Contents

1 Introduction ....................................................... 1
  1.1 Introduction .......................................................... 1
  1.2 Application of Microarray Technology in Drug Discovery ....... 2
  1.3 Aim of the Dissertation ............................................... 3
  1.4 Datasets Used as Case Study ........................................... 3
    1.4.1 Microarray Experiment for a Two-dimensional Testing of Significance - Finding Differentially Expressed Genes .......... 4
    1.4.2 Gene-expression Data for Prediction of Clinical Outcome - Selection of Potential Biomarkers ................................. 5
  1.5 Organization of the Dissertation ....................................... 6

2 Introduction to Microarrays ........................................... 9
  2.1 Microarray Technology ................................................ 9
    2.1.1 cDNA Arrays ...................................................... 9
    2.1.2 Oligonucleotide Arrays ......................................... 11
  2.2 Statistical Analysis of Microarrays ................................... 12
    2.2.1 Pre-processing .................................................... 13
    2.2.2 Significance Analysis of Gene-expression Levels ................. 15
    2.2.3 Supervised and Unsupervised Classification ....................... 16
  2.3 Main Issues in Analyzing Gene-expression Data ..................... 16
    2.3.1 Multiplicity ....................................................... 17
    2.3.2 Resampling-based Inference ..................................... 18
## Classification Methods for Microarrays

### 3 Classification Methods for Microarrays: a Simulation Study

- **3.1 Introduction**  
- **3.2 Class-prediction Methods**  
- **3.3 Simulation Study**  
- **3.4 Results**
  - 3.4.1 Large Training Dataset  
  - 3.4.2 Small Training Dataset  
- **3.5 Discussion**  
- **3.6 Conclusions**

## Performance of Gene-selection and Classification Methods in a Microarray Setting: a Simulation Study

### 4 Performance of Gene-selection and Classification Methods in a Microarray Setting: a Simulation Study

- **4.1 Introduction**  
- **4.2 Methods**
  - 4.2.1 Gene Selection  
  - 4.2.2 Class Prediction  
- **4.3 Simulation Study**
  - 4.3.1 Simulation Setting  
- **4.4 Results**
  - 4.4.1 Normally Distributed Data With a Constant Treatment Effect  
  - 4.4.2 Other Settings  
  - 4.4.3 E- and P-criteria  
- **4.5 Discussion**  
- **4.6 Conclusions**

## Multiplicity Adjustment and Statistical Modelling in the Microarray Setting

### 5 Multiplicity Adjustment: an Overview

- **5.1 Introduction**  
- 5.1.1 Controlling Type I Error  
- 5.1.2 Obtaining $p$-values from Permutations  
- **5.2 Procedures Adjusting for Multiple Testing**
5.2.1 Procedures Controlling the FWER ........................................ 56
5.2.2 Procedures Controlling the FDR ........................................ 57
5.3 Discussion ................................................................. 61

6 Multiple Testing Procedures for the Comparison of Several Treatments With a Common Control in a Microarray Setting 63
6.1 Introduction .............................................................. 63
6.2 Comparing Several Treatments With the Control ................... 65
6.3 Application to the Data .................................................. 66
   6.3.1 Multiple Testing Using Dunnett $p$-values ......................... 66
   6.3.2 Resampling-based Multiple Testing ............................... 67
6.4 A Simulation Study ...................................................... 71
   6.4.1 Simulation Setting .................................................. 71
   6.4.2 Simulation Results ............................................... 72
6.5 Discussion ................................................................. 74

7 A Comparison of Procedures for the Selection of Fudge Factor for the SAM: a Simulation Study 77
7.1 Introduction .............................................................. 77
7.2 Selection of the Fudge Factor in SAM .................................. 78
   7.2.1 Graphical Interpretation of SAM ................................. 78
   7.2.2 Effect of the Fudge Factor on the FDR and Power .............. 81
   7.2.3 Simulation Setting .................................................. 81
   7.2.4 Simulation Results ............................................... 83
7.3 Other Methods for Selecting the Fudge Factor ....................... 90
   7.3.1 Penalized Linear Regression ...................................... 91
   7.3.2 ROC Method .......................................................... 91
   7.3.3 The Empirical Bayes Approach ................................... 93
7.4 Simulation Study ........................................................ 94
   7.4.1 Simulation Setting .................................................. 94
   7.4.2 Simulation Results ............................................... 95
7.5 Discussion ................................................................. 102
III Dose-response Modelling of Microarray Data in Drug Development Experiments 103

8 Testing for Trend in Dose-response Microarray Experiments: a Comparison of Several Testing Procedures 105

8.1 Introduction ................................................................. 105
8.2 Testing for Homogeneity of the Means Under Restricted Alternatives . 107
  8.2.1 Williams’ (1971, 1972) and Marcus’ (1976) Test Statistics . . . . 108
  8.2.2 Likelihood Ratio Test Statistic for Monotonicity
   (Barlow et al. 1972, and Robertson et al. 1988) .................... 109
  8.2.3 The $M$ Test Statistic of Hu et al. (2005) ....................... 110
  8.2.4 A Modification to the $M$ Test Statistic ......................... 110
8.3 Directional Inference .................................................... 110
  8.3.1 Directional Inference in Isotonic Regression .................... 110
  8.3.2 Control of the Directional FDR ................................. 112
8.4 Results ............................................................................. 115
  8.4.1 Number of Significant Findings for Each Statistic Using Different Multiple Testing Adjustments .......................... 115
  8.4.2 Comparison Between $E_{01}^2$, the $M$, and Modified $M$ Test Statistics 119
8.5 Simulation Study .............................................................. 124
  8.5.1 Standard Error Comparison ........................................ 124
  8.5.2 Power Study for a Single Gene Setting .......................... 127
  8.5.3 Power Study Under Multiple Testing Adjustment ............. 127
8.6 Discussion ......................................................................... 128

9 The IsoGene Library in R 131

9.1 Introduction to IsoGene Package .................................... 131
9.2 Testing for Trends: Testing Procedures, Multiplicity and Resampling-based Inference ................................................. 131
9.3 Using IsoGene Library ...................................................... 132
  9.3.1 Data Example ......................................................... 132
  9.3.2 Loading the Library ................................................ 133
9.4 The IsoGene Functions .................................................. 133
  9.4.1 Exploring the Data ..................................................... 133
  9.4.2 Calculating the Test Statistics ................................. 135
9.4.3 Obtaining Raw p-values ......................................... 137
9.4.4 Plot of p-values for a Single Gene ................................. 138
9.4.5 BH/BY-FDR Procedures for Adjusting for Multiple Testing ... 139
9.5 The Help System .................................................. 142

10 Classification of Trends in Dose-response Microarray Experiments
   Using Information Theory Selection Methods 143
   10.1 Introduction .................................................. 143
   10.2 Classification of Trends Using Information Criteria .......... 147
       10.2.1 Classification of Trends Based on Posterior Probabilities 147
       10.2.2 Akaike Weights and Bayesian Posterior Model Probabilities 148
   10.3 Order Restricted Information Criterion ............................. 150
   10.4 Application to the Data ...................................... 152
       10.4.1 Classification of the Trends .............................. 152
   10.5 Simulation Study ............................................ 154
       10.5.1 The Influence of the Initial Inference Step .............. 154
       10.5.2 Correlated Data ....................................... 159
       10.5.3 Evaluation of the Four Information Criteria .............. 162
   10.6 Discussion .................................................. 164

11 A Ratio Test in Dose-response Microarray Experiments: Inference 165
   11.1 Introduction .................................................. 165
   11.2 Ratio Tests .................................................. 166
   11.3 Computation of Isotonic Means and MCTs ...................... 168
       11.3.1 Pool-adjacent-violator-algorithm for the Computation of Isotonic Means 168
       11.3.2 Single and Multiple Contrast Tests ....................... 168
   11.4 Test of Ratios Using Multiple Contrasts ...................... 173
       11.4.1 Multiple Ratios and Their Simultaneous Estimation .......... 173
   11.5 Application to the Data ...................................... 175
   11.6 Discussion .................................................. 178

12 Multiple Confidence Intervals for Selected Ratio Parameters Adjusted for the False Coverage-statement Rate 179
   12.1 Introduction .................................................. 179
   12.2 Confidence Intervals for Single and Multiple Ratios ........... 180
12.3 False Coverage-statement Rate (FCR) ........................................ 181
12.4 FCR-adjusted BH-selected CIs ................................................. 183
12.5 Application to the Data ......................................................... 183
12.6 Discussion ......................................................... 185

IV Selection and Evaluation of Biomarkers in a Pre-clinical Microarray Study 187

13 Introduction to Biomarker Identification in the Drug Discovery Using Microarrays 189
13.1 Introduction ......................................................... 189
13.2 Definitions and Regulatory Perspective on Validation of Surrogate Endpoints and Biomarker ......................................................... 190
13.3 Validation of Surrogate Endpoints in the Clinical Trial Setting: a Joint Model Approach ......................................................... 191
13.4 Selection of Biomarkers in the Microarray Setting ......................................................... 192
  13.4.1 Inference ......................................................... 194
  13.4.2 Therapeutic and Prognostic Biomarkers: a Graphical Interpretation ......................................................... 195
13.5 Evaluation of Biomarkers ......................................................... 197
  13.5.1 Prognostic Biomarkers ......................................................... 197
  13.5.2 Therapeutic Biomarkers ......................................................... 197
13.6 Application to the Data ......................................................... 199
  13.6.1 Treatment Effect on the Response ......................................................... 199
  13.6.2 Treatment Effect on the Gene-Expression: Testing for Therapeutic Biomarkers ......................................................... 200
  13.6.3 The Adjusted Association: Testing for Prognostic Biomarkers ......................................................... 204
13.7 Discussion ......................................................... 205

14 Identification of Biomarker Genes in a Microarray Setting: a Two-Stage Approach 207
14.1 The Two-Stage Modelling Approach ......................................................... 208
  14.1.1 The First-stage Analysis ......................................................... 208
  14.1.2 The Second-stage Analysis ......................................................... 208
14.2 Application to the Data ......................................................... 209
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>14.3</td>
<td>Therapeutic Biomarkers: an Intuitive Discussion</td>
<td>212</td>
</tr>
<tr>
<td>14.4</td>
<td>Discussion</td>
<td>214</td>
</tr>
<tr>
<td>15</td>
<td>Constructing a Joint Biomarker Profile</td>
<td>215</td>
</tr>
<tr>
<td>15.1</td>
<td>Construction of a Joint Biomarker Profile</td>
<td>215</td>
</tr>
<tr>
<td>15.2</td>
<td>Application to the Data</td>
<td>217</td>
</tr>
<tr>
<td>15.2.1</td>
<td>A Joint Prognostic Biomarker Profile</td>
<td>217</td>
</tr>
<tr>
<td>15.2.2</td>
<td>Joint Therapeutic Biomarker Profile</td>
<td>218</td>
</tr>
<tr>
<td>15.3</td>
<td>Discussion</td>
<td>219</td>
</tr>
<tr>
<td>16</td>
<td>Concluding Remarks and Future Research</td>
<td>221</td>
</tr>
<tr>
<td>16.1</td>
<td>Part I: Class Prediction</td>
<td>221</td>
</tr>
<tr>
<td>16.2</td>
<td>Part II: Multiple Testing Adjustment</td>
<td>222</td>
</tr>
<tr>
<td>16.3</td>
<td>Part III: Dose-response Study</td>
<td>222</td>
</tr>
<tr>
<td>16.4</td>
<td>Part IV: Evaluation of Biomarker</td>
<td>224</td>
</tr>
<tr>
<td>16.5</td>
<td>Final Remarks</td>
<td>225</td>
</tr>
</tbody>
</table>

Publications and Reports | 227 |
References | 229 |

A Simulation Results | 245 |
B Simulation Results | 253 |
Samenvatting | 257 |
List of Abbreviations

AIC : Akaike Information Criterion
ANOVA : Analysis of Variance
BIC : Bayesian Information Criterion
BH-FDR : Benjamini and Hochberg procedure for controlling FDR
BY-FDR : Benjamini and Yekutieli procedure for controlling FDR
BC-score : Between-Class scatter score
BW : Between-treatment sum of square and Within-treatment sum of square
CV : Coefficient of Variation
cDNA : complementary DeoxyriboNucleic Acid
DLDA : Diagonal Linear Discriminant Analysis
DQDA : Diagonal Quadratic Discriminant Analysis
EB : Empirical Bayes
ECDF : Empirical Cumulative Distribution Function
EGF : Epidermal Growth Factor
Extval : Extreme-value distribution based gene selection
FCR : False Coverage-statement Rate
FDR : False Discovery Rate
FNR : False Negative Rate
FPR : False Positive Rate
FWER : FamilyWise Error Rate
gFWER : generalized FamilyWise Error Rate
gFDR : generalized False Discovery Rate
kNN : \( k \) Nearest Neighbor
K-L information : Kullback-Leibler information
LRT : Likelihood Ratio Test
LDA : Linear Discriminaiant Analysis
mRNA : messenger Ribonucleic Acid
MED : Minimum Effective Dose
MM : MisMatch
MD-FDR : Mixed Direction False Discovery Rate
MCT : Multiple Contrast Test
ORIC : Order Restricted Information Criterion
PM : Perfect Match
PAM : Prediction Analysis of Microarray
PS : Prediction Strength
PAVA : Pool-Adjacent-Violator-Algorithm
QC : Quality-Control
RMA: robust multichip average approach
ROC : Receiver Operating Characteristic
RBI : Resampling-Based Inference
RF: Random Forest
SAM: Significance Analysis of Microarray
SCT: Single Contrast Test
s.c.: subcutaneous
SPCA: Supervised Principal Component Analysis
SVM: Support Vector Machine
Chapter 1

Introduction

1.1 Introduction

As a result of the global development of genome research, the genomes of more than 240,000 organisms have been sequenced, and well over 61 million genetic sequences have been deposited in international repositories (Benson et al. 2006). However, the biological functions of most of these genes remain unknown, or have been predicted only through homology to genes with functions that are better known. One way to determine the function of genes is through repeated measurements of their ribonucleic acid (RNA) transcripts; for example, knowing that a particular gene is expressed only in cardiac muscle and only under particular conditions implicitly gives us functional knowledge about that gene.

Functional genomics is the study of gene function through parallel expression measurements of a genome. The most common tools used to obtain these measurements include complementary deoxyribonucleic acid (cDNA) microarrays, oligonucleotide microarrays or serial analysis of gene-expression (SAGE). In this dissertation, we focus on the first two techniques, which become important tools for early drug discovery.

The microarray approach involves first isolating RNA or messenger RNA from appropriate biological samples, making the RNA (or a copy of it) fluorescent, hybridizing it to the microarray, washing off the excess, and scanning the microarray under laser light. As a result, massive data are produced from each array. They may contain more than a million measurements of gene-expression per biological sample (for example, the Affymetrix HGU133 plus 2 chip contains more than 1,300,000 different probes).
1.2 Application of Microarray Technology in Drug Discovery

Gene-expression studies provide molecular responses of specific target genes of interest from biological samples, for example, a patient’s blood, tumor tissue or circulating rare cell samples. Efforts are made to answer the question, which genes are expressed in a particular cell type of an organism, at a particular time, under particular conditions. Genes that are consistently up- or down- regulated are of highest interest.

The use of microarray in drug discovery is expanding, with several applications including (Butte 2002): biomarker determination, pharmacology, toxicogenomics, target selectivity, prognostic tests, and disease subclass determination.

- Biomarker determination aims at identifying genes that correlate with and predict another outcome, for example to predict the response to a drug which is related to the mechanism of action.

- In pharmacology, microarrays are used to determine differences in gene-expression in tissues exposed to various doses of compounds.

- In toxicogenomics, the goal is to find gene-expression patterns in a model tissue or organism exposed to a compound and to use the patterns as early predictors of adverse events in humans. For example, if the expression of some genes in the liver is influenced by the metabolism of a particular compound, the microarray technology can be used to find those genes.

- Target selectivity aims at characterizing a compound by the gene-expression pattern it provokes in a target tissue and then at comparing it with other compounds using these patterns. By examining the differences in gene activity between untreated and treated tumor cells - for example, those that were exposed to a compound inhibiting the signalling of the epidermal growth factor receptor (e.g. Iressa) - scientists may understand exactly how different therapies affect tumors and may become able to develop more effective treatments.

- Prognostic tests aim at finding a set of genes that can predict how a patient’s disease will progress. For example, if a combination of gene-expression measurements accurately distinguish metastatic from non-metastatic tumors, microarrays can be used to search for such a pattern.
• For disease subclass determination, microarrays can be used to find multiple subcategories of pathologies in a single clinical diagnosis. In the past, scientists have classified different types of cancers based on the organs in which the tumors originate. With the help of microarray technology, however, they may be able to further classify these types of cancers based on the patterns of gene activity in the tumor cells. Researchers will then be able to design treatment strategies targeted directly to each specific type of cancer.

1.3 Aim of the Dissertation

This dissertation is devoted to analyze the high dimensional microarray data. In particular, the following four types of microarray experiments with the specific questions of interest considered: (1) when gene-expressions are measured for two conditions/treatments, which classification method performs best as a diagnostic tool to classify arrays into two groups? (2) when gene-expression is measured for several treatment groups and one control, which statistical methods can be used to find differentially expressed genes? (3) when gene-expression is measured under a set of increasing doses, what is the dose-response relationship? what is the minimal effect dose (MED) and the shape of dose-response curve? what is the relative difference in means between the highest dose and the control dose? (4) when gene-expression, as well as a clinical outcome, are measured for samples under two treatments, which genes can serve as biomarkers to predict the clinical outcome?

Note that, when answering the questions formulated above, thousands of genes in the array need to be tested simultaneously using a few (biological) arrays/replicates. Thus, the issues related to multiple testing and to inference based on small sample size need to be addressed.

1.4 Datasets Used as Case Study

We consider two microarray experiments. One experiment involves measurements of gene-expression for increasing doses of several compounds. This dataset is used to illustrate the application of microarrays in pharmacology and dose-response experiments (questions 2 and 3 in Section 1.3). The other experiment measures the gene-expression as well as a clinical outcome - the occupancy of a particular receptor in the brain. This dataset is used to illustrate the application of microarray in
Chapter 1. Introduction

biomarker determination (question 4 in Section 1.3). Additionally, simulated data are used to illustrate the application of microarrays in prognostic tests (question 1 in Section 1.3).

1.4.1 Microarray Experiment for a Two-dimensional Testing of Significance - Finding Differentially Expressed Genes

In this microarray experiment, human epidermal squamous carcinoma cell line A431 was grown in Dulbecco’s modified Eagle’s medium, supplemented with L-glutamine (20 mM), Gentamycin (5 mg/ml) and 10% fetal bovine serum. The cells were pre-treated with four compounds (A, B, C, and control), which were stimulated with epidermal growth factor (EGF) (R&D Systems, 236-EG) at different concentrations (0 ng/ml, 1 ng/ml, 10 ng/ml and 100 ng/ml) for 24h. RNA was harvested using RLT buffer (Qiagen). All microarray related steps, including the amplification of total RNAs, labeling, hybridization and scanning were carried out as described in the GeneChip Expression Analysis Technical Manual, Rev.4 (Affymetrix 2004). Biotin-labeled target samples were hybridized to human genome arrays U133 A 2.0 containing probe sets interrogation approximately 22,000 transcripts from the UniGene database (Build 133). Hybridization was performed using 15 µg of complementary RNA for 16 h at 45°C under continuous rotation at 60 rpm. Arrays were stained in Affymetrix Fluidics stations using streptavidin/phycoerythrin staining. Thereafter, arrays were scanned with the Affymetrix scanner 3000, and images were analyzed using the GeneChip Operating System v1.1 (GCOS, Affymetrix). The collected data were quantile normalized in two steps: first within each sample group, and then across all sample groups obtained (Bolstad et al. 2002). After removing the probe sets with absent calls, the resulting data set consists of 48 samples, for four dose levels and three microarrays at each dose level, with 16,998 probe sets. For simplicity, we refer to probe sets as genes throughout the dissertation (Hubbell et al. 2002).

Table 1.1 illustrates the settings of the experiment. One dimension of the data is gene-expression, measured for each gene under four increasing EGF concentration levels (with the first concentration as the control) for a particular compound. The other dimension is gene-expression measured under four conditions (three different compounds and one control compound) for a particular EGF concentration. Hence, the dataset can be used to illustrate testing of dose-response relationship on one hand and to elaborate on comparisons of several treatments with one control on the other
Table 1.1: Number of arrays for case study 1.

<table>
<thead>
<tr>
<th>Setting</th>
<th>Dose (ng/ml)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Control</td>
<td>3</td>
</tr>
<tr>
<td>Treatment A</td>
<td>3</td>
</tr>
<tr>
<td>Treatment B</td>
<td>3</td>
</tr>
<tr>
<td>Treatment C</td>
<td>3</td>
</tr>
</tbody>
</table>

hand.

1.4.2 Gene-expression Data for Prediction of Clinical Outcome - Selection of Potential Biomarkers

In this microarray experiment, male Hanover Wistar rats 63-70 days of age and weighing ± 250 gram were housed at normal conditions in groups of four with normal light cycle (light on at 06:00 and off at 18:00) with free access to food and water. All animals were habituated for one week prior to start of the experiment and handled for two days before treatment and deprived of food for 12 hours preceding the first injection. The rats were treated with either test compound (2.5 or 10 mg / kg injected subcutaneously, s.c.), or a reference compound (20 mg / kg s.c.). For the test compound these doses produce 50% and 90% target occupancy after two hours and for the reference compound the dose gives 50% target occupancy after two hours. Animals were sacrificed two hours after administration of the drugs. Arterial blood was collected by cardiac puncture of the left ventricle for RNA extraction. Brains were dissected and frozen in isomethylbutane at −80°C for ex-vivo receptor binding analysis. As an endpoint, the percentage of target receptors in the brain that were occupied by the drug was measured. RNA was extracted using the RiboPure(tm)-BloodKit from Ambion according to manufacturers instructions. One µg of total RNA was biotin labeled and hybridized to Rat Genome 230 version 2.0 GeneChip arrays according to the instructions of the manufacturer (Affymetrix).

This pre-clinical microarray experiment discussed above consists of two primary endpoints, the receptor occupancy \( Y \), which measures the percentage of receptor occupancy in the brain by the drug and the gene-expression \( X \). Hence, one can ben-
fit by using an easily measurable genetic marker, without the use of the radiotracers, in order to know how much drug should be given to block the right percentage of the drug target. Therefore, a question of interest, in addition to the identification of the differentially expressed genes, is: can we identify some genes to act as biomarkers for the brain receptor occupancy?

1.5 Organization of the Dissertation

The content of the dissertation is organized as follows. In Chapter 2, a brief introduction to the microarray technology and a summary of methods for the analysis of microarray data is given.

The remainder of the dissertation consists of four parts. The first part presents the performance of classification methods for microarrays. In Chapter 3, we compare methods for classification of microarray samples using a simulation study. In Chapter 4, we extend the simulation study to investigate a number of gene selection techniques combined with the classification methods considered in Chapter 3.

The second part focuses on the multiplicity adjustment and statistical modeling in the microarray setting, in which we investigate the problem of comparing several treatments with one control. In Chapter 5, an overview of error rates and various procedures for controlling the error rates are given. In Chapter 6, we compare four procedures for testing several treatments with one control and conduct a simulation study to compare the performance of the four approaches. In Chapter 7, we compare the performance of several methods with respect to power and control of the False Discovery Rate (FDR) in the presence of small variance genes using a simulation study.

The third part focuses the topic on dose-response modelling of microarray data in the early drug development. In Chapter 8, we compare the performance of five test statistics for testing the null hypothesis of homogeneity of means at four dose levels against order restricted alternatives by using a case study, as well by using a simulation study. In Chapter 9, we discuss the implementation of the five test statistics discussed in Chapter 8 in IsoGene R package. In Chapter 10, based on the genes declared significant by the tests used in Chapter 8, we investigate the possible dose-response curve shape using the information theory. Additionally, we conduct simulation studies to compare the performance of several information criteria. In Chapter 11, we formulate the question about testing the dose-response relationship
in a different way, namely, by using a ratio test of means between the highest dose and the control under order restricted constraints. In Chapter 12, to proceed with the inference for the ratio test, construction of simultaneous confidence intervals for the selected parameters is discussed.

The fourth part of the dissertation focuses on selection and evaluation of biomarkers in an early drug discovery study. In Chapter 13, the definition of a biomarker, a surrogate and a true endpoint is given, and we establish the link between the evaluation of a surrogate endpoint in the clinical trial setting and the framework for selecting biomarkers in the microarray setting. We apply a joint model approach to a case study for selecting two types of biomarkers, namely, prognostic and therapeutic biomarkers. In Chapter 14, we consider the use of a two-stage model to screen for prognostic biomarkers, for which the adjusted association between gene-expression and the outcome is not necessarily linear. In Chapter 15, we discuss the construction of a joint biomarker profile based on several genes to enhance the performance of a single gene.

Chapter 16 offers concluding remarks and a perspective for future research.
Chapter 2

Introduction to Microarrays

This chapter provides a concise overview of the DNA microarray technology and data-analytic tasks associated with microarray studies. The latter include pre-processing of the data obtained by the scanning of gene-expression images; significance testing to find differentially expressed genes; and supervised and unsupervised methods for class prediction.

2.1 Microarray Technology

DNA microarrays were first introduced in the mid-1990s and have been the most widely utilized tool for measuring gene-expression since. There are several DNA microarray technologies. Currently, two approaches are prevalent: cDNA arrays and oligonucleotide arrays.

2.1.1 cDNA Arrays

cDNA microarray was first described by Schena et al. (1995) and DeRisi et al. (1997). In spotted DNA arrays, messenger RNA (mRNA) from two different biological samples is reverse-transcribed into cDNA, labeled with green or red dye (i.e., Cy3 or Cy5, respectively), and hybridized to DNA sequences which have been spotted onto a glass slide prior to the hybridization. Corresponding to the dyes and different absorption frequencies, the biological signals in the samples are referred to as channels. After hybridization, a laser scanner measures dye fluorescence of each color at a fine grid
of pixels. Higher fluorescence indicates higher amounts of hybridized cDNA, which in turn indicates a higher gene-expression in the sample. A spot typically consists of a number of pixels. Image analysis algorithms assign pixels to a spot and produce summaries of fluorescence at each spot, as well as summaries of fluorescence in the surrounding unspotted areas (background). The graphical representation of this process is provided in Figure 2.1.

For each spot on the array, a typical output consists of at least four quantities, one of each color (channel) for both the spot and the background. The use of two channels allows for measurement of relative gene-expression across two sources of cDNA, controlling for the amount of spotted DNA. One way of analyzing two channel cDNA arrays is to take the ratios of intensities at each spot. The advantage of the dual channel approach is that it prevents problems in the data that could be caused by variable concentrations of DNA material spotted per DNA sequence. Since both labeled cDNAs compete for the same spot, the relative ratio is still accurate even if the amount of spotted material varies from spot to spot.
2.1.2 Oligonucleotide Arrays

The second approach involves the use of the high-density oligonucleotide arrays. The most widely used oligonucleotide array type is the Affymetrix GeneChip (for brevity Affy), discussed by Lockhart et al. (1996). Details on Affy arrays can be found in Affymetrix Manual (1999).

In Affy arrays, expression of each gene is measured by comparing hybridization of the sample mRNA to a set of probes, composed of 11-20 pairs of oligonucleotides, each of length of 25 base pairs. The first type of probe in each pair is known as perfect match (PM) and is taken from the gene sequence. The second type is known as mismatch (MM) and is created by changing the middle (13th) base of the PM sequence to reduce the rate of specific binding of mRNA for that gene. The goal of MM's is to control for nonspecific binding of mRNA from other parts of the genome.

Figure 2.2: *Affymetrix microarray processing.*

An RNA sample is prepared, labeled with a fluorescent dye, and hybridized to an array. Unlike in two-channel arrays, a single sample is hybridized on a given array. Each array is then scanned, and image is produced and analyzed to obtain a fluorescence intensity value for each probe, measuring hybridization for the corre-
Chapter 2. Introduction to Microarrays

Figure 2.3: Construction of a probe set on an Affymetrix microarray.

sponding oligonucleotide (see Figure 2.2, Affymetrix 1999). For each gene, or probe set (throughout the dissertation, we refer to a probe set as a gene), the typical output consists of two vectors of length 11-20 intensity readings, one for PMs and one for MMs (see Figure 2.3, Affymetrix 1999).

2.2 Statistical Analysis of Microarrays

Microarray expression analysis offers an opportunity to generate functional data on a genome-wide scale and consequently, provide data for the biological interpretation of genes and their functions (Quackenbush 2001). Thus, careful handling and interpretation of microarray expression data is required.

The statistical components of a microarray experiment involve the following steps (Allison et al. 2006):

1) Design. The development of an experimental plan to maximize the quality and quantity of information obtained.

2) Pre-processing. Processing of the microarray image and normalization of the data to remove systematic variation. Other potential preprocessing steps include
transformation of data, data filtering and background subtraction.

(3) Inference and/or classification. Inference entails testing statistical hypotheses (these are usually about which genes are differentially expressed). Classification refers to analytical approaches that attempt to divide data into classes with no prior information (unsupervised classification) or into predefined classes (supervised classification).

(4) Validation of findings. The process of confirming the veracity of the inferences and conclusions drawn in the study.

2.2.1 Pre-processing

Microarrays are powerful in measuring thousands of genes simultaneously, but variability arises throughout the measurement process. The pre-processing steps include the image analysis, quality control of arrays, background subtraction, summarization of intensities for Affy probe sets, and normalization within-and-across arrays.

**Image Analysis**

 Regarding image analysis for cDNA arrays, methods of appropriate quantification of spots on microarrays is a topic of vigorous inquiry. Many image-processing approaches have been developed (Chen *et al.* 1997, Schadt *et al.* 2001, Ekstrom *et al.* 2004, Steinfath *et al.* 2001, and Yang *et al.* 2001), among which the main differences relate to how spot segmentation, distinguishing foreground from background intensities, is carried out.

For Affymetrix arrays several image-processing methods have been developed (Nielsen *et al.* 2005 and Irizarry *et al.* 2003). These methods estimate the amount of RNA from fluorescent array images, while trying to minimize the extraneous variation that occurs owing to technical artifacts.

**Quality Control of Arrays**

Many researchers recognize the need for microarray quality-control (QC) measures that quantify the measurement quality for any particular array. To this aim, several methods have been proposed (for example, using a graphical approach, Chen *et al.* 2004).

A simple quality control procedure can be established at the moment when the spotted image is stored in the database by running a procedure that produces the following items (Amaratunga and Cabrera 2004):

1. An image quality graph could be used to detect specific problems with the
array.

(2) A side-by-side display of boxplots of gene-expression measures for the sequence of arrays, or a set of summaries based on them, could be used to check whether there are any changes between the arrays.

**Background Subtraction**

For cDNA arrays, due to various reasons such as nonspecific binding of the labeled sample to the array substrate and substrate fluorescence, the pixels scanned from the microarray emit a low, but not insubstantial, level of fluorescence that may vary with location. This nonspecific fluorescence is called *background* fluorescence. It is customary to estimate the background intensity and subtract it from the raw spot intensity value to obtain the true spot intensity value.

For oligonucleotide arrays, the background substraction should be carried out as well (Affymetrix 2002). There are several ways to estimate the background intensity (Amaratunga and Cabrera 2004): global background adjustment that averages the intensity of all the pixels not belonging to the spots; spot background adjustment that yields a spot background-adjusted spot intensity values; smoothed background adjustment that applies a smoothing procedure to the values surrounding a spot; and zonal background adjustment that is a variation of the smoothed background adjustment.

**Summarization of Intensities**

For the two types of microarray described above, the spot intensities need to be summarized using specific methods.

For a cDNA microarray with two channels, the gene-expression level of each gene is obtained using an expression ratio of measurements obtained for green and red dyes.

For oligonucleotide arrays, such as those manufactured by Affymetrix, for a probe set, expression level is obtained by combining the $PM$ and $MM$ intensities of the probes. There are several ways in which a composite value can be calculated (Amaratunga and Cabrera 2004) by using: the average difference (i.e., the average of $PM - MM$, Affymetrix MicroArray Suite (MAS) version 5.0); a weighted average difference (Efron *et al.* 2001); perfect matches only (Naef *et al.* 2001); robust multi-chip average approach (RMA, Irizarry *et al.* 2003a); model-based approach (Li and Hung Wong 2001a,b).

**Normalization**

Before multiple microarray measurements can be integrated into a single analysis,
the reported measurements need to be normalized, or modified (possibly corrected), to make them comparable.

Normalization is a matter of adjusting the overall brightness of each scanned microarray image, assuming that the quantity of RNA applied to an array is equal between the arrays (Wu 2001). The methods include: using expression levels of housekeeping genes (Eickhoff 1999); using assumptions that most genes do not change across experiments (Zien et al. 2001); using splines (Li and Hung Wong 2001a); using quantile normalization (Bolstad 2002); using other nonlinear techniques (Ramdas et al. 2001, Tseng et al. 2001); using model based approach (Wu et al. 2004).

For two-channel cDNA microarrays, it is necessary to consider the systematic bias of dyes. An MVA plot (Mean difference of intensities for two dyes Versus Average intensities for two dyes, Yang et al. 2001, Heldermans et al. 2007) can be used to detect the pattern that deviates from the M=0 line. For the case of systematic dye bias, an intensity-dependent normalization can be carried out (Amaratunga and Cabrera 2004). If there is no systematic dye bias, normalization can be done using methods applicable for the oligonucleotide arrays.

2.2.2 Significance Analysis of Gene-expression Levels

Studying which genes are active and which are inactive in different cell types helps scientists to understand both how these cells function normally and how they are affected when various genes do not perform properly.

To answer these questions, significance analysis of gene-expression levels in microarray data is an important task. The task is difficult owing to the high dimensionality of the data (the number of genes). Various statistical approaches are available to test whether two microarray samples are drawn from the same population or distribution. A gene-by-gene analysis (by using either parametric or non-parametric tests) is often the option (Ge et al. 2003, Kerr et al. 2000). In this dissertation, we follow the similar approach and we focus on methods of finding differentially expressed genes with multiple testing adjustment. For example, we consider test-based approaches using a t-test statistic (Dudoit et al. 2000) and a modified t-test statistic (Tusher et al. 2001). Model-based approaches can also be applied by, for example, using an analysis of variance (ANOVA) model (Kerr et al. 2000, Kerr and Churchill 2001) or a linear mixed model (Wolfinger et al. 2001). A more detailed discussion of the multiplicity issue is given in Chapter 5.
2.2.3 Supervised and Unsupervised Classification

Many classification algorithms are used in microarray research. They are used either to discover new categories within a data set (class discovery; \textit{unsupervised} classification) or assign cases to a given category (class prediction; \textit{supervised} classification).

Supervised methods are used to find genes that accurately predict a characteristic of the sample with the information about known groups. These methods might find individual genes as in the case of the nearest neighbor approach (Golub \textit{et al.} 1999), and/or multiple genes, as in the case of decision trees (Quinlan 1992), neural networks (Rumelhart and McClelland 1986), and support vector machines (Furey \textit{et al.} 2000, Brown \textit{et al.} 2000, Chow \textit{et al.} 2001). The methods above, nearest neighbor, classification tree, and support vector machines are studied in detail in Chapter 3 and 4.

Unsupervised methods try to find internal structure or relationships in a data set, instead of trying to determine how to predict a correct answer best. Within unsupervised learning, there are three main classes of techniques: (1) feature determination, or determining genes with interesting properties without specifically looking for a particular a priori pattern, such as principal-components analysis (Alter \textit{et al.} 2000, Raychaudhuri \textit{et al.} 2000); (2) cluster determination, or determining groups of genes or samples with similar patterns of gene-expression, such as nearest neighbor clustering (Golub \textit{et al.} 1999, Ben-Dor \textit{et al.} 2000), self-organizing maps (Tamayo \textit{et al.} 1999, Toronen \textit{et al.} 1999), k-means clustering and one-and two-dimensional hierarchical clustering (Ross \textit{et al.} 2000, Eisen \textit{et al.} 1998); and (3) network determination, or determining graphs representing gene-gene or gene-phenotype interactions using Boolean networks (Liang \textit{et al.} 1998, Wuensche 1998, Szallasi and Liang 1998), Bayesian networks (Friedman \textit{et al.} 2000), and relevance networks (Butte \textit{et al.} 2000, Butte and Kohane 1999, 2000).

2.3 Main Issues in Analyzing Gene-expression Data

In microarray experiments, gene-expression levels of a large of number of genes need to be analyzed simultaneously, while usually only a few replicate samples(arrays) are used. Throughout this dissertation, the analysis of microarray experiments will be subject to the multiplicity issue. Also, we will need to address the issue of small sample size inference.
2.3.1 Multiplicity

Multiple hypotheses testing problem is one of main issues in microarray data analysis. Rejecting true null hypotheses and failing to reject false null hypotheses can both lead to wrong decisions. The error of rejecting a true null hypothesis is the type I error, while the error of failing to reject a false null hypothesis is the type II error. In this dissertation, we mainly focus on the type I error.

For testing a single null hypothesis, the type I error rate is simply the possibility of rejecting this hypothesis when it is actually true. For the case of testing multiple hypotheses, the type I error rate can have a variety of generalizations. The two most commonly used error rates in multiple testing are the Family-Wise Error Rate, abbreviated as FWER, and the False Discovery Rate, abbreviated as FDR.

The FWER is the probability of rejecting at least one true null hypothesis. The weak, exact, and strong control of FWER correspond to the situation where all the null hypotheses are true, an exact set of null hypotheses is true, and any subset of null hypotheses is true, respectively. It is most desirable to have an exact control of FWER, but it is usually impossible to achieve it because the exact set of truly null hypotheses is unknown. Weak control and strong control are possible to obtain, but the former may not be sufficient, and the latter may be conservative. Procedures controlling the FWER, such as the Bonferroni, provide a strong control of FWER, while Storey’s FDR approach (Storey 2001) controls the FWER in a weak sense. In this dissertation, we are not focusing on the ability of a procedure to control strongly the FWER.

Nevertheless, control of the FWER can be too stringent in the microarray setting. In microarray experiments the aim of the study is to find differentially expressed genes, while allowing some false positives. From this point of view, the concept of controlling the FDR becomes popular. The FDR is the expected proportion of false positives among all the rejected null hypotheses. In general, FDR-controlling procedures yield higher power than FWER-controlling procedures, but more type I errors as well (Finner and Roter 2001).

In Chapter 5 we address the multiplicity issue and discuss various procedures for controlling the FWER and the FDR.
2.3.2 Resampling-based Inference

In the microarray setting, often a small number of samples (arrays) is used. Many inference methods that are used for microarray analysis rely on specific assumptions about the distribution of the gene-expression measurements, from which they derive properties of theoretical distributions, on which the inference can be based. By contrast, resampling-based inference (RBI) methods rely on resampling data. Compared with asymptotic tests, RBI has the advantage of being robust and flexible enough to accommodate almost any new statistic, without the need to derive statistic’s distribution. RBI is computationally intensive, but with modern computational tools it is now feasible in many cases and therefore it is widely used (Neuhauser et al. 2005, Barry et al. 2005, Pan 2003, Xu and Li 2003, Landgrebe et al. 2002). Another disadvantage of RBI is that RBI $p$-value distribution can be coarse or granular, and it will often be difficult to obtain significant tests. Thus, the question how to obtain the benefit of combining RBI statistics across genes is an important research topic. In Chapter 5 we give details of resampling-based methods that offer the control of the FWER and the FDR.
Part I

Classification Methods for Microarrays
Chapter 3

Classification Methods for Microarrays: a Simulation Study

3.1 Introduction

As we discussed in Chapter 1, one application area of microarrays is diagnostic tests. Microarrays are used to find gene-expression profiles, which can discriminate between different known cell types or conditions. An important question is what method is most suitable for this purpose? Recently, several studies were conducted aiming at answering this question.

Dudoit et al. (2002) compared the performance of various classification methods for classifying tumors based on gene-expression profiles. They investigated nine methods using three well-known, real-life datasets related to cancer. They found that simple classifiers such as \( k \) nearest neighbor (kNN) and diagonal linear discriminant analysis (DLDA) performed remarkably well as compared to more sophisticated methods like aggregated classification trees.

Lee et al. (2005) conducted a more extensive comparison study of the performance of various classification methods in microarray experiments. In particular, they considered 21 methods and applied them to seven datasets (related mainly to cancer), using three gene-selection techniques. Contrary to the findings of Dudoit et al. (2002),
Chapter 3. Classification Methods for Microarrays: a Simulation Study

Lee et al. (2005) concluded that the more sophisticated classifiers gave better performances than classical methods such as kNN, DLDA, or diagonal quadratic discriminant analysis (DQDA). Additionally, they found that the choice of gene-selection method had much effect on the performance of the classification methods.

Statnikov et al. (2005) conducted a comparison study focusing mainly on the use of support vector machines (SVMs) in microarray experiments. Additionally, they included kNN, back propagation- and probabilistic neural networks. They considered the use of various techniques for ensembling of classifiers to improve performance of individual classifiers, and applied four gene-selection techniques. In the study, 11 cancer-related real-life datasets were used. The authors found that multicategory SVMs were the most effective classifiers. Similarly to Lee et al. (2005), they concluded that using a gene-selection method can significantly improve the performance of non-SVM classifiers and moderately for SVM classifiers. On the other hand, ensemble classifiers did not seem to improve the performance of non-ensemble techniques.

In all the aforementioned studies real-life datasets were used. The main advantage of using such datasets is that they adequately capture the complexity of microarray data. However, even with the use of several datasets, there is only a limited number of settings that can be evaluated. Moreover, studying the influence of the sample size, of the correlation structure between genes, or of the size of the treatment effect on the performance of a classification method is difficult, as the required real-life data are usually not available. Additionally, the set of truly differentially expressed genes is unknown. Although this does not preclude the evaluation of the relative merits of different methods, it complicates the evaluation of their overall performance.

In order to overcome these limitations, we decided to conduct a simulation study with a linear mixed effects simulation model that mimics a microarray context. The parameters of the model were chosen based on an analysis of a real-life experiment.

The use of simulated data allows us to investigate the performance of several classification methods under various, controlled scenarios. As the true configuration of parameters used in the simulation model is known, the overall performance of the methods can be clearly interpreted.

The chapter is organized as follows. In Section 3.2 we shortly describe the investigated class-prediction methods. In Section 3.3, we conduct a simulation study, using a linear mixed effects model to simulate microarray data under different scenarios. We compare the methods discussed in Chapter 3.2 with respect to their ability to discriminate between two classes of biological samples in various experimental set-
tings. Section 3.3 contains details of our simulation study. Results are described in section 3.4. A short discussion and conclusions are presented in sections 3.5 and 3.6, respectively.

3.2 Class-prediction Methods

We focus on the simplest case of discrimination between two different groups of samples. For the purpose of comparison with the results obtained by Dudoit et al. (2002) and Lee et al. (2005), the class-prediction procedures investigated in our study included tree methods, classical discrimination analysis techniques, and a machine learning method. In this section we give a brief description of each of them. For the implementation of the different methods existing R functions were used. In what follows the necessary packages, as well as the parameter settings of the particular functions, are indicated.

Classification tree

A classification tree is a binary recursive partitioning method developed by Breiman et al. (1984). At each step a subset of training samples is split in two, based on the value of one particular gene. The value is chosen to obtain in each partition an as homogeneous set of labels as possible. The subsets remaining at the final stage are assigned to a certain class, the one which is most frequently represented in the subset. In a way, the method has its own gene-selection procedure. It determines which genes to use at each splitting node in order to get the best classification. This feature makes it quite robust to the presence of random noise. The method is implemented by the function `dorpart()`, which is a part of the `IBCLab4` package (Jean Yang 2004).

Bagging

Aggregated trees combine tree classifiers to improve the accuracy of the class prediction. One such method is called bagging (Breiman 1996). Bootstrap replicates (in our case 100) are taken from the training dataset. A tree is constructed for each replicate and the final classification is determined by majority vote. That is, the sample is assumed to belong to the class, to which it is most frequently assigned by the different trees. Bagging is said to be a variance reduction technique, designed to stabilize trees. It is implemented by the function `ipredbagg.factor()` from the `ipred` package (Peters and Hothorn 2007).

Boosting

Boosting, proposed by Schapire and Freund (1999), is another form of aggregating
trees. A series of classification trees is produced for the training dataset, each time with different weights assigned to the samples. The idea is to give samples misclassified in the previous step more weight in the current one. The final outcome is a weighted majority vote of all created trees. It is believed that bagging is much better than boosting in situations with substantial random noise. Boosting is however expected to reduce both the variance and bias of unstable trees. The method is implemented by the functions \texttt{gbm()} and \texttt{gbm.more()} in the \texttt{gbm()} package (Ridgeway 2007). We use a series of 100 trees in the manner described above, and choose for Bernoulli distribution, a shrinkage parameter of 0.001 and 0.5 as the fraction of randomly selected observations for building a tree. The \texttt{gbm()} function is first applied to the data in order to create 100 trees in the manner described above. These are the default values of the \texttt{R} functions. The \texttt{gbm.more()} function is used to create 1000 additional trees.

**Random forest (RF)**

Random forests (Breiman 2001) are formed by a combination of tree predictors. Subsets of samples and genes are obtained by independently drawing samples with replacement from the training dataset and by selecting a number of genes at random. A classification tree is estimated for each of the newly formed datasets. A new sample is allocated to the class with the most votes over all the trees in the forest. The method is implemented by the function \texttt{randomForest()} from the \texttt{randomForest} package (Liaw 2008). The number of samples drawn at random is set at 63\% of the total number, while for the genes the square root of the total number is taken. Both are the default values of the \texttt{R} function. The method was applied with the number of trees equal to 500 and 1000.

**k nearest neighbors (kNN)**

kNN (Ripley 1996) is an intuitive method that classifies unlabeled samples based on their similarity by \( k \) closest samples (using Euclidean distance measure; other similarity measures are also possible, but not discussed in this dissertation) in the training set. The class, to which the sample is assigned, is the one that appears most frequently among the \( k \) neighbors, where \( k \) is chosen based on leave-one-out cross-validation performed on the training dataset. In case of ties, i.e., an equal representation of the two classes among the neighbors, a random assignment is made. The method is implemented by the function \texttt{knn()} from the package \texttt{class} (\texttt{R} standard library). \texttt{knn.cv()} is used to determine the value of \( k \).

**Discriminant analysis**
Linear discriminant analysis (LDA), a classical discriminant method, estimates a linear discriminant function, which is used as a decision boundary. The method is based on the assumption of normally distributed data and equal covariance matrices for the considered classes. Diagonal linear discriminant analysis (DLDA) is a variant of LDA, whereby the covariance matrix is additionally assumed to have a diagonal structure. Diagonal quadratic discriminant analysis (DQDA), also a variant of LDA, assumes diagonal, but not equal covariance matrices for all classes. In a sense, both DLDA and DQDA ignore the correlation structure between expression levels of genes in the microarray data. LDA is implemented by the function `lda()` in the MASS package (R standard library), while `stat.diag.da()` from the sma package (Dudoit, Yang, and Bolstad 2005) is used for DLDA and DQDA. For microarray data, the number of variables (genes) is much larger than the number of observations (arrays), which leads to the singularity of the covariance matrix. Thus, the function `lda()` sets an upper limit on the number of genes as input due to the rank issue. In our setting, the maximum number of genes used is equal to 60 and 100 in Chapter 3 and 4, respectively (that is the sample size in the training set in each chapter).

Support vector machines (SVMs)

Support vector machines, first introduced by Vapnik (2000) in the machine learning theory, are used to solve classification problems. The idea behind the method is the following: the samples are non-linearly mapped to a very high-dimensional feature space. In this space a hyperplane is designed that provides an optimal separation between the two groups (Vapnik 2000). SVMs are characterized by the use of linear, polynomial, spline, and other kernels to solve the optimization problem. In our case the SVM based on a linear kernel is used. The method is implemented by the `svm()` function from the package e1071 (Dimitriadou et al. 2007). The other parameters are set at the default values of the R-function.

3.3 Simulation Study

We assume the setting of an experiment with two-channel cDNA microarrays used in a common reference design. In the experiment, two classes of samples (a treatment and a control group) are compared to the same reference group. No difference between gene-expression levels for the control and reference groups is assumed. On the other hand, a subset of genes is assumed to be differentially expressed in the treatment group as compared to the reference group. Treatment and control samples are labeled with
Chapter 3. Classification Methods for Microarrays: a Simulation Study

the same dye, while the reference samples are labeled with the other dye. The log ratios of the two channels (Control/Reference and Treatment/Reference) are used for classification purposes.

For every setting of interest 100 simulation datasets are created, each containing 100 arrays: 60 are used as a training dataset and the other 40 form a test dataset. One half of the arrays in each set always contain the treatment sample, while the other half is for the control group. On each array we simulate 2000 genes, what corresponds to the number of genes on arrays used in the real-life experiment that we are using as a basis for the simulation study (Van Breda et al. 2005). A linear mixed effects model is utilized to simulate observations subject to various systematic and random effects usually present in real microarray experiments (Kerr et al. 2000, Wolfinger et al. 2001). An observation $Y_{ijk}$, assumed to be the signal intensity for gene $i$ ($i = 1, \ldots, 2000$), array $j$ ($j = 1, \ldots, 100$), and dye $k$ ($k = 1, 2$), is generated by the following model:

$$\log_2(Y_{ijk}) = \mu + A_j + G_i + AG_{ij} + DG_{ik} + TG_i + \varepsilon_{ijk}, \quad (3.1)$$

where $\mu$ is the overall mean, $A_j$ stands for the overall array effect, $G_i$ for the gene effect and $AG_{ij}$ for their interaction. $DG_{ik}$ and $TG_i$ represent, respectively, the gene-specific dye and treatment effects, while $\varepsilon_{ijk}$ is a random error. Some of these effects are fixed, while others are random and drawn from a normal distribution. The details are displayed in Table 3.1. The chosen values are based on the estimated parameters of a ANOVA-model fitted to a real-life cDNA microarray dataset (Van Breda et al. 2005).

Figure 3.1 shows histograms of the log intensity values for two arrays chosen from the real-life dataset. As can be seen from these graphs, for one of them the normal distribution may be a reasonable approximation. For the other a slightly asymmetric distribution with a long tail on the right side might be more appropriate. For the purpose of this simulation study, we restricted ourselves to the use of the normal distribution. In future research we also include asymmetric distributions.

Figure 3.2 presents a normal QQ-plot of the log transformed gene-specific variances of the error terms obtained from the ANOVA-model fitted to the real-life dataset. In this particular case a log-normal distribution seems to be a very good approximation to the distribution of the gene-specific variance of $\varepsilon_{ijk}$. Consequently, it has been used in the simulation model.

The gene-specific dye effect is represented by the difference in the mean of the
Figure 3.1: Histograms with a normal density curve of the log transformed signal measurements for two arrays of the real-life cDNA microarray dataset.

Figure 3.2: Normal QQ-plot of the log transformed gene-specific variances of the error terms obtained from the ANOVA-model fitted to the real-life cDNA microarray dataset.

