Statistical analysis of genomic and proteomic data

Rebecca Elizabeth Walls

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

The University of Leeds
Department of Statistics

October 2007

The candidate confirms that the work submitted is her own and that appropriate credit has been given where reference has been made to the work of others. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.
Abstract

The research described in this thesis can be broadly described by the term “statistical bioinformatics”. The new science of bioinformatics attempts to amalgamate statistical methodology with biological understanding and bring statistical thinking to areas which have previously been void of such. The research contained in this thesis has been loosely split into two parts: firstly, in the major part of this thesis, we consider the development and application of statistical tools for the analysis of genomic data; in part two, we apply our attention to proteomic data. The study of statistical bioinformatics often involves extracting meaning from noisy, high-dimensional datasets, generated from high-throughput micro-technologies; our work fits fully into this framework and in particular, we consider data from microarray and high-performance liquid chromatography experimentation. One should note that the data generated from these different processes are linked through similarity in structure and many analysis techniques have proved fruitful for both types of data.

In Part I, we strive to develop new techniques for the purpose of analysing microarray data. A typical microarray experiment attempts to ascertain which genes display significant differential expression between various samples, often comparing the levels of expression for many thousands of genes simultaneously. The issues of multiple testing associated with such high-dimensional datasets, often combined with insufficient replication, makes the conventional method of significance testing infeasible. The main focus of this work is the exploration of a variety of empirical Bayesian models, which possess highly effective thresholding properties and can be used for the detection of a few pieces of useful information “hidden” in a sequence of noisy data, most of which contains no useful information. We consider the application of these techniques to microarray data, where the problem is to detect the few “interesting” genes amongst many “uninteresting” ones.
In Part II, we turn our attention to data obtained from high-performance liquid chromatography (HPLC) experimentation. This type of experiment has many applications, including separation, identification, purification, and quantification of various compounds dissolved in a mixture. The process yields functional trace data, where the features of interest are “burstlike” peaks, or more specifically, the areas under these peaks. The area under a peak at a particular time point corresponds to the concentration of a particular compound or protein. Currently, the analysis of such data is performed through a combination of manual input and tools from commercial software. One aim of our work is to develop a more automatic approach, which should reduce the amount of time a user has to spend analysing this type of data, as well as achieving objective and accurate results. The extraction and quantification of the peak areas is generally considered to be the main area of biological interest. However, it is noted that some pre-processing of the data is often necessary prior to this stage. Common problems to be dealt with include removing the inherent noise from the data and correcting for drift in the baseline; we provide further discussion of such issues in Part II. We use a vaguelette-wavelet method to obtain the derivative function of the trace and incorporate this idea as we consider new approaches for the robust identification of the start- and endpoints of peaks, which at present is subjective and at the discretion of the experimenter. Further, we consider the use of bootstrapping as a method to obtain confidence intervals for the peak area estimates.
for my mum and dad
Acknowledgements

Firstly, I would like to note my appreciation to Ferenc Marincs, Clare Donnellan, Kay White and Alistair Hay for donating their data and always finding time to discuss all things biological.

I would like to thank my fellow “PhD-ers” and everyone in the Department of Statistics and the Biostatistics Unit, who have all in their unique way contributed to a wonderful experience to be remembered forever.

This work has been funded by the Engineering and Physical Sciences Research Council, for which I am extremely grateful.

This research would not have been possible without the encouragement and guidance of my supervisors: John Kent, whose depth of statistical knowledge is immeasurable, Mark Gilthorpe, who has made me challenge conventional statistical thinking and whose confidence in my abilities has inspired me to push myself beyond my preconceived limits; and Stuart Barber, who has encouraged me to pursue every opportunity, whose door has always been open for sterling advice and without whom I would probably have never started a PhD...

Lastly, but by far from least, I would like to thank my loving family, whose interest in the incomprehensible has never waned and whose support has been unending.
# Contents

Abstract .............................................................. ii  
Dedication ............................................................ v  
Acknowledgements ................................................... vii  
Contents ............................................................... ix  
List of figures ......................................................... xv  
List of tables ......................................................... xvii  

## I Microarray data

### 1 Genomics

1.1 Introduction ....................................................... 3  
1.2 Background biology ............................................. 4  
  1.2.1 Deoxyribonucleic acid .................................. 4  
  1.2.2 Gene expression and hybridisation ..................... 5  
  1.2.3 Genome variation ........................................ 7  
1.3 Microarrays ...................................................... 8  
1.4 The impact of microarray experimentation ................. 10  
1.5 Description of datasets ....................................... 14  
  1.5.1 E.Coli data ............................................... 14  
  1.5.2 HIV data ................................................ 15  
  1.5.3 Lymphoma data ........................................... 15
## CONTENTS

1.5.4 Control-control data ................................................. 16

2 **Analysing microarray data** 17

2.1 Experimental design and replication ................................. 17

2.2 Pre-processing .......................................................... 20

  2.2.1 Background adjustment ........................................... 20

  2.2.2 Transformation ..................................................... 21

  2.2.3 Normalisation ...................................................... 22

  2.2.4 Example ............................................................ 30

2.3 Post-processing analysis .............................................. 37

  2.3.1 Notation ............................................................ 37

  2.3.2 Conventional methods of analysis for identifying differential expression .................................................. 38

  2.3.3 Mixture models .................................................... 46

3 **Empirical Bayesian thresholding for microarray data** 49

3.1 Introduction ............................................................ 49

3.2 Empirical Bayesian thresholding methodology ..................... 50

  3.2.1 The single observation case ...................................... 51

  3.2.2 Parameter estimation .............................................. 54

3.3 Applying the empirical Bayesian method to microarray data ...... 57

3.4 Linear additive models for the estimation of $\sigma_\epsilon^2$ ............. 58

  3.4.1 Some simulation studies ......................................... 60

  3.4.2 Marginal maximum likelihood estimation for $\sigma_\epsilon$ .......... 65

3.5 Examples ............................................................... 66

  3.5.1 HIV data ............................................................ 67

  3.5.2 E.Coli data .......................................................... 70

  3.5.3 Lymphoma data .................................................... 70

3.6 Discussion .............................................................. 73
## 4 Alternative posterior median estimators for sparse sequences

4.1 Introduction ......................................................... 77

4.2 General framework ................................................. 78
   4.2.1 The single observation case .............................. 78
   4.2.2 Parameter estimation ...................................... 80

4.3 Specific cases ...................................................... 81
   4.3.1 Gaussian-Laplace model .................................. 81
   4.3.2 Scaled $t_1$-$t_1$ model ................................... 84
   4.3.3 Laplace-Laplace model .................................... 85

4.4 The Laplace-asymmetric Laplace model ......................... 87
   4.4.1 Motivation .................................................. 87
   4.4.2 The asymmetric Laplace distribution .................... 87
   4.4.3 Methodology for Laplace-asymmetric Laplace model ...... 91

4.5 Incorporating the false discovery rate ......................... 100

4.6 Examples .......................................................... 102
   4.6.1 HIV data ................................................... 102
   4.6.2 E.Coli data ................................................ 108
   4.6.3 Lymphoma data ............................................. 109

4.7 Simulations ........................................................ 112
   4.7.1 Simulation 1: Efficiency of Laplace-Laplace model .... 112
   4.7.2 Simulation 2: Laplace-asymmetric Laplace model in the presence of asymmetric Laplace noise ............................. 115
   4.7.3 Simulation 3: Laplace-asymmetric Laplace model in the presence of normally distributed noise .......................... 118
   4.7.4 Simulation 4: Laplace-Gaussian model in the presence of asymmetric Laplace noise ................................. 120
   4.7.5 Simulation 5: Laplace-Gaussian model in the presence of normal noise .................................................. 123
CONTENTS

4.7.6 Discussion ................................................................. 126

5 Conclusions and directions for future work .......................... 129
  5.1 Conclusions .............................................................. 129
  5.2 Directions for future work ........................................... 133
  5.3 Contributions ......................................................... 135

II HPLC data ........................................................................ 137

6 Proteomics ........................................................................ 139
  6.1 Introduction ............................................................... 139
  6.2 Mass spectrometry ..................................................... 140
  6.3 High performance liquid chromatography (HPLC) .......... 142
  6.4 Description of data ..................................................... 144
  6.5 Common sources of complications ............................... 145

7 Wavelet methodology ......................................................... 149
  7.1 Wavelets and the discrete wavelet transform ................. 149
    7.1.1 Overview of wavelet analysis ................................. 149
    7.1.2 The discrete wavelet representation ....................... 151
    7.1.3 Scaling functions and constructing the mother wavelet .. 153
    7.1.4 Haar’s Wavelets ................................................. 156
    7.1.5 Mallat’s cascade algorithm .................................. 160
    7.1.6 Padding ............................................................. 163
  7.2 Removing noise from a signal ...................................... 164
    7.2.1 Thresholding rules .............................................. 166
    7.2.2 Selecting a threshold .......................................... 167
    7.2.3 Correlated errors ............................................... 173
  7.3 The non-decimated wavelet transform ......................... 173
# CONTENTS

7.4 The vaguelette-wavelet decomposition ........................................ 175
  7.4.1 The wavelet-vaguelette transform ..................................... 176
  7.4.2 The vaguelette-wavelet transform ..................................... 178
  7.4.3 Thresholding the vaguelette-wavelet transform ....................... 179

8 Existing methods for processing spectral data .......................... 181
  8.1 Pre-processing of spectral data ......................................... 182
    8.1.1 Denoising ....................................................... 183
    8.1.2 Baseline correction ........................................... 184
    8.1.3 Normalisation and transformation ................................. 187
    8.1.4 Peak detection and deconvolution ................................ 189
    8.1.5 Alignment ....................................................... 194
    8.1.6 Peak quantification ........................................... 196
  8.2 Post-processing analysis .................................................. 197

9 Novel approaches to analysing HPLC data ............................... 199
  9.1 Denoising ............................................................ 199
  9.2 Baseline estimation .................................................... 204
  9.3 Finding the derivative .................................................. 209
    9.3.1 Constructing the differentiation matrix ............................ 209
    9.3.2 Application to HPLC data ....................................... 211
  9.4 Bootstrap confidence intervals for peak areas ......................... 216
    9.4.1 The general approach .......................................... 217
    9.4.2 Bootstrapping in regression problems ............................. 218
    9.4.3 Bootstrap confidence intervals ................................... 219
    9.4.4 Application to HPLC data ....................................... 221
    9.4.5 Simulations ....................................................... 228

10 Conclusions and further work .............................................. 235
CONTENTS

10.1 Conclusions ................................................. 235
10.2 Directions for further work ................................. 237
10.3 Contributions ................................................ 238

Appendices .................................... 240

A Algebra for Laplace-Gaussian model with a variance parameter .......... 241
A.1 Convolution of normal distribution with a variance parameter and
the Laplace distribution ........................................ 241
A.2 Expression for $\beta(z)$ ........................................ 244
A.3 Posterior probability that the parameter is non-zero ................. 245
A.4 Derivation of the posterior probability distribution of $\mu$ given $Z$ . 246
A.5 Determination of the posterior median .......................... 248

B Algebra for Laplace-asymmetric Laplace model ....................... 250
B.1 Finding the convolution of a Laplace and asymmetric Laplace
distribution ....................................................... 250
B.2 Deriving $\tilde{F}_1(\mu|x)$ ......................................... 251
B.3 Identifying parameter constraints ................................ 253
B.4 Finding the posterior median function .......................... 255
B.5 Non-shrinkage ................................................. 256

C List of published papers ........................................... 257

Bibliography ................................................. 258
# List of figures

