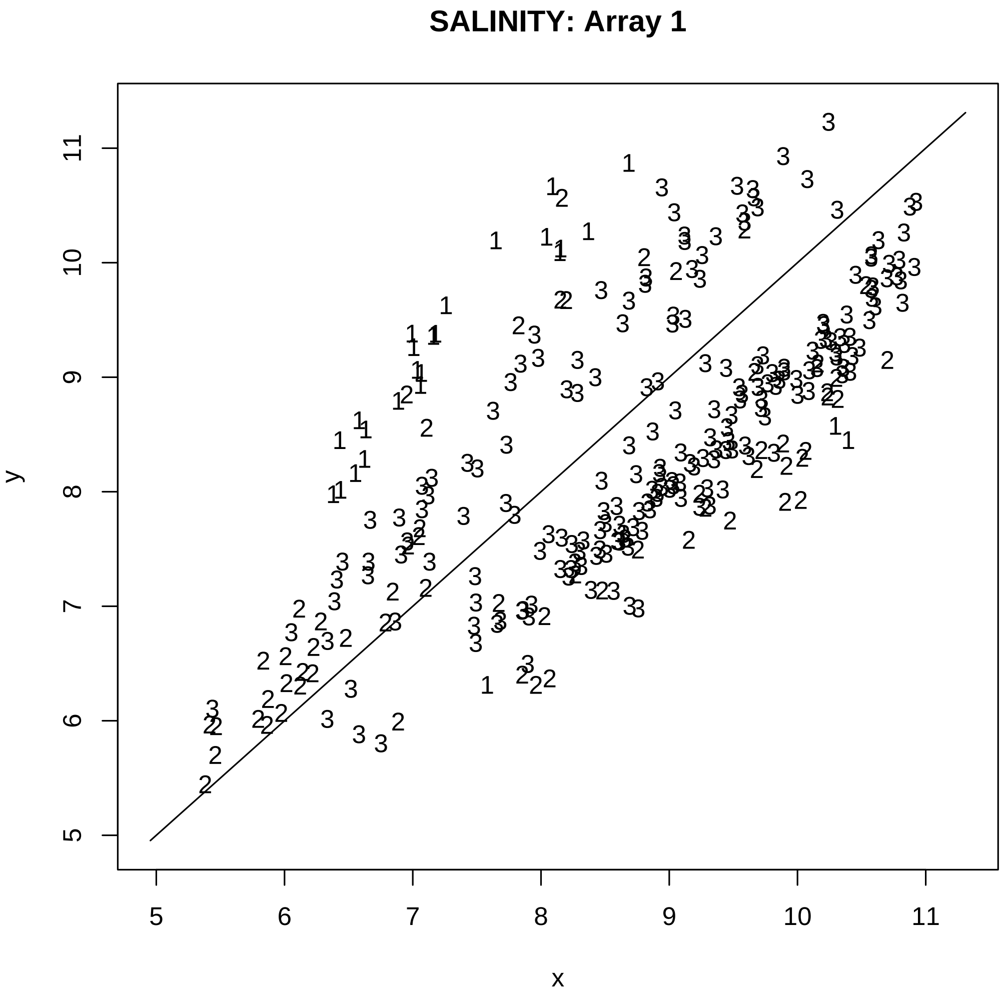


Figure C.1: Genes detected: SALINITY dataset. The symbol ”1” represents the genes detected by the inverse-Gaussian model. Symbols ”1” and ”2” represent the genes detected by the GAM model (including those detected by the inverse-Gaussian model). Symbols ”1”, ”2” and ”3” represent the genes detected by the multiplicative gamma model

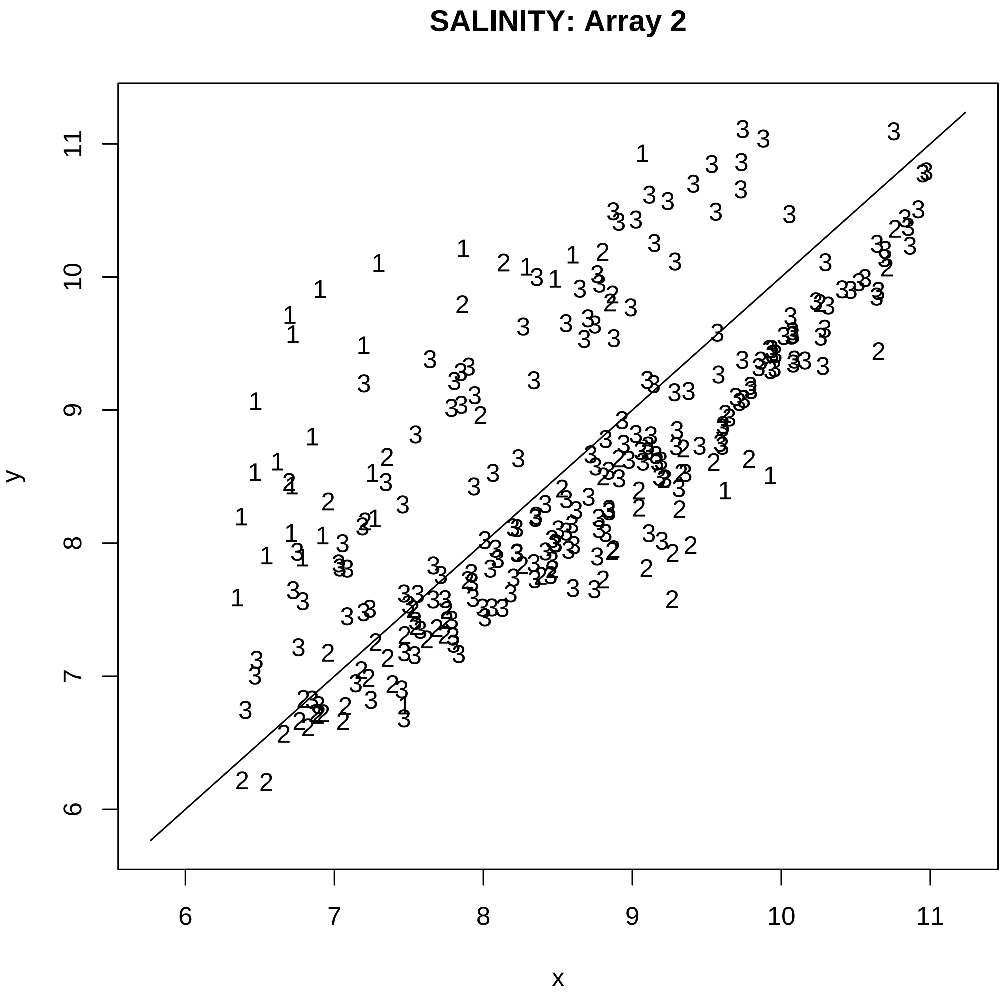


Figure C.2: Genes detected: SALINITY dataset. The symbol "1" represents the genes detected by the inverse-Gaussian model. Symbols "1" and "2" represent the genes detected by the GAM model (including those detected by the inverse-Gaussian model). Symbols "1", "2" and "3" represent the genes detected by the multiplicative gamma model

