Data-adaptive test statistics
for
gene expression arrays

Sachi Nandan Mukherjee
Robotics Research Group
Department of Engineering Science
University of Oxford
December, 2005

This thesis is submitted to the Department of Engineering Science, University of Oxford, for the degree of Doctor of Philosophy. This thesis is entirely my own work, and except where otherwise indicated, describes my own research.
Data-adaptive test statistics for gene expression arrays

Abstract

This thesis proposes a data-adaptive approach to the problem of selecting differentially expressed genes from microarray data. The approach is rooted in machine learning, statistics and decision theory, and makes use of a novel objective function to learn test statistics directly from experimental data. Gene expression arrays or microarrays are a biochemical methodology which allow the expression levels of tens of thousands of genes to be measured simultaneously. In the decade or so since their introduction, microarrays have fundamentally altered the way in which research in molecular biology is conducted, and have found applications in virtually every area of the biological sciences. Microarrays are frequently used to compare gene expression levels between contrasting conditions, such as wild-type and mutant, or healthy and diseased, in the hope of discovering genes which may be of relevance to an underlying biological process or phenomenon. As a consequence, the statistical task (“gene selection”) of selecting genes which are differentially expressed between conditions has emerged as a problem of considerable practical importance. The thesis contributes to the literature on gene selection in two ways. Firstly, gene selection is characterised as a statistical decision problem and analysed within a decision-theoretic framework. This analysis highlights both the difficulties that arise on account of the statistical characteristics of microarray data, as well as some of the shortcomings of existing methods for gene selection. Secondly, motivated by these concerns, a novel, data-adaptive approach for gene selection is proposed. Experiments on real and synthetic data demonstrate the ability of the approach to mitigate many of the problems associated with existing methods and produce accurate results under a range of conditions. Furthermore, the theoretical underpinnings of the methodology introduced are quite general in nature, and the approach is likely to find further applications in bioinformatics and elsewhere.
To my late father
Dr. Ajit Kumar Mukherjee
“Tomar janya”
Publications

Some parts of this thesis have previously appeared elsewhere, as listed below:


Notation

The following notational conventions are used throughout this thesis. Lower-case letters, for example $x$, represent ordinary variables, or instantiations of random variables. In the latter case, the corresponding upper-case letter, for example $X$, represents the random variable of which $x$ is an instantiation. Vectors are denoted in boldface, such that $x$ is a vector, or instantiation of random vector $X$. Matrices are also represented in bold upper-case; it is hoped that it will be clear in context whether a bold upper-case letter represents a matrix or random vector. In some cases, a symbol will refer to a function as well as its output. For example, suppose $T$ is a function of random data $X$, we may write $T = T(X)$. The upper-case letter $P$ represents a probability mass function, while lower-case $p$ represents a probability density function. In both cases, when used with different arguments, these symbols should be understood to refer to different functions, such that $p(x)$, for example, represents the distribution of $x$, and $p(y)$ the distribution of $y$. The symbol $\sim$ is used to mean ‘is distributed according to’, such that $X \sim p$ means that random variable $X$ has $p$ as its distribution. Similarly, $\overset{\text{iid}}{\sim}$ is used to mean ‘are independent and identically distributed according to’. Finally, integrals written as $\int_A \cdot dx$ represent integration with respect to the variable $x$, over the entire set $A$.

The following symbols occur frequently in this thesis:

- $X$ Random sample
- $E$ Expectation operator
- $S$ Sample space
- $a$ Action
- $\mathcal{A}$ Action space
- $\omega$ State of nature
- $\Omega$ Space of all possible states of nature
- $\theta$ Statistical parameter
- $\Theta$ Parameter space
When discussing the selection of differentially expressed genes from microarray data, data obtained under two different biological conditions are represented using the letters $X$ and $Y$, and the subscript ‘$k$’ is used to refer to quantities pertaining to the $k^{th}$ gene in a dataset:

$H_{0k}$  
Null hypothesis for gene $k$

$H_{1k}$  
Alternative hypothesis for gene $k$  

$T_k$  
Value of the test statistic for gene $k$

$\{X_{ik}\}_{i=1}^m$  
$m$ expression levels for gene $k$, under the first condition

$\{Y_{jk}\}_{j=1}^n$  
$n$ expression levels for gene $k$, under the second condition

$\mu_{Xk}, \mu_{Yk}$  
Population means for gene $k$, under each of two conditions

$\bar{X}_k, \bar{Y}_k$  
Sample means for gene $k$, under each of two conditions

$\hat{\sigma}^2_{Xk}, \hat{\sigma}^2_{Yk}$  
Unbiased sample variances for gene $k$, under each of two conditions

$SD_k$  
Unbiased pooled sample standard deviation for gene $k$

$\rho$  
“Reproducibility”

The numbers of genes of various kinds are denoted by the following variables:

$g$  
Total number of genes under study

$g_0$  
Number of non-differentially expressed genes

$g_1$  
Number of differentially expressed genes

$S$  
Total number of genes selected

$S_0$  
Number of non-differentially expressed genes selected

$S_1$  
Number of differentially expressed genes selected
Finally, some standard probability distributions are represented as follows:

- $\mathcal{N}$ Normal distribution
- $\mathcal{G}$ Gamma distribution
- $\mathcal{B}$ Binomial distribution
Acknowledgements

First, I would like to thank my supervisor Stephen Roberts, for his guidance, expertise and patience, but most of all for giving me the freedom to develop my own ideas, and the confidence to “follow my nose”. Thanks also to my co-supervisor Sarah Gurr for help and advice at various times. I have benefited greatly from interacting with various members of the Information Engineering Division at Oxford. There are too many people to mention individually, but I am especially grateful to Nick Hughes, Lyndsey Pickup, lead Rezek, Peter Sykacek and Andrew Zisserman. I am also very grateful to Terry Speed and his group at the Department of Statistics, U. C. Berkeley for welcoming me as a visitor in Spring 2005, and also to Mark van der Laan. I would also like to sincerely thank my examiners for the time and effort they put into giving my work a thorough reading, and for their many excellent, constructive comments: the thesis is much the better for their input.

My DPhil research was funded by the BBSRC, and research and conference trips by St. John’s College, Oxford, the Royal Academy of Engineering, and the International Society for Computational Biology: I am very grateful to all of them.

A huge thank you as always to my mother, brother and the rest of my family, they have been tremendously supportive of me through the years. Finally, I would like to thank my wife Sarah Hirsch. Her love, patience and understanding of my (many) foibles never cease to amaze me, she truly is the most important person in my life. Thank you.
Contents

1 Introduction ........................................... 13
   1.1 Quantitative biology ................................... 13
   1.2 The genome ........................................... 16
   1.3 Post-genomic biology .................................... 19
      1.3.1 The organisation of the genome ................. 19
      1.3.2 Individual sequence variation .................... 21
      1.3.3 Evolutionary sequence variation ................. 21
      1.3.4 Genomic function .................................. 22
   1.4 Motivation of the thesis ................................ 26
   1.5 Contributions of the thesis ............................ 28
   1.6 Outline of the thesis ................................... 30

2 Gene expression and microarrays ......................... 31
   2.1 Gene expression ........................................ 33
   2.2 Microarray technology .................................. 35
      2.2.1 Reverse transcription and hybridisation ........ 38
CONTENTS

2.2.2 cDNA microarrays ........................................... 39
2.2.3 Oligonucleotide arrays ...................................... 42
2.3 Limitations of microarray studies ............................. 44

3 Microarray data analysis ........................................ 47
  3.1 Image processing and normalisation ......................... 49
  3.2 Predictive analysis ........................................... 51
  3.3 Discovery-driven analysis .................................... 52
    3.3.1 Clustering .................................................. 55
    3.3.2 Dimensionality reduction ................................. 56
    3.3.3 Network inference ......................................... 57
  3.4 Analysis of differential expression ......................... 58
    3.4.1 A statistical view of differential expression .......... 60
    3.4.2 Hypothesis testing as a paradigm for differential expression 61
    3.4.3 Test statistics for differential expression .............. 64
    3.4.4 Open questions in differential expression .............. 71

4 A decision-theoretic analysis ................................ 73
  4.1 Statistical decision theory ................................ 75
    4.1.1 Samples, actions and decision procedures ............ 75
    4.1.2 States of nature and the loss function ................ 76
    4.1.3 Risk .......................................................... 77
  4.2 Gene selection as a decision procedure .................... 78
  4.3 Loss functions for gene selection .......................... 81
## CONTENTS

4.3.1 Zero-one loss ...................................... 82
4.3.2 False Discovery Rate ............................... 83
4.3.3 ROC curves .............................. 84

4.4 Risk in gene selection .......................... 87

4.5 Computing risk from sampling distributions .......... 90
  4.5.1 Preliminaries ..................................... 90
  4.5.2 FDR ........................................... 97
  4.5.3 Zero-one loss .................................. 99
  4.5.4 AUC ........................................ 100

4.6 Results ........................................ 101
  4.6.1 Conventional normal model .................... 102
  4.6.2 Larger sample sizes ........................... 104
  4.6.3 Differing variances ............................ 106
  4.6.4 Varying numbers of non-differentially expressed genes .. 110

4.7 Discussion .................................... 110

5 Data-adaptive test statistics .......................... 114
  5.1 Risk-optimality in gene selection .................. 116
  5.2 Reproducibility .................................. 117
    5.2.1 Definition ................................... 118
    5.2.2 Risk and reproducibility ..................... 119
    5.2.3 Violating i.i.d. assumptions .................. 125
    5.2.4 Estimating reproducibility .................... 129
CONTENTS

5.3 Data-adaptive test statistics ................................................. 132
5.4 Results ........................................................................ 135
   5.4.1 Synthetic data .......................................................... 135
   5.4.2 Microarray data .......................................................... 148
5.5 Discussion ................................................................... 151

6 Conclusions ..................................................................... 155
   6.1 Summary of research ...................................................... 156
   6.2 Uncertainty in quantitative molecular biology .................. 158
   6.3 Further work ................................................................. 161

A Hypothesis testing .............................................................. 177

B Distribution of threshold given number of genes selected ...... 182

C Reproducibility ................................................................. 184
List of Figures

1.1 Schematic depiction of a DNA molecule and its base pairs . . . . 17
1.2 Genome, transcriptome and proteome . . . . . . . . . . . . . . 23

2.1 Gene expression and its regulation . . . . . . . . . . . . . . . . . 34
2.2 A generic process diagram for microarray analysis . . . . . . . . 37
2.3 Pseudo-image obtained from a cDNA array . . . . . . . . . . . . 41
2.4 Pseudo-image obtained from an oligonucleotide array . . . . . . 43

3.1 Predictive microarray analysis . . . . . . . . . . . . . . . . . . . . 51
3.2 Discovery analysis of microarray data . . . . . . . . . . . . . . . 53
3.3 Discovery analyses of various kinds . . . . . . . . . . . . . . . . . 54

4.1 ROC curve for gene selection . . . . . . . . . . . . . . . . . . . . . 84
4.2 Comparing gene selection procedures using ROC curves . . . . 86
4.3 Verification of $\beta(s) = \int \beta(\tau)p(\tau \mid s) \, d\tau$ . . . . . 95
4.4 Risk under a conventional Normal model . . . . . . . . . . . . . 103
4.5 Risk plotted against sample size . . . . . . . . . . . . . . . . . . 105
LIST OF FIGURES

4.6 Risk when variances differ ......................... 107
4.7 Risk plotted against number of non-differentially expressed genes 109

5.1 Why reproducibility correlates with risk .................. 120
5.2 Correlation between risk and reproducibility under i.i.d. conditions 127
5.3 Correlation between risk and reproducibility under a hierarchical model .................. 128
5.4 Correlation between risk and reproducibility under a hierarchical model, with a highly variable prior for gene-level means 130
5.5 Data-generating model for synthetic data .................. 136
5.6 Prior distributions for means and variances .................. 137
5.7 Results using synthetic data with identical variance priors .................. 139
5.8 Learned parameters with identical variance priors .................. 140
5.9 Results using synthetic data with differing variances, first case .................. 141
5.10 Results using synthetic data with differing variances, second case 142
5.11 Learned parameters with differing variances .................. 144
5.12 Non-normal data-generating density .................. 145
5.13 Results using synthetic, non-normal data .................. 146
5.14 Learned parameters from non-normal data .................. 147
5.15 Results using benchmark microarray data .................. 149
5.16 Learned parameters from benchmark microarray data .................. 150
5.17 A large number of non-differentially expressed genes is beneficial to gene selection using reproducibility .................. 153


LIST OF FIGURES

6.1 Model uncertainty, small sample size and unlabelled data. . . . . 159
Chapter 1

Introduction

1.1 Quantitative biology

Biology is in the process of becoming a quantitative information science. Increasingly, living organisms are being seen as systems, analogous to engineering systems, which are amenable to mathematical description and analysis (Ideker et al., 2001; Kitano, 2002). The long-term goal of a systems approach in biology is to develop a quantitative, predictive understanding of living organisms, eventually at a level of detail comparable to engineering descriptions of man-made systems.

The practical significance of a truly quantitative, predictive biology would be enormous. For example, research into new drugs remains largely a matter of trial-and-error, with the system-wide effects of new treatments usually revealed only by difficult and expensive clinical trials. Equally, medical practice remains based largely on correlations and accumulated experience, rather than true understand-
In contrast, engineers can approach the “diagnosis” and “treatment” of a man-made system, like a malfunctioning chemical plant, for example, with far greater certainty. The availability of detailed plans, coupled with an understanding of underlying physical phenomena, mean that the root causes of a malfunction in an engineering system can often be deduced from observed behaviour, and the effects of various interventions predicted ahead of time. In principle, a systems understanding of biology would enable a new kind of analytical medicine, in which both diagnosis and treatment would be informed by detailed mathematical models, and as a consequence, would be far more effective, and less unpredictable, than is the case today.

Yet the challenge of developing a predictive understanding of living beings is a formidable one. Biological systems are exquisitely complex, and being evolved rather than human-designed, must be “reverse engineered” to produce system-level descriptions which can in turn be used to make predictions regarding behaviour (Csete and Doyle, 2002). Reverse engineering a complex system is difficult under the best of conditions, but in the domain of biology matters are further complicated by the difficulty inherent in describing or measuring living systems. Indeed, the tasks of cataloguing the components of a biological system, or of measuring system variables, are themselves huge challenges. Thus, while the deep similarities between artificial and biological systems have been appreciated since at least the late 1940s (Wiener, 1948), only in recent times has it started to become possible to exploit this understanding in a practical way.

Biological systems and engineered machines differ in two very fundamental
respects. Firstly, biological systems exhibit organisation and adaptability across a huge range of scales, from metres (the whole organism), down to nanometres (individual cell organelles and macromolecules). Indeed, research in biology has made it very clear that much of the organisation of living systems is at the molecular level. Indeed, from an engineering point of view, a biological system may be regarded as a complex assembly of molecular machines, co-ordinated to give rise to various higher-level behaviours. In contrast, man-made systems tend to be organised at a higher level, and may even be quite homogenous at smaller scales. For example, a robot arm and human arm may appear similar at the macro level, but while the robot arm is essentially homogenous and static at smaller scales, the human arm is made up of billions of cells, each of which is a dynamic, adaptive entity in its own right. Among other things, this remarkable degree of adaptability confers on living beings the ability of “self-maintenance”, whereby damage sustained can be repaired autonomously, using nutrients obtained from the surrounding environment.

A second major difference lies in the fact that living organisms are capable of development. Every human individual starts as a single cell called the zygote, which makes use of nutrients in the surrounding environment to build itself up into a complex organism comprising around $10^{13}$ cells. The process by which a single cell develops into this vast, exquisitely co-ordinated assembly of cells is remarkably autonomous. Given the right conditions, it simply happens, and can therefore reasonably be regarded as a process of “self-assembly”.

These two differences have important implications for the study of biological
systems. The manner in which living systems are organised means that molecular events are often key drivers of macro properties, to the extent that it becomes crucial to understand biology from the molecular level upwards. Indeed, a key trend in biology, throughout the second half of the twentieth century, has been a move towards describing biological phenomena at the molecular level. Equally, the remarkable autonomy with which organisms develop and repair themselves is made possible by the fact that these systems contain coded internal representations of themselves.

1.2 The genome

The processes of self-assembly and self-maintenance alluded to above require vast amounts of information to function correctly. For example, a growing embryo needs a detailed “development plan” if it is to successfully assemble simple nutrients into a fully functioning human being.

The information processing framework within which living organisms function is remarkably consistent across the natural world. Inside every living cell lies a quantity of a substance called deoxyribonucleic acid or DNA which encodes information needed for the development and normal functioning of the cell. An organism’s entire complement of genetic information is called its genome. The genome is, in effect, a coded representation of the organism, and is therefore fundamentally important to understanding biological function.

The structural characteristics of DNA are intimately related to its role as a
CHAPTER 1. INTRODUCTION

repository of biological information. A DNA molecule is double-stranded, resembling a ladder twisted into a helix, as illustrated in Figure 1.1. The ‘sides’ of the ladder are made up of sugar and phosphate, and the ‘rungs’ made up of paired nucleotide bases, joined by hydrogen bonds. Each strand of the molecule comprises a sugar-phosphate backbone and a sequence of nucleotide bases. The hydrogen bonds between the bases bind the two strands together.

Figure 1.1: A schematic depiction of part of a DNA molecule and its base pairs.

The human genome contains around $3 \times 10^9$ pairs of nucleotide bases. Nucleotide bases are themselves complex organic molecules, each made up of a pentose sugar (deoxyribose in the case of DNA), a phosphate group and an organic base. While every nucleotide in a DNA molecule has the same sugar and phosphate groups, the base may be one of four distinct types: Adenine, Thymine, Cytosine and Guanine, which are usually represented by the letters ‘A’, ‘T’, ‘C’ and ‘G’ respectively. The bases exhibit a very specific pattern of bond formation,
‘A’ joining only with ‘T’, and ‘C’ only with ‘G’. As a consequence of this specificity of base pairing, the two strands of a DNA molecule are complementary to one another: wherever ‘A’ occurs in one strand, ‘T’ will occur in the other, and vice versa; wherever ‘C’ occurs in one strand, ‘G’ will occur in the other, and vice versa. The two strands are thus mirror images of each other and share the same information content. The order in which bases occur in a particular DNA molecule is called its sequence. The sequence is a physical encoding of the information content of the molecule, and can be thought of as a long word formed from the alphabet {A, T, C, G}.

The double-helical structure of DNA was discovered in 1953 (Watson and Crick, 1953), and in the decades that followed came a series of remarkable advances in the ability of scientists to perform biological experiments at the molecular level. The central importance of elucidating the full genetic code was widely appreciated from the 1960s onwards, and the basic principles of DNA sequencing were understood in the late 1970s (Sanger et al., 1977), but technological limitations meant that the sequencing of an entire genome remained an unrealistic prospect until a decade later, when a U.S. National Research Council report (National Research Council, 1988) suggested that a “special project” be undertaken with the goal of sequencing the human genome. This led, in 1990, to a massive international collaboration called the Human Genome Project. The bulk of the sequencing was done at five centres in the United States and the United Kingdom, and a “working draft” of the Human Genome was published in 2001, in special issues of the journals Nature and Science (Lander et al., 2001; Venter et al., 2001).
The project was officially completed on April 14th 2003, coincidentally 50 years, to the month, after Watson and Crick’s seminal discovery of the structure of DNA. Automated sequencing has now developed to the point that it is relatively easy to rapidly sequence various genomes, and indeed the full genomes of several hundred organisms have now been sequenced, with many more to follow.

1.3 Post-genomic biology

The advent of whole-genome sequencing has fundamentally changed the way in which biological research is conducted, inspiring a “post-genomic” revolution which has had an impact in virtually every area of the biosciences. In recent years, an enormous amount of work has been done in exploiting the information provided by newly available genome sequences to better understand living systems, and progress continues at a very rapid pace. This Section presents an overview of post-genomic biology. Following Collins et al. (2003), we organise this vast research area into four partially over-lapping themes involving the study of the organisation of the genome, individual sequence variation, evolutionary sequence variation, and overall genomic function.

1.3.1 The organisation of the genome

The genome encodes structurally and functionally important molecules called proteins, as well as many of the regulatory programs which govern their synthesis. Genes are parts of the genome which encode individual proteins, and in addi-
tion can exist in alternative forms called alleles. We may loosely take a gene to be a (not necessarily contiguous) subsequence of the genetic code which specifies the amino-acid sequence of a protein. A gene is further divided into subsequences called exons which explicitly code for proteins, and intervening regions called introns which do not code for proteins. Comprehensive gene sequences are now available for many species of interest, yet the structure and organisation of these sequences remain poorly understood. A comprehensive system-wide analysis must start with a catalogue of system components, or a “parts-list”. The process of understanding the organisation of the genome can be thought of as leading towards the compilation of a genome-level “parts-list”.

Raw sequence information does not, in itself, reveal which parts of the code represent genes. The task of determining which regions of the sequence represent protein-coding genes and other functional components (e.g. splice sites, start and stop codons, promoters etc.) is called gene finding. Computational and statistical methods have come to play a central role in gene finding (Gelfand, 1995; Fickett, 1996). In general, such methods seek to identify genes and functional components on the basis of the statistical properties of relevant subsequences. Hidden Markov Models and Hidden semi-Markov Models in particular have emerged as important tools in gene finding (Krogh et al., 1994; Burge and Karlin, 1997). Research in gene finding is now fairly mature, with catalogues of genes available for most species of interest. However, many genes remain mis-characterised, and gene finding remains an important area of research.

However, only about 1-2 % of the human genome actually codes for proteins
(Lander et al., 2001). The precise function of the remainder of the genome remains remarkably poorly understood, and represents an important area for current and future research (Collins et al., 2003).

### 1.3.2 Individual sequence variation

Each individual has his or her own gene sequence, which may differ from the reference sequence represented by the published human genome. Such variation plays a key role in determining individual phenotypic characteristics, including, for example, susceptibility to various diseases. As a consequence, characterising and understanding individual sequence variation has emerged as a key area of research in genomics. A major international effort, called the “International HapMap Project” (Gibbs et al., 2003) has been initiated with the aim of studying such variation, with a particular emphasis on elucidating the relationship between sequence variation and human disease.

### 1.3.3 Evolutionary sequence variation

Gene sequences from different organisms are related by evolution. As a consequence, subsequences of the genetic code are often conserved between species, with the degree of conservation being related to the closeness of the evolutionary relationship between the species. Understanding sequence variation across species and through evolution is of considerable scientific interest in itself, but can also play an important role in other fields of biological enquiry (Collins et al.,
A key focus of research in this area is molecular phylogenetics or the task of reconstructing evolutionary relationships between species using sequence information. Molecular phylogenetics involves inferring a graph structure depicting the relationships between various species from available sequence data. Phylogenetic inference must take account of noise in observations and the fundamentally random nature of evolution itself and indeed state-of-the-art procedures for phylogenetic inference (Husmeier et al., 2005) are largely probabilistic in nature.

### 1.3.4 Genomic function

The genome is essentially a static information store, but the manner in which its information is utilised is extraordinarily dynamic. Living cells are able to activate and de-activate genes in response to environmental factors and biological signals, and thereby regulate cellular levels of different proteins. The process effecting the transfer of information from genetic sequence to protein product is called gene expression. Sequence information is first transferred into an intermediate form called messenger RNA or mRNA in a process called transcription. The mRNA transcript is then used as a template for protein synthesis, in a process called translation. The task of understanding how genetic information gives rise to biological phenomena, via transcripts, proteins and small molecules is arguably the central challenge of post-genomic biology.

In analogy to the term “genome”, the complete set of possible transcripts in an organism is called its transcriptome, and the complete set of possible proteins is
called the *proteome*. The relationship between these three levels of organisation is shown in Figure 1.2. The pyramid shape of the figure symbolises the fact that each successive layer is more complex than the last. Mechanisms such as the alternative splicing\(^1\) of genes and post-translational modification of proteins mean that a single gene may give rise to many different transcripts, and a single transcript may give rise to many different protein products. Mechanisms of this kind play a central role in the regulation of biological function, but make the study of the transcriptome and especially the proteome extremely challenging.

![Figure 1.2: A schematic depiction of the relationship between the layers of biological organisation represented by the genome, transcriptome and proteome, adapted from de Hoog and Mann (2004).](image)

An important step in understanding genomic function is characterising the functions of individual genes. Around 30,000 genes have now been identified in the human genome, but the biological roles of most of these genes remain un-

\(^1\)Protein-coding exons are spliced together to form mature transcripts. This splicing together of exons may occur in various ways, such that one gene may give rise to several different mRNAs, a phenomenon known as *alternative splicing*. 
known. Quantitative studies of gene expression play an important role in inferring gene function. For example, knowing that a certain gene is activated only in liver cells, and only under certain conditions, implicitly provides information concerning function.

Experimental methods to detect and quantify transcription have been in existence for several decades. However, until recently, such analyses were restricted to the study of a single gene, or at most a small number of genes. During the 1990s, the advent of whole-genome sequencing, coupled with a series of technological advances, started to enable massively parallel assays of gene expression. This led to the development of the gene expression array or microarray by Pat Brown and colleagues in the mid-1990s (Schena et al., 1995). Microarrays allow tens of thousands of mRNA levels to be measured simultaneously. This means that they can be used to measure the mRNA abundances of every gene in an organism in a single experiment, in effect enabling “system-wide” studies of gene expression. In the decade or so since their introduction, microarrays have had a profound effect on research in molecular biology, and have found applications in virtually every area of the biological sciences. This thesis is concerned mainly with statistical methods for microarray analysis. Accordingly, a more detailed look at transcription and microarrays is presented in the next chapter.