1.1 Microarray image showing differentially expressed genes .......................... 11

2.1 Density plots to illustrate the difference between within-array and 
    between-array normalisation for the HIV data ........................................ 24
2.2 Exploratory plots for raw E.Coli data .................................................. 31
2.3 Plots of E.Coli data after a $\log_2$ transformation ................................. 33
2.4 Plots of E.Coli data after a negative square root transformation ................. 35
2.5 MVA plots of lowess-normalised E.Coli data following different 
    transformations ....................................................................................... 36
2.6 Plots of E.Coli data after a vsn normalisation ......................................... 36

3.1 Posterior median function for different mixing weights ......................... 55
3.2 Simulation to compare the estimation power of different variance estimators 66
3.3 Error distribution for HIV data ............................................................... 68
3.4 HIV data before and after thresholding ................................................... 69
3.5 E.Coli data before and after thresholding ............................................... 71
3.6 Lymphoma data before and after thresholding ....................................... 72

4.1 Posterior median functions for Gaussian-Laplace model ....................... 83
4.2 E.Coli data before and after thresholding ............................................... 84
4.3 Fitted densities for control-control and HIV data .................................. 89
4.4 Posterior median functions for the Laplace-ASL model ......................... 96
4.5 Posterior probability of differential expression across a range of observed values .......................................................... 98
4.6 HIV data before and after thresholding using the Laplace-Laplace model ...................................................... 102
4.7 HIV data before and after thresholding using the Laplace-ASL model .......................................................... 107
4.8 E.Coli data before and after thresholding using the Laplace-ASL model ......................................................... 109
6.1 An example of a raw HPLC trace ..................................................................................................................... 143
7.1 The Haar wavelet function $\psi_{j,k}$ for different values of $j$ and $k$ ................................................................. 158
9.1 Multiple spectra and the average HPLC spectra ................................................................................................. 200
9.2 The decimated transform versus non-decimated transform ........................................................................... 202
9.3 Autocorrelation and partial autocorrelation function of residuals ............................................................. 203
9.4 Thresholding in the presence of correlated errors ............................................................................................. 205
9.5 Denoised trace with various estimated baselines and a baseline-corrected trace ............................................. 208
9.6 Comparison of derivative techniques .............................................................................................................. 213
9.7 Comparison of derivative techniques .............................................................................................................. 214
9.8 Comparison of derivative techniques .............................................................................................................. 215
9.9 Some examples of the bootstrapped traces. ....................................................................................................... 223
9.10 The modified bumps function. ........................................................................................................................ 230
9.11 Signal corrupted by increased noise after denoising procedure. ................................................................... 234
List of tables

2.1 Table of $p$-values, point estimates and confidence intervals before any adjustment for multiple testing. ........................................ 44
2.2 Table of $p$-values, point estimates and confidence intervals after adjusting for the false discovery rate. ................................. 46
3.1 Results from a simulation study to test the efficiency of the empirical Bayesian thresholding procedure in the presence of correlated observations. ........................................ 62
3.2 Results from a simulation study to test the efficiency of the empirical Bayesian thresholding procedure in the presence of correlated observations. 63
3.3 Results from a simulation study to test the efficiency of the empirical Bayesian thresholding procedure in the presence of correlated observations. ........................................ 64
3.4 Comparison of $\hat{\sigma}$, values obtained by various estimation methods . . . . 67
3.5 Summary of results ................................................... 74
4.1 Parameter estimates and Akaike’s Information Criterion for control-control and HIV data ........................................... 90
4.2 Laplace-Laplace model: Genes identified as differentially expressed using a cut-off $t_0 = 0.01$ and their corresponding $1 - \omega_{\text{post}}$ values. . . . . . . 104
4.3 Laplace-Laplace model: Two further genes identified as differentially expressed when using a cut-off $t_0 = 0.1$ and their corresponding $1 - \omega_{\text{post}}$ values. ................................................................. 105
4.4 Laplace-Laplace model: Alternative choices for the cut-off $t_0$, the number of differentially expressed genes found using this threshold and the corresponding implied FDR. ......................................................... 105
4.5 Laplace-asymmetric Laplace model: Genes identified as differentially expressed using a cut-off $t_0 = 0.01$ and their associated $1 - \omega_{\text{post}}$ values. 106
4.6 Laplace-asymmetric Laplace model: Four further genes identified as differentially expressed when using a cut-off $t_0 = 0.1$ and their corresponding $1 - \omega_{\text{post}}$ values. ................................. 107
4.7 Laplace-asymmetric Laplace model: Alternative choices for the cut-off $t_0$, the number of differentially expressed genes found using this $t_0$ and the corresponding implied FDR. ........................................ 108
4.8 Laplace-Laplace model: Genes identified as differentially expressed using a cut-off $t_0 = 0.01$ and their corresponding $1 - \omega_{\text{post}}$ values. .... 110
4.9 Laplace-Laplace model: Six further genes identified as differentially expressed when using a cut-off $t_0 = 0.1$ and their corresponding $1 - \omega_{\text{post}}$ values. ....................................................... 110
4.10 Results from a simulation study designed to assess the performance of the symmetric Laplace-Laplace model. ................................. 114
4.11 Results from a simulation study to test the efficiency of the Laplace-asymmetric Laplace model in presence of asymmetric Laplace noise. ... 116
4.12 Results from a simulation study to test the efficiency of the Laplace-asymmetric Laplace model in the presence of asymmetric Laplace noise. 117
4.13 Results from a simulation study to test the efficiency of the Laplace-asymmetric Laplace model in the presence of normally distributed noise. 119
LIST OF TABLES

xix

4.14 Results from a simulation study to test the efficiency of the unadjusted
Laplace-Gaussian model in the presence of asymmetric Laplace noise.

. 121

4.15 Results from a simulation study to test the efficiency of the adjusted
Laplace-Gaussian model in the presence of asymmetric Laplace noise.

. 122

4.16 Results from a simulation study to test the efficiency of the unadjusted
Laplace-Gaussian model in the presence of normally distributed noise.

. 124

4.17 Results from a simulation study to test the efficiency of the adjusted
Laplace-Gaussian model in the presence of normally distributed noise.
5.1

. 125

A table of results to compare the numbers of differentially expressed
genes identified by the four models discussed in this thesis and their
corresponding FDR.

5.2

. . . . . . . . . . . . . . . . . . . . . . . . . . . . 132

A table of results to compare AIC values for the different models
discussed in this thesis. . . . . . . . . . . . . . . . . . . . . . . . . . . . 132

9.1

Times of startpoints, endpoints and peak apexes for HPLC trace. . . . . . 224

9.2

95% percentile and BCa confidence intervals for the areas of the eight
identified peaks (using fixed peak locations).

9.3

. . . . . . . . . . . . . . . 224

95% percentile and BCa confidence intervals for the startpoints (in
minutes) of the eight identified peaks (using variable peak locations). . . . 226

9.4

95% percentile and BCa confidence intervals for the endpoints (in
minutes) of the eight identified peaks (using variable peak locations). . . . 227

9.5

95% percentile and BCa confidence intervals for the areas of the eight
identified peaks (using variable peak locations). . . . . . . . . . . . . . . 228

9.6

Results from simulation study designed to test the efficiency of the
bootstrapping approach with fixed peak locations. The nominal coverage
is 95%. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 229


9.7 Results from simulation study designed to test the efficiency of the bootstrapping approach with variable peak locations. The nominal coverage is 95%. 

9.8 Results from simulation study designed to test the efficiency of the bootstrapping approach with fixed peak locations in the presence of increased noise levels. The nominal coverage is 95%. 

9.9 Results from simulation study designed to test the efficiency of the bootstrapping approach with variable peak locations in the presence of increased noise levels. The nominal coverage is 95%. 

Part I

Microarray data
Chapter 1

Genomics

1.1 Introduction

Genomics, in very broad terms, is the study of an organism’s genetic structure. Whilst such a research area is relatively well established, rapid technological advances of recent years have yielded many new opportunities in which to study genomics in much greater detail. Areas of current interest include sequencing the genome of various organisms and determining the functions of the identified genes. Possessing such knowledge about large segments of a genome has in turn created the possibility of functional genomics, which focuses more on the dynamic aspects of a genetic system, such as how it changes in different biological conditions or over time. In particular, in this thesis we consider microarray experiments, which aim to measure the activity of thousands of genes in a cell simultaneously. However, before presenting any discussion on the statistical techniques involved in analysing the data obtained from microarray experiments, we first provide an initial chapter detailing the relevant biology and nature of such experimentation. Doing so allows the reader to consider the statistics in the context of the problem and yields a greater appreciation of the complications that should be considered when developing tools for such a highly specific statistical analysis. Detailed expositions of the biological
ideas underpinning this thesis should be easily found in any reputable genomics text; Lesk (2007), Jorde et al. (2007) and Brown et al. (2008) are to name but a few.

1.2 Background biology

1.2.1 Deoxyribonucleic acid

Deoxyribonucleic acid, or more commonly referred to as DNA, is the foundation of replicable life. Essentially, DNA is the coded information contained within the threadlike, tubular structures known as chromosomes in the nucleus of a cell. DNA exists to instruct the cell as to how and when to synthesize specific products, such as proteins, which perform and regulate most basic functions. Examples would include forming cellular structures, catalysing biochemical reactions in cells or transporting molecules around the body. A DNA molecule consists of two strands of sugar phosphate, tightly wound around each other to form a right-handed double helix structure. On each sugar phosphate strand resides a linear arrangement of nitrogenous amine bases (adenine, thymine, guanine or cytosine), with bases on one strand paired with bases on the other strand, resembling rungs on a ladder. The bases are joined through hydrogen bonds in a specific manner, in accordance to the complementary base pairing rules of Watson and Crick. Thus, guanine will always be paired with cytosine; similarly, adenine is paired with thymine. A complementary base is the term used to describe the amine base that forms a bonding pair with another amine base (i.e. A with T and G with C).

The particular order of the bases along any one strand of the DNA structure is referred to as its DNA sequence. Briefly, the sequence is broken into triplet subsets of bases (called codons), which each acts as a code for the production of a particular amino acid. A chain of amino acids forms from consecutive codons and this constitutes an individual protein molecule. Further to this, the specific A-T-G-C subsequence of bases that codes
for a specific protein is known as a gene. DNA in the human genome comprises of approximately three billion base pairs, within which exists the coding for many different genes, which themselves can range quite widely in length from a few hundred base pairs to several thousand. Consequently the genes are very dispersed and are separated from each other by sequences of bases whose function remains largely unknown; this is called intergenic DNA. Within a gene, DNA bases may be classified as exons or introns. Exons are DNA bases that eventually code for amino acids in the proteins; introns are sequences that have no coding function. Each gene occupies a specific location (or locus) on a particular chromosome, which allows it to be specifically identified. This provides scope for mapping genes and allows the impact of chromosomal location to be considered.

1.2.2 Gene expression and hybridisation

Gene expression is the term used to describe the process where a gene converts the coded information stored in its DNA sequence into proteins. This process occurs in the cytoplasm of a cell and is imperative to produce essential proteins. Gene expression is regulated; not every gene is expressed in every cell at all times. Instead, a gene will only be expressed when the resulting protein is required or in response to external stimuli or environmental changes. All cells in a particular organism contain the same genetic information, however different genes will be active or highly expressed in brain cells as compared with skin cells, as compared with liver cells.

Measuring the abundance of proteins present in a sample can lead to underestimates of the amounts that particular genes express, as there is no guarantee that every protein will be produced, due to both internal and external complications. Further to this, proteins are notoriously difficult to measure robustly. Hence, it is often the intermediary molecule, messenger ribonucleic acid (mRNA), through which the protein-coding instructions from the gene are transmitted, that is a surrogate for ‘expression’, despite being only a precursor to the protein. The molecule mRNA is similar to DNA, however key differences include
that mRNA is single stranded and will only ever combine with DNA to form a regular double-helix, never with itself; its sugar is ribose, instead of deoxyribose; and the base thymine is replaced by uracil.