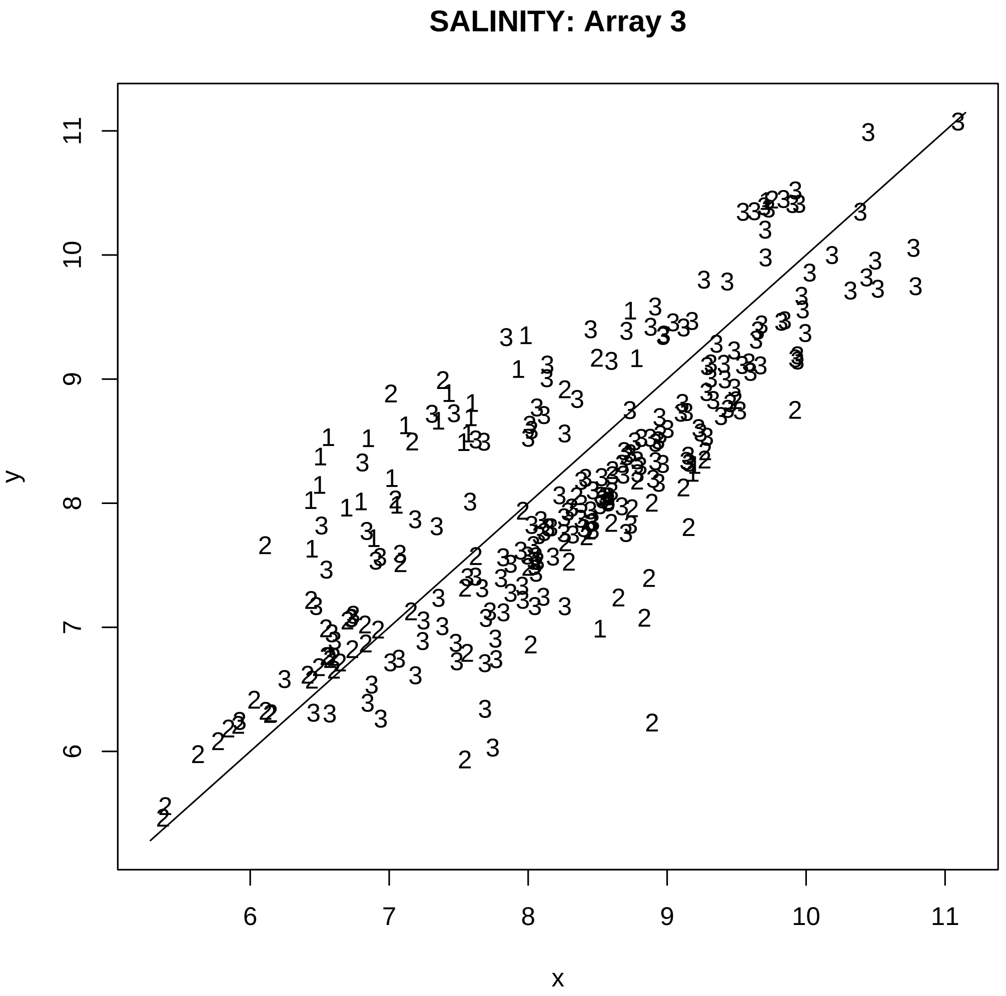


Figure C.3: Genes detected: SALINITY dataset. The symbol "1" represents the genes detected by the inverse-Gaussian model. Symbols "1" and "2" represent the genes detected by the GAM model (including those detected by the inverse-Gaussian model). Symbols "1", "2" and "3" represent the genes detected by the multiplicative gamma model