Table 3.1: Simulation model parameters.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value/Distribution</th>
<th>Parameter</th>
<th>Value/Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu$</td>
<td>$= 9$</td>
<td>$TG_i$</td>
<td>$= 0.5 \times I(i \leq 20, j = \text{odd}, k = 1)$</td>
</tr>
<tr>
<td>$A_j$</td>
<td>$\sim N(0.5, 0.1)$</td>
<td>$DG_{ik}$</td>
<td>$\sim N{1 \times I(k = 1), 0.2}$</td>
</tr>
<tr>
<td>$G_i$</td>
<td>$\sim N(0, 5)$</td>
<td>$\varepsilon_{ijk}$</td>
<td>$\sim N(0, \sigma_i^2)$</td>
</tr>
<tr>
<td>$AG_{ij}$</td>
<td>$\sim N(0, 0.5)$</td>
<td>$\log(\sigma_i^2)$</td>
<td>$\sim N(-2, 0.5)$</td>
</tr>
</tbody>
</table>
normal distribution for $DG_{ik}$. Arrays with even numbers contain control and reference group samples. Arrays with odd numbers contain treatment and reference samples. The latter arrays contain 20 genes that are differentially expressed. We consider two scenarios with respect to the treatment effect. In the first one, for all 20 genes a constant treatment effect of 0.5 on the log scale is assumed. In the second one, we assume 4 groups of 5 genes with treatment effects equal to, respectively, 0.125, 0.250, 0.375 and 0.5 on the log scale.

Heterogeneity is incorporated into model (3.1) by allowing the random error term $\varepsilon_{ijk}$ to have a gene-specific variance. The variances were selected randomly from a log-normal distribution and fixed for all simulated datasets (see Table 3.1).

We consider the case when the random error terms $\varepsilon_{ijk}$ are independent, as well as the case when, for a subset of genes, they are correlated. The error term was drawn from a multivariate normal distribution rather than from a univariate. The covariance matrix of this distribution was adapted in such a way that a predefined correlation was induced between pairs of a number of preselected genes. Three situations were considered: correlation of 0.5 or 0.9 between 10 differentially and 10 non-differentially expressed genes; correlation of 0.9 between 10 non-differentially expressed genes; and correlation of 0.5 between five differentially and five non-differentially expressed genes together with correlation of 0.9 between five differentially and five non-differentially expressed genes.

A typical microarray experiment involves many genes, of which usually only a relatively small subset are differentially expressed. Therefore, one may want to select and base the classification on a subset of, say $p$, genes, majority of which would hopefully differentiate between the compared classes of samples. In this study we employ the same selection method as the one used by Dudoit et al. (2002). More methods for gene selection are considered in Chapter 4. First, for each gene the BW ratio is computed, which is the ratio of the between-treatment sum of squares and the within-treatment sum of squares of the log ratios of the two channels. Second, the $p$ genes with the largest BW ratio are selected and used for classification. In order to study the influence of the choice of $p$ on the performance of various classification methods, we use a range of values: $p = 2, 5, 10, 20, 40, 2000$.

Finally, in reality the microarray experiments may employ only a limited number of arrays for building the prediction rule. For this reason we also considered a scenario, in which the training set contains only 20 microarrays.
3.4 Results

We compare the performance of the various class-prediction methods in terms of the misclassification rate. The rate is calculated based on the number of misclassified arrays divided by total number of arrays (40) in the test dataset. We focus mainly on the results obtained for the case of uncorrelated gene-expression levels. Introduction of correlation between various subsets of differentially and non-differentially expressed genes does not substantially change the results.

3.4.1 Large Training Dataset

In this section we discuss the results obtained when the compared methods were applied to simulated training datasets containing 60 arrays.

**Constant Treatment Effect**

Figure 3.3 presents boxplots for misclassification rates as a function of $p$ for each classification method.

For bagging and boosting results for $p = 2000$ are unavailable due to numerical complexity. One can observe that for all methods, except for the classification tree, there is a clear dependence of the average misclassification rate on $p$. Generally, the smallest rate is observed for $p = 10$. Using too few (2 or 5) or too many (40 or all) genes increases the misclassification. An explanation is that with too few genes it is difficult to discriminate between the classes, even if the genes are truly differentially expressed. On the other hand, as $p$ increases, more and more non-differentially expressed genes enter the subset selected for building the classification rule and make the classification more difficult. A confirmation of the latter pattern can be found in Table 3.2. The table shows the mean and median numbers of the truly differentially expressed genes among the $p$ genes with the largest BW ratios. Generally, for $p = 2, 5,$ and 10, and constant treatment effect, all the selected genes are truly differentially expressed. As $p$ increases, however, the proportion of non-differentially expressed genes among those selected also increases.

Figure 3.4 shows the comparison of the misclassification rates for the different classification methods for $p = 10$ and $p = 2000$. One can observe that with a careful choice of $p$, the methods like RF, SVM, DLDA, and DQDA can achieve lower misclassification rates than the other procedures. Moreover, the rates are relatively stable if
the majority of the selected genes are truly differentially expressed. When all genes are used, the error rates increase (dramatically in the case of kNN).

For each choice of top $p$ genes, the single classification tree is very robust to the number of genes used (Figure 3.3), but the performance is poor. This may be attributable to the fact that the method contains its own mechanism of selecting variables in the tree. However, on average, the rates are higher than the rates that can be obtained with a careful choice of $p$ for other methods (Figure 3.4, left panel). Aggregated trees using bagging and boosting achieve much lower misclassification rates than the single classification tree. Their performance becomes very stable as the number of selected genes increases. Random forests show better performance over the other tree-based methods.
Table 3.2: Mean and median number of truly differentially expressed genes among \( p \) genes with the largest value of the BW criterion.

<table>
<thead>
<tr>
<th></th>
<th>Large Training Dataset</th>
<th>Small Training Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p )</td>
<td>2 5 10 20 40</td>
<td>2 5 10 20 40</td>
</tr>
<tr>
<td>Treat. effect constant (uncorrelated genes)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean</td>
<td>2 4.96 9.38 13.39 15.48</td>
<td>1.46 2.8 4.22 5.87 7.84</td>
</tr>
<tr>
<td>median</td>
<td>2 5 10 13 16</td>
<td>2 3 4 6 8</td>
</tr>
<tr>
<td>Treat. effect prop. decreased (uncorrelated genes)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean</td>
<td>1.95 4.29 6.16 7.78 9.21</td>
<td>0.69 1.37 2.19 2.96 4.01</td>
</tr>
<tr>
<td>median</td>
<td>2 4 6 8 9</td>
<td>1 1 2 3 4</td>
</tr>
<tr>
<td>Treat. effect constant (correlated genes)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean</td>
<td>2 5 9.42 13.7 15.83</td>
<td>1.38 2.71 4.13 5.85 7.82</td>
</tr>
<tr>
<td>median</td>
<td>2 5 10 14 16</td>
<td>2 3 4 6 8</td>
</tr>
</tbody>
</table>

Figure 3.4: Comparison of various classification methods for \( p = 10 \) and \( p = 2000 \).  
TB1: tree (bagging); TB2: tree (boosting); D1: DLDA; D2: DQDA.
Chapter 3. Classification Methods for Microarrays: a Simulation Study

Proportionally Decreasing Treatment Effect

Similar results regarding method comparison are obtained when the treatment effect is proportionally decreased instead of being kept constant (Figure 3.5). RF, SVM, DLDA, and DQDA still achieve lower misclassification rates than the other procedures. However, all misclassification rates increase. From Table 3.2 one can see that the mean and median number of the truly differentially expressed genes among the \( p \) genes selected based on the BW ratio has decreased compared to the previous setting. As genes with the lower treatment effect are introduced, it is more difficult for the BW ratios to select the truly differentially expressed genes. The same pattern of dependence of the average misclassification rate on \( p \) can still be observed. The lowest misclassification rate is achieved for \( p \) equal to 5 or 10, as for these values the majority of the selected genes are truly differentiated.

![Boxplots for misclassification rates](image)

Figure 3.5: Boxplots for misclassification rates (vertical axis) for various classification methods and for different values of \( p \) (horizontal axis). The results for Random Forest are based on 500 trees, except for the last boxplot (all2), where they are based on 1000 trees.
Correlated Genes

Figure 3.6 shows the performance of the classification procedures applied to datasets, in which 10 non- and 10 differentially expressed genes are correlated with the correlation coefficient of 0.9. The treatment effect for this setting is kept constant at 0.5. For other settings assuming correlation between various subsets of differentially and non-differentially expressed genes similar results were observed (data not shown).

The number of genes selected (Table 3.2) and the performance of the classification procedures are quite similar to those observed for the corresponding setting without correlated genes. LDA and SVM do seem to perform a bit better for the optimal value of $p$. They are however still very sensitive to the inclusion of non-differentially expressed genes in the classification procedure.

![Figure 3.6: Boxplots for misclassification rates (vertical axis) for various classification methods and for different values of $p$ (horizontal axis) for the setting with correlated gene-expression levels. The results for Random Forest are based on 500 trees, except for the last boxplot (all2), where they are based on 1000 trees.](image)

3.4.2 Small Training Dataset

When the training dataset contains only 20 arrays, the misclassification rates for all methods substantially increase and do not fall below 20% for any $p$ (Figure 3.7). This
can be attributed to the fact that in this case, the BW criterion has a low power to select truly differentially expressed genes, as can be seen from Table 3.2. Aside from that, similar conclusions regarding the performance of various methods can be drawn as in the case of a large training dataset. For instance, with a suitable choice of $p$, RF and DLDA perform in general better than the other methods (see Figure 3.7).

![Figure 3.7: Boxplots for misclassification rates (vertical axis) for various classification methods and for different values of $p$ (horizontal axis) for the setting with 20 arrays in the training dataset. The results for Random Forest are based on 500 trees, except for the last boxplot (all2), which is based on 1000 trees.](image)

### 3.5 Discussion

Given the complexity of microarray data, one probably should not expect that a single classification method will always outperform all the others. It is therefore paramount to investigate relative merits of different procedures to see in which settings which methods might be expected to work reasonably well.

Simulations allow to overcome limitations related to the use of real-life data. For instance, we were able to investigate the performance of the classification methods in many controlled settings. On the other hand, there are some limitations of a simulation study. An important issue is whether the simulated data adequately capture
the complexity of microarray measurements. The real-life data used as a basis for our simulations could be approximated by data simulated by a linear mixed model. In general, this may not be the case. To deal with this issue, one might consider using in the simulation study, e.g., a non-normal distribution. This is the topic for the future research.

Another issue is related to the computational complexity of such a study. Simulating the data and performing all the classification procedures on them is time consuming. Extending the simulations by simulating data for more genes, more datasets or more settings is therefore not trivial. For this reason, in our study we chose for a limited, yet relevant from a practical point of view (as documented by Van Breda et al. 2005), setting.

The results of our simulation study indicate that the performance of classification methods for microarray experiments is optimal if the majority of genes used for building of a classification rule are truly differentially expressed genes. Thus, gene selection is a very important step. This finding is consistent with the results reported by Lee et al. (2005) and Statnikov et al. (2005). Clearly, further research is needed on the choice and properties of various gene-selection procedures. In next chapter, we will use simulation study to investigate the performance of some gene-selection procedures combined with the classification methods studied in this chapter.

3.6 Conclusions

Our simulations suggest that random forests and DLDA are the methods, which perform best when no correlation between gene-expression levels is present. For the optimal choice of $p$ they achieve low misclassification rates, with DLDA performing slightly better. On the other hand, when all genes are used, the increase of the rates is smallest for these two methods, with RF showing a lower rate than DLDA. To a large extent this is also true for the case when the expression levels for a subset of genes are correlated. The simulation study presented in this chapter was based on block-wise correlation of 20 genes. The influence of the number of genes which are correlated on the misclassification error should be investigated further.

This conclusion overlaps with the results reported by Dudoit et al. (2002) and Lee et al. (2005). Good performance of DLDA, as compared with other methods, was pointed out by Dudoit et al. Lee et al. found that DLDA performs well compared to more sophisticated methods in a homogeneous dataset, and it usually performs better
than other classical classification procedures like LDA. Both Dudoit et al. and Lee et al. reported that aggregating improves the performance of trees; in the study of Lee et al. random forests gave the best results among the tree methods when number of classes was moderate.

Our study also clearly indicates that when the number of arrays available for building of a classification rule is small, the performance of classification procedures substantially deteriorates.

The results of this simulation study were published in Van Sanden et al. (2007).
Chapter 4

Performance of Gene-selection and Classification Methods in a Microarray Setting: a Simulation Study

4.1 Introduction

In the previous chapter we compared the performance of several classification methods using simulated microarray data (Van Sanden et al. 2007). The study supplements prior investigations that were based on the use of real-life data (Dudoit et al. 2002, Lee et al. 2005, and Statnikov et al. 2005). The latter captured the complexity of microarray measurements more adequately, but simulated data are not limited to only those settings, for which data are available. Furthermore, in simulations the true classification, as well as the set of truly differentially expressed genes, is known. Thus the misclassification rate can be accurately determined.

In this chapter we extend the study of classification methods in Chapter 3 in two directions. First of all, the study was limited to the use of one gene-selection
method prior to performing classification. Lee et al. (2005) found that the choice of the gene-selection method had much effect on the performance of the classification procedures. They considered four gene-selection techniques combined with a number of classification methods, and applied them to seven real-life datasets. In this chapter we incorporated 14 well-known selection methods into the simulation study to investigate their influence on 11 classification procedures under several controlled scenarios. There are still other gene selection (linear models (Smyth 2004), the TnoM score (Ben-Dor et al. 2000), random-forest-based approach (e.g., Diaz-Uriarte and Alvarez de Andres 2006)) and classification methods (LogitBoost (Dettling and Bolmann 2003), BagBoosting (Dettling 2004), Regular Discriminant Analysis (e.g., Guo et al. 2007)) that have been recently proposed. However, to keep the study feasible, we decided not include them in the simulations. We also did not consider dimension reduction methods (like MAVE, Antoniadis et al. 2003 or PLS, Nguyen and Rocke 2002, Boulesteix and Strimmer 2007). Secondly, in the Van Sanden et al. (2007) study, the normal distribution has been used as the underlying distribution in the simulation procedure. It has been shown (Purdom and Holmes 2005), however, that microarrays do not always follow the normal distribution. Thus, we considered the symmetric Laplace distribution as a long-tailed alternative. Moreover, we also used the asymmetric Laplace distribution.

To simulate the data, a linear mixed effects simulation model that mimics a microarray context was used. The parameters of the model were chosen based on an analysis of a real-life experiment, in the same way as in Chapter 3.

For the analysis of this experiment, as well as for the simulation of the data, SAS 9.1 was used. The simulated data were analyzed using R.

The chapter is organized as follows. In Section 4.2 we shortly describe the investigated gene-selection and class-prediction methods. Section 4.3 contains details of our simulation study. Results are described in Section 4.4. A short discussion and conclusions are presented in Sections 4.5 and 4.6, respectively.

4.2 Methods

In this section we give a brief description the gene-selection and classification methods considered in the study. In most cases existing R functions were used for the implementation of the different methods. In what follows the necessary packages, as well as the parameter settings of the particular functions, are indicated.
4.2.1 Gene Selection

A typical microarray experiment involves many genes of which usually only a few are differentially expressed. Therefore, one may want to select and base the classification on a subset of, say \( p \), genes, majority of which would hopefully differentiate between the compared classes of samples. An important issue is how to select the genes? In the simulation study we assessed the use of several methods for this purpose.

**Classical Test Statistics**

We considered a basic parametric test statistic, the t-test (Ttest), and a basic non-parametric test statistic, the Wilcoxon Rank Sum test (Wilc). R-functions `stat.t2()` from the `sma` package (Dudoit, Yang, and Bolstad 2005) and `wilcox.test()` from `stats` package (R Core Team) were used for this purpose.

**Significance Analysis of Microarrays (SAM)**

SAM proposed by Tusher et al. (2001) is a method for analysing microarray experiments and detecting significant genes. A score (modified t-statistic) is assigned to each gene based on change in gene-expression relative to the standard deviation augmented by a small positive constant. This constant is used to protect against inflated t-statistics due to small standard errors. Its value is chosen to minimize the coefficient of variation of the test statistic. We discuss the SAM procedure in detail in Chapter 5. The modified t-statistic for the case of two unpaired classes was calculated by the `samr()` function from the `samr` package (Tibshirani).

**Prediction Analysis for Microarrays (PAM)**

PAM fits a nearest shrunken centroid classifier to microarray data. The method, also referred to as soft-thresholding, was introduced by Tibshirani et al. (2002). It provides a list of significant genes, of which expression best characterizes each class. The functions `pamr.train()` and `pamr.listgenes()` from the `pamr` package (Hastie et al.) implement the method.

**Extreme-value Distribution Based Gene Selection (Extval)**

Li et al. (2004) introduced a gene-selection method based on a comparison of the maximum likelihood of a logistic regression model applied to the original data and to permutation datasets. To avoid using computationally intensive procedures they proposed to take advantage of the extreme-value distribution for the log-likelihood ratios. From the use of the distribution a ranking of the genes follows, which can be used to select a predefined number of genes. Li et al. (2004) also suggest two criteria to determine the number of genes to be selected from the list, one based on
the expected values (E-criterion) and one based on p-values (P-criterion). They were both included in our simulation study. The gene-selection method and both criteria were implemented in R (Van Sanden 2006).

**Other Test Statistics**

There are a number of other statistics used for ranking and selection of genes. We considered a couple of well-known methods. Each time the $p$ genes with the largest value of the statistic were selected and used for classification. The BW ratio is the ratio of the between-treatment sum of squares and the within-treatment sum of squares (Dudoit et al. 2002). In a two group setting the BW-ratio reduces to the same statistic as the t-test. The prediction strength (PS) (Xiong et al. 2001) of a certain gene is defined as the ratio of the difference in mean log expression level between the two groups and the sum of the variances of the two classes. The between-class scatter score (BC-score) belongs to the class of correlation scores (Chai and Domenicioni 2004). Most of these scores are designed to handle multi-class problems and reduce to one of the statistics mentioned earlier when applied to a two class problem. The BC-score is obtained by dividing the weighted squared differences of the class means from the overall mean by the sum of the variances of the two classes. All the test statistics and scores were implemented in R (Van Sanden 2006).

**Statistical Impurity Measures**

In contrast to determining a test statistic, we can attempt to find a gene-specific threshold in the expression range. If a measured value for a particular gene is larger (respectively, smaller) than this threshold, the sample is assigned to, for instance, class one (respectively, two). Statistical impurity measures quantify the effectiveness of this method. There are several ways this can be done, leading to multiple impurity measures: twoing rule (Twoingr), information gain (Infgain), Gini index (Gini), max minority (Maxmin), sum minority (Summin), and sum of variances (Sumvar). A full description of them can be found in Murthy et al. (1994) and Su et al. (2003). The methods were implemented in R (Van Sanden 2006).

**Empirical Cumulative Distribution Function (ECDF)**

The method was introduced as a pre-screening technique by Boulesteix et al. (2003). Genes are selected based on how well the ECDF of the two classes is separated. One chooses values $\alpha$ and $\beta$ and retains genes, for which there exists a point where the ECDF is less than $\alpha$ for one class and more than $\beta$ for the other, or more than $1 - \alpha$ for one class and less than $1 - \beta$ for the other. Initially, $\alpha$ is set at 0.1 and $\beta$ at 0.5. The value for $\alpha$ (respectively $\beta$) is increased (respectively decreased) in order to
select the desired number of genes. After the proper values of \( \alpha \) and \( \beta \) are found and the genes are selected, for each of those genes we determine the interval of maximum width satisfying the above mentioned conditions. The width of the intervals is then be used to order the genes. The method was implemented in \( R \) (Van Sanden 2006).

### 4.2.2 Class Prediction

We focused on the simplest case of discrimination between two different groups of samples. The class-prediction procedures investigated in our study included tree methods, classical discrimination analysis techniques, and machine learning methods. The description of the methods was given in Chapter 3.

### 4.3 Simulation Study

#### 4.3.1 Simulation Setting

In Chapter 3, we only considered the normal distribution for the error distribution of the microarray data. To extend our study to a more general situation where the microarray data are not normally distributed, we also consider the use of the Laplace distribution for the error distribution.

**Normally Distributed Data**

We generated 100 normally distributed datasets as described in Section 3.3 of Chapter 3. In contrast to the setting of Section 3.3, each dataset contained 160 arrays: 60 arrays used as the training set and 100 arrays used as the testing set.

**The Laplace Distribution**

For the real-life dataset, which is the basis of the simulation model, the normality assumption for the error terms \( \varepsilon_{ijk} \) in model (3.1) holds reasonably well. Sometimes, however, the distribution of microarray data has longer tails than the normal or is somewhat asymmetric. One of the plots in Figure 3.1 demonstrates these features to some extent. We wish to examine the effect of such a deviation on the performance of the selection and classification methods. Purdom and Holmes (2005) propose the use of the Laplace distribution for microarray data. It is more peaked compared to the normal distribution. We included in the study data simulated from the symmetric and asymmetric Laplace distribution. This was achieved by replacing the normal distribution, as described in Table 3.1, by \( L(0, \sigma_g) \) for the symmetric or \( AL(0, \kappa, \sigma_g) \)
for the asymmetric Laplace distribution. The density function of $\text{AL}(\theta, \kappa, \sigma)$ is given by

$$f(x) = \frac{\sqrt{2}}{\sigma} \frac{\kappa}{1 + \kappa^2} \begin{cases} \exp\left(\frac{-\sqrt{2} \kappa}{\sigma} |x - \theta|\right) & x \geq \theta \\ \exp\left(\frac{-\sqrt{2} \kappa}{\sigma} |x - \theta|\right) & x < \theta \end{cases}$$

while that of the symmetric Laplace distribution is obtained when $\kappa$ is put equal to 1.

The gene specific scale parameter ($\sigma_g$) is the same as for the case with the normal distribution (see Table 3.1). For the skewness parameter $\kappa$ we chose values of 0.5 (skewed to the left) and 1.2 (skewed to the right). They are similar or a bit more extreme than estimates found by Purdom and Holmes (2005) for several microarrays from published microarray experiments. Their estimates ranged from 0.792 to 1.174. To make sure we can study the effect of skewness on the selection and classification methods we made the asymmetry to the left a little bit more severe.

4.4 Results

In this section we present the results of the various selection and classification methods in different settings, starting with data simulated from the normal distribution and with a constant treatment effect. Further on, we investigate changes to the results when considering other distributions or varying treatment effects.

4.4.1 Normally Distributed Data With a Constant Treatment Effect

First, a comparison of the performance of the gene-selection method is made. It can be evaluated in two ways: by the number of genes selected that are actually differentially expressed or by the misclassification rate when combining the selection method with different classification procedures. This rate is calculated based on the number of misclassified arrays divided by total number of arrays (100) in the test dataset.

The gene-selection methods were applied to the simulated training datasets to obtain subsets of genes of varying sizes ($p=2$, 5, 10, 20, 40, 200). Table 4.1 shows the median numbers of the truly differentially expressed genes among the $p=10$ selected genes. Though many methods result in similar numbers, some patterns can be observed. Certain impurity measures, especially Maxmin, do not perform as well
as some of the other methods. ECDF seems by far to be the least desirable method. Furthermore, there is no obvious winner, though SAM appears to perform slightly better than all other methods.

Table 4.1: Median number of truly differentially expressed genes among a subset of size $p=10$ selected by different methods. ($C=constant$ and $P=prop. decreased treatment effect$)

<table>
<thead>
<tr>
<th>$\epsilon_{ijk}$ \sim</th>
<th>$N(0, \sigma^2)$</th>
<th>$N(0, \sigma^2)$</th>
<th>$L(0, \sigma)$</th>
<th>$AL(0, 0.5, \sigma)$</th>
<th>$AL(0, 1.2, \sigma)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ttest</td>
<td>9</td>
<td>7</td>
<td>10</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>Wilc</td>
<td>9</td>
<td>7</td>
<td>10</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>SAM</td>
<td>9</td>
<td>7</td>
<td>10</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>PAM</td>
<td>9</td>
<td>6</td>
<td>10</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>PS</td>
<td>9</td>
<td>7</td>
<td>10</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>BC-score</td>
<td>9</td>
<td>6</td>
<td>10</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>ECDF</td>
<td>4</td>
<td>2</td>
<td>7</td>
<td>3</td>
<td>5.5</td>
</tr>
<tr>
<td>Infgain</td>
<td>9</td>
<td>6</td>
<td>10</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>Twoingr</td>
<td>9</td>
<td>6</td>
<td>10</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>Summin</td>
<td>8</td>
<td>5</td>
<td>10</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>Maxmin</td>
<td>7</td>
<td>4</td>
<td>10</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td>Gini</td>
<td>9</td>
<td>6</td>
<td>10</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>Sumvar</td>
<td>9</td>
<td>6</td>
<td>10</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>Extval</td>
<td>9</td>
<td>7</td>
<td>10</td>
<td>9</td>
<td>10</td>
</tr>
</tbody>
</table>

In the next step we evaluate the gene-selection methods when combined with different classification procedures. For demonstration purposes the results for $p=10$ and 200 (Figure 4.1) were chosen. The figure contains boxplots of the misclassification rates, computed over the 100 simulated datasets. First of all, a comparison between the results for $p=10$ and 200 reveals the importance of proper gene selection. When a lot of noise (genes that are not differentially expressed) is present, the performance of most classification methods (except of the tree methods) weakens. This conclusion is supported by Van Sanden et al. (2007). Furthermore, no gene-selection method is clearly outperforming all the others.
Figure 4.1: Boxplots for the misclassification rates (vertical axis) of various classification methods using $p=10$ (top two rows) and $p=200$ (bottom two rows) genes selected by different selection methods (horizontal axis). [1: Ttest, 2: Wilc, 3: SAM, 4: PAM, 5: PS, 6: BC-score, 7: ECDF, 8: Infgain, 9: Twoingr, 10: Summin, 11: Maxmin, 12: Gini, 13: Sumvar, 14: Extval].
For the classification trees, all gene-selection methods are performing similarly, except for SAM, BC-score, and ECDF. ECDF leads in general to poor results. The other two fail mainly when a relatively low number of genes is selected. The choice of the best gene-selection method seems to depend on the classification method used and on the number of selected genes. When the value of \( p \) becomes larger, the difference between the methods fades.

The results for kNN are similar. SAM, BC-score, and ECDF are not recommendable. There is, however, more variation between the other methods, also for large values of \( p \). The best methods appear to be Wilc, PS, PAM, Ttest (=BW), and Extval.

For the different forms of discriminant analysis and SVM the best methods are also Wilc, PS, PAM, Ttest (=BW) and Extval, while Maxmin, SAM, BC-score, and ECDF perform rather poorly. However, in contrary to the tree methods and kNN, ECDF improves on several methods when \( p \) is quite large (\( p \geq 200 \)), while the performance of Extval weakens.

In general, the radial SVM does not outperform some of the other classification methods. Considering all gene-selection procedures, DLDA is one of the best methods. Random Forrest is not doing so bad either, especially when the number of genes is large (\( p \geq 200 \), data not shown). In that particular setting it is outperforming all other methods. Furthermore, LDA, DQDA, and the radial SVM work reasonably well.

### 4.4.2 Other Settings

The performance of the selection and classification methods was also investigated for data with a proportionally decreased treatment effect, and with other distributional assumptions for the error term \( \varepsilon_{ijg} \). Table 4.1 displays the mean percentage of truly differentially expressed genes for all settings. The comparison between different selection methods appears quite similar under the different assumptions. There is, however, some variability over the settings with different distributional assumptions.

The size of the log-ratios of the two channels depends, besides on the dye and on the treatment effect, on the difference between the error terms of the model that generates the ratios (in model (3.1)). When this difference has the sign opposite to the treatment effect, it makes it more difficult or even impossible to pick up the treatment effect. Table 4.2 contains some basic properties (variance and kurtosis) of \( f_{X,Y}(z) \).
(i.e., the density function for the log ratio of gene-expression in the two channels). When \( X, Y \sim L(0, \sigma) \) with kurtosis 1.5 or \( AL(0, 1.2, \sigma) \) with kurtosis 1.68, \( f_{X−Y}(z) \) has a similar variance but a higher kurtosis than when \( X, Y \sim N(0, \sigma^2) \) with zero kurtosis. Therefore the distribution in question is more peaked with the same spread. It has more weight near zero and in the tails. However, the probability of getting a value in the tails is quite small, and values close to zero have less chance of diminishing the treatment effect to the point where it is not detected anymore. This explains why error terms with a symmetric or right-skewed asymmetric Laplace distribution allow for a better detection of the significant genes than when gene-expression follows a normal distribution. The opposite is true when \( X, Y \sim AL(0, 0.5, \sigma) \): \( f_{X−Y}(z) \) is more peaked than when \( X, Y \sim N(0, \sigma^2) \), but also has a lot more variability (see Table 4.2). The chance of getting values in the tails of this distribution is not so small in this case. It thus explains why fewer of the truly differentially expressed genes were picked up.

Table 4.2: Variance and kurtosis of \( f_{X−Y}(z) \) for normal (N), laplace (L), and asymmetric laplace (AL) distributions.

<table>
<thead>
<tr>
<th>( X, Y \sim )</th>
<th>Variance</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N(0, \sigma^2) )</td>
<td>( 2\sigma^2 )</td>
<td>0</td>
</tr>
<tr>
<td>( L(0, \sigma) )</td>
<td>( 2\sigma^2 )</td>
<td>1.5</td>
</tr>
<tr>
<td>( AL(0, 0.5, \sigma) )</td>
<td>( 4.26\sigma^2 )</td>
<td>2.67</td>
</tr>
<tr>
<td>( AL(0, 1.2, \sigma) )</td>
<td>( 2.14\sigma^2 )</td>
<td>1.68</td>
</tr>
</tbody>
</table>

Figure 4.2 displays median misclassification rates for certain combinations of the gene-selection and classification methods under different settings. The shift between the curves for a certain setting is directly related to the pattern of the variability for the different settings presented in Table 4.1 and just discussed. Only small differences are noticeable when comparing profiles of different settings. However, in general, the results are not substantially or consistently affected by proportionally decreasing the treatment effect, or by simulating data from the symmetric or asymmetric Laplace distribution. Similar results were found for the other combinations of the gene-selection and classification methods (data not shown).
Using DLDA for Classification

Figure 4.2: Misclassification rates for various combinations of gene-selection and classification methods. Setting 1 and 2: normally distributed data with constant (in black) and proportionally decreased (in red) treatment effect; setting 3, 4 and 5: symmetric (in green), asymmetric with skewness parameter $\kappa=0.5$ (in blue) and asymmetric with $\kappa=1.2$ (in black) Laplace distributed data. Upper penal using DLDA for classification, 1: Ttest, 2: Wilc, 3: SAM, 4: PAM, 5: PS, 6: BC-score, 7: ECDF, 8: Infgain, 9: Twoingr, 10: Summin, 11: Mazmin, 12: Gini, 13: Sumvar, 14: Extval; Lower penal using Ttest as gene selection, 1: Tree, 2: Tree (Bagging), 3: Tree(Boosting), 4: Random Forest, 5: kNN, 6: LDA, 7: DLDA, 8: DQDA, 9: Linear SVM, 10: Polynomial SVM, 11: Radial SVM.
4.4.3 E- and P-criteria

Figure 4.3 displays the misclassification rates for various classification methods combined with all considered gene-selection methods, including the E- and P-criteria. For the methods that do not determine the optimal number of genes ($p$) to select, several values of $p$ were considered. The E-criterion generally outperforms the P-criterion, except for some classification methods where both give almost the same result. The E-criterion also works quite well compared to other selection methods. It leads to the lowest median misclassification rate, except when combined with a tree method. In that case its performance is still very close to that of the best choice of gene-selection method and value of $p$.

![Figure 4.3: Misclassification rates of various classification methods (horizontal axis). For every classification method, the dots represent Ttest, Wilc, SAM, PAM, PS, BCScore, ECDF, Infgain, Twoingr, Summin, Maxmin, Gini, Sumvar, and Extval for values of $p=2, 5, 10, 20, 40, 200, 2000$. Stars correspond to the E-criteria and squares to the P-criteria. 1: Tree, 2: Tree (Bagging), 3: Tree(Boosting), 4: Random Forest, 5: kNN, 6: LDA, 7: DLDA, 8: DQDA:, 9: Linear SVM, 10: Polynomial SVM, 11: Radial SVM.](image)

4.5 Discussion

Lai et al. (2005) compared a number of univariate and multivariate gene-selection algorithms across several cancer diagnostic problems. They did not detect any sig-
significant improvement when employing multivariate gene-selection techniques. They argue that this finding could be due to very limited sample size. However, in our previous study (Van Sanden et al. 2007 and Chapter 3) we simulated correlation between certain genes. We also did not find this to have an effect on the performance of the classification methods. With these findings in mind and to keep the size of the study manageable, we decided not to include multivariate methods at this time. It is a topic for further research.

Simulations allow to overcome limitations related to the use of real-life data. For instance, we were able to investigate the performance of the gene-selection and classification methods in many controlled settings. Also, using real-life data we can only obtain a point estimator for the misclassification rate of a certain classification method. To get an estimate of the variance, re-sampling techniques would be necessary. With simulated data we directly obtain the distribution of the misclassification rate, with the precision depending on the number of simulated datasets. Additionally, the true classification, as well as the set of truly differentially expressed genes, are known. Hence we can evaluate the performance of the gene-selection methods directly and study the link between the performance of the classification method and the number of genes used for it that are truly differentially expressed. On the other hand, an important issue is whether the simulated data adequately capture the complexity of microarray measurements. The real-life data used as a basis for our simulations could be approximated by data simulated from a linear mixed model. In general, this may not be the case. To deal with this issue, we considered the use of non-normal distributions, the symmetric and asymmetric Laplace distribution.

Another issue is related to the computational complexity of such a study. Simulating the data and performing all the classification procedures on them is time consuming. Extending the simulations by including more methods or simulating data for more genes, more datasets or more settings is therefore not trivial. For this reason, in our study we chose for a limited, yet relevant from a practical point of view (as documented by Van Breda 2005), setting. In particular, the study was designed for cDNA microarrays. Although it is difficult to predict to what extent the conclusions can be generalized to oligonucleotide arrays, both platforms have common features that we expect to influence the gene-selection and classification methods in a similar way. For instance, in both cases a large number of genes is available for building a classifier and a lot of them are uninformative. It is therefore expected that the conclusions regarding the influence of the number of chosen genes might apply to
oligonucleotide arrays as well.

4.6 Conclusions

Though there are differences between the gene-selection methods, there is not a single one which surpasses all others. Wilc, PS, PAM, Ttest (=BW), and Extval lead to quite good results. The E-criterion is also worth noting. It works well most of the time and does not require predefining the number of genes to select. On the other hand, Maxmin, SAM, and BC-score are not performing as well as some of the other methods. ECDF is found to be the least interesting. The ECDF method was proposed to look for genes that might jointly discriminate between classes. For this reason, it allows for a larger overlap between the classes than the other gene-selection procedures. Its poorer performance seen in our simulations is therefore expected and understandable.

When comparing the effect of different gene-selection methods, it is not just a matter looking at how many truly differentially expressed genes a certain method is able of detecting. The method that leads to the largest percentage of truly differentially expressed genes (SAM) is not the method leading to the lowest misclassification rate. It is important, which particular genes (differentially expressed or noise) are included in the subset on which classification is based.

In general, DLDA and Random Forest give the lowest misclassification rates. DLDA is performing slightly better than Random Forest, except when a large number of genes is used. The radial SVM is clearly the best machine learning method. It is however in most cases not good enough to compete with DLDA.

In our simulations the introduction of longer tailed or asymmetric distributions for the microarray data or specification of a proportionally decreased, instead of a constant, treatment effect did not substantially affect the comparison between the different gene-selection and classification methods.

Results of this simulation study were published in Van Sanden et al. (2008).
Part II

Multiplicity Adjustment and Statistical Modelling in the Microarray Setting
Chapter 5

Multiplicity Adjustment: an Overview

5.1 Introduction

In multiple testing, strong control of the Family-Wise error rate (FWER) can be unnecessarily stringent in microarray settings (Xu and Hsu 2007). The aim of the microarray analysis is to identify many differentially expressed genes while controlling the number of false positives. Hence, we expect more than one false positive, but we wish to control the proportion of false positives as compared to the number of true positives.

In this chapter we discuss several approaches for multiplicity adjustment, namely, using the FWER (Hochberg and Tamhane 1987), the FDR (Benjamini and Hochberg 1995) and two newly developed error rate, the generalized FWER (gFWER, Hommel and Hoffmann 1988) and the generalized FDR (gFDR, Sarkar and Guo 2005b).

5.1.1 Controlling Type I Error

Benjamini and Hochberg (1995) consider the case where there are \( m \) hypotheses (Table 5.1) that need to be tested, among which \( m_0 \) are true null hypotheses and \( m_1 \) are false null hypotheses. Let \( V \) be the number of true hypotheses that we reject and \( R \) the total number of rejected hypotheses. Note that the values of \( m_0 \) and \( m_1 \) are unknown in practice.
The FWER (Hochberg and Tamhane 1987) is defined as the probability to reject erroneously at least one true null hypothesis, i.e., $\text{FWER} = P(V > 0)$. Here the term “family” refers to the collection of hypotheses $H_1, \ldots, H_m$ that is being considered for joint testing. Once the family has been defined, strong control of the FWER (at a joint level $\alpha$) requires that $\text{FWER} \leq \alpha$ for all possible constellations of true and false hypotheses (Lehmann and Romano 2005).

The FDR introduced by Benjamini and Hochberg (1995) is defined as the expected proportion of false rejection among the rejected hypotheses, $\text{FDR} = E(Q)$, where $Q = \frac{V}{R}$ when $R > 0$, and $Q = 0$ otherwise. Approaches based on the control of the FDR have gained their popularity in the microarray setting, because they lead to a higher power as compared to the methods that control the FWER.

Recently, two new error rates were introduced to control the number of false positives. The gFWER proposed by Hommel and Hoffmann (1988), is defined as the probability of rejecting at least $k$ true null hypotheses, i.e., $\text{gFWER} = P(V > k)$. Note that for $k = 0$ $\text{gFWER}$ equals FWER. On the other hand, the gFWER with $k > 1$ is appropriate when one is willing to tolerate one or more false rejections, provided that the total number of false rejections ($k$) is controlled. Controlling the gFWER overcomes the stringency of controlling the FWER, and has the advantage of controlling the number of mistakes as compared to controlling the FWER or FDR (Xu and Hsu 2007).

Recently, a generalization of the FDR, similar in spirit to the way the gFWER generalizes the FWER, was proposed by Sarkar and Guo (2005b). The gFDR is defined the expected proportion of $k$ or more false rejections among all rejections, i.e., $\text{gFDR} = E(V/R)$ when $V > k$, and $\text{gFDR} = 0$ otherwise. The procedures controlling the gFWER and gFDR are not implemented in this dissertation.

Controlling the type I error requires to the calculation of the raw $p$-values of the
test statistic that is used to test the null hypothesis of interest. This can be done by using the distribution of the test statistic under the null hypothesis or by RBI procedure in which the distribution of the test statistic under the null hypothesis is approximated using permutation techniques.

5.1.2 Obtaining p-values from Permutations

Consider a microarray with $m$ genes, and let $t_1, \ldots, t_i, \ldots, t_m$ be the test statistics of primary interest. By permuting the labels of arrays randomly, the (permutation) test statistics of $m$ genes are re-calculated.

Let $T$ be the permutation matrix, i.e.,

$$
T = \begin{pmatrix}
    t_{11} & t_{12} & \cdots & t_{1B} \\
    t_{21} & t_{22} & \cdots & t_{2B} \\
    \vdots & \vdots & \ddots & \vdots \\
    t_{m1} & t_{m2} & \cdots & t_{mB}
\end{pmatrix},
$$

where $B$ is the number of permutations and each element $t_{ib}$ of matrix $T$ is the test statistic for the $i$th gene in the $b$th permutation. The raw $p$-values (Ge et al. 2003) are calculated as

$$
P_i = \frac{\#(b : |t_{ib}| \geq |t_i|)}{B - 1},
$$

where $t_i$ is the observed test statistic for gene $i$.

Alternatively, the $p$-values can be obtained from the joint distribution of permutation test statistics for all the genes (Storey and Tibshirani 2003),

$$
P_i = \frac{\sum_{b=1}^{B} \sum_{j=1}^{m} (|t_{jb}| \geq |t_i|)}{(B - 1) \times m}.
$$

The raw $p$-values obtained by using (5.2) or (5.3) can be used as an input in the procedures for multiple testing adjustment discussed in the next section.

5.2 Procedures Adjusting for Multiple Testing

In this section, we outline several procedures used to control the FWER. These include the Bonferroni (Hochberg and Tamhane 1987), Holm (1979), and maxT (Westfall
and Young 1993) approaches. We also describe procedures controlling the FDR: the Benjamini and Hochberg step-up procedure (BH-FDR), the Benjamini and Yekutieli procedure (BY-FDR), and the significance analysis of microarray (SAM).

5.2.1 Procedures Controlling the FWER

Let $P_i$ be the raw $p$-value for the test statistic $t_i$ for gene $i$ ($i = 1 \ldots m$) and let $H_{0i}$ be the corresponding null hypothesis.

**Bonferroni’s Procedure**

According to the Bonferroni procedure (Hochberg and Tamhane 1987), $H_i$ is rejected if $P_i \leq \alpha/m$, where $\alpha$ is the desired level of the FWER. This ensures that the probability that at least one true hypothesis is rejected (when all are true) is not greater than $\alpha$. The Bonferroni adjustment is a single-step procedure, for which the adjusted $p$-value can be written as $\tilde{P}_i = \min(mP_i, 1)$ and $H_{0i}$ is rejected if $\tilde{P}_i \leq \alpha$.

**Holm’s Procedure**

Holm’s procedure (Holm 1979) is a step-down procedure, in which the raw $p$-values are sorted from the smallest (the most significant) to the largest. Let $P_{(1)} \leq P_{(2)} \cdots \leq P_{(i)} \cdots \leq P_{(m)}$ be the ordered $p$-value and let $H_{(1)}, H_{(2)}, \ldots, H_{(m)}$ be the corresponding null hypotheses.

Holm’s adjusted $p$-values are given by

$$
\tilde{P}_{(1)} = \min \left(mP_{(1)}, 1\right)
$$

$$
\tilde{P}_{(2)} = \min \left(\max(\tilde{P}_{(1)}, (m-1)P_{(2)}), 1\right)
$$

$$
\vdots
$$

$$
\tilde{P}_{(i)} = \min \left(\max(\tilde{P}_{(i-1)}, (m-i+1)P_{(i-1)}), 1\right)
$$

$$
\vdots
$$

$$
\tilde{P}_{(m)} = \min \left(\max(\tilde{P}_{(m-1)}, P_{(m)}), 1\right).
$$

$H_{(i)}$ is rejected if $\tilde{P}_{(i)} \leq \alpha$.

**maxT Procedure**

The maxT procedure, proposed by Westfall and Young (1993) in order to control the FWER, is discussed in the context of microarray analysis by Ge et al. (2003).
The starting point is the observed statistics and the permutation matrix $T$. Let $t_{(1)} \geq t_{(2)} \geq \ldots \geq t_{(m)}$ be the ordered values of the test statistics. The permutation matrix $T$ is sorted based on the original order of the observed statistics (Westfall and Young 1993). For each column of the permutation matrix the adjusted test statistics are calculated in the following way:

\[
\begin{align*}
    u_{m,b} &= |t_{m,b}| & \text{for } i = m, \\
    u_{i,b} &= \max(u_{i+1,b}, |t_{i,b}|) & \text{for } i = m - 1, \ldots, 1.
\end{align*}
\]

Once the adjusted matrix of the test statistics is obtained, the adjusted $p$-values are calculated (over the rows of the matrix):

\[
\hat{P}_{(i)}^* = \frac{\#(b: u_{i,b} \geq |t_{(i)}|)}{B}.
\]

### 5.2.2 Procedures Controlling the FDR

#### The BH-FDR Procedure

Let $P_{(1)} \leq P_{(2)} \ldots \leq P_{(m)}$ be the ordered $p$-values of and let $H_{(1)}, H_{(2)}, \ldots, H_{(m)}$ be the corresponding null hypotheses. The Benjamini-Hochberg procedure (BH-FDR, Yekutieli and Benjamini 1999) consists of rejection $H_{(1)}, H_{(2)}, \ldots, H_{(\ell)}$, where $\ell$ is the largest value of $i$ for which $P_{(i)} \leq \frac{i}{m} \alpha$. The BH-FDR adjusted $p$-values (Ge et al. 2003) are given by

\[
\hat{P}_{(i)} = \min_{k=i, \ldots, m} \left[ \min \left( \frac{m}{i} P_{(i)}, 1 \right) \right].
\]

The null hypothesis $H_{(i)}$ is rejected if $\hat{P}_{(i)} \leq \alpha$. The BH-FDR procedure controls the FDR for positively dependent test statistics (Bejamini and Yekutieli 2001).

#### The BY-FDR Procedure

Yekutieli and Benjamini (2001) proposed a modification of the BY-FDR procedure for controlling the FDR for any joint test statistic distribution. The BY-FDR procedure consists of a modification factor $\sum_{j=1}^{m} 1/j$. Similar to the BH-FDR procedure, $H_{(1)}, \ldots, H_{(\ell)}$ are rejected if $\ell$ is the largest integer, for which

\[
P_{(i)} \leq \frac{i}{m} \sum_{j=1}^{m} 1/j \alpha.
\]

The adjusted $p$-values are given by

\[
\hat{P}_{(i)} = \min_{k=i, \ldots, m} \left[ \min \left( \frac{m \sum_{j=1}^{m} 1/j}{i} P_{(i)}, 1 \right) \right].
\]
Significance Analysis of Microarray (SAM) Procedure

SAM (Tusher et al. 2001) is a procedure widely used in the microarray setting. SAM is a testing procedure, which estimates the FDR by using permutations under the assumption that all null hypotheses are true. The procedure consists of three components: (1) the adjusted test statistics, (2) an approximation of the distribution of the test statistics based on permutations, and (3) the control of the FDR.

For a two-group setting, the modified test statistic in SAM is given by,

\[
t_{k}^{SAM} = \frac{\bar{X}_{k} - \bar{X}_{l}}{s_{k} + s_{0}},
\]

(5.6)

where

\[
\bar{X}_{l} = \frac{\sum_{j=1}^{n_{l}}x_{jl}}{n_{l}}, \quad \bar{X}_{k} = \frac{\sum_{j=1}^{n_{k}}x_{jk}}{n_{k}},
\]

\[
s_{k} = \sqrt{\left(\frac{1}{n_{k}} + \frac{1}{n_{l}}\right) \left(\frac{\sum_{j=1}^{n_{k}}(x_{jk} - \bar{x}_{jk})^{2} + \sum_{j=1}^{n_{l}}(x_{jl} - \bar{x}_{jl})^{2}}{n_{k} + n_{l} - 2}\right)},
\]

and \(s_{0}\) is the fudge factor which is estimated from the data and is discussed later, \(k\) and \(l\) are the index of the two groups of array, and \(j\) is the index of the array.

SAM requires that the test statistic for each permutation is sorted for all the genes, such that the first row of the sorted matrix is the minimum test statistic across permutations, and the last row is the maximum, i.e.,

\[
T^{SAM} = \begin{pmatrix}
  t_{(1)1} & t_{(1)2} & \cdots & t_{(1)B} \\
  t_{(2)1} & t_{(2)2} & \cdots & t_{(2)B} \\
    \vdots & \vdots & \ddots & \vdots \\
  t_{(m)1} & t_{(m)2} & \cdots & t_{(m)B}
\end{pmatrix}.
\]

In \(T^{SAM}\), each element \(t_{(i)b}\) is the sorted test statistic for gene \(i\) in permutation \(b\). The expected values of the observed ordered statistics are approximated by the means of the rows of \(T^{SAM}\), given by \(\bar{t}_{(1)}^{SAM}, \bar{t}_{(2)}^{SAM}, \ldots, \bar{t}_{(m)}^{SAM}\) that are constructed in
The SAM procedure proposed by Tusher et al. (2001) is as follows:

1. Compute order statistics \( t_{(1)}^{SAM} \leq t_{(2)}^{SAM} \leq \cdots \leq t_{(m)}^{SAM} \).

2. Compute the permutation matrix \( T^{\text{SAM}} \).

3. Calculate the expected test statistics \( \bar{t}^{SAM}_{(1)}, \bar{t}^{SAM}_{(2)}, \ldots, \bar{t}^{SAM}_{(m)} \).

4. Plot the \( t_{(1)}^{SAM}, t_{(2)}^{SAM}, \ldots, t_{(m)}^{SAM} \) values versus the \( \bar{t}^{SAM}_{(1)}, \bar{t}^{SAM}_{(2)}, \ldots, \bar{t}^{SAM}_{(m)} \) values (SAM plot).

5. For a fixed threshold \( \lambda \), starting at the origin, and moving up to the right, find the first \( i = i_1 \) such that \( t_{(i)}^{SAM} - \bar{t}_{(i)}^{SAM} > \lambda \). All genes, for which \( t_{(i)}^{SAM} > t_{(i_1)}^{SAM} \), are called “significant positive”. Similarly, start at origin, move down to the left and find the first \( i = i_2 \) such that \( \bar{t}_{(i)}^{SAM} - t_{(i)}^{SAM} > \lambda \). All genes, for which \( t_{(i)}^{SAM} < t_{(i_2)}^{SAM} \), are called “significant negative”. For each \( \lambda \) define the upper cut-point \( C_{up}(\lambda) \) as the smallest \( t_{(i)}^{SAM} \) among the significant positive genes, and similarly define the lower cut-point \( C_{low}(\lambda) \).

6. For a grid of \( \lambda \) values, compute the total number of significant genes (from step 5), and the median number of falsely called genes, i.e., the median number of values among each of the \( B \) sets of \( t_{ib} \), \( i = 1, 2, \ldots, m \) that fall above cut-point \( C_{up}(\lambda) \) and below cut-point \( C_{down}(\lambda) \). Similarly, compute the 90th percentile of the number of falsely called genes.

7. Estimate \( \pi_0 \), the proportion of truly non-differentially expressed genes in the data set, as follows:

   (a) Compute the first and third quantiles of the permuted \( t^{\text{SAM}} \) values, denoted as \( q_{25} \) and \( q_{75} \) (the \( t_{(i)}^{SAM} \) are the values for the original data set; there are \( m \) such values).
(b) Compute \( \hat{\pi}_0 = \# \{ t_i \in (q_{25}, q_{75}) \}/(.5m) \).

(c) Let \( \hat{\pi}_0 = \text{min}(\hat{\pi}_0, 1) \).

8. The median and the 90th percentile of the number of falsely called genes from step 6, are multiplied by \( \hat{\pi}_0 \).

9. Pick a \( \lambda \) and the corresponding number of significant genes.

10. The FDR is estimated by the median (or the 90th percentile) of the number of falsely called genes divided by the number of significant genes.