During expression the DNA splits open along the length of the structure. One strand remains inactive whilst a complementary strand of mRNA forms against the other, in a process called transcription: essentially, copying the message. The term complementary sequence refers to a single strand of DNA that contains the same sequential order of complementary bases as the given strand. Thus, the base sequence along the mRNA strand is identical to the replaced DNA strand, with the exception that uracil occurs instead of thymine. The mRNA strand then separates and moves out into the cytoplasm for protein synthesis to occur. Many copies of mRNA can be made from a single gene; gene expression essentially represents the number of mRNA copies of a particular gene present in a cell.

If there exist two single strands of a DNA molecule, whose sequences are complementary to each other, then there is a natural tendency for them to bind together in order to form one double-stranded DNA molecule. Alternatively, this process can take place between a strand of DNA and a strand of mRNA. This process is known as hybridisation and forms the underlying principle of microarray experimentation. There is no necessity for the strands to have originated from the same source, however the strands must be complementary in sequence for stable bonds to form.

The molecule of mRNA is relatively fragile and can be easily broken down by the action of enzymes that are prevalent in biological solutions. For experimentation purposes, researchers commonly manipulate a form of DNA that possesses the complementary bases of the mRNA, but exists in a more stable state. This form of DNA, known as complementary DNA (cDNA) can be created directly from the sample mRNA through a procedure known as reverse transcription.
1.2.3 Genome variation

It is thought that as little as 0.1% of the genetic make up differs from one individual human to another (Amaratunga and Cabrera, 2004). These slight changes in DNA sequence are referred to as genome variation and often involve only a few bases. For example, individual A may present guanine at the same position where individual B has cytosine; another may be missing a thymine base that is present in someone else. Some variations can be inherited via the reproductive cells; others may be unique to just one individual. Genome variation gives rise to situations where genes that are designed to do the same job may differ slightly from one person to another. The variants or alternative forms of a gene found to occupy the same position on a particular chromosome are known as alleles. We use the term wild-type to refer to the regular and properly-functioning forms of the gene that are thought to exist in nature. The term mutant denotes the atypical allele of a mutated gene.

Genome variations occur as a result of mutations and polymorphisms, which are alterations in the DNA sequence. The variation is referred to as a polymorphism if the difference occurs in at least 1 percent of the population. Further to this, single nucleotide polymorphisms (SNP) refer to polymorphisms that involve only one base. It has been estimated that on average, about 1 in every 1000 bases is a single nucleotide polymorphism and these result in a range of effects, including no effect at all. In cases where less than 1 percent of the population possess the modified gene, then the variation is classified as a mutation. Alterations to the base sequence can result through a variety of means. Common examples include deletion, substitution, insertion, inversion (a small subsequence of bases is removed then reinserted but having being reversed in direction) and translocation (a small subsequence is taken out and reinserted further along the sequence). Variations in the genome may be inherited or acquired. A inherited variant will be present in the DNA of all of the organism’s cells and may be passed on to the next generation. Somatic mutation describes changes that occur spontaneously after
conception as a result of external environmental stimuli or during DNA replication. Such a mutation will be only be present in the affected cells thus, unless these modifications have occurred in the reproductive cells, the variation will not be passed down to future generations.

The majority of variations that occur do not cause any noticeable effect. Sometimes the abnormality lies within the intergenic DNA region. Alternatively even if the variant lies inside the coding region, redundancies in the genetic code allow the same protein to be produced from sequences that are slightly different. For example, GCU, GCC, GCA and GCG all code for the amino acid alanine. Furthermore, cells do have the capabilities to repair certain types of damaged DNA. Generally, polymorphisms that produce harmful results tend not to survive the process of natural selection, thus the lasting ones tend to be harmless. Unfortunately though, mutations, which occur more rarely, can be harmful in so far as they could be the cause of a disease or alternatively, increase a person’s susceptibility to a disease that could result in death, but after the typical reproductive period.

1.3 Microarrays

In the past, scientists have been restricted by a lack of suitable technology to conducting genetic analyses on only a very few genes at any one time. However, recent years have seen the advancement of DNA microarray technology, which allows the examination of thousands of genes simultaneously, with the major aim often being to identify which genes are expressed differently between various samples. The compared samples may have been subjected to different experimental conditions, exposed to external stimuli (a toxic substance or viral infection), be tissues (cells) at different stages of a disease or process, or cells from mutated and wild-type organisms. Microarray experiments allow the observation and measurement of the relative changes in expression for a set of genes,
in response to the differences in condition and/or the activation or expression of other genes.

A DNA microarray aims to measure the abundance of messenger ribonucleic acid (mRNA) present in a set of cells. The concentration of mRNA for a particular gene $i$, by definition, is taken to be the amount that gene $i$ “expresses”. Very simplistically, a typical DNA microarray experiment involves the systematic “grid-like” arrangement of minuscule amounts of cDNA from the genes of interest, called probes, onto a solid support, usually a glass microscopic slide or nylon membrane. The mRNA samples of interest, whose gene sequences are unknown, are labelled, either with a fluorescent dye or radioactively, and then allowed to hybridise to the spots on the array. Matched mRNA molecules will hybridise to any complementary DNA sequence on the slide and the solution will stick to the slide. Any unmatched mRNA will not hybridise and is washed away. The signal emitted at each point on the chip is then measured. If a particular gene is highly expressed in the sample, it produces many molecules of mRNA, which in turn will hybridise to the probes on the microarray and generate a very strong signal - a bright fluorescent area or high level of radioactivity. Genes that are somewhat less expressed produce fewer mRNAs, which results in a weaker signal. If there is no signal, no messenger molecules have hybridised to the probes, indicating that the gene is inactive. By comparing the strengths of the signal for different samples, it is hoped that one might be able to identify any differences between the gene profiles for the various samples. Any spot whose intensity is different between the samples corresponds, by inference, to a gene that is differentially expressed between the conditions.

Whilst many different types of microarrays exist and all operate in a slightly different fashion, the main difference between technologies is that between the spotted 2-channel microarrays and the single-channel microarrays, of which Affymetrix chips are a popular example.

In spotted arrays, complementary DNA or long oligonucleotide probes are printed onto
the substrate. The two mRNA samples of interest are reverse-transcribed into cDNA, labelled with a different fluorescent dye (green cyanine 3 or red cyanine 5) and mixed together, before being washed over the slide and allowed to co-hybridise. After washing, a laser scanner is used to quantify the level of fluorescent signal emitted from each spot on the chip, resulting in two measurements per spot, one for the red channel and another for the green. The two samples of interest may be compared on the same slide, resulting in a direct estimate of differential expression. Alternatively, the samples of interest may be hybridised on different slides along with a reference sample, which will result in an indirect estimate of the difference in expression levels since the measurements are from different hybridisations (Yang and Speed, 2003). Figure 1.1 is an example of a 2-channel microarray chip. The green spots on the chip correspond to genes that are more expressed in the sample labelled with green dye; the red spots correspond to those genes that are more expressed in the sample labelled with red dye; finally, the yellow/brown spots imply that the corresponding genes are expressed to similar extents in both samples. In single-channel arrays, the probes are short oligonucleotides (usually between 5 to 50 base pairs long) and consequently, a gene has to be represented by a combination of multiple probes, usually referred to as the probe set. Furthermore, each chip can only measure the expression of a single sample, thus for inference to be made about differential expression between samples, several separate hybridisations will be necessary. For more details on microarray experimentation, the interested reader is directed to Brown and Botstein (1999) and Nguyen et al. (2002).

1.4 The impact of microarray experimentation

The term genotype describes the pattern of genes that are responsible for a particular trait observable in an individual. Phenotype describes the outwards characteristics of a particular trait, such as eye colour, or an inherited disease, such as cystic fibrosis. The
Figure 1.1: Microarray image showing differentially expressed genes. Green spots: genes that are more expressed in the sample labelled with green dye. Red spots: genes that are more expressed in the sample labelled with red dye. Yellow spots: genes that are expressed to similar extents in both samples. Image from http://www.imbb.forth.gr/people/poirazi/researchEP.html.
two are essentially connected as the phenotype results based on the encoded information in the genotype, however modifications of the phenotype may occur as a result of external factors, such as diet, temperature, and chemical environment, for example.

Genetic diseases are often caused by genes which are inappropriately transcribed – either too much or too little – or which are missing altogether. Such defects are especially common in cancers, which often occur when regulatory genes are deleted or inactivated. An illustration of this is the p53 tumour suppressor gene, which normally codes for a protein that discourages abnormal cell proliferation. A mutation of this gene might cause the resulting protein to lose its ability to stop abnormal cell growth, leading to rapid cell mitosis and the formation of tumours. Microarrays can be used to monitor the expression of genes in response to normal cellular activity, for example, cell transcription, across various biological conditions to identify key differences between groups.

Whilst quite rare, there are a few confirmed diseases that arise from an allelic variation in a single gene, such as haemophilia and Huntington’s disease. More commonly, diseases can arise from the sophisticated interaction of many genes and these have been classified as complex diseases. Each gene may interact with others relatively independently, however, in other cases, how a gene interacts may be dependent on specific combinations of other genes present. It has been observed that in many cases genes do not function in isolation and the expression levels of one gene regulate those of another gene; this is known as gene coregulation.

Individuals may inherit combinations of polymorphisms which makes them more prone to developing a particular disease. The presence of these genetic mutations does not guarantee the development of the disease; the likelihood of this is often heavily dependent on surrounding environment, diet and lifestyle. For example, whilst some individuals have inherited a propensity for coronary heart disease, the problem is further compounded in those that smoke, have a fatty diet and suffer from lack of exercise. Since many complex diseases are often influenced by one’s genetic make up, there is often evident prevalence
of the disease amongst certain groups: families; ethnic groups; male/female. Microarrays are used to identify the key genetic features that may predispose an individual to a complex disease and are particularly useful due to their ability to observe the expression in the same genes in different samples simultaneously.

As a further complication, cancers which appear clinically similar can in fact be genetically heterogeneous. Alizadeh et al. (2000) looked at diffuse large B-cell lymphomas in an aggressive malignancy of mature B-lymphocytes. With this disease, patients who present with the same diagnosis can have very different responses to treatment, summarised in terms of remission achieved and overall survival. It is suspected that this disease is genetically heterogeneous: different sets of missing or damaged genes in different patients can manifest in a disease with homogeneous clinical features, yet these molecular differences have implications on the ability to respond effectively to treatment. Microarrays assist in distinguishing between different patterns of abnormal transcription within a particular disease; such gene profiling can yield a diagnosis based on genotype rather than displayed symptoms, so that each patient receives the most appropriate and effective treatment.

Moreover, gene profiling can have a substantial impact on the invention of therapies targeted to the different varieties of the disease. The ultimate goal of pharmacogenomics would be a preventative, rather than reparative, approach to treatment. It is hoped that in the future, patients can be stratified according to their genome pattern and prevention provided, specific to their genotype, before the disease manifests. An example of this in practice is the prescription of tamoxifen, which has been suggested to prevent breast cancer among women with BRCA1 and BRCA2 gene mutations. Further preventative measures may be prompted in terms of lifestyle choices.

Additionally, genomics plays a part in the development of drugs and treatments. Target molecules that are strongly associated with the disease process can be identified by studying correlations between genome variations and diseases or gene expression data
and diseases. These form the intended sites for drug activity, rather than just tackling the symptoms. Adverse reactions to drugs usually occur as a result of drugs having an affinity with more than just the intended site. The aim is to develop drugs that specifically target just the mutated molecules. Consequently, fewer undesirable side effects are to be expected, resulting in a drug that is more useful in a clinical environment.

1.5 Description of datasets

In the work that follows, we consider four datasets for analysis. Here, we briefly describe the experiments and the form of the data.