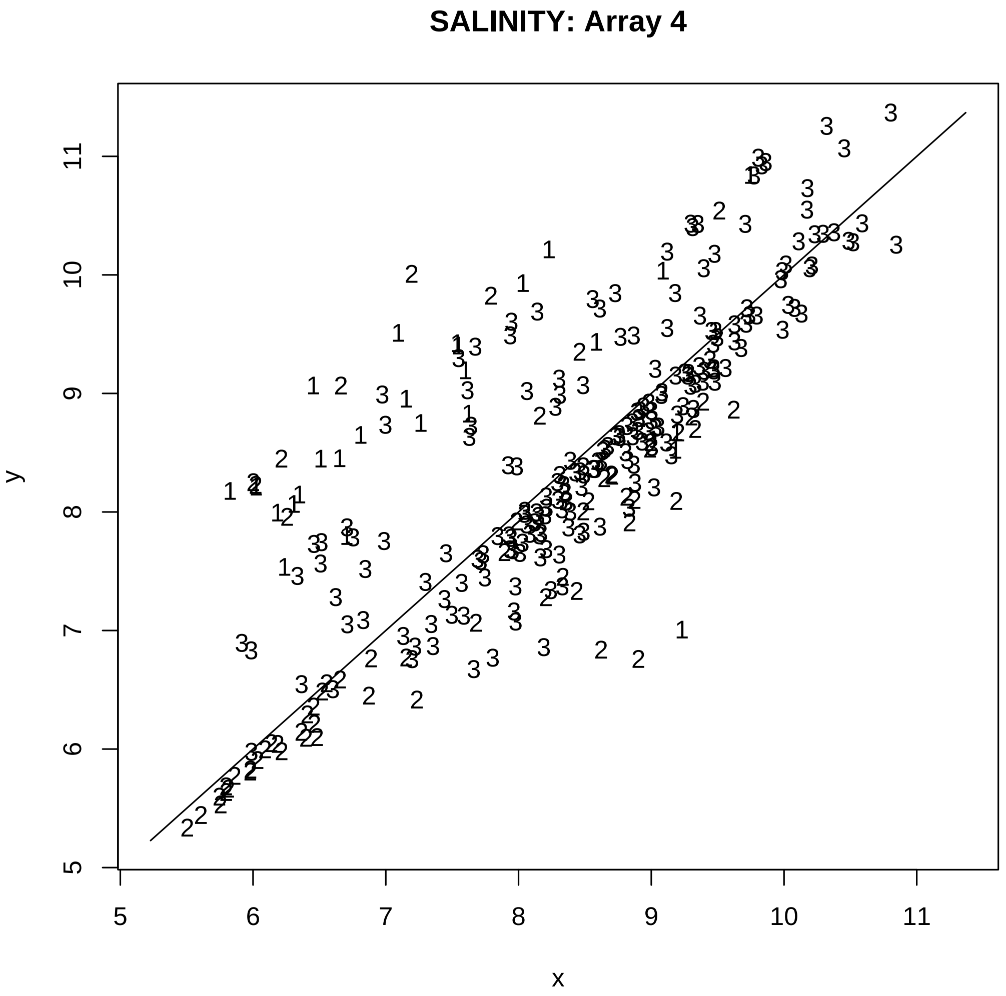


Figure C.4: Genes detected: SALINITY dataset. The symbol "1" represents the genes detected by the inverse-Gaussian model. Symbols "1" and "2" represent the genes detected by the GAM model (including those detected by the inverse-Gaussian model). Symbols "1", "2" and "3" represent the genes detected by the multiplicative gamma model

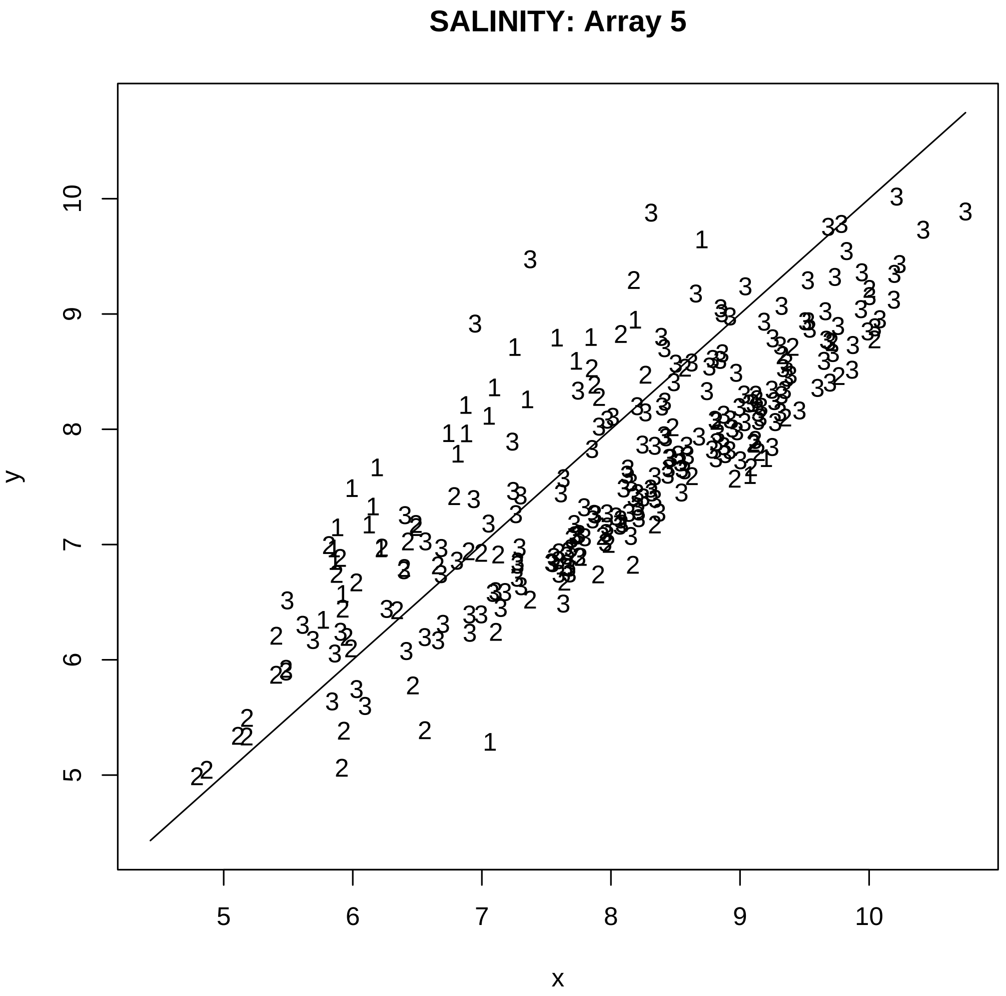


Figure C.5: Genes detected: SALINITY dataset. The symbol "1" represents the genes detected by the inverse-Gaussian model. Symbols "1" and "2" represent the genes detected by the GAM model (including those detected by the inverse-Gaussian model). Symbols "1", "2" and "3" represent the genes detected by the multiplicative gamma model