**Estimation of the SAM Fudge Factor \( s_0 \)**

In the procedure described above, a fudge factor \( s_0 \) in the denominator of the test statistic (5.6) is used. It is calculated as the percentile of the gene-wise standard error distribution that minimizes the coefficient of variation (CV) of the test statistics. This modification is used to overcome bias for genes with expressions close to zero, which have a large value of the test statistic due to a small sample variance. By using an inflated standard error, SAM addresses the problem of the dependence of the value of the test statistic on the variance of expression levels for a particular gene. The calculation of \( s_0 \) is as follows:

1. Let \( s_\alpha \) be the \( \alpha \cdot 100\% \) percentile of \( s_i \) values. Let \( t_i^\alpha = (\bar{X}_1 - \bar{X}_0)/(s_i + s_\alpha) \).

2. Compute the 100 centiles of the \( s_i \) values, denoted by \( q_1 < q_2 \cdots < q_{100} \).

3. For \( \alpha \in (0, 0.05, 0.10, \ldots, 1.0) \)

   (a) compute \( \nu_j = \text{MAD}(t_i^\alpha | s_i \in [q_j, q_{j+1})), j = 1, 2, \ldots, m \), where MAD is the median absolute deviation from the median, divided by .64;

   (b) compute \( \text{cv}(\alpha) = \text{coefficient of variation of the } \nu_j \text{ values} \).

4. Choose \( \hat{\alpha} = \text{argmin}[\text{cv}(\alpha)] \), i.e., \( \hat{\alpha} \) is the quantile of the standard error that minimizes the coefficient of variation of the SAM test statistics.

5. Compute \( \hat{s}_0 = s_{\hat{\alpha}} \).
5.3 Discussion

The procedures controlling the FDR are preferred over those used for control of the FWER in terms of gaining power. The BH-FDR, BY-FDR, maxT and SAM procedures can be combined with RBI to adjust for multiple testing and make free of distributional assumptions. In what follows we investigate the performance of such combinations in terms of the error rate control and power. In Chapter 6 the procedures are applied to the setting of “many-to-one comparisons”, in which several treatments need to be compared with one control group. We compare between asymptotic inference and the RBI, and in particular, we investigate how the SAM procedure controls for the FDR in the presence of truly non-differentially expressed genes with a small variance. In Chapter 7, we address the problem of the choice of the fudge factor. We present the results obtained from a simulation study, in which several methods for choosing the fudge factor are used.

In this chapter, we have discussed the resampling-based procedures in controlling the FWER and FDR. The advantage of such resampling-based procedures is to avoid the distributional assumption of gene-expression data. Approaches, such as $p$-values obtained from permutations adjusted by the BH-FDR and BY-FDR procedure and the SAM procedure, can be used to find differentially expressed genes. However, an issue related to RBI is that, because of the small samples that are typically used in microarray experiments, the RBI $p$-value distributions can be coarse or granular. As a result, it will often be difficult to obtain $p$-values (as in (5.2)) that are below some specified level. To overcome this problem, two approaches discussed in this chapter, i.e., (1) raw $p$-values obtained using (5.3) and adjusted by the BH-FDR procedure and (2) the SAM procedure combine all resampled test statistics across all genes as the null distribution to obtain very small $p$-values. At the same time, both procedures also preserve the correlation between test statistics of all the genes. Thus, one can borrow strength across the genes and derive more powerful rejection regions for testing by assuming a statistic from a mixture of the null and alternative distributions, as well as from the pure null distribution (Efron et al. 2001). However, this is based on two assumptions: that the null distribution of the test statistic is the same for all genes; and that all genes are independent. Unfortunately, neither of the assumptions is necessarily correct.
Chapter 6

Multiple Testing Procedures for the Comparison of Several Treatments With a Common Control in a Microarray Setting

6.1 Introduction

In drug finding experiments, Dunnett’s test is frequently used to compare several treatments with one control. In this chapter, we discuss the situation of many-to-one comparisons in the context of a microarray experiment for drug discovery. The aim of such an experiment is to find genes whose expression levels differentiate between any of treatments and the control, and to find treatments that regulate the expression levels for a set of targeted genes. This type of study is important for finding informative genes, as well as finding potential active compounds (treatments).

To illustrate the setting, we consider the human epidermal squamous carcinoma cell line experiment described in Section 1.4.1, where four experimental conditions (three treatments and one control) in triplicate biological samples with zero concen-
Chapter 6. Many to One Comparisons in a Microarray Setting

The concentration of EGF are considered. The dataset consists of 12 arrays and 16,998 probe sets on each array. We investigate several procedures to make many-to-one comparisons while testing 16,998 genes simultaneously.

The prevalent issue related to gene-expression profiling is the adjustment for the large number of comparisons that need to be made. In order to overcome the conservativeness of the FWER controlling procedures, we apply the FDR controlling procedures. More specifically, we consider the BH-FDR (without and with the use of RBI) and the SAM (Tusher et al. 2001) procedure that control the FDR. The advantage of considering RBI with the BH-FDR procedure and the SAM is that they do not rely on an asymptotic distribution of test statistics. This is especially relevant for the case study, in which the sample size (three arrays per group) is rather small.

For a single gene, Dunnett’s (1955) single step procedure tests for the many-to-one comparisons simultaneously. In order to address the combined two-dimensional testing problem (testing thousands of genes and comparing several treatments with the control simultaneously), we adopt the strategy by testing $k \times m$ hypotheses simultaneously, where $k$ is the number of the many-to-one comparisons and $m$ is the total number of genes. The following four approaches are considered: (1) single step Dunnett’s $p$-values adjusted by the BH-FDR procedure, (2) $p$-values obtained by using permutations, adjusted by the BH-FDR procedure, (3) the SAM procedure without the fudge factor, and (4) the SAM procedure with the fudge factor.

In the last three procedures, we rely on permutations to preserve the structural correlation of $k$ comparisons per gene. This is analogous to Dunnett’s single step procedure, in which a joint multivariate distribution is used for the ordered statistics of many-to-one comparisons. Note that the SAM deals with the multiplicity problem by approximating the FDR, while the other three approaches are based on adjusting $p$-values using the BH-FDR procedure.

The content of this chapter is organized as follows. In Section 6.2 we briefly discuss the methods used for normally distributed data and Dunnett’s procedure for comparing several treatments with one control (Dunnett 1955). Section 6.3 presents results of application of the different methods to the case study data. In Section 6.4 we conduct simulation studies in order to investigate the performance of the four approaches. The chapter ends with a discussion in Section 6.5.
6.2 Comparing Several Treatments With the Control

In the considered case study multiple comparisons arise from comparing several treatments with a control and from testing thousands of hypotheses (genes) simultaneously. To identify genes differentiating between the treatments and the control, the use of an analysis of variance (ANOVA) type of model can be considered. Kerr et al. (2001) formulated a general ANOVA model for the log-transformed gene-expression measurements. Wolfinger et al. (2001) discussed a “gene by gene” modeling approach, in which gene-specific linear mixed models are used to determine the significance and magnitude of treatment effects independently for each gene. Let \( X_{ijk} \) be the \( i \)th gene-expression of array \( j \) in treatment group \( k \), \( i=1,...,m, \ j=1,...,n, \ k=0,...,3 \). The gene-specific linear model is given by

\[
X_{ijk} = \mu_{ik} + \varepsilon_{ijk}; \quad \varepsilon_{ijk} \sim N(0, \sigma_i^2), \tag{6.1}
\]

where \( \mu_{ik} \) is the mean expression level for treatment \( k \) for gene \( i \), and \( \mu_{i0} \) is the mean expression level for the control group for gene \( i \). Inference can be made using the estimates of the means with their variabilities.

In order to assess the effect of treatments, the following hypotheses are to be tested:

\[
H_{0i} : \mu_{i0} = \mu_{i1} = \mu_{i2} = \mu_{i3} \quad \text{no treatment effect},
H_{1i} : \mu_{ik} \neq \mu_{il} \quad \text{for at least one pair of} \ k \ \text{and} \ l \ (k \neq l). \tag{6.2}
\]

The null hypothesis can be tested using an F-test for the one-way ANOVA model (6.1). In particular, we focus on the comparison of the treatments versus the control group. Hence, the alternative hypotheses of primary interests, as considered by Dunnett (1995), are

\[
H_{01i} : \mu_{i0} - \mu_{i1} = 0, \quad H_{11i} : \mu_{i0} - \mu_{i1} \neq 0,
H_{02i} : \mu_{i0} - \mu_{i2} = 0, \quad H_{12i} : \mu_{i0} - \mu_{i2} \neq 0,
H_{03i} : \mu_{i0} - \mu_{i3} = 0, \quad H_{13i} : \mu_{i0} - \mu_{i3} \neq 0. \tag{6.3}
\]

The test statistics for the hypotheses can be written as

\[
t_{ik} = \frac{\bar{X}_{ik} - \bar{X}_{i0}}{s_i \sqrt{\frac{1}{n_k} + \frac{1}{n_0}}} \quad i = 1,...,m, \ k = 1,2,3, \tag{6.4}
\]
where $s_i$ is the pooled variance of gene $i$, $\bar{X}_{ik}$ is the estimated mean gene-expression for the $k$th treatment group for gene $i$, $n_k$ is the number of arrays for treatment $k$, and $n_0$ is the number of arrays in the control group.

For comparing the treatment means with the control mean, Dunnett (1955, 1964) proposed the following set of $(1-\alpha)$ level simultaneous confidence intervals for gene $i$ ($i = 1, \ldots, m$) and the comparison $k$ ($k = 1, 2, 3$):

$$\mu_{ik} - \mu_{i0} \in \bar{X}_{ik} - \bar{X}_{i0} \pm |t|^\alpha_{k,\nu,\rho} s_i \sqrt{1/n_k + 1/n_0},$$

(6.5) where $|t|^\alpha_{k,\nu,\rho}$ is the two-sided upper $\alpha$ point of the $k$-variate equicorrelated $t$-distribution with common correlation $\rho$ (equal to $n_k/(n_k + n_0)$) and $\nu$ (equal to $\sum_{k=0}^{3} n_k - 2$) degrees of freedom. The values of $|t|^\alpha_{k,\nu,\rho}$ have been tabulated in Bechhofer and Dunnett (1988). From this set of confidence intervals, statements $\mu_{ik} > \mu_{i0}$ ($\mu_{ik} < \mu_{i0}$) can be made for treatment $k$ satisfying $T_{ik} > |t|^\alpha_{k,\nu,\rho}$ ($T_{ik} < |t|^\alpha_{k,\nu,\rho}$). The probability of all such decisions being correct is no less than the confidence level $1 - \alpha$.

As compared to the Bonferroni procedure, the Dunnett’s procedure is more powerful, which takes the correlation of the test statistics into account and tests against a limited number of alternatives. Note that Dunnett’s (1955) procedure is a single-step procedure, in which the null hypotheses (6.3) for gene $i$ are tested simultaneously. The permutation $p$-values and the SAM approach are obtained by using the methods described in Section 5.1.2 and Section 5.2.2.

### 6.3 Application to the Data

#### 6.3.1 Multiple Testing Using Dunnett $p$-values

First, we present the results using the Dunnett $p$-values. The single-step testing scheme consists of testing all $3 \times 16,998$ genes simultaneously. By testing the three null hypotheses in (6.3), gene-wise Dunnett $p$-values for the three comparisons are obtained using the Dunnett’s procedure and the BH-FDR procedure is used to control the FDR. Table 6.1 (second column) presents the results using this approach.

Among the $3 \times 16,998$ tests, the null hypothesis is rejected for $4292$ (=$958 + 2 \times 749 + 3 \times 612$) tests, identifying 2885 genes as significant for at least one comparison between the treatments and the control. The number of genes with one significant comparison is 958; there are 749 genes with two significant comparisons, and 612
Table 6.1: Number of significant genes identified using (1) Dun: Dunnett adjusted p-values adjusted by the BH-FDR procedure, (2) BH: permutation p-values adjusted by the BH-FDR procedure, (3) -s0: the SAM procedure without the fudge factor, and (4) +s0: the SAM procedure with the fudge factor.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Dun</th>
<th>BH</th>
<th>-s0</th>
<th>+s0</th>
</tr>
</thead>
<tbody>
<tr>
<td># sign genes</td>
<td>2882</td>
<td>3586</td>
<td>5223</td>
<td>613</td>
</tr>
<tr>
<td># Comp 1*</td>
<td>958</td>
<td>1555</td>
<td>2262</td>
<td>262</td>
</tr>
<tr>
<td># Comp 2*</td>
<td>749</td>
<td>1244</td>
<td>1514</td>
<td>232</td>
</tr>
<tr>
<td># Comp 3*</td>
<td>612</td>
<td>787</td>
<td>1447</td>
<td>119</td>
</tr>
</tbody>
</table>

1*: # of genes declared significant for one treatment compared with the control
2*: # of genes declared significant for two treatments compared with the control
3*: # of genes declared significant for all the three treatments compared with the control

genes with all three significant comparisons between the three treatments and the control.

The Dunnett p-values obtained in this section are valid only if the distributional assumption in (6.5) holds. In order to overcome this problem, one can use RBI technique, which does not require any assumption about the distribution of the gene-expression. In the next section we present the results for permutation p-values adjusted by the BH-FDR procedure and by the SAM procedure.

6.3.2 Resampling-based Multiple Testing

The permutation p-values are obtained by using 1000 random permutations of the 12 sample labels and by calculating the test statistics for the newly formed treatment groups. Adjusting the permutation p-values by the BH-FDR procedure leads to 3586 significant genes, with 1555 genes with one significant comparison, 1244 genes with two significant comparisons, and 787 genes with all three significant comparisons.

The second resampling-based approach is the SAM. The choice of fudge factor for the SAM procedure is made based on the algorithm provided in Section 5.2.2. Figure 6.1 plots the CVs versus the quantiles of standard error using the human epidermal squamous carcinoma cell line data. Based on the graph, the 60%th quantile
Figure 6.1: Plot of quantiles of the standard error vs. the CV (coefficient of variation) of the test statistics: selection of the fudge factor in the SAM procedure.

is chosen. Figure 6.2 illustrates the selection of threshold $\lambda$ to control the FDR. Panel $a$ shows the relationship between the FDR and $\lambda$, which allows to choose the $\lambda$ with a desired level of the control of the FDR. For instance, to control the FDR at 0.05, the required $\lambda$ is 0.98. Panels $b$ and $c$ display the number of significant findings and false positives in function of $\lambda$. The last panel shows the observed and expected $t$-test statistics, where the genes beyond the band (i.e., the absolute difference between the expected and the observed test statistics is larger than a certain $\lambda$) are considered to be significant. With $\lambda$ equal to 0.98, the number of significant genes is 613 and the median number of false positives is 35.

The analysis without fudge factor $s_0$ leads to 5224 genes declared to be significant (Table 6.1, the fourth column: $-s_0$), while the analysis with fudge factor $s_0$ reduces the number of significant findings to 613 genes (Table 6.1, the fifth column: $+s_0$). Note that the SAM procedure without the fudge factor and the permutation $p$-values adjusted by the BH-FDR procedure differ only in the procedure adjusting the multiplicity while they use the same test statistics.

Figure 6.3 shows a plot of $t$-test statistics against the gene-specific standard error.
Figure 6.2: \textit{SAM(}+s_0\textit{)} plots: a. FDR vs. \( \lambda \); b. the number of significant genes vs. \( \lambda \); c. the number of false positives vs. \( \lambda \); d. observed vs. expected test statistics. FDR 50\% and FDR 90\% denote the estimation of FDR using the median and 90th percentile of the number of falsely positive genes (namely, FP 50\% and FP 90\%), respectively.

Red points in the outer zone represent tests, for which the null hypothesis is rejected by the SAM procedure with the fudge factor. These significant tests, together with the tests represented by black points, are rejected by the SAM procedure without the fudge factor. Grey points in the middle zone indicate the tests, for which the null hypothesis is not rejected. Note that the use of the fudge factor affects a large number of genes in this data set.

As we observed in Figure 6.3, a large number of genes with small variance are declared significant without the adjustment of the fudge factor in the case study. For the small sample size, the issue of protecting the inflated test statistics via the fudge factor and the proportion of genes with small variance become important in
Figure 6.3: Plot of $t$-test statistics vs. their standard errors. The grey zone in the middle contains the non-significant tests; red points in the outer zone are tests declared significant by the SAM procedures with and without the fudge factor; black points are tests declared significant only by the SAM procedure without the fudge factor.

this microarray setting. In the next chapter, we focus on conducting a simulation study to investigate the effect of fudge factor on the control of the FDR and power obtained using various procedures.

Note that the Dunnett $p$-values adjusted by the BH-FDR procedure and the permutation $p$-values adjusted by the BH-FDR procedure find the number of significant genes in-between the numbers found by the SAM procedure without and with the
fudge factor. There is similarity among the four approaches: the number of genes with (one, two, and three) significant comparisons is decreasing in the same direction, with the largest number of genes with just one significant comparison and the smallest number of genes with all three comparisons.

### 6.4 A Simulation Study

In this section, we perform a simulation study to investigate the performance of the four approaches used in the analysis used in the case study (see Section 6.3). In particular, we focus on the control of the FDR and the power of the procedures.

#### 6.4.1 Simulation Setting

We generate the data with the same number of genes (16,998) per array and use the gene-specific variances as observed in the case study. We assume that about 10% of the total genes (i.e., 1700 genes) are truly differentially expressed. Gene-expression levels are generated according to the model

\[ x_{ijk} \sim N(\mu_{ik}, \hat{\sigma}_i^2), \]

\[ i = 1, \ldots, 16,998, \]

\[ j = 1, \ldots, n_k, \]

\[ k = 0, 1, 2, 3, \]

where \( \hat{\sigma}_i^2 \) is the estimated gene-specific variance from the data. The means for the treatment groups are specified in the following way:

\[
\mu_{ik} = \begin{cases} 
\delta_{ik} \times \hat{\sigma}_i, & i \leq 1700 \text{ and } k = 1, \\
\delta_{ik} \times \hat{\sigma}_i, & i \leq 1309 \text{ and } k = 1, 2, \\
\delta_{ik} \times \hat{\sigma}_i, & i \leq 714 \text{ and } k = 1, 2, 3, \\
0, & \text{otherwise},
\end{cases}
\]  

(6.6)

where \( \delta_{ijk} \times \hat{\sigma}_i \) represents the mean of the treatment group for the differentially expressed genes. It is assumed that \( \delta_{ik} \sim U(2.8, 4.5) \). Among the 1700 truly differentially expressed genes, 319 genes are assumed to be differentially expressed for only one treatment; 595 genes are assumed to be differentially expressed for two treatments; and 714 genes are assumed to be differentially expressed for all three treatments. In total, four settings with sample size of three, four, five, and six arrays per treatment group are considered. For each setting 100 data sets are generated.

We choose \( \delta_{ik} \sim U(2.8, 4.5) \) for the differentially expressed genes because the calculated \( t \)-test statistics calculated are in the range of \( 2.8 \times \sqrt{3/2} \) and \( 4.5 \times \sqrt{3/2} \). A \( t \)-test statistic value in that range is declared significant (without multiple testing). However, under multiple testing adjustment, it is difficult to obtain a critical value
Chapter 6. Many to One Comparisons in a Microarray Setting

for rejection of the null hypothesis. Hence, we hope that the range generated above is large enough to be declared significant with a multiplicity adjustment. The range is obtained by considering $t_{ik} = \frac{\hat{\mu}_{ik}}{(\hat{\sigma}_i \times \sqrt{2/3})} = \frac{\delta_{ik} \times \hat{\sigma}_i}{(\hat{\sigma}_{ik} \times \sqrt{2/3})} = \frac{\delta_{ik} \times \sqrt{3/2}}{\hat{\sigma}_i}$ and using the pooled variance $\hat{\sigma}_i = \hat{\sigma}_{ik}$ in accordance with Dunnett’s procedure.

6.4.2 Simulation Results

Using the simulated data, we compare the number of significant genes identified by the four procedures. The two columns of Figure 6.4 show the distribution of the FDR and power with 100 simulated datasets. For both SAM procedures (with and without the fudge factor), the FDR is well controlled at around 5%. For each simulated dataset, the fudge factor selection is based on the automatic calculation (minimizing CV of standard error) described in Section 5.2.2.

With three arrays per group, the power obtained for the Dunnett $p$-values and the permutation $p$-values adjusted by the BH-FDR (32% and 34%, respectively) is much lower than the power for the SAM approach with (54%) and without (72%) the fudge factor. Note that for the small sample size the permutation approach yields a slightly higher median value of the FDR than 0.05. Increasing the sample size to four, five, or six arrays per group substantially improves the power of the four approaches. For the SAM procedure with the fudge factor, the power increases to 85% for sample size of 16 arrays, to 96% for 20 arrays, and to almost 99% for 24 arrays. For permutation $p$-values adjusted by the BH-FDR procedure, the power increases to 72% for 16 arrays, to 93% for 20 arrays, and to almost 98% for 24 arrays.

In Table 6.2 we present the number of genes with significant comparisons between the treatments and the control. As expected, the same ordering of the number of significant findings for each procedure can be observed. The number of genes with one, two, and three significant comparisons increases in a similar way as compared to the true number used for simulation. For example, for the SAM procedure with the fudge factor, and three arrays per group, the mean number of genes with one significant comparison is large, i.e., 560.5 genes, among which on average only 205 genes have a true difference in expression for one treatment; among 368.5 genes with two significant comparisons, on average 209 genes have a true difference in expression for two treatments; and among 271.5 genes with three significant comparisons, on average 269 genes have a true difference in expression for three treatments. We can conclude from the table that, although the number of significant findings decreases
Figure 6.4: Simulation results: $U \sim (2.8, 4.5)$. The boxplots in each panel of the figure present the results for Dunnett approach (Dun), the permutation approach (BH), the SAM procedure without the fudge factor ($SAM - s_0$), and the SAM procedure with fudge factor ($SAM + s_0$). Plots in the left panel show the FDR achieved for three, four, five, and six arrays per group. Plots in the right panel show the power.

as the sample size increases, the average number of truly significant findings corresponds to the number simulated in the setting. Note that results for genes with three significant comparisons contain almost no false positives.

The SAM procedure without the fudge factor seems to outperform the SAM procedure with the fudge factor with respect to power. This is because no non-differentially
expressed genes with a small standard error were generated in this simulation. In such a case that there is no need to use the fudge factor, so the use of the fudge factor diminishes the power of the procedure.

6.5 Discussion

The aim of the microarray experiment presented in this chapter was to find genes whose expression levels differentiate between any of treatments and the control. These genes are useful as indicators for the active treatment effect. In terms of multiplicity adjustment, such an experiment requires an adjustment for comparisons within a gene (treatment versus control) and an adjustment for performing tests for many genes simultaneously. In this chapter, we considered four approaches, which combine the two-dimensional testing problem simultaneously. The analysis of the case study presented in Section 6.3 reveals substantial differences between different methods for multiplicity adjustment. The SAM procedure with the fudge factor led to the smallest number of significant findings as compared to the other three procedures, while the SAM procedure without the fudge factor resulted in the largest number of significant discoveries. This difference motivated the simulation study in Section 6.4.

The performance of Dennett’s approach adjusted by using the BH-FDR procedure, permutation $p$-values approach using the BH-FDR adjustment, the SAM method with and without the fudge factor were compared in the simulation study for the case where non-differentially expressed genes with small variance were not present. We have shown that with a small sample size (three observations per treatment group), the SAM approach without the fudge factor performed better with respect to power. When sample size increased, the FDR obtained for the four approaches was well-controlled at the desired level and the power obtained by all approaches became comparable.

However, as we argue in Chapter 5, the main issue related to the SAM procedure is not only how the procedure controls the FDR but, more importantly, how the procedure controls the FDR when the small variance genes are present. Moreover, the SAM procedure requires a selection of the value of the fudge factor and fixing the value during the permutations. In this chapter, the fudge factor $s_0$ was chosen in order to minimize the CV of the SAM test statistics. Other methods, such as the penalized regression (Wu 2005), Receiver Operating Characteristics Curve (ROC curve, Broberg 2003), Empirical Bayes (EB, Efron and Tibshirani 2002) are available
for this purpose as well. In the next chapter, we present two simulation studies. The first one is conducted in order to investigate the performance of the SAM procedure when small variance genes are present, while in the second simulation study, we compare the performance of several selection methods for the fudge factor.

The contents of chapter were summarized in the manuscript (Lin et al. 2007c).
Table 6.2: Simulation results. The mean number of genes with significant comparisons using (1) Dun: Dunnett p-values adjusted by the BH-FDR procedure, (2) BH: permutation p-values adjusted by the BH-FDR procedure, (3) -s₀: the SAM procedure without the fudge factor, and (4) +s₀: the SAM procedure with the fudge factor. The mean number of genes with the corresponding number of truly significant comparisons is given in parentheses.

<table>
<thead>
<tr>
<th>N</th>
<th>simulated # of genes</th>
<th>Dun</th>
<th>BH</th>
<th>-s₀</th>
<th>+s₀</th>
</tr>
</thead>
<tbody>
<tr>
<td>391</td>
<td>1*</td>
<td>334(120)</td>
<td>442(111)</td>
<td>619.5(280)</td>
<td>560.5(205)</td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>595</td>
<td>197.5(127)</td>
<td>254.5(136)</td>
<td>529.5(357)</td>
<td>368.5(209)</td>
</tr>
<tr>
<td></td>
<td>714</td>
<td>166.5(165)</td>
<td>120(117)</td>
<td>387.5(382)</td>
<td>271.5(269)</td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>391</td>
<td>523(296)</td>
<td>623(321)</td>
<td>581(360)</td>
<td>605(334.5)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>595</td>
<td>495(394)</td>
<td>610(445)</td>
<td>628(519)</td>
<td>564.5(443)</td>
</tr>
<tr>
<td></td>
<td>714</td>
<td>495(492.5)</td>
<td>477(473)</td>
<td>603.5(598)</td>
<td>534(529.5)</td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>391</td>
<td>484(357)</td>
<td>569(369)</td>
<td>552.5(380)</td>
<td>562(378)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>595</td>
<td>573(516)</td>
<td>632(544)</td>
<td>624(570)</td>
<td>620(560)</td>
</tr>
<tr>
<td></td>
<td>714</td>
<td>639.5(635)</td>
<td>632(629)</td>
<td>683(677)</td>
<td>670(666)</td>
</tr>
<tr>
<td>24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>391</td>
<td>457.5(377.5)</td>
<td>546.5(383)</td>
<td>541.5(385)</td>
<td>544.5(385)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>595</td>
<td>592(564.5)</td>
<td>622(575)</td>
<td>616(585)</td>
<td>616(581)</td>
</tr>
<tr>
<td></td>
<td>714</td>
<td>690(687)</td>
<td>683(679)</td>
<td>708(703)</td>
<td>701(698)</td>
</tr>
</tbody>
</table>

1: number of genes with one significant comparison
2: number of genes with two significant comparisons
3: number of genes with three significant comparisons
Chapter 7

A Comparison of Procedures for the Selection of Fudge Factor for the SAM: a Simulation Study

7.1 Introduction

One of the main components of the SAM procedure is the adjustment of the test statistic. The introduction of the fudge factor to the test statistic aims at deflating the large value of test statistics due to small standard error of gene-expression. However, the values of the modified test statistic decrease simultaneously for all the genes, and in particular affect the truly differentially expressed genes with a small standard error. We noticed that the fudge factor does not effectively improve the power and the control of the FDR as compared to the SAM procedure without the fudge factor.

In this chapter, we study how the SAM procedure modifies the test statistics and include some other methods, which adjust the test statistic in a similar way as the SAM procedure. Namely, we considered the penalized $t$-test statistic (Wu 2005), selection of the fudge factor using receiver operating characteristics curve (ROC, Broberg 2003) and Empirical Bayes (EB, Efron and Tibshirani 2002). We compare
the performance of these methods together with the original SAM procedure in terms of power and the control of the FDR using a simulation study.

The contents of the chapter is organized as follows. In Section 7.2 we use the human epidermal squamous carcinoma cell line data described in Section 1.4.1 to illustrate the SAM procedure and conduct a simulation study to investigate the choice of the fudge factor in the SAM procedure. In Section 7.3 we describe another three methods selecting the fudge factor. In Section 7.4 we discuss the use of simulation data to study these methods, and compare the results in terms of the power and control of the FDR. The chapter ends with a discussion in Section 7.5.

7.2 Selection of the Fudge Factor in SAM

7.2.1 Graphical Interpretation of SAM

In order to investigate the effect of the fudge factor, we decompose the number of the the true null hypotheses $m_0$ (in Table 5.1) into two numbers: $m_0^0$ for truly non-differentially expressed genes with a moderate to large variance and $m_0^1$ for truly non-differentially expressed genes with a relatively small variance. Accordingly, the number of the falsely rejected true hypotheses $V$ is decomposed into $V^0$ and $V^1$. Consequently, the FDR can also be decomposed as $\text{FDR} = FDR^0 + FDR^1$, and can be estimated as $V^0/R + V^1/R$, where $R$ is the total number of genes declared significant.

To show the effect of the fudge factor, we compare the values of the usual $t$-test statistic and SAM test statistic and investigate how the SAM test statistics are affected by the fudge factor. Figure 7.1 shows the effect sizes (numerator of $t$-test statistics) versus the absolute values of SAM $t$-test statistics without (Figure 7.1a) and with the fudge factor (Figure 7.1b) using the human epidermal squamous carcinoma cell line data. We observe that a large number of genes have small effect sizes but large test statistic values, which are lying along the zero vertical line (Figure 7.1a). With the introduction of the fudge factor, the values of the test statistics dramatically decrease (Figure 7.1b). This also addresses the need for using the fudge factor to protect against genes with large test statistics, but due to small variances.

Figure 7.2 illustrates how the fudge factor affects genes with different variances. The two axes of Figure 7.2a represent the numerator (the absolute value of the effect size) and denominator (the standard error) of the $t$-test statistics. The angle $\alpha$ (between the y-axis and the solid line) for three genes with a small, medium, and
large standard error ($s_1 < s_2 < s_3$) and a corresponding small, medium, and large effect size ($\Delta_1 < \Delta_2 < \Delta_3$) is shown. The test statistic of three genes is equal to $t_1 = \Delta_1/s_1 = t_2 = \Delta_2/s_2 = t_3 = \Delta_3/s_3 = \cot(\alpha)$. When fudge factor $s_0$ is added to the denominator (extending the standard errors, $s_1$, $s_2$, and $s_3$ horizontally), the new angles are $\beta_1 + \alpha$, $\beta_2 + \alpha$, and $\beta_3 + \alpha$, respectively. The three newly formed angles are between the y-axis and the dotted line ($\alpha + \beta_1$), the short dashed line ($\alpha + \beta_2$), and the long dashed line ($\alpha + \beta_3$), respectively. Thus, the SAM test statistics for the three genes become $\cot(\alpha + \beta_1)$, $\cot(\alpha + \beta_2)$, and $\cot(\alpha + \beta_3)$, respectively. The values of SAM test statistics are illustrated by the cotangent function in Figure 7.2b. Panel $b$ shows the same $t$-test statistic value corresponding to angle $\alpha$ in the left plot, and the three decreased test statistic values $t_{1_{\text{SAM}}}^{\text{SAM}} < t_{2_{\text{SAM}}}^{\text{SAM}} < t_{3_{\text{SAM}}}^{\text{SAM}}$ after adding the fudge factor in the right plot.

Let $s_{(1)}, s_{(2)}, \ldots, s_{(m)}$ be the order standard errors in the microarray experiment with $m$ genes. Let $s^{(0)}, s^{(1)}, \ldots, s^{(100)}$ be centiles of $s_{(1)}, s_{(2)}, \ldots, s_{(m)}$ and let fudge factor $s_0 = s^{(q)}$. It is easy to see that for gene $i$, the SAM test statistic with fudge factor ($t_i^{\text{SAM}}$) and the $t$-test statistic ($t_i$) have the following relationship:
Chapter 7. Choice of Fudge Factor: a Simulation Study

Figure 7.2: Graphical interpretation of the SAM test statistic: a. SAM test statistics; b. cotangent function.

\[
\begin{align*}
    &t_{i}^{SAM} < (1/2)t_i \quad \text{if } s_{(i)} < s^{(q)}, \\
    &t_{i}^{SAM} = (1/2)t_i \quad \text{if } s_{(i)} = s^{(q)}, \\
    &t_{i}^{SAM} > (1/2)t_i \quad \text{if } s_{(i)} > s^{(q)}. 
\end{align*}
\]

Hence, the SAM test statistic with a fudge factor is smaller than 1/2 of the t-test statistic for genes with standard error smaller than the fudge factor. Moreover, the ratio between the SAM test statistic without and with the fudge factor is \(s_i/(s_i + s_0)\) since \(|t_i^{SAM} = s_i/(s_i + s_0) \times (\Delta_i/s_i)\). Depending on \(s_i\), the SAM test statistic decreases with the ratio of \(s_i/(s_i + s_0)\). From the illustration above, it is easy to note that the fudge factor has bigger effect on genes with small standard error than those with larger standard errors.
7.2.2 Effect of the Fudge Factor on the FDR and Power

The SAM procedure introduces the following dilemma. Assuming that $m_1^1 > 0$, the number of non-differentially expressed genes, for which the variance of genes is relatively small, the analysis without correcting the test statistic by the fudge factor $s_0$ is expected to lead to significant findings for these genes. This implies that the FDR among this subset of small variance genes will not be controlled. As a result, the overall FDR will not be controlled. On the other hand, the analysis, in which the test statistics are corrected using the fudge factor, is expected to solve the false positive finding problem for small variance genes, but at the same time to reduce the power. Thus, the question is how the FDR is controlled for small variance genes with the use of the SAM procedure. We use a simulation study to investigate this issue.

7.2.3 Simulation Setting

Similar to the simulations described in section 6.4.1, 100 data sets, with microarrays containing 16,998 genes, among which 10% of genes (1700 genes) are truly differentially expressed, are generated. In order to study the effect of the fudge factor on genes with a small treatment effect and a small variance, we simulate four settings with different proportion of non-differentially expressed genes with a small variance: (1) 0, (2) 5% (850 genes), (3) 10% (1700 genes), and (4) 20% (3400 genes) of the total number of genes on the array.

In order to achieve the effect of large test statistics for genes with small variance, we simulate the means of treatment groups from normal distribution $N(0, 0.3)$ for the four treatment groups. As a result, $E(\mu_1) = E(\mu_2) = E(\mu_3) = E(\mu_4)$, it ensures the homogeneity of expected means for the four treatment groups, but generates the possibility of small mean differences. For the four settings, the number of small variance genes varies from zero, 850, 1700, to 3400. The variance of these genes is set equal to 0.0036, which is 5% quantile of the standard error in the case study data. In total 12 arrays (three arrays per groups) are generated for four treatments.

As an overview, the number of truly differentially expressed genes ($m_1^1$), truly non-differentially expressed genes with small variances ($m_0^1$) and moderate variances ($m_0^0$) for the four settings is given in Table 7.1.

To illustrate the test statistic values for each subset of genes simulated, Figure 7.3 plots the test statistic values vs. their standard error for one simulated data under the setting of 850 small variance genes. As expected, a lot of small variance genes
Table 7.1: Number of genes in each subset simulated under four settings.

<table>
<thead>
<tr>
<th>$m_0^0$</th>
<th>$m_0^1$</th>
<th>$m_1$</th>
<th>$m$</th>
</tr>
</thead>
<tbody>
<tr>
<td>15298</td>
<td>0</td>
<td>1700</td>
<td>16998</td>
</tr>
<tr>
<td>14448</td>
<td>850</td>
<td>1700</td>
<td>16998</td>
</tr>
<tr>
<td>13598</td>
<td>1700</td>
<td>1700</td>
<td>16998</td>
</tr>
<tr>
<td>11898</td>
<td>3400</td>
<td>1700</td>
<td>16998</td>
</tr>
</tbody>
</table>

$m_0^1$ shows inflated test statistics (in red) lying at the left of vertical dashed line (i.e., the standard error of 0.06). Truly differentially expressed genes (in black) are mostly lying outside the band of non-differentially expressed genes (in grey). This example confirms our simulation setting is realistic.

Figure 7.3: Plot of the test statistics vs. the standard error for one simulated data under the setting of 850 small variance genes.
7.2.4 Simulation Results

Figure 7.4 compares the results for the four settings described above. Figure 7.4a shows the relationship between the power and the FDR, where the power is estimated as the mean true discovery proportion and the FDR is estimated as the mean false discovery proportion across 100 simulated datasets. The four lines (representing four settings) show the power and the FDR obtained using no fudge factor, and 1%, . . . , 100% centile of the standard error distribution as the fudge factor. The result of automatic choice of SAM is indicated using the capital letter (A, B, C, D) for the four settings, respectively. For the last three settings, where the number of non-differentially expressed genes with small variances increases, the FDR is no longer controlled at the desired level. Also, the power of the procedure decreases, because a larger quantile of standard error is used as the fudge factor.

Let us decompose the FDR into two parts, i.e., \( \text{FDR} = \text{FDR}^0 + \text{FDR}^1 \), where \( \text{FDR}^0 \) is estimated as the mean proportion of false positives among \( m_0 \) non-differentially expressed genes with moderate to large variances, while \( \text{FDR}^1 \) can be estimated as the mean proportion of false positives among \( m_1 \) non-differentially expressed genes with small variances. From Figure 7.4c, we can see that the \( \text{FDR}^0 \) for SAM procedure is maintained around 5% regardless of the setting. However, the \( \text{FDR}^1 \) is not controlled at the desired level, implying that SAM procedure fails to remove genes with a small variance from the significant comparison list. The problem of not controlling the FDR remains unless a much higher quantile of standard error is used as the fudge factor. However, that results in a great loss of power.

On the other hand, the approach based on permutation \( p \)-values adjusted by the BH-FDR procedure (indicated by 1,2,3,4 in Figure 7.4) seems to yields lower power as compared to the SAM without the fudge factor. From Figure 7.4b, we can see that the \( \text{FDR}^1 \) is slightly higher than 0.05 for the first two settings (indicated by 1 and 2), while it is getting close to 0.05 for the last two settings (indicated by 3 and 4). When the proportion of small variance gene increases, the problem with controlling the \( \text{FDR}^1 \) becomes more severe. Since the permutation approach adjusted by the BH-FDR procedure is not intended to protect against the small variance genes, it yields high value of the \( \text{FDR}^1 \), similarly to the SAM approach without the fudge factor.

In order to examine the variability in the estimated FDR and power, Figure 7.5 shows the boxplots of the FDR (panel a), \( \text{FDR}^0 \) (panel b), and \( \text{FDR}^1 \) (panel c) obtained for the SAM with and without the fudge factor and for permutation \( p\)-
Chapter 7. Choice of Fudge Factor: a Simulation Study

Figure 7.4: The setting of three arrays per group. a: FDR vs. power; b: FDR₀ (small variance genes) vs. power; c: FDR₁ vs. power. A, B, C, and D denote results for the SAM procedure with the fudge factor under the four settings of different proportions of the small variance genes. 1, 2, 3, and 4 denote the results for the permutation approach adjusted by the BH-FDR procedure under the four settings.

values with the BH-FDR adjustment. Figure 7.6 shows the boxplots of the power of the three approaches. The same conclusion can be drawn from these figures as from Figure 7.4, but additionally the distribution of the FDR can be examined. Note that the variability of the FDR and power for the SAM procedure with the fudge factor seems to be larger than for the other two procedures.
Figure 7.5: The setting of three arrays per group. a: Boxplots of the FDR for the SAM procedure without the fudge factor (-s0) and with the fudge factor (+s0), and for permutation p-values adjusted by the BH-FDR procedure; b: boxplots of the FDR0; c: boxplots of the FDR1.
Figure 7.6: The setting of three arrays per group. Boxplots of power for the SAM procedure without the fudge factor ($-s_0$) and with the fudge factor ($+s_0$), and for permutation p-values adjusted by BH-FDR procedure. Setting 1, 2, 3, and 4: $m^1_0 = 0, 850, 1700, \text{and } 3400$, respectively.
Figure 7.7: The setting of four arrays per group. a: FDR vs. power; b: FDR0 (small variance genes) vs. power; c: FDR1 vs. power. A, B, C, and D denote results for the SAM procedure with the fudge factor under the four settings of different proportions of the small variance genes. 1, 2, 3, and 4 denote the results for the permutation approach adjusted by the BH-FDR procedure under the four settings.

Similar conclusions can be drawn from the simulation studies using a larger sample size (four, five, or six arrays per group). As sample size (as well as the number of small variance genes) increases, the SAM procedure shows similar patterns for the case of three arrays per group. More importantly, the SAM procedure tends to maintain the FDR closer to the desired level as the sample size increases, at the price of losing power. Figure 7.7 shows the results for the FDR and power for the setting of four.
arrays per group.

Figure 7.8: The setting of four arrays per group. a: Boxplots of the FDR for the SAM procedure without the fudge factor (−s0) and with the fudge factor (+s0), and for permutation p-values adjusted by the BH-FDR procedure; b: boxplots of the FDR0 (small variance genes); c: boxplots of the FDR1.

The distribution of the FDR (including FDR0 and FDR1) and power is shown by the boxplots in Figure 7.8 and Figure 7.9, respectively. The power of both the SAM procedure without the fudge factor and permutation p-values adjusted by the
Figure 7.9: The setting of four arrays per group. Boxplots of power for the SAM procedure without the fudge factor (-s0) and with the fudge factor (+s0), and for permutation p-values adjusted by BH-FDR procedure. Setting 1, 2, 3, and 4: m1 = 0, 850, 1700, and 3400, respectively.

BH-FDR approach becomes similar as the sample size increases. Similar results were obtained for the simulation with five and six arrays per group and shown in the Appendix A.
7.3 Other Methods for Selecting the Fudge Factor

In this section, we consider other methods for the selection of the fudge factor. We focus on finding differentially expressed genes using microarray data from two groups/conditions, namely the control and the treatment group.

Let $X_{ijk}$ be the gene-expression of the gene for array $j$ in treatment group $k$, $i = 1,...,m, j = 1,...,n, k = 0,1$. Let $z_{ijk}$ be an indicator variable which take the value of 1 if gene $i$ in the microarray is obtained under the treatment and zero otherwise, i.e.,

$$
\begin{align*}
    z_{ijk} &= 0 & k &= 0 \text{ control,} \\
    z_{ijk} &= 1 & k &= 1 \text{ treatment.}
\end{align*}
$$

The gene-specific linear model is given by

$$
X_{ijk} = \beta_{0i} + \beta_{1i} z_{ijk} + \varepsilon_{ijk}; \quad \varepsilon_{ijk} \sim N(0, \sigma_i^2).
$$

(7.1)

Here, $\beta_{1i}$ is the effect of treatment on gene $i$, with $\beta_{0i}$ the corresponding to the mean of the control group for gene $i$. The difference between the two groups for gene $i$ can be tested by testing the null hypothesis of $\beta_{1i} = 0$. Using the least-squares fitting we obtain the following test statistic:

$$
\hat{t}_i = \frac{\hat{\beta}_{1i}}{s(\hat{\beta}_{1i})} = \frac{\bar{x}_{i0} - \bar{x}_{i1}}{\sqrt{\sum_{j=0}^{n_0}(x_{ij} - \bar{x}_{i0})^2 + \sum_{j=1}^{n_1}(x_{ij} - \bar{x}_{i1})^2}} \times \sqrt{\frac{n_0 + n_1 - 2}{n_0 + n_1}} \frac{n_0 n_1}{n_0 + n_1},
$$

(7.2)

where $n_0$ and $n_1$ are the number of arrays in the control and treatment group, respectively.

The test statistic in (7.2) is a two sample $t$-statistic, where the pooled variance estimator is defined as:

$$
\hat{s}_i^2 = \frac{1}{n_0 + n_1 - 2} \left\{ \sum_{j=1}^{n_0} (x_{ij0} - \bar{x}_{i0})^2 + \sum_{j=1}^{n_1} (x_{ij1} - \bar{x}_{i1})^2 \right\}.
$$

As we discussed in Chapter 5, the modification used by the SAM procedure is the introduction of the fudge factor to the denominator. The modified $t$ test statistic is given in (5.6). The fudge factor is chosen as a certain quantile of the standard error distribution for all the genes, which minimizes the CV of the SAM test statistics as discussed in Section 5.2.2 (Tusher et al. 2001).
7.3.1 Penalized Linear Regression

Wu (2005) proposed the penalized linear regression as a method for the detection of differentially expressed genes. It is defined as the result of the following optimization algorithm (Tibshirani 1996, Efron et al. 2004):

$$\min_{\beta_0, \beta_1} \sum_{k=0}^{1} (x_{ijk} - \beta_0 - \beta_1 z_{ijk})^2 + \varphi_1 |\beta_1|.$$ (7.3)

According to Wu (2005), the optimal parameters for the penalized model can be obtained as:

$$\hat{\beta}_1 = \text{sign}(\bar{x}_{i0} - \bar{x}_{i1})(|\bar{x}_{i0} - \bar{x}_{i1}| - \varphi)_+$$

where the subscript plus means positive part (\(z_+ = z\) if \(z > 0\) and zero otherwise), and \(\hat{\beta}_{0i} = \bar{x}_i - \frac{n_0}{n} \hat{\beta}_1\)

where \(\bar{x}_i = \sum_j x_{ij}/n,\) \(n = n_0 + n_1,\) and \(\varphi = (1/2)(1/n_0 + 1/n_1)\phi_1.\)

The penalized t-test statistic can be written as,

$$t^*_i = \text{sign}(\bar{x}_{i0} - \bar{x}_{i1})\frac{(|\bar{x}_{i0} - \bar{x}_{i1}| - \varphi)_+}{\sqrt{(1/n_0 + 1/n_1)s^2_i + \varphi^2/(n-2)}}$$ (7.4)

The shrinkage parameter \(\varphi\) is chosen to minimize the correlation between the penalized t-test statistics in (7.4) and \(s_i\) for all genes. It is easy to note that \(\varphi\) is applied to both the nominator and denominator and shrinks the absolute mean difference in the numerator by \(\varphi\). The denominator is shrunken by \(\varphi^2/(n_0 + n_1 - 2)\). Thus, the advantage of the penalized t-statistic is that it shrinks the absolute mean differences to zero, deflating large values of test statistics due to the small standard error.

Wu (2005) argued that for a given value of \(\varphi\), the penalized t-test statistic has the sam property as the SAM t-test statistic. Wu proposed to replace the SAM test statistic with the penalized t-test statistic and to control the FDR using the SAM procedure discussed in Section 5.2.2.

7.3.2 ROC Method

Broberg (2003) proposed a method to jointly optimize the number of genes that are falsely declared positive and the number of genes that are falsely declared negative
by optimizing over a range of FDR and fudge factor values. Broberg (2003) utilized
the receiver operating characteristic curve (ROC curve), which plots false positive
rate (FPR) against false negative rate (FNR). FPR is defined as the number of false
positives among the significant genes divided by the total number of genes. Similarly,
FNR is defined as the number of false negatives divided by the total number of genes.
The true-positive rate (TPR) is defined as the number of true positives divided by
the total number of genes. And the true-negative rate (TNR) is the number of true
negatives divided by the total number of genes.

For each combination of the FDR and of the fudge factor $s_0$, the procedure cal-
culates the number of false positives and false negatives. The criterion to be used for
the choice of the fudge factor is to minimize $C = \sqrt{\text{FPR}^2 + \text{FNR}^2}$. Due to the large
amount of calculations, the optimization of $C$ is performed over a lattice of (FDR,
$s_0$) values. Lovell et al. (1996) and Genovese and Wasserman (2001) present other
versions of the minimizing criterion. For example, one of the criteria incorporates
weight $\omega$ in order to express the relative importance of FPR compared to FNR, what
results in $C = \sqrt{\omega^2 \text{FPR}^2 + \text{FNR}^2}$.

**Estimation of Proportion of Truly Differentially/Non-differentially Ex-
pressed Genes**

In the ROC procedure, the $p$-values are first obtained using the permutation method
(5.2) described in Section 5.1.2. The estimated proportion $\hat{p}(\alpha)$ of genes considered
differentially expressed is defined as

$$
\hat{p}(\alpha) = \frac{\#\{i : p(i) \leq \alpha\}}{m}.
$$

Hence, $p(\alpha)$ is the relative frequency of genes that have a $p$-value less than $\alpha$.

The ROC method uses the following estimator for the proportion of truly non-
differentially expressed genes:

$$
\hat{p}_0 = \frac{\#\{i : q_{25} \leq t(i) \leq q_{75}\}}{m/2},
$$

where $q_{25}$ and $q_{75}$ are the 25% and 75% percentile of the values of the SAM test
statistics, respectively.
Estimation of the FPR and FNR

According to Storey (2001), the false positive probability estimator can be defined as \( \hat{\text{FPR}} = \hat{p}_0 \alpha \), which represents the proportion of non-differentially expressed genes multiplied by the probability that such a gene produces a significant result. The estimator of the false negative probability equals \( \hat{\text{FNR}} = 1 - \hat{p}_0 (1 - \alpha) - \hat{p}(\alpha) \).

Using the estimates \( \hat{\text{FNR}} \) and \( \hat{\text{FPR}} \), corresponding to the significance level \( \alpha \), we can obtain the fudge factor \( s_0 \) that minimizes \( C \). Then, the SAM procedure is carried out with the corresponding value of \( s_0 \).

### 7.3.3 The Empirical Bayes Approach

Efron et al. (2001) and Efron and Tibshirani (2002) model the distribution of the adjusted test statistics \( t_i \) in (5.6), \( i = 1, \ldots, m \), as a mixture of two components. One component corresponds to the differentially expressed genes, and the other to the non-differentially expressed genes. Denoting the density of the former by \( f_1 \) and the latter by \( f_0 \), the mixture density of the test statistics is given by

\[
 f(t) = \pi_0 f_0(t) + (1 - \pi_0) f_1(t),
\]

where \( \pi_0 \) is the prior probability that a gene is differentially expressed. Applying Bayes’ rule to (7.5) results in the posterior probability that a gene with test statistic \( t \) is differentially expressed, i.e.,

\[
 p_i(t) = 1 - \pi_0 \frac{f_0(t)}{f(t)}.
\]

Following Efron et al. (2001) and Efron and Tibshirani (2002), a gene is called differentially expressed if its posterior probability (7.6) is larger than or equal to 0.9. The FDR for the resulting rejection region \( \Gamma = \{ t : p_i(t) \leq 0.9 \} \) is then estimated using a resampling approach (see Section 5.1.2) by

\[
 \hat{\text{FDR}}(\Gamma) = \hat{\pi}_0 \frac{\# \{ t_i^b \in \Gamma \}/B}{\max \{ \# \{ t_i \in \Gamma \}, 1 \}},
\]

where \( B \) is total number of permutations.

In this empirical Bayes approach, both \( f \) and \( f_0 \) have to be estimated. Efron and Tibshirani (2002) estimate the density \( f \) of the observed test statistics by a Poisson regression with offset \( \log f_0 \), where \( f \) is modeled by a natural cubic spline with five degrees of freedom. Storey (2001) and Efron et al. (2001) proposed the following estimator for \( \pi_0 \):
Chapter 7. Choice of Fudge Factor: a Simulation Study

1. For $\tau = 0, 0.01, \ldots, 0.95$, compute $\hat{\pi}_0(\tau) = \#\{p_i(t) > \tau\} / ((1 - \tau)m)$.

2. Fit a natural cubic spline $h$ with three degrees of freedom through the data points $(\tau, \hat{\pi}_0(\tau))$, where each data point is weighed by $1 - \tau$.