Transcriptional profiling using microarrays has proved to be an effective approach to system-wide biological investigation, but the complex relationship between transcripts and functional proteins means that such studies inevitably provide only a limited view of the complete picture. It has become clear that post-
translational modifications of proteins, such as phosphorylation, glycosylation and methylation, play a key role in many areas of biological function (de Hoog and Mann, 2004). The large-scale study of protein levels, or proteomics (de Hoog and Mann, 2004) has therefore emerged as a major area of research in post-genomic biology. Key technologies in proteomics include mass spectrometry (Aebersold and Mann, 2003), and various array-based systems (Patterson and Aebersold, 2003).

The study of genomic function is further complicated by the fact that genes and proteins do not usually function in isolation, but rather as parts of complex molecular networks. Information in molecular networks flows “vertically” from gene through transcript to protein, and also “horizontally”, for example through protein-protein interactions. Clearly, a full understanding of biological function requires analyses which are integrated both vertically and horizontally, taking full account of the relationship between various levels of genomic organisation and of molecular networks. Yet, as noted by Collins et al. (2003), such analyses are “far more complex than any problem that molecular biology, genetics or genomics has yet approached”. Accordingly, recent efforts in integrative, post-genomic biology have focused on either simple model organisms (Banerjee and Zhang, 2002), or relatively well-understood systems, such as growth-factor signalling, in mammals (Wiley et al., 2003).
1.4 Motivation of the thesis

The transcriptome is intermediate in complexity between the genome and proteome, and technologies for transcriptional profiling are now reasonably mature. Furthermore, while transcription itself represents only one part of overall gene function, transcriptional regulation is sufficiently complex as to provide fertile ground for biological discovery. For these reasons, the study of transcription has proved to be a key focus of research in post-genomic biology.

As noted above, microarray technology in particular has been widely adopted, and is now routinely used in all areas of the biosciences. Microarrays are very often used to compare gene expression levels between contrasting conditions, such as wild-type and mutant, or healthy and diseased, in the hope of discovering genes which may be of biological relevance. For example, knowing that mRNA corresponding to a specific set of genes is greatly over-represented in tissues affected by a certain form of cancer may provide information regarding molecular mechanisms underlying the cancer. Indeed, comparative experiments of this kind have become a dominant mode of use for microarrays (Stoughton, 2005). The task of discovering genes which are differentially expressed between conditions has therefore emerged as a key statistical problem in microarray data analysis, and is the main topic of this thesis. We will refer to this task as gene selection.

The genome-wide nature of microarray studies means that microarrays represent a powerful tool for quantitative molecular biology. However, the complexity of the relationship between transcription and overall system behaviour means that
comparative microarray experiments must typically be followed up by more detailed studies using other, complementary genomic technologies, including proteomics and more sensitive transcriptional assays such as real-time RT-PCR\textsuperscript{2}. The results of gene selection play an important role in guiding further experimental work of this kind. Furthermore, the considerable expense involved in such studies mean that errors in gene selection can be extremely costly in terms of wasted resources and missed opportunities. Accurate gene selection is therefore critically important in the context of microarray-based studies.

However, microarray data have a number of characteristics which make their analysis difficult. Firstly, the data display a remarkable mismatch between dimensionality and sample size. A dataset may contain expression levels for tens of thousands of genes, but will typically have no more than ten samples for each gene. Secondly, microarray data are highly variable, on account of both experimental noise and biological variability. Thirdly, the statistical distributions underlying microarray data tend to be poorly characterised. Taken together, these factors make the reliable selection of differentially expressed genes a challenging task.

Gene selection is usually cast as a problem in statistical hypothesis testing (Pan, 2002; Speed, 2003). We will argue that while hypothesis testing is a good paradigm for the study of differential expression, conventional hypothesis testing procedures are not sufficiently flexible to deal with the model uncertainty that is inherent in the microarray domain. The practical importance of accurate gene

\textsuperscript{2}Real time reverse transcription polymerase chain reaction (Bustin, 2002)
selection and the difficulties inherent in analysing microarray data motivate the need for a principled approach to gene selection which is capable of producing reliable results under practically realistic conditions of small sample size, high-dimensionality and model uncertainty. These requirements - of principle, robustness and computing in the presence of uncertainty - provide the primary motivation for this thesis.

A second, broader, motivation is provided by the importance of hypothesis testing problems in the analysis of post-genomic data. The need for robust analysis under statistically difficult conditions is by no means confined to the microarray domain. It is therefore hoped that the work presented here may prove useful in post-genomic data analysis more generally.

The choice of differential expression as a problem domain is motivated by two factors. Firstly, as noted above, gene selection is a practically important problem, so methods proposed here may be of immediate utility to the bioinformatics community. Secondly, transcriptional analysis is relatively mature and differential expression in particular has been widely studied, making it a good domain in which to pursue a fuller, principled analysis.

### 1.5 Contributions of the thesis

This thesis addresses the concerns raised above by first analysing the conventional hypothesis testing approach to gene selection within a decision-theoretic framework, and then exploiting the understanding so obtained to develop a data-
adaptive procedure for gene selection. The original contributions of the thesis can therefore be organised very naturally into two themes.

The first theme concerns the development of a decision-theoretic framework for gene selection. To the best of our knowledge, this work is the first explicit application of statistical decision theory to the differential analysis of microarrays. Decision theory provides a sound theoretical framework for our analysis, allowing us to deal in a principled manner with the inherently random nature of gene selection, while taking full account of factors such as small sample size and high-dimensionality.

The second theme involves learning effective test statistics from microarray data. Here, we propose a novel scheme by which test statistics may be learned directly from experimental data, using a simple notion of reproducibility in selection results as the learning criterion. Reproducibility, as we define it, can be computed without any knowledge of the ‘ground-truth’, but takes advantage of some of the properties of microarray data to provide a useful proxy for decision-theoretic risk. We are therefore able to use reproducibility to indirectly minimise risk under the true data-generating distribution, and thereby deal with model uncertainty in a simple but effective manner. Our work marries a computationally-intensive, learning-based approach to a classical hypothesis testing paradigm, resulting in an overall procedure which is substantially more robust than conventional hypothesis testing methods under conditions relevant to microarray data analysis.
1.6 Outline of the thesis

This thesis is organised as follows. The next chapter, Chapter 2, provides an introductory look at transcription and microarray technology. Chapter 3 presents a broad overview of existing methods for the statistical analysis of microarray data. The first three chapters together lay the ground for the original work presented in Chapters 4 and 5. Chapter 4 proposes a decision-theoretic analysis of gene selection, and includes a comparative study of three widely-used procedures, under a number of practically relevant conditions. Chapter 5 then presents a data-adaptive procedure for gene selection, along with empirical results examining the ability of the procedure to discover truly differentially expressed genes from real and simulated microarray data. Finally, Chapter 6 draws conclusions and discusses possible directions for further work.
Chapter 2

Gene expression and microarrays

At the level of an individual organism, the genome is essentially a static entity, in that the gene sequence of an individual does not vary significantly through time or between cells. Yet cells differ in every way imaginable, structurally and functionally, and combine to form tissues and organs which are mostly highly specialised. Furthermore, cells display sophisticated control mechanisms at molecular scales, operating within millisecond timeframes. A pertinent question then is this: how does such heterogeneity and adaptability arise from a single code?

A large part of the answer lies in the process of gene expression by which information is transferred from genes to proteins. In a particular cell, at any given time, only a small fraction of the information available in the genome is actually used to make proteins. Furthermore, the quantities of individual proteins produced vary through time and between cells, thereby giving rise to dynamic behaviour from a static genome. This dynamic utilisation of genetic information
CHAPTER 2. GENE EXPRESSION AND MICROARRAYS

plays a central role in the processes of differentiation and development. As cells of different kinds develop, through repeated divisions of a common precursor cell, individual genes are switched on and off in response to developmental programs, leading to the structural and functional specialisation observed in fully developed cells and tissues. Equally, the responses of individual cells to environmental circumstances take place largely through the selective synthesis of proteins. Thus, when a cell detects the presence of a particular chemical substance, or receives a biological signal via a hormone or signalling molecule, it can, in response, synthesise required proteins by activating the appropriate genes, in effect adapting its chemical contents and physical structure to changing conditions.

The fundamental biological roles played by the dynamic utilisation of genetic information place gene expression at the heart of many of the most interesting and important questions in biology and medicine. This Chapter provides an introduction to the biology of gene expression, and to the microarray technology which has, in recent years, allowed gene expression to be investigated on a system-wide scale. Further details regarding gene expression can be found in textbooks such as Alberts et al. (2002) and Strachan and Read (1999), while a number of excellent review papers (including Duggan et al., 1999; Schulze and Downward, 2001; Stoughton, 2005) and books (Baldi and Hatfield, 2002; Speed, 2003) provide a more detailed treatment of microarrays.
2.1 Gene expression

Genomic information is encoded in biochemical form as deoxyribonucleic acid or DNA. As we saw previously, DNA is a polymer having four distinct types of nucleotide bases, Adenine, Thymine, Cytosine and Guanine, signified by the letters ‘A’, ‘T’, ‘C’ and ‘G’, such that a DNA molecule can be thought of as a biochemical encoding of a word formed from the alphabet \{A, T, C, G\}.

Proteins are large, complex molecules which play a central role in biological function. They are important structural constituents of living beings as well as active agents in control and signalling processes. Proteins are built up from smaller units called amino-acids. Amino-acids are small organic molecules, containing a carboxyl group (COOH) and an amino group (NH$_2$) both attached to the same carbon atom. Amino-acids are joined together by covalent carbon-nitrogen bonds (known as ‘peptide-bonds’) to form long chains called polypeptides, which in turn fold into three-dimensional configurations to form proteins. A given amino-acid sequence may be able to fold into a number of stable spatial configurations, such that there is a many-to-one relationship between proteins and polypeptides.

Each amino-acid is specified by a three-letter DNA code called a codon. Each position in the codon can be occupied by one of four possible nucleotide bases, so there are $4^3 = 64$ possible codons. Since the total number of different amino-acids is only twenty, there is some redundancy in the coding scheme which maps codons to amino-acids, such that more than one codon may correspond to the same amino-acid. For example ‘AUU’, ‘AUC’ and ‘AUA’ all encode the amino
acid isoleucine. Remarkably, the same twenty amino-acids occur in every living being. Indeed, even the mapping from triplet codons to amino-acids is conserved across species.

Figure 2.1: Gene expression and its regulation. Regulatory mechanisms exist at every stage, from transcription through translation to activation.

Protein synthesis starts with the copying of genetic information from DNA into a molecule called messenger RNA, or mRNA. This is known as transcription. Like DNA, RNA has four bases, but has a base called Uracil in place of Thymine. During transcription, DNA strands separate and a molecule of mRNA is formed with a sequence complementary to that of the DNA. This molecule of mRNA is called a transcript, and contains the same information as its DNA template. Following transcription, the mRNA molecule moves into the cytoplasm, and is ‘read’ by a polypeptide making device called a ribosome. The ribosome
assembles\textsuperscript{1} the required amino-acids and links them together to form a polypeptide chain, as specified by the mRNA transcript. The formation of a polypeptide chain from a mRNA transcript is called \textit{translation}. Transcription and translation together constitute gene expression.

At a conceptual level, the pathway from genes to proteins is simple enough: DNA passes its information to mRNA which is then used by ribosomes to make polypeptides, which fold into proteins. However, these processes are subject to a remarkable amount of regulation and control. Figure 2.1 summarises the various levels at which protein synthesis is regulated. The biological details of these regulatory mechanisms will not concern us here, but it is worth emphasising the sheer complexity of the processes underlying transcription and translation.

\section{2.2 Microarray technology}

Gene expression plays a central role in every aspect of biological function. As a consequence, quantitative studies of gene expression are vitally important in understanding biological phenomena.

As we have seen, proteins are the main active agents in cellular function. As such, measurements of protein levels represent a natural way to characterise cellular state. However, proteins are inherently unstable, and furthermore, may exist in many alternative forms due to post-translational modifications (PTMs). As a consequence of these factors, protein levels have proved technically difficult to

\textsuperscript{1}Another type of RNA called \textit{transfer RNA} or \textit{tRNA} is intimately involved in ribosomal protein synthesis, but its role will not concern us here.
measure on a large-scale.

In contrast, transcript abundances are far more amenable to large-scale, quantitative study (Stoughton, 2005), and furthermore correlate fairly well with protein levels. The study of transcription has therefore been a major focus of research in molecular biology for many decades (Alberts et al., 2002). Prior to the advent of microarrays, analyses of transcriptional activity were highly labour-intensive, and individual experiments were restricted to the study of a single gene, or at most a small number of genes. Large-scale assays of gene expression only started to become possible in the 1990s, as the availability of comprehensive sequence information, coupled with a series of technological advances, started to allow investigators to mount thousands of unique DNA molecules in an orderly fashion on glass slides and other surfaces (Schena et al., 1995; Lockhart et al., 1996).

There are now many different microarray platforms available, but they are all based upon the same underlying concept. Known genetic material is immobilised at specific locations on a surface, such as a glass slide or silicon wafer, and then allowed to interact with unknown, but fluorescently tagged, DNA representations of mRNA obtained from biological samples of interest. The unknown samples are called targets and the known samples immobilised on the surface are called probes\(^2\). The specificity of binding between nucleic acids causes unknown targets to bind preferentially to complementary probes. This means that following probe-

\[^2\text{In this thesis, we follow the convention adopted by Duggan et al. (1999) and others in the January 1999 supplement to Nature Genetics ("The Chipping Forecast"). It is worth noting, however, that some authors use the term ‘probe’ to mean the unknown sample and ‘target’ to refer to the arrayed material.}\]
target binding, the intensity of fluorescence emitted at each location on the surface can be used to quantify the abundance of each target mRNA which was present in the biological sample.

Figure 2.2: A generic process diagram for microarray analysis, after Stoughton (2005).

The two most widely-used types of microarray are cDNA and oligonucleotide arrays. The probes in cDNA arrays are made up of large fragments of DNA, whereas oligonucleotide arrays have as their probes short DNA subsequences called oligonucleotides. These two technologies differ in a number of important ways, but the basic steps involved in their use are quite similar, and are summarised as a generic process diagram in Figure 2.2.

This Section presents an overview of microarray technology. We first describe two biochemical processes - reverse transcription and hybridisation - which un-
derpin microarray technology, and go on to discuss cDNA and oligonucleotide arrays in more detail. We then look briefly at some of the limitations of microarray technology.

2.2.1 Reverse transcription and hybridisation

The correspondence between DNA and RNA sequences is such that any RNA sequence corresponds to a specific DNA sequence and vice versa. Given a quantity of mRNA extracted from a tissue of interest, it is possible to synthesise a DNA molecule whose sequence is complementary to that of the extracted mRNA. This process is called reverse transcription and the synthesised DNA called a complementary DNA or cDNA. Thus, DNA can be transcribed into RNA, and RNA reverse transcribed into DNA:

\[
\begin{array}{c}
\text{DNA} \\
\text{RNA}
\end{array}
\]

\[
\begin{align*}
\text{Transcription} & \quad \text{Reverse transcription} \\
\text{DNA} & \quad \text{RNA}
\end{align*}
\]

Hybridisation refers to the binding together of two single strands of DNA or RNA. DNA is double stranded, with the two strands held together by hydrogen bonds between paired nucleotide bases. Subjecting a DNA molecule to high temperatures causes these bonds to be broken and the two strands to come apart. The temperature at which the strands separate is usually 65° C or more, and is known as the melting temperature. If the temperature is lowered to below the melting temperature, the individual strands bind back to their counterparts. Furthermore,
if DNA is brought to its melting temperature, a single stranded RNA, which has a
sequence complementary to one strand of the DNA, will hybridise to that strand,
and do so in such a way that prevents the binding back of the remaining strand of
DNA when the temperature is lowered.

2.2.2 cDNA microarrays

The probes in cDNA microarrays are made up of DNA molecules generated from
known sequences corresponding to genes of interest. cDNA arrays were intro-
duced in Schena et al. (1995).

cDNA arrays are usually prepared in the laboratory. First, a set of genes to be
investigated is decided upon, following which a subsequence of the coding region
of each gene, typically around 1,000 bases in length, is used to synthesise a DNA
probe for that gene. The DNAs are amplified using the polymerase chain reaction
(PCR) and then printed onto a glass slide or nylon membrane using a high-speed
robot. Each probe is typically only a few hundred micrometres in size, such that
around 30,000 probes can be placed on a standard glass slide. Probes are placed in
specific, known locations on the slide, and the quantity of DNA in each location
is made approximately equal. The slide is then subjected to chemical and heat
treatment which causes the DNA sequences to become attached to the glass slide
and denatured. Typically, each location on the slide corresponds to a single gene,
but in some experimental designs, each gene may be represented by several spots
on the same slide.

Once the slide has been prepared, mRNA is extracted from a biological sam-
ple of interest. Reverse transcription is used to produce cDNA representations of the extracted mRNA. These cDNA targets are labelled with a fluorescent dye and incubated with the microarray under conditions suitable for hybridisation. On account of the specificity of hybridisation reactions, complementary probes and targets gradually find each other and hybridise. However, targets may also bind to non-complementary probes, reducing the specificity of the assay. A number of experimental factors, including the time allowed for hybridisation, nucleic acid concentration, and complexity of the sample affect the specificity achieved. The time allowed for hybridisation is quite important, since any given target sequence will pair and dissociate many times during the hybridisation reaction, stabilising as the system reaches equilibrium (Stoughton, 2005). As a consequence, hybridisation times of several hours or more are usually needed to achieve good specificity.

Following hybridisation, the slide is washed to remove unhybridised cDNA, and then scanned using a confocal laser, to produce images of the luminous emission of the tagged target molecules.

Bias is a serious concern at every step described so far, such that the brightness observed at a given location on the slide may relate only approximately to the abundance of mRNA in the original sample. A particular problem associated with the use of long DNA fragments as probes is variation in hybridisation properties between spots. Indeed, left unchecked, such variation would render invalid comparisons between genes, as it would become impossible to determine whether the relative brightness of two spots was truly due to different levels of expression, or simply a consequence of inter-spot variability. To mitigate this problem, cDNA
obtained from the biological sample of interest is usually mixed with a reference cDNA, which is labelled with a different dye, and the mixture allowed to competitively hybridise to the probes. The slide is then scanned and photographed at two different wavelengths, one corresponding to each of the dyes used. This gives rise to two images, and following image analysis, two sets of real-valued spot intensities. The ratio of spot intensities for each probe is then used to quantify mRNA abundance in the biological sample of interest, relative to the corresponding abundance in the reference sample. This procedure is in effect a matched control method, which allows variations in the properties of individual spots to cancel each other out. The two images obtained from an experiment of this type are usually presented as a composite pseudo-image. Red and green are used to colour spot intensities derived from the two scanned images, such that each spot in the combined image has a colour somewhere between red and green, depending on the relative expression level of the gene to which it corresponds. Figure 2.3 shows a pseudo-image of this kind, obtained from a cDNA array.
Since cDNA arrays are prepared in the laboratory, the investigator has considerable flexibility in the choice of probes, and the precise manner in which they are spotted onto the array. This can be a major advantage, particularly when smaller, customised arrays are needed. Low cost is also a major advantage of cDNA arrays, and a factor which has contributed to their popularity in academic laboratories. However, cross-hybridisation is a serious problem with these arrays. This occurs when target mRNA binds to probes which are similar but not exactly identical to the truly complementary probe, thus producing false detections.

### 2.2.3 Oligonucleotide arrays

Oligonucleotide arrays have as their probes short segments of DNA called oligonucleotides, and were first introduced in Lockhart et al. (1996). Oligonucleotide arrays have been developed and commercialised by Affymetrix Inc. (Santa Clara, CA, USA). The description presented here focuses on Affymetrix arrays.

In an oligonucleotide array, each gene is represented by a probe consisting of several oligonucleotides, rather than a single longer DNA sequence, as is the case in cDNA arrays. The basic idea is that a few short subsequences of the coding region of a gene are sufficient to specify it, rendering the use of a longer sequence unnecessary. Each short subsequence is accompanied by a control probe which differs by only one base from the main probe. The entire set of subsequences and controls for a particular gene is called a probe set. Decisions regarding the length of oligonucleotide to be used, and the precise set of oligonucleotides with which to characterise a given gene, are themselves important technical questions.
The oligonucleotides used by Affymetrix each comprise 25 nucleotides, and the company uses sophisticated in-house procedures to design probe sets for their arrays.

Once the probe sets have been designed, they are synthesised in situ, by photolithography onto silicon wafers. This manufacturing process was inspired by methods used for producing microchips, and for this reason, oligonucleotide arrays produced in this manner are also known as DNA chips. The arrays are then shipped to laboratories, where they are used to monitor mRNA levels in a process broadly similar to that of cDNA arrays, but using specialised equipment supplied by Affymetrix. An important difference between the two approaches is that on account of the much lower inter-spot variability of oligonucleotide arrays, a reference sample is not needed. Thus, experiments using oligonucleotide arrays involve only one channel, and produce absolute measures of mRNA abundance. Figure 2.4 shows the output from an oligonucleotide array.

The main advantages of oligonucleotide arrays are that they are less likely to
be corrupted by cross-hybridisation, require a smaller amount of mRNA from the biological sample of interest and have a higher dynamic range. However, oligonucleotide arrays are available for only a limited number of popular genomes, and are both more expensive and in many ways less flexible than cDNA arrays.

2.3 Limitations of microarray studies

Transcriptional profiling using microarrays has proved to be a powerful and practical approach to the investigation of system-wide behaviour in biology. However, microarray technology is not a panacea. This Section looks at some of the limitations of microarray-based studies.

Transcription itself is only one of several levels of genomic organisation. As a consequence, studies of transcription, while often highly informative, do not by themselves provide a full description of cellular state. For example, studying mRNA levels alone cannot provide information regarding post-translational modifications of proteins or protein-protein interactions (de Hoog and Mann, 2004).

A more subtle limitation of microarray studies arises on account of the fact that observed transcript abundances are measured at the level of a population average across a large number of cells, rather than at the level of a single cell. This means that variations in gene expression between individual cells within a tissue cannot be observed by conventional microarray profiling. Advances in single-cell expression profiling have revealed that such variation may be of considerable biological importance (Levsky et al., 2002).
These fundamental biological limitations mean that microarray data should be regarded as one of many possible views of a biological system. Microarray experiments should therefore be treated as one stage in an iterative experimental process, guiding further investigation, but not necessarily providing definitive answers in themselves. Wherever possible, these data should also be integrated with other sources of information, including genomic, proteomic and single-cell data. Indeed, the integrative analysis of genomic data from multiple sources has emerged as a key trend in recent years (Rhodes and Chinnaiyan, 2005; Schadt et al., 2005).

Microarray data are also highly variable, on account of both experimental factors and the complexity of underlying biological processes. The microarray experimental process, as detailed above, introduces variability at every stage, from tissue collection to image acquisition. Good experimental protocol can minimise such variation, but cannot entirely remove it. The effect of underlying biological variability is equally important. Tissue samples which are treated as homogenous in the context of a particular experiment are themselves subject to variation (Baldi and Hatfield, 2002), such that even if data could be obtained via a hypothetical “perfect” measurement process, some variability would certainly remain.

Large sample sizes would be expected to mitigate the effects of variability, but unfortunately, the cost associated with performing microarray experiments means that the number of arrays used is generally very small, usually less than ten, and almost never more than a few dozen. In contrast, the number of genes under study may run into the tens of thousands, resulting in a remarkable - and in many ways
characteristic - mismatch between dimensionality and sample size.

Thus, biological complexity, coupled with high noise levels and small sample sizes, mean that it can be very difficult to draw sound conclusions from microarray experiments. Procedures for the analysis of microarray data should therefore be principled, in order to inspire confidence in results obtained, and sufficiently robust to deal effectively with uncertainty under statistically difficult conditions. Indeed, it is these requirements which motivate the original contributions of this thesis. In the following Chapter, we look in more detail at microarray data analysis, before going on to present original work in Chapters 4 and 5.
Chapter 3

Microarray data analysis

Microarrays are, as we have seen, an important technology in quantitative molecular biology, capable of providing information regarding molecular mechanisms underlying key events in health and disease. However, the biological processes underlying transcription are complex, and indeed transcription itself represents only one layer of biological complexity. Furthermore, microarray data are highly variable, on account of both experimental process and underlying biology, and are characterised by a remarkable mismatch between dimensionality and sample size. As a consequence of these factors, the task of obtaining meaningful biological insights from microarray data has proved to be a major challenge, and has in turn motivated a great deal of recent research in computer science and statistics.

Statistical and machine learning methods in particular have come to play a central role in dealing with the uncertainty inherent in microarray studies. Indeed, in recent years, the statistical analysis of microarray data has emerged as a major
topic of study in its own right (Quackenbush, 2001; Baldi and Hatfield, 2002; Sebastiani et al., 2003; Speed, 2003; Husmeier et al., 2005), and is an increasingly vital aspect of microarray-based research. The original work presented in this thesis forms a part of this growing body of research; the current Chapter provides a review of existing work in the field.

The rapid pace of progress in this area has given rise to a large and growing literature on the development and application of statistical and machine learning methods to microarray data. We organise this body of work into two very broad themes: predictive analysis and discovery-driven analysis. By “predictive analysis” we mean analyses whose explicit goal is to predict some target value or label from microarray data. Supervised methods, including classification and regression, play a key role in analyses of this kind. By “discovery-driven analysis” we mean analyses whose goal is to better understand a system in a more general sense, and not necessarily to predict a specific quantity. Relevant problems include tissue and gene clustering, dimensionality reduction, the inference of biological networks and the analysis of differential expression, the latter problem being the main focus of the original work presented in later chapters. Clearly, the two modes of enquiry we have called “predictive” and “discovery-driven” overlap to a certain extent. For example, biological categories discovered via cluster analysis may subsequently be used in a predictive system. Equally, a model devised for the purpose of a very specific predictive task may prove useful in understanding a biological system.