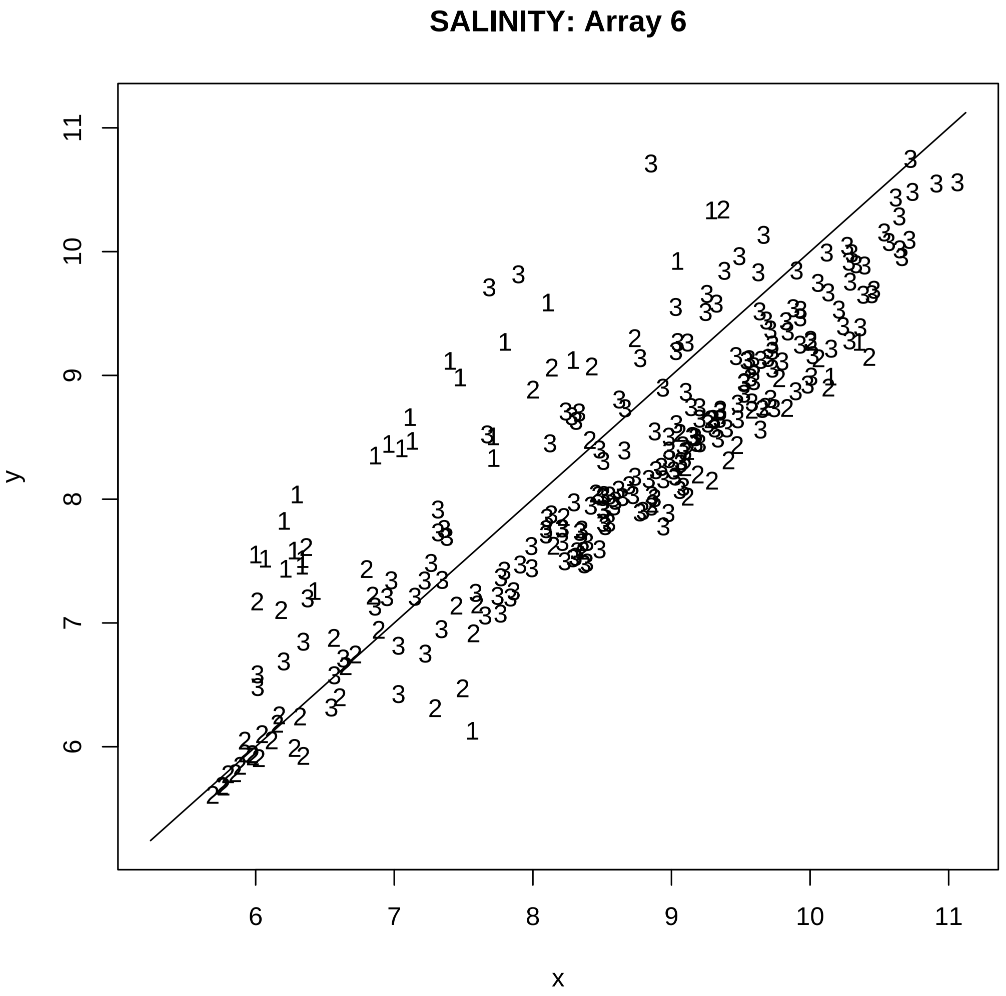


Figure C.6: Genes detected: SALINITY dataset. The symbol "1" represents the genes detected by the inverse-Gaussian model. Symbols "1" and "2" represent the genes detected by the GAM model (including those detected by the inverse-Gaussian model). Symbols "1", "2" and "3" represent the genes detected by the multiplicative gamma model

Appendix D

DROUGHT dataset: Plots of the gene detected

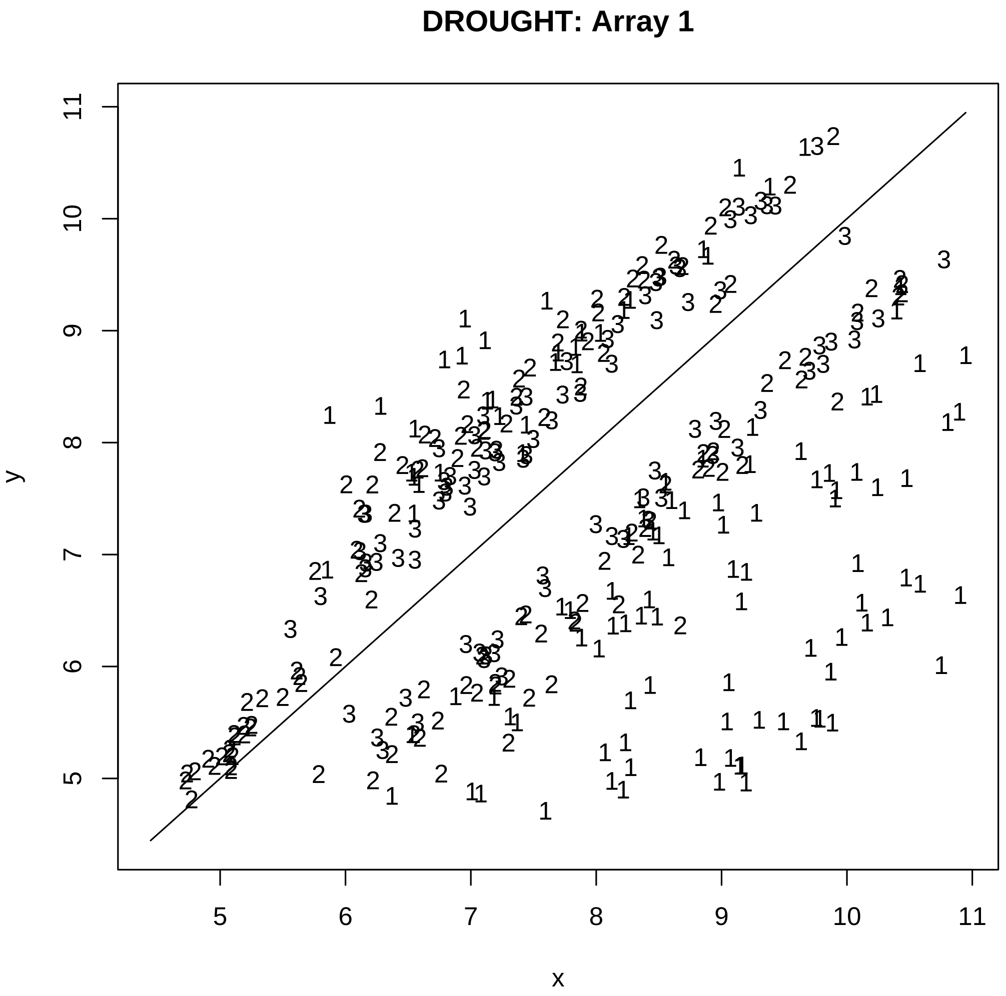


Figure D.1: Genes detected: DROUGHT dataset. The symbol ”1” represents the genes detected by the inverse-Gaussian model. Symbols ”1” and ”2” represent the genes detected by the GAM model (including those detected by the inverse-Gaussian model). Symbols ”1”, ”2” and ”3” represent the genes detected by the multiplicative gamma model