3. Estimate $\pi_0$ by $\min\{h(1), 1\}$.

Efron et al. (2001) suggest to specify the optimal choice of the fudge factor by running the EB procedure for several values of $s_0$, and by selecting the value of $s_0$ that leads to the largest number of differentially expressed genes.

### 7.4 Simulation Study

In this section we investigate the performance of the fudge factor, obtained by using the penalized $t$-test (T), the ROC (R), and the EB (E), the SAM procedure (S) and permutation $p$-values adjusted by the BH-FDR procedure (B) using a simulation study.

#### 7.4.1 Simulation Setting

In this simulation, we use a slightly different setting by generating 10,000 genes on each array. In total, 100 data sets are generated with 16 arrays under two conditions (control and treatment with eight arrays each). We assume that 10% of the 10,000 genes are truly differentially expressed. The mean gene-expression for the two groups is defined as follows:

$$
\mu_{ik} = \begin{cases} 
\delta_{ik} \times \hat{\sigma}_i & i \leq 1000 \text{ and } k = 1, \\
0 & \text{otherwise.}
\end{cases}
$$

(7.7)

It is assumed that $\delta_{ik} \sim U(2.8, 4.5)$, while $\sigma_i$ is randomly sampled from the estimated standard error distribution observed in the case study analyzed in Section 6.3. Similarly as the setting in Section 7.2.3., the means of treatment groups for genes with small variance are generated from normal distribution $N(0, 0.3)$ for the four treatment groups. For the four settings, the number of small variance genes varies from zero, 500, 1000, and 2000. The variance of these genes is chosen to be equal to 0.0036, which is 5% quantile of the variances in the observed case study data and is expected to be small to inflate the test statistics.
7.4.2 Simulation Results

The FPR vs. FNR

First, we study the relationship between estimated mean FPR and FNR of the five methods under the four settings across 100 simulation datasets, where the FPR and FNR are defined in section 7.3.2. As we observe in Figure 7.10, the joint optimization of FPR and FNR is obtained for the SAM procedure, except that in the first setting (as with no small variance genes) the SAM and permutation approach show similar results. As the proportion of small variance genes increases (i.e., $m_0^1 = 500, 1000,$ and 2000 genes corresponding to 5%, 10%, and 20% of the total number of genes, respectively), the FPR and FNR rates increase. The permutation approach yields higher FPR as compared to the other procedures, since the procedure does not adjust the test statistic to protect against the small variance genes. On the other hand, the penalized $t$ approach tends to yield a higher FNR, because the penalized $t$-test statistic downplays the truly differentially expressed genes as well as non-differentially expressed genes with small variance. In-between these methods, coincidentally, the ROC and EB methods show similar results, but the difference between them seems to increase as the number of the small variance genes increases.

Proportion $P_i$ of non-differentially expressed genes with small variance declared significant vs. Power

Secondly, we study the relationship between the proportion $P_i$ of non-differentially expressed genes with small variance declared significant and power (Figure 7.11). Let $P_i$ defined as

$$P_i = 1 - V^1/m_0^1,$$

where $V^1$ is the number of genes falsely declared significant among non-differentially expressed genes with a small variance. The value of $P_i$ lies between 0 and 1. As $P_i$ gets closer to 1, a smaller proportion of small variance genes is declared as false positives. From Figure 7.11 we can see that as the number of the small variance genes increases (note that in the first setting no small variance genes are present, so $P_i = 1$), both the SAM and the penalized $t$ approaches control $P_i$ proportion better than the EB and ROC methods (between 0.18 and 0.3 under the four settings). However, the SAM and penalized $t$ approach differ in the way that the former increases $P_i$ at the price of losing power, while the latter decreases $P_i$ but gaining power. Among the
Figure 7.10: FPR rate vs. FNR rate using the five methods under the four settings. Black line: setting 1 \((m_0^1=0)\); red line: setting 2 \((m_0^1=500)\); green line: setting 3 \((m_0^1=1000)\); blue line: setting 4 \((m_0^1=2000)\). B: permutation p-values adjusted by the BH-FDR procedure, S: SAM, R: ROC, E: EB, and T: penalized t.

Five methods, the permutation approach (with no correction of the t-test statistics) is the worst in terms of controlling \(P_i\), but always maintains the power almost at 1.

In Figure 7.12 we compare the distribution of \(P_i\) values. We note that the variability of \(P_i\) values differs for different methods. For the SAM and ROC approaches, \(P_i\) value decreases as the number of small variance genes \((m_1^1)\) increases. The variability of \(P_i\) values for the SAM procedure decreases, while the variability of for the ROC
Figure 7.11: Proportion $Pi$ of non-differentially expressed genes with small variance declared significant vs. Power using the five methods under the four settings. Black line: setting 1 ($m_0^1=0$); red line: setting 2 ($m_0^1=500$); green line: setting 3 ($m_0^1=1000$); blue line: setting 4 ($m_0^1=2000$). B: permutation p-values adjusted by BH procedure, S: SAM, R: ROC, E: EB, and T: penalized t.

increases with the increasing number of small variance genes. For the penalized t the variability increases with increasing $m_0^1$. 

Figure 7.12: Boxplot of $\pi$ for the four simulated settings. Black: setting 1 ($m_1^0=0$); red: setting 2 ($m_1^0=500$); green: setting 3 ($m_1^0=1000$); blue: setting 4 ($m_1^0=2000$). S: SAM, B: permutation $p$-values adjusted by the BH-FDR procedure, E: empirical Bayes, R: ROC, and T: penalized $t$ approach.

FDR vs. Power

From Figure 7.13a we can see that for all the five methods the (overall) FDR is not well controlled at the desired level of 0.05, which is estimated as the mean proportion of false discoveries across 100 simulated datasets. The SAM procedure seems to yield the lowest FDR, while the ROC and EB approaches yield a similar, high value of the FDR. More importantly, taking SAM as a reference procedure, we can make two sets
of comparisons: (1) between the SAM, penalized $t$, and ROC methods that utilize the same approach to control the FDR, but differ in the choice of the fudge factor, (2) and between the SAM, EB, and permutation approaches, that differ not only in the choice of the fudge factor, but also the approach used to control the FDR.

For the first set, as $m_{01}$ increases, the FDR (Figure 7.13a) increases for the penalized $t$ (from 0.1, to 0.22 and to 0.52) and ROC methods (from 0.4, to 0.57, and to 0.7), and the power increases for the penalized $t$ approach (from 0.1, to 0.21, and to 0.4) and it is maintained for the ROC approach (about 0.6). The SAM procedure behaves differently: it maintains the FDR at around 0.2 at the price of losing the power (from 0.9, to 0.7, and to 0.5). Using the decomposition of FDR into the FDR$^0$ (shown in Figure 7.13b) and FDR$^1$ (presented in Figure 7.13c), we observe different patterns for the FDR$^0$ for the three methods. For the small variance genes, the FDR$^0$ is controlled at 0.05 for SAM and ROC methods under the four settings, but not for the penalized $t$, for which it increases as the number of small variance genes increases. The penalized $t$-test statistic is modified by adding a constant to both the denominator and numerator, which results in declaring more positive for non-differentially expressed genes with a moderate or large variance as significant. On the other hand, the penalized $t$ approach reduces the FDR$^1$, in particular when 5% or 10% non-differentially genes with small variance are present. However, it does not do so for a larger proportion of such genes (i.e., 20%). In all, we can conclude that the SAM procedure is most successful in controlling the overall FDR at the level about 0.2, as compared to other approaches, it does not control the FDR at the desired level of 0.05 though. The main component of FDR for the SAM and ROC approaches is FDR$^1$. The penalized $t$ method has a tendency to control the FDR$^1$, but it loses the control of the FDR$^0$. We also note that the variability of the FDR and FDR$^0$ for the penalized $t$ approach is large as shown in Figures 7.14a and 7.14b.

For the second set of comparisons, from Figure 7.13a we can observe that the FDR for the three methods is not controlled at the desired level of 0.05. The EB method yields the largest FDR among the three approaches and a constant power (around 60%). The SAM approach shows a decreasing FDR (0.25, 0.24, and 0.18, respectively) and decreasing power (90%, 70%, and 50%, respectively) as the proportion of non-differentially expressed genes with small variance increases (from 5% to 10%, and to 20%). For the permutation approach, the FDR (0.35, 0.5, and 0.7) increases, but the power is consistently close to 1. In Figure 7.13b, the FDR$^0$ for the three methods is around 0.05. This confirms the results from the simulation study described in Section
7.2. Hence, the \( \text{FDR}^1 \) shows the same pattern as the overall FDR.

\[ \begin{align*}
\text{a. FDR vs. Power} & \quad \text{b. FDR}^0 \text{ vs. Power} \\
\text{c. FDR}^1 \text{ vs. Power} \end{align*} \]

**Figure 7.13**: a. FDR vs. power, b. FDR\(^0\) vs. power, c. FDR\(^1\) vs. power under the four settings. Black line: setting 1 \((m_0^1=0)\); red line: setting 2 \((m_0^1=500)\); green line: setting 3 \((m_0^1=1000)\); blue line: setting 4 \((m_0^1=2000)\). S: SAM, B: permutation \(p\)-values adjusted BH-FDR, E: empirical Bayes, R: ROC, and T: penalized t approach.

Figure 7.14 shows the power of the five methods. We can conclude that the EB and ROC methods yield not only a similar power, but also a similar FDR. Both methods produce rather consistent values of the FDR and power regardless of the proportion of non-differentially expressed genes with small variance. The permutation approach performs slightly better than the EB and ROC methods, yielding a lower FDR and
a higher power. The worst performance with respect to the power is observed for the penalized \( t \) approach. This method tends to control FDR\(^1\), but to lose control of FDR\(^0\). Finally, for SAM, the conclusion is similar as in the simulation study in Section 7.3. It can be regarded as the best method to choose the fudge factor for controlling FDR, but at the cost of losing power.

Figure 7.14: a. Boxplot of FDR, b. Boxplot of FDR\(^0\), c. Boxplot of FDR\(^1\), d. Boxplot of power under the four settings. Black line: setting 1 \((m^1_0=0)\); red line: setting 2 \((m^1_0=500)\); green line: setting 3 \((m^1_0=1000)\); blue line: setting 4 \((m^1_0=2000)\). S: SAM, B: permutation p-values adjusted BH-FDR, E: empirical Bayes, R: ROC, and T: penalized \( t \) approach.
Chapter 7. Choice of Fudge Factor: a Simulation Study

7.5 Discussion

The question of interest in this chapter is the capability of the SAM procedure to control the FDR for truly non-differentially expressed genes with small variance. We have shown in the first simulation study that the overall FDR cannot be controlled even when the proportion of truly non-differentially expressed genes with a small variance is relatively small (i.e., 5% of the total number of genes in the array). Moreover, we have shown that when the FDR is decomposed into the FDR$^0$ and FDR$^1$, there is no problem to control the FDR$^0$ (regardless the proportion of non-differentially expressed genes with small variance). However, the FDR$^1$, associated with the small variance genes, is not well controlled. When the proportion of non-differentially expressed genes with a small variance increases, the SAM procedure with the fudge factor either loses power or does not control the FDR at the desired level, as compared to the SAM procedure without the fudge factor. Thus, the automatic selection of the fudge factor does not guarantee the power and the control of the FDR in the SAM approach.

We have also investigated several methods for selecting the fudge factor. We have shown that none of these methods controls the FDR at the desired level. The SAM procedure is most successful in controlling the FDR but at the price of losing power. The penalized $t$ approach shows poor performance in the control of both the FDR and power, with a poor control over truly non-differentially expressed genes with a moderate or large variance. Coincidentally, the EB and ROC approaches, which differ in the methods used to control the FDR and to choose the fudge factor, yield similar and consistent results with respect to the power and FDR, regardless of the proportion of non-differentially expressed genes with a small variance.

Note that the proportion of non-differentially expressed genes with a small variance ($m_1$) is unknown in the case study. Mixture modelling approach (Delmar et al. 2003, Manda et al. 2007) have been proposed to estimate this proportion. None of the methods considered in the second simulation study takes this proportion into consideration. The possibility of choosing the fudge factor with the information of the proportion of small variance genes and of modifying the test statistics for genes with different variances. The aim of such an approach would be to achieve a good balance between the control of the FDR and power.

The contents of this chapter were summarized in the manuscript (Lin et al. 2008).
Part III

Dose-response Modelling of Microarray Data in Drug Development Experiments
Chapter 8

Testing for Trend in Dose-response Microarray Experiments: a Comparison of Several Testing Procedures

8.1 Introduction

Investigation of a dose-response relationship is of primary interest in many drug-development studies. Typically, in dose-response experiments the outcome of interest is measured at several (increasing) dose levels, and the aim of the analysis is to establish the form of the dependence of the response on dose (Agresti 1997). The response can be either the efficacy of a treatment or the risk associated with the exposure to the treatment (in toxicology studies). In a typical dose-response study subjects are randomized to several dose groups, among which there is usually a control group. Ruberg (1995a, 1995b) and Chuang-Stein and Agresti (1997) formulated four main questions usually asked in dose-response studies: (1) Is there any evidence of the drug effect? (2) For which doses is the response different from the response in the control group? (3) What is the nature of the dose-response relationship? and (4) What is the optimal dose?
Within the microarray setting, a dose-response experiment has the same structure as described above. The response is the gene-expression at a certain dose level. The dose-response curve, similarly to the dose-response studies, is assumed to be monotone, i.e., the gene activity increases or decreases as the dose level increases. The direction of the relationship is usually unknown in advance.

In this chapter we focus on the first question: is there any evidence of the drug effect? To answer this question, we test for the null hypothesis of homogeneity of means (no dose effect) against an ordered alternative. We compare several testing procedures, that take into account the order restriction of the means with respect to the increasing doses and that adjust for multiple testing. In particular, we discuss the testing procedures of Williams (Williams 1971 and 1972), Marcus (Marcus 1976), the global likelihood ratio test (LRT, Barlow et al. 1972, and Robertson et al. 1988), and the $M$ (Hu et al. 2005) statistic. Moreover, we propose a novel procedure based on a modification of the estimator of standard error of the $M$ statistic.

Williams (1971, 1972) proposed a step-down procedure to test for the dose effect. The tests are performed sequentially from the comparison between the isotonic mean of the highest dose and the sample mean of the control, to the comparison between the isotonic mean of the lowest dose and the sample mean of the control. The procedure stops at the dose level where the null hypothesis (of no dose effect) is not rejected. Marcus (1976) proposed a modification of the Williams procedure, in which the sample mean of the control was replaced by the isotonic mean of the control. A global likelihood ratio test, discussed by Bartholomew et al. (1961), Barlow et al. (1972), and Robertson et al. (1988), uses the ratio between the variance calculated under the null hypothesis and the variance calculated under an ordered alternative. Recently, Hu et al. (2005) proposed a test statistic that was similar to Marcus’ statistic, but with the variance estimator calculated under the ordered alternative. The degrees of freedom of the $M$ statistic (the difference between the number of observations and the number of dose levels) were fixed for all the genes and all the arrays. We propose a modification for the variance estimator of the $M$ statistic. Namely, the difference between the number of observations and the unique number of isotonic means is used as the degrees of freedom for the variance estimator.

Our goal is to compare the performance of the five test statistics. To this aim we apply them to the human epidermal squamous carcinoma cell line data (Section 1.4.1), in which gene-expression was measured under four EGF doses (three higher doses and one control) for the control compound. The microarray experiment for
the control compound consists of 12 arrays, each containing 16,998 genes, with three arrays available for each of four EGF concentration levels. When applied to the case study, the five test statistics are adjusted for multiple testing by using resampling-based procedures that control either the FWER or the FDR. Following the results of the analysis of the case study, we conduct a simulation study to further investigate the performance of the five test statistics.

The chapter is organized as follows. In Section 8.2 we review the five test statistics. Directional inference to testing the isotonic regression and the multiplicity issue are discussed in Section 8.3. In Section 8.4, we compare the results of the analysis of the case study using the five tests, discussed in Section 8.2. A simulation study conducted to investigate the performance of variance estimators and the power of the five test statistics is presented in Section 8.5. Section 8.6 completes the chapter with a short discussion.

### 8.2 Testing for Homogeneity of the Means Under Restricted Alternatives

In this section, we review several procedures for testing the homogeneity of the means against order restricted alternatives. In particular we focus on four existing procedures: Williams’ (Williams 1971 and 1972), Marcus’ (Marcus 1976), the global likelihood ratio test (Bartholomew 1961, Barlow et al. 1972, and Robertson et al. 1988), and the $M$ (Hu et al. 2005) statistic. Additionally, we introduce a modification to the degrees of freedom of the $M$ statistic.

In the microarray experiment, for each gene, the following ANOVA model is considered:

$$Y_{ij} = \mu(d_i) + \varepsilon_{ij}, \ i = 0, 1, \ldots, K, \ j = 1, 2, \ldots, n_i, \ (8.1)$$

where $Y_{ij}$ is the $j$th gene-expression at the $i$th dose level, $d_i$ ($i = 0, 1, \ldots, K$) are the $K+1$ dose levels, $\mu(d_i)$ is the mean gene-expression at each dose level, and $\varepsilon_{ij} \sim N(0, \sigma^2)$.

The null hypothesis of no dose effect is given by

$$H_0 : \mu(d_0) = \mu(d_1) = \cdots = \mu(d_K). \ (8.2)$$

A one-sided alternative hypothesis of a positive dose effect for at least one dose level
Testing for Trend in Dose-response Microarray Experiments

(i.e., an increasing trend) is specified by

\[ H_{\text{Up}}^1 : \mu(d_0) \leq \mu(d_1) \leq \cdots \leq \mu(d_K), \]  

(8.3)

with at least one strict inequality. When testing the effect of a drug for a positive outcome the researcher can specify a positive effect as the desirable alternative. However, in the current microarray setting, it seems reasonable to assume that the gene-expression levels may increase or decrease in response to increasing doses, but with the direction of the trend not known in advance. Thus, we must also consider an additional alternative:

\[ H_{\text{Down}}^1 : \mu(d_0) \geq \mu(d_1) \geq \cdots \geq \mu(d_K), \]  

(8.4)

with at least one strict inequality. Testing \( H_0 \) against \( H_{\text{Down}}^1 \) or \( H_{\text{Up}}^1 \) requires estimation of the means under both the null and the alternative hypotheses. Under the null hypothesis, the estimator for the mean response \( \hat{\mu} \) is the sample mean. Let \( \hat{\mu}_0, \hat{\mu}_1, \ldots, \hat{\mu}_K \) be the maximum likelihood estimates for the means (at each dose level) under the ordered alternative. Barlow \textit{et al.} (1972) and Robertson \textit{et al.} (1998) showed that \( \hat{\mu}_0, \hat{\mu}_1, \ldots, \hat{\mu}_K \) are given by the isotonic regression of the observed means.

8.2.1 Williams' (1971, 1972) and Marcus' (1976) Test Statistics

Williams’ procedure defines \( H_0 \) as the null hypothesis, and \( H_{\text{Up}}^1 \) or \( H_{\text{Down}}^1 \) as the one-sided alternative. Williams’ (1971, 1972) test statistic was suggested for a setting, in which \( n_i \) observations are available at each dose level. As all dose levels are compared with the control level, the test statistic is given by

\[ t_i = \frac{\hat{\mu}_i^* - \bar{y}_0}{\sqrt{2s^2/r}} \]  

(8.5)

Here, \( \bar{y}_0 \) is the sample mean at the first dose level (control), \( \hat{\mu}_i^* \) is the estimate for the mean at the \( i \)th dose level under the ordered alternative, \( r \) is the number of replications at each dose level, and \( s^2 \) is an estimate of the variance. For \( \hat{\mu}_i^* \), Williams (1971, 1972) used the isotonic regression of the observed response with respect to dose (Barlow \textit{et al.} 1972). Williams’ test procedure is a sequential procedure. In the first step, \( \hat{\mu}_K^* \) is compared to \( \bar{y}_0 \). If the null hypothesis is rejected, \( \hat{\mu}_{K-1}^* \) is compared to \( \bar{y}_0 \), etc.
Marcus (1976) proposed a modification to Williams’ test statistic that replaced \( \bar{y}_0 \) with \( \hat{\mu}^* \), the estimate of the first dose (control) mean under ordered restriction. Marcus’ test statistic performs closely to Williams’ in terms of power (Marcus 1976). Note that, for \( K = 1 \), Williams’ and Marcus’ test statistics reduce to the two-sample t-test.

### 8.2.2 Likelihood Ratio Test Statistic for Monotonicity

*(Barlow et al. 1972, and Robertson et al. 1988)*

Williams’ and Marcus’ procedures are step-down procedures, i.e., the comparison between a lower dose and control is tested only if the test of a higher dose vs. control is significant. The underlying assumption is that there is a monotone dose-response relationship with a known direction.

Testing the equality of ordered means using likelihood ratio tests (when response is assumed to be normally distributed) was discussed by Barlow et al. (1972) and Robertson et al. (1988). Both authors considered the likelihood ratio test, in which the variance under the null and the alternative were compared. The likelihood ratio test statistic is given by

\[
\Lambda_{01}^2 = \frac{\hat{\sigma}_H^2}{\hat{\sigma}_0^2} = \frac{\sum_{ij}(y_{ij} - \hat{\mu}^*_j)^2}{\sum_{ij}(y_{ij} - \bar{y})^2},
\]

where \( \hat{\sigma}_H^2 \) and \( \hat{\sigma}_0^2 \) are the estimates for the variance under the null and the alternative hypothesis, respectively. And \( \hat{\mu} = \frac{\sum_{ij} y_{ij}}{\sum_i n_i} \) is the overall mean. The null hypothesis is rejected for a “small” value of \( \Lambda_{01}^2 \). Equivalently, \( H_0 \) is rejected for large value of \( E_{01}^2 \), where

\[
E_{01}^2 = 1 - \Lambda_{01}^2 = \frac{\sum_{ij}(y_{ij} - \bar{y})^2 - \sum_{ij}(y_{ij} - \hat{\mu}^*_j)^2}{\sum_{ij}(y_{ij} - \bar{y})^2}.
\]

Estimating the parameters using isotonic regression requires the knowledge of the direction of the trend. In practice, the direction of the trend is often not known in advance. In such a case one can maximize the likelihood twice: for a monotone decreasing trend and for a monotone increasing trend, and choose the trend with a higher likelihood. In practice, we can calculate \( E_{01}^2 \) for each direction and choose the higher value of \( E_{01}^2 \) (Barlow et al. 1972). A resampling-based approach, as described in Section 5.1.2, can be used to approximate the null distribution for the test statistic, so that two-sided \( p \)-values are obtained for inference.
8.2.3 The $M$ Test Statistic of Hu et al. (2005)

Recently, Hu et al. (2005) proposed the following test statistic $M$ to test for a monotonic trend:

$$M = \frac{\hat{\mu}_K^* - \hat{\mu}_0^*}{\sqrt{\sum_{i=0}^{K} \sum_{j=1}^{n_i} (y_{ij} - \hat{\mu}_i^*)^2 / (n - K)}}. \quad (8.8)$$

where $n$ is the total number of arrays.

Hu et al. (2005) discussed a setting, in which the comparison of primary interest is the difference between the highest dose level ($K$) and the control dose. The numerator of the $M$ test statistic is the same as that of Marcus’ statistic, while the denominator is an estimate of the standard error under an ordered alternative. This is in contrast to Williams’ and Marcus’ approaches that use the unrestricted means to derive the estimate for the standard error.

Hu et al. (2005) evaluated the performance of the $E_{01}$ and $M$ test statistics by comparing the ranks of genes obtained by using both statistics, and reported similar findings for simulated and real-life data sets.

8.2.4 A Modification to the $M$ Test Statistic

For the variance estimate, Hu et al. (2005) used $n - K$ degrees of freedom (see equation (8.8)). However, the unique number of isotonic means is not fixed, but changes across the genes. For that reason, we propose a modification to the standard error estimator used in the $M$ statistic by replacing it with $\sqrt{\sum_{i=0}^{K} \sum_{j=1}^{n_i} (y_{ij} - \hat{\mu}_i^*)^2 / (n - I)}$, where $I$ is the unique number of isotonic means for a given gene. Such a modification is expected to improve the standard error estimates across all the genes.

The five test statistics are implemented in the R IsoGene package, which is discussed in detail in Chapter 9.

8.3 Directional Inference

8.3.1 Directional Inference in Isotonic Regression

The five test statistics discussed in Section 8.2 should be calculated assuming a particular direction of the ordered alternative. However, the direction of the test is unknown in advance. In this section, we address the issue of how to obtain the two-sided $p$-value
from the five testing procedures, and how to determine the direction of the trend from two-sided p-value afterwards.

We focus on the two possible directions of the alternatives: $H_{1U}$ defined in equation (3) and $H_{1Down}$ defined in equation (4). Let $p^{Up}$ and $T^{Up}$ denote the p-value and the corresponding test statistic computed to test $H_0$ vs. $H_{1U}$, and let $p^{Down}$ and $T^{Down}$ denote the p-value and the corresponding test statistic computed to test $H_0$ vs. $H_{1Down}$. Barlow et al. (1972) showed that, for $K > 2$, a χ² statistic for testing $H_0$ may actually yield $p^{Up} < \alpha$ and $p^{Down} < \alpha$. However, $p = 2 \min(p^{Up}, p^{Down})$ is always a conservative p-value for the two-sided test of $H_0$ vs. either $H_{1U}$ or $H_{1Down}$.

Hu et al. (2005) adapted the approach by taking the larger of the likelihoods of $H_{1U}$ or $H_{1Down}$, i.e., the larger of $T^{Up}$ and $T^{Down}$ is used as the test statistic for the two-sided inference. In contrast to Hu et al. (2005), we obtain two-sided p-values by taking $p = \min(2 \min(p^{Up}, p^{Down}), 1)$, where $p^{Up}$ and $p^{Down}$ are calculated for $T^{Up}$ and $T^{Down}$ using permutations to approximate the null distribution of these test statistics. We use $p^{Up}$ and $p^{Down}$ to determine the direction of the trend, as described below.

After rejecting the null hypothesis against the two-sided test there is still a need to determine the direction of the trend. The direction can be inferred by the following procedure. If $p^{Up} \leq \alpha/2$, then reject $H_0$ and declare $H_{1U}$; if $p^{Down} \leq \alpha/2$, then reject $H_0$ and declare $H_{1Down}$. The validity of this directional inference is based on the following property: under $H_{1U}$, $p^{Down}$ is stochastically larger than $U[0,1]$; and under $H_{1Down}$, $p^{Up}$ is stochastically larger than $U[0,1]$ (proof not given here). Thus, the probability of falsely rejecting $H_0$ is $\leq \alpha$, and the probability of declaring a wrong direction for the trend is $\leq \alpha/2$. It is also important to note that the event $p^{Up} < \alpha/2$ and $p^{Down} < \alpha/2$ may be observed. Under $H_0, H_{1U}$, or $H_{1Down}$, this event is unlikely. However, it is likely if the treatment has a large and non-monotone effect. An example of this unique situation, in which the null hypothesis can be rejected for both directions, is given in Section 8.4.1.

In order to illustrate whether the property needed for directional inference applies to the five test statistics, we conduct a simulation study to investigate the distribution of the $p^{Up}$ and $p^{Down}$ values. For each simulation, data are generated under $H_{1U}$: the means are assumed to be equal to $(1, 2, 3, 4)/\sqrt{5}$ for the four doses, respectively, and the variance is equal to $\sigma^2 = 1$. The test statistics $T^{Up}$ and $T^{Down}$ are calculated for the two possible alternatives $H_{1U}$ and $H_{1Down}$. Their corresponding $p^{Up}$- and $p^{Down}$-values are obtained using 10,000 permutations.
Figure 8.1 shows the cumulative distribution of $p_{Up}$ and $p_{Down}$. Clearly, the simulations show that the cumulative distribution of $p_{Down}$ (the $p$-value of the test statistics calculated assuming the wrong direction, dotted line in Figure 8.1) is stochastically higher than $U[0,1]$ (solid line in Figure 8.1), which is the distribution of the $p$-values under the null hypothesis. Moreover, the distribution of $p_{Up}$ (the $p$-value for the test statistics calculated assuming the right direction, dashed line in Figure 8.1) is, as expected, stochastically smaller than $U([0,1])$. Similar results (not shown) are obtained when the data are generated under $H_{Down}$. The results imply that all the five test statistics possess the property required for the directional inference: under $H_{Up}$, the distribution of $p_{Down}$ is stochastically greater than $U[0,1]$.

Figure 8.2 shows the values of test statistics, which were calculated under $H_{Up}$ and $H_{Down}$, for data generated under $H_{Up}$. The five test statistics are calculated for testing $H_0$ vs. $H_{Down}$ (the x-axis of each test statistic in Figure 8.2). The behavior of Marcus’, $M$, and the modified $M$ statistics is similar as they all use the difference between the highest and the lowest isotonic mean. The maximum value of the test statistics (when calculated assuming the wrong direction) is equal to zero. In contrast, Williams’ test statistic for testing $H_0$ vs. $H_{Down}$ (shown on the x-axis of the panel b) can be positive or negative, because the sample mean of control group is used instead of the isotonic mean. Note that we reject the null hypothesis in favor of $H_{Down}^+$ for negative values of the test statistic. Further, the value of the test statistics for testing $H_0$ vs. $H_{Up}$ (the y-axis of Figure 8.2) is higher than the value of the test statistics calculated for testing $H_0$ vs. $H_{Down}^+$ (the x-axis of Figure 8.2).

### 8.3.2 Control of the Directional FDR

When the FDR controlling procedures are used to adjust for multiple testing in the microarray setting, the set of two-sided $p$-values computed for each gene is adjusted by using the BH-FDR or BY-FDR procedure described in Section 5.2.2. A discovery in this case is a rejection of $H_0$ for some gene; a false discovery is to reject $H_0$ when $H_0$ is true. As mentioned before, in a microarray dose-response experiment we are also interested in the direction of the dose-response trend.

Benjamini and Yekutieli (2005) provide a framework for addressing the multiplicity problem when attempting to determine the direction of multiple parameters: a discovery is to declare the sign of a parameter as either being positive or negative. Three types of false discoveries are possible: declaring a zero parameter either as
negative or as positive, declaring a negative parameter as positive, and declaring a positive parameter as negative. The FDR corresponding to these discoveries is termed the Mixed Directional FDR (MD-FDR). In the current setting, the MD-FDR is the expected value of the number of genes, for which \( H_0 \) is true, that are erroneously declared to have either a positive or negative trend plus the genes with a monotone trend but with a wrong direction of the declared trend, divided by the total number
of genes declared to have a trend. Benjamini and Yekutieli (2005) prove that if \( p \)-values pose the directional property described in Section 8.3.1, then applying the BH procedure at level \( q \) to the the set of two-sided \( p \)-values computed for each gene, and declaring the direction of the trend corresponding to the smaller one-sided \( p \)-value, controls the MD-FDR at level \( q/2 \cdot (1 + m_0/m) \), where \( m \) is the total number of genes and \( m_0 \) is the number of genes, for which \( H_0 \) holds.

In general, directional inference is a more general setting than hypotheses testing (Benjamini and Yekutieli, 2005). Nevertheless, as a false discovery is made based
on the $p$-value that is stochastically larger than $U[0, 1]$, then the resampling-based methods that control the FDR (Yekutieli and Benjamini, 1999) also control the MD-FDR. This is achieved by simply applying the resampling-based procedure to test $H_0$, and if $H_0$ is rejected, declaring the direction of the trend according to the minimum one-sided $p$-value. For each rejected null hypothesis it is also advisable to examine if the larger $p$-value is $\leq \alpha$. If this is the case, this may serve as an indication of a non-monotone dose-response relationship.

8.4 Results

In this section, we present results of an application of the five testing procedures to the case study. We compare the performance of each of five test statistics in combination with the Bonferroni, Holm, maxT, and FDR-BH multiple-testing adjustment procedures. In Section 8.4.1 we examine the number of significant genes for all the testing procedures. In Section 8.4.2 we make a comparison between the global likelihood ratio test $\bar{E}_{01}$ and the two t-test type statistics: $M$ and the modified $M$.

8.4.1 Number of Significant Findings for Each Statistic Using Different Multiple Testing Adjustments

The testing procedures discussed in the previous sections are applied to the case study data. For each test statistic, $p^{Up}$ and $p^{Down}$ are obtained based on the permutation matrix, in which the null distribution of the test statistics ($T^{Up}$ and $T^{Down}$, respectively) are approximated using 1000 permutations. The inference is made based on the two-sided $p$-values obtained using the method described in the simulation study of Section 8.3.1.

Table 8.1 shows the number of rejected hypotheses for several multiplicity adjustment methods and for the five test statistics that are tested at the significance level of 0.05. Figure 8.3 shows the adjusted $p$-values for the five test statistics. Clearly, the adjusted $p$-values for the maxT, Bonferroni, and BY-FDR procedure are larger than the adjusted $p$-values obtained for the BH-FDR procedure. For instance, for $\bar{E}_{01}$, without adjusting for multiple testing, we reject the null hypothesis for 5457 genes. With the Bonferroni, Holm, and BH-FDR adjustment procedures we obtain the same number of significant genes, i.e., 1814. Using the maxT method for controlling the FWER obtains the least number of discoveries with only 224 genes declared
significant.

Table 8.1: Number of rejected null hypotheses for various testing procedures at the significance level of 0.05.

<table>
<thead>
<tr>
<th>Method</th>
<th>$E_{01}^2$</th>
<th>Williams</th>
<th>Marcus</th>
<th>$M$</th>
<th>Modified $M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unadjusted</td>
<td>5457</td>
<td>5238</td>
<td>5465</td>
<td>5449</td>
<td>5451</td>
</tr>
<tr>
<td>maxT</td>
<td>224</td>
<td>215</td>
<td>223</td>
<td>265</td>
<td>251</td>
</tr>
<tr>
<td>Bonferroni</td>
<td>1814</td>
<td>1592</td>
<td>1669</td>
<td>1755</td>
<td>1745</td>
</tr>
<tr>
<td>Holm</td>
<td>1814</td>
<td>1592</td>
<td>1669</td>
<td>1755</td>
<td>1745</td>
</tr>
<tr>
<td>BH-FDR</td>
<td>3613</td>
<td>3209</td>
<td>3533</td>
<td>3562</td>
<td>3567</td>
</tr>
<tr>
<td>BY-FDR</td>
<td>1814</td>
<td>1592</td>
<td>1669</td>
<td>1755</td>
<td>1745</td>
</tr>
</tbody>
</table>

Note that the number of significant genes obtained for each test statistic for a given multiple testing adjustment is similar. For example, for the BH-FDR adjustment, we find 3613, 3562, and 3567 significant genes for $E_{01}^2$, $M$, and the modified $M$ statistic, respectively. This method yields more liberal results as compared to the other multiple testing adjustment procedures. For that reason, the FDR adjustment for multiplicity is commonly used within the microarray framework (Ge et al. 2003, Tusher et al. 2001, Storey and Tibshirani 2003). Moreover, the BH-FDR procedure controls for the MD-FDR (as discussed in Section 8.3.2). Therefore, in what follows, we use the BH-FDR procedure to investigate the performance of the considered test statistics.

As we argue in Section 8.3.1, there is a possibility (although unlikely), that the null hypothesis is rejected for both directions (i.e., $p_{Up} \leq \alpha/2$ and $p_{Down} \leq \alpha/2$). For the analysis discussed above, the null hypothesis is rejected for only five genes when using Marcus’ statistic with $p_{Up}$ and $p_{Down}$ smaller than the rejection threshold (with multiple testing adjustment), suggesting a non-monotonic trend. The five genes are shown in Figure 8.4.

As can be observed from Figure 8.4, for the five genes the data reveal a non-monotonic pattern. For Marcus’ statistic, the large values of $T_{Up}$ and $T_{Down}$ are obtained from the large difference between the isotonic mean of the highest and control doses, relative to the variance calculated under the unrestricted alternative. Instead, $E_{01}^2$, $M$, and the modified $M$ use the variance estimator calculated under the ordered alternative, that results in smaller test statistic values. Hence, using these test statis-
Figure 8.3: Adjusted $p$-values using the Bonferroni, BH(FDR) and maxT procedures for the five test statistics.

In particular, for the five genes, the estimates of $\sigma^2$ (for Williams’ and Marcus’ test statistic) calculated under the unrestricted alternative are equal, respectively, to 0.0414, 0.0075, 0.0204, 0.0145, and 0.0232. They are smaller than the estimates of $\sigma^2$ for $E^2_{01}$, $M$, and the modified $M$ procedures calculated under the ordered alternative $H^U_{1p}$, that are equal, respectively, to 0.2995, 0.1788, 0.3277, 0.3317, and 0.2437, and
Figure 8.4: Five genes rejected by Marcus’ statistics with both $p_{\text{Up}}$ and $p_{\text{Down}}$ values smaller than the rejection threshold. Solid line: the isotonic means obtained for testing $H_0$ against $H_{1\text{Up}}$. Dashed line: the isotonic means obtained for testing testing $H_0$ against $H_{1\text{Down}}$. Under $H_{1\text{Down}}$, that are equal, respectively, to 0.2608, 0.1868, 0.4679, 0.4401, and 0.2065.
8.4.2 Comparison Between $\bar{E}_{01}^2$, the $M$, and Modified $M$ Test Statistics

Although in our case study, the number of significant genes obtained for the five testing procedures is very similar, there are some discrepancies. In this section, we investigate the subset of genes not commonly found by $\bar{E}_{01}^2$, $M$, and the modified $M$ statistics, respectively.

First we compare genes identified as significant or non-significant by $M$ and $\bar{E}_{01}^2$. The logarithm of two-sided $p$-values for these genes is shown in Figure 8.5. Among the total of 16,998 genes, 3420 genes are found significant for monotonic trends for both statistics. However, 193 genes are found to be significant for $\bar{E}_{01}^2$ and non-significant for $M$-test statistic, while for 142 genes the reversed order is observed. These genes account for 8.9% (193 + 142)/(3420 + 193 + 142) of the total significant findings for both test statistics, which is not negligible.

Similar to Hu et al. (2005), we compare the ranking of $M$ and $\bar{E}_{01}^2$ of all the genes. In both Hu et al. (2005) and our example the correlation of the ranks is equal to 0.99. Based on this observation, Hu et al. (2005) concluded that the two statistics perform similarly. However, in our data, the correlation of ranks of 142 genes found significant only for the $M$ statistic (panel c of Figure 8.6) is 0.92, while the correlation of ranks of 193 genes significant only for $\bar{E}_{01}^2$ (panel b) is 0.85. Both are somewhat lower than the correlation for genes in panel a (3420 genes significant for both statistics, correlation of 0.98) and in panel d (genes non-significant by either statistic, correlation of 0.99). The discrepant conclusions (rejecting the null only for one of statistics) can be explained by the fact that the $M$ statistic looks for the mean difference between the highest dose and the control. On the other hand, $\bar{E}_{01}^2$ is a global test for the monotonic trend.

The logarithm of the two sided $p$-values for the genes identified as significant or non-significant by the $M$ and modified $M$ statistics is shown in Figure 8.7. Among the total of 16,998 genes, 3478 genes are found significant for monotonic trends by both tests. However, 86 genes are found to be significant by the $M$ statistic and non-significant by the modified $M$ test, while for 89 genes the reverse is true. These genes account for about 4.8% (86 + 89)/(86 + 89 + 3478) of the total significant findings for both test statistics.

The overall correlation between the ranks of genes obtained for $M$ and the modified $M$ test statistics is 0.99. The correlation between genes in each panel of Figure 8.8
Figure 8.5: Logarithm of p-values (two-sided) for $E^2_{01}$ and $M$. Panel a: 3420 genes declared significant by both $E^2_{01}$ and $M$ statistics; panel b: 142 genes are declared significant by $M$ statistic only; panel c: 193 genes in are declared significant from $E^2_{01}$ only; panel d: 13,244 genes are not declared significant by either statistic.

is also very high, with 0.97 (in panel b) for genes declared significant only by the modified $M$, 0.98 (in panel c) for genes declared significant only by $M$, 0.99 (in panel a) for genes declared significant by both of the test statistics, and 0.998 (in panel c) for genes declared significant by neither of the test statistics. The difference between the two statistics lies in the adjustment of the degrees of freedom in the standard error estimator of the modified $M$ test statistic. Nevertheless, the discrepancy found
Figure 8.6: Correlation between $E_{01}^2$ and $M$. Panel a: correlation (0.98) between rankings of 3420 genes declared significant both from $E_{01}^2$ and $M$. Panel b: correlation (0.92) between rankings of 142 genes declared significant only from $M$. Panel c: correlation (0.85) between rankings of 193 genes declared significant only from $E_{01}^2$ and $M$. Panel d: correlation (0.99) between rankings of 13,244 genes not declared significant from $E_{01}^2$ and $M$.

is not substantial.
Figure 8.7: Logarithm of $p$-values (two-sided) for the $M$ and the modified $M$. Panel a: 3478 genes are declared significant by both $M$ and the modified $M$ statistics; panel b: 86 genes are declared significant by $M$ statistic only; panel c: 89 genes are declared significant by the modified $M$ only; panel d: 13,345 genes are not declared significant by either statistic.
Figure 8.8: Correlation between $M$ and the modified $M$. Panel a: correlation (0.99) between rankings of 3478 genes declared significant by both from $M$ and the modified $M$. Panel b: correlation (0.97) between rankings of 89 genes declared significant by only the modified $M$. Panel c: correlation (0.98) between rankings of 86 genes declared significant by only $M$ and Panel d: correlation (0.998) between rankings of 13,345 genes not declared significant by either the $M$ or modified $M$. 
8.5 Simulation Study

We conduct a simulation study to investigate the performance of the five test statistics. In Section 8.5.1, we compare the three estimators for the variance of Williams’ and Marcus’ (which is the same), $M$, and modified $M$ test statistics. In Section 8.5.2, we investigate the power of the five statistics for a single gene, while in Section 8.5.3, the power of the tests with the multiple testing adjustment is evaluated.

8.5.1 Standard Error Comparison

As a base for the simulations, the ANOVA model (8.1) is assumed. With four dose levels, the order-restricted alternative hypothesis (8.3) can be classified into seven possible trends. Table 8.2 defines the mean structure and assumed parameter values for these seven models, and for the null model ($H_0$) used in the simulations. The scale parameter $\psi$ controls the magnitude of the isotonic means. The larger $\psi$, the larger distance between the means. In this set of simulations it is chosen to equal 1 and 3 based on the settings considered by Marcus (1976).

For each model, $L = 10,000$ datasets are generated. Each dataset contains three arrays per each of four dose levels, i.e., 12 arrays (observations) in total are generated, with variance $\sigma^2=1$.

Table 8.2: Simulation settings: $\mu_i$ is the mean response of dose level $i$, $i = 1, 2, 3, 4$, and $\psi = 1$ or 3.

<table>
<thead>
<tr>
<th>Model</th>
<th>Mean Structure</th>
<th>$\mu_1$</th>
<th>$\mu_2$</th>
<th>$\mu_3$</th>
<th>$\mu_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$g_1$</td>
<td>$\mu_1 = \mu_2 = \mu_3 &lt; \mu_4$</td>
<td>(1 1 1)</td>
<td>1 1 2</td>
<td>$\times 2\psi/\sqrt{3}$</td>
<td></td>
</tr>
<tr>
<td>$g_2$</td>
<td>$\mu_1 &lt; \mu_2 = \mu_3 = \mu_4$</td>
<td>(1 1 2)</td>
<td>2 2 2</td>
<td>$\times \psi$</td>
<td></td>
</tr>
<tr>
<td>$g_3$</td>
<td>$\mu_1 &lt; \mu_2 = \mu_3 = \mu_4$</td>
<td>(1 2 2)</td>
<td>2 2 2</td>
<td>$\times 2\psi/\sqrt{3}$</td>
<td></td>
</tr>
<tr>
<td>$g_4$</td>
<td>$\mu_1 &lt; \mu_2 &lt; \mu_3 &lt; \mu_4$</td>
<td>(1 2 2 3)</td>
<td>2 2 2 3</td>
<td>$\times \psi/\sqrt{2}$</td>
<td></td>
</tr>
<tr>
<td>$g_5$</td>
<td>$\mu_1 = \mu_2 &lt; \mu_3 &lt; \mu_4$</td>
<td>(1 1 2 3)</td>
<td>2 2 2 3</td>
<td>$\times 2\psi/\sqrt{11}$</td>
<td></td>
</tr>
<tr>
<td>$g_6$</td>
<td>$\mu_1 &lt; \mu_2 &lt; \mu_3 = \mu_4$</td>
<td>(1 2 3 3)</td>
<td>2 2 2 3</td>
<td>$\times 2\psi/\sqrt{11}$</td>
<td></td>
</tr>
<tr>
<td>$g_7$</td>
<td>$\mu_1 &lt; \mu_2 &lt; \mu_3 &lt; \mu_4$</td>
<td>(1 2 3 5 4)</td>
<td>2 2 2 3 5</td>
<td>$\times \psi/\sqrt{5}$</td>
<td></td>
</tr>
<tr>
<td>Null</td>
<td>$\mu_1 = \mu_2 = \mu_3 = \mu_4$</td>
<td>(0 0 0 0)</td>
<td>0 0 0 0</td>
<td>$\times \psi$</td>
<td></td>
</tr>
</tbody>
</table>

The performance of the standard error estimators for the Williams and Marcus,
and the modified $M$ test statistics is evaluated.

For Williams' statistic the estimator is $\hat{\sigma}_1 = \sqrt{\frac{3}{2}S^2} = \sqrt{\frac{3}{2}\hat{\sigma}_1}$, where

$$
\hat{\sigma}_1 = \sqrt{\sum_{i=0}^{3} \sum_{j=1}^{3} (y_{ij} - \bar{y}_i)/(12 - 4),}
$$

and where $y_{ij}$ is the gene-expression at dose level $i$ and array $j$, while $\bar{y}_i$ is the sample mean of gene-expression levels at dose $i$.

The estimator of the $M$ statistic, proposed by Hu et al. (2005), is given by

$$
\hat{\sigma}_2 = \sqrt{\sum_{i=0}^{3} \sum_{j=1}^{3} (y_{ij} - \hat{\mu}_i^*)^2/(12 - 4)},
$$

Moreover, we consider the standard error estimate of the modified $M$, denoted as

$$
\hat{\sigma}_3 = \sqrt{\sum_{i=0}^{3} \sum_{j=1}^{3} (y_{ij} - \hat{\mu}_i^*)^2/(12 - I)},
$$

where $I$ is the number of unique isotonic mean levels obtained in the isotonic regression model.

First, we evaluate the mean squared error (MSE) of $\hat{\sigma}_1$, $\hat{\sigma}_2$, and $\hat{\sigma}_3$. The squared bias is estimated by $\hat{\theta}^2_{\sigma} = (\tilde{\sigma} - \sigma)^2$, with $\tilde{\sigma} = \sum_{j=1}^{L} \hat{\sigma}_j/L$. The empirical variance is estimated by $\hat{\theta}^2_{\sigma} = \sum_{j=1}^{L} (\hat{\sigma}_j - \tilde{\sigma})^2/L$, leading to the simulation estimate of the MSE given by $\text{MSE}_\sigma = \hat{\theta}^2_{\sigma} + \hat{\theta}^2_{\sigma}$.

Table 8.3 shows the squared bias, variance, and the MSE estimates of the three standard error estimators under the null hypothesis and under the seven alternative hypotheses. The smallest MSE values are obtained for $\hat{\sigma}_3$. Note that, although $\hat{\sigma}_3$ tends to have the highest squared bias, its mean square error is the smallest due to the small variability of this estimator.
Table 8.3: Squared bias, variance and MSE for $\hat{\sigma}_1$, $\hat{\sigma}_2$, and $\hat{\sigma}_3$. The numbers in the table are on $10^{-3}$ scale.

<table>
<thead>
<tr>
<th></th>
<th>$\sigma_1$</th>
<th>$\sigma_2$</th>
<th>$\sigma_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bias$^2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$g_1$</td>
<td>0.739</td>
<td>1.593</td>
<td>0.954</td>
</tr>
<tr>
<td>$g_2$</td>
<td>1.079</td>
<td>0.831</td>
<td>1.382</td>
</tr>
<tr>
<td>$g_3$</td>
<td>0.805</td>
<td>1.548</td>
<td>1.02</td>
</tr>
<tr>
<td>$g_4$</td>
<td>1.185</td>
<td>0.122</td>
<td>1.984</td>
</tr>
<tr>
<td>$g_5$</td>
<td>1.115</td>
<td>0.199</td>
<td>1.982</td>
</tr>
<tr>
<td>$g_6$</td>
<td>0.917</td>
<td>0.304</td>
<td>1.751</td>
</tr>
<tr>
<td>$g_7$</td>
<td>0.706</td>
<td>0.117</td>
<td>1.843</td>
</tr>
<tr>
<td>Null</td>
<td>0.739</td>
<td>1.086</td>
<td>2.062</td>
</tr>
<tr>
<td>Variance</td>
<td>$\sigma_1$</td>
<td>$\sigma_2$</td>
<td>$\sigma_3$</td>
</tr>
<tr>
<td>$g_1$</td>
<td>60.143</td>
<td>61.702</td>
<td>52.806</td>
</tr>
<tr>
<td>$g_2$</td>
<td>61.001</td>
<td>61.883</td>
<td>53.308</td>
</tr>
<tr>
<td>$g_3$</td>
<td>60.254</td>
<td>61.787</td>
<td>52.608</td>
</tr>
<tr>
<td>$g_4$</td>
<td>57.691</td>
<td>58.198</td>
<td>51.585</td>
</tr>
<tr>
<td>$g_5$</td>
<td>59.264</td>
<td>59.172</td>
<td>52.394</td>
</tr>
<tr>
<td>$g_6$</td>
<td>60.8</td>
<td>60.795</td>
<td>53.321</td>
</tr>
<tr>
<td>$g_7$</td>
<td>60.131</td>
<td>60.513</td>
<td>54.102</td>
</tr>
<tr>
<td>Null</td>
<td>60.143</td>
<td>59.686</td>
<td>51.092</td>
</tr>
<tr>
<td>MSE</td>
<td>$\sigma_1$</td>
<td>$\sigma_2$</td>
<td>$\sigma_3$</td>
</tr>
<tr>
<td>$g_1$</td>
<td>60.881</td>
<td>63.295</td>
<td>53.76</td>
</tr>
<tr>
<td>$g_2$</td>
<td>62.08</td>
<td>62.714</td>
<td>54.69</td>
</tr>
<tr>
<td>$g_3$</td>
<td>61.059</td>
<td>63.335</td>
<td>53.628</td>
</tr>
<tr>
<td>$g_4$</td>
<td>58.876</td>
<td>58.321</td>
<td>53.569</td>
</tr>
<tr>
<td>$g_5$</td>
<td>60.379</td>
<td>59.371</td>
<td>54.376</td>
</tr>
<tr>
<td>$g_6$</td>
<td>61.717</td>
<td>61.099</td>
<td>55.072</td>
</tr>
<tr>
<td>$g_7$</td>
<td>60.837</td>
<td>60.63</td>
<td>55.945</td>
</tr>
<tr>
<td>Null</td>
<td>60.881</td>
<td>60.772</td>
<td>53.154</td>
</tr>
</tbody>
</table>
8.5.2 Power Study for a Single Gene Setting

Another simulation study is conducted to evaluate the power of the five test statistics for a single gene setting. Similarly, as in the study presented in Section 8.5.1, datasets of 12 arrays are generated under the seven order-restricted models and the null model (Table 8.2). For each alternative model 5000 datasets are generated with an increasing and a decreasing trend, respectively. For the null model, 10,000 datasets in total are simulated for the comparison of the error rates. The isotonic means of the seven alternatives are specified in Table 8.2 with variance $\sigma^2 = 1$.