1.5.1 E.Coli data

The E.Coli dataset originates from a cDNA experiment using “homemade” spotted arrays. The metJ gene encodes a protein that regulates expression of the methionine biosynthesis genes of E.Coli. This microarray experiment looks at how mutating (inactivating) the metJ gene affects the expression profile of the met regulon. The method for this inactivation and further details can be found in Marincs et al. (2004). The RNA was isolated from both the wild-type (non-mutated) and the mutant strains, from which cDNA was synthesised using two different fluorescent nucleotides (green for the sample containing the wild-type gene and red for the treated sample containing the mutant gene). These labelled cDNAs were mixed and hybridised to the array. After the hybridisation the arrays were washed and scanned to produce the raw data.

We have data for $m = 3$ chips and $n = 76$ genes, along with an additional 102 “blank” spots used to estimate the background effect. There were $r = 6$ replicated spots of each gene per chip.
1.5.2 HIV data

We consider the publicly available HIV spike-in dataset of van’t Wout et al. (2003) and discussed further in Gottardo et al. (2006) and McLachlan et al. (2006). Here, \( m = 4 \) cDNA experiments were conducted to compare expression levels for \( n = 7680 \) genes from uninfected CD4-T cell lines with genes from cells at time \( t = 24 \) hours after infection with the HIV-1 virus. Furthermore, 12 HIV-1 genes were spiked in as positive controls, as these were known in advance to be differentially expressed in infected versus uninfected cells. For each condition, the same pooled RNA preparations were used on all slides; hence these should be considered technical replicates rather than individual samples. For more information on types of replication, one should refer to Section 2.1.

1.5.3 Lymphoma data

The lymphoma dataset consists of a small subset of the publicly available data from the previously mentioned cDNA 2-channel experiment by Alizadeh et al. (2000), which explores the classification of diffuse large B-cell lymphomas into distinct groups with homogeneous clinical behaviours by gene expression profiling. Gene expression levels were measured using a specialised microarray, the “Lymphochip”, which contains those genes known to be preferentially expressed in lymphoid cells or those known to be of immunological or oncological significance. The experimental samples were labelled with a fluorescent red dye and hybridised onto the chips along with a common reference mRNA sample, made up from a pool of nine different lymphoma cell lines and labelled with a fluorescent green dye. The dataset contains measurements for \( n = 9216 \) genes and \( m = 4 \) replicates.
1.5.4 Control-control data

This dataset is derived from an experiment in which a control sample was hybridised to itself. The control samples were RNA from unchallenged FLO-1 cells and we have observations for $n = 18641$ genes and $m = 3$ replicates. The key aspect of such an experiment is that no differential expression should be observed: every gene should be expressed to the same extent in both samples (as the samples are essentially the same). Thus, what we observe should simply be the noise realised in the experiment. These data are currently unpublished.
Chapter 2

Analysing microarray data

The analysis of microarray experiments can be simplistically broken down into three phases: experimental design, measurement and pre-processing, and post-processing analysis. All three of these phases offer new challenges for statisticians to tackle. In this thesis, we consider the problem of identifying differential expression, which would fall into the post-processing category. Thus, the development of optimal experimental designs and novel pre-processing methods is outside the scope of this work. Time restrictions prevent a review of all of the work in the microarray field; however, for completeness, a brief overview is provided of some of the more pertinent issues from the first two stages.

2.1 Experimental design and replication

Design is an important, yet often overlooked, element of the experimental process. Suppose we wish to conduct an experiment to compare just two biological conditions: the most simple experimental design would be to obtain a direct comparison of the expression levels of the two samples by hybridising both samples simultaneously onto the same array. However, for more than two conditions, this design becomes infeasible. Instead,
a common reference design is often adopted for 2-channel microarray experiments (Sterrenburg et al., 2002). This involves pairing each sample of interest with a reference sample, which could be extracted from tissue of wild-type organisms, a control sample of no biological interest, or a pool of all the samples of interest. Such a design will result in an indirect estimate of differential expression as the expression levels for the samples of interest result from different hybridisations. However, Wit et al. (2005) pointed out that this type of design is very inefficient as half of all resources are given to measuring a condition of no scientific interest; a matter with which various other authors concur (Vinciotti et al., 2005; Kerr and Churchill, 2001; Landgrebe et al., 2006; Khanin and Wit, 2005; Wit and McClure, 2004). Many alternatives, for example, loop designs, have been proposed in the past that could be applicable to microarray experimentation (Bagchi and Cheng, 1993; John and Williams, 1995), however these are restricted by certain limitations. For example, Kerr and Churchill (2001) showed that loop designs are only optimal for experiments involving a relatively small number of samples of interest. Wit et al. (2005) strove to find practical and efficient designs for 2-channel microarray experiments. They considered a class of symmetric designs, called interwoven loop designs, which are shown to perform relatively well in comparison to near-optimal designs yielded by simulated annealing. Rosa et al. (2005) provides examples of how the modelling of microarray data should be influenced by the individual study design.

Regardless of the specific design, replication is an essential part of any microarray experiment, in order to assess and eliminate any inconsistencies in the data from future analyses. Many researchers in the field are now recognising the importance of replication in microarray experiments (Lee et al., 2000; Pritchard et al., 2001); however in practice, the inclusion of many replicates is often severely compromised by the cost of commercially produced chips. Biological replication involves comparing independent samples that have originated from distinct biological sources and provides an estimate of the biological variation present in such a system. Technical replication attempts to provide information on the natural and systematic variation that occurs when performing the
experiment and exists at a variety of levels. These could include multiple tissue samples from each individual subject, several extractions from each tissue sample, numerous hybridisations of each cDNA sample onto different arrays, or replicate spots of the same gene spotted across each array. Dye-swap designs are a common approach to incorporating technical replication. This involves labelling and hybridising the samples in duplicate, half of the experiments with the fluorescent dyes as assigned and half with the fluorescent dye assignment reversed. This is an attempt to accommodate any biases that are inherent in the labelling and hybridisation processes, for example, if some genes are more efficiently labelled with one or other of the dyes.

Within-array replicate spots are a common feature in many experiments, yet the treatment of this form of replication is quite varied. At the very least, much of the literature supports the practice of averaging the measurements for replicate spots prior to any formal statistical analysis being carried out. However, Smyth et al. (2005) claimed that averaging the replicate spot information results in a loss of valuable information and used the replicates to estimate the strength of correlation between the spots for a particular gene, with the aim being to improve the precision with which gene-specific variances can be estimated. The inclusion of replicate spots has proven a useful tool for quality assessment: any discrepancies between the spots suggest the presence of a local artifact or other spatial effects. Various researchers have considered the repeatability of measurements across replicate spots and have used this to construct measures by which to assess spot quality (Beissbarth et al., 2000), and as a basis on which to identify and reject outlying spots (e.g. Tseng et al., 2001). There exist relatively few studies that have attempted to improve the assessment of differential expression through the incorporation of between-spot information, however Baggerly et al. (2001) provides an example of this type of work.
2.2 Pre-processing

2.2.1 Background adjustment

Currently, microarray technologies do not produce robust and reliable datasets on a consistent basis. The raw data essentially consists of the signal intensities extracted from the scanned images of the arrays using specialized imaging software. The intensities of the signal imply the amount of sample DNA bound to the array, however poor quality images, often resulting from fluctuations in within-spot intensities and fabricated artifacts (for example, dirt, fingerprints or hairs on the slide), can lead to these measurements being substantially corrupted. Furthermore, spot shapes are frequently irregular, leading to problems discriminating between spot foreground and background. Additionally, the scanners used to determine the amount of sample bound to each spot often have an intensity saturation level: any pixels with an intensity that exceeds a certain threshold can not be properly quantified and are truncated to the corresponding threshold value. This common affliction is known as saturation and gives rise to biased estimators of gene expression. The amount of bias is clearly positively dependent on the proportion of saturated pixels.

Theoretically, the intensities of those pixels in the image not corresponding to spots should be zero. In practice, these pixels actually emit a low, but not insubstantial, level of fluorescence that may vary with location. This effect is due to a variety of causes, including the non-specific binding of the labelled samples to the array substrate and substrate fluorescence. Thus it is assumed that the measured spot intensity contains a certain amount of this non-specific background fluorescence and that the observed intensity is actually an additive combination of the true spot intensity plus the background fluorescence. To adjust for any background effect, an estimate of the background intensity is made from the data and this value is subtracted from the raw spot intensities in order to produce a set of background-corrected spot intensity values.
An estimate of the background intensity can be made in a variety of ways: a number of background measurements (readings taken from pixels between those of the spots) may be averaged to give an overall background intensity, which is then in turn subtracted from all spot intensities, or a measurement of the local background intensity may be subtracted from each spot individually. Alternatively, in the E.Coli dataset presented throughout this work, 102 “blank” spots were randomly positioned on the array in order to generate an estimate of the background intensity.

In theory, the background intensity estimate should always be less than the spot intensity for all spots, however in practice, this does not always prove to be the case. Negative adjusted spot intensities can be problematic if certain transformations are required as part of later normalisation procedures. This undesirable property can be avoided by the use of a suitable “fudge factor”; a simple approach is to set a threshold $t$, so that for each spot $i$

$$\text{background-corrected intensity}_i = \max(\text{spot intensity}_i - \text{background intensity}_i, t),$$  \hspace{1cm} (2.1)

where background intensity$_i$ may be a global or spot-specific background estimate and $t$ is taken to be a low percentile of the spot intensity values.

### 2.2.2 Transformation

A transformation of the data is often an integral part of pre-processing and usually precedes any more formal pre-processing procedures. The need for some transformation arose in order to ease a visual inspection of the data, however striving for normality and homogeneity of the variance are better justifications for such an action, as these are important for some post-processing analysis techniques. By convention, researchers in the biological domain often use a log$_2$ transformation. This is an attractive choice for a non-statistician as it makes the post-processing analysis of fold changes simpler. However, we
question whether this choice is perhaps somewhat arbitrary. In Section 2.2.4 we show that this transformation is not always the most appropriate choice and there exist more robust alternatives, which in turn, can improve the success of normalisation.

### 2.2.3 Normalisation

The problem of inconsistent data quality is further compounded in comparative studies involving multiple arrays. It was discovered by early researchers in the microarray field that substantial differences in intensity measures were arising for samples that were treated exactly alike. These differences can usually be linked to systematic effects inherent in the experimental process, such as mRNA preparation, the concentration and amount of DNA put on the microarrays, labelling efficiency, hybridisation efficiency, lack of spatial homogeneity of the hybridisation on the slide, scanner settings, saturation effects and background fluorescence; all of which contribute to the challenge of analysing microarray data. Furthermore, there is a dye bias present in almost all 2-channel experiments. Due in part to the physicochemical properties of the dyes (it is known that the green dye binds more effectively to the sample), the intensities from the green channel tend to be generally higher than those from the red channel. The magnitude of the difference is often dependent on the overall intensity. Other factors that contribute to this bias are the efficiency of dye incorporation, the scanning properties of the dyes, fluctuations in processing procedures or the settings of the scanner. Clearly, the effects of any systematic sources of variation need to be removed to improve the comparability between channels. This stage of pre-processing is known as normalisation. Successful normalisation should adjust the data in such a way that subsequent analysis only reveals true biological differences pertinent to the scientific question being addressed.

The most appropriate type of normalisation scheme to use differs according to the type of microarray platform employed. Since the data presented in this work focuses on 2-channel cDNA arrays, we present some discussion of the normalisation methods most relevant to
the 2-channel platform. These methods primarily concentrate on balancing any disparities in fluorescence intensities between the two coloured dyes (within-array normalisation). Having completed this satisfactorily, if the experiment involves multiple arrays, one must also consider the possible scale heterogeneity introduced through replicate arrays (between-array normalisation), before robust comparisons can be made. The effects of these two distinct types of normalisation on the data are illustrated in Figure 2.1. Prior to any normalisation, there is considerable variation between the red and green channels and the multiple arrays. Following within-array normalisation, the similarity between the red and green intensities for each array is increased, however discrepancies between arrays still remain. Between-array normalisation improves the comparability of the multiple arrays. For a discussion of pre-processing techniques for single-channel arrays, the reader should see Bolstad et al. (2003).