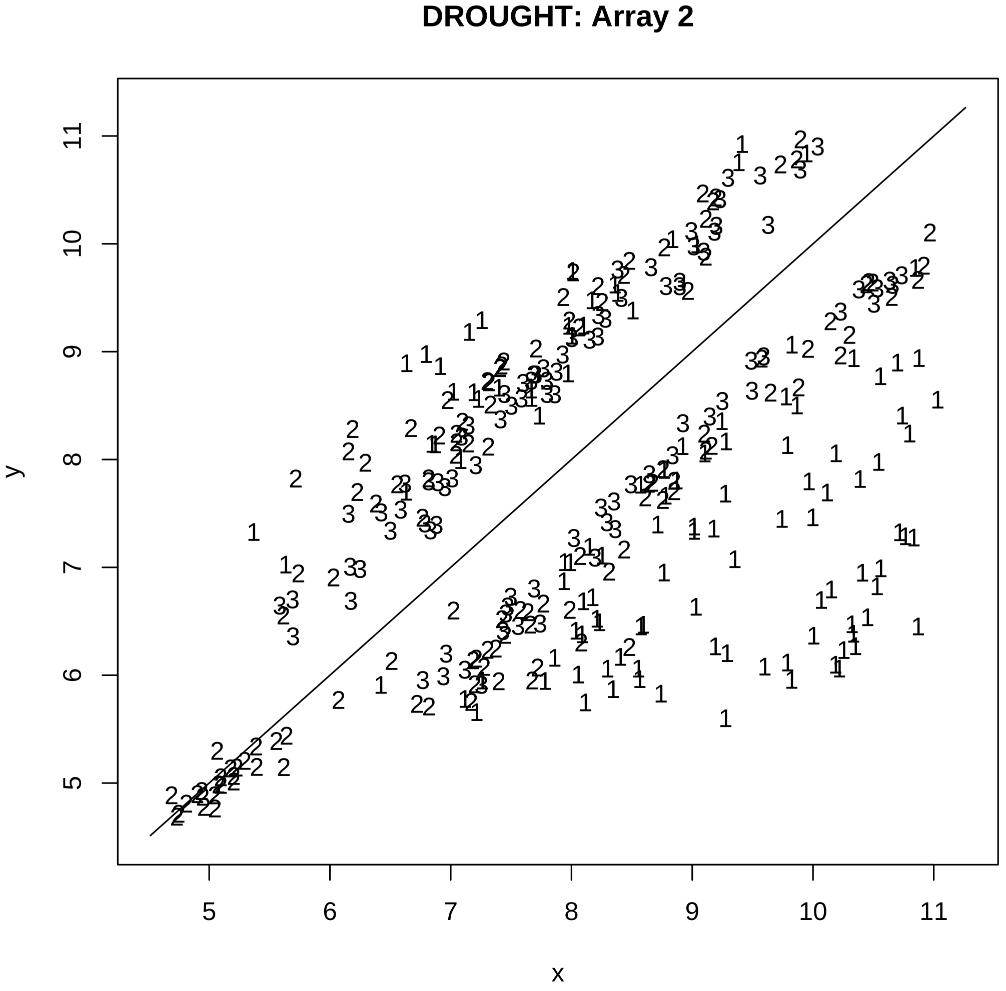


Figure D.2: Genes detected: DROUGHT dataset. The symbol "1" represents the genes detected by the inverse-Gaussian model. Symbols "1" and "2" represent the genes detected by the GAM model (including those detected by the inverse-Gaussian model). Symbols "1", "2" and "3" represent the genes detected by the multiplicative gamma model

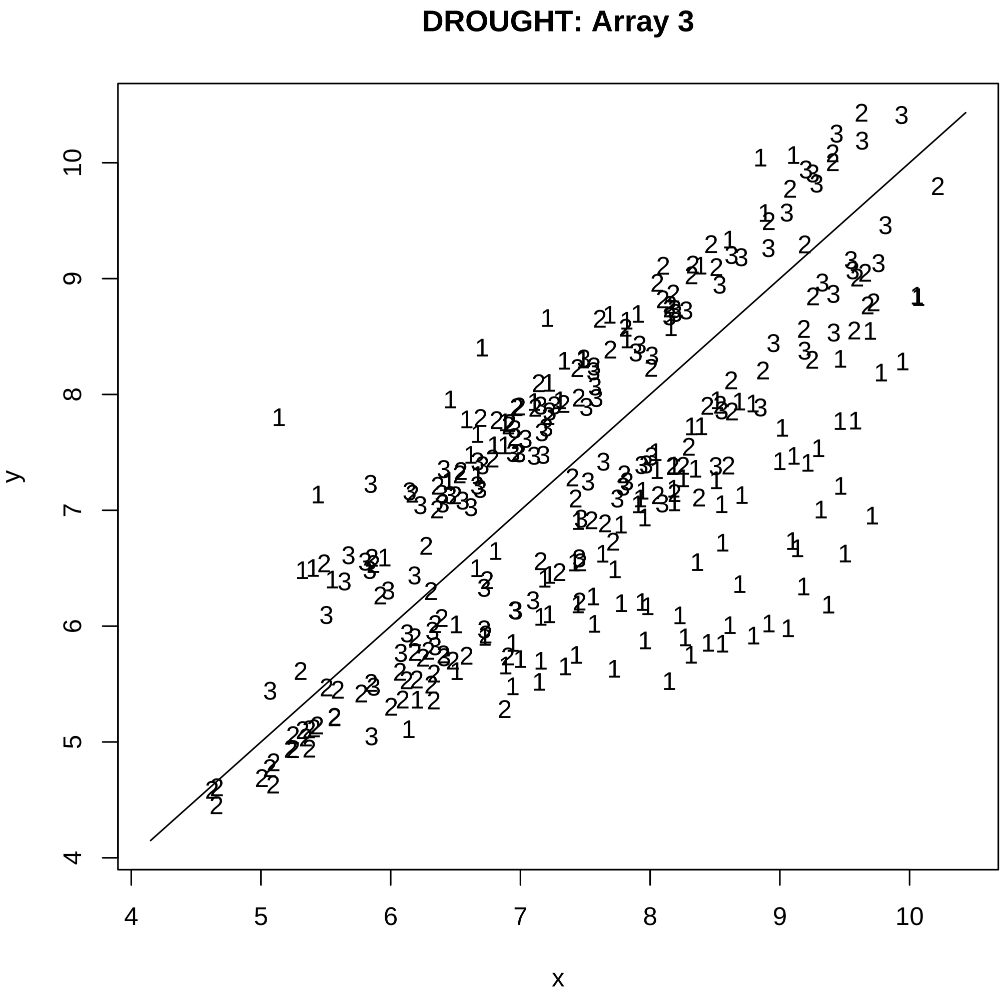


Figure D.3: Genes detected: DROUGHT dataset. The symbol ”1” represents the genes detected by the inverse-Gaussian model. Symbols ”1” and ”2” represent the genes detected by the GAM model (including those detected by the inverse-Gaussian model). Symbols ”1”, ”2” and ”3” represent the genes detected by the multiplicative gamma model