For each dataset and each test, $p$-values are obtained from 10,000 permutations. The results are summarized by the proportion of significant tests (with permutation-based $p$-values $\leq 0.05$) that correctly classify the increasing or decreasing trend. For each $\psi$, the power and type I error are shown in Table 8.4. The standard error estimate of the power can be obtained by $\sqrt{\hat{p}(1 - \hat{p})}/10000$ (Marcus 1976) where $\hat{p}$ is the estimate for the power.

The estimated Type I error probability is around 5% for all the tests. The power of the tests depends on the alternative. In general, regarding $E_{2,1}$, Williams’ and Marcus’ tests, we arrive at the same conclusion as Marcus (1976), that the tests yield similar power. We can additionally observe that $M$ and the modified $M$ tests perform similarly as the other three. Hence, for a single gene setting, no test is uniformly better across the set of the considered alternative hypotheses.

8.5.3 Power Study Under Multiple Testing Adjustment

We have also investigated the power of the considered test statistics when dealing with the multiple testing problem. Microarrays with 5000 genes per microarray are generated. For each of the seven alternative models (see Table 8.2) a set of 100 genes (1400 genes in total) with an increasing and a decreasing trend is included. For the remaining 3600 genes no dose effect is assumed (the null model). The $p$-values for the considered test statistics are obtained using 10,000 permutations, and the multiplicity adjustment is provided by using the BH-FDR procedure.

In total, 100 datasets are generated for settings with $\psi = 1$ and $\psi = 3$. Table 8.5 shows the power and FDR with their simulation-based standard error estimates.

For $\psi = 1$ the power of all the tests is very low. Moreover, the FDR is not controlled at the desired level of 5%. This could be related to the multiplicity adjustment procedure: the total number of rejected hypothesis is small, and the proportion of
Table 8.4: Power of the five test statistics for a single gene setting when data are generated under the eight models in Table 8.2.

<table>
<thead>
<tr>
<th></th>
<th>$E^{2}_{01}$</th>
<th>Williams</th>
<th>Marcus</th>
<th>$M$</th>
<th>Modified $M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$g_1$</td>
<td>0.2261</td>
<td>0.1882</td>
<td>0.2173</td>
<td>0.2299</td>
<td>0.1996</td>
</tr>
<tr>
<td>$g_2$</td>
<td>0.2772</td>
<td>0.2196</td>
<td>0.2404</td>
<td>0.2371</td>
<td>0.2331</td>
</tr>
<tr>
<td>$g_3$</td>
<td>0.2245</td>
<td>0.2189</td>
<td>0.199</td>
<td>0.2259</td>
<td>0.2096</td>
</tr>
<tr>
<td>$g_4$</td>
<td>0.2602</td>
<td>0.2943</td>
<td>0.2706</td>
<td>0.3046</td>
<td>0.3177</td>
</tr>
<tr>
<td>$g_5$</td>
<td>0.3271</td>
<td>0.2684</td>
<td>0.2873</td>
<td>0.3134</td>
<td>0.301</td>
</tr>
<tr>
<td>$g_6$</td>
<td>0.2662</td>
<td>0.2454</td>
<td>0.2345</td>
<td>0.2604</td>
<td>0.2819</td>
</tr>
<tr>
<td>$g_7$</td>
<td>0.2953</td>
<td>0.2866</td>
<td>0.2744</td>
<td>0.3053</td>
<td>0.3231</td>
</tr>
</tbody>
</table>

$\psi = 1$

<table>
<thead>
<tr>
<th></th>
<th>$E^{2}_{01}$</th>
<th>Williams</th>
<th>Marcus</th>
<th>$M$</th>
<th>Modified $M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$g_1$</td>
<td>0.9739</td>
<td>0.9369</td>
<td>0.961</td>
<td>0.9669</td>
<td>0.9169</td>
</tr>
<tr>
<td>$g_2$</td>
<td>0.9761</td>
<td>0.9058</td>
<td>0.9289</td>
<td>0.9462</td>
<td>0.887</td>
</tr>
<tr>
<td>$g_3$</td>
<td>0.9772</td>
<td>0.9773</td>
<td>0.9678</td>
<td>0.9773</td>
<td>0.9416</td>
</tr>
<tr>
<td>$g_4$</td>
<td>0.9787</td>
<td>0.9914</td>
<td>0.9873</td>
<td>0.993</td>
<td>0.994</td>
</tr>
<tr>
<td>$g_5$</td>
<td>0.9871</td>
<td>0.9624</td>
<td>0.9761</td>
<td>0.9844</td>
<td>0.9822</td>
</tr>
<tr>
<td>$g_6$</td>
<td>0.9684</td>
<td>0.9706</td>
<td>0.9579</td>
<td>0.9747</td>
<td>0.9856</td>
</tr>
<tr>
<td>$g_7$</td>
<td>0.9803</td>
<td>0.9826</td>
<td>0.978</td>
<td>0.9883</td>
<td>0.9936</td>
</tr>
</tbody>
</table>

$\psi = 3$

<table>
<thead>
<tr>
<th></th>
<th>$E^{2}_{01}$</th>
<th>Williams</th>
<th>Marcus</th>
<th>$M$</th>
<th>Modified $M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$g_1$</td>
<td>0.9112</td>
<td>0.8905</td>
<td>0.9112</td>
<td>0.8905</td>
<td>0.8928</td>
</tr>
</tbody>
</table>

wrong rejections is not well estimated, i.e, the FDR is not well estimated.

With $\psi = 3$ the power of the test statistics is greatly improved and the FDR is well controlled. $E^{2}_{01}$ seems to provide a slightly higher power compared to the other tests. This can be explained by the fact that $E^{2}_{01}$ yields good power under each of seven alternative models. Note that the power obtained using the modified $M$ test statistic is comparable. When multiplicity is taken into account, $E^{2}_{01}$, $M$, and the modified $M$ have a higher power compared to Williams’ and Marcus’ tests (0.9112, 0.8905, and 0.8928 compared to 0.8454 and 0.8477, respectively).

8.6 Discussion

In this chapter, we evaluate several test statistics for testing a monotonic trend in the relationship of gene-expression and doses in a microarray context. In particular,
Table 8.5: Power study of the five test statistics under multiple testing adjustment.

<table>
<thead>
<tr>
<th>( \psi = 1 )</th>
<th>( E^2_{01} )</th>
<th>Williams</th>
<th>Marcus</th>
<th>( M )</th>
<th>Modified ( M )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power</td>
<td>0.0354</td>
<td>0.0287</td>
<td>0.0289</td>
<td>0.0306</td>
<td>0.0309</td>
</tr>
<tr>
<td>SE(Power)</td>
<td>(0.0049)</td>
<td>(0.0046)</td>
<td>(0.0046)</td>
<td>(0.0048)</td>
<td>(0.0048)</td>
</tr>
<tr>
<td>FDR</td>
<td>0.1944</td>
<td>0.2077</td>
<td>0.2135</td>
<td>0.1835</td>
<td>0.1907</td>
</tr>
<tr>
<td>SE(FDR)</td>
<td>(0.0507)</td>
<td>(0.0579)</td>
<td>(0.0568)</td>
<td>(0.0534)</td>
<td>(0.0534)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( \psi = 3 )</th>
<th>( E^2_{01} )</th>
<th>Williams</th>
<th>Marcus</th>
<th>( M )</th>
<th>Modified ( M )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power</td>
<td>0.9112</td>
<td>0.8454</td>
<td>0.8477</td>
<td>0.8905</td>
<td>0.8928</td>
</tr>
<tr>
<td>SE(Power)</td>
<td>(0.0074)</td>
<td>(0.0099)</td>
<td>(0.0096)</td>
<td>(0.0082)</td>
<td>(0.0079)</td>
</tr>
<tr>
<td>FDR</td>
<td>0.0404</td>
<td>0.0424</td>
<td>0.0426</td>
<td>0.0399</td>
<td>0.0401</td>
</tr>
<tr>
<td>SE(FDR)</td>
<td>(0.0053)</td>
<td>(0.0054)</td>
<td>(0.0053)</td>
<td>(0.0052)</td>
<td>(0.0052)</td>
</tr>
</tbody>
</table>

we consider Williams’ step down procedure (Williams 1971, 1972), Marcus’ procedure (Marcus 1976), likelihood ratio statistic (Robertson et al. 1988), \( M \) (Hu et al. 2005), and the modified \( M \) test statistic. Directional inference using these statistics is discussed for the situation when the direction of the trend is unknown in advance. To avoid inference based on asymptotic theory, we consider the use of permutation tests. Accordingly, several multiplicity adjustment methods including the MD-FDR are applied. The BH-FDR procedure that controls the FDR provides the most powerful approach as compared to the other methods (Tusher et al. 2001, Ge et al. 2003, Storey and Tibshirani 2003).

For the analysis discussed above, we observe comparable results for the five test statistics. However, a difference in the results between \( M \) and \( \bar{E}^2_{01} \) is observed. Modifying the number of degrees of freedom for the \( M \) statistic improves the MSE of the estimate of the standard error. However, the simulation study investigating power of the five test statistics under a multiple testing adjustment shows that the \( M \) and the modified \( M \) have a similar power.

As we argue in Section 8.3.1, a two sided inference can result in rejecting the null hypothesis in both directions \( p_{UP}^* < \alpha/2 \) and \( p_{Down}^* < \alpha/2 \). This implies, as illustrated in Section 8.3.1, a non-monotone dose-response relationship. The difference between the four t-type test statistics (Williams’, Marcus’, \( M \), and the modified \( M \)) is due to the estimates of the standard error. Williams and Marcus used the unbiased
estimator calculated under the unrestricted ordered alternative, while \( M \) and the modified \( M \) use an estimator calculated under the ordered alternative. Williams’ and Marcus’ tests tend to reject genes when the difference calculated for the numerator exists and the standard error calculated under the unrestricted alternative is small. In particular, when the true means follow a simple tree (i.e., \( \mu_1 \leq [\mu_2, \mu_3, \mu_4] \), where no ordering of \( \mu_2 \), \( \mu_3 \), and \( \mu_4 \) is required), a unimodal partial ordering (i.e., \( \mu_1 \leq \mu_2 \leq \mu_3 \geq \mu_4 \)) or a simple loop (i.e., \( \mu_1 \leq [\mu_2, \mu_3] \leq \mu_4 \), where \( \mu_2 \) and \( \mu_3 \) can be unequal) (Robertson et al. 1988), Williams’ and Marcus’ tests are more likely to reject the null hypothesis of homogeneity of means (no dose effect) in favor of the simple ordered alternative (\( H_{UP}^U \) or \( H_{DOWN}^U \)) than \( M \) and the modified \( M \) test statistics. We have shown that for a single gene the power of the four t-type test statistics is comparable (Table 8.4). However, the power, obtained after adjusting for multiple testing, for \( M \) and the modified \( M \) is higher than that obtained for Williams’ and Marcus’ tests.

For a single gene the power obtained for \( E_{01}^2 \) is comparable to the power obtained for the four t-type test statistics. Moreover, after adjustment for multiplicity, the power obtained for \( E_{01}^2 \) is only slightly higher than for the \( M \) and the modified \( M \) tests (shown in Table 8.5). In our opinion, if the question of primary interest is the comparison between the highest and the lowest dose levels, \( E_{01}^2 \), \( M \), and the modified \( M \) tests are comparable (in terms of the control of the FDR and power). However, if the question of primary interest is to detect a monotone trend, the global test \( E_{01}^2 \) is to be preferred.

We focus on testing the null hypothesis against a simple ordered alternative. Whenever the null hypothesis is rejected, the primary interest is to identify the dose-response curve shape. For a dose-response experiment with \( K+1 \) dose levels, there is a finite number of isotonic models, which can be fitted to the data. For example, for an experiment with four dose levels there are seven upward monotone models (given in Table 8.2) and seven downward monotone models, which can be fitted to the data. The testing procedures discussed in this chapter allow us to identify genes, for which the dose response curve is monotone, but not to identify the dose-response curve shape. The latter can be done using a model selection procedure, based on information criteria. Such a procedure is presented in the following chapter.

The results presented in this chapter were published in Lin et al. (2007a).
Chapter 9

The IsoGene Library in $R$

9.1 Introduction to IsoGene Package

In this chapter, we present an $R$ package called IsoGene to implement the statistical methods discussed by Lin et al. (2007), where the primary interest is testing for a monotonic relationship between gene-expressions and doses in a microarray experiment.

The main IsoGene package functions are IsoRawp() and IsoTestBH(), which calculate the raw $p$-values using permutations and adjust them using the BH- and BY-FDR procedures. The supporting functions IsoGene1() andIsoGenem() are used to calculate the five test statistics from isotonic regression for one gene and all the genes, respectively. The remaining functions IsopvaluePlot(), IsoBHplot(),IsoPlot1() and IsoPlot2() are used to display the data and to show the results of testing procedures.

9.2 Testing for Trends: Testing Procedures, Multiplicity and Resampling-based Inference

Details on the five test statistics for testing the null hypothesis of no dose effect against order restricted alternatives are given in Section 8.2.

For adjusting for multiple testing, only the BH-FDR procedure (Benjamini and Hochberg 1995) and BY-FDR procedure (Benjamini and Yekutieli 2001) are consid-
Chapter 9. IsoGene R

tered in library IsoGene(). The matrix of the values of the test statistics for each gene and permutation is referred as the permutation matrix under the null distribution (see Section 5.1.2).

This matrix is used to calculate the one-sided \( p \)-values for the inference. In the first step the one-sided raw (unadjusted for multiple testing) \( p^{U_p} \)-values are calculated using (5.2) or (5.3) based on the test statistic \( T^{U_p} \).

For \( p^{Down} \)-values, expect of \( \hat{E}_{01}^{2} \), for which the test statistic value \( t_i \) is always between 0 and 1 and can be obtained in the same way as \( p^{U_p} \)-values,

\[
p^{Down} = \frac{\#(b : t_{ib} \leq t_i)}{B} \text{ or } p^{Down} = \sum_{b=1}^{B} \sum_{j=1}^{m} \frac{|t_{jb}| \leq |t_i|}{B \times m}
\]

should be used with \( t_{ib} \) and \( t_{jb} \) the test statistic values obtained for gene \( i \) and \( j \) from permutation \( b \). This is because under the decreasing trend, we reject the four test statistics (namely, Williams’, Marcus’, the \( M \) and modified \( M \)) with large negative values.

Based on the \( p \)-values, various methods adjusting the type I error can be applied, such as the Bonferroni, Holm, Hochberg, and BH-FDR and BY-FDR (Reiner et al. 2003 and Ge et al. 2003).

9.3 Using IsoGene Library

9.3.1 Data Example

We use the data analyzed in Chapter 8 as a case study to illustrate the use of the IsoGene library.

A dataframe with the log2 transformed gene intensities is loaded into R environment. The first ten genes and first six samples are displayed, where the row names of the genes show the probe ID, X1, X1.1 and X1.2 are the three arrays for dose zero, while X2, X2.1 and X2.2 are the arrays for the first dose.

```r
> load("data.Rdata")
> data[1:10,1:6]
   X1  X1.1  X1.2  X2  X2.1  X2.2
31637_s_at  6.923109  7.024719  7.170328  7.219297  7.076908  7.404949
32402_s_at  5.107275  5.092935  5.255918  5.312913  4.893855  4.596591
33646_g_at  5.913526  6.026197  5.141728  5.828770  5.269202  5.461664
```

9.3.2 Loading the Library

To load the IsoGene package into R, a binary zip-package of IsoGene program (for Windows) needs to be installed. IsoGene library requires R packages Multtest and ff, which need to be installed as well. Once the packages are installed, they are available for use after being loaded in memory, which is usually done by the user:

```r
> library(IsoGene)
```

The functions included in the package can be listed using the R help system:

```r
> help(IsoGene)
```

First, IsoPlot1() and IsoPlot2() can be used to explore the data. Second, IsoGene1() and IsoGene2() can be used to calculate the test statistics. Third, IsoRawp() provides the output for two-sided or one-sided p-values ($p_{U}^p$ or $p_{D}^down$). Based on the p-values obtained, one can choose one test statistic and multiplicity adjustment method for inference by using IsoTestBH(). Finally, IsopvaluePlot() can be useful for examining both of $p_{U}^p$- or $p_{D}^down$-values, and in particularly, as a post hoc procedure it can be used to examine genes with both small $p_{U}^p$- and $p_{D}^down$-values.

9.4 The IsoGene Functions

9.4.1 Exploring the Data

IsoPlot1() and IsoPlot2() are two plotting functions that can be used to explore the data. Scatterplots for the second gene in the dataset (data[2,]) can be produced by

```r
> x <- c(rep(1,3), rep(2,3), rep(3,3), rep(4,3))
> gene1 <- data[2,]
```
Library IsoGene aims to identify genes with a monotonic trend in the expression levels with respect to the increasing doses using several test statistics. They include the global likelihood ratio test (Bartholomew 1961, Bartholomew et al. 1972 and Robinson et al. 1988), Williams (1971, 1972), Marcus (1976), M (Hu et al. 2005) and the modified M (Lin et al. 2007). The p-values of the five test statistics are obtained using permutation and they are adjusted using BH (Benjamini and Hochberg 1995) and BY (Benjamini and Yekutieli 2004) procedures for controlling FDR.

Details

Package: IsoGene
Type: Package
Version: 1.0
Date: 2007-05-02
License: Free

The package includes the following functions:

- `IsoGene1` calculates the five test statistics testing both increasing and decreasing alternatives for a single gene
- `IsoGenem` calculates the five test statistics testing both increasing and decreasing alternatives for all the genes in the data set
- `IsoRawp` obtains the raw (one-sided and two-sided) p-values using permutations
- `IsoTestBH` a BY procedure to adjust p-values while controlling FDR
- `IsoBHPlot` plot of adjusted BH and BY p-values
- `IsoPlot1` plot of data and sample means
- `IsoPlot2` plot of data and isotonic means in an increasing or decreasing trend
- `IsopvaluePlot` plot of p-values obtained using permutation under increasing or decreasing alternatives

Author(s)
Lin et al.

Maintainer: Dan Lin <dan.lin@uhasselt.be>

References

See Also
<multtest>, `IsoGene1`, `IsoGenem`, `IsoRawp`, `IsoTestBH`, `IsoBHPlot`, `IsoPlot1`, `IsoPlot2`.

Figure 9.1: The main help file of IsoGene package.

```r
> par(mfrow=c(1,2))
> IsoPlot1(x, y=gene1)
> IsoPlot2(x, y=gene1)
```

The left panel in Figure 9.2 shows the original data points (as circles) and sample means (as pluses) for each dose. The right panel in Figure 9.2 shows the increasing isotonic regression model (blue solid line) fitted on the data. The fitted monotonic line does not indicate the significance of the test, but simply shows a more likely increasing (or decreasing) trend.
Figure 9.2: The data points are plotted as circles, while sample means as pluses. The right panel additionally plots the fitted increasing isotonic regression model (blue solid line).

9.4.2 Calculating the Test Statistics

The five test statistics described in Chapter 8 can be obtained by using the function `IsoGene1()` for a single gene and using the function `IsoGenem()` for all the genes simultaneously. The following R codes illustrate the input and output generated by these two functions:

```r
> stat1 <- IsoGene1(x, gene1)
> stat1
$ E2.up [1] 0.2010957
$ Williams.up [1] 0.6712589
$ Marcus.up [1] 1.3646790
$ M.up [1] 1.0004640
$ ModM.up [1] 1.0611520
$ E2.dn [1] 0.0098105
```
The first 10 objects are the values calculated for the five test statistics under increasing and decreasing trends. The last object indicates the higher likelihood of isotonic regression with "u" meaning an increasing trend or "d" meaning a decreasing trend.

We use the first 10 genes as an example to illustrate the use of function `IsoGenem()`:

```r
> statm <- IsoGenem(x, data[1:10,])
> statm
$E2.up
  2   3   4   5   6
0.00000000 0.20109571 0.50774178 0.24835414 0.00000000
  7   8   9  10  11
0.16263545 0.43080221 0.00000000 0.06367646 0.00000000
$Williams.up
[1] -1.1298186 0.6712589 2.4888850 0.8911883 -0.6520746
[6] 0.5582301 2.1412458 -0.5774471 0.8895008 -1.6655641
$Marcus.up
[1] 0.0000000 1.3646791 2.4888850 1.3929232 0.0000000
[6] 1.4262255 2.1412458 0.0000000 0.8895008 0.0000000
$M.up
[1] 0.0000000 1.0004635 2.0194520 0.9386711 0.0000000
[6] 0.7196721 1.6954778 0.0000000 0.4960717 0.0000000
$ModM.up
[1] 0.0000000 1.0611518 2.1419523 1.0494662 0.0000000
[6] 0.8046179 1.7983258 0.0000000 0.5261635 0.0000000
$E2.dn
  2   3   4   5   6
0.275992755 0.009810531 0.000000000 0.002919082 0.139779913
  7   8   9  10  11

The output from IsoGenem has the same structure as the one for IsoGene1, but each object contains the values of the test statistics and the likely direction of the isotonic regression model for all the genes.

### 9.4.3 Obtaining Raw p-values

As discussed above, we use permutations to obtain the raw p-values for the five test statistics. The function `IsoRawp()` can be used in the following way:

```r
> rawp <- IsoRawp(x=x, y=data, niter=1000, seed=1234)
```

The four arguments in this function need to be specified, with no default pre-specified values. `x` is the explanatory variable indicating the dose levels for all the samples in the data. `data` is the data frame of the gene expression. `niter` defines the number of permutations used to approximate the null distribution and `seed` determines the random seed used to generate the permutations. The output item `rawp` contains four objects with p-values for the five test statistics: the first one contains the two-sided p-values, the second contains the one-sided p-values, the third contains p\textsuperscript{U\textsubscript{P}}-values, and the last one contains p\textsuperscript{D\textsubscript{own}}-values. Below we print a part
of the object with two-sided \textit{p}-values for illustration:

\begin{verbatim}
> rawp.twosided=rawp[[1]]
> rawp.twosided[1:10,]
\end{verbatim}

<table>
<thead>
<tr>
<th>Probe.ID</th>
<th>E2</th>
<th>Williams</th>
<th>Marcus</th>
<th>M</th>
<th>ModM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>31637</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>2</td>
<td>32402</td>
<td>0.129</td>
<td>0.225</td>
<td>0.124</td>
<td>0.123</td>
</tr>
<tr>
<td>3</td>
<td>33646</td>
<td>0.003</td>
<td>0.004</td>
<td>0.003</td>
<td>0.002</td>
</tr>
<tr>
<td>4</td>
<td>34063</td>
<td>0.487</td>
<td>0.379</td>
<td>0.500</td>
<td>0.467</td>
</tr>
<tr>
<td>5</td>
<td>33494</td>
<td>0.071</td>
<td>0.185</td>
<td>0.035</td>
<td>0.063</td>
</tr>
<tr>
<td>6</td>
<td>34031</td>
<td>0.082</td>
<td>0.220</td>
<td>0.086</td>
<td>0.103</td>
</tr>
<tr>
<td>7</td>
<td>34449</td>
<td>0.357</td>
<td>0.445</td>
<td>0.432</td>
<td>0.400</td>
</tr>
<tr>
<td>8</td>
<td>34478</td>
<td>0.472</td>
<td>0.516</td>
<td>0.535</td>
<td>0.518</td>
</tr>
<tr>
<td>9</td>
<td>35436</td>
<td>0.151</td>
<td>0.116</td>
<td>0.148</td>
<td>0.150</td>
</tr>
<tr>
<td>10</td>
<td>36711</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

The first output object from \texttt{rawp} is a matrix with six columns, where the first column indicates the probe ID. Columns from the second to the sixth are \textit{p}-values for each of the five test statistics, respectively. The remaining three output objects (\texttt{rawp[[2]]}, \texttt{rawp[[3]]}, \texttt{rawp[[4]]}) are structured in the same way.

### 9.4.4 Plot of \textit{p}-values for a Single Gene

For a single gene, the function \texttt{IsopvaluePlot()} can be used to show the \textit{p}^{Up} and \textit{p}^{Down}-values for a given test statistic:

\begin{verbatim}
IsopvaluePlot(x, y, niter, seed, stat = c("E2", "Williams", "Marcus", "M", "ModifM"))
\end{verbatim}

We use one gene as an example to illustrate how \textit{p}^{Up} and \textit{p}^{Down}-values (in the upper and lower panels of Figure 9.3) are obtained. In Figure 9.3, the observed test statistics are drawn as the dashed line, and the values of the test statistics obtained from permutations are spread over the x-axis. For this gene, the \textit{p}^{Up} is much smaller as compared to the \textit{p}^{Down} since \textit{T}^{Up} \gg \textit{T}^{Down}, which implies a possible increasing trend in the data.
> gene1 <- data[2,]
> IsopvaluePlot(x, gene1, niter=1000, seed=123, stat="E2")

![Gene 31637_s_at:p-value^up=0](image1)

![Gene 31637_s_at:p-value^down=0.761](image2)

**Figure 9.3:** The $p^{\text{Up}}$ and $p^{\text{Down}}$-values using $E_{01}^2$ for an example gene. The dashed line is the observed test statistic value. In the upper panel, the dashed line (at the right) is larger than most of the test statistics from permutations, which results in a small $p^{\text{Up}}$-value. In the lower panel, the dashed line (close to zero) is smaller than most of the test statistics from permutations, which results in a large $p^{\text{Down}}$-value.

**9.4.5 BH/BY-FDR Procedures for Adjusting for Multiple Testing**

With the two-sided $p$-values, the user needs to select one of the five test statistics, the FDR level, and the type of multiplicity adjustment (BH-FDR or BY-FDR) to obtain the list of significant genes:
The following example shows the use of the global likelihood ratio test $E_{01}^2$, the FDR level of 0.05 and the BH-FDR procedure controlling the FDR:

```r
> E2.BH <- IsoTestBH(rp, FDR=0.05, type="BH", stat="E2")
> E2.BH[1:10,]
```

<table>
<thead>
<tr>
<th>Probe.ID</th>
<th>row.name</th>
<th>raw p-values</th>
<th>BH adjusted p values</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>s_at</td>
<td>0.000</td>
<td>0.000000000</td>
</tr>
<tr>
<td>2</td>
<td>g_at</td>
<td>0.003</td>
<td>0.015647131</td>
</tr>
<tr>
<td>3</td>
<td>36711_at</td>
<td>0.000</td>
<td>0.000000000</td>
</tr>
<tr>
<td>4</td>
<td>37079_at</td>
<td>0.001</td>
<td>0.007115111</td>
</tr>
<tr>
<td>5</td>
<td>37117_at</td>
<td>0.011</td>
<td>0.042679297</td>
</tr>
<tr>
<td>6</td>
<td>37152_at</td>
<td>0.003</td>
<td>0.015647131</td>
</tr>
<tr>
<td>7</td>
<td>38158_at</td>
<td>0.003</td>
<td>0.015647131</td>
</tr>
<tr>
<td>8</td>
<td>38241_at</td>
<td>0.000</td>
<td>0.000000000</td>
</tr>
<tr>
<td>9</td>
<td>39248_at</td>
<td>0.003</td>
<td>0.015647131</td>
</tr>
<tr>
<td>10</td>
<td>39249_at</td>
<td>0.008</td>
<td>0.033642751</td>
</tr>
</tbody>
</table>

Here we show only the first ten genes declared significant by using $E_{01}^2$ test. The output results in a matrix of five columns: the first column indicates the probe ID, the second column is the corresponding row number of significant genes in the original dataset, the third column is the unadjusted/raw p-value, and the last column is the adjusted p-value using the requested “BH” procedure. The order of the list of genes found significant is based on the row number. Moreover, the function IsoBHPlot() can be used to visualize the number of significant findings for the BH-FDR and BY-FDR procedures for the specified test statistic:

```r
IsoBHPlot(rp, FDR=c(0.05,0.1), stat=c("E2","Williams","Marcus","M","ModifM"))
```

Figure 9.4 shows the unadjusted (solid blue line) and the BH-FDR (dotted and dashed red line) and BY-FDR (dashed green line) adjusted p-values for $E_{01}^2$. It is obtained using the function IsoBHPlot():

```r
> IsoBHPlot(rp, FDR=0.05, stat="E2")
```
Figure 9.4: The unadjusted (solid blue line), and the BH-FDR (dotted and dashed red line) and BY-FDR (dashed green line) adjusted p-values for $E_{01}^2$. 
9.5 The Help System

In this section, we show an example of the help files available in the IsoGene package. For example, using the codes below, the help file for the function `IsoTestBH()` can be obtained and shown in Figure 9.5.

```r
> help(IsoTestBH)
```

![Help text for IsoTestBH()](image)

The `IsoTestBH()` function in the IsoGene package is described as follows:

### Description

Test of monotonic trends using the five test statistics with BH or BY adjustment

### Usage

```r
IsoTestBH(rp, FDR, type = c("BH", "BY"), stat = c("E2", "Williams", "Marcus", "M", "ModifM"))
```

### Arguments

- `rp`: raw p-value matrix with each row for one gene and 6 columns, the first column contains the Probe ID, the second to the sixth columns are raw p-values for the five test statistics.
- `FDR`: desired FDR to control.
- `type`: choose BH or BY procedure to control FDR.
- `stat`: choose one of the five test statistics to use.

### Details

The input raw p-values to this function can be the one-sided or the two-sided ones which are obtained using function `raw.p`. The results using one-sided p-values and FDR controlling at alpha/2 is equivalent to using two-sided p-values and FDR controlling at alpha.

### Value

A list of significant genes while controlling FDR is obtained, with 4 columns: the first column is the probe ID, the second column is the row id, the third column is the raw p-values of the significant genes and the last column is the adjusted p-values of significant genes using BH or BY procedure.

### Note

This function only allows one type of FDR adjustment, either BH or BY. For other type of adjustment, use function `mt.rawp2adjp` in package `multtest`.

### Author(s)

Lin et al.

### See Also

`mt.rawp2adjp`, `IsoRawp`

### Examples