Global versus intensity-dependent normalisation schemes

Many normalisation schemes exist; some of the issues relevant to the choice of scheme are discussed in Walls et al. (2005). Most schemes can be classified simply into one of two groups: global or intensity-dependent normalisation.

Global or linear normalisation schemes are based around the assumption that the spot intensities for a pair of channels are linearly related with no intercept. A result of this assumption is that it should be possible to correct any lack of comparability by simply adjusting every single spot intensity by the same amount, regardless of its intensity level. The amount is called the normalising factor. Global schemes are the most widely used by scientists, despite their lack of sophistication, as they can be implemented mathematically and computationally with ease. Here we describe the schemes in the context of within-array normalisation, where the aim is to balance the red and green intensities and eliminate dye bias. However, many of these described scale adjustment methods can be extended to correct for between-array differences.
Figure 2.1: Density plots to illustrate the difference between within-array and between-array normalisation for the HIV data. Top left panel: Without any normalisation. Top right panel: After within-array normalisation. Bottom panel: After between-array normalisation.
An example of global normalisation includes normalisation by the sum: given that the total mRNA content should be roughly the same across samples, normalisation by the sum constrains the sum of the red intensities to be equal to the sum of green intensities. This is achieved by dividing all of the red and green intensities by their respective sums, forcing both the sums to be equal to one. There exist variations on this type of normalisation, including scaling the intensities so that the mean or median are the same across channels or arrays, or using a subset of the genes on the array, as opposed to the entire set. Various alternative methods exist in the literature for the purpose of normalising expression ratios: these include log-centering, rank invariant methods (Tseng et al., 2001) and Chen’s ratio statistic (Chen et al., 1997). However, all are limited by a lack of consideration for any systematic biases.

Intensity-dependent normalisation schemes encompass those methods in which the normalising factor is not a constant, but a nonlinear normalisation function of the intensity level: $Y \rightarrow f(Y)$. The need for more sophisticated approaches such as these arose from the problem of how to normalise datasets where the features of the data imply that the normalising factor needed to adjust the low-intensity measurements is not the same as the factor needed to adjust high-intensity measurements.

Define $M_i = R_i - G_i$ and $A_i = (R_i + G_i)/2$; in other words, $M_i$ is the difference between red and green intensities for each spot $i$ and $A_i$ is the arithmetic mean for each spot. It is worth noting at this stage that when the intensities have previously undergone a logarithmic transformation, $M_i$ is equivalent to the log-ratio $\log(R_i/G_i)$.

For each spot $i$ we can calculate

$$M_i^* = M_i - c(A_i),$$

(2.2)

where $c(\cdot)$ is the normalising function. Replacing the original $M_i$ by the residuals, $M_i^*$, and using the definitions of $A_i$ and $M_i$, the normalised red and green intensities are
obtained by

\[ \hat{R}_i = A_i + \frac{M_i^*}{2}, \quad \hat{G}_i = A_i - \frac{M_i^*}{2}. \]

There are various methods that fit into this family of normalisation techniques. A widely used approach is that of Yang et al. (2001), who proposed fitting a locally weighted linear regression (LOWESS) (Cleveland, 1979) curve to the \( M_i \) values on a \( \log_2 \) scale. Berger et al. (2004) discussed how to choose the bandwidth parameter necessary for this type of regression so as to optimise the efficiency of the procedure. On commercially produced chips, each subgrid of probes is robotically printed by a different print-tip. The lowess normalisation may be applied locally to each grid on the microarray (Yang et al., 2002b), which can help overcome inconsistencies in the print-tips and spatial variation on the slide surface.

Kepler et al. (2002) proposed a similar generalised approach based on an alternative local regression method. In addition to generating normalised intensities, this method also estimated the expression-level dependent variance.

The success of intensity-dependent schemes rely on choosing an appropriate invariant gene set: the subset of genes from which the normalisation function will be estimated. The genes within the invariant set act as “anchor points” from which to gauge the scale of adjustment necessary, hence they should not exhibit differential expression across samples; furthermore, to avoid the need for any extrapolation of the normalising function, the expression levels of the invariant gene set should span the entire range of observed expression levels. When it is reasonable to assume that a large majority of genes are not differentially expressed, then the entire set of genes of the microarray can be used as the invariant gene set, as most schemes will sufficiently tolerate a small number of outliers. Alternatively, researchers may incorporate certain control or housekeeping genes with known expression behaviour for the purpose of creating an invariant gene set.
Fixed effects ANOVA models have been proposed as a strategy for normalisation (Kerr et al., 2000) as they have the potential to normalise with respect to spatial effects arising from location of spots, chip effects and dye biases. A limitation of such a model is the assumption that a linear function is appropriate for this type of data: any non-linear intensity-dependent biases (namely the gene specific dye effects inherent in 2-channel microarray experiments) cannot be accommodated. Furthermore, the adoption of a fixed effects model does not take into account the multiple sources of variation. Wolfinger et al. (2001) extended the ANOVA model to include random effects, which accommodated the array effects and array-treatment interactions. Further, the model described by Wolfinger et al. (2001) can be adapted, depending on the specifics of the experimental design, to incorporate extra levels of technical replication. We take these ideas into consideration when developing our own linear model in Section 3.4.

More sophisticated normalisation methods are beginning to emerge in the literature. The interested reader should consult Xuan et al. (2005), who used an iterative non-linear regression method; Wang et al. (2004), who used wavelet regression as a means for recovering the regression function from the data; and Wang et al. (2005), who proposed a robust semi-parametric approach, implemented through a two-way semi-linear model. This approach is particularly attractive in certain cases as it is not built upon the assumption that the majority of genes are not differentially expressed.

**Variance-stabilising normalisation methods**

It has been noted (see Rocke and Durbin (2001), for example) that the variance of the raw intensities often increases with their mean. Such heterogeneity is clearly undesirable and a variety of variance-stabilisation transformations have been suggested in an attempt to bring the distribution of the gene intensities closer to Gaussian. These include a simple logarithmic transformation (Smyth et al., 2003; Hoyle et al., 2002) through to the Log-Linear Hybrid transformation of Holder et al. (2001) (also known as Linlog in Cui et al.
The datasets used for illustration in this work have been subjected to a variance stabilising
normalisation scheme known as “vsn”, proposed by Huber et al. (2002). The method is
intensity-dependent and calibrates for channel and array variations through both shifting
and scaling. The variance stabilisation is incorporated through a ‘generalised logarithmic’
(glog) transformation (Munson, 2001; Durbin et al., 2002; Huber et al., 2002), which
strives to achieve constant variance across all genes.

The glog transformation is motivated by the following 2-component model for the
measurement error of gene expression arrays (Rocke and Durbin, 2001):

\[
y_i = \alpha + \mu_i e^{\eta_i} + \epsilon_i, \quad i = 1, \ldots, n
\]

where \( y_i \) is the measured expression level for a single gene, \( \alpha \) is an offset term to
accommodate the mean background, \( \mu_i \) is the true expression level for gene \( i \), and \( \eta_i \)
and \( \epsilon_i \) are additive and multiplicative error terms, respectively. The components \( \eta_i \) and \( \epsilon_i \)
are assumed to be independent and normally distributed with zero means and variances
\( \sigma^2_\eta \) and \( \sigma^2_\epsilon \) respectively.

For low expression levels i.e. the \( \mu_i \) close to zero, the observed intensity will be controlled
by the first term in the model (2.3) in that it can be written as \( y_i \approx \alpha + \epsilon_i \). Thus, \( y_i \) is
approximately normal with mean \( \alpha \) and variance \( \sigma^2_\epsilon \).

At high expression levels, \( \mu_i \) is large and the middle term of (2.3) dominates the model.
The observed intensity can be consequently written as \( y_i \approx \mu_i e^{\eta_i} \), and is approximately
log-normally distributed with variance \( \mu_i^2 \sigma^2_\eta \), where \( \sigma^2_\eta = e^{\sigma^2_\eta}(e^{\sigma^2_\eta} - 1) \). Here, it is clear
that the variance of \( y_i \) varies linearly with \( \mu_i^2 \); on a logarithmic scale, \( \log(y_i) \approx \log(\mu_i) + \eta_i \)
and the variance of \( \log(y_i) \) is constant. It is this key idea that underpins such a
normalisation scheme.
For moderate expressions levels, the observed intensity forms an intermediary between the two extremes and behaves as a mixture of a normal and log-normal distribution with variance $\mu^2_S^2 +\sigma^2_\eta$. Once again, the mean-variance dependency is evident.

The glog transformation can be written as

$$\hat{y} = \log[(y - \hat{\alpha}) + \sqrt{(y - \hat{\alpha})^2 + \hat{c}}],$$

(2.4)

where $\hat{c} = \sigma^2_\epsilon/S^2_\eta$.

Applied to microarray data, the glog transformation is equivalent to a linear transformation of the measurements in the low-intensity range and coincides with the natural logarithm of the measurements in the high-intensity range, and for data originating from the model (2.3) yields symmetric, transformed gene intensities, which have stabilised variance. However, it possesses further advantages over other transformations such as the logarithm and negative square root: there is no singularity at zero and the function remains smooth and real-valued, even in the negative domain, which can be important after background subtraction. The parameters can be estimated through a robust variant of maximum likelihood but some replication is necessary. The vsn procedure has been implemented in the arrayMagic package for the Bioconductor project, found at [http://www.bioconductor.org](http://www.bioconductor.org) (Gentleman and Carey, 2002).

An alternative approach to the variance-stabilisation problem is to consider each gene intensity to be a Poisson random variable, with some parameter $\lambda$. It is known that the variance of the Poisson random variable changes with the mean; suggestions have been made as to how to transform these type of observations into a sequence of random variables with near constant variance and a distribution closer to normality. One of the first such transformations to appear was that of Anscombe (1948), who suggested the mapping $Y \rightarrow 2\sqrt{Y + 3/8}$. Fryzlewicz and Nason (2004) proposed a new class of transformations for this purpose, known generically as Haar-Fisz transforms. Later, Fryzlewicz et al.
(2007) introduced the Data-Driven Haar-Fisz transform which used a similar approach but additionally estimated the mean-variance relation as part of the stabilisation process. This transform is suitable when there is a monotone mean-variance relationship however, unlike the work of Huber et al. (2002) where the model (2.3) has to be specified, it is not necessary to know the precise form of the relationship.

Motakis et al. (2006) described the application of the Data-Driven Haar-Fisz transform within a single-channel microarray context. The transform relies on an input sequence of positive random variables, \( X_i \), with some mean \( \mu_i \) and a variance \( \sigma_i^2 \); further, there should be some monotone (non-decreasing) relationship between the mean and variance i.e. \( \sigma_i^2 = h(\mu_i) \). In order for the transformation to perform at its best the underlying \( \mu_i \) should form a piecewise constant sequence (Fryzlewicz et al., 2007). How to organise microarray data into this suitable format is not intuitively obvious at first; Motakis et al. (2006) suggested ordering replicate gene intensity sets according to their increasing mean observed value, where the mean is estimated across replicates. Simulations suggested that the Data-Driven Haar-Fisz transform achieved good variance stabilisation whilst producing intensities that were closer to being normally distributed than other established methods. However, applying this type of approach to microarray data is still in its infancy and no generalisation to two-colour microarray data has yet been proposed. Furthermore, the method is currently limited as the calibration of multiple slides, a major aim of normalisation, is not addressed. Consequently, we do not adopt the practice of Motakis et al. (2006) in this thesis.