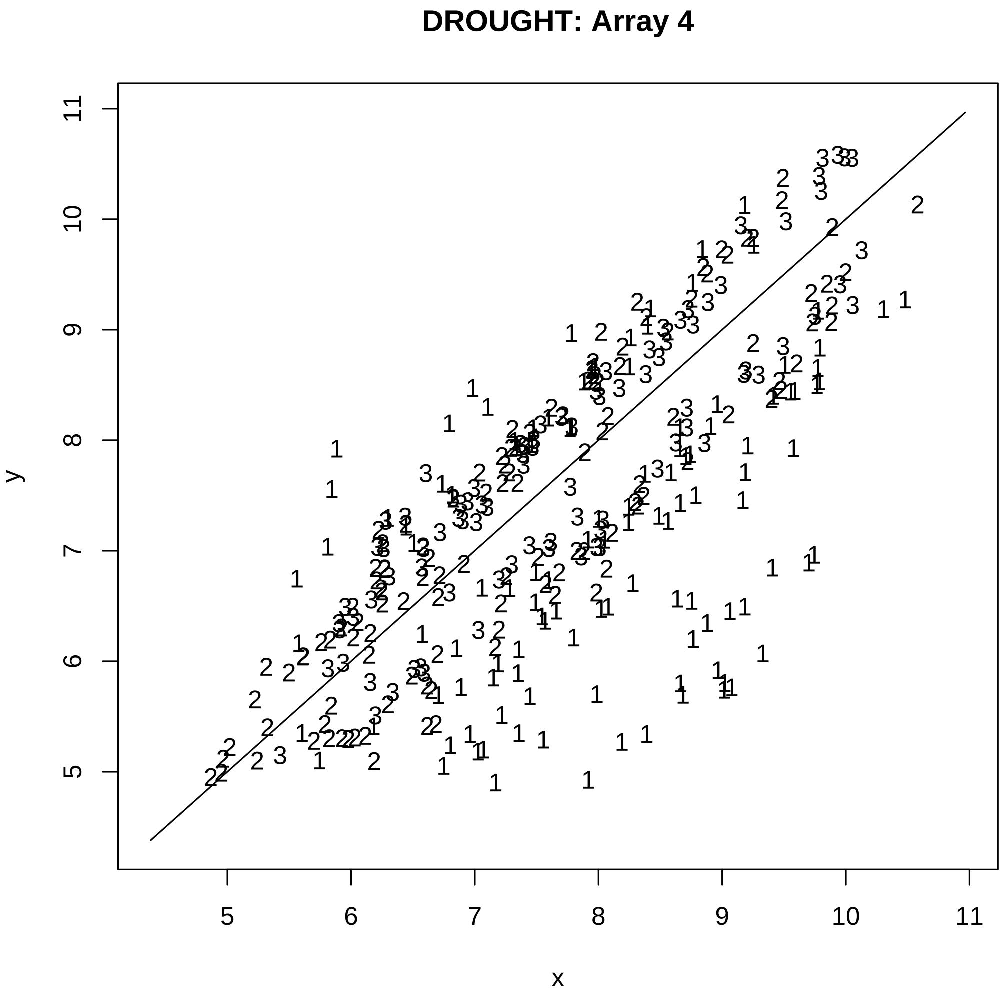


Figure D.4: Genes detected: DROUGHT dataset. The symbol "1" represents the genes detected by the inverse-Gaussian model. Symbols "1" and "2" represent the genes detected by the GAM model (including those detected by the inverse-Gaussian model). Symbols "1", "2" and "3" represent the genes detected by the multiplicative gamma model

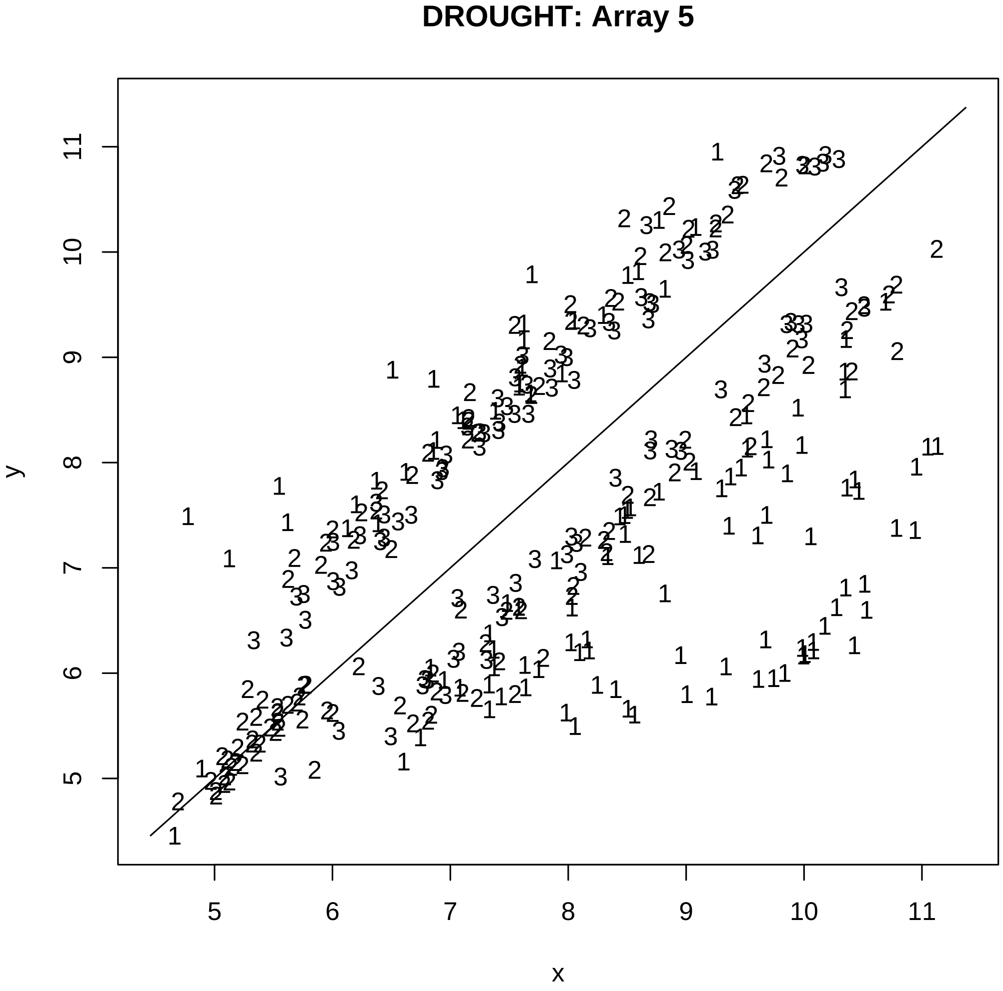


Figure D.5: Genes detected: DROUGHT dataset. The symbol "1" represents the genes detected by the inverse-Gaussian model. Symbols "1" and "2" represent the genes detected by the GAM model (including those detected by the inverse-Gaussian model). Symbols "1", "2" and "3" represent the genes detected by the multiplicative gamma model