The remainder of this Chapter is organised as follows. We begin by briefly
discussing image analysis and normalisation for microarray data, and go on to look at each of predictive analysis and discovery-driven analysis in turn. The main focus of this thesis is the analysis of differential expression; we close the Chapter with a review and discussion of this particular topic.

3.1 Image processing and normalisation

The scanning of a hybridised microarray produces an image (or, in the case of a cDNA array, a pair of images) containing thousands of spots, with the intensity of each spot providing a measure of the expression level of the specific gene to which the spot corresponds. Such images are not usually studied directly, but first transformed into numerical values, which are subsequently used to perform computational and statistical analyses of various kinds.

Image analysis for microarrays involves first segmenting the array image into individual spots, then calculating the intensity of each spot, and finally associating each intensity value with the identity of the gene to which it corresponds. Experimental variability in the positions, intensities and morphologies of spots make image analysis for microarrays a challenging problem in its own right, discussed in the context of cDNA arrays in Yang et al. (2002a), Smyth et al. (2003) and Lawrence et al. (2004), and in the context of Affymetrix arrays in Affymetrix Inc. (2002).

Intensity values obtained from microarray images are sometimes subjected to log-transformation (e.g. Yang et al., 2001) with the aim of rendering them ap-
proximately Normally distributed, and consequently better suited to analysis by conventional statistical methods. However, raw microarray data have been shown to follow a distribution which only approaches log-normality at high expression levels (Durbin et al., 2002; Huber et al., 2002). As a consequence, low-abundance transcripts can pose problems for log-transformation, motivating a need for transformations, such as those proposed by Durbin et al. (2002) and Huber et al. (2002), based on more realistic noise models.

Microarray data also exhibit considerable systematic variation, on account of a number of factors including dye-biases and experimental procedure. Normalisation procedures are widely used to remove systematic variation, and allow valid comparisons of mRNA abundances to be made within an array, and between different arrays. Standard methods for the normalisation for cDNA microarray data are described in Yang et al. (2001) and Yang et al. (2002b), while methods for Affymetrix data are discussed in Schadt et al. (2001), Affymetrix Inc. (2002), Bolstad et al. (2003) and Liu et al. (2005).

Image analysis and normalisation are clearly interesting topics of research in their own right, and there remain many open questions regarding optimal approaches for these tasks. One interesting possibility would be to unify image analysis and normalisation with downstream data analysis, such that substantive biological questions could be addressed by the direct analysis of image data. However, in this thesis, we follow current convention, and take real-valued data obtained from standard image analysis and normalisation procedures as the starting point for data analysis.
Thus, we assume that following image analysis and normalisation, a microarray experiment in which $g$ genes are investigated on $n$ arrays gives rise to $n$ $g$-dimensional vectors $X_i$, collectively comprising a matrix $D$ of size $g \times n$:

$$D = [X_1 X_2 \ldots X_n]$$

(3.1)

### 3.2 Predictive analysis

In predictive analyses, each microarray data vector $X_i$ is associated with an output or target value $L_i$, and the goal of the analysis is to predict output or target values for new data. This is illustrated schematically in Figure 3.1. Examples of problems of this kind include tissue classification, discussed in more detail below, in which the output values are phenotypic class labels, and microarray-based regression (Segal et al., 2003), in which the output is a real-valued response variable.

![Figure 3.1](image)

Figure 3.1: A schematic depiction of the predictive analysis of microarray data.

The problem of tissue classification has attracted a great deal of interest in the literature. In problems of this kind, each microarray data vector $X_i$ has a “label” $L_i$, which captures some property of the tissue sample from which the data vector
was obtained, and the aim of the analysis is to automatically categorise previously unseen microarray data vectors. Provided labelled training data is available, this task falls very naturally into the framework of supervised learning.

An early example of microarray-based tissue classification is the work of Golub et al. (1999), in which the authors developed a system to distinguish subtypes of leukaemia on the basis of expression data alone. This work was influential, but the classification scheme used was rather ad hoc. More recently, a number of powerful methods from the machine learning literature have been brought to bear on microarray classification, including support vector machines (Furey et al., 2000), neural networks (Khan et al., 2001) and Bayesian methods (Mallick et al., 2005).

In many cases, it can be useful to construct a classifier based on a subset of the total number of features or dimensions in the data. The process of selecting a subset of dimensions for classification is called feature selection. On account of the high-dimensionality of microarray data, feature selection methods can sometimes be beneficial in tissue classification. Examples of feature selection applied to microarray-based classification include Golub et al. (1999), Mukherjee (2001), Weston et al. (2001) and Xing et al. (2001).

### 3.3 Discovery-driven analysis

In most cases, microarray experiments are performed with the aim of obtaining a better understanding of a particular biological system or phenomenon. As we have
seen, microarrays themselves represent only one of several available perspectives on a biological system of interest, and furthermore are themselves highly variable in nature. As a consequence, individual microarray-based studies must usually be followed up by more accurate analyses of transcription, as well as other, complementary genomic assays. In this sense, the majority of microarray studies can be regarded as exploratory in nature, and may be thought of as representing one stage in an ongoing, iterative experimental process, as illustrated in Figure 3.2.

![Figure 3.2: A schematic depiction of the role played by data analysis in discovery-driven microarray studies: data analysis is one stage in an ongoing, iterative experimental process.](image)

Discovery-driven microarray studies vary widely with respect to a number of factors, including the availability of prior knowledge regarding the system of interest, the number of arrays under study and the nature and size of follow-up plans. The diversity of discovery-driven studies is reflected in the diversity of data-analytic problems in the area. Cluster analysis, dimensionality reduction, the inference of networks and analysis of differential expression each have a role to play, as illustrated schematically in Figure 3.3. “Coarse-grained” analyses such as cluster analysis, dimensionality reduction and statistical tests for differential expression tend to come into their own when data is scarce, and/or prior knowl-
edge relatively poor. Such analyses lay the ground for further work, by focusing
attention on particular genes, or groups of genes. In contrast, when data is more
plentiful, or the system of interest better characterised, it becomes possible to tease
out finer-grained structure, by making use of methods, such as Bayesian networks,
capable of inferring the structure and connectivity of biological networks.

Figure 3.3: A schematic depiction of the high-level relationship between various
types of discovery-driven analysis for microarray data.

The choice of a specific method for analysis depends also upon the goals and
design of the experiment. For example, in some cases, transcript levels are mea-
sured under contrasting conditions, such as healthy and diseased, or wild-type and
mutant. In such experiments, the tissues to be analysed are usually carefully cho-
sen so as to minimise confounding variation. Under such conditions, analyses of
differential expression can go a long way towards finding sets of genes likely to
be relevant to the underlying biological process, and therefore worthy of further
investigation. In other cases, the experimental design may be more exploratory,
and the set of conditions under which transcript levels are measured more diverse. Then, unsupervised learning procedures, like clustering and dimensionality reduction, can be useful in uncovering hidden structure in the data, and suggesting directions for further research.

We look at each of clustering, dimensionality reduction and network inference below, and then discuss the analysis of differential expression, which is the main topic of this thesis, in the following Section.

3.3.1 Clustering

Clustering refers to the task of grouping observations into subsets or “clusters”, such that observations within clusters are in some sense more closely related to one another than to observations assigned to different clusters (Hastie et al., 2001). Clustering has emerged as an extremely popular approach in the analysis of microarray data, with a wide variety of methods in common use (reviewed in Tibshirani et al., 1999). Clustering procedures have been used to group together genes (Eisen et al., 1998), arrays (Golub et al., 1999), or both genes and arrays at once (Alon et al., 1999).

Early applications of clustering in microarray data analysis (Eisen et al., 1998; Alon et al., 1999; Golub et al., 1999) used algorithms like k-means and agglomerative hierarchical clustering. Such methods are intuitive and simple to implement, but suffer from a number of limitations, including the hard assignment of cluster memberships and the lack of a principled approach to model order selection. In contrast, modern probabilistic formulations of clustering (e.g. Roberts et al.,
1998) are principled and flexible, allowing, for example, the number of clusters to be inferred directly from data, and objects to belong to multiple clusters at once. In light of these strengths, a number of authors, including Barash and Friedman (2002) and Heard et al. (2005), have advocated the use of probabilistic methods for the clustering of gene expression data.

### 3.3.2 Dimensionality reduction

*Dimensionality reduction* refers to the task of mapping data vectors from a high-dimensional space into a lower-dimensional space while preserving as much information as possible. Widely-used methods include principal components analysis (Jolliffe, 1986) and independent components analysis (Roberts and Everson, 2001). Such methods can play an important role in the analysis of high-dimensional microarray data, by finding compact yet informative data representations. This can ameliorate some of the deleterious effects associated with high dimensionality, enabling, for example, more fine-grained downstream modelling, using computationally expensive or statistically complex methods which might have been inappropriate for the original data. Dimensionality reduction methods can also be used to map gene expression data down to two or three dimensions, thereby allowing the data to be studied visually.

Principal components analysis or PCA seeks to find a linear projection of the data which preserves as much variance as possible; PCA has been applied to microarray data by a number of authors, includingAlter et al. (2000) and Yeung and Ruzzo (2001). Independent components analysis or ICA is a generalisation
of PCA in which a projection is learned whose basis vectors are as statistically independent as possible. Analysis of microarray data using ICA can therefore help to uncover a relatively small number of independent biological processes which underlie observations across a large number of genes (Lee and Batzoglou, 2003).

### 3.3.3 Network inference

Biological systems are equipped with sophisticated - and often extremely complex - molecular control and signalling mechanisms. Genes and their products form regulatory and signalling networks in which molecular components influence one another to implement high-level functions such as feedback control and signal transduction. The task of characterising biological networks of this kind and understanding their properties is therefore of central importance in quantitative systems biology.

Detailed quantitative models have been built and analysed for a few relatively well-understood networks (e.g. Schoeberl et al., 2002). However, most networks of interest are too large, or too vaguely characterised to permit detailed mathematical modelling (Husmeier, 2003). The statistical inference of networks from microarray data represents a slightly less detailed, but nonetheless highly informative, approach to the study of biological networks, and has, in recent years, attracted a great deal of attention in the machine learning community, with Bayesian methods emerging as the dominant methodological approach (Segal et al., 2003; Friedman, 2004; Husmeier et al., 2005).
3.4 Analysis of differential expression

Quantitative biology is in its infancy, and most biological processes remain rather poorly understood. An important first step in understanding the molecular basis of a phenomenon of interest is to investigate changes in molecular behaviour across contrasting conditions, such as wild-type and mutant, or healthy and diseased. In some cases, there may be sufficient prior knowledge regarding the phenomenon to suggest a small set of genes and gene products likely to be of interest, but in most cases, there tends to be little detailed knowledge of this kind, such that system-wide microarray studies are needed to ensure that nothing of importance is missed.

In a typical experiment, tissues are collected under two or more contrasting conditions, and microarrays used to measure transcript levels in each case. Comparative experiments of this kind are now used in virtually every area of biological enquiry, and are arguably a dominant mode of use for microarrays (Stoughton, 2005).

In comparative microarray experiments, genes which are truly up- or down-regulated between conditions are said to be “differentially expressed”. The potential biological relevance of differentially expressed genes has meant that the statistical task of discovering which of the genes under study is truly differentially expressed has emerged as a key problem in microarray data analysis. This problem, which we refer to as gene selection, is the main topic of this thesis.

Accurate gene selection is important for two main reasons. Firstly, in some cases, analyses of differential expression may provide direct biological insights. For example, knowing that a set of genes tends to be consistently up-regulated in
tissues affected by a certain form of cancer may provide information regarding molecular mechanisms underlying the cancer. Secondly, gene selection plays a key role in providing researchers with short-lists of genes worthy of further investigation. Microarray experiments are only one of a wide range of assays available. As noted above, they are large-scale, relatively cheap, but highly variable and, by their very nature, biologically incomplete. Other assays such as RT-PCR\(^1\) and various proteomic methods may be highly informative, but are too expensive and difficult to apply to every single gene or gene product, leading to the need to filter out sets of genes likely to be of biological relevance.

Thus, a failure to identify truly differentially expressed genes can lead to missed opportunities, and equally, false identifications can lead to wasted resources. Yet the uncertainty inherent in microarray data and the small sample sizes typical of comparative studies have meant that consistently accurate gene selection has proved difficult to achieve. In this Section, we discuss existing work in gene selection. We first place the gene selection task within a statistical framework, showing that gene selection can be described very naturally as a hypothesis testing problem. We then look at a number of test statistics for gene selection, and finally discuss some open questions.

\(^1\)Real time reverse transcription polymerase chain reaction (Bustin, 2002). This is essentially a highly accurate method for quantifying mRNA abundance.
3.4.1 A statistical view of differential expression

Consider an experiment in which microarrays are used to measure gene expression levels under two conditions of biological interest. The total number of genes under study is \( g \), and there are \( m \) arrays in one condition, and \( n \) in the other. As before, if we assume that each gene is represented only once on an array, each array can be thought of as giving rise to a vector of \( g \) real-valued expression levels. We treat each such vector as a random vector, and denote the data obtained under the first condition as \( \{X_i\}_{i=1}^m \), and the data obtained under the second condition as \( \{Y_j\}_{j=1}^n \). Then, the complete set of data, obtained from all \( (m + n) \) arrays, is a random matrix \( D \) of size \( g \times (m + n) \):

\[
D = [X_1X_2 \ldots X_m \ Y_1Y_2 \ldots Y_n]
\]  

We assume that the random vectors \( X_i \) are independent and identically distributed (i.i.d.) according to some multivariate density \( p_X \) with mean vector \( \mu_X \), and that \( Y_j \) are i.i.d. under a density \( p_Y \) with mean vector \( \mu_Y \):

\[
X_i \overset{iid}{\sim} p_X, \quad E[X_i] = \mu_X \tag{3.3}
\]

\[
Y_j \overset{iid}{\sim} p_Y, \quad E[Y_j] = \mu_Y \tag{3.4}
\]

The data for a single gene may be regarded as a univariate sample of size \((m + n)\). The data for the \( k^{th} \) gene is denoted \( D_k \), and its constituent random
variables \( X_{ik} \) and \( Y_{jk} \):

\[
D_k = X_{1k}, X_{2k}, \ldots, X_{mk}, Y_{1k}, Y_{2k}, \ldots, Y_{nk}
\]  

(3.5)

Let the \( k^{th} \) components of mean vectors \( \mu_X \) and \( \mu_Y \) be denoted by \( \mu_{Xk} \) and \( \mu_{Yk} \) respectively. Then, a gene \( k \) is said to be a \textit{differentially expressed gene}, if \( \mu_{Xk} \neq \mu_{Yk} \) and a \textit{non-differentially expressed gene} otherwise (Pan, 2002; Speed, 2003). In the interests of brevity, in this thesis we use the abbreviations ‘DEG’ for differentially expressed gene and ‘NDEG’ for non-differentially expressed gene. If \( \mu_{Xk} < \mu_{Yk} \), gene \( k \) is said to be \textit{up-regulated} in condition \( Y \), relative to condition \( X \), while if \( \mu_{Xk} > \mu_{Yk} \), it is said to be \textit{down-regulated} in condition \( Y \).

### 3.4.2 Hypothesis testing as a paradigm for differential expression

The gene selection task falls very naturally into a hypothesis testing framework\(^2\). Indeed, most of the existing work in gene selection poses the problem within a hypothesis testing paradigm. Following existing work (e.g. Pan, 2002; Speed, 2003), we define null and alternative hypotheses for the \( k^{th} \) gene in the following way:

\[
H_{0k} : \mu_{Xk} = \mu_{Yk}
\]  

(3.6)

\[
H_{1k} : \mu_{Xk} \neq \mu_{Yk}
\]  

(3.7)

---

\(^2\)Hypothesis testing is reviewed, at an introductory level, in Appendix A.
where, $H_{0k}$ and $H_{1k}$ denote respectively the null and alternative hypotheses for gene $k$. A gene $k$ is then said to be a DEG if it satisfies the alternative hypothesis $H_{1k}$, and a NDEG otherwise.

A hypothesis testing procedure for differential expression must decide whether each of the genes under study satisfies the null or alternative hypothesis. This is done by first defining a suitable test statistic $T$, and then using the value of the test statistic for each gene, in combination with a suitably defined threshold, to decide between the competing hypotheses. We discuss possible choices for $T$ below, and assume for the moment that $T$ is some real-valued function of the data, having the property that larger values indicate greater evidence against the null hypothesis. Let $T_k$ be a random variable representing the value of the test statistic $T$ for gene $k$:

$$T_k = T(D_k)$$

If $T$ is used to compute $g$ test statistics $T_1, T_2 \ldots T_g$, one for each gene, a set of genes to be selected as ‘differentially expressed’ may be obtained by either (i) ranking the genes under the test statistic, and selecting a pre-determined number of top-ranked genes, or (ii) by imposing a threshold $\tau$ on the values $T_k$, such that only genes satisfying $T_k > \tau$ are selected. The test statistic $T$ plays the role of summarising the information contained in the data, with the aim of discriminating between data drawn under the null and alternative hypotheses.

It is interesting to compare hypothesis testing with supervised and unsuper-
vised learning as a framework for gene selection. The substantive goal of gene selection is to decide whether or not each gene is truly differentially expressed. This amounts to assigning a ‘label’ (in the language of hypothesis testing one of either \(H_{0k}\) or \(H_{1k}\)) to each gene \(k\), on the basis of random data \(D_k\). This is roughly analogous to the assignment of labels to data in classification problems. However, in contrast with classification, gene selection is an unsupervised problem, insofar as we do not usually know which of the genes under study are truly differentially expressed. This means that we cannot treat gene selection as a standard problem in supervised learning, and are unable to use empirical estimates of accuracy to learn a function which discriminates between DEGs and NDEGs.

Feature selection bears some similarity with gene selection, insofar as both tasks involve selecting subsets of genes. However, the two tasks have rather different goals. The aim of gene selection is simply to find all genes whose population means are distinct between tissue classes; in contrast, feature selection algorithms aim to find genes which best explain class labels. To illustrate the distinction between the two tasks, consider a hypothetical scenario in which a single gene fully explains class labels, but a hundred genes are nonetheless consistently up-regulated in one set of tissues. A gene selection procedure will aim to identify all the up-regulated genes, while a feature selection algorithm should return the single explanatory gene. The distinction is biologically important: all hundred genes may have effects of interest to the investigator, despite the fact that a single gene captures the class information.

Equally, gene selection, although unsupervised in nature, differs from canonici-
cal problems in unsupervised learning. Unsupervised learning aims to find structure - essentially any salient structure - in data. In contrast, gene selection is focused specifically on finding genes having a difference in underlying means. Gene selection may therefore be likened to an unsupervised classification problem in which the “classes” to be distinguished between are characterised by properties of the underlying model, namely the fact that in one “class” population means under the conditions are identical and in the other these means are distinct.

Hypothesis testing is concerned precisely with problems of this kind, in which data is “unlabelled”, but the property which objects of interest should satisfy is characterised in terms of population parameters. Thus, the hypothesis testing paradigm is a very appropriate one for gene selection, and indeed, in the existing literature, gene selection is usually cast as a problem in statistical hypothesis testing (Pan, 2002; Speed, 2003).

### 3.4.3 Test statistics for differential expression

A number of test statistics for gene selection have been put forward in the literature (reviews include Pan, 2002; Troyanskaya et al., 2002; Cui and Churchill, 2003; Speed, 2003). We divide test statistics for gene selection into three categories: (i) Intuitive, *ad hoc* statistics, (ii) Classical test statistics, including the t-statistic and rank-sum statistic, (iii) Microarray-specific test statistics, including those based on Bayesian models.
(i) Intuitive, ad hoc statistics.

The most obvious approach to selecting differentially expressed genes is via a direct comparison of sample means in the two conditions. Let the sample means for gene \( k \) in each of the two conditions be \( \bar{X}_k \) and \( \bar{Y}_k \) respectively:

\[
\bar{X}_k = \frac{1}{m} \sum_{i=1}^{m} X_{ik} \quad (3.9)
\]

\[
\bar{Y}_k = \frac{1}{n} \sum_{j=1}^{n} Y_{jk} \quad (3.10)
\]

A naive test statistic for gene selection may then be constructed simply by taking the difference of the two means. Let \( T_k \) represent the value of the test statistic for the \( k^{th} \) gene:

\[
T_k = \bar{X}_k - \bar{Y}_k \quad (3.11)
\]

This test statistic is called difference of means.

A method similar to difference of means, which is widely used as a first-pass analysis, is to consider the extent of the change in expression levels between conditions in the form of a ratio. This approach is usually called fold analysis and the extent of the change the fold change. As noted by Baldi and Long (2001), the difference of means function, when applied to log-transformed microarray data, can be regarded as roughly equivalent to a log-space fold analysis.

These approaches are simple and intuitive, but suffer from a number of drawbacks. Firstly, they take no account of variance, such that highly variable genes
which have large gene-level differences $\bar{X}_k - \bar{Y}_k$ in means are treated the same as those genes which have equally large differences in means but lower variability. Secondly, in contrast with classical test statistics like the t-statistic, these functions do not have null sampling distributions\(^3\) which are independent of unknown parameters, which means that P-values and other measures of significance cannot be easily computed.

(ii) Classical test statistics.

The t-test is a canonical “two-sample” hypothesis test (DeGroot and Schervish, 2002) and has been widely used in the differential analysis of microarray data (e.g. Arfin et al., 2000; Tanaka et al., 2000). The gene-level statistic is a t-statistic:

$$T_k = \frac{\bar{X}_k - \bar{Y}_k}{\left(\frac{1}{m} + \frac{1}{n}\right)^{\frac{1}{2}} \hat{SD}_k}$$

where, $\hat{SD}_k$ represents the pooled sample standard deviation for the $k^{th}$ gene:

$$\hat{SD}_k = \sqrt{\frac{(m-1)\hat{\sigma}^2_{X_k} + (n-1)\hat{\sigma}^2_{Y_k}}{m + n - 2}}$$

The variables $\hat{\sigma}^2_{X_k}$ and $\hat{\sigma}^2_{Y_k}$ represent the unbiased sample variances for the $X$
and $Y$ data respectively:

$$
\hat{\sigma}^2_{Xk} = \frac{1}{m-1} \sum_{i=1}^{m} (X_{ik} - \bar{X}_k)^2 \quad (3.14)
$$

$$
\hat{\sigma}^2_{Yk} = \frac{1}{n-1} \sum_{j=1}^{n} (Y_{jk} - \bar{Y}_k)^2 \quad (3.15)
$$

If the data are Normally distributed and population variances equal in both classes, the sampling distribution of the t-statistic under the null hypothesis is known to be a t-distribution with $df = (m + n - 2)$ degrees of freedom. This distribution can therefore be used to very easily compute P-values for the t-statistic. If variances are expected to differ between conditions, a situation known in the statistical literature as the Behrens-Fisher problem (DeGroot and Schervish, 2002), the t-statistic may still be used, but the degrees of freedom $df$ should then be calculated using the following expression, first proposed by Welch (1938):

$$
df = \frac{(\hat{\sigma}^2_{Xk}/m + \hat{\sigma}^2_{Yk}/n)^2}{(\hat{\sigma}^2_{Xk}/m)^2/(m-1) + (\hat{\sigma}^2_{Yk}/n)^2/(n-1)} \quad (3.16)
$$

The main advantages of the t-statistic are its simplicity and the ease with which P-values can be computed. However, if the underlying data-generating distribution is significantly non-normal, the t-statistic may not give good results. Furthermore, doubts have been expressed regarding the ability of the t-statistic to accurately select differentially genes when sample sizes are small (Cui and Churchill, 2003).
The Wilcoxon-Mann-Whitney rank-sum test or simply rank-sum test (DeGroot and Schervish, 2002) is a classical non-parametric statistic which has been applied to the problem of gene selection (Troyanskaya et al., 2002). The rank-sum approach makes use of order statistics to construct a measure of how likely it is that the distributions underlying data in two classes are the same. Consider the \((m+n)\) observations \(X_{1k} \ldots X_{mk}, Y_{1k} \ldots Y_{nk}\) for gene \(k\). Let these observations be placed in ascending order, and each observation assigned a rank from 1 to \((m+n)\), corresponding to its position in the ordering. Also, let the random variable \(W_k\) denote the sum of the rank positions of the \(m\) observations \(X_{1k}, X_{2k} \ldots X_{mk}\). If both \(X\) and \(Y\) data are identically distributed, the observations \(X_{1k}, X_{2k} \ldots X_{mk}\) will tend to be spread quite evenly among the ordering of all \((m+n)\) observations. In particular, it can be shown that if the data in the two classes are identically distributed, the expected value of \(W_k\) is \(m(m+n+1)/2\). Then, the deviation of the observed value of \(W_k\) from this value provides a measure of evidence against the hypothesis of identical class distributions, leading to the test statistic given below:

\[
T_k = W_k - m(m + n + 1)/2 \tag{3.17}
\]

This statistic is called the Wilcoxon-Mann-Whitney or rank-sum statistic.

The main advantage of the rank-sum approach is that it makes no assumption concerning the form of the underlying distribution. However, if the true data-generating distribution does not deviate significantly from distributional assumptions, the rank-sum test may be less effective than a suitable parametric method.
Indeed, Thomas et al. (2001) applied the rank-sum test to a widely-analysed leukaemia dataset and found that it did not uncover any differentially expressed genes at all. Furthermore, when the rank-sum statistic is used to test means, an implicit assumption is that under the null hypothesis the distributions of the $X$ and $Y$ data are identical. Thus, if class variances are unequal, or if the form of the class distributions differ, the rank-sum test may not be an appropriate choice.