```r
set.seed(1234)
rp <- data.frame(paste("g",1:100), matrix(runif(500,0,0.1), 100, 5))
sign <- IsoTestBH(rp, FDR = 0.05, type = "BH", stat = "E2")
```

*Figure 9.5: The help file for the function `IsoTestBH()`.*

The `IsoGene` package can be downloaded at [https://r-forge.r-project.org/projects/isogene/](https://r-forge.r-project.org/projects/isogene/).

The contents of this chapter were summarized in the manuscript (Lin et al. 2008).
Chapter 10

Classification of Trends in Dose-response Microarray Experiments Using Information Theory Selection Methods

10.1 Introduction

The second question of interest in the dose response study is the nature (or the curve shape) of the dose-response relationship. This question is closely related to the problem of determination of the minimum effective dose (MED), that is the smallest dose at which the mean response is shifted from the mean of dose zero (Ruberg 1989, Tamhane et al. 1996). Several testing procedures were proposed for finding the MED. For example, Williams (1971, 1972) proposed a step-down procedure, in which tests are performed sequentially from the highest to the lowest dose level. The procedure stops at the first dose level, for which the null hypothesis (of no dose effect) is not rejected. As a result, the MED is the first dose above that dose level, at which the testing stopped. Other test procedures, proposed by Tamhane et al. (1996), are based
on contrasts among the sample means of gene-expression at different dose levels. Note that the Williams procedure assumes monotonicity in the dose-response relationship, while the tests based on contrasts of the sample means do not require this assumption. In the microarray setting, the testing procedures mentioned above are additionally subject to the multiple testing problem. To avoid the multiple testing procedures for determining the MED and to identify the shape of the dose-response curve, we propose to classify possible dose-response trends using model selection based on information theory. Assuming a monotone relationship, the dose-response curve could be either linear, nonlinear, concave or convex. Furthermore, for an experiment with $K + 1$ dose levels, there is a fixed number of monotonic models that can be fitted.

For instance, in a dose-response experiment with four dose levels, upon the establishment of a monotonic relationship between gene-expression and doses, there is a set of seven models, shown in Table 10.1 and Figure 10.1, that can be fitted to the data. Each model can be associated with a MED. For example, $g_1$ is a model with two parameters, and the MED is the last dose level, while $g_2$ is also a model with two parameters, but the MED is the third dose level.

Table 10.1: The set of seven possible monotonic dose-response models (with different number of parameters) for an experiment with four dose levels. $\mu_i$ is the mean response at dose level $d_i$. MED: minimum effective dose.

<table>
<thead>
<tr>
<th>Model</th>
<th>Mean Structure</th>
<th>MED</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>$g_1$</td>
<td>$\mu_1 = \mu_2 = \mu_3 &lt; \mu_4$</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>$g_2$</td>
<td>$\mu_1 = \mu_2 &lt; \mu_3 = \mu_4$</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>$g_3$</td>
<td>$\mu_1 &lt; \mu_2 = \mu_3 = \mu_4$</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>$g_4$</td>
<td>$\mu_1 &lt; \mu_2 = \mu_3 &lt; \mu_4$</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>$g_5$</td>
<td>$\mu_1 = \mu_2 &lt; \mu_3 &lt; \mu_4$</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>$g_6$</td>
<td>$\mu_1 &lt; \mu_2 &lt; \mu_3 = \mu_4$</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>$g_7$</td>
<td>$\mu_1 &lt; \mu_2 &lt; \mu_3 &lt; \mu_4$</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>
The set of the seven possible upward monotonic dose-response curves for an experiment with four dose levels.

The aim is to choose the best model among these possible monotonic trends. To this purpose we propose to use the information model selection theory proposed by Burnham and Anderson (2002). For a given set of candidate models, the Order Restricted Information Criterion (ORIC), the Akaike Information Criterion (AIC), and the Bayesian Information Criterion (BIC) are used to calculate the posterior probability of each model in the set. The model with the highest posterior probability is selected. This allows us to identify both the shape of the dose-response curve and the MED level.

We use the dose-response data (three higher EGF doses and one control EGF dose for the control compound) analyzed in Chapter 8 as case study to illustrate the dose-response identification approach. After the initial testing testing procedure (using the $E_{01}^2$ statistic), there were 3613 genes for which the null hypothesis of homogeneity of means in gene-expression is rejected (see Table 8.1 in Chapter 8). However, due to a different seed used for the permutations in the current analysis, we obtained 3499 genes.

Figure 10.2 shows examples of data for four genes. Two genes in Figure 10.2a
and Figure 10.2c reveal pattern $g_7$ with four parameters and an increasing and a decreasing trend, respectively; gene-expression in Figure 10.2b reveals pattern $g_1$ with two monotonic means, while gene-expression in Figure 10.2d reveals pattern $g_4$ with three monotonically decreasing means.

![Graphs showing trend patterns](image)

**Figure 10.2:** Dose-response relationship in a microarray experiment. Circles are observed gene-expressions, pluses are the sample means at each dose level. Solid line: isotonic regression.

The contents of the chapter is organized as follows. In the first step we use the $E_0^2$ test and select 3499 genes, for which the null hypothesis is rejected. This list of genes is used for classification in the second step of the analysis. In Section 10.2 we
address the problem of trend classification within the framework of model selection. In Section 10.3, the ORIC is discussed. We apply the proposed methods to the human epidermal squamous carcinoma cell line data as the case study in Section 10.4. Section 10.5 describes the results of a simulation study conducted to compare the influence of the use of the four information criteria (likelihood, AIC, BIC, and ORIC). Section 10.6 concludes the chapter with a discussion.

10.2 Classification of Trends Using Information Criteria

10.2.1 Classification of Trends Based on Posterior Probabilities

In this section we address the problem of trend classification (or of the identification of the dose-response curve shape). For each gene declared significant in the first testing step, the set of models \{g_1, g_2, g_3, g_4, g_5, g_6, g_7\}, given in Table 10.1 and shown in Figure 10.1, is the set of seven possible models with increasing trends for an experiment with four dose levels. Analogously, a set of seven models with decreasing trends is considered as well, with isotonic means given as in Table 10.1, but with negative values. Note that a model selection procedure that leads to a selection of the best model from the set of all candidate models will allow to identify both dose-response curve shape and the MED. On the other hand, using hypothesis testing for the determination of the MED will not always allow us to identify the shape of the dose-response curve.

Burnham and Anderson (2002) defined the posterior probability of the model given the data as

$$P(g_i|D) = \frac{P(D|g_i)P(g_i)}{\sum_{r=1}^{R} P(D|g_r)P(g_r)} \quad i = 1, \ldots, R. \quad (10.1)$$

Here, \(P(D|g_i)\) and \(P(g_i)\) are the likelihood and the prior probabilities of the \(i\)th model, respectively, and \(R\) is the number of all the possible models (\(R=7\) for our case study). Note that if the observed dose-response curve is monotonic and we use a non-informative prior, i.e., \(P(g_i) = 1/R\), the isotonic regression model obtained will have the highest posterior probability. However, the posterior model probabilities (10.1) do not take the complexity of the model into account. In what follows we focus on the model selection procedures based on information criteria, that take into account
both the goodness-of-fit and the model complexity. In other words, we focus on the question whether the isotonic regression model can be further simplified.

10.2.2 Akaike Weights and Bayesian Posterior Model Probabilities

Akaike Weights

The model selection theory discussed by Burnham and Anderson (2002, 2004) allows to incorporate the need to balance between goodness-of-fit and model complexity within the model selection procedure. The starting point for Burnham and Anderson’s model selection theory is the Kullback-Leibler information (K-L) given by

\[ I(f, g) = \int f(x) \log \frac{f(x)}{g(x|\theta)} dx. \]  

(10.2)

Here, \( f \) represents the density function of the true and unknown model, \( g \) represents the density function of the model that is used to approximate \( f \), and \( \theta \) is the unknown parameter to be estimated. The K-L information is interpreted as the loss of information when the true model \( f \) is approximated by the model \( g(x|\hat{\theta}) \), where \( \hat{\theta} \) is the parameter estimate for the unknown parameter \( \theta \). For a given set of candidate models \( \{g_1, g_2, \ldots, g_R\} \), one can compare the K-L information for each model and select the model that minimizes the information loss across the considered set of models (Burnham and Anderson 2002, 2004, Poeter and Anderson 2005). However, in practice \( I(f, g) \) cannot be computed since the true model \( f \) is unknown.

Akaike (1973, 1974) made the link between the K-L information and the likelihood theory and showed that the expected Kullback-Leibler information can be expressed as

\[ \hat{E}(K-L) = \log \ell(\theta|D) - M, \]  

(10.3)

where \( \ell(\theta|D) \) is the likelihood and \( M \) is the number of parameters in the model. The well-known Akaike’s Information Criterion (AIC) is given by

\[ AIC = -2\log \ell(\theta|D) + 2M. \]  

(10.4)

Akaike’s approach allows for model selection that takes into account both goodness-of-fit and model complexity. Because the individual \( AIC \) values are not interpretable, as they contain arbitrary constants and are much affected by sample size, for a given
set of $R$ models, Burnham and Anderson (2004) proposed to rescale the $AIC$ to

$$\Delta AIC_i = AIC_i - AIC_{\text{min}}, \quad i = 1, \ldots, R, \quad (10.5)$$

with $AIC_{\text{min}}$ being the smallest $AIC$ value across the set of $R$ models. The $AIC$ differences, $\Delta AIC_i$, are interpreted as the information loss when model $g_i$, rather than the best model $g_{\text{min}}$, is used to approximate $f$. Some simple rules of thumb are used in assessing the relative merits of the models in the set (Burnham and Anderson 2002): models with $\Delta AIC_i \leq 2$ have substantial support (evidence); those with $4 \leq \Delta AIC_i \leq 7$ have considerably less support; and models with $\Delta AIC_i > 10$ have essentially no support.

Akaike (1981) advocated the use of $\exp(-\frac{1}{2} \Delta AIC_i)$ for the relative likelihood of the model given the data:

$$\ell(g_i|D) \propto e^{-\frac{1}{2} \Delta AIC_i}. \quad (10.6)$$

Note that the model likelihood $\ell(g_i|D)$ takes into account both goodness-of-fit and model complexity, while $P(g_i|D)$ takes into account only goodness-of-fit. Similar to the posterior probabilities in (10.1), Akaike (Burnham and Anderson 2002) defined Akaike’s weights by

$$P_A(g_i|D) = \frac{\exp(-\frac{1}{2} \Delta AIC_i)p(g_i)}{\sum_{r=1}^R \exp(-\frac{1}{2} \Delta AIC_r)p(g_r)}. \quad (10.7)$$

Akaike’s weight $P_A(g_i|D)$ can be interpreted as the weight of evidence that model $g_i$ is the best $K-L$ model given a set of $R$ models and given that one of the models in the set must be the best $K-L$ model. Note that for non-informative prior probabilities $P(g_i) = 1/R$, Akaike’s weights can be interpreted as the posterior probabilities of the models.

**Bayesian Posterior Model Probabilities**

The Bayesian Information Criterion ($BIC$), proposed by Schwarz (1978), is given by

$$BIC = -2\log(\ell(\theta|D)) + M\log(n), \quad (10.8)$$

where $n$ is the number of arrays in the data. The $BIC$ uses a higher penalty on the number of observations than the $AIC$, which penalizes on the number of parameters in the model. Therefore, the $BIC$ leads to the selection of less complicated models.
The posterior model probabilities are given by
\[ P_B(g_i|D) = \frac{\exp\left(-\frac{1}{2} \Delta BIC_i\right)p(g_i)}{\sum_{r=1}^R \exp\left(-\frac{1}{2} \Delta BIC_r\right)p(g_i)}. \] (10.9)

Burnham and Anderson (2002) show that for prior probabilities
\[ P(g_i) = B \exp\left(\frac{1}{2} \Delta BIC_i\right) \exp\left(-\frac{1}{2} \Delta AIC_i\right) \] (10.10)
it follows that \[ P_B(g_i|D) = P_A(g_i|D), \] where \( B \) is a constant.

### 10.3 Order Restricted Information Criterion

Anraku (1999) proposed an information criterion for parameters under an order restriction (ORIC). In contrast to the AIC, the ORIC method is particularly suitable for detecting the configuration of the isotonic means under an order restriction.

In contrast to the AIC and BIC, which penalize on the number of parameters in the model and/or on the sample size, the ORIC takes into account the level probability for the number of parameters under each order restricted model in set \( R \).

The ORIC proposed by Anraku (1999) is given by
\[ ORIC = -2\log(\ell\theta|D) + \sum_{i=0}^K iP(i, K, w_i), \]
with \( P(i, K, w_i) \) denoting the level probability that given \( K \) doses under \( H_0 \) the isotonic regression will result in \( i \) unique isotonic means (Robertson et al. 1988), \( w_i = n_i/\sigma_i \) where \( n_i \) is the number of arrays at dose \( i \) \((i = 1, \ldots, K)\), and \( \sigma_i \) is the variance at dose \( i \). Note that \( \sum_{i=1}^K P(i, K, w_i) = 1 \).

For the case that \( w_i = \cdots = w_K \), or \( K = 2 \), it follows that \( ORIC = -2\log(\ell\theta|D) + \sum 1/i \) (Robertson et al. 1988).

The posterior model probabilities (with non-informative priors) are given by
\[ P_{OR}(g_i|D) = \frac{\exp\left(-\frac{1}{2} \Delta ORIC_i\right)}{\sum_{r=1}^R \exp\left(-\frac{1}{2} \Delta ORIC_r\right)}. \] (10.11)

Under simple order alternatives, i.e., \( H_{1P}^U \) or \( H_{1D}^D \) (see (8.3) and (8.4), respectively), in the setting of four doses (control and three higher doses) and equal number of arrays per dose, the level probabilities \( P(i, K, w_i) \), given by Robertson et al. (1988), are equal to 0.25, 0.45833, 0.25, and 0.04167 for \( i = 0, 1, 2, \) and 3, respectively. The
smallest value of the ORIC for the model in the set of R possible models indicates the best configuration of the parameters under order restriction.

Table 10.2 lists the value of the penalty used by the AIC, BIC, and ORIC. Note that the relative magnitude of penalty used by the ORIC for models \( g_1, g_2, g_3 \) (two parameters) and models \( g_4, g_5, g_6 \) (three parameters) is larger as compared to that used for model \( g_7 \) (four parameters). Thus, due to the small penalty difference between models \( g_4, g_5, g_6 \), and model \( g_7 \), the latter is more likely classified as the best model in the set of \( R \) possible models.

Table 10.2: Comparison of penalty values for the AIC, BIC and ORIC for models \( g_1 \) to \( g_7 \); \( n \) is the total number of arrays.

<table>
<thead>
<tr>
<th>models</th>
<th>parameters</th>
<th>AIC</th>
<th>BIC</th>
<th>ORIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>( g_1, g_2, g_3 )</td>
<td>2</td>
<td>( 2 \times 2 \times 2 \times \log(n) )</td>
<td>( 2 \times 1.66666 )</td>
<td></td>
</tr>
<tr>
<td>( g_4, g_5, g_6 )</td>
<td>3</td>
<td>( 2 \times 3 \times 3 \times \log(n) )</td>
<td>( 2 \times 1.91666 )</td>
<td></td>
</tr>
<tr>
<td>( g_7 )</td>
<td>4</td>
<td>( 2 \times 4 \times 4 \times \log(n) )</td>
<td>( 2 \times 2.08334 )</td>
<td></td>
</tr>
</tbody>
</table>

So far we discussed the model selection using the information theory to classify the dose-response curve shape and identify the MED level. An alternative approach to determine the MED is to use a hypothesis testing procedure, such as Williams’ (1971, 1972) or Marcus’ (1976) procedure, that identify the MED based on hypothesis testing in a step-down fashion. However, identification of the MED using these procedures does not imply the shape of the dose-response curve. On the other hand, model selection procedures based on information theory can address two objectives at the same time - the determination of the MED and the classification of trends. Once the best model from the candidate set is selected, one can identify the MED from the selected model.

Note that the methods discussed above are conditional on the result of testing for a monotonic trend, that is essential. The model selection procedure is applied to those genes, for which there is an evidence of a monotone relationship between gene-expression and dose. Hence, the best monotone model must be in the model set, as required. Moreover, the initial testing step reduces the computational time for the model selection, as the direction of the trend (upward or downward trend) is known from the initial step.
10.4 Application to the Data

10.4.1 Classification of the Trends

Based on the 3499 genes found to be significant in the initial inference step (by using the $E_{01}^2$ test), the classification of trends for these genes is to be obtained. Using the four information criteria (likelihood, $AIC$, $BIC$, and $ORIC$) the genes are classified into the seven curve shapes shown in Figure 10.1 (see Table 10.3). Using the likelihood-based posterior probabilities defined in (10.1), 1710 genes (48.85%) are classified as $g_7$ - the isotonic regression model with four parameters. As shown in Table 10.3, when the $AIC$ and $BIC$ criteria are used to calculate the posterior probabilities, the number of genes that are classified as $g_1$ (isotonic regression with two parameters) increases from 344 for the likelihood to 1528 and 1648, respectively. The same pattern is observed for models $g_2$ and $g_3$ (both are isotonic regression models with two parameters). For the $ORIC$, the number of genes classified to models from $g_1$ to $g_6$ decreases as compared to the $AIC$ and $BIC$. On the other hand, 816 genes are classified as $g_7$.

Table 10.3: Classification of genes into dose-response trends: the number of genes classified to each model.

<table>
<thead>
<tr>
<th>Model</th>
<th>Likelihood</th>
<th>$AIC$</th>
<th>$BIC$</th>
<th>$ORIC$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$g_1$</td>
<td>344</td>
<td>1528</td>
<td>1648</td>
<td>1348</td>
</tr>
<tr>
<td>$g_2$</td>
<td>25</td>
<td>307</td>
<td>369</td>
<td>221</td>
</tr>
<tr>
<td>$g_3$</td>
<td>14</td>
<td>106</td>
<td>126</td>
<td>86</td>
</tr>
<tr>
<td>$g_4$</td>
<td>343</td>
<td>370</td>
<td>337</td>
<td>253</td>
</tr>
<tr>
<td>$g_5$</td>
<td>885</td>
<td>823</td>
<td>715</td>
<td>655</td>
</tr>
<tr>
<td>$g_6$</td>
<td>178</td>
<td>170</td>
<td>149</td>
<td>120</td>
</tr>
<tr>
<td>$g_7$</td>
<td>1710</td>
<td>195</td>
<td>155</td>
<td>816</td>
</tr>
</tbody>
</table>

Figure 10.3 shows the data for gene 3467. Based on the likelihood and $ORIC$, the best model for the gene is $g_7$ (solid line). Using the $AIC$, the gene is classified as $g_6$ (dotted line). For both models, the MED is estimated to be the second dose level. Using the $BIC$, the model is further reduced and the gene is classified as $g_2$ (dashed line), and the MED level is estimated to be the third dose level.
Figure 10.3: Classification of trend for gene 3467: the best models according to the likelihood, AIC, BIC, and ORIC. Solid line is the isotonic regression model $g_7$ (four parameters) selected by the likelihood and ORIC, dotted line is $g_6$ (three parameters) selected by the AIC, and dashed line is $g_5$ (two parameters) selected by the BIC.

Note that, although the three models ($g_1$, $g_2$, and $g_3$) have the same number of parameters, the effective dose levels are not the same. The MED for model $g_1$ is the highest dose level, while the effective dose levels for $g_2$ and $g_3$ are the third and the second dose levels, respectively. In general, the AIC and BIC criteria favor the two-parameter models as compared to the three-parameter models $g_4$, $g_5$, $g_6$, and as compared to the four-parameter model $g_7$, which is favored by the likelihood.
Table 10.3 shows that the \textit{AIC} and \textit{BIC} classify most of genes as \(g_1\) and \(g_5\), the \textit{ORIC} classifies most of genes as \(g_1\) and \(g_7\), while the likelihood criterion classifies genes mostly as \(g_7\).

### 10.5 Simulation Study

Three simulation studies are conducted to investigate the performance of the model selection procedure. In particular, two questions are of interest: (1) is the initial inference step of filtering out genes with monotonic trends necessary? (2) when the \textit{AIC}, \textit{BIC}, and \textit{ORIC} are used to reduce the complexity of the models, does the selection favor some models over others? In the first study, we compare the proposed model selection methods for classification of dose response curve shapes with and without the initial step of testing monotonic trends. In the second study, we introduce the correlation for gene-expression and compare the results of classification with the first results of the simulation study. In the third study we evaluate whether the model selection criteria are sensitive to the shape of the seven possible dose-response relationships.

#### 10.5.1 The Influence of the Initial Inference Step

In this study, we investigate two different approaches for classification of genes into the seven models under ordered alternative, under the null model and the non-monotonic models. The first approach is to classify all the genes into different models using the four information criteria. Hence, the classification is done without an initial inference step. The second approach is divided into two steps. In the first step the null hypothesis (8.2) is tested against the ordered alternatives (8.3) or (8.4). After the initial step of filtering out non-significant genes, in the second step, monotonic genes are further classified into one of the seven monotonic models. The first approach avoids the problem of hypothesis testing and multiplicity. The second approach zooms on at the genes of interest with monotonic trends, and controls the FDR.

The performance of the two approaches is evaluated by the misclassification error and the correct classification for each model. In the simulations, 100 data sets with 12 microarrays (three arrays for each of four dose levels) and 17,000 genes on each array, are randomly generated. Among all the genes, 250 genes are generated under each of the seven models with an increasing and decreasing trend, what accounts for
3500 genes. Another 3500 genes are generated under a non-monotonic trend. Finally, the remaining 10,000 genes are generated under the null model. The sample means used for simulation are given in Table 10.4, with scale factor $\psi = 1$ and the variance $\sigma^2 = 0.01$.

Table 10.4: Means for simulated data.

<table>
<thead>
<tr>
<th>$\mu_1$</th>
<th>$\mu_2$</th>
<th>$\mu_3$</th>
<th>$\mu_4$</th>
<th>$\psi = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$g_1$:</td>
<td>(1 1 1 2)</td>
<td>$\times 2\psi/\sqrt{3}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$g_2$:</td>
<td>(1 1 2 2)</td>
<td>$\times \psi$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$g_3$:</td>
<td>(1 2 2 2)</td>
<td>$\times 2\psi/\sqrt{3}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$g_4$:</td>
<td>(1 2 2 3)</td>
<td>$\times \psi/\sqrt{2}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$g_5$:</td>
<td>(1 1 2 3)</td>
<td>$\times 2\psi/\sqrt{11}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$g_6$:</td>
<td>(1 2 3 3)</td>
<td>$\times 2\psi/\sqrt{11}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$g_7$:</td>
<td>(1 2 3.5 4)</td>
<td>$\times \psi/\sqrt{5}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>null:</td>
<td>(0 0 0 0)</td>
<td>$\psi$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>non-mon:</td>
<td>(1 -1 2 -2)</td>
<td>$\times \psi/\sqrt{5}$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The results of the simulations (the correct classification and misclassification error rates) are summarized in Table 10.5 (the first approach: classification without an initial inference step) and Table 10.6 (the second approach: an initial inference step followed by the classification step).

The misclassification and correct classification rates are computed as follows. Suppose $N_{X_i}$ genes are classified as model $g_i$. Among these $N_{X_i}$ genes, $N_{Y_i}$ genes are truly generated under $g_i$. The proportion of correct classification for model $g_i$ is thus equal to $N_{Y_i}/250$, for the non-monotonic model the proportion is equal to $N_{Y_i}/3500$, and for the null model it is equal to $N_{Y_i}/10000$. The proportion of misclassification is equal to $(N_{X_i} - N_{Y_i})/N_{X_i}$. This error rate is defined as the relative proportion, mimicking the concept of the FDR: the proportion of misclassifications (false positives) among the number of classifications (total discoveries). From Tables 10.5 and 10.6, it is easy to note that the correct classification rate obtained for the two approaches is quite similar using the $AIC$ and $BIC$, while the likelihood criterion under the second approach yields a higher correct classification rate than under the first approach. This is because the testing procedure filters out non-significant genes and controls the
FDR in testing. On the other hand, the misclassification rates for the seven models under the order alternatives in the second approach are much lower than those in the first approach using the $AIC$, $BIC$, and $ORIC$. The initial testing step in the second approach selects the genes, for which the null hypothesis was rejected in favor of the monotonic ordered alternatives, and allows them to be correctly classified into one of the seven models. This greatly improves the misclassification error. On the other hand, in the first approach, a large number of non-monotonic genes are misclassified into one of the seven models under the order restriction after penalization using the $AIC$, $BIC$, and $ORIC$, what results in a high misclassification error. Thus, the second approach is recommended to ensure a correct classification of genes of interest.

We also observe that for $g_7$, both the misclassification error and the correct classification proportion using the $ORIC$ are higher than those using the $AIC$ and $BIC$. However, the opposite results can be observed for models $g_4$, $g_5$, and $g_6$. This is because the penalty used by the $ORIC$ favors $g_7$, as noted in Section 10.3. As the number of genes classified as $g_7$ ($N_{X7}$) increases, the number of correct classification ($N_{Y7}$) increases as expected. Thus, the proportion of correct classification $N_{Y7}/250$ also increases. On the other hand, the misclassification error $(N_{X7} - N_{Y7})/N_{X7}$ increases at the same time, due to an increase in $N_{Y7}$. 

Table 10.5: Misclassification and correct classification rates. First approach: classification without initial inference step. L=Likelihood, A=AIC, B=BIC, O=ORIC, NA=Not Available.

<table>
<thead>
<tr>
<th>model</th>
<th>Error</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L</td>
<td>A</td>
</tr>
<tr>
<td>$g_1$</td>
<td>NA</td>
<td>0.7113</td>
</tr>
<tr>
<td>$g_2$</td>
<td>NA</td>
<td>0.6968</td>
</tr>
<tr>
<td>$g_3$</td>
<td>NA</td>
<td>0.7125</td>
</tr>
<tr>
<td>$g_4$</td>
<td>0.9421</td>
<td>0.2780</td>
</tr>
<tr>
<td>$g_5$</td>
<td>0.8847</td>
<td>0.0266</td>
</tr>
<tr>
<td>$g_6$</td>
<td>0.8674</td>
<td>0.2378</td>
</tr>
<tr>
<td>$g_7$</td>
<td>0.7884</td>
<td>0.0710</td>
</tr>
<tr>
<td>$g_1$</td>
<td>NA</td>
<td>0.7118</td>
</tr>
<tr>
<td>$g_2$</td>
<td>NA</td>
<td>0.6962</td>
</tr>
<tr>
<td>$g_3$</td>
<td>NA</td>
<td>0.7103</td>
</tr>
<tr>
<td>$g_4$</td>
<td>0.8835</td>
<td>0.098</td>
</tr>
<tr>
<td>$g_5$</td>
<td>0.8799</td>
<td>0.0273</td>
</tr>
<tr>
<td>$g_6$</td>
<td>0.8495</td>
<td>0.2415</td>
</tr>
<tr>
<td>$g_7$</td>
<td>0.7882</td>
<td>0.0713</td>
</tr>
<tr>
<td>null</td>
<td>NA</td>
<td>0</td>
</tr>
<tr>
<td>non.mon</td>
<td>0.7606</td>
<td>0.0398</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>model</th>
<th>L</th>
<th>A</th>
<th>B</th>
<th>Error</th>
<th>O</th>
<th>L</th>
<th>A</th>
<th>B</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>g₁</td>
<td>0.022</td>
<td>0.0291</td>
<td>0.0320</td>
<td>0.0263</td>
<td>0.3369</td>
<td>0.9999</td>
<td>1</td>
<td>0.9995</td>
<td></td>
</tr>
<tr>
<td>g₂</td>
<td>0.0272</td>
<td>0.0310</td>
<td>0.0356</td>
<td>0.0261</td>
<td>0.2505</td>
<td>0.9960</td>
<td>0.99810</td>
<td>0.9885</td>
<td></td>
</tr>
<tr>
<td>U</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>g₃</td>
<td>0.0253</td>
<td>0.0300</td>
<td>0.0327</td>
<td>0.0265</td>
<td>0.333</td>
<td>0.9999</td>
<td>0.9999</td>
<td>0.9993</td>
<td></td>
</tr>
<tr>
<td>p</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>g₄</td>
<td>0.6112</td>
<td>0.2188</td>
<td>0.2160</td>
<td>0.2191</td>
<td>0.4033</td>
<td>0.7842</td>
<td>0.7903</td>
<td>0.6741</td>
<td></td>
</tr>
<tr>
<td>g₅</td>
<td>0.5093</td>
<td>0.0149</td>
<td>0.0108</td>
<td>0.0182</td>
<td>0.5041</td>
<td>0.9834</td>
<td>0.9895</td>
<td>0.8513</td>
<td></td>
</tr>
<tr>
<td>g₆</td>
<td>0.5106</td>
<td>0.2307</td>
<td>0.2638</td>
<td>0.0661</td>
<td>0.4995</td>
<td>0.9830</td>
<td>0.9890</td>
<td>0.8469</td>
<td></td>
</tr>
<tr>
<td>g₇</td>
<td>0.6796</td>
<td>0.0710</td>
<td>0.0503</td>
<td>0.3350</td>
<td>0.965</td>
<td>0.7176</td>
<td>0.6526</td>
<td>0.9545</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>g₁</td>
<td>0.0255</td>
<td>0.0456</td>
<td>0.0480</td>
<td>0.0427</td>
<td>0.3336</td>
<td>0.9999</td>
<td>1</td>
<td>0.9992</td>
<td></td>
</tr>
<tr>
<td>g₂</td>
<td>0.0216</td>
<td>0.0289</td>
<td>0.0341</td>
<td>0.0236</td>
<td>0.2507</td>
<td>0.9966</td>
<td>0.9984</td>
<td>0.9890</td>
<td></td>
</tr>
<tr>
<td>o</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>g₃</td>
<td>0.0270</td>
<td>0.0300</td>
<td>0.0322</td>
<td>0.0269</td>
<td>0.3399</td>
<td>0.9999</td>
<td>0.9999</td>
<td>0.9996</td>
<td></td>
</tr>
<tr>
<td>w</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>g₄</td>
<td>0.5220</td>
<td>0.0238</td>
<td>0.0210</td>
<td>0.0259</td>
<td>0.4972</td>
<td>0.9792</td>
<td>0.9861</td>
<td>0.8352</td>
<td></td>
</tr>
<tr>
<td>g₅</td>
<td>0.5141</td>
<td>0.0144</td>
<td>0.0113</td>
<td>0.0179</td>
<td>0.4987</td>
<td>0.9825</td>
<td>0.9886</td>
<td>0.8447</td>
<td></td>
</tr>
<tr>
<td>n</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>g₆</td>
<td>0.5097</td>
<td>0.2338</td>
<td>0.2649</td>
<td>0.0660</td>
<td>0.5025</td>
<td>0.985</td>
<td>0.9900</td>
<td>0.8492</td>
<td></td>
</tr>
<tr>
<td>g₇</td>
<td>0.6798</td>
<td>0.0713</td>
<td>0.0523</td>
<td>0.3991</td>
<td>0.9970</td>
<td>0.7102</td>
<td>0.6483</td>
<td>0.9544</td>
<td></td>
</tr>
<tr>
<td>null</td>
<td>NA</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>NA</td>
<td>0.8796</td>
<td>0.9390</td>
<td>0.8796</td>
<td></td>
</tr>
<tr>
<td>non.mon</td>
<td>0.7396</td>
<td>0.2449</td>
<td>0.1337</td>
<td>0.2449</td>
<td>0.9989</td>
<td>0.9988</td>
<td>0.9988</td>
<td>0.9988</td>
<td></td>
</tr>
</tbody>
</table>
10.5.2 Correlated Data

In this simulation study, we introduce the correlation between gene-expression levels for different genes. The same settings as described in the previous section is used to generate 100 datasets. Additionally, in each dataset, for 250 genes under each model ($g_1$ to $g_7$, the null and non-monotonic model), i.e., for the total of 2250 genes, a correlation of 0.5 (or 0.9) is assumed at each dose level.

The first approach, (i.e., the classification without the initial inference step, Table 10.7) results in a higher misclassification error rate and a similar correct classification rate as compared to the second approach (i.e., the initial inference step followed by classification step, Table 10.8) using $AIC$, $BIC$, and $ORIC$. Moreover, a similar pattern in the misclassification error and in the correct classification rate can be observed as compared to the previous study. However, the misclassification error rate increases slightly for models $g_1$–$g_3$ (two parameter models), and substantially for models $g_4$–$g_7$ (three and four parameter models). Similarly, the correct classification rate decreases slightly for models $g_1$–$g_3$, and more substantially for models $g_4$–$g_7$. In general, we reach the same conclusion as in the previous simulation study that the performance of the second approach is satisfactory in terms of both the misclassification rates and the correct classification rates.

In Appendix B, we also include the results for correlation equal to 0.9 in Table B.1 (the first approach) and Table B.2 (the second approach).
Table 10.7: Misclassification and correct classification rates. First approach: classification without the initial inference step. The setting induces the correlation (0.5) for gene-expression of 250 genes under each model and each dose level. L=Likelihood, A=AIC, B=BIC, O=ORIC, NA=Not Available.

<table>
<thead>
<tr>
<th></th>
<th>L</th>
<th>A</th>
<th>B</th>
<th>O</th>
<th></th>
<th>L</th>
<th>A</th>
<th>B</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>g1</td>
<td>NA</td>
<td>0.7287</td>
<td>0.6828</td>
<td>0.8677</td>
<td>NA</td>
<td>0.9611</td>
<td>0.9728</td>
<td>0.9456</td>
<td></td>
</tr>
<tr>
<td>g2</td>
<td>NA</td>
<td>0.7379</td>
<td>0.7053</td>
<td>0.8649</td>
<td>NA</td>
<td>0.9335</td>
<td>0.9504</td>
<td>0.9017</td>
<td></td>
</tr>
<tr>
<td>g3</td>
<td>NA</td>
<td>0.7287</td>
<td>0.6835</td>
<td>0.8677</td>
<td>NA</td>
<td>0.9583</td>
<td>0.9695</td>
<td>0.9409</td>
<td></td>
</tr>
<tr>
<td>g4</td>
<td>0.9193</td>
<td>0.3493</td>
<td>0.3142</td>
<td>0.3838</td>
<td>0.0016</td>
<td>0.7160</td>
<td>0.7125</td>
<td>0.6185</td>
<td></td>
</tr>
<tr>
<td>g5</td>
<td>0.8791</td>
<td>0.1546</td>
<td>0.1286</td>
<td>0.1547</td>
<td>0.0017</td>
<td>0.7581</td>
<td>0.7288</td>
<td>0.6737</td>
<td></td>
</tr>
<tr>
<td>g6</td>
<td>0.8396</td>
<td>0.3505</td>
<td>0.3570</td>
<td>0.2626</td>
<td>0.0020</td>
<td>0.7585</td>
<td>0.7295</td>
<td>0.6691</td>
<td></td>
</tr>
<tr>
<td>g7</td>
<td>0.8072</td>
<td>0.2044</td>
<td>0.1682</td>
<td>0.4918</td>
<td>0.8826</td>
<td>0.4445</td>
<td>0.3911</td>
<td>0.7254</td>
<td></td>
</tr>
<tr>
<td>g1</td>
<td>NA</td>
<td>0.7400</td>
<td>0.7590</td>
<td>0.9374</td>
<td>NA</td>
<td>0.9578</td>
<td>0.9680</td>
<td>0.9425</td>
<td></td>
</tr>
<tr>
<td>g2</td>
<td>NA</td>
<td>0.7387</td>
<td>0.7067</td>
<td>0.8649</td>
<td>NA</td>
<td>0.9334</td>
<td>0.9498</td>
<td>0.9030</td>
<td></td>
</tr>
<tr>
<td>g3</td>
<td>NA</td>
<td>0.7265</td>
<td>0.6805</td>
<td>0.8669</td>
<td>NA</td>
<td>0.9613</td>
<td>0.9732</td>
<td>0.9461</td>
<td></td>
</tr>
<tr>
<td>g4</td>
<td>0.8906</td>
<td>0.2094</td>
<td>0.1736</td>
<td>0.2636</td>
<td>0.0022</td>
<td>0.8605</td>
<td>0.8532</td>
<td>0.7440</td>
<td></td>
</tr>
<tr>
<td>g5</td>
<td>0.8794</td>
<td>0.1499</td>
<td>0.1292</td>
<td>0.1510</td>
<td>0.0018</td>
<td>0.7597</td>
<td>0.7310</td>
<td>0.6715</td>
<td></td>
</tr>
<tr>
<td>g6</td>
<td>0.8695</td>
<td>0.3527</td>
<td>0.3573</td>
<td>0.2671</td>
<td>0.0019</td>
<td>0.7646</td>
<td>0.7340</td>
<td>0.6782</td>
<td></td>
</tr>
<tr>
<td>g7</td>
<td>0.8082</td>
<td>0.2094</td>
<td>0.1689</td>
<td>0.4942</td>
<td>0.8782</td>
<td>0.4422</td>
<td>0.3898</td>
<td>0.7195</td>
<td></td>
</tr>
</tbody>
</table>

null   | NA | 0.0000 | 0.0001 | 0.0000 | NA | 0.6164 | 0.7053 | 0.0954 |
non.mon| 0.7617 | 0.0436 | 0.0090 | 0.0096 | 1.0000 | 0.9875 | 0.9292 | 0.4311 |
Table 10.8: Misclassification and correct classification rates. Second approach: the initial inference step followed by the classification step. The setting induces the correlation (0.5) for gene-expression of 250 genes under each model and each dose level. L=Likelihood, A=AIC, B=BIC, O=ORIC, NA=Not Available.

<table>
<thead>
<tr>
<th>Error</th>
<th>L</th>
<th>A</th>
<th>B</th>
<th>O</th>
<th>L</th>
<th>A</th>
<th>B</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>g1</td>
<td>0.1369</td>
<td>0.1775</td>
<td>0.2158</td>
<td>0.1551</td>
<td>0.3336</td>
<td>0.9256</td>
<td>0.9332</td>
<td>0.9181</td>
</tr>
<tr>
<td>g2</td>
<td>0.1971</td>
<td>0.2987</td>
<td>0.3462</td>
<td>0.2585</td>
<td>0.2151</td>
<td>0.9038</td>
<td>0.9147</td>
<td>0.8850</td>
</tr>
<tr>
<td>g3</td>
<td>0.1364</td>
<td>0.1662</td>
<td>0.1893</td>
<td>0.1337</td>
<td>0.3271</td>
<td>0.9308</td>
<td>0.9462</td>
<td>0.9232</td>
</tr>
<tr>
<td>g4</td>
<td>0.6210</td>
<td>0.2632</td>
<td>0.2460</td>
<td>0.2533</td>
<td>0.3551</td>
<td>0.7123</td>
<td>0.7164</td>
<td>0.6065</td>
</tr>
<tr>
<td>g5</td>
<td>0.5189</td>
<td>0.1000</td>
<td>0.0845</td>
<td>0.0856</td>
<td>0.4599</td>
<td>0.7394</td>
<td>0.6852</td>
<td>0.6422</td>
</tr>
<tr>
<td>g6</td>
<td>0.5376</td>
<td>0.3383</td>
<td>0.3722</td>
<td>0.2174</td>
<td>0.5606</td>
<td>0.7862</td>
<td>0.7267</td>
<td>0.7530</td>
</tr>
<tr>
<td>g7</td>
<td>0.7104</td>
<td>0.2408</td>
<td>0.2049</td>
<td>0.4673</td>
<td>0.8523</td>
<td>0.4373</td>
<td>0.3741</td>
<td>0.6712</td>
</tr>
<tr>
<td>g8</td>
<td>0.2060</td>
<td>0.7592</td>
<td>0.7586</td>
<td>0.7607</td>
<td>0.3419</td>
<td>0.9417</td>
<td>0.9528</td>
<td>0.9229</td>
</tr>
<tr>
<td>g9</td>
<td>0.1919</td>
<td>0.2944</td>
<td>0.3308</td>
<td>0.2505</td>
<td>0.2109</td>
<td>0.8839</td>
<td>0.9058</td>
<td>0.8325</td>
</tr>
<tr>
<td>g10</td>
<td>0.1083</td>
<td>0.1535</td>
<td>0.1595</td>
<td>0.1345</td>
<td>0.3907</td>
<td>0.9531</td>
<td>0.9531</td>
<td>0.9340</td>
</tr>
<tr>
<td>g11</td>
<td>0.8707</td>
<td>0.1310</td>
<td>0.1352</td>
<td>0.0943</td>
<td>0.4915</td>
<td>0.8529</td>
<td>0.8570</td>
<td>0.7116</td>
</tr>
<tr>
<td>g12</td>
<td>0.5530</td>
<td>0.1318</td>
<td>0.1021</td>
<td>0.1047</td>
<td>0.4297</td>
<td>0.7608</td>
<td>0.7429</td>
<td>0.6315</td>
</tr>
<tr>
<td>g13</td>
<td>0.5397</td>
<td>0.2889</td>
<td>0.3279</td>
<td>0.2026</td>
<td>0.4606</td>
<td>0.7621</td>
<td>0.7508</td>
<td>0.6342</td>
</tr>
<tr>
<td>g14</td>
<td>0.6972</td>
<td>0.2038</td>
<td>0.1575</td>
<td>0.4860</td>
<td>0.8880</td>
<td>0.4847</td>
<td>0.3861</td>
<td>0.7679</td>
</tr>
<tr>
<td>null</td>
<td>NA</td>
<td>0.0000</td>
<td>0.0002</td>
<td>0.0000</td>
<td>NA</td>
<td>0.8660</td>
<td>0.9244</td>
<td>0.8660</td>
</tr>
<tr>
<td>non-mon</td>
<td>0.7802</td>
<td>0.3170</td>
<td>0.2038</td>
<td>0.3170</td>
<td>0.8020</td>
<td>0.8020</td>
<td>0.8020</td>
<td>0.8020</td>
</tr>
</tbody>
</table>
10.5.3 Evaluation of the Four Information Criteria

In the third simulation study, 100 data sets are generated, each with 12 microarrays (three arrays for each of four dose levels) containing 5000 genes. The proportion of genes generated under different models follows the results obtained for the $AIC$ and shown in Table 10.3. The isotonic means at different dose levels are assumed as in Table 10.4 with $\psi=1$ and $\sigma^2 = 0.01$. The results using the four information criteria are presented in Table 10.9. As we observe, except of model $g_7$, the classification based on the $AIC$ and $BIC$ achieves better results compared to the classification based on the likelihood. The former reduce not only the misclassification error, but also its standard error. However, as compared to the $AIC$ and $BIC$, the $ORIC$ achieves better classification for model $g_7$, while it performs worse for models $g_1$-$g_6$.

For example, 2184 genes (i.e., 43% of 5000 genes in total in each data set) are generated under $g_1$. Using the likelihood criterion for classification, 25% of these genes are misclassified as $g_4$, 24.87% are misclassified as $g_5$, and 16.5% as $g_7$. The total misclassification error rate (number of correct classifications/2184) is 66.46% (=25%+24.87%+16.5%) with the standard error 0.0105. For the $AIC$ and $BIC$, the error rates are reduced to 0.01% and 0.004%, respectively. Similar patterns are observed for all the models except of $g_7$. For model $g_7$, the misclassification is due to the erroneous classification to model $g_6$ that is similar to $g_7$.

To enhance the distinction between $g_6$ and $g_7$, we modified the isotonic means for model $g_7$ to $(1, 2, 3, 5) \times \psi/\sqrt{5}$, that makes them substantially different from those for model $g_6$. After the modification, the results of this simulation for models $g_1$-$g_6$ are the same as shown in Table 10.9. The results for model $g_7$ are summarized in the last row of Table 10.9 (“Mod $g_7$”). It shows a substantially improved performance for model $g_7$ using the $AIC$ and $BIC$. The misclassification errors are 0.0123 and 0.0242 using the $AIC$ and $BIC$, respectively. Thus, when model $g_7$ is substantially different from the other models, the $AIC$ and $BIC$ perform as good as the likelihood.
Table 10.9: Simulation results on misclassification rates based on the likelihood (L), AIC (A), BIC (B) and ORIC (O). Misclassification error in % (standard error).

<table>
<thead>
<tr>
<th>True Model</th>
<th>Misclassification Models</th>
<th>L</th>
<th>A</th>
<th>B</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>$g_1$</td>
<td>Total error</td>
<td>66.46</td>
<td>0.01</td>
<td>0.00</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.0105)</td>
<td>(0.0002)</td>
<td>(0.0001)</td>
<td>(0.0005)</td>
</tr>
<tr>
<td>$g_4$</td>
<td></td>
<td>25.10</td>
<td>0.00</td>
<td>0.00</td>
<td>0.02</td>
</tr>
<tr>
<td>$g_5$</td>
<td></td>
<td>24.87</td>
<td>0.01</td>
<td>0.00</td>
<td>0.03</td>
</tr>
<tr>
<td>$g_7$</td>
<td></td>
<td>16.50</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>$g_2$</td>
<td>Total error</td>
<td>74.69</td>
<td>0.36</td>
<td>0.14</td>
<td>1.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.0212)</td>
<td>(0.0027)</td>
<td>(0.0018)</td>
<td>(0.005)</td>
</tr>
<tr>
<td>$g_5$</td>
<td></td>
<td>25.10</td>
<td>0.20</td>
<td>0.06</td>
<td>0.37</td>
</tr>
<tr>
<td>$g_6$</td>
<td></td>
<td>24.93</td>
<td>0.15</td>
<td>0.07</td>
<td>0.30</td>
</tr>
<tr>
<td>$g_7$</td>
<td></td>
<td>24.66</td>
<td>0.01</td>
<td>0.00</td>
<td>0.44</td>
</tr>
<tr>
<td>$g_3$</td>
<td>Total error</td>
<td>66.72</td>
<td>0.01</td>
<td>0.00</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.0443)</td>
<td>(0.0009)</td>
<td>0.00</td>
<td>0.0013</td>
</tr>
<tr>
<td>$g_4$</td>
<td></td>
<td>25.07</td>
<td>0.01</td>
<td>0.00</td>
<td>0.01</td>
</tr>
<tr>
<td>$g_6$</td>
<td></td>
<td>25.19</td>
<td>0.01</td>
<td>0.00</td>
<td>0.01</td>
</tr>
<tr>
<td>$g_7$</td>
<td></td>
<td>16.46</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>$g_4$</td>
<td>Total error</td>
<td>49.59</td>
<td>2.07</td>
<td>1.38</td>
<td>15.83</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.0202)</td>
<td>(0.0058)</td>
<td>(0.0048)</td>
<td>(0.0145)</td>
</tr>
<tr>
<td>$g_5$</td>
<td></td>
<td>49.59</td>
<td>2.07</td>
<td>1.38</td>
<td>15.83</td>
</tr>
<tr>
<td>$g_5$</td>
<td>Total error</td>
<td>49.90</td>
<td>1.78</td>
<td>1.18</td>
<td>15.26</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.0154)</td>
<td>(0.0039)</td>
<td>(0.0033)</td>
<td>(0.0102)</td>
</tr>
<tr>
<td>$g_6$</td>
<td></td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>$g_7$</td>
<td></td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>$g_7$</td>
<td>Total error</td>
<td>50.06</td>
<td>1.67</td>
<td>1.05</td>
<td>15.42</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.0307)</td>
<td>(0.0098)</td>
<td>(0.0075)</td>
<td>(0.0241)</td>
</tr>
<tr>
<td>$g_2$</td>
<td></td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>$g_7$</td>
<td></td>
<td>50.06</td>
<td>1.67</td>
<td>1.05</td>
<td>15.42</td>
</tr>
<tr>
<td>$g_7$</td>
<td>Total error</td>
<td>0.29</td>
<td>28.08</td>
<td>34.61</td>
<td>4.41</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.0032)</td>
<td>(0.0261)</td>
<td>(0.025)</td>
<td>(0.0121)</td>
</tr>
<tr>
<td>$g_4$</td>
<td></td>
<td>0.00</td>
<td>0.11</td>
<td>0.26</td>
<td>0.01</td>
</tr>
<tr>
<td>$g_5$</td>
<td></td>
<td>0.00</td>
<td>0.06</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>$g_6$</td>
<td></td>
<td>0.29</td>
<td>27.92</td>
<td>34.28</td>
<td>4.40</td>
</tr>
<tr>
<td>Mod $g_7$</td>
<td>Total error</td>
<td>0.00</td>
<td>1.23</td>
<td>2.42</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.0032)</td>
<td>(0.0261)</td>
<td>(0.0250)</td>
<td>(0.0121)</td>
</tr>
<tr>
<td>$g_4$</td>
<td></td>
<td>0.00</td>
<td>0.61</td>
<td>1.22</td>
<td>0.00</td>
</tr>
<tr>
<td>$g_5$</td>
<td></td>
<td>0.00</td>
<td>0.61</td>
<td>1.20</td>
<td>0.00</td>
</tr>
</tbody>
</table>
10.6 Discussion

In this chapter, we have investigated the issue of classifying dose response curve shapes of the genes that tested positively for a monotonic trend. After the initial inference step, the set of all possible models is considered and a non-informative prior is given to these models. The likelihood, $AIC$, $BIC$, and $ORIC$ are used to assess the goodness-of-fit and complexity of the models. The Akaike weights, representing the posterior probabilities of the models, are calculated from the $AIC$ of the models; Bayesian posterior probabilities are obtained from the $BIC$; and the posterior probability obtained from the $ORIC$ takes order restriction into account. We have shown that using the likelihood to calculate the posterior probabilities of models leads to the selection of the isotonic regression model, that maximizes the likelihood under the order constraints. For the setting of four doses, the $ORIC$ tends to classify genes as model $g_7$ as opposed to models $g_4$, $g_5$, and $g_6$ that have one parameter less. And, as expected, using the $AIC$ and $BIC$ leads to a selection of simpler models with less parameters (compared with the model selection based on the likelihood). Moreover, we have shown that after a classification is made, the mean structure of the selected model can be used to estimate the minimum effective dose level.

From our simulation study, we note that the initial inference step of testing monotonic trends reduces the misclassification error by filtering out the genes with non-monotonic trends. Moreover, the classification based on the $AIC$ and $BIC$ leads to a smaller classification error when compared to the use of likelihood. The $ORIC$ performs better for model $g_7$, but worse for the other models in the set of $R$ possible models. In general, the simulated data without and with correlation of gene-expression levels yield similar results. However, as expected, the latter increases the misclassification error and decreases the correct classification rate.

The results presented in this chapter were published in Lin et al. (2007b).
Chapter 11

A Ratio Test in Dose-response Microarray Experiments: Inference

11.1 Introduction

In Chapter 8, we discussed the five test statistics in testing the null hypothesis of homogeneity of means against order restricted alternatives. A rejection of the null hypothesis implies a significant monotone trend of gene-expression with respect to doses. In the case study in Section 10.1 we showed that 3499 genes were found to be significant when the LRT (the $E_{01}^2$ test statistic) was used to test the null hypothesis (8.2) against the order restricted alternative (8.3) or (8.4) and the BH-FDR procedure was used to control the FDR. Among the significant genes, 1600 genes exhibited increasing trends and 1899 genes showed decreasing trends. However, the rejection of the null hypothesis (8.2) does not indicate the magnitude for which gene-expression increases or decreases. In this chapter, we wish to search for genes, for which the mean gene-expression increases by $100 \times \delta\%$ from the control for at least one (perhaps the highest) dose, where $\delta$ is a relative biological difference of interest.

Genes can be tested for a pre-specified $\delta$ and can be ranked based on the significance of the ratio test. Genes found significantly with a large $\delta$ are of higher interest to the biologists. In the case study, the experiment includes four EGF doses (three
doses and a control) for the control compound and three arrays per dose level. Figure 11.1 shows four genes, for which $\hat{\mu}_K^*/\hat{\mu}_0^* = 2.1$ (panels a and b) and $\hat{\mu}_K^*/\hat{\mu}_0^* = 0.48$ (panels c and d). The estimated mean expression for the control dose $\hat{\mu}_0^*$ is different for gene a and b (gene c and d), and clearly, the estimated mean expression for the highest dose $\hat{\mu}_K^*$ is different for gene a and b (gene c and d). The aim of the study is to identify genes, for which the increase (decrease) of the mean gene-expression from baseline is at least 110% regardless of mean gene-expression $\hat{\mu}_0^*$ at the zero dose of EGF.

The four t-type test statistics (Williams’, Marcus’, M and the modified M), considered in Chapter 8, can be used to test the mean expression levels between the highest and the control dose in a step-down fashion. In particular, Bretz (2006) proposed the use of Multiple Contrasts Tests (MCTs) for Williams’ and Marcus’ tests, when the question of interest in the dose-response study is formulated as $\mu(d_K)/\mu(d_0) = \gamma$, where $\gamma$ is the ratio of interest and is to be estimated. In this chapter, we consider the use of MCTs in the microarray setting.

The content of the chapter is organized as follows. In Section 11.2 we formulate the null and alternative hypotheses of the ratio test. In Section 11.3 we summarize the Pool Adjacent Violator Algorithm (PAVA) to obtain the order restricted mean response in isotonic regression and introduce the MCTs; and in particular, Williams’ and Marcus’ tests are discussed. In Section 11.4, we formulate the test of ratio $\mu(d_K)/\mu(d_0)$ in terms of the MCTs. In Section 11.5 we show the results of the application to the case study. We complete the chapter with a discussion in Section 11.6.

### 11.2 Ratio Tests

As mentioned in the Section 11.1, we mainly focus on the inference about the ratio of the mean gene-expression (under an order restriction) at the highest dose and the control in a dose-response microarray experiment. For this purpose we formulate one-sided hypotheses about an increasing and decreasing trend for each gene, respectively, as follows:

\[
H_0^{U} : \frac{\mu(d_K)}{\mu(d_0)} \leq (1 + \delta) \quad \text{vs.} \quad H_1^{U} : \frac{\mu(d_K)}{\mu(d_0)} > (1 + \delta) \quad \text{and} \quad \mu(d_0) \leq \ldots \leq \mu(d_K),
\]

\[
H_0^{D} : \frac{\mu(d_K)}{\mu(d_0)} \geq (1 - \delta) \quad \text{vs.} \quad H_1^{D} : \frac{\mu(d_K)}{\mu(d_0)} < (1 - \delta) \quad \text{and} \quad \mu(d_0) \geq \ldots \geq \mu(d_K).
\]  

(11.1)
Figure 11.1: Panels a and b: example of two genes with an increasing trend where \( \hat{\mu}_K / \hat{\mu}_0 = 2.1 \), but the estimated mean expression of the highest dose \( \hat{\mu}_K \) for gene in panel b is larger than that for gene in panel a. Panels c and d: example of two genes with a decreasing trend: \( \hat{\mu}_K / \hat{\mu}_0 = 0.48 \), but the estimated mean expression at the highest dose \( \hat{\mu}_K \) for gene in panel c is smaller than that for gene in panel d.

The relative mean difference \( \delta \) quantifies the biological importance of the trend with respect to the increasing doses. Note that for \( \delta = 0 \), the hypotheses in (11.1) correspond to the testing problem for an order restricted alternative. However, for \( \delta > 0 \), rejecting the null hypothesis in (11.1) implies that the mean gene-expression at the highest dose increases or decreases by \( 100 \times \delta \% \) compared to the mean expression.
for the control. Note that formulating the hypotheses as ratios implies that \( \delta \) is independent of the gene-expression for the control.

11.3 Computation of Isotonic Means and MCTs

In this section, we formulate the ratio test and introduce the expression of Williams’ and Marcus’ tests in terms of MCT proposed by Bretz (1999, 2006).

11.3.1 Pool-adjacent-violator-algorithm for the Computation of Isotonic Means

Any procedure for testing the null hypothesis of the homogeneity of means against an order restricted alternative requires the calculation of the isotonic means with respect to dose. In order to obtain the isotonic means, the mostly widely used technique is PAVA (Ayer et al. 1955, Barlow et al. 1972, Robertson et al. 1988). If sample means for neighboring doses are not in a monotonic restricted order, the method non-parametrically amalgamates the means, until the amalgamated means are completely ordered. The result of the algorithm can be linked to the following analytical expression using max-min formulas. Given \( n_0, n_1, \ldots, n_K \) observations (arrays) and sample means \( \bar{y}_0, \bar{y}_1, \ldots, \bar{y}_K \), at doses \( d_0, \ldots, d_K \), respectively, and assuming normally distributed data, the maximum likelihood estimates \( \hat{\mu}_i^* \), subject to the simple order restriction \( \mu(d_0) \leq \mu(d_1) \leq \cdots \leq \mu(d_K) \), are given by

\[
\hat{\mu}_i^* = \max_{0 \leq u \leq i \ \text{min} \ i \leq v \leq K} \frac{\sum_{i=u}^v n_i \bar{y}_i}{\sum_{i=u}^v n_i}, \quad i = 0, \ldots, K,
\]

where \( \bar{y}_i = \sum_{j=1}^{n_i} y_{ij} / n_i \) is the sample mean for dose \( i = 0, 1, \ldots, K \).

11.3.2 Single and Multiple Contrast Tests

The standardized statistic of a single contrast test (SCT) is defined as,

\[
T^{sc} = \frac{\sum_{i=0}^K c_i \bar{y}_i}{s \sqrt{\sum_{i=0}^K \frac{c_i^2}{n_i}}}
\]

where the weights \( c_i \) are the contrast coefficients fulfilling the condition \( \sum_{i=0}^K c_i = 0 \). \( T^{sc} \) is univariate central \( t \)-distributed with \( \nu = \sum_{i=1}^K (n_i - 1) \) degrees of freedom.
MCT was first described by Mukerjee et al. (1986). The test can be used to test for trend under order restrictions. The main reason for developing such a test was to obtain a test with a similar power behavior as the LRT, but easier to use.

MCT seeks to locate several contrast vectors in SCTs (Bretz 1999), as good as possible, in the set of all possible combinations of contrast coefficients. The resulting test, $T^{MC}$, is the maximum over $r$ of such single contrast tests $T^{SC}$ defined in (11.3):

$$ T^{MC} = \max\{T_1^{SC}, T_2^{SC}, \ldots, T_r^{SC}\}, $$

where $r$ is the number of SCT in MCT.

The contrast matrix for $T^{MC}$ can be written as

$$ C^{MC} = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_r \end{pmatrix} = \begin{pmatrix} c_{10} & c_{11} & \cdots & c_{1K} \\ c_{20} & c_{21} & \cdots & c_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ c_{r0} & c_{r1} & \cdots & c_{rK} \end{pmatrix}. $$

Each row of the contrast matrix $C^{MC}$ corresponds to a contrast vector $c$ of a SCT.

Assuming the normality of the response, the joint distribution of $T^{MC}$ is a central $r$-variate $t$-distribution with $\nu$ degrees of freedom and correlation matrix $R = \{\rho_{l,m}\}_{l,m}^r$, $l,m = 1, \ldots, r$ (Bretz 2006). The entries of $R$ consist of the correlation between each two of the $r$ contrast vectors, $c_l$ and $c_m$ say. Correlation $\rho_{l,m} = \text{Corr}(\sum_{i=0}^K c_{li}\bar{y}_i, \sum_{i=0}^K c_{mi}\bar{y}_i)$ under $H_0$ can be computed by

$$ \rho = \frac{\sum_{i=0}^K c_{li}c_{mi}/n_i}{\sqrt{(\sum_{i=0}^K c_{li}^2/n_i)(\sum_{i=0}^K c_{mi}^2/n_i)}}. $$

Multiple contrasts can be defined for many different test statistics. Somerville (1997, 1999) provides a list of several multiple comparison procedures (not necessarily designed for order restricted testing), which can be formulated as MCTs, such as Dunnett’s (1955, 1964) many-to-one approach, or Tukey’s (1953) all-pair comparisons. According to Robertson et al. (1988), the global LRT statistic can be expressed as the maximum of an infinite number of contrast statistics as well.

**MCT for an Order Restriction Alternative**

Under an order restricted alternative, the multiple contrast test can be constructed by decomposing the alternative $H_1^{Up}: \mu(d_0) \leq \mu(d_1) \leq \ldots \leq \mu(d_3)$ into $2^K - 1$
single contrasts (Bretz 1999). For the case of \( K = 3 \), the alternative hypothesis can be decomposed into all the possible scenarios as follows,

\[
H_{1}^{\text{Up}} = \bigcup_{i=1}^{7} H_{1(i)}^{\text{Up}}
\]

with

- \( H_{1(1)}^{\text{Up}} : \mu(d_0) = \mu(d_1) = \mu(d_2) < \mu(d_3) \),
- \( H_{1(2)}^{\text{Up}} : \mu(d_0) < \mu(d_1) = \mu(d_2) = \mu(d_3) \),
- \( H_{1(3)}^{\text{Up}} : \mu(d_0) = \mu(d_1) < \mu(d_2) = \mu(d_3) \),
- \( H_{1(4)}^{\text{Up}} : \mu(d_0) < \mu(d_1) = \mu(d_2) < \mu(d_3) \),
- \( H_{1(5)}^{\text{Up}} : \mu(d_0) < \mu(d_1) < \mu(d_2) < \mu(d_3) \),
- \( H_{1(6)}^{\text{Up}} : \mu(d_0) = \mu(d_1) < \mu(d_2) < \mu(d_3) \),
- \( H_{1(7)}^{\text{Up}} : \mu(d_0) < \mu(d_1) < \mu(d_2) = \mu(d_3) \).