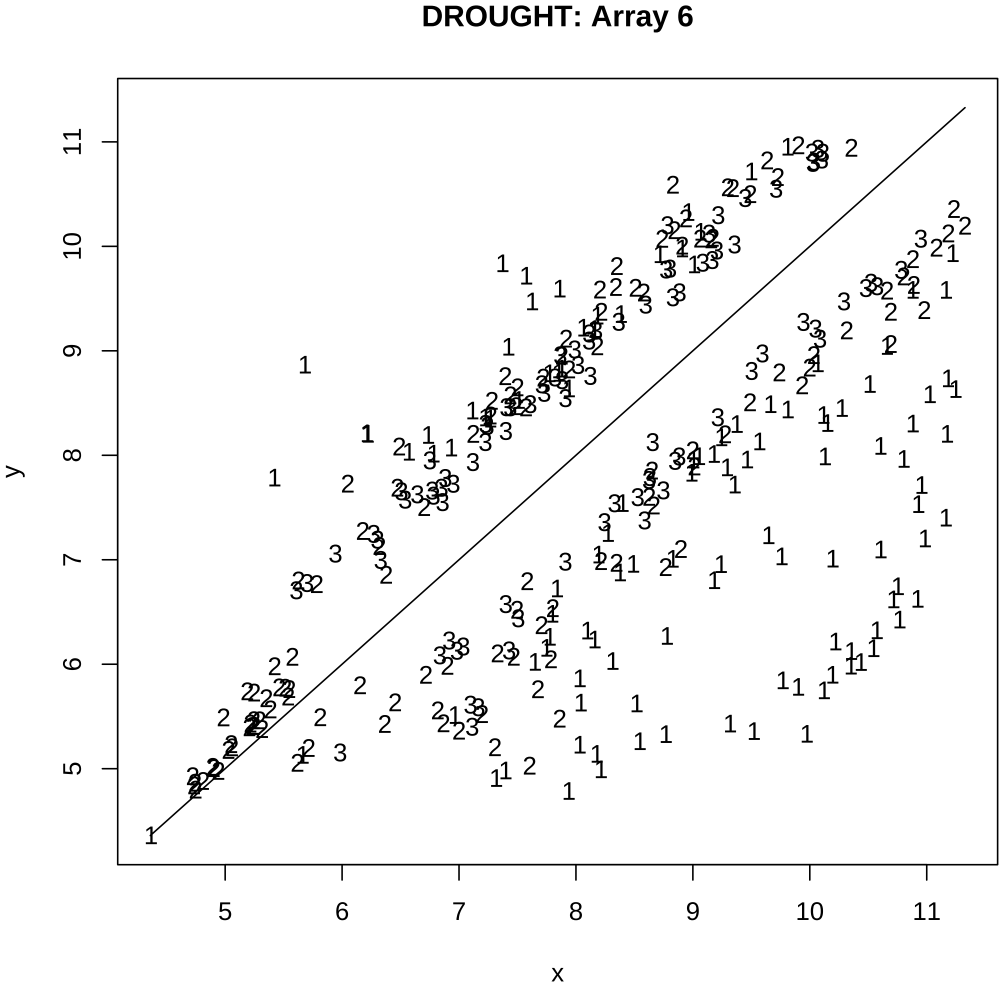


Figure D.6: Genes detected: DROUGHT dataset. The symbol "1" represents the genes detected by the inverse-Gaussian model. Symbols "1" and "2" represent the genes detected by the GAM model (including those detected by the inverse-Gaussian model). Symbols "1", "2" and "3" represent the genes detected by the multiplicative gamma model

References

Banerjee, A. & Bhattacharyya, G. 1979, 'Bayesian results for the inverse Gaussian distribution with an application', *Technometrics* **21**, 247–251.

Benjamini, Y. & Hochberg, Y. 1995, 'Controlling the False Discovery Rate : a practical and powerful approach to multiple testing', *Journal of the Royal Statistical Society* **57**, 289–300.

Berger, J. 1980, *Statistical decision theory : Foundations, concepts and methods*, Springer-Verlag New York Inc.

Betro, B. & Rotondi, R. 1991, 'On Bayesian inference for the inverse Gaussian distribution', *Statistics and Probability Letters* **11**, 219–224.

Bressan, R., Zhang, C., Zhang, H., Hasegawa, P., Bohnert, H. & Zhu, J. 2002, 'Learning from the Arabidopsis experience. The nect gene research paradigm', *Plant Physiology* **127**, 1354–1360.

Bussemaker, H., Li, H. & Siggia, E. 2001, 'Regulatory element detection using correlation with expression', *Nature Genetics* **27**, 167–171.

Chhikara, R. 1986, *Inverse Gaussian distribution: theory, methodology and applications*, New York: M. Dekker.

Conlon, E., Liu, X., Lieb, J. & Liu, J. 2003, 'Integrating regulatory motif discovery and genome-wide expression analysis', *Proc' Nat'l Acad. Sci. USA* **100**(6), 3339–3344.

Cox, D. & Hinkley, D. 1974, *Theoretical Statistics*, Chapman and Hall, London.

Cui, H., Kerr, K. & Churchill, G. 2002, Data transformation for cDNA microarray data. Submitted.

Cui, X. 2004, Statistical tests for differential expression in cDNA microarray. Submitted to Genome Biology.

Dudley, R. & Haughton, D. 2002, 'Asymptotic normality with small relative errors of posterior probabilities of half-spaces', *The Annals of Statistics* **30**(5), 1311–1344.

Dudoit, S., Fridlyand, J. & Speed, T. 2002, 'Comparison of discrimination methods for the classification of tumors using gene expression data', *Journal of the American Statistical Association* **97**(457), 77–87.

Dudoit, S., Shaffer, J. & Boldrick, J. 2003, 'Multiple hypothesis testing in microarray experiments', *Statistical Sciences* **18**(1), 71–103.

Dudoit, S., van der Laan, M. & Pollard, K. 2004, 'Multiple testing. Part I. Single-step procedures for control of general Type I error rates', *Statistical Applications in Genetics and Molecular Biology* **3**(1). Article 13.

Efron, B. 2003, 'Robbins, Empirical Bayes and microarrays', *Annals of Statistics* **31**(2), 366–378.

Efron, B., Storey, J. D. & Tibshirani, R. 2001, Microarrays, Empirical Bayes methods and False Discovery Rate, Technical Report 218, Department of Statistics, Stanford university.