(iii) Microarray-specific test statistics.

Significance Analysis of Microarrays or SAM (Tusher et al., 2001) is a recently proposed statistic designed for the specific purpose of analysing differential expression in microarray data. The SAM statistic resembles the t-statistic, but has a constant term $C$ added to the denominator:

$$T_k = \frac{\bar{X}_k - \bar{Y}_k}{(\frac{1}{m} + \frac{1}{n})^{\frac{1}{2}} \hat{SD}_k + C}$$

(3.18)

The use of the term $C$ is intended to reduce the false positive rate by moderating the effect of small values of the sample standard deviation term $\hat{SD}_k$. When selecting differentially expressed genes using a conventional t-statistic, small values of $\hat{SD}_k$ are typically a major contributor to the false positive rate. This problem arises when non-differentially expressed genes happen to have small values of $\hat{SD}_k$, leading to large absolute values of the t-statistic for these genes, which in turn causes them to be mistakenly selected. Adding a small constant to the denominator of the t-statistic has the effect of downplaying the importance of small values of $\hat{SD}_k$, and can thereby reduce the number of false positives.
The term $C$ may be set in a number of ways. In the original SAM paper, (Tusher et al., 2001), it is determined from the entire dataset by minimising the coefficient of variation of the overall statistic in moving windows across the data. An alternative procedure which is sometimes used is to set $C$ equal to the 5th or 10th percentile of observed pooled standard deviations.

Baldi and Long (2001) proposed a regularised test statistic for differential expression via a Bayesian argument. Starting from an explicit data model, the authors derived the following gene-level statistic:

$$T_k = \frac{\bar{X}_k - \bar{Y}_k}{\sqrt{\frac{\hat{SD}^2 + (m+n-1)\hat{SD}^2}{v_0 + m + n - 2}}}$$  \hspace{1cm} (3.19)

where, $\hat{SD}$ represents a global estimate of standard deviation derived from all $g$ genes. The parameter $v_0$ is a variance hyperparameter and is intended to be set by the user.

This statistic is similar in spirit to the SAM statistic discussed above. Setting $C_1 = \frac{v_0}{v_0 + m + n - 2}\hat{SD}$ and $C_2 = \frac{(m+n-1)}{v_0 + m + n - 2}$ we may re-write Equation 3.19 in the following way:

$$T_k = \frac{\bar{X}_k - \bar{Y}_k}{\sqrt{C_1 + C_2\hat{SD}^2_k}}$$  \hspace{1cm} (3.20)

Here, the terms $C_1$ and $C_2$ are constant across genes. $C_1$ in particular plays broadly the same role as the constant term in the SAM statistic, smoothing the gene-level variance and thereby reducing the false positive rate.
A drawback of this method is the fact that the parameter \( v_0 \) must be set by the user. This problem has been partially addressed by treating the problem of differential expression within an “empirical Bayes” framework (Lönnstedt and Speed, 2002), in which hyperparameter values are determined from the data.

### 3.4.4 Open questions in differential expression

Thus, several test statistics have been put forward for gene selection, including simple, intuitive statistics, classical test statistics and novel, microarray-specific statistics. Yet there has been relatively little attention paid to important questions regarding the effectiveness of test statistics under various conditions. There is, as yet, no clear notion of what it means for a test statistic to be effective in selecting differentially expressed genes, nor any formal framework within which various methods can be assessed.

As a consequence, several questions of practical importance have remained unresolved, or only partially addressed in the literature. Classical theory (Lehmann, 1986) suggests that the t-statistic ought to be optimal whenever data is Normally distributed and condition-specific variances equal, yet some authors (e.g. Baldi and Long, 2001; Cui and Churchill, 2003) have suggested that the t-statistic may be unsuitable for microarray data at small sample sizes. If data are Normally distributed, is there any truly substantive problem associated with using a t-statistic at small sample sizes? Equally, simple methods like difference of means are regarded as naive due to the fact that they do not take account of standard deviation, yet the question as to whether a standard deviation term is at all beneficial, and if
so, under what conditions, has remained unaddressed. The practical importance of decisions made on the basis of the results of gene selection mean that questions of this kind are of much more than purely theoretical interest.

A second, and more general, set of concerns relates to the suitability, in the microarray context, of a conventional approach to hypothesis testing. Microarray data are usually highly variable and furthermore, the complexity of the data-generating process means that there tends to be considerable uncertainty regarding underlying models. Yet, using a conventional test statistic amounts to using a fixed function to rank and select genes. Flexible, learning-based models have proved to be effective in many areas of practical data analysis, and it may be that such an approach could be useful in gene selection. Thus, while we have argued that hypothesis testing is a reasonable paradigm for the analysis of differential expression, it is not clear whether or not conventional procedures for hypothesis testing are sufficiently flexible for the microarray domain.

We address these two sets of concerns in the remainder of this thesis. First, in the following Chapter, we propose a decision-theoretic framework for gene selection, which enables us to formally understand the effectiveness of test statistics for gene selection in terms of loss and risk. We then take advantage of some of the insights provided by our analysis to put forward, in Chapter 5, a novel, adaptive procedure for gene selection.
Chapter 4

A decision-theoretic analysis of gene selection

In recent years, the statistical task of selecting differentially expressed genes from microarray data has attracted a great deal of attention. Hypothesis testing has emerged as a standard paradigm for gene selection, and a number of test statistics have been proposed in the literature. However, relatively little attention has been paid to understanding the ability of hypothesis testing procedures to accurately select differentially expressed genes under various conditions. To take just one example, there has been some discussion in the bioinformatics community regarding the suitability of the t-statistic when variances differ across genes in a systematic manner. If differentially expressed genes are expected to have generally higher or lower variances than other genes, does the t-statistic remain an appropriate choice? Questions such as this have serious implications for bioinfor-
matics users, but have remained largely unresolved.

The aim of this Chapter is to better understand the ability of statistical procedures to discover truly differentially expressed genes. To this end, we characterise gene selection as a statistical decision problem, and carry out our analysis within a decision-theoretic framework. The original contributions of this Chapter are threefold. Firstly, our work is, to the best of our knowledge, the first explicitly decision theoretic study of the analysis of differential expression. Secondly, we present results which allow risk in gene selection to be computed directly from the sampling distributions of test statistics. This effectively removes the need for costly, and possibly intractable, brute-force simulation. Finally, we take advantage of the theory developed to obtain results comparing three widely-used test statistics under various conditions. However, we emphasise at the outset that the approach taken here is theoretical and cannot be directly applied to real-world microarray datasets, where true sampling distributions and states of nature are unknown.

The remainder of this Chapter is organised as follows. We begin by presenting some background material on statistical decision theory, defining terms and notation which will be used throughout the remainder of the thesis. We then go on to develop a decision-theoretic view of gene selection, showing how risk may be computed directly from relevant sampling distributions, and finally present results.
4.1 Statistical decision theory

Statistical decision theory is a branch of statistics concerned with questions regarding the effectiveness and optimality of statistical procedures. In this Section, we present a short introduction to some key concepts in statistical decision theory. Further details can be found in standard reference works including Berger (1980) and Berger (1985) and, at an introductory level, in Casella and Berger (2002).

4.1.1 Samples, actions and decision procedures

A collection of random variables $D = X_1, X_2 \ldots X_n$ is called a random sample if each of $X_1, X_2 \ldots X_n$ are statistically independent and distributed according to the same probability function $p$. The set of all possible random samples is called the sample space, and is denoted by $S$. The function $p$ is called the true data-generating distribution. We shall assume that the data $X_i$ are real-valued, and that the function $p$ is a probability density function. However, the definitions given below can be easily modified for the discrete case.

Statistical decision theory is concerned with the making of decisions on the basis of random data. The complete set of possible decisions in a given problem is called the action space, and denoted by $\mathcal{A}$. A decision procedure is a function or rule $\delta : S \mapsto \mathcal{A}$ which maps random samples to actions, such that if $a \in \mathcal{A}$ is an action taken in response to data $D$, we may write $a = \delta(D)$.

This notion of decision-making is highly general, and applies to statistical procedures as diverse as classification, regression and hypothesis testing. In the
case of classification, the action space is a finite set of class labels; in regression, the action space is a set containing all possible values of the response variable.

### 4.1.2 States of nature and the loss function

In any practical application, actions taken by a decision procedure may be associated with real-world costs. For example, in a classification problem there is typically some cost associated with mislabelling a sample, while in regression there may be some cost associated with a prediction which deviates significantly from the true value of the response variable. Clearly, to determine the cost incurred by a particular action we require some notion of the underlying truth, in light of which the action can be assessed.

In a decision problem, the information needed to determine costs associated with various possible actions is called the *state of nature*. We denote a particular state of nature by \( \omega \), and the set of all possible states of nature by \( \Omega \). For example, the assigning of a label to a sample by a classifier is an action, while the identity of the class to which the sample truly belongs may be regarded as a state of nature.

The loss incurred by taking action \( a \) when the true state of nature is \( \omega \) is given by a real-valued function \( L(\omega, a) \), called the *loss function*. The loss function can be thought of as quantifying the cost associated with a decision, in light of the underlying truth. Negative loss is referred to as *gain or utility*. 
4.1.3 Risk

Loss depends on the action taken, which in turn depends on random data. Loss is therefore a function of random variables, and is itself a random variable. The random nature of loss becomes clear if the loss function is written out explicitly in terms of the random sample as $L(\omega, a) = L(\omega, \delta(D))$. The expected value of the loss function under the true data-generating distribution is called expected loss or risk:

$$R(\omega, \delta) \overset{def}{=} \int_S L(\omega, \delta(D))p(D \mid \omega) \, dD \quad (4.1)$$

where, as before, $S$ is the sample space.

$R(\omega, \delta)$ is called the risk function. Note that the risk function depends on both decision procedure $\delta$ and state of nature $\omega$, and that the probability density $p$ under which the expectation is taken is conditional upon the state of nature.

The concept of risk plays a key role in the assessment of statistical procedures and in providing a framework within which competing procedures can be compared. Loss captures the cost incurred by a decision made on the basis of a particular sample of data, while risk provides a sense of the cost likely to be incurred when using a particular decision procedure. Thus, the loss function does not in itself address the random nature of the decision making process, but the risk function does.
4.2 Gene selection as a decision procedure

Given a random matrix \( D = [X_1 \ldots X_m \ Y_1 \ldots Y_n] \) of gene expression data derived under two biological conditions, a gene selection procedure must determine whether or not each of the genes under study satisfies the null hypothesis of identical means in the two conditions. We have seen already that this task can be characterised as a hypothesis testing problem. If \( \mu_{Xk} \) and \( \mu_{Yk} \) are the \( k^{th} \) components of condition-specific mean vectors \( \mu_X \) and \( \mu_Y \) respectively, the null and alternative hypotheses for gene \( k \) are \( H_{0k} : \mu_{Xk} = \mu_{Yk} \) and \( H_{1k} : \mu_{Xk} \neq \mu_{Yk} \).

The sample space \( S \) in gene selection is simply the space of all possible gene expression datasets \( D \). Assuming real-valued expression levels, we may write \( S \subseteq \mathbb{R}^{g(m+n)} \), where, as before, \( g \) is the total number of genes under study, and \( m \) and \( n \) are the number of samples in each of the two biological conditions.

A gene selection procedure must decide whether to accept or reject the null hypothesis for \textit{each} of the \( g \) genes under study. At the level of a single gene, the possible actions are to accept or reject the null hypothesis; we denote these actions by \( a_0 \) and \( a_1 \) respectively. Then, a gene selection procedure can be regarded as a decision procedure \( \delta \), which given data \( D \) returns an \textit{action vector} \( a \) of gene-level actions, such that \( a = \delta(D) \). To avoid confusion with the univariate actions \( a_0 \) and \( a_1 \), we depart slightly from convention and use a superscript to denote the individual components of action vector \( a \), such that \( a^k \) represents the \( k^{th} \) component of vector \( a \). Then, \( a^k = a_i \) denotes the acceptance of hypothesis \( H_i \) for gene \( k \). We may therefore think of the action space for gene selection as a \( g \)-dimensional
generalisation of a univariate action space \( \{a_0, a_1\} \):

\[
\mathcal{A} = \{a_0, a_1\}^g
\]  

(4.2)

What are the states of nature in gene selection? If we assume that the loss associated with a particular action is determined entirely by which hypothesis is actually true, we need only consider two possible states of nature for each gene: \( \omega_0 \) corresponding to the case in which the null hypothesis \( H_0 \) is true, and \( \omega_1 \) corresponding to the case in which the alternative hypothesis \( H_1 \) is true.

Accounting for all \( g \) genes together, we may think of a \( g \)-component vector \( \omega \) as representing the overall state of nature. The components \( \omega^k \) of vector \( \omega \) represent the states of nature of individual genes, such that \( \omega^k = \omega_i \) signifies that \( H_i \) is true for gene \( k \). Then, the set \( \Omega \) of all possible states of nature is simply:

\[
\Omega = \{\omega_0, \omega_1\}^g
\]  

(4.3)

How is action vector \( a \) determined using a univariate test statistic? Suppose \( T_k \) is a random variable representing the value of the test statistic for gene \( k \). We define a vector \( T \) whose components are the gene-level values \( T_k \):

\[
T = [T_1 \ T_2 \ldots T_g]^T
\]  

(4.4)

Given the \( g \) test statistics \( T_1, T_2 \ldots T_g \), we may obtain an action vector \( a \) in one of two ways: (i) by ranking the genes under the test statistic, and selecting a
pre-determined number of top-ranked genes, or (ii) by imposing a threshold \( \tau \) on the values \( T_k \), such that only genes satisfying \( T_k > \tau \) are selected. We consider each of these cases below.

(i) Selecting a pre-determined number of genes. In this approach, genes are ranked under the test statistic (or its absolute value), and the genes occupying the top \( s \) positions under the ranking are selected.

Let the gene-level test statistics \( T_1, T_2 \ldots T_g \) be placed in descending order, and let \( \text{rank}(T_k) \) represent the rank position of gene \( k \). Then, the process of selecting the top \( s \) genes under the statistic \( T \) is a decision procedure \( a = \delta(D) \), with the components \( a^k \) of action vector \( a \) specified as follows:

\[
a^k = \begin{cases} 
  a_1 & \text{if } \text{rank}(T_k) \leq s \\
  a_0 & \text{otherwise}
\end{cases}
\quad (4.5)
\]

The test statistic \( T \), and number \( s \) of genes to be selected, are together sufficient to fully define this procedure. We denote a gene selection procedure defined in this manner by \( \delta_{T,s} \).

(ii) Setting an explicit threshold. In this approach, a threshold \( \tau \) is determined, perhaps in correspondence with a suitably defined notion of significance, such that only genes whose test statistic values exceed \( \tau \) are selected. The com-
ponents $a^k$ of action vector $a$ are then specified as follows:

$$a^k = \begin{cases} 
  a_1 & \text{if } T_k > \tau \\
  a_0 & \text{if } T_k \leq \tau 
\end{cases} \quad (4.6)$$

We denote a gene selection procedure specified by test statistic $T$ and threshold $\tau$ by $\delta_{T,\tau}$.

### 4.3 Loss functions for gene selection

A loss function $L : \Omega \times A \mapsto \mathbb{R}$ quantifies the cost associated with taking a particular action under a specific state of nature. A loss function for gene selection is therefore a real-valued function $L(\omega, a)$ of an action vector $a$ and state of nature vector $\omega$. In this Section we look at loss functions for gene selection, as well as a performance measure - the area under the ROC curve - which, while not actually a valid loss or utility function, will nonetheless prove to be a useful construct in the analysis of gene selection.

Table 4.1: Summary table for gene selection; adapted from Benjamini and Hochberg (1995).

<table>
<thead>
<tr>
<th></th>
<th>Selected</th>
<th>Not selected</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEGs</td>
<td>$S_1$</td>
<td>$g_1 - S_1$</td>
<td>$g_1$</td>
</tr>
<tr>
<td>NDEGs</td>
<td>$S_0$</td>
<td>$g_0 - S_0$</td>
<td>$g_0$</td>
</tr>
<tr>
<td>(Total)</td>
<td>$S$</td>
<td>$g - S$</td>
<td>$g$</td>
</tr>
</tbody>
</table>

The total number of genes under study is denoted by $g$; let $g_1$ be the number of genes which are truly differentially expressed, and $g_0$ be the number of genes
which are not, such that \( g = g_0 + g_1 \). The number of genes selected is \( S \); let the number of truly differentially expressed genes among them be \( S_1 \), and the number of non-differentially expressed genes be \( S_0 \), such that \( S = S_0 + S_1 \). In other words, \( S_1 \) is the number of true positives, and \( S_0 \) the number of false positives. Table 4.1 summarises the relationships between these various quantities.

To simplify our presentation, we also define a gene-level function \( l(\omega^k, a^k) \) which takes on the value 0 if the action \( a^k \) is correct and 1 otherwise:

\[
l(\omega^k, a^k) = \begin{cases} 
0 & \text{if } \omega^k = \omega_i \text{ AND } a^k = a_i \\
1 & \text{if } \omega^k = \omega_i \text{ AND } a^k = a_j \text{ AND } i \neq j 
\end{cases} \tag{4.7}
\]

4.3.1 Zero-one loss

Zero-one loss is a simple, intuitive way of quantifying cost in two-action decision problems, and is widely used in the analysis of classifiers as well as hypothesis tests. We define a zero-one loss function \( L_{01} \) for gene selection by averaging the function \( l(\omega^k, a^k) \) across all \( g \) genes:

\[
L_{01}(\omega, a) \overset{\text{def}}{=} \frac{1}{g} \sum_{k=1}^{g} l(\omega^k, a^k) \tag{4.8}
\]

The loss function \( L_{01} \) captures the cost incurred by a gene selection procedure in the simplest possible way, by counting the total number of Type I and Type II errors. In terms of the variables of Table 4.1, the number of false positives is \( S_0 \), and the number of false negatives is \( g_1 - S_1 \). Using these variables, \( L_{01} \) may also
be expressed as follows:

\[ L_{01} = \frac{g_1 - S_1 + S_0}{g} \]  \hspace{1cm} (4.9)

4.3.2 False Discovery Rate

False Discovery Rate or FDR (Benjamini and Hochberg, 1995) is a popular method for controlling error rates in multiple testing, and has been widely adopted as a basis for adjusting P-values in microarray data analysis (Dudoit et al., 2003; Ge et al., 2003; Reiner et al., 2003; Storey and Tibshirani, 2003). Indeed, it is probably fair to say that in the context of P-value adjustments for high-dimensional genomic applications, a consensus has emerged in favour of FDR and its variations.

The False Discovery Rate in gene selection is defined as the proportion of non-differentially expressed genes among those genes selected (Dudoit et al., 2003). It provides a good overall measure of error in gene selection, naturally taking account of both Type I and Type II errors. We formally express the False Discovery Rate as a loss function \( L_{FDR} \), by making use of the gene-level function \( l(\omega^k, a^k) \), and summing over only those genes for which the action \( a_1 \) is taken:

\[ L_{FDR}(\omega, a) \overset{def}{=} \frac{\sum_{k:a^k=a_1} l(\omega^k, a^k)}{\sum_{k:a^k=a_1} 1} \]  \hspace{1cm} (4.10)

The numerator of this expression is equivalent to the total number \( S_0 \) of false positives, while the denominator represents the total number \( S \) of genes selected.
$L_{FDR}$ may therefore also be expressed in terms of these variables:

$$L_{FDR} = \frac{S_0}{S} \quad (4.11)$$

### 4.3.3 ROC curves

*Receiver operating characteristic* or *ROC curves* are plots of true positive against false positive rates, and are widely used in the analysis of classifiers and other two-action decision procedures. ROC curves provide an ‘at a glance’ summary of error rates across a range of thresholds, making explicit the trade-off between Type I and Type II errors that is inherent in any two-action decision problem.

![ROC curve diagram](image)

**Figure 4.1**: ROC curve for gene selection. ‘TPR’ and ‘FPR’ stand for true positive rate and false positive rate respectively.

As we have seen, gene selection procedures compute a numerical score or test statistic $T_k$ for each gene $k$, and proceed to select a set of genes either by
choosing a certain number of top-ranked genes or by imposing a threshold on the score itself. In the context of ROC analysis there is no substantive distinction between gene selection based on a fixed number of genes and the setting of an explicit threshold. For simplicity, in discussing ROC curves we will not make any distinction between the two approaches and use ‘threshold’ as a catch-all term for both.

Figure 4.1 shows a ROC curve for gene selection. The curve is a plot of sample true positive against sample false positive rates, parameterised by a threshold $\tau$. The sample true positive rate is simply the proportion of DEGs selected, i.e. $S_1/g_1$, while the sample false positive rate is the proportion of NDEGs (mistakenly) selected, i.e. $S_0/g_0$. Each $x$-value represents a false positive rate at a specific threshold, and the corresponding $y$-value represents the true positive rate at the same threshold. From left to right the curve goes from the most conservative possible selection regime, in which no genes are selected at all, to the most liberal, in which every gene is selected. The leftmost point of a ROC curve is always located at $(0, 0)$, and the rightmost point at $(1, 1)$. This is due to the fact that if no genes are selected at all, there can be no positives, so both true and false positive rates must be zero, whereas if every gene is selected the true and false positive rates must both be unity.

ROC curves are especially useful in comparing selection procedures. If the ROC curve for one procedure lies entirely above the curve for a competing procedure, then the first procedure has a lower rate of error than the second, no matter what the threshold. This situation is illustrated in Figure 4.2: there is no threshold
Figure 4.2: Comparing gene selection procedures using ROC curves.

The area under the curve captures, as a single number, the ability of a selection procedure to accurately select differentially expressed genes across a range of thresholds. This quantity is called the area under the ROC curve or $AUC$. A standard result in ROC analysis is the equivalence between the area under the ROC curve and the empirical probability that ranking scores for ‘positives’ exceed those for ‘negatives’ (Hanley and McNeil, 1982; Cortes and Mohri, 2004). In the present context, this corresponds to the empirical probability that the values of test statistics for differentially expressed genes exceed those for non-differentially expressed genes. We make use of this relationship to express $AUC$ in decision-
theoretic terms, as a function of state of nature \( \omega \) and a vector \( T \) of test statistics:

\[
AUC(\omega, T) = \frac{\sum_{j: \omega^j = \omega_0} \sum_{i: \omega^i = \omega_1} I_{T_i > T_j}(T_i, T_j)}{\sum_{j: \omega^j = \omega_0} \sum_{i: \omega^i = \omega_1}} \frac{1}{g_0 g_1} \tag{4.12}
\]

\[
= \frac{\sum_{j: \omega^j = \omega_0} \sum_{i: \omega^i = \omega_1} I_{T_i > T_j}(T_i, T_j)}{g_0 g_1} \tag{4.13}
\]

where, \( I_A \) denotes the indicator function for a set \( A \).

\( AUC \) quantifies the ability of a test statistic to tell apart differentially expressed genes and non-differentially expressed genes. It is a good measure of the usefulness of a test statistic for gene selection, and one which is independent of threshold or number of genes selected. We note however that (negative) \( AUC \) is not a true loss function since it is a function of the state of nature and a vector of test statistics, rather than the state of nature and vector of actions. Thus, while a true loss or utility function has as its domain the set \( \Omega \times \mathcal{A} \), \( AUC \) has domain \( \Omega \times \mathbb{R}^q \).

### 4.4 Risk in gene selection

The notion of loss is inherently data-dependent. If the loss for a procedure \( \delta_1 \) is lower than for procedure \( \delta_2 \), we can conclude only that \( \delta_1 \) is more effective than \( \delta_2 \) for the data at hand. In order to draw a more general conclusion regarding the relative effectiveness of the two procedures, we must take account of the random nature of loss.
Risk is the average cost incurred when a decision procedure is applied to data obtained in a particular state of nature. A risk function $R$ for gene selection is therefore a function of state of nature vector $\omega$, and gene selection procedure $\delta$, and is obtained by taking the expectation of the loss function under the true data-generating distribution. Writing the loss function explicitly in terms of the data matrix $D$, as $L(\omega, \delta(D))$, and integrating under the conditional density $p(D \mid \omega)$, we get:

$$R(\omega, \delta) = \int_S L(\omega, \delta(D))p(D \mid \omega) \, dD \quad (4.14)$$

The quantity $AUC$ is also derived from random data, and is itself a random quantity. This becomes clear if $AUC$ is written explicitly in terms of the random matrix $D$. $AUC$ is a function of state of nature $\omega$ and a vector $T$ of test statistics; if the vector $T$ is written as a function $T(D)$ of data matrix $D$, $AUC$ may be expressed as $AUC(\omega, T(D))$. In analogy to risk, the expected value of $AUC$ under the true data-generating distribution may then be expressed as follows:

$$E[AUC] = \int_S AUC(\omega, T(D))p(D \mid \omega) \, dD \quad (4.15)$$

We call this latter quantity the expected area under the ROC curve or $E[AUC]$. Thus, while $AUC$ offers a sample-dependent measure of the ability of a test statistic to rank genes correctly, $E[AUC]$ is a measure of ranking quality which depends only on the test statistic and underlying data-generating distribution. As discussed previously, while the (negative) area under the ROC curve is not a true loss func-
tion, it can be thought of as roughly analogous to a loss function; \( E[AUC] \) can be thought of as roughly analogous to risk.