Every true dose-response relation will fall in exactly one of these sub-alternatives (as far as under \( H_{1}^{\text{Up}} \)). Each of the sub-hypothesis can be seen as a single contrast test. The maximum of this multiple contrast test will determine the configuration of the isotonic means.

Note that these seven sub-alternatives are identical to the seven possible dose-response curve shapes given in Table 10.1. In Chapter 10, we classified genes into one of seven models (from \( g_1 \) to \( g_7 \)) based on the posterior model probability using different information criteria. In this chapter, by means of defining all possible sub-alternatives as a set of single contrast tests, the maximum test statistic value of MCT will find the “best” dose-response relationship. Therefore, multiple contrast test can be an alternative approach to identify the dose-response curve shape based on the parametric test.

In what follows, we discuss Williams- and Marcus-type MCTs (Bretz 1999). They are two established multiple comparison procedures, which can be represented as MCTs.
A Williams-type Multiple Contrast Test

In Williams’ procedure (Williams 1971, 1972), the amalgamated mean for the highest dose \( \mu(d_K) \) can be expressed as follows (Bretz 1999),

\[
\hat{\mu}^*_K = \max_{1 \leq u \leq K} \sum_{i=u}^{K} n_i \bar{y}_i / \sum_{i=u}^{K} n_i \\
= \max \left\{ \frac{n_1 \bar{y}_1 + n_2 \bar{y}_2 + \ldots + n_K \bar{y}_K}{n_1 + n_2 + \ldots + n_K} , \frac{n_K-1 \bar{y}_{K-1} + n_K \bar{y}_K}{n_{K-1} + n_K} \right\} \\
= \max \left\{ \begin{bmatrix} 0 & \ldots & 0 & 1 \\ 0 & \ldots & \frac{n_K-1}{n_{K-1} + n_K} & \frac{n_K}{n_{K-1} + n_K} \end{bmatrix} \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \bar{y}_K \end{bmatrix} \right\} \\
= \max C \bar{y}_{(-0)},
\]

where \( \bar{y}_{(-0)} = (\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_K)' \). Note that arrays of the control group are not included in the amalgamation process. In Williams’ test we have

\[
\hat{\mu}^*_K - \bar{y}_0 = \max \{ C \bar{y}_{(-0)} - \bar{y}_0 \bar{1} \} \\
= \max \{ [-1 + C] \bar{y} \} \\
= \max C^{Wl} \bar{y},
\]

where \( \bar{y} = (\bar{y}_0, \bar{y}_{(-0)})' \).

Hence, the Williams-type MCT matrix \( C^{Wl} \) can be expressed as

\[
C^{Wl} = \begin{bmatrix} -1 & 0 & \ldots & 0 & 1 \\ -1 & 0 & \ldots & \frac{n_K-1}{n_{K-1} + n_K} & \frac{n_K}{n_{K-1} + n_K} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -1 & \frac{n_K-1}{n_{K-1} + n_K} & \ldots & \frac{n_K}{n_{K-1} + n_K} & \frac{n_K}{n_{K-1} + n_K} \end{bmatrix}. \tag{11.6}
\]

The maximum contrast (11.5) consists of comparisons of the control with the weighted average over the last \( i \) treatments \( (i = 1, \ldots, K) \). The contrast matrix \( C^{MC} \) (11.5) of MCT is given by (11.6). For the experiment with three dose levels and the control, the multiple contrast matrix for Williams’ MCT is given by

\[
C^{Wl}_{K=3} = \begin{bmatrix} -1 & 0 & 0 & 1 \\ -1 & 0 & 1/2 & 1/2 \\ -1 & 1/3 & 1/3 & 1/3 \end{bmatrix}. \tag{11.7}
\]
Williams’ multiple contrast test takes the order restriction of the means into account through the contrast definition following the construction of the isotonic estimates (11.2). As pointed out by Bretz (1999), Williams’ t-type test statistic is not identical to Williams’ MCT because of different variance estimators. The variance in MCT adopts the completely studentized statistic by making use of the mean square error, i.e., \[ \sqrt{\sum_{i=0}^{K} \sum_{j=0}^{n_i} (y_{ij} - \bar{y}_i)^2 / (n - 1)}. \] Note that, Williams (1971, 1972) used the denominator of the t-test, i.e., \[ \sqrt{\sum_{i=0}^{K} \sum_{j=0}^{n_i} (y_{ij} - \bar{y}_i)^2 / (n - K)}. \]

### A Marcus-type Multiple Contrast Test

Similarly to the Williams-type MCT, Marcus-type MCT can be derived (Bretz 1999, 2006). The amalgamated mean for the control dose \( \mu(d_0) \) is given by

\[
\hat{\mu}_0^* = \min_{0 \leq v \leq K} \frac{\sum_{i=0}^{v} \sum_{j=0}^{n_i} n_i \bar{y}_i / n_i}{\sum_{i=0}^{v} n_i}.
\]

Therefore,

\[
\hat{\mu}_K^* - \hat{\mu}_0^* = \max \left\{ 0, \max_{0 \leq i,j \leq K} \left\{ \frac{n_j \bar{y}_j + \ldots + n_K \bar{y}_K}{n_j + \ldots + n_K} - \frac{n_0 \bar{y}_0 + \ldots + n_i \bar{y}_i}{n_0 + \ldots + n_i} \right\} \right\}. \tag{11.8}
\]

The difference \( \hat{\mu}_K^* - \hat{\mu}_0^* \) can be represented as a simple maximum term in (11.8). A natural way of applying the MCT principle to Marcus’ approach is to identify each element of (11.8) as a contrast. However, a closed form expression for the resulting contrast matrix \( C_{MCT} \) for Marcus’ test depends on the number of dose levels.

The multiple contrast matrix for Marcus’ MCT for \( K = 3 \) is given by

\[
C_{Mar}^{K=3} = \begin{pmatrix}
-1 & 0 & 0 & 1 \\
-1 & 0 & 1/2 & 1/2 \\
-1 & 1/3 & 1/3 & 1/3 \\
-1/2 & -1/2 & 0 & 1 \\
-1/3 & -1/3 & -1/3 & 1 \\
-1 & -1 & 1 & 1
\end{pmatrix}. \tag{11.9}
\]

It is easy to note that Williams contrasts \( C_{W}^{K=3} \), given in (11.7), are all contained in Marcus’ contrasts \( C_{Mar}^{K=3} \) (from the first to the third row), given in (11.9). These are suitable for concave dose-response shapes, as the higher dose groups are being
pooled and compared to the control. Two of Marcus’ contrasts (given by the fourth and fifth row of the $C_{MK}^{Mar}$) seem to be suitable to detect convex shapes, as they take the average over the lower doses.

The estimation of multiple contrast tests can be based on the multivariate $t$-distribution (Bretz 1999). Bretz proved the theorem that for $T^{MC}$ given in (11.4) with $C^{MC}$ given in (11.5), where $\text{rank}(C^{MC})=K$, the corresponding distribution function can be expressed as a $K$-variate central $t$-distribution.

11.4 Test of Ratios Using Multiple Contrasts

We have discussed the link between Williams’ and Marcus’ MCTs and the corresponding testing procedures. The link between the ratio test and the MCT can be established in a similar way. In this section, we first introduce the test statistics for multiple ratio tests in the form of MCT and the procedure for simultaneous estimation of ratios.

11.4.1 Multiple Ratios and Their Simultaneous Estimation

The multiple contrast test for ratios of linear combinations of means $\mu(d_i)$ for doses $d_i, i = 0, ..., K$, is given by

$$
\gamma_l = \frac{\sum_{i=0}^{K} c_{li} \mu(d_i)}{\sum_{i=0}^{K} d_{li} \mu(d_i)} = \frac{c_l^t \mu}{d_l^t \mu}, \quad l = 1, ..., r,
$$

(11.10)

where $c_l$ and $d_l$ are the vectors of contrast coefficients for the numerator and denominator of the $l$th ratio. We are interested in the vector of parameters $\gamma = (\gamma_1, ..., \gamma_r)$, where $r$ is the number of ratios and is equivalent to the number of SCTs in MCTs (11.4).

Marcus-type Multiple Ratio Test

Similarly as we defined the Marcus-type contrast matrix $C^{Mar}$ for mean difference $\hat{\mu}_K - \hat{\mu}_0$, the Marcus-type contrast matrix can be defined for the ratio test $\mu(d_K)/\mu(d_0)$ as follows:

$$
\frac{\mu(d_K)}{\mu(d_0)} = \max(\gamma_1, \gamma_2, \ldots, \gamma_r)
$$

(11.11)
For $K=3$, the contrast coefficients for the Marcus-type multiple contrasts, given in (11.9), can be obtained for the numerator and denominator of the ratio tests, respectively, by

$$C = \begin{pmatrix}
0 & 0 & 0 & 1 \\
0 & 0 & 1/2 & 1/2 \\
0 & 1/3 & 1/3 & 1/3 \\
0 & 0 & 0 & 1 \\
0 & 0 & 1 & 1 \\
0 & 0 & 1 & 1
\end{pmatrix} \quad \text{and} \quad D = \begin{pmatrix}
1 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 \\
1/2 & 1/2 & 0 & 0 \\
1/3 & 1/3 & 1/3 & 0 \\
1 & 1 & 0 & 0
\end{pmatrix}.$$ 

**Simultaneous Estimation of Multiple Ratios**

For Marcus-type ratio test, the null hypothesis about the ratio (11.11) can be tested using multiple contrast tests:

$$\frac{\mu(d_K)}{\mu(d_0)} = \max(\gamma_1, \gamma_2, \ldots, \gamma_r) = \max(\frac{c_1'\mu}{d_1'\mu}, \frac{c_2'\mu}{d_2'\mu}, \ldots, \frac{c_r'\mu}{d_r'\mu}).$$

Moreover, a reparametrization of the ratio-based hypotheses in (11.1) can be expressed as (Paulson 1942, Laster and Johnson 2003):

$$H_0^U: \mu(d_K) - \gamma \mu(d_0) \leq 0 \ \text{vs.} \ H_1^U: \mu(d_K) - \gamma \mu(d_0) > 0 \quad (11.12)$$

under $\mu(d_0) \leq \mu(d_1) \leq \ldots, \mu(d_K)$.

To test the hypotheses in (11.12) by replacing linear combinations of the means, the corresponding test statistics $L_l$ can be written as (Zerbe et al. 1982, Young et al. 1997, Dilba et al. 2005)

$$L_l = (c_l - \gamma_l d_l)'\hat{y}, \ l = 1, \ldots, r,$$

where the vectors $c_l = (c_{l0}, \ldots, c_{lK})'$ and $d_l = (d_{l0}, \ldots, d_{lK})'$ are vectors of real-value coefficients in the numerator and denominator of the contrasts associated with the $l$th ratio, defined in (11.10). Under the normality assumption and the null hypotheses $H_0^U$ and $H_0^D$, defined in (11.12), $L_l$ is distributed as $N(0, \sigma^2_{L_l})$, where

$$\sigma^2_{L_l} = Var(L_l) = \sigma^2(c_l - \gamma_l d_l)'M(c_l - \gamma_l d_l),$$

and $Var(\hat{\mu}) = \sigma^2 M$, with $M$ being a diagonal matrix containing the reciprocals of the sample sizes $n_i$. Let $s^2$ be the unbiased pooled variance estimator of the common
variance \( \sigma^2 \) based on \( \nu = \sum_{i=0}^{K}(n_i - 1) \) degrees of freedom. Hence, \( s_{L_i}^2 \) is the variance estimator of \( \sigma_{L_i}^2 \):

\[
s_{L_i}^2 = s^2(c_i - \gamma_id_i)'M(c_i - \gamma_id_i).
\]

Because \( s_{L_i} \) is distributed as \( (\sigma_{L_i}^2\nu^{-1}\chi^2(\nu))^{1/2} \) and is independent of \( L_i \), the test statistic

\[
t_l(\gamma_l) = L_l/s_{L_l}
\]

follows a \( t \)-distribution with \( \nu \) degrees of freedom (Dilba et al. 2005). Jointly, the vector \( t = (t_1, ..., t_r)' \) follows a central \( r \)-variate \( t \)-distribution with \( \nu \) degrees of freedom and correlation matrix \( R = [\rho_{ij}] \), where

\[
\rho_{ij} = \frac{(c_i - \gamma_id_i)'M(c_i - \gamma_id_i)}{\sqrt{(c_i - \gamma_id_i)'M(c_i - \gamma_id_i)}} \sqrt{(c_j - \gamma jd_j)'M(c_j - \gamma jd_j)}
\]

\( 1 \leq i \neq j \leq r \). Let \( Mt_r(\nu, R) \) denote a central \( r \)-variate \( t \)-distribution having \( \nu \) degrees of freedom and correlation matrix \( R \). Based on the theorem (3.1) discussed by Bretz (1999), the inference of multiple ratio contrast tests can be based on the \( K \)-variate \( t \)-distribution (Young 1997).

The multiple ratio parameters are estimated by plugging in the maximum likelihood estimates of the ratios \( \hat{\gamma} \) in the correlation matrix \( R \) (Dilba et al. 2005). This method is called the “plug-in” approach. Details on the method are given in Section 12.2.

As a result, the maximum value of the multiple estimated ratio parameters \( \hat{\gamma} \) is the estimator for \( \mu(d_K)/\mu(d_0) \). Thus, the corresponding smallest \( p \)-value of the MCT is obtained for each gene, which leads to a significance test for the ratio parameters.

### 11.5 Application to the Data

While testing the null and alternative hypotheses, defined in (11.1), the value of \( \delta \) can be chosen according to the biological interest. For genes with an increasing trend, the alternative will focus on the differences that are larger than \( 100 \times \delta\% \) of the control. For genes with a decreasing trend, the differences smaller than \( 100 \times \delta\% \) of the control are of interest. When testing \( \delta = 0 \), in fact, a two-sided test is made. Nevertheless, in this case, we can categorize significant genes into increasing or decreasing trends.

For each gene in the case study dataset, an MCT with Marcus’ contrasts is applied as discussed in Section 11.3.2, and BH-FDR procedure is used to control the
FDR. For testing $\delta = 0$, we obtain 2024 genes with statistically significant increasing trends, and 2222 genes with decreasing trends. The number of significant genes equals 4244 (=2024+2222) and is slightly different from the number of 3499 obtained using the $E_{01}^2$ in Chapter 10. The difference is due to the fact that the inference for MCTs is based on the multivariate $t$-distribution described in Section 11.4.1, while the $E_{01}^2$ test is carried out based on the null distribution approximated by using permutations.

Moreover, different values of $\delta$ (0.05, 0.1, 0.15, and 0.2) are also used for testing $H_{0U}^*$ ($H_{0D}^*$) against one-sided alternatives $H_{1U}^*$ for genes with an increasing trend ($H_{1D}^*$ for genes with a decreasing trend). For an increasing trend, 934 genes are found significant for $\delta=0.05$, 429 genes for $\delta=0.1$, 247 genes for $\delta=0.15$, and 142 genes for $\delta=0.2$. As expected, as $\delta$ increases, the number of significant genes decreases. The same is true for genes with a decreasing trend (results shown in Table 11.1). Note that genes significant for a larger $\delta$ are always the subset of the genes significant for a smaller $\delta$.

Table 11.1: Number of significant genes tested under $H_{1U}^*$ and $H_{1D}^*$ with $\delta = 0$, 0.05, 0.1, 0.15, 0.2.

<table>
<thead>
<tr>
<th>$\delta$</th>
<th>0</th>
<th>0.05</th>
<th>0.1</th>
<th>0.15</th>
<th>0.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_{1U}^*$ Rejected</td>
<td>2879</td>
<td>934</td>
<td>429</td>
<td>247</td>
<td>142</td>
</tr>
<tr>
<td>$H_{1D}^*$ Rejected</td>
<td>3387</td>
<td>968</td>
<td>330</td>
<td>133</td>
<td>66</td>
</tr>
</tbody>
</table>

The genes presented in Figure 11.2 and Figure 11.3 are found to be statistically significant when the $E_{01}^2$ test was used. However, the $E_{01}^2$ does not distinguish between genes $a$ and $b$ ($\delta = 0.2$) and genes $c$ and $d$ ($\delta = 0.1$). This distinction can be achieved using the ratio test. Moreover, using the t-type tests (Williams', Marcus', the $M$, and modified $M$ tests) allows us to test the null hypothesis $H_0'': \mu(d_3) - \mu(d_0) = \vartheta$ vs. $H_{1U}''$: $\mu(d_3) - \mu(d_0) > \vartheta$. However, for a specific shift in mean gene-expression between the largest dose and the control, one needs to test $H_0'': \mu(d_3) - \mu(d_0) = \vartheta$ vs. $H_{1U}''$: $\mu(d_3) - \mu(d_0) > \vartheta$. The value of $\vartheta$ is gene specific and depends on the expression level. For example, the increase of 20% for gene $a$ in Figure 11.2 implies the estimated difference of $\hat{\mu}_3^* - \hat{\mu}_0^*$ equal to about 5.6, while the increase of 20% for gene $b$ in Figure 11.2 implies the estimated difference of $\hat{\mu}_3^* - \hat{\mu}_0^*$ equals to about 4.2, because the estimated mean for the control dose $\hat{\mu}_0^*$ of gene $b$ is larger than
that of gene $a$. In contrast, using the ratio test allows us to test the null hypothesis $H_0^*: \mu(d_3)/\mu(d_0) = 1 + \delta$ vs. $H_1^{UP}: \mu(d_3)/\mu(d_0) > 1 + \delta$ in (11.1) with same value of
\( \delta = 0.2 \) for all the genes with increasing trends. The same conclusion can be drawn for genes \( a, b, c, \) and \( d \) (shown in Figure 11.3) with decreasing trends for testing
\[ H_0^* : \frac{\mu(d_3)}{\mu(d_0)} = 1 - \delta \text{ vs. } H_1^{\text{Down}} : \frac{\mu(d_3)}{\mu(d_0)} < 1 - \delta, \]
genes \( a \) and \( b \) are statistically significant for \( \delta = 0.2 \), and genes \( c \) and \( d \) are statistically significant for \( \delta = 0.1 \).

### 11.6 Discussion

In this chapter, we focused on testing the ratio of the mean gene-expression difference between the highest dose and the control using the MCT in the microarray setting. It is often expected that the increase of the dose implies a particular percentage increase/decrease of the mean expression level. Thus, this biological difference is pre-determined and the ratio test becomes of special interest. It avoids testing for gene-specific difference in mean gene-expression between two doses, which depends on the gene-expression level. Based on the significance of ratios, it is also possible to rank genes, which are found significant using the five test statistics (i.e., the \( LRT \), Williams', Marcus', the \( M \) and modified \( M \) tests, discussed in Chapter 8).

In this chapter, we have shown representation of the two mean gene-expression level difference under an order restriction using multiple contrasts, i.e., the Marcus-type MCT. In order to answer the testing problem defined in (11.1), we transformed the ratio test into the MCT corresponding to Marcus-type multiple ratio contrasts (Bretz 1999, 2006). This method efficiently utilizes the multivariate \( t \)-distribution as the basis for the inference on the multiple ratio contrast tests.

By choosing a set of \( \delta \) values in advance, we have performed the ratio tests for each \( \delta \). Sequential increasing of \( \delta \) values reduces the number of significant findings. Genes found with a larger value of \( \delta \) are of great interest to the scientists for further investigation.
Chapter 12

Multiple Confidence Intervals for Selected Ratio Parameters Adjusted for the False Coverage-statement Rate

12.1 Introduction

The ratio test, discussed in Chapter 11, was used in order to select a subset of genes, for which a significant increase (decrease) of $100 \cdot \delta\%$ in the mean gene-expression is detected. Once a subset of genes is selected for a given value of $\delta$, the primary interest is to construct simultaneous confidence intervals (CIs) for the parameters of the selected genes, taking into account the multiplicity issue. Ignoring the multiplicity issue by constructing $(1 - \alpha) \times 100\%$ interval for each gene is not appropriate. However, using the Bonferroni adjustment is too conservative and leads to too wide confidence intervals. Benjamini and Yekutieli (2005) introduced the concept of the False Coverage-statement Rate (FCR) to ensure the non-coverage rate for a set of selected parameters.

In this chapter, we use the FCR method proposed by Benjamini and Yekutieli (2005) for the parameters which were selected by the ratio test introduced in the
Chapter 12. An application of the FCR Adjusted CIs for a Ratio Test

Chapter 11.

The contents of this chapter is organized as follows, Section 12.2 introduces the methods of constructing the confidence interval for a single ratio and multiple ratios. Section 12.3 discusses the concept of the FCR when constructing multiple confidence intervals for a set of selected parameters. Section 12.4 summarizes the FCR-adjusted BH-FDR procedure for constructing the confidence intervals, while controlling the FCR. In Section 12.5 we present the results of an application to the case study. The chapter ends with a discussion in Section 12.6.

12.2 Confidence Intervals for Single and Multiple Ratios

Following the definitions in Section 11.4.1, for a single ratio, say $\gamma = c'\mu/d'\mu$, given in (11.10), a confidence interval can be constructed using Fieller’s theorem (Fieller 1954). A two-sided $(1-\alpha)\times100\%$ CI for the parameter $\gamma$ is the solution of the inequality

$$|t(\gamma)| = \frac{|(\gamma d - c)'\bar{y}|}{s[\gamma^2d'Md - 2\gamma c'Md + c'Mc]^{1/2}} \leq \frac{1}{2}t_{\alpha/2}(\nu),$$  \hspace{1cm} (12.1)

where $t_{\alpha/2}(\nu)$ is the $1-\alpha/2$ quantile of $t$-distribution with $\nu = \sum_{i=1}^{K} (n_i - 1)$ degrees of freedom. The inequality in (12.1) can be expressed as a quadratic inequality in $\gamma$:

$$A\gamma^2 + B\gamma + C \leq 0,$$  \hspace{1cm} (12.2)

where

$$A = (d'y)^2 - \tilde{t}^2s^2d'Md,$$

$$B = -2[(c'y)(d'y) - \tilde{t}^2s^2c'Md],$$

$$C = (c'y)^2 - \tilde{t}^2s^2c'Mc$$

and $\tilde{t} = t_{\alpha/2}(\nu)$.

Depending on the value of the leading coefficient $A$, and the discriminant $B^2 - 4AC$, there are three possible solutions to the inequality in (12.2) (Kendall 1999). If $A > 0$, then it can be shown that also $B^2 - 4AC > 0$ and there are two solutions to equation (12.2). Consequently, the CI is a finite interval lying between the two roots of (12.2). The other two cases, when $A \leq 0$ and $B^2 - 4AC \leq 0$, result in either a region containing all values lying outside the finite interval defined by the two roots of (12.2), or containing the entire $\gamma$-axis, which is commonly referred as Fieller’s problem. If
the denominator of the ratio test, defined in (11.10), is significantly different from 0, the last two cases occur only with a small probability.

For \( r \) multiple ratios \( \gamma = (\gamma_1, \ldots, \gamma_r)' \), many combinations of these types (\( A \leq 0 \) and/or \( B^2 - 4AC \leq 0 \)) exist and it is difficult to fully describe the geometry of all such regions. Dilba et al. (2005) discussed the “plug-in” method to approximate the simultaneous CIs.

Let \( Q_l(\gamma_l, t_{1-\alpha/2}) = A_l\gamma_l^2 + B_l\gamma_l + C_l \) denote a quadratic function for the ratio parameter \( \gamma_l \) derived on the basis of critical point \( t_{1-\alpha/2} \) of \( t \)-distribution and by solving inequality (12.1).

By plugging the maximum likelihood estimators of the ratio parameters \( \hat{\gamma}_l \) (\( l = 1, \ldots, r \)), defined in (11.10), into the correlation matrix \( R(\gamma) \), we obtain \( R(\hat{\gamma}) = [\hat{\rho}_{ij}] \), where \( \hat{\rho}_{ij} = h(\hat{\gamma}_i, \hat{\gamma}_j) \), \( 1 \leq i \neq j \leq r \) (see Section 11.4.1). Let \( c_{1-\alpha} \{ R(\hat{\gamma}) \} \) denote the two-sided equicoordinate critical point associated with \( R(\hat{\gamma}) \). The approximate simultaneous CIs are obtained by solving

\[
Q_l \{ \gamma_l, c_{1-\alpha} \{ R(\hat{\gamma}) \} \} \leq 0, \ l = 1, \ldots, r.
\]

(12.3)

The one-sided simultaneous CIs can be obtained similarly (Dilba et al. 2005).

### 12.3 False Coverage-statement Rate (FCR)

Benjamini and Yekutieli (2005) argued that the issue of parameter selection and multiplicity is often ignored when it comes to the construction of multiple confidence intervals for a selected subset of parameters. They emphasized the concern about the average false coverage of confidence intervals following the selection. They introduced the concept of the FCR-adjusted multiple confidence intervals for selected parameters. The FCR was defined as the expected proportion of constructed confidence intervals not covering their respective parameters among all the confidence intervals constructed, where the proportion is 0 if no parameter is selected.

For a \( 1 - \alpha \) confidence interval for a single parameter \( \gamma \), which does not involve a selection procedure, the probability of constructing a non-covering confidence interval is at most \( \alpha \):

\[
P\{ \gamma \notin CI, CI \ constructed \} \leq Pr\{ \gamma \notin CI \} \leq \alpha
\]

When inference about multiple parameters is needed in an experiment without testing, the marginal probability of non-covering confidence intervals for the multiple parameters is also at most \( \alpha \). However, when facing both the multiplicity and
selection, guaranteeing the expected marginal coverage over selected parameters at
$1 - \alpha$ is not equivalent to assuring that the expected proportion of non-coverage for
the parameters is $\alpha$.

Consider a procedure for constructing a selection of multiple confidence intervals,
based on a vector of $m$ parameter estimators $t$. The selection procedure is given
by $S(t) \subseteq \{1, ..., m\}$. The selection procedure is followed by the construction of a
confidence interval for each $\gamma_i, i \in S(t)$. For example, the Bonferroni adjusted critical
point for a two-sided simultaneous confidence interval is defined as $t_{\alpha/2m}(\nu)$.

Let $V_{CI}$ be the number of constructed intervals not covering their respective pa-
rameters. The FCR of the selective CI procedure is $FCR = E_t(Q_{CI})$, where $Q_{CI}$ is
defined as follows,

$$Q_{CI} = \begin{cases} 
V_{CI}/R_{CI} & \text{if } R_{CI} > 0, \\
0 & \text{otherwise},
\end{cases}$$

where $R_{CI}$ is the number of CIs resulting from the selection procedure. For a single
parameter ($m = 1$), the FCR equals the probability of constructing a single non-
covering confidence interval.

The FCR-adjustment can be applied to any selection rule. Benjamini and Yekutieli
(2005) showed that, for independently distributed test statistics and most selection
rules, this procedure of adjusting for the FCR at level $q$ is roughly equivalent to
constructing a marginal CI with confidence level $1 - R_{CI} \cdot q/m$ for the $R_{CI}$ selected
parameters.

It is easy to note that there is a connection between the concept of the FCR-
controlling CIs and the FDR-controlling BH procedure. In the BH-FDR procedure
described in Section 5.2.2, we reject $R$ null hypotheses for which the sorted $p$-values
are less and equal to $R \cdot q/m$ while controlling the FDR at level $q$. The proposed
FCR approach above ensures the confidence level $1 - R_{CI} \cdot q/m$ for the CIs of the $R_{CI}$
parameters.

In the following section, we describe the procedure for constructing FCR-adjust
BH-selected CIs (Benjamini and Yekutieli 2005). The constructed CIs using the
proposed procedure are dual to the BH-FDR testing procedure.
12.4 FCR-adjusted BH-selected CIs

The construction of the FCR-adjusted BH-selected confidence intervals, proposed by Benjamini and Yekutieli (2005), consists of four steps:

1. Sort the $p$-values used for testing the $m$ hypotheses regarding the parameters, $P_1 \leq \ldots \leq P_m$.
2. Calculate $R_{CI} = \max\{i : P_{(i)} \leq i \cdot q/m\}$, where $q$ is the FDR level to be controlled.
3. Select the $R_{CI}$ parameters, for which $P_{(i)} \leq R_{CI} \cdot q/m$, corresponding to the rejected hypotheses.
4. Construct a $1 - R_{CI} \cdot q/m$ CI for each parameter selected.

The first three steps are related to the BH-FDR testing procedure that controls the FDR, while the last step is related to the construction of the FCR-adjusted confidence intervals. The length of the constructed CIs increases as the number of parameters considered ($m$) increases, but decreases as the number of selected parameters ($R_{CI}$) increases (Benjamini and Yekutieli 2005).

Constructing the FCR-adjusted confidence intervals is directly linked to the BH-FDR procedure that controls the FDR. The FCR-adjusted BH-selected procedure ensures the control of the FCR as well as the control of the FDR. If the null hypotheses are rejected in the testing step, the FCR-adjusted CIs will not cover their null parameter values. This property only holds if the confidence intervals correspond to the $p$-values, meaning one(two)-sided $p$-values should correspond to the one(two)-sided CIs for selected parameters.

12.5 Application to the Data

In this section we construct the FCR-adjusted BH-selected CIs for the ratio parameters of genes found significant by the ratio test with $\delta = 0.2$ in Section 11.5. The number of rejected hypotheses for $\delta = 0.2$ is 142 for the increasing trend and 66 for the decreasing trend. The FCR-adjusted $1 - R_{CI} \cdot q/m$ CI for the selected ratios is constructed, where $R_{CI} = 208$ and $m = 3499$.

Figure 12.1 shows the confidence intervals for selected genes using the Bonferroni, FCR adjusted, and without any multiplicity adjustment (namely, unadjusted).
Figure 12.1: Example of CIs constructed by the Bonferroni, FCR-adjusted, and the adjusted method for selected genes found significant by the ratio test with $\delta = 0.2$. 

Chapter 12. An application of the FCR Adjusted CIs for a Ratio Test
Figure 12.2: Length of confidence intervals for genes found significant by the ratio test with $\delta = 0.2$, using the unadjusted (dashed line with pluses), Bonferroni (solid line), and FCR-adjusted (dotted line with stars) approaches.

Figure 12.2 shows the width of the CIs for the 208 significant genes. The unadjusted CIs (dotted line with pluses shown in Figure 12.2) are always the shortest, while the Bonferroni CIs are the widest (solid line).

12.6 Discussion

In the microarray setting, the control of the FDR is well addressed when testing thousands of genes simultaneously. The multiplicity issue in the construction of simultaneous CIs for the selected genes needs to be addressed in this context as well. In this chapter, our aim was to construct simultaneous CIs for the ratio parameters for genes found significant by the ratio tests in Section 11.5.

We applied the procedure proposed by Benjamini and Yekutieli (2005) to adjust the simultaneous confidence levels to ensure the FCR. The FCR-adjusted BH-selected
procedure links the control of FCR in constructing CIs with the control of FDR in testing the hypotheses. We compared the length of CIs obtained using FCR adjustment, no adjustment, and the Bonferroni adjustment. The constructed FCR-adjusted BH-selected CIs did not cover the null value of parameter tested, while the Bonferroni-adjusted CIs were too wide. The use of the FCR-adjusted CIs needs to be highlighted as an analogue to the control of the FDR in adjusting for multiple testing in the microarray setting.

The contents of Chapters 11 and 12 were summarized in the manuscript (Lin et al. 2008).
Part IV

Selection and Evaluation of Biomarkers in a Pre-clinical Microarray Study
Chapter 13

Introduction to Biomarker Identification in the Drug Discovery Using Microarrays

13.1 Introduction

Biomarkers play a number of stage-specific roles in guiding drug development (Lesko 2007). It is of particular importance for pharmaceutical industries that the reliability of biomarkers can be used to help make decisions to terminate a drug development program early and to allow companies to reprioritize and focus on the alternative molecules with greater potential for success. Moreover, in the selection of dose(s) from dose-ranging trials in phase II, biomarkers can causally link to clinical efficacy or, in some cases, to safety endpoints. Furthermore, biomarkers lay the foundation for the possible development of a diagnostic test to identify responders to a given drug, non-responders, or perhaps those at risk for adverse events.

The rapid development of microarray technology makes it possible to search for genomic biomarkers, i.e., genes, which expression could be used as a biomarker and, perhaps, as a surrogate. In a microarray experiment, there are many, say $m$, potential biomarkers (the genes) for one outcome. Thus, there is a need for selection of genes, for which gene-expression might serve as a biomarker for the outcome. In this chapter we focus on the methods that are used to identify and evaluate biomarkers.
The case study used in this chapter is the male Hanover Wistar rat dataset discussed in Section 1.4.2, where the rats were randomized into three treatment arms (the test compound with doses 2.5 and 10 mg/kg s.c., and the reference compound) with the corresponding sample size eight, seven, and nine, respectively. Gene-expression of 15,521 probe sets, along with the primary endpoint, the brain occupancy, were measured. In the analysis of the case study, we make two comparisons between the treatment groups: (1) the low dose of the test compound vs. the reference compound; and (2) the high dose of the test compound vs. the reference.

The chapter is organized as follows. In Section 13.2 introduces definitions of a biomarker, a clinical endpoint, and a surrogate endpoint. Methods of the evaluation of surrogate endpoints in the clinical trial setting are reviewed in Section 13.3. Based on the clinical trial framework, in Section 13.4 we apply the methods to the microarray setting for selection of two types of biomarkers. In Section 13.5 we discuss the evaluation of biomarkers. Finally, we apply the methods to the case study data in Section 13.6. The chapter is concluded with a discussion in Section 13.7.

13.2 Definitions and Regulatory Perspective on Validation of Surrogate Endpoints and Biomarker

According to Biomarkers Definition Working Group (2001), biomarker is an objectively measured indicator of normal biological or pathogenic processes, or of pharmacologic responses to a therapeutic intervention.

Clinical endpoint (also called true endpoint in this dissertation) is a characteristic or variable that reflects how a patient feels, functions, or survives.

Surrogate endpoint is a biomarker that is intended to substitute for a clinical endpoint.

In randomized clinical trials, the main interest is to assess the effect of treatment \( Z \) on the primary clinical (“true”) endpoint \( Y \). However, in some cases, the endpoint may be difficult to use due to, e.g., high measurement costs or a long observation time (for example, if the primary endpoint is time to event). In these cases, the use of a surrogate endpoint \( X \) is of interest if it can be measured earlier, easier, and/or more precisely than the clinical endpoint, and thus when it can shorten and/or reduce cost of a clinical trial (Burzykowski, Molenberghs, and Buyse 2005).
Before any biomarker, or any other endpoint, can be used as a surrogate for a true endpoint, it should be validated. The validation of a candidate surrogate endpoint is not straightforward. Statistical approaches to the validation of surrogate endpoint have recently become a topic of intensive research (Burzykowski, Molenberghs, and Buyse 2005).

The most important reason for biomarker development is not to achieve a new surrogate endpoint, although surrogate endpoint - where feasible - does offer many advantages. More importantly, the FDA set up one of the critical path initiatives to examine the significant body of collective biomarker information to build a new framework for biomarker and surrogate endpoint use in drug development (Lesko 2007). However, in this process, there will be an attempt to identify widely used, disease-specific biomarkers and to determine the gap that separates them from becoming surrogate endpoints.

13.3 Validation of Surrogate Endpoints in the Clinical Trial Setting: a Joint Model Approach

We focus on the meta-analytic approach defined by Buyse et al. (2000). That is, we consider the situation when a potential surrogate is evaluated using data from multiple, say \( n \), trials. We further assume that the distribution of the true endpoint \( Y \) and surrogate endpoint \( X \) come from an exponential family and that true treatment effects on the endpoints are given by

\[
\begin{align*}
    g[E(X_{ij} | Z_{ij} = 1)] - g[E(X_{ij} | Z_{ij} = 0)] &= \alpha_i, \\
    g[E(Y_{ij} | Z_{ij} = 1)] - g[E(Y_{ij} | Z_{ij} = 0)] &= \beta_i,
\end{align*}
\]  

(13.1)

where \( g() \) denotes an appropriate link function, and \( X_{ij} \) and \( Y_{ij} \) denote the true endpoint and the surrogate endpoint, respectively, for patient \( j \) in trial \( i \). For normally distributed \( X_{ij} \) and \( Y_{ij} \), \( g() \) is an identity link. Thus, the joint model above can be written as

\[
\begin{align*}
    X_{ij} &= \mu_X + \alpha_i Z_{ij} + \varepsilon_{X_{ij}}, \\
    Y_{ij} &= \mu_Y + \beta_i Z_{ij} + \varepsilon_{Y_{ij}},
\end{align*}
\]  

(13.2)

where the individual-level error terms \( \varepsilon_{X_{ij}} \) and \( \varepsilon_{Y_{ij}} \) are zero-mean normally distributed with variance-covariance matrix
Chapter 13. Introduction of Biomarker Identification

\[ \Sigma = \begin{pmatrix} \sigma_{XX} & \sigma_{XY} \\ \sigma_{XY} & \sigma_{YY} \end{pmatrix}. \] (13.3)

The joint model (13.2) leads to the following conditional distribution of \( Y_{ij} \), given \( X_{ij} \) and \( Z_{ij} \) (Burzykowski, Molenberghs, and Buyse 2005):

\[ Y_{ij} \sim N\{\mu_y - \sigma_{YX} \sigma_{XX}^{-1} \mu_X + (\beta_i - \sigma_{YX} \sigma_{XX}^{-1} \alpha_i)Z_{ij} + \sigma_{YX} \sigma_{XX}^{-1} X_{ij}; \sigma_{YY} - \sigma_{YX}^2 \sigma_{XX}^{-1}\}. \] (13.4)

We focus on the surrogacy measure at the individual trial level. The association between both endpoints after adjusting for the treatment effect is captured by

\[ R^2_{\text{indiv}} = R^2_{\varepsilon_{Y_i} | \varepsilon_{X_i}} = \frac{\sigma_{XY}^2}{\sigma_{XX} \sigma_{YY}}. \] (13.5)

Buyse and Molenberghs (1998) proposed the adjusted association as a measure for the individual-level surrogacy. The adjusted association \( \rho \) is the squared root of \( R^2_{\text{indiv}} \), given in (13.5). Indeed, \( \rho = 1 \) indicates perfect surrogacy, in the sense that, given the surrogate endpoint, a perfect prediction of the true endpoint is possible.

Recently, Alonso and Molenberghs (2006) proposed a new approach to the evaluation of surrogate endpoints, based on the information theory. In this approach, the focus of attention is placed on the reduction of the uncertainty about the true endpoint induced by the surrogate endpoint. For a single trial setting and, in the case when the two endpoints are normally distributed, the information theory approach requires from a "good" surrogate to explain a "large amount" of the variability of the true endpoint. Furthermore, Alonso and Molenberghs (2006) showed that \( R^2_{\text{indiv}} \), given in (13.5), is an information theoretic measure for association.

13.4 Selection of Biomarkers in the Microarray Setting

Biomarker identification has been extended to the microarray experiments. The aim of such experiments can be to discover important genes or genes pathways that links with variations of phenotype endpoints, such as chemical and drug toxicity and efficacy of drugs (Tan et al. 2005), diagnosis and prognosis of diseases (Ma et al. 2007, Alon et al. 2006, Spang et al. 2001, West et al. 2001), and survival status (Dave et al. 2004, Rosenwalk et al. 2003). Thus, these phenotype endpoints can be binary (disease
groups), categorical (disease subgroups), ordinal (toxicity levels), continuous (efficacy measures), or survival-type. In this chapter, we focus on the brain occupancy of drug, which is a continuous response.

Various methods have been developed to identify predictive genes for the clinical outcome. They mainly consist of three steps: (1) technical screening of genes (dealing with missing data, genes without much variation, etc.), (2) supervised gene screening, and (3) predictive model building using the top genes on a ranked list. The second step of the analysis is directly linked with selection of potential biomarkers. In what follows, we consider the model used in the clinical trial setting for this purpose.

Consider a pre-clinical experiment, in which treatment effects are estimated for gene-expression level and the response. Let \( X_{ij} \) be the \( i \)th gene-expression (\( i = 1, \ldots, m \)), of the \( j \)th subject (\( j = 1, \ldots, n \)). We denote the measurement for the response of primary interest as \( Y_j \). Subjects in the experiment are randomized to two treatment groups, active drug and placebo. Let \( Z_j \) be an indicator variable, which takes value 1 if the \( j \)th subject was randomized to the active drug group, and 0 otherwise. Finally, denote by \( \hat{\alpha}_i \) and \( \hat{\beta} \) the maximum likelihood estimates of the treatment effects for the \( i \)th gene-expression and the response, respectively.

Following Buyse and Molenberghs (1998), we assume a gene-specific joint model for gene-expression \( X_{ij} \) and the response \( Y_j \), i.e.,

\[
\begin{pmatrix}
X_{ij} \\
Y_j
\end{pmatrix}
\sim N
\left[
\begin{pmatrix}
\mu_{X_i} + \alpha_i Z_j \\
\mu_Y + \beta Z_j
\end{pmatrix},
\Sigma_i
\right],
\tag{13.6}
\]

where \( \Sigma_i \) is given by

\[
\Sigma_i = \begin{pmatrix}
\sigma_{ii} & \sigma_{iv} \\
\sigma_{iv} & \sigma_{yy}
\end{pmatrix}.
\]

Parameters \( \alpha_i \) are gene-specific fixed treatment effects and \( \mu_{X_i} \) are gene-specific fixed intercepts. The correlation coefficient between gene-expression and the response is given by

\[
\rho_i = \frac{\sigma_{iv}}{\sqrt{\sigma_{ii} \sigma_{yy}}}. 
\tag{13.7}
\]

Note that \( \rho_i \) is the gene-specific correlation coefficient between gene-expression and the response after adjusting the treatment effect.

The joint model (13.6) implies the following conditional model (Burzykowski, Molenberghs, and Buyse 2005):
Chapter 13. Introduction of Biomarker Identification

\[ Y_j = \bar{\mu}_i + \tilde{\beta}_i Z_j + \tilde{\gamma}_i X_{ij} + \tilde{\varepsilon}_{ij}, \]  

(13.8)

where

\[ \tilde{\beta}_i = \beta - \sigma_i Y \sigma_{ii}^{-1} \alpha_i, \]

\[ \tilde{\gamma}_i = \sigma_i Y \sigma_{ii}^{-1}, \]

and

\[ \tilde{\varepsilon}_{ij} \sim N(0, \sigma_Y^2 - \sigma_{ii}^2 \sigma_{ii}^{-1}). \]

Note that the joint model (13.6) and the conditional model (13.8) are identical to the joint model (13.2) and conditional model (13.4) used in the evaluation of a surrogate endpoint in a multiple clinical trial setting.

13.4.1 Inference

For a microarray with \( m \) genes, there are \( m \) null hypotheses to be tested, which implies that an adjustment for multiple testing should be applied. Thus, we will use the BH-FDR procedure, discussed in Chapter 5, to adjust the asymptotic \( p \)-values for the parameters of interest and to control the FDR at the level of 0.05. In what follows, we discuss the null hypotheses to be tested when identifying two types of biomarkers: \textit{therapeutic} and \textit{prognostic}. We elaborate on the difference between the therapeutic and prognostic biomarkers in Section 13.4.2.

Testing for a Therapeutic Biomarker

A typical analysis of DNA microarrays allows monitoring expression levels of thousands of genes simultaneously, and identifying differentially expressed genes. This type of genes can potentially be \textit{therapeutic biomarkers}, for which the treatment effect on the gene-expression can be predictive for the treatment effect on the response. In the first step, we test, which genes are differentially expressed using the joint model defined in (13.6). Hence, for each gene, we test the hypotheses

\[ H_{0i}^A : \alpha_i = 0, \]

\[ H_{1i}^A : \alpha_i \neq 0. \]  

(13.9)

Testing the treatment effect upon the response consists of testing

\[ H_0 : \beta = 0, H_1 : \beta \neq 0. \]  

(13.10)
Note that the case, in which both $H^A_{0i} : \alpha_i = 0$ and $H_0 : \beta = 0$ are rejected, implies that the gene is a potential therapeutic biomarker.

**Testing for a Prognostic Biomarker**

Within the clinical trial setting, the adjusted association is a measure for individual-level surrogacy (Buyse and Molenberghs 1998), which is a measure of a linear association between the surrogate and the true endpoints. Within the microarray setting, we test whether or not a gene can serve as a prognostic biomarker, which can be used to predict the response. Thus, one needs to test the hypotheses

\[
egin{align*}
H^B_{0i} : \rho_i &= 0, \\
H^B_{1i} : \rho_i &\neq 0,
\end{align*}
\tag{13.11}
\]

where $\rho_i$ is defined in (13.7). A gene is declared an up-regulated prognostic biomarker if the null hypothesis in (13.11) is rejected and $\hat{\rho}_i > 0$, and as a down-regulated prognostic biomarker when $\hat{\rho}_i < 0$. In case that $H^A_{0i} : \alpha_i = 0$, $H_0 : \beta = 0$, and $H^B_{0i} : \rho_i = 0$ are rejected, the gene is declared as a potential prognostic/therapeutic biomarker.

**13.4.2 Therapeutic and Prognostic Biomarkers: a Graphical Interpretation**

Figure 13.1 shows scatterplots of three hypothetical examples of the relationship between gene-expression ($X$) and the response ($Y$). Circles represent values for one treatment group and pluses measurements for the other treatment group. In all examples the treatment effect upon response is significant. The upper three panels present the scatterplot of the gene-expression versus the response, while the lower three panels show the scatterplot of the residuals (after adjusting for treatment effects) for both the response and gene-expression. In panel a, the gene is not differentially expressed, but there is a linear association of gene-expression with the response. Note that the linear pattern remains after adjusting for treatment effect, as shown in panel d. We term a gene with this pattern a prognostic biomarker. Panel b shows an example, in which the gene is differentially expressed, the two treatment groups are clearly separated, but the association between gene-expression and the response does not have a linear appearance, which can be seen also in panel e. We term a gene with this type of relationship a therapeutic biomarker. Panel c shows a combination of the two previous
patterns. Gene-expression is differentially expressed and the treatment effect upon response is significant, the two treatment groups are clearly separated with respect to gene-expression and the response and the association between the gene-expression and the response can be summarized by a straight line. This can also be seen from panel f, which shows the same example after adjusting for treatment effects. We term this type of relationship as a *prognostic/therapeutic biomarker*.

Figure 13.1: *Biomarker types in microarray experiment, when the response is differentially expressed.* Pluses and circles represent the two treatment groups, respectively. Upper row (panel a, b, and c): scatterplots for the response (Y) versus gene-expression (X). Lower row (panel d, e, and f): scatterplots for the residuals after adjusting for treatment effects. Column 1 (panel a and d): an example of a prognostic biomarker. Column 2 (panel b and d): an example of a therapeutic biomarker. Column 3 (panel c and f): an example of a prognostic/therapeutic biomarker.
13.5 Evaluation of Biomarkers

13.5.1 Prognostic Biomarkers

A prognostic biomarker, where $\rho_i$ is found to be significant, can be evaluated using an estimate $\hat{\rho}_i$, which measures the linear association between gene-expression and the response, after accounting for treatment effects. Indeed, this is the square root of “individual $R^2$” surrogacy measure proposed in the context of randomized clinical trials by Buyse et al. (2000).

13.5.2 Therapeutic Biomarkers

For therapeutic biomarkers, the adjusted association $\rho_i$ is not applicable, because there is no association between the gene-expression and the response, or the association is not linear (i.e., cannot be summarized by a straight line).

In this case we follow the approach of Alonso and Molenberghs (2006) and propose a measure for therapeutic biomarker, the relative deviance reduction. The total variability of the response, the deviance, without any information about the gene-expression level can be measured by

$$D(Y) = \sum_{j=1}^{n} (Y_j - \hat{\mu})^2,$$  \hspace{1cm} (13.12)

where $\hat{\mu} = 1/n \sum_{j=1}^{n} Y_j$ and $j = 1, \ldots, n$ indicates the arrays. For a therapeutic biomarker, because gene-expression is differentially expressed, one can use the gene-expression level in order to predict the response level. While a linear regression model is not an appropriate model for this type of a biomarker, a regression tree model (Venables and Ripley 1994), in which the gene-expression is the only predictor, can capture the structure of the data shown in Figure 13.2.

Moreover, because the gene is differentially expressed, we can restrict the tree to have only two terminal nodes (two final homogenous groups of the response), in which the cutoff point (or the split point) is determined only by the gene-expression level. An example of the cutoff point is shown as the vertical line in Figure 13.2. Let $k$ denote the number of terminal nodes in the tree and let $D(Y|X, k = 2)$ denote the sum of deviances for the terminal nodes,
Figure 13.2: A regression tree model for a hypothetical example with two terminal nodes. The blue line in the plot indicates the split point in the regression tree. \( D(Y) \) represents the total variability in the response \( Y \), while \( D_1(Y|X) \) and \( D_2(Y|X) \) represent the variability within each of the terminal nodes.

\[
D(Y|X, k = 2) = D_1(Y|X) + D_2(Y|X) = \sum_{Y_j \in k_1} (Y_j - \hat{\mu}_1)^2 + \sum_{Y_j \in k_2} (Y_j - \hat{\mu}_2)^2, \tag{13.13}
\]

where \( D_1(Y|X) \) and \( D_2(Y|X) \) denote the deviance in each of the terminal nodes, \( k_1 \) and \( k_2 \) denote the sets of subject indices corresponding to the two terminal nodes, and \( \hat{\mu}_1 \) and \( \hat{\mu}_2 \) are the mean response in the two terminal nodes. The reduction in the deviance, \( D(Y) - D(Y|X, k = 2) \), measures the gain in prediction of the response level using gene-expression, as compared to the case where the gene-expression is not used. In other words, the reduction in deviance measures whether information about the gene-expression is relevant for predicting the response level. The relative deviance reduction, \( R^2_D \), is given by

\[
R^2_D = \frac{D(Y) - D(Y|X)}{D(Y)} = \frac{D(Y) - D_1(Y|X) - D_2(Y|X)}{D(Y)},
\]
hence,

\[ R^2_D = \frac{\sum_{j=1}^{n} (Y_j - \hat{\mu})^2 - \left[ \sum_{Y_j \in k_1} (Y_j - \hat{\mu}_1)^2 + \sum_{Y_j \in k_2} (Y_j - \hat{\mu}_2)^2 \right]}{\sum_{j=1}^{n} (Y_j - \hat{\mu})^2}. \]  

(13.14)

Similar to \( R^2 \) in a linear regression, \( R^2_D \) measures the proportion of variability explained by the regression tree model. It is easy to see that \( R^2_D \), as a measure of association, is equivalent to the variance reduction factor discussed by Alonso et al. (2003). Moreover, in the case when the model has two terminal nodes and we fit a regression model \( Y_j = \beta_0 + \beta_1 X_{ij} \), one can easily see that the \( R^2 \) of the regression model and the \( R^2_D \) of the regression tree are equal (i.e., \( R^2 = R^2_D \)) for gene \( i \).

Following Alonso and Molenberghs’ (2006) information approach, it is easy to see that both \( R^2_{\text{indiv}} \) (i.e., the square of \( \rho \) (13.7)) and \( R^2_D \) belong to the family of information theoretic association measures. This is a crucial point, as it implies that, although prognostic and therapeutic biomarkers are evaluated using different validity measures (i.e., \( \rho \) and \( R^2_D \), respectively), both measures can be interpreted in the same way. Both \( R^2_{\text{indiv}} \) and \( R^2_D \) measure the proportion of the total variability explained by the model.

### 13.6 Application to the Data

In this section, we present the results of an application of the methods described in Section 13.3 and 13.4 to the case study using the two comparisons between the treatment groups: (1) the low dose of the test compound vs. the reference compound; and (2) the high dose of the test compound vs. the reference.

#### 13.6.1 Treatment Effect on the Response

In the first stage of the analysis, the joint model, given in (13.6), is used to test the null hypothesis \( H_0 : \beta = 0 \) against the alternative \( H_1 : \beta \neq 0 \) for the comparison between the low dose of the test compound and the reference, and for the comparison between the high dose of the test compound and the reference, respectively. Boxplots for the distribution of the response in Figure 13.3 show that there is no significant effect of the low dose as compared to the reference compound. Indeed, the \( t \)-test statistic is equal to 0.41 with \( p \)-value of 0.59. The null hypothesis is not rejected. On
the other hand, the effect of the high dose is highly significant with the test statistic equal to 8.68 and $p$-value equal to $5 \times 10^{-7}$.

Figure 13.3: Boxplot of the brain occupancy distribution for low and high doses of the test compound vs. the reference

13.6.2 Treatment Effect on the Gene-Expression: Testing for Therapeutic Biomarkers

In this step, we test the hypotheses, defined in (13.9), based on the asymptotic $p$-values from the joint model, using the BH-FDR procedure to control the FDR for the two comparisons. For the comparison of the low dose of the test compound vs. the reference, there are 31 genes declared significant with the FDR level of 0.05, while the comparison of the high dose of the test compound vs. the reference yields only seven significant genes.

As discussed in Section 13.5.2, the differentially expressed genes are potential therapeutic biomarkers, given that the treatment effect upon the response is present. In Section 13.6.1, we have shown that there was no significant treatment effect of the low dose of the test compound on the response. Thus, the differentially expressed genes cannot be used as therapeutic biomarkers for the low dose of the test compound, while seven genes declared significant for the comparison of the high dose of the test compound and the reference are potential therapeutic biomarkers.
We use $R^2_D$ from the regression tree to measure the quality of therapeutic biomarkers. Table 13.1 lists the genes found significant with their estimated treatment effect $\alpha_i$, the test statistic for $\alpha_i$, the corresponding $p$-value, and $R^2_D$. The left panel of the table shows the list of differentially expressed genes for the comparison of the low dose of the test compound with the reference. It is easy to note from the left panel that the $R^2_D$ for the 31 genes is very low, given that the treatment effect upon the response is absent. In the right panel of the table (for the comparison of the high dose of the test compound vs. the reference), the highest $R^2_D$ obtained is 0.84, which is the maximum value achievable in this data set. Note that there are five genes declared significant expressed for both comparisons. This suggests that these five genes may potentially be good therapeutic biomarkers, because the levels of gene-expression and the response are targeted by increasing the dose of the test compound.
Table 13.1: Differentially expressed genes for the comparisons of the test compound low and high doses vs. the reference: gene ID, treatment effect upon gene-expression $\alpha_i$, the $t$-test statistic, $p$-value, and $R^2_D$. Genes in bold are differentially expressed in the both comparisons.

<table>
<thead>
<tr>
<th>Gene.ID</th>
<th>Low Dose vs. Reference</th>
<th>High Dose vs. Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\alpha$</td>
<td>$\alpha$ stat</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>-0.8726</td>
</tr>
<tr>
<td>2</td>
<td>102</td>
<td>-1.2522</td>
</tr>
<tr>
<td>3</td>
<td>1234</td>
<td>1.1453</td>
</tr>
<tr>
<td>4</td>
<td>1795</td>
<td>0.459</td>
</tr>
<tr>
<td>5</td>
<td>2290</td>
<td>-0.4815</td>
</tr>
<tr>
<td>6</td>
<td>2861</td>
<td>0.4069</td>
</tr>
<tr>
<td>7</td>
<td>2896</td>
<td>0.7927</td>
</tr>
<tr>
<td>8</td>
<td>3596</td>
<td>0.773</td>
</tr>
<tr>
<td>9</td>
<td>3881</td>
<td>1.7606</td>
</tr>
<tr>
<td>10</td>
<td>4034</td>
<td>0.9118</td>
</tr>
<tr>
<td>11</td>
<td>5202</td>
<td>0.8654</td>
</tr>
<tr>
<td>12</td>
<td>5298</td>
<td>0.7115</td>
</tr>
<tr>
<td>13</td>
<td>5568</td>
<td>1.0129</td>
</tr>
<tr>
<td>14</td>
<td>5741</td>
<td>2.6529</td>
</tr>
<tr>
<td>15</td>
<td>6157</td>
<td>0.497</td>
</tr>
<tr>
<td>16</td>
<td>6433</td>
<td>1.4587</td>
</tr>
<tr>
<td>17</td>
<td>6680</td>
<td>2.0613</td>
</tr>
<tr>
<td>18</td>
<td>6855</td>
<td>-1.5682</td>
</tr>
<tr>
<td>19</td>
<td>6899</td>
<td>1.8199</td>
</tr>
<tr>
<td>20</td>
<td>7622</td>
<td>1.1361</td>
</tr>
<tr>
<td>21</td>
<td>8004</td>
<td>2.5878</td>
</tr>
<tr>
<td>22</td>
<td>8030</td>
<td>0.7868</td>
</tr>
<tr>
<td>23</td>
<td>8299</td>
<td>1.0036</td>
</tr>
<tr>
<td>24</td>
<td>8852</td>
<td>1.7843</td>
</tr>
<tr>
<td>25</td>
<td>9857</td>
<td>1.0682</td>
</tr>
<tr>
<td>26</td>
<td>10332</td>
<td>1.1019</td>
</tr>
<tr>
<td>27</td>
<td>11041</td>
<td>0.751</td>
</tr>
<tr>
<td>28</td>
<td>11554</td>
<td>0.8067</td>
</tr>
<tr>
<td>29</td>
<td>13452</td>
<td>0.3581</td>
</tr>
<tr>
<td>30</td>
<td>13639</td>
<td>0.6598</td>
</tr>
<tr>
<td>31</td>
<td>15249</td>
<td>0.4252</td>
</tr>
<tr>
<td>32</td>
<td>5319</td>
<td>1.9918</td>
</tr>
<tr>
<td>33</td>
<td>10103</td>
<td>2.7211</td>
</tr>
</tbody>
</table>
Figure 13.4: a. Gene 11041, differentially expressed for the comparison of the test compound low dose vs. the reference, $R_D^2 = 0.22$. b. Gene 5741, differentially expressed for the comparison of the test compound high dose vs. the reference, $R_D^2 = 0.84$.

Figure 13.4 shows two genes from the list of differentially expressed genes with the highest $R_D^2$ for the two comparisons. Figure 13.4b shows the scatterplot for gene 5741 that is differentially expressed for the comparison of the test compound high dose vs. the reference. The total variability of the response is 2798. The regression tree for the gene with two terminal nodes splits the data into two homogenous groups. The split point, determined by the gene-expression, is equal to 4.8. This means that observations, for which the expression level is below 4.8, form the first group, and the expression levels higher than 4.8 form the second group. The deviances are equal to 320.9 and 118.8 for the first and the second group, respectively. Note that the two groups correspond to the two treatment groups. For this gene $R_D^2 \times 100\% = 84.29\%$. This will not always be the case. Such a situation is illustrated in the left panel of Figure 13.4, which shows the scatterplot for gene 11041, which is differentially expressed for the comparison of the test compound low dose vs. the reference, for which the treatment effect upon the response is not statistically significant. For this gene the optimal split point is 4.85; the five observations at the left of the split line form one group and the other twelve observations form the second group. The total variability in the response is not much reduced by forming these two groups, what results in a low value of $R_D^2$ equal to 22.21%.