### 2.2.4 Example

Figure 2.2 shows some exploratory plots of the raw E.Coli data. The clustering of similar characters in the scatterplot suggests that the intensities for the replicated spots for each gene have similar measures. This is to be expected and any departure from this may have implied some form of location bias. Whilst the outliers appear sufficiently apparent to
Figure 2.2: Exploratory plots for raw E.Coli data. Top left panel: Scatterplot of data. Points denoted by the same character and colour correspond to spot replicates. The black line is the identity line $y = x$ and the blue line is a linear model fitted to the data. Top right panel: MVA plot with $M = 0$ line. Bottom left panel: Histogram of data. Bottom right: Normal Q-Q plot of data.
identify them simply through a visual inspection, the departure of the fitted linear model from the identity line is what we aim to improve upon through normalisation.

A further concern with raw microarray data is that the magnitudes of such measurements are often heavily skewed, with many small intensities followed by a very long tail, as illustrated by the histogram in Figure 2.2. It is possible that as much as 75% of the data may lie in the lowest 10% range of intensities.

The inspection of an MVA (minus vs. average) plot is a useful tool in assessing the need for normalisation of the two channels. We construct a scatterplot of $M_i$ against $A_i$, and if there is no systematic dye bias, the points on the MVA plot should be scattered around the $M = 0$ line. Furthermore, the funnel shaped distribution of the points around the identity line, as observed in Figure 2.2, is a common feature to many datasets and suggests some variance heterogeneity or in other words, non-constant variance. Specifically, as the mean intensity increases so does the variance.

Figure 2.3 shows plots of the logarithmically transformed E.Coli data ($Y \rightarrow \log_2(Y)$). Logging the intensities has substantially reduced the skewness of the distribution, but has not eliminated it completely. Moreover, working with logged intensities has other advantages: the variation of logged intensities tends to be less dependent on the magnitude of the values and thus improves variance estimation. However, in the case of the E.Coli data, there is still evidence of some systematic dependency, even after the transformation; such a feature was also noted by Yang et al. (2002b) and Yang et al. (2002a) for other datasets.

A variety of power transformations have been experimented with in the past and have proved useful for certain microarray datasets (Amaratunga and Cabrera, 2001; Tusher et al., 2001). Specific to the E.Coli dataset, a transformation of the form $Y \rightarrow \frac{1}{\sqrt{Y}}$ shows considerable improvements over a logarithmic transform: the histogram in Figure 2.4 shows a definite reduction in skewness, with the resulting distribution of a more desirable, symmetrical form. Stronger power transformations overcompensated for the skewness,
Figure 2.3: Plots (as in Figure 2.2) for E.Coli data after a $\log_2$ transformation.
yielding transformed data which were skewed in the opposite direction. The fitted line in the scatterplot is a better approximation to the desired $y = x$ line than those observed previously. The problematic funnel shape is less apparent in the MVA plot and the points appear to be more randomly dispersed than in those plots previously, suggesting that this particular transformation has stabilised the variance somewhat. These improvements indicate that an adjustment, even one as simple as transformation, has made a definite attempt to correct for some of the incorporated bias.

We performed lowess normalisation of the E.Coli data, following both a logarithmic transformation (performed in Bioconductor) and a negative square root transformation (implemented manually in R). Figure 2.5 gives MVA plots of the normalised intensities. Yang et al. (2002b) and Yang et al. (2002a) claimed that this type of normalisation should eliminate any systematic dependency that the $\log_2$ ratios have on intensity. However, this appears not to be the case for the E.Coli data, where there is still some evidence of the non-constant variance, even after normalisation. This feature is less apparent when the normalisation scheme is teamed with the alternative negative square root transformation. Figure 2.6 shows plots of the E.Coli data after the vsn normalisation of Huber et al. (2002), which yields better results than the lowess scheme. Based on these exploratory plots, we suggest that alternative transformations to the conventional $\log_2$ can lead to superior outcomes post normalisation. Furthermore, we emphasise the need for visual inspection of the data before and after normalisation to ensure that the most appropriate method of normalisation has been employed. The consequences of blindly choosing a poor normalisation scheme can have a substantial impact on future analyses, as data with true biological meaning may be lost.
Figure 2.4: Plots (as in Figure 2.3) for E.Coli data after a negative square root transformation.
Figure 2.5: Left panel: MVA plot of E.Coli data after lowess normalisation following a log$_2$ transformation. Right panel: MVA plot of E.Coli data after lowess normalisation following a negative square root transformation. In both plots the dashed lines are lowess curves fitted to the data.

Figure 2.6: MVA plot of E.Coli data after a vsn normalisation.
2.3 Post-processing analysis

2.3.1 Notation

Suppose that in a particular 2-channel microarray experiment, we compare \( n \) genes under two experimental conditions (treatment/control, diseased/non-diseased). We have \( m \) replicate chips and on each chip, \( r \) replicate spots. We begin with the raw measurements \( T_{ijk} > 0 \) and \( C_{ijk} > 0 \) \((i = 1, \ldots, n, j = 1, \ldots, m, k = 1, \ldots, r)\) for the treatment and control samples respectively. It is thought that these observed intensities \( T_{ijk} \) and \( C_{ijk} \) are actually each an additive combination of the true spot intensity plus some background fluorescence. Hence to adjust for this, an estimate of the background intensity is made from the data and this value is subtracted from the raw spot intensities to give a set of background-corrected spot intensity values, denoted \( T'_{ijk} \) and \( C'_{ijk} \) respectively.

The chosen normalisation procedure is applied to the set \{\((T'_{ijk}, C'_{ijk})\}\}. Let \( T'_{ijk} \) and \( C'_{ijk} \) denote the background-corrected, normalised intensities for the \( i \)th gene, \( j \)th chip and \( k \)th spot for the treatment and control samples respectively, where we assume that \( T'_{ijk} \) and \( C'_{ijk} \) are on some approximately logarithmic scale. Then we can compare the two samples through the difference

\[
X_{ijk} = T'_{ijk} - C'_{ijk}.
\]

We call \( X \) a contrast variable. When the vsn normalisation procedure is used, the contrast \( X \) will be simply the difference between expression in the two samples in the low-intensity range and equivalent to a log-ratio of expression in the high-intensity range. All subsequent analysis will be based solely on the set of \( X_{ijk} \) values.
2.3.2 Conventional methods of analysis for identifying differential expression

The recent development of microarray technology has resulted in a flood of literature regarding determination of the differentially expressed genes. Essentially, the problem comes down to identifying those genes with $T_{ijk}^*$ significantly different to $C_{ijk}^*$ or in other words, a contrast value significantly different from zero. We touch upon some of these existing approaches to the problem and their relative merits, or perhaps more importantly, their limitations, in order to draw comparisons with the new procedures described in Chapters 3 and 4.

The simplest method employed to select differentially expressed genes is on the basis of a fold change criterion. Let $\bar{X}_i$ be the contrast for gene $i$, averaged over the replicates. Those genes with $|\bar{X}_i|$ greater than some threshold value are said to be differentially expressed (Schena et al., 1996; DeRisi et al., 1997). For example, many biological scientists choose to adopt a two-fold difference as the cut-off: any gene with expression in one class over two-fold greater or less than the expression in the other class is concluded as differentially expressed. Whilst the choice of threshold is flexible making the test more or less conservative depending on the user, this method essentially involves some arbitrary selection based on no sound statistical grounds, making its gross overuse in the field a great concern. Moreover, the fold-changes are of limited value as the intensities are subject to some biological, experimental and measurement error, and unless the error distribution can be derived, it is difficult to assess whether a particular contrast is worth noting or whether it has occurred by chance.

Standard significance testing is another commonly adopted approach to the problem, which allows the assessment of significance of the observed differences whilst accounting for all the different sources of variation, which are not necessarily the same from gene to gene. Generally, the null hypothesis would be that the experimental conditions have some
common distributional parameter (for example, the mean or median of the normalised intensities), which leads to the \( t \)-test and its variations. Many problems exist with this form of assessment, namely variance estimation and multiple testing.

**Improving variance estimation**

Modelling the variability of gene expression usually takes one of two forms: either a constant global variance across all genes or individual gene-specific variances. The former, in most cases, is too stringent an assumption, however a carefully chosen normalisation scheme (such as vsn) can make this assumption reasonable. The latter has low power as each gene-specific variance estimate is usually based on only a limited number of replicates, due to the expense of commercially produced chips. This will result in an estimate of the sample standard deviation which is unreliable and also makes distributional assumptions of normality difficult to justify.

The problem of limited replication can lead to significant underestimates of the variances, which consequently leads to inflated \( t \)-statistics and an increase in the number of false positives. Lonnstedt and Speed (2002) proposed discarding those genes with a small fold-change and a very small standard deviation as an ad-hoc solution to the problem. Many researchers (Efron et al., 2001; Tusher et al., 2001; Broberg, 2003; Baldi and Long, 2001) have considered the use of a “modified” \( t \)-statistic, which incorporates an offset standard deviation. These methods have proved successful, however are not motivated by a model and have no associated distributional theory.

In Chapter 3, we compare our new method to the that of Tusher et al. (2001), known as Significant Analysis of Microarrays (SAM), who used a test statistic of the form,

\[
d_i = \frac{\bar{x}_{iT} - \bar{x}_{iC}}{s_1 \sqrt{1/m_1 + 1/m_2 + s_0}}
\]

(2.6)
where $\bar{x}_{iT}$ and $\bar{x}_{iC}$ are mean levels of gene expression for gene $i$ in treatment and control conditions respectively; $s_i$ is the pooled standard deviation of repeated measures for gene $i$; $m_1$ and $m_2$ are the number of replicates in each condition/class and $s_0$ is the offset, estimated by minimising the coefficient of variation of the $d_i$. Permutation methods were employed to estimate the null distribution of the modified test statistic, after which cutoff points $c_1$ and $c_2$ were derived to create a critical region of the form $d_i < c_1$ or $d_i > c_2$, where $c_1 < c_2$.

Shrinkage-based variance estimators have recently begun to manifest in the literature, in response to the need for an intermediate approach to variance estimation. These methods aim to borrow strength across groups of genes and many do so through a Bayesian framework. Examples of this type of work include Baldis and Long (2001), Broet et al. (2002) and Lewin et al. (2005), which all generated gene-specific variances, being the weighted average of the empirical variance and a prior variance estimate. Lonnstedt and Speed (2002) used a parametric empirical Bayesian approach within a 2-channel application and, using a mixture of normals and a conjugate prior, derived an expression for the posterior odds of differential expression for each gene. Smyth (2004) extended the work of Lonnstedt and Speed (2002) to accommodate for all of the commonly used microarray platforms, resetting the model in the context of general linear models with arbitrary coefficients and contrasts of interest. This resulted in the posterior odds statistic being reformulated in terms of a moderated $t$-statistic. Instead of relying solely on single-gene estimated variances, these were combined with a global variance estimator in a weighted average. Essentially, this is equivalent to the gene-specific sample variances being shrunk towards a pooled estimate, allowing more robust inferences to be made when the number of replicates is small. This method can be found in the \texttt{limma} package in Bioconductor and we compare the results of our new method to the “\texttt{limma}” approach in Chapter 3. All of the above described methods result in the use of “regularised” $t$-statistics. Delmar et al. (2005) used the EM-algorithm to fit a mixture model to the distribution of gene-specific variances. The entire set of genes
was partitioned into latent groups, based on homogeneity of their variances. Essentially, genes that shared membership of the same group, shared a similar variance. A particular gene was then assigned a new variance, based on all the genes in that particular latent class. Estimating the variance in this way can be shown to have increased power over gene-specific variances, because of the large number of genes in any one class.