The expression for risk and the expected area under the ROC curve in Equations 4.14 and 4.15 above involve taking averages under a data-generating density \( p(D \mid \omega) \), conditioned on a state of nature vector \( \omega \). But what does it mean to say that the state of nature is \( \omega \), and that the matrix \( D \) of gene expression data is distributed according to \( p(D \mid \omega) \)? It means that \( p(D \mid \omega) \) is the limiting distribution of experimentally observed gene expression levels obtained from the particular biological system under study. An example will clarify this statement. Suppose the phenomenon under study is the metabolism of galactose by yeast, and that the data in the two conditions represent yeast gene expression levels in galactose-rich and galactose-poor growth media respectively. In the context of this particular biological question, each of the genes under consideration is either truly differentially expressed or not. The state of nature \( \omega \) captures this underlying truth. Suppose that the entire microarray experiment is performed a large number of times, and that the observed data from the series of experiments form a sequence \( D_1, D_2, D_3, \ldots \). When we say a dataset \( D \) is distributed according to \( p(D \mid \omega) \), we mean that if the number of experiments in a hypothetical series of this kind were to grow large, the empirical distribution defined by the results of those experiments would converge to the probability density function \( p(D \mid \omega) \).

To put this another way: if we drew samples from the density \( p(D \mid \omega) \), we would obtain a distribution of values that would, in the limit, be identical to the distribution obtained by experiment. Conditioning upon \( \omega \) makes the density specific to a
particular biological question, in this case the comparison between galactose-rich and galactose-poor conditions. Thus, the notion of a data-generating distribution conditioned upon a state of nature is intended to capture, at least in a conceptual sense, both experimental and biological variability, in the context of a specific biological system.

### 4.5 Computing risk from sampling distributions

The aim of this Section is to show how risk in gene selection can be computed directly from the sampling distributions of test statistics. We first introduce a few useful expressions, and then look in turn at each of FDR loss, zero-one loss and the area under the ROC curve.

#### 4.5.1 Preliminaries

**Sampling distributions.** The information contained in the data is summarised for the decision making procedure by the test statistic. As a consequence, the sampling distributions of the test statistic under the null and alternative hypotheses play a central role in computing risk. If we assume that the data for every non-differentially expressed gene (NDEG) is i.i.d., then the sampling distributions for the univariate test statistics corresponding to those NDEGs are identical. We call the sampling distribution for NDEGs $p_0$:

$$ p_0(t) \overset{def}{=} p(T_k = t \mid \omega^k = \omega_0) \quad (4.16) $$
Similarly, we assume the data for truly differentially expressed genes (DEGs) is i.i.d. The sampling distribution for the test statistics corresponding to the DEGs is \( p_1 \):

\[
p_1(t) \overset{\text{def}}{=} p(T_k = t \mid \omega^k = \omega_1) \quad (4.17)
\]

The cumulative distribution functions (cdfs) corresponding to \( p_0 \) and \( p_1 \) are \( \phi_0 \) and \( \phi_1 \) respectively:

\[
\phi_0(t) \overset{\text{def}}{=} \int_{-\infty}^{t} p_0(u) \, du \quad (4.18)
\]
\[
\phi_1(t) \overset{\text{def}}{=} \int_{-\infty}^{t} p_1(u) \, du \quad (4.19)
\]

**True and false positive rates.** The *true positive rate* \( \beta \) is the probability of selecting a DEG, while the *false positive rate* \( \alpha \) is the probability of (mistakenly) selecting a NDEG:

\[
\beta \overset{\text{def}}{=} P(a^k = a_1 \mid \omega^k = \omega_1) \quad (4.20)
\]
\[
\alpha \overset{\text{def}}{=} P(a^k = a_1 \mid \omega^k = \omega_0) \quad (4.21)
\]

At a given threshold \( \tau \), the probability \( \beta(\tau) \) of a DEG being selected is simply
$P(T_k > \tau \mid \omega^k = \omega_1)$, which can be expressed in terms of the cdf $\phi_1$ as follows:

\[
\beta(\tau) = P(T_k > \tau \mid \omega^k = \omega_1) \tag{4.22}
\]

\[
= 1 - P(T_k \leq \tau \mid \omega^k = \omega_1) \tag{4.23}
\]

\[
= 1 - \phi_1(\tau) \tag{4.24}
\]

If genes are ranked under the absolute value of a test statistic, this expression must be modified slightly to take account of negative values of the statistic. When $T_k > 0$, $|T_k| = T_k$, and $P(|T_k| > \tau \mid \omega^k = \omega_1) = 1 - \phi_1(\tau)$ as shown above. But when $T_k \leq 0$, $|T_k| = -T_k$, and $P(|T_k| > \tau \mid \omega^k = \omega_1)$ is equivalent to $P(T_k \leq -\tau \mid \omega^k = \omega_1)$, or $\phi_1(-\tau)$, such that the overall true positive rate is:

\[
\beta(\tau) = 1 - \phi_1(\tau) + \phi_1(-\tau) \tag{4.25}
\]

Similarly, the false positive rate $\alpha$ can be written as a function of threshold $\tau$, by making use of the cdf $\phi_0$:

\[
\alpha(\tau) = 1 - \phi_0(\tau) \tag{4.26}
\]

If genes are ranked under the absolute value of the test statistic the false positive rate is given by the following expression:

\[
\alpha(\tau) = 1 - \phi_0(\tau) + \phi_0(-\tau) \tag{4.27}
\]
All of the procedures studied in this Chapter rank genes under the absolute value of a test statistic, so we will subsequently use Equations 4.25 and 4.27 for the true and false positive rates.

**Selecting a pre-determined number of genes.** As noted by Lönnstedt and Speed (2002), in many cases the number of genes to be selected depends on factors like the scale of the experiment, follow-up plans and so on, and should therefore be regarded as an integral part of the experimental set-up. In the remainder of this Chapter we will concentrate on the case in which the number of genes to be selected is pre-determined. We note, however, that all of the results presented below can be adapted very easily for the case in which an explicit threshold is used.

The true positive rate \( \beta(\tau) \) expresses the probability of correctly selecting a DEG in terms of a threshold \( \tau \); in analogy with \( \beta(\tau) \), we may think of \( \beta(s) \) as the true positive rate in terms of the number \( s \) of genes selected\(^1\). From the definition of the true positive rate given in Equation 4.20 and the decision rule for the case where a fixed number of genes are selected given in Equation 4.5, we define \( \beta(s) \) as follows:

\[
\beta(s) \overset{\text{def}}{=} P(\text{rank}(T_k) \geq s \mid \omega^k = \omega_1) \quad (4.28)
\]

This quantity may be expressed as the expectation of \( \beta(\tau) \) under the condi-

\(^1\)Note that when the number of genes is pre-determined, it is no longer a random variable, and is therefore denoted by lower-case \( s \), instead of \( S \).
tional density \( p(\tau \mid s) \):

\[
\beta(s) = \int_0^\infty \beta(\tau) p(\tau \mid s) \, d\tau \tag{4.29}
\]

This latter expression can be derived from the definition of \( \beta(s) \). First, we express \( \beta(s) \) in terms of an order statistic \( T^{(s')} \):

\[
\beta(s) \overset{def}{=} P(\text{rank}(T_k) \geq s \mid \omega^k = \omega_1) \tag{4.30}
= P(T_k \geq T^{(s')} \mid \omega^k = \omega_1) \tag{4.31}
\]

where, \( T^{(s')} \) is the \( s' \)th order statistic of the complete set of \( g \) test statistics, comprising \( g_1 \) test statistics from DEGs, and \( g_0 \) statistics from NDEGs, and \( s' = g - s + 1 \). In other words, \( T^{(s')} \) is the \( s \)th largest test statistic \( T_k \).

Let the pdf of \( T^{(s')} \) be \( f \), such that \( T^{(s')} \sim f \). Assuming, for simplicity, the one-sided case of non-negative \( T_k \), we may then write:

\[
\beta(s) = \int_0^\infty P(T_k \geq u \mid \omega^k = \omega_1) f(u) \, du \tag{4.32}
\]

But the pdf of \( T^{(s')} \) is equivalent to \( p(\tau \mid s) \). That is, if \( p(\tau \mid s) \) is written as a function \( q_s(\tau) \) of \( \tau \), we may write \( f = q_s \). Then from Eq. 4.32:

\[
\beta(s) = \int_0^\infty P(T_k \geq u \mid \omega^k = \omega_1) q_s(u) \, du \tag{4.33}
\]
Figure 4.3: Computational verification of $\beta(s) = \int \beta(\tau)p(\tau \mid s) \, d\tau$. The true value of $\beta(s)$ and the quantity $\int \beta(\tau)p(\tau \mid s) \, d\tau$, i.e. the right-hand side of Equation 4.29, are plotted against the number $s$ of genes selected, for values of $s$ ranging from 1 to the total number of genes $g = 1025$.

Substituting $\tau$ for $u$:

$$\beta(s) = \int_0^\infty P(T_k \geq \tau \mid \omega^k = \omega_1) q_s(\tau) \, d\tau \quad (4.34)$$

And replacing $q_s(\tau)$ with $p(\tau \mid s)$:

$$\beta(s) = \int_0^\infty P(T_k \geq \tau \mid \omega^k = \omega_1) p(\tau \mid s) \, d\tau \quad (4.35)$$

Finally, if we use a definition of $\beta(\tau)$ as $P(T_k \geq \tau \mid \omega^k = \omega_1)$, we get the expression shown in Equation 4.29:

$$\beta(s) = \int_0^\infty \beta(\tau)p(\tau \mid s) \, d\tau$$

We further verified Equation 4.29 by means of a simple “brute force” simula-
tion in which we calculated the left-hand and right-hand sides of the Equation by computational means, in particular:

(i) We generated 100 datasets under an i.i.d. Normal model. The model is fully described in Section 4.6; we used the same parameters as in the ‘conventional normal’ case described in that Section, namely $\mu_{k(0)}, \mu_{Xk(1)} = 0$, $\sigma^2_{k(0)}, \sigma^2_{k(1)} = 2, \mu_{Yk(1)} = 3, g_0 = 1000, g_1 = 25$, and $m, n = 5$.

(ii) We calculated $\beta(s)$ directly, by calculating the average proportion of true positives obtained, for values of $s$ ranging between 1 and $g = 1025$. We call this ‘true $\beta(s)$’. The statistic used was difference of means.

(iii) Finally, we also calculated the right-hand side of Equation 4.29, namely $\int \beta(\tau)p(\tau \mid s) \, d\tau$, by taking the average of the function $\beta(\tau)$ (as given in Equation 4.25) under the empirically observed distribution of the order statistic $T^{(s')}$, which, as discussed above, is equivalent to $p(\tau \mid s)$.

Figure 4.3 shows the results obtained; true $\beta(s)$ and the right-hand side of Equation 4.29, i.e. the quantity $\int \beta(\tau)p(\tau \mid s) \, d\tau$, are virtually identical at values of $s$ ranging from 1 to the total number of genes $g = 1025$.

Finally, we suggest the following unnormalised density function as an approximation to $p(\tau \mid s)$, presenting the full derivation in Appendix B:

$$p(\tau \mid s) \propto \frac{1}{\sqrt{2\pi v(\tau)}} e^{-[s-m(\tau)]^2/[2v(\tau)]}$$  (4.36)
where, $m(\tau)$ and $v(\tau)$ are functions of threshold $\tau$ and are given by:

\[ m(\tau) = g_1 \beta(\tau) + g_0 \alpha(\tau) \]  
(4.37)

\[ v(\tau) = g_1 \beta(\tau)(1 - \beta(\tau)) + g_0 \alpha(\tau)(1 - \alpha(\tau)) \]  
(4.38)

### 4.5.2 FDR

We call the risk function obtained by averaging the FDR loss function $L_{FDR}$ under the true data-generating distribution the **FDR risk function** $R_{FDR}$:

\[ R_{FDR}(\omega, \delta) = E[L_{FDR}(\omega, \delta(D)) p(D | \omega)] \]  
(4.39)

\[ = E[L_{FDR}] \]  
(4.40)

Recall that $L_{FDR}$ can also be expressed as a fraction $S_0/s$ where $S_0$ is the number of NDEGs selected and $s$ the total number of genes selected. The risk function $R_{FDR} = E[L_{FDR}]$ is therefore equivalent to $E[S_0/s]$:

\[ R_{FDR} = E \left[ \frac{S_0}{s} \right] \]
\[ = E \left[ \frac{s - S_1}{s} \right] \]
\[ = 1 - \frac{E[S_1]}{s} \]  
(4.41)

If the expected number of true positives $E[S_1]$ can be computed, the exact value of the FDR risk function follows from Equation 4.41. Now, the true positive
rate $\beta(s)$ is the probability of selecting a DEG when $s$ genes are selected. Since we have assumed that the data for every DEG is i.i.d., it follows that the probability of each DEG being selected is identical to $\beta(s)$, and that the distribution over the number of DEGs selected is Binomial. If the number of genes selected equals or exceeds the total number of DEGs (i.e. if $s \geq g_1$), the probability parameter of the Binomial is $\beta$, and the number of Bernoulli trials $g_1$, such that $S_1 \sim B(\beta(s), g_1)$.

Accounting also for the $s < g_1$ case, we get the distribution of $S_1$:

$$S_1 \sim B \left( \beta(s) \times \max \left( \frac{g_1}{s}, 1 \right), \min(s, g_1) \right) \quad (4.42)$$

Then, for any value of $s$, $E[S_1] = \beta(s)g_1$. From Equations 4.25, 4.29 and 4.41 we may then express the FDR risk function in terms of the cdf $\phi_1$:

$$R_{FDR} = 1 - \frac{\beta(s)g_1}{s} \quad (4.43)$$

$$= 1 - \frac{g_1}{s} \int_{0}^{\infty} \left[ 1 - \phi_1(\tau) + \phi_1(-\tau) \right] p(\tau \mid s) \, d\tau \quad (4.44)$$

We compute the final integral numerically, by drawing samples from the approximation to $p(\tau \mid s)$ given in Equation 4.36. Note also that in general sampling distributions are not known in real-world applications, so the approach taken here cannot be applied to experimental data.
4.5.3 Zero-one loss

We call the risk function obtained by averaging the zero-one loss function $L_{01}$ under the true data-generating distribution the \textit{zero-one risk function} $R_{01}$:

\[
R_{01}(\omega, \delta) = \int S L_{01}(\omega, \delta(D)) p(D | \omega) \, dD
\]

\[
= E[L_{01}]
\]

(4.45)

(4.46)

Recall that $L_{01}$ can also be expressed in terms of total number of genes $g$, total number of DEGs $g_1$, number of DEGs selected $S_1$ and number of NDEGs selected $S_0$ as $(g_1 - S_1 + S_0)/g$. The risk function $R_{01} = E[L_{01}]$ is therefore equivalent to $E[(g_1 - S_1 + S_0)/g]$:

\[
R_{01} = E \left[ \frac{g_1 - S_1 + S_0}{g} \right]
\]

(4.47)

\[
= E \left[ \frac{g_1 + s - 2S_1}{g} \right]
\]

(4.48)

\[
= \frac{g_1}{g} + \frac{s}{g} - \frac{2}{g} E[S_1]
\]

(4.49)

Using $E[S_1] = \beta(s)g_1$, and Equations 4.25, 4.29 and 4.49, the zero-one risk function may then be expressed directly in terms of cdfs $\phi_0$ and $\phi_1$:

\[
R_{01} = \frac{g_1}{g} + \frac{s}{g} - \frac{2}{g} \beta(s)g_1
\]

(4.50)

\[
= \frac{g_1}{g} + \frac{s}{g} - \frac{2g_1}{g} \int_0^\infty [1 - \phi_1(\tau) + \phi_1(-\tau)] p(\tau | s) \, d\tau
\]

(4.51)
4.5.4 AUC

The area under the ROC curve was expressed in Equation 4.13 as a function of state of nature \( \omega \) and a vector \( T \) of test statistics. If genes are ranked under the absolute value of the test statistic, this expression must be modified slightly, as shown below:

\[
AUC(\omega, T) = \sum_{j: \omega^j = \omega_0} \sum_{i: \omega^i = \omega_1} I[T_i > |T_j|] \frac{T_i}{g_0 g_1} (4.52)
\]

where, \( I_A \) denotes the indicator function for a set \( A \).

In this case, the area under the ROC curve is equivalent to the empirical probability that the absolute values of test statistics for DEGs exceed those for NDEGs. Then, the expected area under the curve is simply the probability \( P(|T_{NDEG}| < |T_{DEG}|) \), where \( T_{DEG} \) is a random variable representing the value of the test statistic for a DEG, and \( T_{NDEG} \) a random variable representing the value of the test statistic for a NDEG. We can therefore write \( E[AUC] \) directly in terms of the null sampling density \( p_0 \), and alternative cdf \( \phi_1 \), as follows:

\[
E[AUC] = P(|T_{NDEG}| < |T_{DEG}|) \quad (4.53)
\]

\[
= \int_{-\infty}^{\infty} P(|T_{DEG}| > |u|) p_0(u) \, du \quad (4.54)
\]

\[
= \int_{-\infty}^{\infty} (1 - \phi_1(|u|) + \phi_1(-|u|)) p_0(u) \, du \quad (4.55)
\]
4.6 Results

In this Section, we present results obtained using the analysis of risk developed above. We previously divided test statistics for gene selection into three categories, “naive”, “classical”, and “microarray-specific”. Here, we analyse the performance of three widely-used test statistics, each a representative of one category. In particular, we look at the t-statistic as an example of a classical two-sample test statistic; the SAM statistic, as an example of a microarray-specific, moderated statistic; and difference of means, as an example of a “naive” statistic which ignores variance altogether. For each of these statistics, we compute risk under FDR and zero-one loss functions, as well as the expected area under the ROC curve. These performance measures are used to compare the three procedures under various conditions relevant to gene selection, including differing variances, small to moderate sample sizes and the presence of large numbers of non-differentially expressed genes.

We assume a Normal model for suitably transformed expression values. The data for NDEGs is taken to be i.i.d. under a Normal distribution with identical population mean \( \mu_k(0) \) in both conditions, and population variance \( \sigma^2_{k(0)} \):

\[
X_{ik}, Y_{jk} | \omega_k = \omega_0 \sim \mathcal{N}(\mu_k(0), \sigma^2_{k(0)}) \quad (4.56)
\]

(Here, the subscript ‘(0)’ indicates that the parameters \( \mu_k(0) \) and \( \sigma^2_{k(0)} \) pertain to NDEGs.)

The data for DEGs is also taken to be Normally distributed, but with distinct
means $\mu_{X(k)}$ and $\mu_{Y(k)}$ \((\mu_{X(k)} \neq \mu_{Y(k)})\) in the two conditions, and variance $\sigma_{k(1)}^2$:

\[
X_{ik} \mid \omega^k = \omega_1 \overset{iid}{\sim} \mathcal{N}(\mu_{X(k)}, \sigma_{k(1)}^2) \quad (4.57)
\]

\[
Y_{jk} \mid \omega^k = \omega_1 \overset{iid}{\sim} \mathcal{N}(\mu_{Y(k)}, \sigma_{k(1)}^2) \quad (4.58)
\]

Note that in this formulation population variances may differ between DEGs and NDEGs, but are the same in both conditions for a particular gene.

### 4.6.1 Conventional normal model

We begin by examining the performance of the test statistics under a conventional Normal model, with equal population variances $\sigma_{k(1)}^2$ and $\sigma_{k(0)}^2$ for DEGs and NDEGs. The sample size in microarray experiments is typically quite small; accordingly we set the sample size to 5 in each condition \((m = n = 5)\). The sample size enters into our assessment via the sampling distributions shown in Equations 4.16 - 4.19, which in turn impact on the expressions for risk and area under the ROC curve shown in Equations 4.44, 4.51 and 4.55. Figure 4.4 shows risk under FDR and zero-one loss functions, and the expected area under the ROC curve, computed under this model. Specific parameter values used here are as follows: $\mu_{k(0)} = 0$, $\sigma_{k(0)}^2 = 2$, $\mu_{X(k)} = 0$, $\sigma_{k(1)}^2 = 2$, and $\mu_{Y(k)} = 3$. The total number of NDEGs $g_0 = 1000$ and the total number of DEGs $g_1 = 25$, while the sample sizes are $m = n = 5$. Risk is calculated assuming that 25 genes are selected, i.e. $s = 25$. 

Figure 4.4: Risk and expected area under the ROC curve, under a conventional Normal model.
Risk, under both loss functions, is highest for the t-statistic, followed by SAM and difference of means. The expected area under the ROC curve echoes this finding: the t-statistic has the smallest value of $E[AUC]$. The relatively poor performance of the t-statistic here may be partly due to small-sample effects on the standard deviation term. The standard deviation term in the denominator of the statistic is highly variable at small sample sizes, with the effect that purely by chance, many non-differentially expressed genes end up having relatively small estimated standard deviations, and correspondingly high absolute values of the statistic. This in turn causes these genes to be assigned high ranks under the statistic, resulting in a large number of false positives. SAM has a small constant term added to the standard deviation in its denominator, which makes it far less sensitive in this regard. The difference of means statistic does not have a standard deviation term at all, and as a consequence does not suffer as acutely from deleterious small-sample effects. However, it is also worth noting that the very fact that the data-generating model used here is entirely i.i.d. benefits difference of means since there is essentially no mismatch between the true model and the difference of means model.

4.6.2 Larger sample sizes

If small-sample effects are to blame for the poor performance of the t-statistic in the previous experiment, we should expect it to do better at larger sample sizes, not just in absolute terms, but relative to the other methods. Figure 4.5 shows risk under FDR and zero-one loss functions, and the expected area under the ROC
(a) Risk under FDR loss  
(b) Risk under zero-one loss  
(c) Expected area under the ROC curve

Figure 4.5: Varying sample size. Risk, under FDR and zero-one loss functions, and expected area under the ROC curve, are plotted against total sample size.
curve, plotted against total sample size \( m + n \). Both sample sizes are equal, i.e. \( m = n \), and the \( x \)-axis shows the total sample size, i.e. \( m + n \). Also, in order to provide the statistics with a sufficiently challenging problem at larger sample sizes, we set both variances to 5, i.e. \( \sigma^2_{k(0)}, \sigma^2_{k(1)} = 5 \). All other parameters are as stated in the previous section, specifically, \( \mu_{k(0)} = 0, \mu_{Xk(1)} = 0, \mu_{Yk(1)} = 3, g_0 = 1000, g_1 = 25 \) and \( s = 25 \).

The difference in risk and expected area under the ROC curve between the t-statistic on the one hand, and SAM and difference of means on the other, does indeed decrease with increasing sample size. We note however, that total sample sizes exceeding 20 or so are quite unusual in microarray studies, and that the set-up in the previous experiment is far more typical.

### 4.6.3 Differing variances

Subsets of genes which are perturbed in comparative experiments may have distinct underlying variances in comparison with other genes. For example, suppose a specific group of transcription factors is up-regulated in a particular biological condition. It is then highly plausible that expression levels for these transcription factors may have different population variances compared with the other genes. This scenario of variances differing between DEGs and NDEGs is quite distinct from the well-studied Behrens-Fisher problem (DeGroot and Schervish, 2002) of variances differing between conditions. In order to better understand the effects of systematic differences of this kind, we compute risk and expected area under the ROC curve while allowing population variances to differ between DEGs and
Figure 4.6: Differing variances. Risk, under FDR and zero-one loss functions, and expected area under the ROC curve, are plotted against the log-ratio of population variances of DEGs to NDEGs, i.e. $\log_2(\sigma^2_{k(1)}/\sigma^2_{k(0)})$. 
NDEGs. Specifically, the population variance $\sigma_{k(1)}^2$ for DEGs is allowed to vary in the range $[1, 4]$, while the corresponding value $\sigma_{k(0)}^2$ for NDEGs is held constant at 2. Figure 4.6 shows the three performance measures plotted against the log-ratio $\log_2(\sigma_{k(1)}^2/\sigma_{k(0)}^2)$ of the population variances of DEGs to NDEGs. All other parameters are as stated for the ‘conventional Normal model’, specifically, $\mu_{k(0)} = 0, \mu_{X_{k(1)}} = 0, \mu_{Y_{k(1)}} = 3, g_0 = 1000, g_1 = 25, m = 5, n = 5$ and $s = 25$.

Risk, under both loss functions, is generally higher for the t-statistic. But what is noticeable is that the t-statistic does particularly badly when the variance for DEGs exceeds the variance for NDEGs. Indeed, the difference in risk between the t-statistic and the other methods increases dramatically with the ratio $\sigma_{k(1)}^2/\sigma_{k(0)}^2$. We attribute this phenomenon to the effect of the standard deviation term in the statistic. We can think of the sample means and standard deviation in the t-statistic as playing the role of ‘features’ in discriminating between DEGs and NDEGs. Then, the very fact that sample standard deviations tend to be higher for DEGs when $\sigma_{k(1)}^2 > \sigma_{k(0)}^2$ means that DEGs end up being penalised by the statistic, leading to higher error rates.

Looking at the relative performance of difference of means and SAM, what is interesting is that when the variance for DEGs exceeds the variance for NDEGs, difference of means does relatively well, but when the reverse is true, SAM is more effective. As a consequence, while difference of means does well overall, there is no clear ‘winner’ across the range of variance ratios: if variances differ, the ‘best’ statistic to choose depends to a certain extent on the underlying variance ratio.
Figure 4.7: Varying numbers of non-differentially expressed genes. Risk, under FDR and zero-one loss functions, and expected area under the ROC curve, are plotted against the total number $g_0$ of non-differentially expressed genes, while the number of differentially expressed genes is held constant at 25.
4.6.4 Varying numbers of non-differentially expressed genes

In many microarray experiments, especially those pertaining to higher organisms, the total number of genes under study may be very large indeed. Furthermore, when an experimental perturbation is very specific in nature, it may be the case that the number of DEGs is vastly exceeded by the number of NDEGs. To better understand the behaviour of gene selection procedures under such conditions, we allow the number $g_0$ of NDEGs to vary in the range $[500, 10000]$, while holding the number of DEGs constant at $g_1 = 25$. Figure 4.7 shows risk under FDR and zero-one loss functions, and the expected area under the ROC curve, plotted against the total number $g_0$ of NDEGs. All other parameters are as stated for the conventional Normal model, specifically, $\mu_{k(0)} = 0$, $\sigma^2_{k(0)} = 2$, $\mu_{Xk(1)} = 0$, $\sigma^2_{k(1)} = 2$, $\mu_{Yk(1)} = 3$, $g_1 = 25$, $m = 5$, $n = 5$ and $s = 25$. The difference in risk, under both loss functions, between the t-statistic on the one hand, and SAM and difference of means on the other, seems to grow larger with $g_0$.