13.6.3 The Adjusted Association: Testing for Prognostic Biomarkers

Figure 13.5: Examples of two genes with the highest adjusted association $\rho$ from the joint model (13.6). The left panel a and c show the scatter plot of the gene-expression and the response, and the right panel b and d show the relationship of the gene-expression and the response after adjusting for the treatment effect, i.e., the residuals resulting from model (13.6). The observations from the reference group are denoted by pluses and the observations from the test compound high dose are denoted by circles. The adjusted association for gene 6392 and 2535 equals to 0.964 and -0.956, respectively.

For the adjusted association, the hypotheses, defined in (13.11), are tested. Using
the BH-FDR procedure for controlling the FDR of the asymptotic $p$-values obtained from the joint model, given in (13.6), there are no genes found to be significant at the FDR level of 0.05. The linear association between gene-expression and the response is not statistically significant in this dataset. However, for illustration, we present in Figure 13.5 two genes with the highest adjusted association for the case of the comparison of the test compound high dose vs. the reference. The left panel of the figure shows the scatterplot of the gene-expression and the response, and the right panel shows the relationship of the gene-expression and the response after adjusting for the treatment effect, i.e., the residuals resulting from model (13.6). The adjusted association $\rho$ obtained for the two genes is 0.964 and -0.958, respectively. However, due to the small number of observations, the power for detecting a significant association is low after the adjustment for multiple testing, although the observed association seems high.

13.7 Discussion

In this chapter we discussed methods for identifying biomarkers in the drug discovery microarray experiments. The applied approach focused on modelling the association between the gene-expression and the response, after adjusting for the treatment effect. The link was established between biomarker selection in the microarray framework and the evaluation of surrogate endpoints in the clinical trial setting. The purpose of finding biomarkers is not just limited to classify microarray samples into groups, but to predict the clinical outcome, either continuous, categorical, or survival type.

We considered two types of biomarkers, namely, prognostic and therapeutic biomarker. The prognostic biomarker is defined by the presence of the linear association between gene-expression and the response and can be evaluated using the adjusted association, i.e., the correlation coefficient adjusted for the treatment effect. The therapeutic biomarker is defined by the association between the treatment effects upon the gene-expression and upon the response and can be evaluated using the relative reduction in deviance. Both types of biomarkers can be used to predict the continuous response based on the gene-expression.

In the experiment of male Hanover Wistar rats, the response - the occupancy of target receptors in the brain - was measured under administration of multiple drugs (the test compound at low and high doses, and the reference compound). Two comparisons of drugs were considered between the reference compound and the test
compound low dose, and between the reference and the test compound high dose. This setting enabled us to search for dose-dependent alterations in gene-expression. The high dose of the test compound yields seven differentially expressed genes as compared to 31 differentially expressed genes for the low dose of the test compound. Although the number of differentially expressed genes decreased, the high dose of the test compound gave rise to the treatment effect on the response (with contrast to no effect of the low dose on the response). The high dose of the test compound activated seven genes as potential therapeutic biomarkers, and in particular, among them five genes were differentially expressed for the low dose of the test compound as well. However, the biological relevance of these genes is still to be examined by the laboratory scientists.

Moreover, the association measure of prognostic biomarkers in one comparison can be examined in the other comparison, which confirms the validity of the potential genes as prognostic biomarkers. However, in our case study, the adjusted association was not significant after multiplicity adjustment. This may be due to the fact that there was a small number of observations (i.e., eight, seven, and nine arrays for the test compound low dose, high dose, and the reference, respectively).
In Chapter 13, the joint model (13.6) was used to identify therapeutic and prognostic genomic biomarkers. In this chapter, we propose a two-stage approach to model the adjusted association between gene-expression and the response in a non-parametric way. As their relationship is not necessarily captured by a straight line, in this case, a Spearman’s correlation $r_s$ (based on ranks) can be considered to summarize a non-linear relationship between $X$ (gene-expression) and $Y$ (the response).

We consider the use of the SAM procedure for testing the Spearman’s correlation, while controlling the FDR. The advantage of the SAM procedure is its flexibility in incorporating any form of the test statistic and the use of permutations to approximate the null distribution of the test statistic.

The contents of this chapter is organized as follows. Section 14.1 introduces the two-stage modelling approach. In Section 14.2 we present the results of the application of the method to the case study. In Section 14.3 we discuss the prediction of treatment effect upon the response using therapeutic biomarkers. The chapter ends with a
discussion in Section 14.4.

14.1 The Two-Stage Modelling Approach

14.1.1 The First-stage Analysis

Following the definition of $X_{ij}$, $Y_j$, and $Z_j$ given in Section 13.4, we consider a gene-specific linear model of the form proposed by Wolfinger et al. (2001),

$$X_{ij} \sim N(\mu X_i + \alpha_i Z_j, \sigma^2_{X_i}).$$  \hspace{1cm} (14.1)

For each gene we wish to test the hypotheses $H^0_{A_i} : \alpha_i = 0$ versus $H^1_{A_i} : \alpha_i \neq 0$. Whenever the null hypothesis $H_0$ is rejected (after adjusting for multiple testing), the ith gene is declared differentially expressed. Hence, it can be used as a therapeutic biomarker. Similarly, the inference about the treatment effect for the response is based on the model

$$Y_j \sim N(\mu_Y + \beta Z_j, \sigma^2_Y).$$  \hspace{1cm} (14.2)

14.1.2 The Second-stage Analysis

The second stage of the analysis consists of estimating the adjusted association. Let $\tilde{X}_{ij}$ and $\tilde{Y}_j$ be the residuals from the models in the first stage, given by (14.1) and (14.2), respectively:

$$\begin{pmatrix} \tilde{X}_{ij} \\ \tilde{Y}_j \end{pmatrix} = \begin{pmatrix} X_{ij} - \hat{X}_{ij} \\ Y_j - \hat{Y}_j \end{pmatrix}. \hspace{1cm} (14.3)$$

We assume that the relationship between the residuals can be either linear or non-linear. In the former case we test coefficient $\gamma_i$ in the following linear regression model:

$$\tilde{Y}_j \sim N(\mu_{\tilde{Y}_i} + \gamma_i \tilde{X}_{ij}, \sigma^2_{\tilde{Y}_j}).$$  \hspace{1cm} (14.4)

In case the null hypothesis $H^0_{C_i} : \gamma_i = 0$ is rejected, once can conclude that there is a significant linear association between the response and gene-expression.

In the latter case, we test for a non-linear, but monotone, association between the response and gene-expression using Spearman’s correlation $r_{si}$ for gene $i$. Let $d_j$ be the rank-difference between $\tilde{Y}_j$ and $\tilde{X}_{ij}$. Spearman’s correlation $r_{si}$ is given by

$$r_{si} = 1 - \frac{6 \sum d_j^2}{n(n^2 - 1)},$$  \hspace{1cm} (14.5)
where \( n \) is the number of arrays. The test statistic for \( r_s \) can be written as follows (Glasser and Winter 1961):
\[
t_{r_{si}} = \frac{r_{si}}{\sqrt{(1 - r_{si}^2)/(n - 2)}}
\]
(14.6)
and can be approximated by a \( t \)-distribution when \( n > 20 \).

Note that the linear correlation \( \gamma_i \) in model (14.4) is analogous to the adjusted association obtained from the joint model (13.7).

Based on the test statistic for Spearman’s correlation \( r_s \), given in (14.6), or the linear correlation \( \gamma \), defined in (14.4), the SAM procedure can be applied to test for significance.

### 14.2 Application to the Data

In this section, we present the results of the two-stage approach applied to male Hanover Wistar rat data used in Chapter 13. In the first stage of the analysis, models (14.1) and (14.2) are fitted for the gene-expression and the response, respectively. We obtain 39 genes (for the comparison of the test compound low dose vs. the reference) and 10 genes (for the comparison of the test compound high dose vs. the reference) significant for testing \( H_{0i}^A : \alpha_i = 0 \) using the SAM procedure, while controlling the FDR at 0.05. We find that 27 genes out of the 39 genes and 6 genes out of the 10 genes are found significant using the joint model (13.6) in Chapter 13.
Table 14.1: Differentially expressed genes for the high dose of the test compound (as compared to the reference compound). Gene IDs in bold (i.e., six genes) are found in common with the results in Table 13.1 using the joint model (13.6) in Chapter 13.

<table>
<thead>
<tr>
<th>gene.id</th>
<th>α</th>
<th>t.stat</th>
<th>$R^2_D$</th>
</tr>
</thead>
<tbody>
<tr>
<td>12205</td>
<td>2.148</td>
<td>3.949</td>
<td>0.8429</td>
</tr>
<tr>
<td><strong>5319</strong></td>
<td>1.992</td>
<td>4.299</td>
<td>0.7828</td>
</tr>
<tr>
<td>3881</td>
<td>1.847</td>
<td>4.026</td>
<td>0.7318</td>
</tr>
<tr>
<td><strong>5741</strong></td>
<td>2.927</td>
<td>5.16</td>
<td>0.8429</td>
</tr>
<tr>
<td><strong>7622</strong></td>
<td>1.508</td>
<td>4.245</td>
<td>0.8429</td>
</tr>
<tr>
<td>8004</td>
<td>2.847</td>
<td>4.047</td>
<td>0.7318</td>
</tr>
<tr>
<td><strong>10103</strong></td>
<td>2.721</td>
<td>4.973</td>
<td>0.8429</td>
</tr>
<tr>
<td>1833</td>
<td>1.545</td>
<td>3.612</td>
<td>0.8429</td>
</tr>
<tr>
<td><strong>5568</strong></td>
<td>1.515</td>
<td>4.235</td>
<td>0.8429</td>
</tr>
<tr>
<td><strong>6899</strong></td>
<td>1.974</td>
<td>4.672</td>
<td>0.8429</td>
</tr>
</tbody>
</table>

Given that the significant treatment effect upon the response is only present for the comparison of the test compound high dose vs. the reference, the 10 genes declared significant (Table 14.1) are considered as potential therapeutic biomarkers.
In the second stage of the analysis, tests defined by (14.4) and (14.6) are applied to the residuals obtained from the first stage. The SAM procedure was applied to test for the significance of the linear correlation $\gamma_i$ and of the Spearman’s correlation $r_s$. After adjusting for multiple testing, using FDR=0.05, none of the null hypotheses is rejected. We therefore conclude that none of the genes can serve as a prognostic biomarker for the response. Table 14.2 shows a list of 20 genes with the highest Spearman’s correlation.

Table 14.2: Spearman’s correlations for 20 genes with the highest estimated correlation for the comparison of the low dose of the test compound and the reference. The standard error of the Spearman’s correlations is given in parentheses. Genes in bold have high Spearman’s correlation for both comparisons.

<table>
<thead>
<tr>
<th>Gene.id</th>
<th>Low dose vs. Ref</th>
<th>High dose vs. Ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 6524</td>
<td>-0.8284(0.1446)</td>
<td>-0.6794(0.1961)</td>
</tr>
<tr>
<td>2 4554</td>
<td>-0.8137(0.1501)</td>
<td>-0.4559(0.2378)</td>
</tr>
<tr>
<td>3 6527</td>
<td>0.7990(0.1553)</td>
<td>0.3441(0.2509)</td>
</tr>
<tr>
<td>4 13518</td>
<td>0.7917(0.1553)</td>
<td>0.7118(0.1877)</td>
</tr>
<tr>
<td>5 12684</td>
<td>-0.7843(0.1602)</td>
<td>-0.5853(0.2167)</td>
</tr>
<tr>
<td>6 10541</td>
<td>0.7745(0.1633)</td>
<td>0.5853(0.2167)</td>
</tr>
<tr>
<td>7 12422</td>
<td>-0.7745(0.1633)</td>
<td>-0.2147(0.2610)</td>
</tr>
<tr>
<td>8 11186</td>
<td>0.7598(0.1679)</td>
<td>0.8441(0.1433)</td>
</tr>
<tr>
<td>9 10775</td>
<td>0.7407(0.1735)</td>
<td>0.4871(0.2334)</td>
</tr>
<tr>
<td>10 2205</td>
<td>-0.7549(0.1693)</td>
<td>-0.3735(0.2479)</td>
</tr>
<tr>
<td>11 3250</td>
<td>0.7525(0.1701)</td>
<td>0.5529(0.2227)</td>
</tr>
<tr>
<td>12 8978</td>
<td>-0.7500(0.1708)</td>
<td>-0.6147(0.2108)</td>
</tr>
<tr>
<td>13 6626</td>
<td>-0.7402(0.1736)</td>
<td>-0.5529(0.2227)</td>
</tr>
<tr>
<td>14 13217</td>
<td>0.7353(0.1750)</td>
<td>0.4824(0.2341)</td>
</tr>
<tr>
<td>15 9872</td>
<td>-0.7353(0.1750)</td>
<td>-0.3088(0.2542)</td>
</tr>
<tr>
<td>16 3908</td>
<td>-0.7328(0.1757)</td>
<td>-0.2500(0.2588)</td>
</tr>
<tr>
<td>17 2375</td>
<td>0.7304(0.1764)</td>
<td>0.3647(0.2489)</td>
</tr>
<tr>
<td>18 4548</td>
<td>0.7255(0.1777)</td>
<td>0.5735(0.2189)</td>
</tr>
<tr>
<td>19 12143</td>
<td>0.7206(0.1790)</td>
<td>0.7147(0.1869)</td>
</tr>
<tr>
<td>20 11650</td>
<td>-0.7083(0.1823)</td>
<td>-0.7294(0.1828)</td>
</tr>
</tbody>
</table>
14.3 Therapeutic Biomarkers: an Intuitive Discussion

For a given treatment effect upon the response, the relative deviance reduction $R_D^2$ depends on the variability of the response, of the gene-expression, and on the treatment effect upon the gene-expression $\alpha_i$. For gene $i$, let $k_1$ and $k_2$ denote the two terminal nodes formed by the regression tree, which maximizes the value of deviance reduction, i.e., $D(Y) - D(Y|X) = D(Y) - D_1(Y|X) - D_2(Y|X)$, where $D_1(Y|X)$ and $D_2(Y|X)$ are the deviance in each group of observation formed by the regression tree. As the treatment effect $\alpha_i$ gets larger to the point that $\min(X_{ij}|Z_j = 1) > \max(X_{ij}|Z_j = 0)$, the split point in the regression tree forms exactly the two treatment groups. Thus, the two terminal nodes $k_1$ and $k_2$ correspond to the treatment indicator $Z$, and

$$R_D^2 = \frac{D(Y) - D(Y|X)}{D(Y)} = \frac{D(Y) - D_1(Y|Z = 1) - D_2(Y|Z = 0)}{D(Y)}$$

is maximized. This implies that there is a threshold level for $\alpha_i$, say $\alpha_i^*$, for which

$$\begin{cases} 
R_D^2 < R_{D\text{MAX}}^2, & \alpha_i < \alpha_i^*; \\
R_D^2 = R_{D\text{MAX}}^2, & \alpha_i \geq \alpha_i^*.
\end{cases}$$

![Figure 14.1: Gene-expression vs. the response for gene 7622 in panel a and gene 10103 in panel b.](image)
Figure 14.1 shows the scatterplot of the gene-expression vs. the response for gene 7622 and gene 10103. Gene 7622 is differentially expressed and the regression tree splits the sample into two treatment groups, while gene 10103 has a larger variance and larger treatment effect for gene-expression, resulting in an equally good $R_D^2 = 0.84$. We will use these two genes as examples to illustrate the relationship between $R_D^2$ and the treatment effect on the gene-expression, and on the variance of the gene-expression.

Figure 14.2: Plot of treatment effect vs. $R_D^2$ for gene 7622 and gene 10103. The two vertical dotted lines are the threshold treatment effects necessary to reach the maximum $R_D^2$ for the two genes. The observed threshold for treatment effects for gene 7622 and gene 10103 are indicated by a plus and a star, respectively.

Figure 14.2 shows the threshold plot for gene 10103. The observed treatment effect, $\hat{\alpha}_{10103} = 2.721$ (shown by a star), is located to the right of the threshold level, $\alpha^*_{10103} = 2.045$ (indicated by the dotted line). This is the threshold treatment effect to reach the maximum $R_D^2 = 0.84$. On the other hand, for gene 7622 the variance estimate of gene-expression is equal to 0.1815 and is smaller than the variance estimator (0.6659) for gene 10103. The observed $\hat{\alpha}_{7622} = 1.508$ (shown by a plus) is sufficient to reach the maximum $R_D^2$. Thus, genes with a smaller variance of the
expression levels require a smaller treatment effect in order to reach the maximum $R^2_D$. Note that the relationship between the treatment effect and $R^2_D$ is not exactly monotonically increasing. We observe that, when the treatment effect is too small, $R^2_D$ can decrease, because the grouping of samples resulting from the regression tree does not corresponding to the two treatment groups.

14.4 Discussion

In this chapter, we considered a two-stage approach for finding two types of biomarkers: therapeutic and prognostic. The advantage of such an approach is that it allows to model a non-linear association (using Spearman’s correlation) between the response and the gene-expression for the prognostic biomarkers. Nevertheless, no prognostic biomarkers were found after multiple testing adjustment in the case study. This is the same conclusion as we obtained from the joint model in Chapter 13.

The number of differentially expressed genes found as the therapeutic biomarkers for the comparison of the high dose of the test compound vs. the reference using SAM procedure (10 genes) is similar to that obtained (6 genes) using the joint model (13.6) in Chapter 13. Five genes were found in common using both approaches.

Moreover, we discussed the treatment effect on gene-expression to obtain the maximum value of $R^2_D$ for the therapeutic biomarkers. The value $R^2_D$ depends on the variability of the gene-expression, the response, and on the treatment effect upon the gene-expression. Given the response and a specific gene, a sufficient large treatment effect to separate observations into the two treatment group, the $R^2_D$ obtained using the regression tree is maximized. Thus, seven genes (among 10 differentially expressed genes) reached the maximum $R^2_D$ value (equal to 0.84).
Chapter 15

Constructing a Joint Biomarker Profile

The gene-specific approach, described in Chapters 13 and 14, allows to identify individual genes as biomarkers. In this chapter we focus on the question how to combine information from expression levels of individual genes into one variable, for which the joint profile may enhance the prediction of the response. We term it a “joint biomarker profile”.

The contents of this chapter is organized as follows. In Section 15.1 we introduce the method of supervised principal component analysis (SPCA) to construct the joint biomarker profile. Section 15.2 presents the results of the application to the case study. The chapter ends with a discussion in Section 15.3.

15.1 Construction of a Joint Biomarker Profile

The construction is based on the supervised principal components (SPCA) method, proposed recently by Bair et al. (2006). We use the method to construct a gene profile, that can be used in order to predict a quantitative response. The SPCA relies on the underlying assumption that there is a latent variable $U(X)$ (the gene profile), which is associated with the response variable $Y$. Note that, in contrast to the setting considered in the previous chapters, in which the biomarker (gene-expression) is observed, the joint biomarker will be a latent variable that needs to be constructed.
As for the case of individual genomic biomarkers, we will distinguish between joint prognostic and therapeutic biomarker profiles.

Following the method proposed by Bair et al. (2006), for each type of joint biomarker profile, the SPCA method consists of four steps:

- **Step 1:** Fit the gene-specific joint model (13.6) and estimate $\hat{\alpha}$ and $\hat{\rho}$.
- **Step 2:** For each type of a joint biomarker profile, form a reduced gene-expression matrix, consisting of only those genes, for which the gene-specific estimates ($\hat{\alpha}$ and $\hat{\rho}$) exceed a threshold level.
- **Step 3:** Let $X_I$ and $X_{II}$ be the reduced expression matrices for joint prognostic and therapeutic biomarker profiles, respectively. For each matrix, compute the first principal component $U(X_I)$ and $U(X_{II})$, respectively.
- **Step 4:** For each type of the joint biomarker profile, use the first principal component in order to calculate the corresponding surrogacy measure: $r_s$ for the joint prognostic biomarker profile and $R^2_D$ for the joint therapeutic biomarker profile.

The first three steps ensure that the selection of the gene subset (i.e., supervised gene screening step), from which the gene profile is calculated. We hope that the constructed joint biomarker profile maximizes the adjusted association (for prognostic biomarkers) or the relative reduction in deviance (for therapeutic biomarker biomarkers). As a consequence, for a given dataset, $U(X_I)$ and $U(X_{II})$ are the “best” joint biomarker profiles for the prognostic and therapeutic biomarkers, respectively.

For the joint prognostic biomarker we assume the following underlying model for the response and latent variable $U(X_I)$:

$$Y_i = \beta_0 + \beta_1 U(X_I)_i + \beta_2 Z_i + \varepsilon_i. \quad (15.1)$$

This model corresponds to the conditional model (13.8). As the validity measure we can use $R^2_{indiv}$ based on the regression model, or its square root $\rho$. Alternatively, Spearman’s correlation $r_s$ can be used.

For therapeutic biomarker genes, similar to the single-gene case, we use regression tree with two terminal nodes, that can be expressed as

$$Y_i = \gamma_0 + \gamma_1 I_i[U(X_{II})] + \varepsilon_i. \quad (15.2)$$
where $I_i[U(X_{II})]$ is an indicator variable, that depends on the split point ($\eta$) defining the two terminal nodes in the regression tree,

$$I_i[U(X_{II})] = \begin{cases} 
1 & U(X_{II}) > \eta, \\
0 & U(X_{II}) \leq \eta.
\end{cases}$$

15.2 Application to the Data

15.2.1 A Joint Prognostic Biomarker Profile

In the analysis presented in Section 14.2, no individual prognostic biomarkers were found after the multiple testing adjustment by using the SAM procedure.

Figure 15.1: Panel a and c: Plot of the response vs. the first PCA using the top 10 and 20 genes, respectively. Panel b and d: Plot of the response vs. the first PCA using the top 10 and 20 genes after adjusting the treatment effect (i.e., the residuals resulting from model (14.3)), respectively. The Spearman’s correlation using the top 10 and 20 genes equals $r_{s10}=0.91$ and $r_{s20}=-0.94$, respectively.
The joint prognostic biomarker profile can still be constructed based on genes with the highest Spearman’s correlation. We use top 10 and 20 genes with the highest Spearman’s correlation, obtained from the two-stage analysis, to construct a joint profile. Figure 15.1 shows the relationship between the response and the first principal component of the gene-expressions based on the top 10 and 20 genes. The Spearman’s correlation equals 0.91 and -0.94 using the top 10 and 20 genes, respectively. Although the observed Spearman’s correlation shown using these genes seems high, the validity of joint profile is still to be evaluated. For instance, the significance of the correlation obtained can be tested by approximating the null distribution of the test statistic using permutations.

15.2.2 Joint Therapeutic Biomarker Profile

As we discussed in Section 14.2, there were 10 potential therapeutic biomarkers for the comparison of the test compound high dose vs. the reference (see Table 14.1). We can use the 10 differentially expressed genes to construct a joint therapeutic biomarker profile.

![Figure 15.2: Plot of the response vs. the first PCA using 10 differentially expressed genes for the comparison of the test compound high dose vs. the reference.](image)

Figure 15.2: Plot of the response vs. the first PCA using 10 differentially expressed genes for the comparison of the test compound high dose vs. the reference.
Figure 15.2 shows a scatterplot of the first principal component of the gene-expression and the response. The $R^2_D$ achieved using these genes equals 0.84. The joint therapeutic biomarker profile does not improve the quality obtained for an individual biomarker. This is due to the fact that the maximum $R^2_D$ is reached using a single gene.

15.3 Discussion

In this chapter, we discussed the method for constructing the joint prognostic and therapeutic biomarker profile. For the therapeutic biomarker, the joint profile was constructed using 10 differentially expressed genes. As the individual therapeutic biomarker reached the maximum $R^2_D$ (0.84), the joint profile yielded the same maximum $R^2_D$ value. The gain in constructing the joint profile was not evident in this case study.

With no individual prognostic biomarkers found, the joint prognostic biomarker profile can still be constructed. Based on the top 10 and 20 genes, the Spearman’s correlations (i.e., 0.91 and -0.94) were obtained. However, the significance of the joint profile needs to be tested. The main question is whether the correlation of 0.9 is a “large” value and can be obtained using an independent dataset. The cross-validation of the joint biomarker profile is the topic for future research. The validity of joint biomarker profile needs to be determined further. The selection of the optimal number of genes to be used and the validity of the joint prognostic biomarker profile also deserves future research.

The contents of Chapters 13, 14, and 15 were summarized in the manuscript (Lin et al. 2008).
Chapter 16

Concluding Remarks and Future Research

In this dissertation we considered four microarray settings related to the testing of thousands of genes simultaneously. Each setting covers a topic on the applications of microarray in the drug discovery.

16.1 Part I: Class Prediction

The first part of the dissertation was an application of microarrays to diagnosis. Specifically, in Chapters 3 and 4, we evaluated the performance of various classification and gene-selection methods using simulations. The merits of simulated data enable us to control the experimental settings and to examine the selected genes. We observed that the gene-selection method that leads to the largest percentage of truly differentially expressed genes does not necessarily lead to the lowest misclassification rate. The best choice of the classification method does not appear to depend strongly on the gene-selection procedure. In general, gene-selection methods, such as Wilcoxon, prediction strength (PS), prediction analysis of microarray (PAM), between-within (BW), and extreme-value distribution based gene selection (Extval) lead to quite good results as compared to the other investigated methods. DLDA and random forests perform better than the other classification methods in our setting.

In the simulation study, we have considered the univariate gene-selection methods.
There are many existing multivariate gene-selection procedures, in which the correlation of genes are taken into account. Performance of multivariate gene-selection methods deserves further research.

16.2 Part II: Multiple Testing Adjustment

The second part of the dissertation focused on an application of microarrays to pharmacology, in which we determined differences in gene-expression in samples exposed to various doses of compounds. To this aim, in Chapter 5, 6, and 7, we discussed several approaches (Dunnett asymptotic p-values adjusted by the BH-FDR procedure, permutation p-values adjusted by the BH-FDR procedure, the SAM with and without the fudge factor) for comparing several treatments with one control. We found that under the small sample size setting (three arrays per group), the SAM approach without the fudge factor led to a higher power as compared to the SAM procedure with the fudge factor. None of the studied methods controlled the FDR when small variance genes were present. Moreover, in a simulation study we investigated the choice of the fudge factor used by different methods like, for example, the penalized t, EB, and ROC. The results of these methods were not satisfactory, because none of them controlled the FDR at the desired level due to the increasing proportion of the small variance genes.

The drawback of these methods is that the correction made to the test statistic weakens the significance of truly differentially expressed genes while guarding against the truly non-differentially genes with small variances. Thus, the estimation of the proportion of small variance genes is of great importance. Mixture models can be considered to estimate this proportion. How to find a unified approach to correct the test statistics using the information obtained about the proportion of small variance genes deserves more research.

16.3 Part III: Dose-response Study

The third part of the dissertation focused on an application of microarrays to finding monotone gene-expression patterns under a set of increasing doses. To this goal, in Chapter 8, we investigated an order restricted dose-response relationship by testing the homogeneity of means against order restricted alternatives. In particular, we drew our attention to calculation of the two-sided p-values and the directional inference for
five test statistics. When rejecting the null hypothesis, the direction of the trend needs to be determined based on the $p$-values for an increasing and a decreasing trend. We argued that, due to the standard error estimation used in Williams’ and Marcus’ tests, these tests are more likely to reject the null hypothesis of no dose effect than the $M$ and modified $M$ when there is a non-monotone dose-response relationship.

In Chapter 10, we classified genes found significant for the order restricted alternatives into possible shapes of dose-response curves using the information theory. The number of dose levels in our experiment was limited to four, thus, the information criteria, such as $AIC$, $BIC$ and $ORIC$ led to the selection of the dose response curve from seven possible order restricted curves. However, as the number of doses in the experiment increases, the number of possible dose response curves, which can be fitted on the data, also increases. We suggest to consider parametric models to study the dose-response relationship. For example, a four parameter logistic model can be used to explore a sigmoidal shape. Once, a set of all possible models is fitted, the information criteria can be used to assess the choice of the best dose-response curve.

In Chapter 11, we formulated a ratio test of the means under the order restricted alternatives between the highest dose and the control. By efficient use of MCTs, the main question of interest was translated into classical multiple ratio contrast tests. The inference about such ratios is based on a multivariate $t$-distribution, while the estimation of the multiple ratios can be challenging and is beyond the scope of this dissertation. In this chapter, we only considered the “plug-in” method by replacing the unknown ratio parameters with the ratio estimator from the data in the correlation matrix of the test statistics proposed by Dilba et al. (2005).

Chapter 12 provided an application of the FCR in the microarray setting. Based on the selected genes from the ratio test, we constructed FCR-adjusted confidence intervals for the ratio parameters, which control the non-coverage rate of the null parameters for the null hypothesis and the control of the FDR using the BH-FDR procedure. The goal of this application was to emphasize the need for adjusting multiplicity in constructing CIs in the microarray setting.

It is of great importance to extend the order restricted inference to multiple dimensions in the microarray setting. We considered the multiple testing problems using the human epidermal squamous carcinoma cell line data, described in Section 1.4.1, in two dimension separately, i.e., for comparisons of several treatments with one control, and for dose-response relationship. In target selectivity of drug development, we should consider the question: how does the dose-response relationship in
one treatment differ from those for other treatments? A model that takes into account the marginal effect of two factors have been studied by Robertson et al. (1988). However, due to the complexity of gene-expression patterns, the interaction between the monotonic dose-response relationship and treatments is inevitable. Our future research will focus on modelling the order restricted constraint in the ANOVA framework, on defining a suitable test for the interaction, and on deriving the appropriate inference for the test statistic.

16.4 Part IV: Evaluation of Biomarker

The fourth part of the dissertation focused on an application of microarrays in biomarker determination. As the promising microarray technology emerges, genes show their great potential as genetic biomarkers for the clinical outcome. Not only biomarkers can be used as features for class prediction as we discussed in Chapter 3 and 4, but also they can be used to predict various clinical measurements. For example, continuous responses as the brain occupancy in our case study, or survival measurements. The use of biomarkers is not only limited to prediction of the response, but can also be considered for the prediction of treatment effects. However, the validation and evaluation of biomarkers remains an important question.

In Chapters 13 and 14, we established the link between the evaluation of surrogate endpoints in the clinical trial setting with the selection of biomarkers in the microarray setting. The aim was to identify two types of potential biomarkers: therapeutic and prognostic. We modelled the relationship between gene-expression and the response for two comparisons, namely, the low dose of the test compound vs. the reference and the high dose of the test compound vs. the reference. No therapeutic biomarkers existed because of the absence of the treatment effect for the first comparison. Some potential therapeutic biomarkers, which were found for the second comparison, deserved biological confirmation. Due to small sample size in the case study data, no prognostic biomarkers were found after adjusting for multiple testing.

In Chapter 15, the biomarker framework was extended by defining the concept of a joint biomarker profile using several genes. The idea of a joint profile is to enhance the information from individual biomarkers. However, it is not always possible for the joint therapeutic biomarker profile to achieve a higher predictive power if a single therapeutic biomarker reaches the maximum $R^2_D$.

For our future research, to avoid selecting genes in the two comparisons separ-
rately, a model, which takes into account the three groups simultaneously, should be considered.

In the case study, we considered a continuous response. Methods for selection of biomarkers for binary, categorical, survival response should be explored. We need to define/find different validity measures for biomarkers using various types of responses. Moreover, the quality and stability of the joint biomarker profile should be further investigated. Simulation studies can be set up to investigate the performance of joint biomarker profile.

Any individual potential biomarker or a joint biomarker profile should be validated using independent datasets. It is worth noting that multi-trial microarray data are lacking. A meta-analysis of microarray data is of great importance for increasing statistical power and achieving consistent biomarker selection.

16.5 Final Remarks

To end this dissertation, we are quoting Butte (2002) to tell the unending story of microarray data analysis for drug discovery:

“The use of microarrays in basic and applied research in drug discovery is only going to increase, but as these data sets grow in size, it is important to recognize that untapped information and potential discoveries might still be present in existing data sets. It should be clear that any set of microarray measurements could be analysed and re-analysed in many different ways. In the application of functional genomics to drug discovery, to extract the most information from microarrays, an open mind always needs to be kept with regard to the choices of analytical methods, using supervised and unsupervised techniques, and methods yet to come.

In the future, truly showing a return on investment from functional genomics will depend on taking findings beyond the microarray stage and integrating them with the rest of the discovery pipeline. The list of genes resulting from a microarray analysis should not be viewed as an end in itself; its real value increases only as that list moves through biological validation, ranging from the numerical verification of expression levels with alternative techniques, to ascertaining the meaning of the results, such as finding common promoter regions or biological relationships between the genes. However, tools that link these genes back to known biological pathways, as well as discovering new pathways, are in their infancy. Tools that can automatically indicate the importance of particular findings have yet to be invented. Until they come into
being, the analysis of microarray datasets in a vacuum devoid of biological knowledge will be less rewarding.”
Publications and Reports


227


References

*Affymetrix Santa Clara, CA.*


*Affymetrix Santa Clara, CA.*


229


Li, C and Hung Wong, W (2001b) Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biology*, **2**(8), research 0032.1–0032.11.


Williams, D.A. (1971) A test for differences between treatment means when several dose levels are compared with a zero dose control, *Biometrics, 27*, 103–117.


Appendix A

Simulation Results

This appendix supplements the simulation in Section 7.2. Figure A.1 and Figure A.4 show the relationship between the FDR (including FDR\(^0\) and FDR\(^1\)) and power for the settings of five and six arrays per group, respectively. Figure A.2, Figure A.3, Figure A.5, and Figure A.6 plot the distribution of the FDR (including the FDR\(^0\) and FDR\(^1\)) and power using boxplots under the settings of five and six arrays per group, respectively.
Figure A.1: The setting of five arrays per group. a: FDR vs. power; b: FDR\(^0\) (small variance genes) vs. power; c: FDR\(^1\) vs. power. A, B, C, and D denote results for the SAM procedure with the fudge factor under the four settings of different proportions of the small variance genes. 1, 2, 3, and 4 denote the results for the permutation approach adjusted by the BH-FDR procedure under the four settings. Black line: setting 1 \((m_0^1=0)\); red line: setting 2 \((m_0^1=500)\); green line: setting 3 \((m_0^1=1000)\); blue line: setting 4 \((m_0^1=2000)\).
Figure A.2: The setting of five arrays per group. a: Boxplots of the FDR for the SAM procedure without the fudge factor ($-s_0$) and with the fudge factor ($+s_0$), and for permutation p-values adjusted by the BH-FDR procedure; b: boxplots of the FDR$^0$ (small variance genes); c: boxplots of the FDR$^1$. Black line: setting 1 ($m_0^1=0$); red line: setting 2 ($m_0^1=500$); green line: setting 3 ($m_0^1=1000$); blue line: setting 4 ($m_0^1=2000$).
Figure A.3: The setting of five arrays per group. Boxplots of power for the SAM procedure without the fudge factor (−s0) and with the fudge factor (+s0), and for permutation p-values adjusted by BH-FDR procedure. Setting 1, 2, 3, and 4: m^1_0 = 0, 850, 1700, and 3400, respectively. Black line: setting 1 (m^1_0=0); red line: setting 2 (m^1_0=500); green line: setting 3 (m^1_0=1000); blue line: setting 4 (m^1_0=2000).
Figure A.4: The setting of six arrays per group. a: FDR vs. power; b: FDR^0 (small variance genes) vs. power; c: FDR^1 vs. power. A, B, C, and D denote results for the SAM procedure with the fudge factor under the four settings of different proportions of the small variance genes. 1, 2, 3, and 4 denote the results for the permutation approach adjusted by the BH-FDR procedure under the four settings. Black line: setting 1 (m_0^1=0); red line: setting 2 (m_0^1=500); green line: setting 3 (m_0^1=1000); blue line: setting 4 (m_0^1=2000).
Figure A.5: The setting of six arrays per group. a: Boxplots of the FDR for the SAM procedure without the fudge factor ($-s_0$) and with the fudge factor ($+s_0$), and for permutation p-values adjusted by the BH-FDR procedure; b: boxplots of the FDR$^0$ (small variance genes); c: boxplots of the FDR$^1$. Black line: setting 1 ($m_0^1=0$); red line: setting 2 ($m_0^1=500$); green line: setting 3 ($m_0^1=1000$); blue line: setting 4 ($m_0^1=2000$).
Figure A.6: The setting of six arrays per group. Boxplots of power for the SAM procedure without the fudge factor (−s₀) and with the fudge factor (+s₀), and for permutation p-values adjusted by BH-FDR procedure. Setting 1, 2, 3, and 4: \( m_0^1 = 0, 850, 1700, \) and \( 3400, \) respectively. Black line: setting 1 (\( m_0^1=0 \)); red line: setting 2 (\( m_0^1=500 \)); green line: setting 3 (\( m_0^1=1000 \)); blue line: setting 4 (\( m_0^1=2000 \)).
Appendix B

Simulation Results

In this appendix we include the results of simulation study in Section 10.5.2 for the setting of correlated data (correlation equal to 0.9). Table B.1 presents the results of the misclassification and correct classification rates for the first approach, i.e., classification without the initial inference step. Table B.2 presents the results of the misclassification and correct classification rates for the second approach, i.e., the initial inference step followed by classification step.
Table B.1: Misclassification and correct classification rates. First approach: classification without the initial inference step. The setting induces the correlation (0.9) for gene-expression of 250 genes under each model and each dose level. L=Likelihood, A=AIC, B=BIC, O=ORIC, NA=Not Available.

<table>
<thead>
<tr>
<th></th>
<th>Error</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L</td>
<td>A</td>
</tr>
<tr>
<td>g1</td>
<td>NA</td>
<td>0.7288</td>
</tr>
<tr>
<td>g2</td>
<td>NA</td>
<td>0.7383</td>
</tr>
<tr>
<td>U</td>
<td>g3</td>
<td>NA</td>
</tr>
<tr>
<td>P</td>
<td>g4</td>
<td>0.9294</td>
</tr>
<tr>
<td></td>
<td>g5</td>
<td>0.8804</td>
</tr>
<tr>
<td></td>
<td>g6</td>
<td>0.8473</td>
</tr>
<tr>
<td></td>
<td>g7</td>
<td>0.8071</td>
</tr>
<tr>
<td>g1</td>
<td>NA</td>
<td>0.7401</td>
</tr>
<tr>
<td>g2</td>
<td>NA</td>
<td>0.7389</td>
</tr>
<tr>
<td>D</td>
<td>g3</td>
<td>NA</td>
</tr>
<tr>
<td>O</td>
<td>g4</td>
<td>0.8916</td>
</tr>
<tr>
<td>W</td>
<td>g5</td>
<td>0.8983</td>
</tr>
<tr>
<td>N</td>
<td>g6</td>
<td>0.8971</td>
</tr>
<tr>
<td></td>
<td>g7</td>
<td>0.8087</td>
</tr>
<tr>
<td>null</td>
<td>NA</td>
<td>0.0000</td>
</tr>
<tr>
<td>non-mon</td>
<td>0.7617</td>
<td>0.0437</td>
</tr>
</tbody>
</table>
Table B.2: Misclassification and correct classification rates. Second approach: the initial inference step followed by classification step. The setting induces the correlation (0.9) for gene-expression of 250 genes under each model and each dose level. L=Likelihood, A=AIC, B=BIC, O=ORIC, NA=Not Available.

<table>
<thead>
<tr>
<th></th>
<th>Error</th>
<th></th>
<th></th>
<th></th>
<th>Correct</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L</td>
<td>A</td>
<td>B</td>
<td>O</td>
<td></td>
<td>L</td>
<td>A</td>
<td>B</td>
<td>O</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>g1</td>
<td>0.1369</td>
<td>0.1751</td>
<td>0.1872</td>
<td>0.1524</td>
<td>0.3333</td>
<td>0.9218</td>
<td>0.9327</td>
<td>0.9139</td>
<td></td>
</tr>
<tr>
<td>g2</td>
<td>0.1941</td>
<td>0.3168</td>
<td>0.3619</td>
<td>0.2749</td>
<td>0.2186</td>
<td>0.9114</td>
<td>0.9262</td>
<td>0.8929</td>
<td></td>
</tr>
<tr>
<td>g3</td>
<td>0.1362</td>
<td>0.1559</td>
<td>0.1864</td>
<td>0.1304</td>
<td>0.3279</td>
<td>0.9436</td>
<td>0.9474</td>
<td>0.9244</td>
<td></td>
</tr>
<tr>
<td>g4</td>
<td>0.6131</td>
<td>0.2412</td>
<td>0.2391</td>
<td>0.2353</td>
<td>0.3432</td>
<td>0.7161</td>
<td>0.7087</td>
<td>0.5946</td>
<td></td>
</tr>
<tr>
<td>g5</td>
<td>0.5229</td>
<td>0.1009</td>
<td>0.0773</td>
<td>0.0809</td>
<td>0.4633</td>
<td>0.7322</td>
<td>0.7125</td>
<td>0.6431</td>
<td></td>
</tr>
<tr>
<td>g6</td>
<td>0.5421</td>
<td>0.3353</td>
<td>0.3685</td>
<td>0.2266</td>
<td>0.5800</td>
<td>0.7899</td>
<td>0.7265</td>
<td>0.7530</td>
<td></td>
</tr>
<tr>
<td>g7</td>
<td>0.7162</td>
<td>0.2440</td>
<td>0.2144</td>
<td>0.4796</td>
<td>0.8290</td>
<td>0.4295</td>
<td>0.3663</td>
<td>0.6558</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.1971</td>
<td>0.7594</td>
<td>0.7576</td>
<td>0.7597</td>
<td>0.3458</td>
<td>0.9424</td>
<td>0.9611</td>
<td>0.9271</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.1998</td>
<td>0.2607</td>
<td>0.2939</td>
<td>0.2323</td>
<td>0.2007</td>
<td>0.8730</td>
<td>0.9020</td>
<td>0.8217</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.1113</td>
<td>0.1451</td>
<td>0.1620</td>
<td>0.1350</td>
<td>0.3792</td>
<td>0.9536</td>
<td>0.9536</td>
<td>0.9307</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.8672</td>
<td>0.1401</td>
<td>0.1342</td>
<td>0.1050</td>
<td>0.5106</td>
<td>0.8755</td>
<td>0.8681</td>
<td>0.7190</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.5520</td>
<td>0.1483</td>
<td>0.1153</td>
<td>0.1081</td>
<td>0.4224</td>
<td>0.7686</td>
<td>0.7464</td>
<td>0.6392</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.5446</td>
<td>0.2787</td>
<td>0.3278</td>
<td>0.2068</td>
<td>0.4384</td>
<td>0.7810</td>
<td>0.7592</td>
<td>0.6310</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.6959</td>
<td>0.1820</td>
<td>0.1583</td>
<td>0.4871</td>
<td>0.9033</td>
<td>0.5042</td>
<td>0.4011</td>
<td>0.7758</td>
<td></td>
</tr>
<tr>
<td>null</td>
<td>NA</td>
<td>0.0000</td>
<td>0.0002</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.8659</td>
<td>0.9245</td>
<td>0.8659</td>
<td></td>
</tr>
<tr>
<td>non-mon</td>
<td>0.7802</td>
<td>0.3171</td>
<td>0.2035</td>
<td>0.3171</td>
<td>0.8019</td>
<td>0.8019</td>
<td>0.8019</td>
<td>0.8019</td>
<td></td>
</tr>
</tbody>
</table>
Functionele genomica is het onderzoek van de functie van genen op basis van simultane meting van duizenden genen van een genoom. De meest gebruikte hulpmiddelen om deze metingen uit te voeren, omvatten complementaire DNA (cDNA) microarrays, oligonucleotide microarrays of seriële analyse van genexpressies (SAGE). In deze thesis concentreren we ons op de eerste twee technieken die belangrijke hulpmiddelen zijn voor het onderzoek naar nieuwe geneesmiddelen.

Het gebruik van microarrays in het onderzoek naar geneesmiddelen breidt uit. Toepassingen zijn (Butte 2002): (1) voorspellende testen, om een groep van genen te vinden die nauwkeurig de ene ziekte van de andere kan onderscheiden, om de subklasse van een ziekte te bepalen, om meervoudige subcategorieën van tumoren in een enkele diagnose te vinden; (2) target selectiviteit, om de eigenschappen van een chemische stof te bepalen op basis van de veranderingen in genexpressie in het weefsel dat men wil onderzoeken en om zo ook verschillende chemische stoffen te vergelijken; (3) pharmacologie, om de verschillen in genexpressie te bepalen in weefsels die blootgesteld worden aan verschillende dosissen van een chemische stof; (4) biomarker bepaling, om genen te vinden die correleren met voortekenen van ziekteprogressie, maar eenvoudiger zijn om te meten en op te volgen in klinische proeven.

Deze thesis is toegewijd aan de analyse en verwerking van microarray data, in het bijzonder in de vier toepassingsgebieden hierboven vermeld en de volgende vier types van microarray experimenten waarin volgende vragen worden beschouwd: (1) wanneer genexpressie gemeten wordt voor twee behandelingen, welke classificatie methode presteert dan het best om de arrays in twee groepen te classificeren? (2) wanneer genexpressie gemeten wordt voor verschillende behandelingsgroepen en één controle groep, welke statistische methode kan dan gebruikt worden? (3) wanneer genexpressie gemeten wordt voor toenemende dosissen, wat is dan de relatie tussen genexpressie en dosis? Wat is de minimale dosis om effect te hebben (MED) en wat
is de vorm van dosis-responsie curve? Wat is het relatieve verschil tussen de hoogste dosis en de controle dosis? (4) wanneer zowel genexpressie als klinische resultaten gemeten worden voor stalen die twee behandelingen kregen, welke genen kunnen dan dienen als biomarkers om de klinische resultaten te voorspellen?

In hoofdstuk 2 starten we met een korte introductie in de microarray technologie en geven we een samenvatting van de methodes om microarray data te analyseren. Twee types microarrays, cDNA arrays en oligonucleotide arrays, worden besproken. De statistische analyse van microarrays omvat de verwerking van de gescande genexpressie beelden, de significantie testen om differentieel uitgedrukte genen te vinden en gecontroleerde en ongecontroleerde methodes voor klasse voorspelling.

Het vervolg van deze thesis is opgebouwd uit vier delen. Elk deel behandelt de toepassing van microarrays in de eerste fasen van het onderzoek naar nieuwe geneesmiddelen.

Het eerste deel van de thesis is een toepassing van microarrays voor het maken van voorspellingen. Meer specifiek, in de hoofdstukken 3 en 4 hebben we, gebruikmakend van simulatie studies, de prestaties geëvalueerd van verschillende classificatie en gen selectie methodes om arrays te classificeren in twee groepen. De voordelen van gesimuleerde data is dat ze ons toelaten om de instellingen te beheren en de geselecteerde genen te onderzoeken. We hebben gemerkt dat de gen selectie methodes die leiden tot het hoogste percentage van werkelijk differentiële genen, niet noodzakelijk leiden tot de laagste missclassificatie ratio, en dat de beste keuze van de classificatie methode niet sterk afhankelijk blijkt te zijn van de gen selectie procedure. In het algemeen leiden Wilcoxon, Predictive Strength (PS), Prediction Analysis of Microarray (PAM), Between and Within ratio (BW) en Extreme-value distribution based gene selection (Extval) tot vrij goede resultaten onder de onderzochte gen selectie methodes en DLDA en Random Forest presteren beter dan de andere classificatie methodes.

Het tweede deel van de thesis is een toepassing van microarrays in farmacologie. Dat wil zeggen, om verschillen in genexpressies te bepalen in stalen die blootgesteld worden aan verschillende dosissen van chemische stoffen. Daarom hebben we in de hoofdstukken 5, 6 en 7 de verschillende aanpakken besproken (Dunnett p-vaardes aangepast door de BH-FDR procedure, permutatie p-vaardes aangepast door BH-FDR procedures, SAM met en zonder fudge factor) om vergelijkingen van verschillende behandelingen te testen met één controle en om, gebruik makend van simulaties, de prestaties van deze methodes te vergelijken in termen van FDR controle en onderscheidingsvermogen. Alle vier de methodes die we voorstelden combineren
het veelvoudige testen van “meervoudig naar één” vergelijkingen binnen en tegenover genen tegelijk, en voornamelijk in de laatste drie methodes steunen we op de hersampling gebaseerde conclusie om de correlatie te behouden van de test statistieken voor de “meervoudig naar één” vergelijkingen. De hersampling gebaseerde methode kan nuttig zijn in deze opstelling omdat de verbinding distributie (multivariatie t, Dunnett 1955) van de teststatistieken onder meervoudige vergelijkingen tussen verschillende behandelingen met de controle wordt bemoeilijkt door het testen van duizenden genen tegelijk. Het probleem van het regelen van de meervoudige testen is moeilijker wanneer de verbinding distributie van de teststatistieken onbekend is.

In onze studie hebben we opgemerkt dat bij kleine steekproefgrootte (drie arrays per groep), de SAM aanpak zonder fudge factor leidt tot een groter onderscheidingsvermogen in vergelijking tot SAM mt fudge factor, terwijl alle methodes de FDR niet controleren wanneer genen met lage varianties aanwezig zijn. Bovendien hebben we simulatie studies uitgevoerd om de keuze van de fudge factor te bepalen, gebruikt door de verschillende methodes, bijvoorbeeld de penalized t, EB, en ROC - maar de resultaten van deze methodes waren voor geen enkele methode voldoende om de FDR op het gewenste niveau te controleren. Dit is te wijten aan de verhoogde proportie van genen met een lage variantie. Het nadeel van deze methodes is dat de correctie van de test statistieken het significantieniveau van de echte differentiële expressiegenen vermindert, terwijl we beschermen tegen het selecteren van echte niet-differentiële genen met kleine varianties.

Het derde deel van deze thesis is een toepassing van microarrays in targetselec-tie die focust op het vinden van het genexpressie patroon onder toenemende dosissen. Met dit doel hebben we in hoofdstuk 8 een dosis-responsie relatie onderzocht. Bij het vergelijken van de werking van de vier bestaande test statistieken voor het testen van de homogeniteit van de gemiddelden ten opzichte van order restrictive alternatieven, hebben we een aanpassing van de standaard fout van de MS test statistiek voorgesteld, terwijl onze simulatie studie een gelijk onderscheidingsvermogen getoond heeft voor de globale waarschijnlijkheidstest (\(E_{01}^2\)), de M en aangepaste M test. In hoofdstuk 9 hebben we de implementatie aangetoond van de test statistieken en meervoudige test procedures in R. De R bibliotheek IsoGene is beschikbaar voor R gebruikers om test statistieken en procedures voor de controle van de FDR te kiezen. In hoofdstuk 10 classificeren we de genen verder in mogelijke dosis-responsi grafieken, gebruik makend van de informatietheorie. Het aantal dosissen die beschouwd worden in dit experiment is beperkt tot vier. De informatiecriteria zoals AIC, BIC en
ORIC leiden tot de selectie van de beste dosis-responsie curven van de zeven mogelijke curven. Echter, als het aantal dosissen in het experiment stijgt, zullen er teveel mogelijke dosis-responsie curven zijn die fitten voor de gegevens. Het is aangeraden om parametrische modellen te overwegen om de dosis-responsie relatie te bestuderen. Bijvoorbeeld: een vier-parameter logistisch model kan gebruikt worden om een sigmoidale vorm te onderzoeken. Zodra alle mogelijke modellen gefit zijn, kunnen de informatie criteria gebruikt worden om de keuze van de beste dosis-responsie curve te bepalen. In hoofdstuk 11 hebben we de onderzoeks vraag opnieuwgeformuleerd op een alternatieve manier, namelijk een verhoudingstest van de relatieve gemiddelden tussen de hoogste dosis en de controle. Gebaseerd op de geselecteerde genen hebben we in hoofdstuk 12 simultane betrouwbaarheidsintervallen geconstrueerd voor de verhoudingsparameters, die rekening houden met de False Coverage statement Rate (FCR) die garandeert dat de non-covering rate van de nulparameters getest zijn volgens de BH procedure in de geselecteerde genen.

Het vierde deel van de thesis is een toepassing van microarrays in biomarker bepaling. Aangezien veelbelovende microarray tools opduiken, tonen genetische biomarkers hun potentieel als surrogaat eindpunten voor de klinische resultaten. De evaluatie van biomarkers blijft echter een belangrijk vraagstuk. In hoofdstukken 13 en 14 hebben we een link gelegd tussen de evaluatie van de surrogate eindpunten in de klinische testopstellingen en de selectie van biomarkers in de microarray setting en hebben we zowel het joint model als de twee-stage modellen toegepast om potentieel twee types van biomarkers te identificeren: therapeutische en voorspellende. Aangezien ze dienen om de werkelijke eindpunten te voorspellen in termen van respectievelijk correlatie en behandelingseffect, zijn de twee types van biomarkers geëvalueerd, gebruikmakend van de aangepaste associatie vanuit het perspectief van de individuele test niveaus en de vermindering in afwijking $R^2_D$ vanuit het perspectief van de informatie theorie. In hoofdstuk 15 wordt het biomarker framework uitgebreid om een joint biomarker profiel te construeren, gebruik makend van de geselecteerde genen. Het idee van een joint profiel is de informatie van de individuele potentiële biomarkers te versterken. Het is echter niet altijd mogelijk om een hogere voorspellende kracht te bekomen indien één enkele therapeutische biomarker de maximale $R^2_D$ bereikt, terwijl de correlatie meting verzwakt ten gevolge van de steekproefgrootte en de meervoudige test aanpassing.

In de toekomst zal een echt resultaat van de investering in functionele genomics afhangen van bevindingen die verder gaan dan microarrays, en deze te integreren met de rest van de ontwikkelingen. De lijst van genen die resulteren uit microarray analyse
moet niet als een eindpunt gezien worden op zichzelf. Hun echt belang vergroot enkel als de lijst biologisch wordt gevalideerd, gaande van de numerische verificatie van expressie niveaus met alternatieve technieken, om zich te verzekeren van de betekenis van de resultaten, zoals het vinden van gemeenschappelijke promotors regionen van biologische relaties tussen genen. Middelen die deze genen linken met hun biologisch pad, alsook het ontdekken van nieuwe paden, zijn echter in hun beginstadium. Middelen die automatisch de belangrijkheid kunnen aantonen van bepaalde bevindingen, moeten nog uitgevonden worden. Tot zolang zal de analyse van microarray data sets verstoken blijven van biologische kennis, en minder lonend zijn (Butte 2002).