In Manda et al. (2007), we pursued a closely related idea to that of Delmar et al. (2005) but via a Bayesian mixture model framework. The model classifies genes according to their similarities in variance, resulting in a small number of latent classes, each containing numerous genes that share a similar variance, which can thus be estimated from a larger number of replicates than simply those per gene i.e. all of the replicates of all the genes in that latent class. The variance for a particular gene is then calculated through a weighted probabilistic assignment, that is, as a weighted average of the variances from all latent classes; one effect of this is that very large or small variances that are incompatible with the overall distribution are decreased and increased respectively. The Bayesian hierarchical approach possesses some advantages over the EM-algorithm implementation of Delmar et al. (2005): many sources of variability can be modelled simultaneously through a common framework and all unknown quantities are treated in a consistent manner to give fully probabilistic information on all unobserved variables, even their functions. Moreover, traditional methods, such as the EM-algorithm, have difficulties coping with some of the more complex non-linear models used to explore noisy biological data.

Adjusting for the false discovery rate

High dimensional datasets are an unavoidable consequence of testing tens of thousands of genes simultaneously on a single chip, which invariably leads to the problem of testing multiple hypotheses. Ignorance of the issues associated with making multiple inferences from multiple single-inference procedures will result in a increased false positive rate
Chapter 2. Analysing microarray data

(rejecting the null hypothesis when in fact it is true).

A simple approach to combating multiple testing is to make an adjustment to the \( p \)-values, such as the Bonferroni procedure. The Bonferroni approach provides an example of controlling the familywise error rate (FWER), which is defined as the probability of yielding at least one false positive out of all the hypotheses tested, simply by chance alone. If \( m \) hypothesis tests are performed, each test is controlled so that the probability of a false positive for that particular test is less than or equal to \( \alpha / m \) for some chosen value of \( \alpha \). It then follows that the overall FWER is less than or equal to \( \alpha \). More powerful Bonferroni-inspired procedures now exist (Simes, 1986; Hommel, 1988; Hochberg, 1988; Rom, 1990) which target the \( p \)-values in a similar manner. However, all these types of scheme lose substantial power to detect a specific hypothesis and become implausible when the number of tested hypotheses grows large.

A widely used alternative approach is to adjust \( p \)-values in order to control the false discovery rate (FDR), which is useful in situations where, instead of knowing the probability of obtaining at least one false positive, it would be more helpful to know the total number of false positives relative to the total number of significant items. The FDR is defined to be the false positive rate amongst all rejected (null) hypotheses; in other words, the total number of null hypotheses that are rejected when in fact true, divided by the number of rejected hypotheses. Whereas Bonferroni-type methods attempt to control the probability of making a false-positive conclusion, Benjamini and Hochberg (1995) provided a sequential \( p \)-value method to control the expected proportion of false positives.

Suppose we have a set of \( m \) independent hypotheses and the corresponding observed \( p \)-values, \( p_1 \leq p_2 \leq \ldots \leq p_m \), of which those corresponding to null hypotheses are assumed to be independent and uniformly distributed. The FDR method provides an estimate of a threshold \( t \), defined as
\[ \hat{t} = \arg \max_{1 \leq t \leq m} \{ t : p(t) \leq \alpha t/m \}. \] (2.7)

Benjamini and Hochberg (1995) proved by induction that by rejecting the null hypotheses corresponding to the \( t \)th smallest, observed \( p \)-values, the FDR, on average, is controlled under some pre-chosen level \( \alpha \). If \( \hat{t} \) does not exist, no null hypotheses are rejected.

In the context of microarray data, the multiple testing issue is complicated further due to gene coregulation, which implies that the hypotheses being tested are not independent. Benjamini et al. (1997) showed through simulation that the above described procedure will control the FDR for equally positively correlated normally distributed test statistics, however Benjamini and Yekutieli (2001) went further to show that in fact, the procedure controls the FDR in families with positively dependent test statistics; for all other cases of dependency, a small modification to the procedure can be made to control the FDR, however the resulting procedure will be more conservative.

**Example**

Consider an experiment to investigate how the gene expression levels for a particular gene vary across four progressive stages of a degenerative disease, labelled “Normal”, “Oesophagitis”, “BE” and “OA”. Measurements of gene expression levels were collected for 31, 29, 27 and 33 patients respectively (Donnellan et al., 2008). As the data were severely non-normal, two-sample pairwise Wilcoxon tests were used to assess the extent of any location differences between the disease status groups. The resulting \( p \)-values, estimates of location differences and confidence intervals for these point estimates are presented below in Table 2.1.

In this case, the negative correlation structure between the test statistics theoretically invalidates the Benjamini-Hochberg procedure (Benjamini and Yekutieli, 2001). However in practical situations where there are more pairwise contrasts than independent
Table 2.1: Table of p-values, point estimates and confidence intervals before any adjustment for multiple testing.

<table>
<thead>
<tr>
<th></th>
<th>Normal</th>
<th>Oesophagitis</th>
<th>BE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oesophagitis</td>
<td>0.027</td>
<td>3.694</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.1, 25.4)</td>
<td>-</td>
</tr>
<tr>
<td>BE</td>
<td>0.067</td>
<td>1.542</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.0, 6.3)</td>
<td>29.269</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(6.8, 152.5)</td>
</tr>
<tr>
<td>OA</td>
<td>0.471</td>
<td>-1.500</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-22.3, 2.4)</td>
<td>10.027</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.4, 128.7)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(3.4, 151.6)</td>
</tr>
</tbody>
</table>

comparisons to be made, the main problem with the Benjamini-Hochberg procedure is that it is likely to be too conservative rather than not enough so. Thus, adopting the approach of Benjamini and Yekutieli (2001) seems unnecessarily cautious.

Using the described Benjamini-Hochberg procedure with $\alpha = 0.05$, we compare each of the $p(t)$ to $0.05t/6$ for $1 \leq t \leq 6$, starting with $p(6)$ for ease. We find $i$ to be equal to two as

$$p(2) = 0.003 \leq \frac{0.05 \cdot 2}{6} = 0.017.$$  (2.8)

Therefore we reject the null hypotheses corresponding to the two smallest $p$-values.

Whilst the philosophy behind the two-sample Wilcoxon test is a comparison of the medians of two distributions $X$ and $Y$, a common misconception is that the estimate of the location difference is crudely given by $\text{median}(X) - \text{median}(Y)$. This statistic is
commonly rejected as it is possible for two distributions to have the same median value but come from substantially different distributions. A more robust estimate of the difference in location is given by the Hodges-Lehmann estimator (Hodges and Lehmann, 1963).

Suppose that we have the set of observations \(x_1, x_2, \ldots, x_n\) from the distribution \(X\) and \(y_1, y_2, \ldots, y_m\) from the distribution \(Y\). Then the Hodges-Lehmann difference estimator \(\hat{\Delta}\) is defined as

\[
\hat{\Delta} = \text{median}\{d_{i,j}\}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, m
\]  

(2.9)

where \(d_{i,j} = y_j - x_i\) and \(\{d_{i,j}\}\) is the set of all differences between every possible pair of \(x_i\) and \(y_j\).

Only in the case where the two sample distributions are different in location only, will the estimator \(\hat{\Delta}\) be equal to the difference between the sample medians, due to the differences \(d_{i,j}\) being distributed symmetrically around the location difference.

The confidence interval for a Hodges-Lehmann estimator can be calculated in a similar fashion to that of a confidence interval for a median. For large samples, the rank of the lower confidence limit can be approximated by

\[
R_l \approx \frac{mn}{2} - z_{(1-\alpha)/2} \sqrt{\frac{mn(m + n + 1)}{12}},
\]  

(2.10)

with the corresponding upper confidence limit having the rank

\[
R_u \approx mn - R_l + 1.
\]  

(2.11)

The values corresponding to \(R_l\) and \(R_u\) in the ranked list of \(d_{i,j}\) (some rounding of \(R_l\) and \(R_u\) will probably be necessary) are the confidence limits for \(\hat{\Delta}\).
Table 2.2 shows \( p \)-values after an adjustment has been made for the false discovery rate. These can simply be compared to the prechosen significance level, \( \alpha \). Aligned with the multiple testing issue, constructing 95% confidence intervals for each of the individual tests is incongruous and hence some correction is necessary. Consider the ordered \( p \)-values \( p(1) \leq p(2) \leq \ldots \leq p(m) \) with their corresponding null hypotheses \( H(i) \). For the null hypotheses \( H(i) \), we construct a \( 100(1 - \frac{i}{m}) \)% confidence interval. Doing this for all \( i = 1, \ldots, m \) will result in a familywise confidence level greater than or equal to \( 1 - \alpha \).

Table 2.2: Table of \( p \)-values, point estimates and confidence intervals after adjusting for the false discovery rate.

<table>
<thead>
<tr>
<th></th>
<th>Normal</th>
<th>Oesophagitis</th>
<th>BE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oesophagitis</td>
<td>0.054</td>
<td>3.694</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>(0.0, 29.2)</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>BE</td>
<td>0.081</td>
<td>&lt;0.001</td>
<td>29.269</td>
</tr>
<tr>
<td></td>
<td>(0.1, 6.6)</td>
<td>(4.7, 245.4)</td>
<td></td>
</tr>
<tr>
<td>OA</td>
<td>0.471</td>
<td>0.062</td>
<td>0.009</td>
</tr>
<tr>
<td></td>
<td>(-22.3, 2.4)</td>
<td>(-0.5 136.5)</td>
<td>(1.8, 227.1)</td>
</tr>
</tbody>
</table>

### 2.3.3 Mixture models

An intuitive approach to modelling gene expression data is to assume that the overall dataset consists of two smaller subsets: one of those containing the genes that are not differentially expressed and the other, those that are. Formally, the gene-specific summary measures (e.g. \( t \)-statistics or \( p \)-values) are assumed to come from a mixture distribution
Chapter 2. Analysing microarray data

with two components: one distribution corresponding to a group of differentially expressed genes and the other corresponding to a group of non-differentially expressed genes. However, a common problem with mixture models is that many rely on permutation methods to attain null distributions. This can be inefficient when the number of replicates is small, a problem that it pertinent in microarray applications.

Efron et al. (2001) first introduced this approach in a microarray context, who discussed the use of a non-parametric empirical Bayesian approach for making simultaneous inferences about particular genes from high-density oligonucleotide chips. The previously mentioned work of Lonnstedt and Speed (2002) is another example of this type of approach, as is Pan et al. (2003). Efron and Tibshirani (2002) also considered a two-component mixture as a means to model gene expression data, summarised by a rank-based non-parametric two-sample test statistic. A similar approach was adopted by Broet et al. (2002), who extended the two-component model to a fully Bayesian hierarchical model, with an unknown number of mixture components. The number of components was treated as a random variable and estimated alongside the other parameters, in the manner of Richardson and Green (1997).

A recent paper by McLachlan et al. (2006) described an alternative empirical Bayesian approach, also incorporating the two-component mixture model in a manner similar to Efron et al. (2001). Here, the adopted test statistics were transformed to corresponding \( z \)-scores of the form

\[
z_i = \Phi^{-1}(1 - p_i)
\]

where \( p_i \) is the corresponding \( p \)-value for the original test statistic for gene \( i \) and \( \Phi \) is the \( N(0, 1) \) distribution function. It was assumed that the null and non-null distributions of the \( z \)-scores each corresponded to a single normal, which they fitted with computational ease using the EM-algorithm. A limitation of this method is the use of \( p \)-values as a
basis for inference: such an approach can be limiting as useful information that is lost in forming the $p$-values may not be recovered.