4.7 Discussion

In this Chapter, we presented an analysis of gene selection in which statistical decision theory was used to better understand the behaviour of gene selection procedures under various conditions. We began by placing the task of selecting differentially expressed genes within the framework of classical decision theory. This approach stands in contrast to the somewhat ad hoc nature of existing work on the comparison of gene selection procedures (e.g. Pan, 2002; Broberg, 2003) and of-
fers several important advantages. It shows explicitly that performance measures such as loss and the area under the ROC curve should properly be regarded as random variables, and that a suitably defined notion of risk can be used to take account of the random nature of these measures. Furthermore, a decision-theoretic analysis shows unambiguously that the risk associated with gene selection depends upon both the statistical procedure used and the underlying data-generating distribution. An important implication, to which we will return in the next chapter, is that the ability of any particular method to accurately select differentially expressed genes must depend upon statistical characteristics of the underlying biological system.

Existing analyses of gene selection have also been overly reliant on brute-force simulation. Yet the sheer size of microarray datasets limits the scope of simulation studies, as only a small subset of conditions can possibly be explored if the effects of comparing thousands of genes are to be accounted for by computational means alone. For example, consider addressing the ‘differing variances’ problem by simulation. Obtaining accurate estimates of risk or expected area under the ROC curve would then require the sampling of several hundred high-dimensional datasets for each value of the variance ratio. In contrast, we computed risk values directly from relevant sampling distributions. As a consequence, we were able to very easily study various conditions of interest with little computational overhead.

However, a limitation of our approach was the assumption of i.i.d. expression levels for DEGs and NDEGs. While these assumptions facilitated our theoretical approach to the calculation of error rates in gene selection, they are biologically
unrealistic. The i.i.d. assumptions made here also provide an explanation for the good performance of the simple difference of means statistic. Under i.i.d. conditions, especially when null and alternative variances are equal, difference of means is in some sense the perfect statistic, as it is based upon the true data-generating model.

Our results, comparing the t-statistic, SAM, and difference of means, explicitly took account of small sample size. In contrast, classical results concerning the optimality of hypothesis testing procedures (e.g. Lehmann, 1986) are largely asymptotic in nature. Yet, very small sample sizes are actually quite typical of microarray experiments. Such sample sizes are considerably smaller than those found in conventional problems in statistics and machine learning, and may limit the utility of asymptotic results in the study of gene selection procedures. Indeed, our approach resulted in several insights which are not immediately apparent from classical asymptotic results. For example, our results suggested that the t-statistic may be an unsuitable choice for gene selection when sample sizes are small. Previous work (e.g. Baldi and Long, 2001; Cui and Churchill, 2003) has drawn attention to the potential unsuitability of the t-statistic for small sample size microarray data, and our results add further weight to these arguments, albeit under i.i.d. assumptions. Also, we were able to show that the t-statistic does badly when the population variances for differentially expressed genes are high relative to those of non-differentially expressed genes. However, it is worth noting that a key advantage of the t-statistic is that its null sampling distribution can be specified without reference to unknown population parameters, such that exact P-values can
be computed very easily, even at small sample sizes. Thus, in applications where
the explicit and precise control of Type I error is especially important, it may be
necessary to trade off good ranking performance against accurate estimation of
P-values, and at times the t-statistic may therefore be a good choice.
Chapter 5

Data-adaptive test statistics for gene selection

The risk associated with a gene selection procedure in a specific biological context depends inevitably upon the underlying data-generating distribution. As a consequence, individual test statistics may be well-suited to data drawn under certain conditions, but may perform poorly under other conditions. As we have seen, decision theory may be used to quantify the effectiveness of gene selection procedures under specific conditions. However, in practice, uncertainty regarding the underlying data-generating distribution means that it is difficult to make a good choice of test statistic on the basis of theoretical results alone. Equally, the small sample size and high-dimensionality which are characteristic of microarray data exacerbate the effects of model uncertainty by making it difficult to reliably characterise underlying distributions or verify modelling assumptions. All of this
can mean that, in practice, results obtained using conventional hypothesis testing procedures may not always be very reliable.

The underlying problem is that using a conventional test statistic for gene selection amounts to using a fixed function to rank genes, one which is selected on a priori grounds, and is not in any way responsive to the data. In a domain as uncertain as microarray analysis, this approach is simply not flexible enough to produce reliable results. Ideally then, we would like to choose a test statistic which is optimal for the data at hand; in other words, a statistic which minimises risk under the true data-generating distribution.

In this Chapter, we present a novel approach to gene selection in which test statistics are learned from data. As we shall see, certain features of the gene selection problem preclude a conventional approach to learning via the empirical minimisation of risk. We therefore propose a proxy for risk called “reproducibility”. Reproducibility, as we define it, can be computed without any knowledge of the “ground truth”, but takes advantage of some of the special properties of microarray data to provide a useful guide to risk under the true data-generating distribution. This allows us to exploit reproducibility to indirectly minimise risk, and thereby obtain results substantially more robust than conventional methods.

The remainder of this Chapter is organised as follows. Since our aim is to learn test statistics for gene selection, we start by considering risk-optimality in gene selection, looking at what it would entail to learn a risk-optimal statistic, and why it is difficult to do so. We define reproducibility, presenting theoretical results which show its validity as a proxy or surrogate for risk, and show how it can be
used to learn test statistics for gene selection. We go on to present experimental results on real and simulated microarray data and finally discuss some of the finer points and shortcomings of our approach.

5.1 Risk-optimality in gene selection

The risk function for gene selection was presented in Equation 4.14, and is reproduced below:

\[ R(\omega, \delta) = \int_S L(\omega, \delta(D))p(D | \omega) dD \]  

(5.1)

As before, \( \omega \) represents the underlying state of nature, \( \delta \) the gene selection procedure, \( S \) the sample space and \( D \) a microarray data matrix drawn under true data-generating distribution \( p \).

In the context of a specific biological problem, the state of nature \( \omega \) and underlying data-generating distribution \( p(\cdot | \omega) \) can be regarded as fixed but unknown. If the number of genes to be selected is pre-determined, risk depends only on the choice of test statistic \( T \). Then, we would ideally like to choose a test statistic \( T^* \) (from a suitably defined family \( T \)) which minimises risk:

\[ T^* \overset{def}{=} \arg \min_{T \in T} R(\omega, \delta_{T,s}) \]  

(5.2)

\[ = \arg \min_{T \in T} \int_S L(\omega, \delta_{T,s}(D))p(D | \omega) dD \]  

(5.3)

where, \( \delta_{T,s} \) denotes a gene selection procedure specified by test statistic \( T \) and
CHAPTER 5. DATA-ADAPTIVE TEST STATISTICS

117

a number $s$ of genes to be selected.

Although, as a matter of theory, the risk function can always be computed directly from its definition, in practical data analysis, the density $p$ is usually unknown, precluding any direct computation of risk, or its minimisation. Yet the need to choose statistics or functions which minimise risk arises in many areas of data analysis, including, for example, classification and regression. In these applications, while the true data-generating distribution is usually unknown, the availability of class labels or target values means that the loss function can at least be evaluated for observed data. This, in turn, permits the use of computational methods like cross-validation to estimate and empirically minimise risk.

In contrast to classification and regression, in which the loss function can be evaluated, in gene selection, we do not usually know which genes in a study are truly differentially expressed, and as a result cannot even evaluate the loss function for the data available. As a result, risk in gene selection cannot be minimised directly, nor by conventional empirical means.

5.2 Reproducibility

The approach to learning test statistics presented here is based on the use of a simple notion of reproducibility in selection results as a proxy for risk. While, for reasons outlined above, risk in gene selection cannot itself be computed in practice, reproducibility can be computed very easily, and requires no knowledge of the true data-generating distribution or state of nature. In this Section, we define
reproducibility, and examine its relationship with risk.

5.2.1 Definition

Consider two sets of data $D_a$ and $D_b$, drawn independently under the true data-generating distribution. Each of the two datasets contains expression levels for the same set of genes, indexed, as before, by $k = 1..g$. A test statistic $T$ produces two potentially distinct rankings of these $g$ genes from the two datasets. Let $L_a$ and $L_b$ represent sets containing the indices of the $s$ highest-ranked genes obtained from datasets $D_a$ and $D_b$ respectively.

Reproducibility $\varrho$ is then defined as the number of genes in common between the sets $L_a$ and $L_b$:

$$\varrho = |L_a \cap L_b|$$

where $|\cdot|$ denotes cardinality.

Our central claim is that, subject to certain conditions, reproducibility, as defined here, provides a useful guide to risk in gene selection, such that it may be used as a proxy for risk in the learning of test statistics.

$L_a$ and $L_b$ are sets containing the indices of genes selected by applying the gene selection procedure $\delta_{T,s}$ to datasets $D_a$ and $D_b$ respectively. For example, if $a_{i(b)}^1$ are the gene-level actions obtained from dataset $D_b$ using $\delta_{T,s}$, $L_b = \{i : a_{i(b)}^1 = a_1\}$. 
5.2.2 Risk and reproducibility

Prior to an analysis of the relationship between risk and reproducibility, it may be helpful to consider an intuitive argument which provides a good sense of why, how, and under what conditions the two quantities are related. Consider a scenario in which the total number of genes runs into the thousands, with the majority being non-differentially expressed genes or NDEGs. Suppose also that some proportion of the genes selected using a given test statistic are false positives. Then, provided a good number of these false positives are chosen more or less at random from the large pool of NDEGs, the variability in their identities will tend to be high, compared with the corresponding variation among the DEGs selected. Hence, the greater the proportion of DEGs among those selected, the more agreement there will tend to be between sets of results, and the higher the reproducibility.

Figure 5.1 illustrates this informal argument by analogy with the drawing of balls from hats. The process of selecting genes from microarray data is inherently random, and can be thought of as roughly analogous to the drawing of balls from hats via some random process. In the figure, blue balls represent DEGs, and yellow balls NDEGs. Two sets of selection results, $L_a$ and $L_b$, are obtained by drawing balls from the two hats: in each case, the proportion of false positive NDEGs depends on the risk associated with the selection procedure. Since the blue balls are drawn from a relatively small pool compared with the yellow balls, the balls which happen to be in common between the two sets of results will tend to be mostly composed of blue balls. Blue balls represent DEGs, and the reproducibility score $\varrho$ is simply the number of balls in common between the two
sets of results. Thus, the more blue balls there are in the sets, or the lower the risk, the higher the reproducibility score will tend to be.

Figure 5.1: Why reproducibility correlates with risk. Blue balls represent DEGs, and yellow balls NDEGS. Two sets of selection results are obtained by drawing balls from the two hats: in each case, the proportion of falsely selected NDEGs depends on the risk associated with the selection procedure. Since the blue balls are drawn from a relatively small pool compared with the yellow balls, the balls which happen to be in common between the two sets of results will tend to be mostly composed of blue balls, which represent DEGs. As a consequence, the more blue balls there are in the sets, or the lower the risk, the higher the reproducibility score will tend to be.

In order to present this argument more formally, we first express reproducibility in terms of risk, and then examine the conditions under which the two are (negatively) correlated. Recall that under a False Discovery Rate (FDR) notion of loss,
as defined in Equation 4.10, risk $R_{FDR}$ is equivalent to the expected proportion of false positives among the set of genes selected (see Section 4.5.2 for details). If $S_1$ is a random variable representing the number of true positives among the $s$ genes selected, $R_{FDR}$ may be expressed in the following form:

$$R_{FDR} = E\left[\frac{s - S_1}{s}\right] \quad (5.5)$$

We will use the FDR loss function throughout this Chapter, and in the interest of notational simplicity will subsequently omit the subscript ‘FDR’, referring to risk simply as $R$.

Assuming that all DEGs are i.i.d., and that all NDEGs are also i.i.d., the distribution over the number $S_1$ of true positives is Binomial. This was shown earlier in Equation 4.42, which is reproduced below:

$$S_1 \sim B\left(\beta(s) \times \max\left(\frac{g_1}{s}, 1\right), \min(s, g_1)\right) \quad (5.6)$$

where, as before, $\beta(s)$ is the true positive rate, and $g_1$ the total number of DEGs.

It was shown previously, in Equation 4.43, that $\beta(s)$ and FDR risk $R$ are related by the expression $\beta(s) = (1 - R)s/g_1$. Substituting this expression into Equation 5.6 above, and taking the expectation of $S_1$, we get:

$$E[S_1] = (1 - R)s \quad (5.7)$$
Now, let $S_{1(a)}$ and $S_{1(b)}$ represent the number of true positives obtained from the two datasets $D_a$ and $D_b$, using the same test statistic $T$, and with the total number of genes selected being $s$ in each case. Then, each of $S_{1(a)}$ and $S_{1(b)}$ is Binomially distributed according to Equation 5.6.

Reproducibility $\varrho$ is the number of genes in common between two sets of results. Let $\varrho$ comprise $\varrho_1$ DEGs and $\varrho_0$ NDEGs, such that $\varrho = \varrho_1 + \varrho_0$. How are $\varrho_1$ and $\varrho_0$ distributed, given the numbers $S_{1(a)}$ and $S_{1(b)}$ of true positives in the two sets? Consider taking one DEG at a time from set $L_b$ and trying to find it in set $L_a$. This is in effect a series of Bernoulli trials, with the number of ‘successes’ being the number $\varrho_1$ of DEGs in common between the two sets. Under the i.i.d. assumptions made above, every DEG has the same chance of appearing in the result-set, so the distribution over $\varrho_1$ is itself Binomial:

$$\varrho_1 \sim B\left(\frac{S_{1(a)}}{g_1}, S_{1(b)}\right)$$

(5.8)

Note that it has been assumed (without loss of generality) that $S_{1(a)} > S_{1(b)}$.

A similar argument can be made for the NDEGs, such that the distribution over $\varrho_0$ is also Binomial:

$$\varrho_0 \sim B\left(\frac{s - S_{1(b)}}{g_0}, s - S_{1(a)}\right)$$

(5.9)

where, as before, $g_0$ is the total number of NDEGs.
Taking the expectation of $\varrho$:

\begin{align*}
E[\varrho] &= E[\varrho_1] + E[\varrho_0] \\
&= E \left[ \frac{S_{1(a)}S_{1(b)}}{g_1} \right] + E \left[ \frac{(s - S_{1(b)})(s - S_{1(a)})}{g_0} \right] \\
&= E[S_{1(a)}S_{1(b)}] + E[(s - S_{1(b)})(s - S_{1(a))}] \\
&= E[S_{1(a)}]E[S_{1(b)}] + E[(s - S_{1(b)})(s - S_{1(a))}] \\
&= (1 - R)s^2 + (1 - R)s^2 - 2(1 - R)s^2 \\
&= 0
\end{align*}

Since $S_{1(a)}$ and $S_{1(b)}$ are independent, $E[S_{1(a)}S_{1(b)}] = E[S_{1(a)}]E[S_{1(b)}]$, and from Equation 5.7 $E[S_{1(a)}] = E[S_{1(b)}] = (1 - R)s$. Substituting in the expression above (for details see Appendix C):

\begin{align*}
E[\varrho] &= \frac{(1 - R)^2 s^2}{g_1} + \frac{s^2}{g_0} + \frac{(1 - R)^2 s^2}{g_0} - 2 \frac{(1 - R)s^2}{g_0} \\
&= \frac{(1 - R)^2 s^2}{g_1} + \frac{s^2}{g_0} + \frac{(1 - R)^2 s^2}{g_0} - 2 \frac{(1 - R)s^2}{g_0}
\end{align*}

Now, suppose the risk values associated with two test statistics $T_1$ and $T_2$ are $R_1$ and $R_2$ respectively. Suppose also that $T_1$ has lower risk than $T_2$. Then, the difference $\Delta R = R_1 - R_2$ between $R_1$ and $R_2$ is negative. Using Equation 5.12 and simplifying, the corresponding difference $\Delta E[\varrho] = E[\varrho_1] - E[\varrho_2]$ in expected reproducibility is given by:

\begin{align*}
\Delta E[\varrho] &= s^2(-\Delta R) \left[ (2 - (R_1 + R_2)) \left( \frac{1}{g_1} + \frac{1}{g_0} \right) - \frac{2}{g_0} \right] \\
&= s^2(-\Delta R) \left[ (2 - (R_1 + R_2)) \left( \frac{1}{g_1} + \frac{1}{g_0} \right) - \frac{2}{g_0} \right]
\end{align*}

Examining this expression it is clear that $E[\varrho]$ and $R$ are negatively correlated if and only if the term within square brackets is positive. Thus, a necessary and sufficient condition for negative correlation is:

\begin{align*}
(2 - (R_1 + R_2)) \left( \frac{1}{g_1} + \frac{1}{g_0} \right) - \frac{2}{g_0} &> 0 \\
(2 - (R_1 + R_2)) \left( \frac{1}{g_1} + \frac{1}{g_0} \right) - \frac{2}{g_0} &> 0
\end{align*}
Cancelling $g_0$ and simplifying (for details see Appendix C) yields the condition:

\[
\frac{R_1 + R_2}{2} < \frac{g_0}{g_0 + g_1}
\] (5.15)

The FDR risk associated with totally random selection is $g_0/(g_0 + g_1)$, since the average proportion of false positives obtained purely by chance must equal the overall proportion of NDEGs. Condition 5.15 can therefore be interpreted as meaning that the average of the risk of statistics $T_1$ and $T_2$ must be lower than the risk associated with entirely random gene selection. Thus, expected reproducibility is negatively correlated with true risk for any pair of test statistics which are in this sense ‘better than random’.

Equation 5.13 may also be re-written in the following form:

\[
\Delta E[\varrho] = -s^2 \Delta R \left[ \frac{2 - R_1 - R_2}{g_1} - \frac{R_1 + R_2}{g_0} \right]
\] (5.16)

Now, suppose the number $g_1$ of DEGs is held constant, and the number of NDEGs allowed to vary. The coefficient of $\frac{1}{g_0}$ is $(R_1 + R_2)$. Since $R_1$ and $R_2$ are expected proportions of false positives, $(R_1 + R_2) > 0$. Then, for a given difference $\Delta R$ in risk, the ‘response’ in terms of reproducibility must grow with $g_0$. In other words, a large proportion of nulls is beneficial to reproducibility, as suggested by the informal argument with which this Section began.
5.2.3 Violating i.i.d. assumptions

The theoretical results presented above show that risk and expected reproducibility are negatively correlated under i.i.d. conditions. However, these i.i.d. assumptions are very strong, and biologically unrealistic. We therefore carried out a simple empirical study to examine the impact of violations of this assumption on the correlation between risk and reproducibility.

In order to do so, we generated data under a hierarchical Normal model, with prior distributions for gene-level means and variances. Data in \( X \) and \( Y \) conditions were drawn from Normal densities with a Normal prior for the means and Gamma priors for the variances. Data for NDEGs were drawn as \( X_{ik}, Y_{jk} \sim \mathcal{N}(\mu_{k(0)}, \sigma_{k(0)}^2) \), with \( \sigma_{k(0)}^2 \sim \mathcal{G}(a_0, b_0) \). For DEGs: \( X_{ik} \sim \mathcal{N}(\mu_{Xk(1)}, \sigma_{k(1)}^2) \) and \( Y_{jk} \sim \mathcal{N}(\mu_{Yk(1)}, \sigma_{k(1)}^2) \), with \( \mu_{Xk(1)} \) fixed and \( \mu_{Yk(1)} \sim \mathcal{N}(\mu_{Yk(1)}', \sigma_{Yk(1)}'^2) \), and \( \sigma_{k(1)}^2 \sim \mathcal{G}(a_1, b_1) \). This model is also used to generate synthetic data in the Results section below, and is illustrated in Figure 5.5 in that Section as a probabilistic graphical model.

Our basic strategy was to use simulation to look at the behaviour of the reproducibility score under conditions of increasing inter-gene variability in underlying parameters. The general computational procedure was a straightforward “brute-force” simulation, comprising three steps at each iteration: (i) generate a pair of datasets (under either an i.i.d. or hierarchical model), (ii) compute reproducibility following Equation 5.4, that is as the size of the intersection between sets of true positives obtained from the two datasets, and (iii) compute loss under the FDR loss function. These three steps were repeated for 200 iterations, and the results
obtained used to empirically calculate risk and expected reproducibility. In all cases, the statistic used was a t-statistic.

We first looked at the relationship between risk and reproducibility under i.i.d. conditions, generating data under the Normal model discussed in the previous Chapter. Data in X and Y conditions were drawn from Normal densities with fixed parameters. In the NDEG case, $X_{ik}, Y_{jk} \overset{iid}{\sim} \mathcal{N}(\mu_{k(0)}, \sigma_{k(0)}^2)$; in the DEG case, $X_{ik} \overset{iid}{\sim} \mathcal{N}(\mu_{Xk(1)}, \sigma_{k(1)}^2)$ and $Y_{jk} \overset{iid}{\sim} \mathcal{N}(\mu_{Yk(1)}, \sigma_{k(1)}^2)$. In the same manner as the ‘differing variances’ results presented in the previous Chapter, the population variance $\sigma_{k(0)}^2$ for NDEGs was held constant at 2, while the corresponding variance $\sigma_{k(1)}^2$ for DEGs was allowed to vary in the range $[1, 4]$, providing a steady change in risk, against which to compare reproducibility. Other parameter values used here were as follows: $\mu_{k(0)} = 0, \mu_{Xk(1)} = 0, \mu_{Yk(1)} = 3, g_0 = 1000, g_1 = 25, m = 5, n = 5$ and $s = 25$.

Figure 5.2 shows results obtained under this i.i.d. model, with risk and expected reproducibility plotted against the log-ratio of the population variances of DEGs to NDEGs, i.e. $\log_2(\sigma_{k(1)}^2/\sigma_{k(0)}^2)$. Error bars show the standard deviations of reproducibility and risk (i.e. of loss). Note also that in the plots reproducibility is scaled by the number of genes selected $s$, and therefore lies in the interval $[0, 1]$.

Figure 5.3 shows results obtained using the hierarchical model described above. Panel (a) shows results obtained using a moderately variable prior for gene-level variances, while Panel (b) shows results obtained using a highly variable prior for variances. Note that since the variances are now themselves random quantities, the horizontal axes in these figures are log-ratios of expected variance.
Figure 5.2: Correlation between risk and reproducibility under i.i.d. conditions.

Panel (c) shows the distributions from which gene-level variances were drawn for NDEGs, and for comparison, the fixed value $\sigma^2_{k(0)} = 2$ used in the i.i.d. experiment described above. Panel (d) shows the same priors in terms of standard deviation. Panel (e) shows the prior used for the mean $\mu_{Yk(1)}$. The specific parameters used in the ‘moderately variable’ case were as follows: $\mu_{k(0)} = 0$, $\mu_{Xk(1)} = 0$, $a_{(0)} = 1$, $b_{(0)} = 2$, $\mu'_{Yk(1)} = 3$, $\sigma^2_{Yk(1)} = 0.5$. For DEGs, the variance of the NDEG Gamma prior defined by $a_{(0)}$ and $b_{(0)}$ was retained, but its expected value allowed to vary over ten steps in the range $[1, 4]$, that is from $0.5\sigma^2_{k(0)}$ to $2\sigma^2_{k(0)}$. (The precise series of values obtained in this way for the hyperparameters $a_{(1)}$ and $b_{(1)}$ were $a_{(1)} = 0.25, 0.44, 0.69, 1, 1.36, 1.78, 2.25, 2.78, 3.36, 4$ and $b_{(1)} = 4, 3, 2.4, 2, 1.71, 1.5, 1.33, 1.2, 1.09, 1$).

In the ‘highly variable’ case, all parameters were the same as above, except for the variance hyperparameters which were $a_{(0)} = 1$, $b_{(0)} = 5$ in the NDEG
Figure 5.3: Correlation between risk and reproducibility under a hierarchical model.
case. These parameters give a mean value of 5 for NDEG variances; the expected value of DEG variances was therefore allowed to vary over ten steps in the range $[2.5, 10]$. (This yielded hyperparameter values for $a^{(1)}$ as shown above, and $b^{(1)} = 10, 7.5, 6, 5, 4.29, 3.75, 3.33, 3, 2.73, 2.5$).

Finally, we performed these experiments once again, but with a more variable prior for the mean $\mu_{Y_k^{(1)}}$. These results are shown in Panels (a) and (b) of Figure 5.4 below. The prior for the mean is shown in Panel (c), along with the prior used in the previous experiment for comparison. All parameter values were the same as for the previous experiment, except for the variance hyperparameter for the mean, which was $\sigma^2_{Y_k^{(1)}} = 2$.

Overall, the negative correlation between risk and reproducibility appears to hold up well under the hierarchical model. Furthermore, there is little noticeable change in the standard error of observed reproductibilities. This suggests that reproducibility may be a useful guide to risk even when the i.i.d. assumptions underpinning our formal results are violated.