Efron, B. & Tibshirani, R. 1986, *An introduction to the bootstrap*, CRC Press.

Gelman, A. & Tuerlinckx, F. 2000, 'Type S error rates for classical and Bayesian single and multiple comparison procedures', *Computational Statistics* **15**, 373–390.

Genovese, C. & Wasserman, L. 2002*a*, A large sample approach to False Discovery Rates, Technical report, Department of Statistics, Carnegie Mellon University, Pittsburgh.

Genovese, C. & Wasserman, L. 2002*b*, 'Operating characteristics and extensions of the FDR procedure', *Journal of the Royal Statistical Society* **64**, 499–518.

Haab, B., Dunham, M. & Brown, P. 2001, 'Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions', *Genome Biology* **2**(2), research 0004.1–0004.13.

Ibrahim, J., Chen, M. & Gray, R. 2002, 'Bayesian models for gene expression with DNA microarray data', *Journal of the American Statistical Association* **97**(457), 88–99.

Ideker, T., Thorsson, V., Siegel, A. & Hood, L. 2000, 'Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data', *Journal of Computational Biology* **6**, 805–817.

Keles, S., Van der Laan, M., Dudoit, S., Xing, B. & Eisen, M. 2003, Supervised detection of regulatory motifs in DNA sequences, Technical report, Division of Biostatistics, U. of California, Berkeley.

Kendziorski, C., Newton, M., Lan, H. & Gould, M. 2002, On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles, Technical Report 166, University of Wisconsin, Department of Biostatistics.

Kerr, M., Afshari, C., Bennett, L., Bushel, P., Martinez, J., Walker, N. & Churchill, G. 2002, 'Statistical analysis of a gene expression microarray experiment with replication', *Statistica Sinica* **12**, 203–217.

Kerr, M. & Churchill, G. 2001a, 'Experimental design for gene expression microarrays', *Biostatistics* **2**, 183–202.

Kerr, M., Churchill, G. & Martin, M. 2001b, 'Analysis of variance for gene expression microarray data', *Journal of Computational Biology* **7**, 819–837.

Littel, R., Milliken, G., Stroup, W. & Wolfinger, R. 1996, *SAS system for mixed models*, SAS Institute Inc.

Liu, J., Gupta, M., Liu, X., Mayerhofere, L. & Lawrence, C. 2004, Statistical models for biological sequence motif discovery. To appear in Case Study in Bayesian Statistics, 6.

Liu, J., Neuwald, A. & Lawrence, C. 1995, 'Bayesian models for multiple local sequence alignment and Gibbs sampling strategies', *Journal of the American Statistical Association* **90**, 1156–1170.

Liu, X., Brutlag, D. & Liu, J. 2002, 'An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments', *Nature Biotechnology* **20**, 835–839.

Motoaki, S. & Mari, N. 2002, 'Monitoring the expression profiles of 7000 Arabidopsis genes under drought, cold and high salinity stresses using a full-length cDNA microarray', *The Plant Journal* **31**(3), 279–292.

Muller, P., Parmigiani, G., Robert, C. & Rousseau, J. 2004, 'Optimal sample size for multiple testing: the case of gene expression microarrays', *Journal of the American Statistical Association* . To appear.

Newton, M., Kendziorski, C., Richmond, C., Blattner, F. & Tsui, K. 2001, 'On differential variability of expression ratios : improving statistical inference about gene expression changes from microarray data.', *Journal of Computational Biology* **8**, 37–52.

Newton, M., Noueiry, A., Sarkar, D. & Ahlquist, P. 2003, Detecting differential gene expression with a semiparametric hierarchical mixture method, Technical Report 1074, University of Wisconsin-Madison, Department of statistics.

Nguyen, D., Arpart, A., Wang, N. & Carroll, R. 2002, 'Dna microarray experiments : biological and technological aspects', *Biometrics* **58**, 701–717.

Oswald, T., Avery Colin, M. & McCarty, M. M. 1944, 'Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Induction pf transformation by a Desoxyribonucleic Acid fraction isolated from Pneumococcus Type III', *Journal of Experimental Medicine* **79**(1), 137–160.

Scholten, D., Miron, A., Merchant, F., Miller, A., Miron, P., Iglehart, J. & Gentleman, R. 2004, 'Analyzing factorial designed microarray experiments', *Journal of Multivariate Analysis* **90**, 19–43.

Shaffer, J. 1999, 'A semi-Bayesian study of Duncan's Bayesian multiple comparison procedure', *Journal of Statistical Planning and Inference* **82**, 197–213.

Smyth, G. 2004, 'Linear models and empirical bayes methods for assessing differential expression in microarray experiments', *Statistical Applications in Genetics and Molecular*

Biology **3**(1), article 3.

Storey, J. 2002, 'A direct approach to False Discovery Rates', *Journal of the Royal Statistical Society* **64**, 479–498.

Storey, J. 2003, 'The positive False Discovery Rate : a Bayesian interpretation and the q-value', *Annals in Statistics* **31**(6), 2013–2035.

Storey, J., Taylor, J. & Siegmund, D. 2004, 'Strong control, conservative point estimation and simultaneous conservative consistency of False Discovery Rates: a unified approach', *Journal of the Royal Statistical Society* **66**, 187–205.

Van Der Laan, J., Dudoit, S. & Pollard, K. 2004, 'Multiple testing. Part II. Step-down procedures for control of the Family-Wise error Rate', *Statistical Applications in Genetics and Molecular Biology* **3**(1). Article 14.

Weaver, R. 2002, *Molecular Biology*, second edition edn, McGraw-Hill.

Westfall, P., Johnson, W. & Utts, J. 1997, 'A Bayesian perspective on the Bonferroni adjustment', *Biometrika* **84**, 419–427.

Wolfinger, R., Gibson, G., Wolfinger, E., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C. & Paules, R. 2001, 'Assessing gene significance from cDNA microarray expression data via mixed models', *Journal of computational Biology* **8**, 625–637.

Wu, H., Kerr, K., Cui, X. & Churchill, G. 2003, *The analysis of gene expression data: methods and software*, Springer, N.Y., chapter MAANOVA: A Software Package for the Analysis of Spotted cDNA Microarray Experiments.

Yang, Y., Dudoit, S., Luu, P., Lin, D., Peng, V., Ngai, J. & Speed, T. 2002, ‘Normalization for cDNA microarray data : a robust composite method addressing single and multiple slide systematic variation’, *Nucleic Acids research* **30**(15).