### 5.2.4 Estimating reproducibility

We use the bootstrap to estimate reproducibility from a microarray dataset. The bootstrap (Efron, 1979; Efron and Tibshirani, 1993; Davison and Hinckley, 1997) is a widely-used technique in computational statistics in which quantities of interest are estimated by resampling an original dataset. Datasets obtained by resampling are called ‘bootstrap datasets’, and each iteration in which such a dataset is constructed from the original data is called a ‘bootstrap iteration’.
Figure 5.4: Correlation between risk and reproducibility under a hierarchical model, with a higher variability prior for gene-level means.
Algorithm 1 Bootstrap estimation of reproducibility.

(1) Sample arrays with replacement from dataset $D$ to obtain a pair of bootstrap datasets $D^*_a$ and $D^*_b$, each having as many samples and genes as $D$.
(2) Apply test statistic $T$ to each of $D^*_a$ and $D^*_b$, to obtain two ranked lists of genes. Select the top $s$ genes in each case to obtain sets $L^*_a$ and $L^*_b$.
(3) Compute reproducibility $\varrho = |L^*_a \cap L^*_b|$.
(4) Repeat steps (1)-(3) $B$ times.
(5) Return the average of the $B$ reproducibility scores obtained.

We treat reproducibility $\varrho$ as the quantity to be estimated, and use the bootstrap as shown in Algorithm 1: The algorithm takes as input a test statistic $T$, dataset $D$, and number $s$ of genes to be selected. At each bootstrap iteration, entire arrays are resampled from the original dataset $D$, to generate a pair of bootstrap datasets $D^*_a$ and $D^*_b$. The arrays are resampled uniformly at random and with replacement, such that every array has the same chance of being selected, and may be selected more than once, or not at all. The bootstrap datasets $D^*_a$ and $D^*_b$ are resampled independently, and each made to contain as many arrays as the original dataset $D$. To take an illustrative example, if the original dataset is $D = [X_1 X_2 X_3 X_4 Y_1 Y_2 Y_3 Y_4]$, a pair of bootstrap datasets $D^*_a$ and $D^*_b$ obtained at a particular iteration might look like:

$$D^*_a = [X_3 X_3 X_1 X_2 Y_3 Y_1 Y_1 Y_2]$$
$$D^*_b = [X_2 X_1 X_3 Y_1 Y_4 Y_2 Y_4]$$

Genes are selected from each of the datasets $D^*_a$ and $D^*_b$, by ranking under...
test statistic $T$, and then selecting the top $s$ genes in each case. This produces, at each iteration, a pair of sets of results $L_a^*$ and $L_b^*$, from which the size $|L_a^* \cap L_b^*|$ of the intersection between $L_a^*$ and $L_b^*$ is computed. Finally, the average value of $|L_a^* \cap L_b^*|$ obtained over $B$ bootstrap iterations (in all the experiments presented here, $B = 200$) is returned as an estimate of reproducibility $\hat{\rho}$. We will refer to this estimate as $\hat{\rho}(T, D, s)$.

### 5.3 Data-adaptive test statistics

Consider again the risk-optimal statistic $T^*$ of Equation 5.2:

$$T^* = \arg \min_{T \in \mathcal{T}} R(\omega, \delta_{T,s})$$

So far, we have defined a notion of reproducibility, and have the assurance that asymptotically, reproducibility is negatively correlated with risk, for any pair of test statistics having lower risk than purely random selection. We have also a simple procedure by which reproducibility can be estimated from data. We are therefore in a position to use reproducibility as a criterion for learning a test statistic from data, by optimising over estimated reproducibility in place of risk:

$$\hat{T}^* = \arg \max_{T \in \mathcal{T}} \hat{\rho}(T, D, s) \quad (5.17)$$

We refer to the learned statistic as $\hat{T}^*$ to distinguish it from the truly optimal member $T^*$ of the family $\mathcal{T}$. Note also that since reproducibility is negatively
correlated with expected loss, the minimisation of risk has been replaced by a maximisation of reproducibility.

How should the family $T$ of test statistics be chosen? Clearly, this is an important issue. We would like to give the learning procedure a large set of possible statistics from which to choose, but must be careful to restrict the set sufficiently to satisfy the ‘better than random’ condition of Equation 5.15. We suggest the family of statistics defined as follows:

$$T = \left\{ T : T = \frac{\bar{X}_k - \bar{Y}_k}{\lambda_1 \times \hat{SD}_k + \lambda_2}, \lambda_1 \in \{0, 1\}, \lambda_2 \in [0, 5] \right\}$$

(5.18)

As before, $\bar{X}_k - \bar{Y}_k$ is the difference of sample means between conditions, and $\hat{SD}_k$ the sample pooled standard deviation. The $\lambda$’s are scalar parameters, common to all genes. The family of statistics $T$ is a generalisation of the t-statistic. The t-statistic itself as well as difference of means and SAM are special cases, as shown in Table 5.1.

<table>
<thead>
<tr>
<th>STATISTIC</th>
<th>$\lambda_1$</th>
<th>$\lambda_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>t-statistic</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Difference of means</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>SAM</td>
<td>1</td>
<td>Small positive values</td>
</tr>
</tbody>
</table>

Small values of $\lambda_2$ can play a role in mitigating deleterious effects associated with the standard deviation term $\hat{SD}_k$; equally, the results presented in the previous Chapter suggested that it may at times be useful to ignore standard deviation altogether. Accordingly, we use the parameter $\lambda_1$ to allow the standard deviation
term to be ‘switched off’, and allow $\lambda_2$ to take on small positive values. Allowing $\lambda_2$ to grow very large would also allow the standard deviation term to be ignored, but would leave us with a very large range of values from which to sample $\lambda_2$. The approach taken here is motivated by a desire to use either small values of $\lambda_2$ or to ignore standard deviation altogether. Thus, a reasonable alternative to the two-parameter scheme used here would be to sample $\lambda_2$ from a wider range, but with most of the probability mass concentrated at very small and very large values.

**Algorithm 2** Reproducibility-based optimisation of parameters

Initialise $\hat{\rho}^* \leftarrow 0$, $\lambda_1^* \leftarrow 1$, $\lambda_2^* \leftarrow 0$

1. Sample $\lambda_1$ from $\{0, 1\}$ (with equal probability for 0 and 1); sample $\lambda_2$ from a uniform distribution over the range $[0, 5]$
2. If $\lambda_1 = 0$, set $\lambda_2 \leftarrow 1$
3. Calculate $\hat{\rho} = \hat{\rho}(T_{\lambda_1, \lambda_2}, D, s)$ using Algorithm 1, where $T_{\lambda_1, \lambda_2}$ represents the member of $T$ defined by parameters $\lambda_1$ and $\lambda_2$
4. If $\hat{\rho} > \hat{\rho}^*$, set $\hat{\rho}^* \leftarrow \hat{\rho}$, $\lambda_1^* \leftarrow \lambda_1$, $\lambda_2^* \leftarrow \lambda_2$
5. Repeat (1) to (4) $N$ times
6. Return $\lambda_1^*$, $\lambda_2^*$

We use a simple stochastic optimisation scheme to learn the $\lambda$’s, as shown in Algorithm 2. The parameters are sampled, reproducibility is calculated using the bootstrap procedure of Algorithm 1, and the parameters giving rise to highest reproducibility returned after $N = 500$ iterations. Note that when $\lambda_1 = 0$, $\lambda_2$ is forced to 1. This has no effect on the ranking of genes, since when $\lambda_1 = 0$ the denominator is constant across genes, such that the precise value of $\lambda_2$ has no substantive effect, and the ranking procedure is in any case equivalent to difference of means.
CHAPTER 5. DATA-ADAPTIVE TEST STATISTICS

5.4 Results

In this Section, we present empirical results examining the ability of the proposed data-adaptive method to correctly select differentially expressed genes from microarray data. Some of the data analysed below is not Normally distributed, so in addition to the statistics discussed in the previous Chapter (t-statistic, difference of means and SAM) we also present results for the non-parametric Wilcoxon-Mann-Whitney rank-sum statistic (DeGroot and Schervish, 2002).

5.4.1 Synthetic data

We first present results using synthetic data. These offer the important advantage that the ‘ground-truth’ is known a priori, allowing an accurate assessment of results and an objective comparison of methods.

We generated data using a hierarchical model, following a procedure similar to Lönstedt and Speed (2002). Data in each of the conditions were drawn from Normal densities with prior distributions for the means (a Normal) and variances (a Gamma). Thus, the data for NDEGs is distributed as $X_{ik}, Y_{jk} \sim \mathcal{N}(\mu_{k(0)}, \sigma_{k(0)}^2)$, with $\sigma_{k(0)}^2 \sim \mathcal{G}(a_{(0)}, b_{(0)})$. For DEGs: $X_{ik} \sim \mathcal{N}(\mu_{Xk(1)}, \sigma_{Xk(1)}^2)$ and $Y_{jk} \sim \mathcal{N}(\mu_{Yk(1)}, \sigma_{Yk(1)}^2)$, with $\mu_{Xk(1)}$ fixed and $\mu_{Yk(1)} \sim \mathcal{N}(\mu'_{Yk(1)}, \sigma'_{Yk(1)}^2)$, and $\sigma_{Yk(1)}^2 \sim \mathcal{G}(a_{(1)}, b_{(1)})$. Figure 5.5 shows these distributions as a pair of probabilistic graphical models, whose nodes represent random variables and whose edges represent conditional dependencies.

In all the experiments presented in this Section, the total number of NDEGs
was \( g_0 = 1000 \), the total number of DEGs was \( g_1 = 25 \) and the sample size was 5 in each condition such that \( m, n = 5 \). Where required, the number of genes selected was \( s = 25 \). In all cases the null mean and alternative mean for the \( X \) data were set to zero, i.e. \( \mu_{k(0)} = 0 \) and \( \mu_{Xk(1)} = 0 \), while the hyperparameters for the alternative mean \( \mu_{Yk(1)} \) were \( \mu_{Yk(1)}' = 3 \) and \( \sigma_{Yk(1)}'^2 = 0.5 \). The prior defined by these parameters is shown in Panel (a) of Figure 5.6. Panel (b) of the figure shows two priors over gene-level variances which will be used below, while Panel (c) shows the same priors in terms of standard deviation. The use of these variance priors varies from experiment to experiment and is described in more detail in the context of individual experiments presented below. In general, our aim in choosing parameters for these simulations was to ensure sufficient inter-gene variability in means and variances, while providing an overall level of variation which was such that the simulations were neither so easy that all methods succeeded, nor so
Figure 5.6: Priors for means and variances.

Identical variance priors.

In the first set of experiments, variances for DEGs and NDEGs were drawn from the same prior distribution, with hyperparameters \(a(\cdot) = 1\), \(b(\cdot) = 2\). This distribution is shown in Panels (b) and (c) of Figure 5.6 above as a solid blue curve. All other parameters were as stated above.

A total of 200 datasets were generated under the model, and the sampled data were analysed to obtain 200 sets of gene selection results for each test statistic. These results were used to obtain the plots shown in Figure 5.7; “DOM” refers to difference of means. Risk, shown in Panel (a), was estimated by averaging the FDR loss function over all 200 sets of data. The boxplot of the number of DEGs discovered shown in Panel (b) was obtained from the distribution of outcomes over 200 iterations. The expected ROC curves shown in Panel (c) were computed by averaging sample ROC curves obtained from the 200 sets of results, while the reproducibility scores shown in Panel (d) are the averages of observed bootstrap
reproducibility estimates. Finally, Panel (e) shows a detailed view of the lower left-hand region of the ROC curve.

Risk is highest for the rank-sum and difference of means statistics. The relatively poor performance of difference of means is probably due to the fact that the data are non-i.i.d., such that gene-level estimates of standard deviation are indeed useful. The performance of the rank-sum statistic is also unsurprising, given that the data are Normally distributed. SAM and the data-adaptive method slightly outperform the t-statistic. Figure 5.8 shows the values of parameters $\lambda_1$ and $\lambda_2$ learned from the data, in the form of histograms. Note that since $\lambda_2$ is forced to 1 when $\lambda_1 = 0$, we show only values of $\lambda_2$ for the $\lambda_1 = 1$ case. The data-adaptive approach almost always chooses to make use of the standard deviation term, and also moderates this term by learning small values of $\lambda_2$.

The bootstrap reproducibility scores of the five test statistics, as shown in Panel (d), correctly reflect the corresponding ordering under risk or utility, with the exception of the relative ordering of the difference of means and rank-sum statistics. The reproducibility score for difference of means is much higher than rank-sum, and indeed close to the t-statistic, but as noted above the true utility of difference of means in this experiment is actually similar to rank-sum.

**Differing variances.** We discovered previously that the effect on gene selection procedures of variances differing between DEGs and NDEGs depends on whether DEGs have higher variances than NDEGs or vice versa. Accordingly, we generated data under two distinct models with differing variances.

In the first case, the variance hyperparameters in the data-generating model
Figure 5.7: Identical variance priors. Results on synthetic microarray data, with identical variance priors for DEGs and NDEGs. “DOM” refers to difference of means.
were set so that DEGs had generally higher variances than NDEGs. Specifically, we set the variance hyperparameters as follows: \( a_{(0)} = 1 \), \( b_{(0)} = 2 \), and \( a_{(1)} = 4 \), \( b_{(1)} = 1 \), such that the expected value of the NDEG variance \( \sigma^2_{k(0)} \) was 2, while the expected value of the DEG variance \( \sigma^2_{k(1)} \) was 4. The priors so defined are shown in Panels (b) and (c) of Figure 5.6; the blue curves represent the NDEG case, while the red, dashed curves represent the DEG case. All other parameters values were as stated previously.

Results are shown in Figure 5.9; as before, the plots shown are obtained from an analysis of 200 datasets. As expected from theoretical results, the t-statistic does badly under these conditions, and is outperformed by SAM. Difference of means and the data-adaptive method are the most effective procedures in this experiment, although, as shown by the boxplots in Panel (b), the data-adaptive method’s slightly better best-case performance is accompanied by greater variability than difference of means. The reproducibility scores shown in Panel (d) correctly reflect the true utility ordering of the five statistics in this case, although
Figure 5.9: Synthetic data, differing variances, first case. Here, DEGs have higher variances than NDEGs. “DOM” refers to difference of means.
Figure 5.10: Synthetic data, differing variances, second case. Here, DEGs have lower variances than NDEGs. “DOM” refers to difference of means.
the adaptive method has a slightly higher reproducibility score, which is not really borne out by the other plots.

In the second case, variance hyperparameters were set to give DEGs generally lower variances than NDEGs. Specifically, we swapped the variance hyperparameter settings of the previous experiment, setting \( a_{(0)} = 4, b_{(0)} = 1 \), and \( a_{(1)} = 1, b_{(1)} = 2 \), such that the expected value of \( \sigma^2_{k(0)} \) was 4, while the expected value of \( \sigma^2_{k(1)} \) was 2. Thus, in this experiment, the red, dashed curves in Panels (b) and (c) of Figure 5.6 represent the NDEG case while the blue curves represent the DEG case. All other parameter values were as stated previously.

The results obtained are shown in Figure 5.10. These conditions favour the t-statistic, and indeed it now does well. The performance of SAM and the data-adaptive method are comparable to the t-statistic, while rank-sum does worse than either. As expected from theoretical results, difference of means does very badly, since it is unable to exploit the discriminatory information contained in gene-level standard deviations. The reproducibility scores shown in Panel (d) correctly reflect the true utility ordering of statistics in this case, with the exception, as was the case in the first set of experiments with identical variance priors, of the relative ordering of the difference of means and rank-sum statistics. Rank-sum ought to score higher than difference of means, but in fact the reverse is the case.

Figure 5.11 below shows the values of parameters \( \lambda_1 \) and \( \lambda_2 \) learned from data with differing variance priors. Panels (a) and (b) show values learned in the first case in which \( E[\sigma^2_{k(0)}] < E[\sigma^2_{k(1)}] \), favouring difference of means over the t-statistic, while Panels (c) and (d) show values learned in the second case, in which...
Figure 5.11: Parameters learned from data with differing variance priors. Panels (a) and (b) show values learned in the ‘first case’; Panels (c) and (d) are the values learned in the ‘second case’.

\[ E[\sigma^2_{k(1)}] < E[\sigma^2_{k(0)}], \]

favouring the use of a standard deviation term. We find that in the first case, the data-adaptive approach chooses to ignore the standard deviation term in most cases, and otherwise, when \( \lambda_1 = 1 \), chooses relatively high values of \( \lambda_2 \). In contrast, in the second case, the algorithm always chooses to use the standard deviation term, and also chooses smaller values of the moderating term \( \lambda_2 \).

**Non-normality.** It is reasonable to expect that suitably pre-processed microarray data are at best only approximately Normally distributed. As such, the robust-
ness of gene selection procedures to deviations from Normality may at times be of considerable practical importance. We therefore performed experiments using data generated from a mixture of a Normal model with equal variance priors as described previously, and a uniform distribution, such that the overall density had heavier tails than a Normal.

![Figure 5.12: Non-normal data-generating density.](image)

More specifically, if the mean for a gene $k$ under a particular condition is $\mu_k$, and the corresponding standard deviation is $\sigma_k$, we draw two of five samples uniformly from the interval $[\mu_k - 5\sigma_k, \mu_k + 5\sigma_k]$ and the remaining three samples from a Normal as described previously. The effect of drawing samples in this manner is illustrated in Figure 5.12. The blue curve is a standard Normal pdf, while the red, dashed curve is the pdf obtained by mixing the same Normal with a uniform distribution as described above. The overall effect of the mixing process is a distinct fattening of the tails of the Normal.

Since our aim here is to understand the effect of non-normality itself, we
Figure 5.13: Synthetic data, non-normality. Here, data were generated from a mixture of Normal and uniform distributions. “DOM” refers to difference of means.
endeavoured to retain roughly the same overall level of variance as in previous experiments. Accordingly, we set the variance hyperparameters as follows: $a_{(0)}, a_{(1)} = 0.25$ and $b_{(0)}, b_{(1)} = 4$, giving $E[\sigma^2_{k(0)}], E[\sigma^2_{k(1)}] = 1$ (rather than 2 as previously). When combined with the mixing process described above this resulted in an overall data variance of $\sim 2.46$, which is similar to the other experiments. All other parameters for the hierarchical Normal model were as described previously.

![Histogram of $\lambda_1$](image1.png)  
(a) Histogram of $\lambda_1$  

![Histogram of $\lambda_2$ (when $\lambda_1 \neq 0$)](image2.png)  
(b) Histogram of $\lambda_2$ (when $\lambda_1 \neq 0$)  

Figure 5.14: Parameters learned from non-normal data.

Results obtained are shown in Figure 5.13. Risk is lowest for the t-statistic, SAM and the data-adaptive method, while rank-sum does worse than either. Difference of means is by far the least effective statistic in this case, discovering on average less than a third as many true positives as the t-statistic, SAM and the adaptive method. Figure 5.14 shows the parameter values learned in this case. Reflecting the poor performance of difference of means, the adaptive method makes use of the standard deviation term in almost every case.

The reproducibility scores, shown in Panel (d) of Figure 5.13, mostly reflect
the true utility ordering of the statistics, but again the relative ordering of rank-sum and difference of means is incorrect, with the bias in the same direction as before, in favour of difference of means.

5.4.2 Microarray data

We applied our method to data from a benchmark microarray study conducted by GeneLogic Inc. (Gaithersburg, MD, U.S.A.). Biological samples were obtained from a single acute myeloid leukaemia (AML) cell line, such that it was known \textit{a priori} that there was no differential expression on account of underlying biology. Known concentrations of mRNA corresponding to just 11 of the 12,626 genes under study were then added to the hybridisation mix prior to performing the microarray experiment.

The subset of the GeneLogic dataset used here consists of Affymetrix oligonucleotide arrays in which spike-in\footnote{In microarray studies of this kind, genes to which known concentrations of mRNA are added are said to be ‘spiked-in’. The studies themselves are usually called ‘spike-in studies’.} concentrations are the same for all spiked-in genes on a particular array (for further details see Antonellis \textit{et al.} (2002), in particular Table 1 of that report). Data were preprocessed using the Robust Multichip Average (Bolstad \textit{et al.}, 2003).

We started with a total of twelve arrays, three each with spike-ins at 12.5, 25, 50 and 75 picomolar (pM) concentrations. We treated the 12.5 pM and 25 pM arrays as one group (the $X$ data), and the 50 pM and 75 pM arrays as a second group ($Y$ data). We then resampled the arrays with replacement, at each iteration
Figure 5.15: Results on benchmark microarray data. Data were obtained from a spike-in study conducted by GeneLogic Inc. using Affymetrix oligonucleotide arrays. “DOM” refers to difference of means.
taking three arrays from the first group and three from the second group. This provided us with a series of iterations each of which had six arrays in two classes, with 11 genes known to be differentially expressed between them. Figure 5.15 shows the results obtained; the total number of resampling iterations was 100.

![Histograms of λ_1 and λ_2](image)

(a) Histogram of λ_1  
(b) Histogram of λ_2 (when λ_1 ≠ 0)

Figure 5.16: Parameters learned from benchmark microarray data.

Difference of means and the adaptive approach are very effective in this case, with rank-sum also doing well, but displaying greater variability and slightly higher risk than these two statistics. The t-statistic does very badly indeed. Why does the t-statistic perform so poorly here? Since each class contains two distinct spike-in concentrations, we conjecture that the explanation lies in the resulting bi-modality of the data. The reproducibility scores shown in Panel (d) of Figure 5.15 correctly reflect the true utility ordering of the statistics in this experiment.

Figure 5.16 shows the parameter values learned by the adaptive method. In most cases the algorithm chooses to switch off the standard deviation term, which seems a good choice in light of the poor performance of the t-statistic. Further-
more, when the standard deviation term is activated, learned values of the moder-
ating term $\lambda_2$ are relatively high, suggesting that in these cases the algorithm may
be attempting to combat the bi-modality of the data by downplaying the effect of
standard deviation.

5.5 Discussion

The data-adaptive approach to gene selection proposed in this Chapter can be
thought of as exploiting a non-parametric proxy for risk to learn a moderated
parametric statistic. The notion of reproducibility itself is non-parametric, inso-
far as the relationship between reproducibility and risk does not depend on any
specific data-generating distribution. However, the family of statistics used for
learning was based on the t-statistic, and contained the t-statistic and difference
of means as special cases. As a consequence, faced with data generated under
various hierarchical Normal models, the data-adaptive method was able to match
or outperform the t-statistic, but equally, when confronted with non-Normal or bi-
modal data, it was able to do as well or better than the non-parametric rank-sum
statistic.

We were able to show that for ‘better than random’ test statistics, reproducibil-
ity is negatively correlated with true risk. However, our formal results were ob-
tained under a biologically unrealistic assumption of i.i.d. expression levels for
DEGs and NDEGs. Accordingly, we sought to empirically examine the correla-
tion between risk and reproducibility under non-i.i.d conditions, and found that
the correlation actually holds up well in practice. We found also that in experiments using synthetic data reproducibility-based learning was able to successfully choose from a parameterised family of statistics, learning generally sensible parameter values. Furthermore, in most cases, the rank ordering of various statistics under bootstrap reproducibility broadly corresponded with true utility. A subtle error which occurred in three different experiments concerned the relative reproducibility scores for the difference of means and rank-sum statistics. In each of these three experiments (identical variance priors; differing variances, second case; and non-normality) bootstrap reproducibility for difference of means was unjustifiably high. We conjecture that this effect may be due to the relative simplicity of the difference of means statistic, which may be causing relatively low variability in selection results from bootstrapped data. This is an issue which warrants further investigation. However, the generally good results of the data-adaptive method suggest that this issue may not be a major concern in practice.

A further assumption we made was that the statistics involved in the learning process were better than random. The set of statistics $T$, as defined by Equation 5.18, is sufficiently restricted that every member of the family will outperform random gene selection under any plausible data-generating distribution, but the family as a whole offers considerable flexibility nonetheless.

Theoretical results suggested that reproducibility should be most effective when the overall proportion of differentially expressed genes is relatively small. We confirmed this prediction empirically, by drawing data under a conventional Normal model, and varying the number of NDEGs while holding the number of
Figure 5.17: A large number of non-differentially expressed genes is beneficial to gene selection using reproducibility. The difference $\Delta R_{FDR}$ in risk between the t-statistic and the data-adaptive approach is plotted against the total number $g_0$ of non-differentially expressed genes, with the total number of differentially expressed genes held constant. The data-adaptive approach does better relative to the t-statistic as $g_0$ grows larger.

DEGs constant. Figure 5.17 shows the difference $\Delta R_{FDR}$ in FDR risk between the t-statistic and the data-adaptive approach, plotted against the total number $g_0$ of non-differentially expressed genes. Clearly, the data-adaptive approach does better, relative to the t-statistic, as $g_0$ grows larger. It is worth noting, however, that even when there are only as many NDEGs as DEGs, our method does slightly better than the t-statistic.

In practice, can we be sure that there are indeed only a relatively small number of genes which are truly differentially expressed? In many cases, we can be fairly confident that this is the case, as for example in experiments using large genome-wide arrays, but involving very specific biological perturbations. Equally, in some cases, such as studies using a small, custom array, or involving very drastic perturbations which may affect a large proportion of genes under investigation, it should
be clear that our method may not be applicable. In general, since we need only to verify that a relatively small number of genes are DEGs, a rough check via an examination of standard exploratory plots, such as the ‘MA plot’ (Dudoit et al., 2002; Bolstad et al., 2003), should be sufficient. If there appears to be a great deal of activity, with a large proportion of genes being differentially expressed, our method should not be used.

Finally, we may draw a comparison between the data-adaptive scheme proposed here and empirical Bayes approaches to gene selection (e.g. Lönnstedt and Speed, 2002). Empirical Bayes methods are also data-adaptive, in that they determine the hyperparameters of a Bayesian model from the data. A key difference from our method, however, is that Empirical Bayes methods require the specification of a full Bayesian model. If the true data-generating distribution deviates significantly from the specified model, such methods are likely to perform poorly.
Chapter 6

Conclusions

Biological systems are remarkably complex and are, for the most part, profoundly difficult to observe at the molecular level. Furthermore, such observations as can be made tend to be incomplete and subject to considerable uncertainty. Drawing robust conclusions from quantitative data in molecular biology can therefore be fraught with difficulty. As a consequence, in recent years, statistical data analysis has emerged as a key element of post-genomic biology.

Problems in the nascent field of statistical bioinformatics are technically challenging as well as practically important, and are motivating much exciting research at the intersection of molecular biology on the one hand, and statistics and machine learning on the other. The challenging conditions which characterise genomic data analysis mean that the field presents real opportunities to make methodological advances which are also of immediate practical relevance. The work presented in this thesis has sought to address one particular problem
in genomic data analysis, namely the analysis of differential expression in gene microarray data.

The remainder of this Chapter is organised as follows. We first provide a brief summary of the work presented in the main body of the thesis. We then take a step back from microarray analysis itself, and look more generally at the nature of uncertainty in quantitative molecular biology. Finally, we suggest a few specific directions in which it may prove useful to extend our work.

### 6.1 Summary of research

Microarrays are the most widely-used and mature of the high-throughput technologies which have revolutionised the biological sciences in recent years, and comparative experiments involving contrasting tissue types are arguably their most well-established mode of use. The statistical task of discovering genes which are differentially expressed between conditions, which we have called “gene selection”, is therefore a key task in statistical bioinformatics. However, the characteristics of microarray data mean that it has proven difficult to reliably select truly differentially expressed genes, and a number of important issues have been left unaddressed or only partially resolved in the literature. There has been a lack of a statistically principled framework within which to analyse the ability of gene selection methods to accurately select differentially expressed genes. Furthermore, existing methods for gene selection have mostly been formulated as classical hypothesis tests, and have tended to use a fixed function, or test statistic, to rank and
then select genes.

In this thesis, we sought to place the gene selection task within a decision-theoretic framework. This allowed us to better understand the ability of the conventional approach to gene selection to deal with statistical characteristics relevant to the microarray domain, including small sample sizes, high noise levels, and large numbers of non-differentially expressed genes. The results of our analysis led us to argue that a hypothesis testing approach based on fixed, inflexible test statistics may not be sufficiently robust to produce consistently accurate results under varied conditions. Motivated by these concerns, we put forward a novel, data-adaptive scheme for gene selection, which involved learning effective test statistics directly from data. We presented theoretical and empirical evidence to support our claim that the flexibility offered by the data-adaptive approach leads to good performance under diverse conditions.

We argued that the effectiveness of our method is partly due to the fact that it uses a non-parametric proxy for risk to learn a member of a parametric family of statistics. This allows it to be robust to variations in the underlying model but at the same time generally more effective than the wholly non-parametric rank-sum statistic. More generally, our approach can be thought of as marrying a computationally-intensive, learning-inspired procedure to a conventional hypothesis testing paradigm, in effect bringing together some aspects of machine learning and classical statistics.
6.2 Uncertainty in quantitative molecular biology

A running theme in this thesis has been the need to deal carefully with issues arising as a consequence of the statistical characteristics of microarray data. These characteristics included:

- Uncertainty regarding underlying distributions
- A mismatch between small sample size and high dimensionality, and
- The fact that we do not usually know which genes in a dataset are truly differentially expressed

In many ways, it is the very fact that gene selection is affected by all three of these characteristics at once that makes it a difficult problem. Indeed, it was the confluence of these factors that provided much of the motivation for the work presented here.

Conditions of this kind are by no means restricted to gene selection. Generalising away from microarray analysis itself, we can describe the three conditions mentioned above as model uncertainty, small sample size and unlabelled data. The coming together of these three factors is illustrated by means of a Venn diagram in Figure 6.1. In the microarray domain, and in quantitative molecular biology more generally, it is not unusual to encounter all three conditions together.

Model uncertainty arises as a consequence of the fact that biological systems and processes of interest tend to be a priori quite poorly understood. Furthermore, in many cases, small sample effects exacerbate the effects of model uncertainty, by
making estimation difficult. This is especially true when the number of variables under study is large and underlying covariance structure complex, as is usually the case in system-wide studies involving thousands, or tens of thousands, of genes or gene products.

![Diagram of Model uncertainty, small sample size and unlabelled data.](image)

Figure 6.1: Model uncertainty, small sample size and unlabelled data.

Of course, small sample size is not a universal feature of quantitative data in molecular biology. Some assays, such as multi-parameter flow cytometry (De Rosa et al., 2001), for example, measure a small number of variables at relatively large sample sizes. Unsurprisingly, data of this kind have proved highly amenable to analysis using standard machine learning methods (Sachs et al., 2005). However, truly system-wide assays, such as microarrays and various emerging methods in proteomics (de Hoog and Mann, 2004) do tend to have small sample sizes. These methods tend to be time-consuming and expensive, effectively limiting the number of experiments which can be performed. As these emerging technologies reach maturity, experiments will become easier and cheaper, but the very
fact that these studies tend to involve very large numbers of variables means that, in system-wide biology, sample sizes will continue to be very small relative to dimensionality, possibly for many years to come.

Relatively few problems in molecular biology (or indeed in science more generally) are of a narrowly predictive nature, and truly supervised problems are perhaps less common in scientific than in engineering applications. As a consequence, unlabelled data is arguably the norm in exploratory analyses of biological data. Indeed, even in the case of network inference, there is no meaningful analogue to “training data”: in a system of active research interest, we do not usually know ahead of time which network topology is correct.

Taken together, these factors mean that data analytic problems in many areas of quantitative molecular biology differ in fundamental ways from canonical problems in statistics and machine learning. Standard problems in these latter fields tend to be characterised by one or two of model uncertainty, small sample size and unlabelled data, but very rarely all three at once. This difference motivates a need for novel statistical methods which are well-suited to the needs of post-genomic biology, and capable of producing reliable results under the conditions alluded to above.

The shift towards truly integrative, system-level studies in biology will only accelerate the trends identified here, and increase the need for appropriate statistical methods. Indeed, it is likely that substantive methodological advances in statistics and machine learning will be needed if “systems biology” is to live up to its considerable promise.
6.3 Further work

We bring the thesis to a close with a discussion of some key directions for further research. The work presented here focused on the specific problem of selecting differentially expressed genes from microarray data. However, a number of problems in genomic data analysis can be cast within a broadly similar hypothesis testing framework. Some examples include hypothesis testing problems in microarray time series data (e.g. Tai and Speed, 2004) and high-throughput proteomics (e.g. Gustafsson et al., 2004). The decision-theoretic analysis of gene selection presented in Chapter 4 could be modified very easily to address problems of this kind. The results concerning the relationship between risk and various sampling distributions were quite general in nature, and could be extended to study other hypothesis testing problems in high-throughput molecular biology.

The notion of reproducibility, which underpinned our approach to gene selection, is also fairly general, and the conditions under which a data-adaptive approach may be beneficial are not unique to gene selection. The results regarding the relationship between reproducibility and risk are not tied to the gene selection problem, and should remain relevant wherever the combinatorial argument presented in Chapter 5 holds. It should therefore prove relatively straightforward to extend our data-adaptive approach to other hypothesis testing problems. In each case, a suitable family of statistics from which to learn would have to be defined, perhaps by generalising an existing parametric statistic, in analogy to the generalisation of the t-statistic used here.
A second important direction for further work in data-adaptive test statistics is in the estimation of P-values. A test statistic learned under the scheme proposed here will not, in general, have a known null sampling distribution from which to compute P-values. However, there is now a rich literature on the estimation of null distributions by computationally intensive means (e.g. Good, 2000; Pollard and van der Laan, 2003). It should be possible to take advantage of such methods to explicitly control error rates under a test statistic learned using the data-adaptive approach.

Finally, a more speculative possibility would be to extend our work to problems in unsupervised learning. As noted above, the argument which connects reproducibility to risk under the true data-generating distribution is not specific to gene selection, but it is not even really specific to hypothesis testing. Consider a clustering problem in which objects are known to belong to one of two clusters. If we think of hypotheses as roughly analogous to clusters, and genes as analogous to objects to be clustered, it is easy to see that the gene selection problem with which we have been concerned bears a resemblance to binary clustering. This analogy suggests the interesting possibility of extending the notion of reproducibility-based learning to a more general setting. It is possible that reproducibility could be used as a proxy for a suitably-defined notion of risk in clustering, perhaps allowing the learning of distance metrics which minimise the chance of incorrect cluster assignments.


C. Cortes and M. Mohri. AUC Optimization vs. Error Rate Minimization. In


Appendix A

Hypothesis testing

A collection of random variables $X = X_1, X_2 \ldots X_n$ with $X_i \sim p$ is called a random sample. The true data-generating distribution $p$ is a statistical representation of the population from which the sample is drawn; its parameters are known as population parameters.

A hypothesis is a statement regarding a population parameter. For example, if $X_1, X_2 \ldots X_n$ is a random sample from a population with mean $\mu$, the statement “$\mu = 0$” is a hypothesis regarding the parameter $\mu$. More generally, if $\Theta$ denotes the complete set of possible values a parameter $\theta$ may take, a hypothesis regarding $\theta$ takes the form $\theta \in A$, with $A \subseteq \Theta$. The set $\Theta$ is known as the parameter space.

A hypothesis test is a decision procedure whose aim is to determine, on the basis of a random sample, which of two competing hypotheses should be accepted.

1In practice, population parameters are usually unknown. They may be estimated from a random sample, but these estimates are conceptually distinct from the true values. For example, if we draw samples $X_1, X_2 \ldots X_n$ from a Normal distribution with mean $\mu$, the parameter $\mu$ is a population parameter, whereas the sample mean $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ is an estimate of $\mu$. 

177
The two competing hypotheses are called the *null hypothesis* and *alternative hypothesis*, and are denoted by $H_0$ and $H_1$ respectively. The ‘default’ hypothesis, signifying the absence of any interesting effect, is usually called the null hypothesis. The null and alternative hypotheses are complementary: if the null hypothesis is $H_0 : \theta \in \Theta_0$, with $\Theta_0 \subseteq \Theta$, the alternative hypothesis is $H_1 : \theta \in \Theta_0^c$, where the set $\Theta_0^c$ is the complement of $\Theta_0$, such that $\Theta_0 \cup \Theta_0^c = \Theta$ and $\Theta_0 \cap \Theta_0^c = \emptyset$. We will subsequently refer to $\Theta_0^c$ as $\Theta_1$.

The sample space $S$ in a hypothesis test is the space of all possible random samples $X$, while the action space $A$ is the set $A = \{a_0, a_1\}$, where $a_0$ represents the action of accepting the null hypothesis $H_0$, and $a_1$ that of accepting the alternative $H_1$. Since the two hypotheses are complementary, a test procedure must accept exactly one of them; the hypothesis which is not accepted is said to be rejected by the test.

A test statistic $T = T(X)$ is a real-valued function of a random sample which is used to perform a hypothesis test. The function $T$ summarises the information contained in the sample, and is used to make a decision regarding which of the two hypotheses to accept. Test statistics are often designed to have the property that larger values indicate greater evidence against the null hypothesis. This property is assumed for the remainder of this Section.

A decision rule must define a mapping from samples to actions; however, a test statistic $T$ in itself defines only a mapping from the sample space to the real line. In order to specify a decision rule using $T$, we need also to define an appropriate mapping from the real line to the action space. In other words, we must specify
which values of $T$ should lead us to decide upon actions $a_0$ and $a_1$, i.e. to accept or reject the null hypothesis $H_0$. Given our assumption that larger values of $T$ indicate greater evidence against $H_0$, we may choose a single value $\tau$, such that if $T$ exceeds $\tau$ we decide upon action $a_1$ and accept $H_1$, and otherwise decide upon $a_0$, accepting $H_0$. Then, the overall decision rule $\delta$ is simply:

$$
\delta(X) = \begin{cases} 
a_1 & \text{if } T(X) > \tau 
a_0 & \text{otherwise}
\end{cases}
$$ (A.1)

The value $\tau$ is called the critical value or threshold of the test.

It is always the case that one of the competing hypotheses is actually true and the other false, since the parameter of interest $\theta$ must lie in exactly one of the sets $\Theta_0$ and $\Theta_1$. Two types of errors may therefore occur in a hypothesis test: the alternative hypothesis may be accepted when in fact the null is true, or the null may be accepted when the alternative is true. The first kind of error, is called a Type I error or false positive and the second kind is called a Type II error or false negative. Table A.1 summarises these two kinds of error.

The critical value of the test procedure affects the chance of committing both Type I and Type II errors. If the critical value is set very high, there will be only a small chance of mistakenly rejecting the null hypothesis and committing a Type I error, yet there will be a good chance of mistakenly accepting the null and com-

---

2Since the alternative hypothesis is usually an ‘interesting’ possibility, its acceptance is regarded as a ‘positive’, while the acceptance of the default null hypothesis as a ‘negative’. This is why the erroneous acceptance of the alternative hypothesis is called a false positive, while the erroneous acceptance of the null is called a false negative.
mitting a Type II error. Conversely, if the critical value is very low, it becomes more likely that the null will be mistakenly rejected, but unlikely that it will be mistakenly accepted. The choice of critical value thus involves an inherent trade-off between the two types of error. The conventional procedure for dealing with this trade-off is to first specify an acceptable level of Type I error and then determine a critical value corresponding to the specified error rate. The pre-determined acceptable level of Type I error is called the *significance level* of the test.

Table A.1: Two types of errors in hypothesis testing, after Casella and Berger (2002).

<table>
<thead>
<tr>
<th>TRUTH</th>
<th>Accept $H_0$</th>
<th>Reject $H_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$</td>
<td>Correct decision</td>
<td>Type I error or false positive</td>
</tr>
<tr>
<td>$H_1$</td>
<td>Type II error or false negative</td>
<td>Correct decision</td>
</tr>
</tbody>
</table>

However, accepting or rejecting the null hypothesis at a pre-determined significance level may not always be the best way to report the result of a hypothesis test. Suppose a hypothesis test rejects the alternative hypothesis at significance level $\alpha = 0.01$. It may be the case that the test would have accepted $H_1$ at $\alpha = 0.012$, or that it would only have accepted $H_1$ at a significance level as high as $\alpha = 0.05$. These possibilities correspond to quite different levels of evidence in favour of $H_1$, yet merely reporting the rejection of $H_1$ at $\alpha = 0.01$ does not in any way distinguish between them. An alternative approach is to map the observed value $t$ of the test statistic to the probability $P(T \geq t \mid \theta \in \Theta_0)$ of the statistic taking a value at least as large as $t$, under the null hypothesis. This probability is called the *P-value*. The P-value is equivalent to the probability of committing a Type I
error if the critical value is set to exactly $t$. P-values are widely used in reporting the results of hypothesis tests, although conventional significance levels like 0.01 and 0.05 are still regarded as important benchmarks in many areas of science and medicine.
Appendix B

Distribution of threshold given number of genes selected

The general relationship between threshold $\tau$ and number of genes selected $s$ is probabilistic. Applying Bayes theorem:

$$p(\tau \mid s) = \frac{P(s \mid \tau)p(\tau)}{P(s)} \quad \text{(B.1)}$$

The likelihood term $P(s \mid \tau)$ represents the distribution over the number of genes selected at a given threshold $\tau$. The number of genes selected is simply the sum of the number of NDEGs and DEGs selected ($S_0$ and $S_1$ respectively) at
threshold $\tau$. Given $\tau$, $S_1$ and $S_0$ are Binomially distributed:

\begin{align*}
s \mid \tau &= S_1 \mid \tau + S_0 \mid \tau \quad \text{(B.2)} \\
S_1 \mid \tau &\sim \mathcal{B}(\beta(\tau), g_1) \quad \text{(B.3)} \\
S_0 \mid \tau &\sim \mathcal{B}(\alpha(\tau), g_0) \quad \text{(B.4)}
\end{align*}

where, as in the main text, $\beta(\tau)$ and $\alpha(\tau)$ are the true and false positive rates at threshold $\tau$, and $g_1$ and $g_0$ the total number of DEGs and NDEGs respectively. Twice making use of the Normal approximation to the Binomial and adding the two Normals so obtained, we approximate $P(s \mid \tau)$ as follows:

\begin{align*}
P(s \mid \tau) &\approx \mathcal{N}(m(\tau), v(\tau)) \quad \text{(B.5)} \\
m(\tau) &= g_1\beta(\tau) + g_0\alpha(\tau) \quad \text{(B.6)} \\
v(\tau) &= g_1\beta(\tau)(1 - \beta(\tau)) + g_0\alpha(\tau)(1 - \alpha(\tau)) \quad \text{(B.7)}
\end{align*}

Further assuming uniform priors for $\tau$ and $s$, we get the unnormalised density function for $p(\tau \mid s)$ which was shown in Equation 4.36 in the main text:

\begin{align*}
p(\tau \mid s) &\propto \frac{1}{\sqrt{2\pi \ v(\tau)}} e^{-[s - m(\tau)]^2/[2v(\tau)]} \quad \text{(B.8)}
\end{align*}

Principally due to the properties of the Normal approximation to the Binomial, the approximation is very accurate for the moderate to large values of $g_0$, $g_1$ and $s$ encountered in gene selection.

183
Appendix C

Reproducibility

Consider two sets of gene expression data $D_a$ and $D_b$ drawn independently under the same data-generating distribution. Each of the two datasets contains expression levels for the same set of genes. A test statistic $T$ produces two potentially distinct rankings of these genes from the two datasets. Let $L_a$ and $L_b$ represent sets containing the indices of the $s$ highest-ranked genes obtained from datasets $D_a$ and $D_b$ respectively. Reproducibility $\varrho$ is then defined as the number of genes in common between the sets $L_a$ and $L_b$:

$$\varrho = |L_a \cap L_b|$$  \hspace{1cm} (C.1)

where $|\cdot|$ denotes cardinality.

Let reproducibility $\varrho$ comprise $\varrho_1$ DEGs and $\varrho_0$ NDEGs, such that $\varrho = \varrho_1 + \varrho_0$. Also, let $S_{1(a)}$ and $S_{1(b)}$ represent the number of true positives in each of $L_a$ and

184
How are $\varrho_1$ and $\varrho_0$ distributed, given the numbers $S_{1(a)}$ and $S_{1(b)}$ of true positives in the two sets? Consider taking one DEG at a time from set $L_b$ and trying to find it in set $L_a$. This is in effect a series of Bernoulli trials, with the number of ‘successes’ being the number $\varrho_1$ of DEGs in common between the two sets. Assuming all DEGs are i.i.d., and all NDEGs are also i.i.d., every DEG has the same chance of appearing in the result-set, so the distribution over $\varrho_1$ is Binomial:

$$\varrho_1 \sim \mathcal{B}(S_{1(a)}/g_1, S_{1(b)}) \quad \text{(C.2)}$$

where, as in the main text, $g_1$ is the total number of DEGs. Note that we have assumed (without loss of generality) that $S_{1(a)} > S_{1(b)}$.

We can make a similar argument for the NDEGs, such that the distribution over $\varrho_0$ is also Binomial:

$$\varrho_0 \sim \mathcal{B} \left( \frac{s - S_{1(b)}}{g_0}, s - S_{1(a)} \right) \quad \text{(C.3)}$$

where, as in the main text, $g_0$ is the total number of NDEGs.
Taking the expectation of $q$:

\[
E[q] = E[q_1] + E[q_0] \quad \text{(C.4)}
\]

\[
= E\left[\frac{S_{1(a)}S_{1(b)}}{g_1}\right] + E\left[\frac{(s - S_{1(b)})(s - S_{1(a)})}{g_0}\right] \quad \text{(C.5)}
\]

\[
= \frac{E[S_{1(a)}S_{1(b)}]}{g_1} + \frac{s^2}{g_0} + \frac{E[S_{1(a)}S_{1(b)}]}{g_0} - \frac{s}{g_0}E[S_{1(a)}] - \frac{s}{g_0}E[S_{1(b)}] \quad \text{(C.6)}
\]

\[
= \frac{E[S_{1(a)}S_{1(b)}]}{g_1} + \frac{s^2}{g_0} + \frac{E[S_{1(a)}S_{1(b)}]}{g_0} - \frac{s}{g_0}E[S_{1(a)}] - \frac{s}{g_0}E[S_{1(b)}] \quad \text{(C.7)}
\]

The random variables $S_{1(a)}$ and $S_{1(b)}$ are numbers of true positives and are both distributed as follows (Equation 4.42 in the main text):

\[
S_{1(a)}, S_{1(b)} \sim B\left(\beta(s) \times \max\left(\frac{g_1}{s}, 1\right), \min(s, g_1)\right) \quad \text{(C.8)}
\]

where, $\beta(s)$ is the true positive rate when $s$ genes are selected.

The expected values of $S_{1(a)}$ and $S_{1(b)}$ are therefore both equal to $\beta(s)g_1$. It was shown in Equation 4.43 in the main text that true positive rate $\beta(s)$ and risk $R$ under a FDR loss function are related by the expression $\beta(s) = (1 - R)s/g_1$. We may therefore express $E[S_{1(a)}]$ and $E[S_{1(b)}]$ in terms of risk $R$ and the number $s$ of genes selected:

\[
E[S_{1(a)}] = (1 - R)s \quad \text{(C.9)}
\]

\[
E[S_{1(b)}] = (1 - R)s \quad \text{(C.10)}
\]
Since $S_{1(a)}$ and $S_{1(b)}$ are also independent, we may write:

\[
E[S_{1(a)}S_{1(b)}] = E[S_{1(a)}]E[S_{1(b)}] = (1 - R)^2 s^2 \quad (C.11)
\]

Substituting Equations C.9, C.10 and C.12 in Equation C.7, we get the following expression, as shown in Equation 5.12 in the main text:

\[
E[\varrho] = \frac{(1 - R)^2 s^2}{g_1} + \frac{s^2}{g_0} + \frac{(1 - R)^2 s^2}{g_0} - 2 \frac{(1 - R)s^2}{g_0} \quad (C.13)
\]

Now, suppose two test statistics $T_1$ and $T_2$ have corresponding risk $R_1$ and $R_2$. Suppose also that statistic $T_1$ has lower risk than $T_2$. Then, the difference $\Delta R$ between $R_1$ and $R_2$ ($= R_1 - R_2$) is negative. From Equation C.7, the corresponding difference $\Delta E[\varrho] = E[\varrho_1] - E[\varrho_2]$ in expected reproducibility is given by:

\[
\Delta E[\varrho] = E[\varrho_1] - E[\varrho_2] = \frac{(1 - R_1)^2 s^2}{g_1} + \frac{s^2}{g_0} + \frac{(1 - R_1)^2 s^2}{g_0} - 2 \frac{(1 - R_1)s^2}{g_0} - \frac{(1 - R_2)^2 s^2}{g_1} - \frac{s^2}{g_0} - \frac{(1 - R_2)^2 s^2}{g_0} + 2 \frac{(1 - R_2)s^2}{g_0} \quad (C.14)
\]
Cancelling $\frac{s^2}{g_0}$ and collecting terms:

$$\Delta E[\rho] = \frac{s^2}{g_1} [(1 - R_1)^2 - (1 - R_2)^2]$$
$$+ \frac{s^2}{g_0} [(1 - R_1)^2 - (1 - R_2)^2] - 2 \frac{s^2}{g_0} [(1 - R_1) - (1 - R_2)]$$
$$= s^2 [(1 - R_1) - (1 - R_2)]$$
$$\times \left[ \frac{(1 - R_1) + (1 - R_2)}{g_1} + \frac{(1 - R_1) + (1 - R_2)}{g_0} - \frac{2}{g_0} \right]$$
$$= s^2 (-\Delta R) \left[ (2 - (R_1 + R_2)) \left( \frac{1}{g_1} + \frac{1}{g_0} \right) - \frac{2}{g_0} \right]$$

(C.16)

(C.17)

(C.18)

The latter expression was presented in the main text as Equation 5.13. Examining this expression we can see that $E[\rho]$ and $R$ are negatively correlated if and only if the term within square brackets is positive. Thus, a necessary and sufficient condition for negative correlation is:

$$(2 - (R_1 + R_2)) \left( \frac{1}{g_1} + \frac{1}{g_0} \right) - \frac{2}{g_0} > 0$$

(C.19)

$$\Rightarrow (2 - (R_1 + R_2)) \left( \frac{g_0 + g_1}{g_1 g_0} \right) > \frac{2}{g_0}$$

(C.20)

Cancelling $g_0$ and rearranging, we get the condition shown in the main text:

$$\frac{2 - (R_1 + R_2)}{2} > \frac{g_1}{g_0 + g_1}$$

(C.21)

$$\Rightarrow \frac{R_1 + R_2}{2} < 1 - \frac{g_1}{g_0 + g_1}$$

(C.22)

$$\Rightarrow \frac{R_1 + R_2}{2} < \frac{g_0}{g_0 + g_1}$$

(C.23)
We may also rewrite Equation C.18 as follows, to get the expression given in Equation 5.16 in the main text:

\[
\Delta E_{[2]} = s^2 (-\Delta R) \left[ (2 - (R_1 + R_2)) \left( \frac{1}{g_1} + \frac{1}{g_0} \right) - 2 \right] \quad (C.24)
\]

\[
= -s^2 \Delta R \left[ \frac{2 - R_1 - R_2}{g_1} + \frac{2 - R_1 - R_2}{g_0} \right] \quad (C.25)
\]

\[
= -s^2 \Delta R \left[ \frac{2 - R_1 - R_2}{g_1} - \frac{R_1 + R_2}{g_0} \right] \quad (C.26)
\]