Gusnanto et al. (2005) proposed a linear mixed model for describing the log-ratio of a particular gene, which incorporated a random effect component corresponding to differential expression. This random effect was assumed to follow a mixture distribution with three normal components, corresponding to elevation, reduction and no change in expression level. Non-linear shrinkage of the observations followed, with a proportion of observations being thresholded to zero.
Chapter 3

Empirical Bayesian thresholding for microarray data

3.1 Introduction

Johnstone and Silverman (2004) have recently introduced “EBayesThresh”, a highly effective empirical Bayesian technique for the detection of a few pieces of useful information ‘hidden’ in an array of noisy data, most of which contains no useful information. Following Schimek and Schmidt (2004), we explore the application of this empirical Bayesian technique to microarray data, where the problem is to detect the few ‘interesting’ genes among many ‘uninteresting’ ones. The purpose of the microarray experiment is to identify which genes have significantly different expression between different biological conditions, precisely the kind of problem that EBayesThresh is designed for.

The chapter proceeds as follows: in Section 3.2 we describe this empirical Bayesian thresholding procedure and then discuss how the methodology can be adapted for use within the context of microarray experiments in Section 3.3. The method relies on a
variance parameter $\sigma^2$. Estimation of the parameter by a marginal maximum likelihood approach reportedly leads to instability of the method. Instead we propose using a residual sum of squares approach, based on a linear additive model for replication as an alternative strategy for estimating a common variance. This approach is discussed in detail in Section 3.4. Three examples of the method in practice are illustrated in Section 3.5 and the results from our new method are compared to those obtained by various other existing approaches in the discussion in Section 3.6.

This work has recently appeared in print; the reader is directed to Walls et al. (2007).

### 3.2 Empirical Bayesian thresholding methodology

Here we describe an empirical Bayesian thresholding method, originally proposed by Johnstone and Silverman (2002, 2004, 2005), for the estimation of sparse sequences corrupted by normal noise. The term sparse sequence is meant in this context as consisting of a few large (positive or negative) signal values interspersed amongst many zero values, all subject to corruption by stochastic noise, which Johnstone and Silverman assumed to be normally distributed with unit variance. In this chapter, we extend the work to include noise models with variance $\sigma^2$, whilst alternative distributions are considered for modelling the noise in Chapter 4. The aim of the method described is to identify which elements of the sequence are zero values corrupted by noise and which are the true signal values. This is done by estimating a threshold value: any observations with an absolute value below the threshold are set to zero and those observations with an absolute value above the threshold value are said to be the true non-zero signal values. Since the method is adaptive, it is designed to deal with varying degrees of sparsity. Johnstone and Silverman (2005) consider signals with a wide range of sparse behaviour, from very sparse sequences containing as few as 0.5% non-zero values through to less sparse sequences containing 50% non-zero values.
The method is described below; in Section 3.2.1 we develop the theory for one observation before considering multiple observations and parameter estimation in Section 3.2.2.

3.2.1 The single observation case

Consider a single observation $Z$ which can be modelled by the form

$$Z = \mu + \epsilon,$$  \hspace{1cm} (3.1)

where $\mu$, an unknown mean, follows some prior density and $\epsilon$ is a normally distributed $N(0, \sigma^2)$ noise element. The method is designed for the estimation of sparse sequences, hence an assumption made \textit{a priori} is that there is a strong chance that $\mu = 0$. This feature is incorporated into the prior model through a mixture “density”

$$f_{\text{prior}}(\mu) = (1 - \omega)\delta_0(\mu) + \omega \gamma(\mu).$$ \hspace{1cm} (3.2)

where $\omega$ is a mixing weight, $0 \leq \omega \leq 1$; the $\delta_0$ component is a point mass at zero (of course, the presence of $\delta_0$ means that the above is not a true density); and the non-zero part, $\gamma$ is assumed to be a fixed, unimodal, symmetric density. Thus, with probability $(1 - \omega)$, $\mu$ is zero and with probability $\omega$, $\mu$ comes from the density $\gamma$.

For the choice of $\gamma$, we use a Laplace density (a double exponential distribution) with scaling parameter $a > 0$

$$\gamma_a(u) = \frac{1}{2a} \exp(-a|u|).$$ \hspace{1cm} (3.3)

We use this choice for $\gamma$ because of its heavier-than-normal tails and because it leads to tractable calculations.
Assuming that $\omega$, $a$ and $\sigma^2_\epsilon$ are known, it can be shown that the posterior distribution for $\mu$ conditional on $Z = z$ takes the form

$$f_{\text{post}}(\mu|Z = z) = (1 - \omega_{\text{post}})\delta(\mu) + \omega_{\text{post}}f_1(\mu|z),$$

(3.4)

where

$$f_1(\mu|Z = z) = f(\mu|Z = z, \mu \neq 0)$$

and can be derived explicitly as

$$f_1(\mu|Z = z) = \begin{cases} \frac{e^{az}\phi((\mu-z)/\sigma_\epsilon-a\sigma_\epsilon)}{\sigma_\epsilon\{e^{-a^2}\Phi(z/\sigma_\epsilon-a\sigma_\epsilon)+e^{a^2}\Phi(z/\sigma_\epsilon+a\sigma_\epsilon)\}} & \text{if } \mu < 0 \\ \frac{e^{az}\phi((\mu-z)/\sigma_\epsilon+a\sigma_\epsilon)}{\sigma_\epsilon\{e^{-a^2}\Phi(z/\sigma_\epsilon-a\sigma_\epsilon)+e^{a^2}\Phi(z/\sigma_\epsilon+a\sigma_\epsilon)\}} & \text{if } \mu > 0. \end{cases}$$

In particular,

$$\omega_{\text{post}}(z) = P(\mu \neq 0|Z = z) = \frac{1 + \beta(z)}{\omega^{-1} + \beta(z)}$$

(3.5)

in terms of

$$\beta(z) = \beta(z, a) = \frac{a\sigma_\epsilon}{2} \left\{ \frac{\Phi(z/\sigma_\epsilon-a\sigma_\epsilon)}{\phi(z/\sigma_\epsilon-a\sigma_\epsilon)} + \frac{\Phi(z/\sigma_\epsilon+a\sigma_\epsilon)}{\phi(z/\sigma_\epsilon+a\sigma_\epsilon)} \right\} - 1$$

where $\Phi(\cdot)$ is the cdf of the standard normal distribution and $\Phi(\cdot) = 1 - \Phi(\cdot)$. If the value of $\sigma_\epsilon$ becomes very small, the value of $\beta(z/\sigma_\epsilon)$ will be numerically infinite. Theoretically, this should not pose a problem for subsequent calculations. However, computational problems with the method may be encountered occasionally when this particular situation arises.
Let $\hat{\mu}(z;\omega,a,\sigma_\epsilon)$ be the posterior median of the distribution (3.4) for $\mu$ given $Z = z$ (from now on denoted simply by $\hat{\mu}(z;\omega)$). For a fixed $\omega \in (0,1)$, the function $\hat{\mu}(z;\omega)$ will be a monotonic function of $z$ with a thresholding property (seen in Figure 3.1), in that there exists $t(\omega) > 0$ such that $\hat{\mu}(z;\omega) = 0$ if and only if $|z| \leq t(\omega)$. An observation $Z$ yields a non-zero $\hat{\mu}$ if $|Z|$ exceeds the threshold. Otherwise the observation $Z$ yields a zero value for $\hat{\mu}$. In the application presented here, we use the posterior median thresholding rule, where $\hat{\mu}(z;\omega)$ is taken as the estimate of $\mu$. Depending on the particular application, other thresholding rules may be preferred. For example, one alternative possibility would be to use the posterior mean of $\mu$ given $Z = z$. In the context of the problem described here, the posterior median seems more appropriate because, whilst the posterior mean rule implements a certain degree of shrinkage, it does not possess the same attractive thresholding property that the posterior median does, and will inevitably produce estimates of the zero $\mu$ which are close to zero, yet irrefutably non-zero.

To find the posterior median $\hat{\mu}(z;\omega)$, suppose that $z > 0$ and let

$$
\tilde{F}_1(\mu|z) = \int_0^\infty f_1(u|z)du
= \frac{\exp(-az)\Phi((\mu - z)/\sigma_\epsilon + a\sigma_\epsilon)}{\exp(-az)\Phi(z/\sigma_\epsilon - a\sigma_\epsilon) + \exp(az)\Phi(z/\sigma_\epsilon + a\sigma_\epsilon)} \quad \text{for } \mu \geq 0.
$$

Then $\hat{\mu}(z;\omega)$ can be obtained from the following properties

$$
\hat{\mu}(z;\omega) = 0 \quad \text{if } \omega_{\text{post}}(z)\tilde{F}_1(0|z) \leq \frac{1}{2}

\tilde{F}_1(\hat{\mu}(z;\omega)|z) = \{2\omega_{\text{post}}(z)\}^{-1} \quad \text{otherwise.}
$$

If $\omega_{\text{post}}(z) \leq \frac{1}{2}$ then the median will be necessarily zero and evaluating $\tilde{F}_1(0|z)$ is unnecessary. In the case where $z < 0$, the antisymmetry property $\hat{\mu}(-z;\omega) = -\hat{\mu}(z;\omega)$ can be used. The posterior median estimate of $\mu$ can be derived explicitly to be
\[ \hat{\mu} = \max[0, z - a\sigma^2 - \sigma\Phi^{-1}\{\min(1, y_0)\}] \]

where

\[ y_0 = a^{-1}\sigma^{-1}\phi(z/\sigma - a\sigma)(w^{-1} + \beta(z)). \]

Ideally, a threshold should adapt to the sparsity of the signal to be estimated, in that the threshold should increase as the signal becomes more sparse. The empirical Bayesian method described automatically encompasses this adaptive property: a small mixing weight \( \omega \) will correspond to a large threshold \( t(\omega) \) and conversely, a larger mixing weight will yield a smaller threshold. Figure 3.1 shows the resulting posterior median functions for two different values of \( \omega \) with the scaling parameter \( a = 0.5 \) and \( \sigma = 1 \). In the left panel, \( \omega = 0.01 \) and the threshold is chosen to be 3.68. Clearly a large range of observed \( Z \) values would yield a zero estimate of \( \mu \). Furthermore, the posterior median function possesses a shrinkage property in that those \( Z \) estimated to be non-zero are shrunk non-linearly towards zero. In the right panel, \( \omega = 0.45 \) which corresponds to a threshold of 1.80. Clearly illustrated, a smaller range of observed \( Z \) values would yield a zero estimate of \( \mu \) and the data are not shrunk so severely in this case.

3.2.2 Parameter estimation

Now suppose that we have a sequence of \( n \) i.i.d. observations \( Z_i \) (\( i = 1, \ldots, n \)), each from a distribution with a unique mean parameter \( \mu_i \) and subject to normal noise. The model (3.1) can be extended accordingly and written as

\[ Z_i = \mu_i + \epsilon_i, \quad (3.6) \]
Figure 3.1: The solid line is the posterior median function for a Laplace prior with mixing weight \( \omega = 0.01 \) (left) and \( \omega = 0.45 \) (right). The dotted line in each plot is a horizontal line at zero; the dashed line is the \( y = x \) diagonal line and the dot-dashed vertical line illustrates the position of the estimated threshold value.

where the \( \mu_i \) are independent means from the prior distribution (3.2) and the \( \epsilon_i \) are independent \( N(0, \sigma^2) \) random variables.

Clearly, there are three key parameters to estimate: \( a, \sigma^2_\epsilon \) and \( \omega \). Here we use a constrained form of maximum likelihood for estimating \( \omega \) and \( a \) and discuss possible alternatives for the estimation of \( \sigma^2_\epsilon \). Given particular values for these parameters and the sequence of \( n \) observations, the Bayesian procedure can be applied to each observation \( Z_i \), resulting in an estimate of the corresponding \( \mu_i \).

Let \( g_a \) be the convolution of \( \gamma_a(\cdot) \) with the normal density \( N(0, \sigma^2_\epsilon) \), which takes the explicit form

\[
g_a(z) = \frac{a}{2} \exp \left( \frac{(a\sigma_\epsilon)^2}{2} \right) \left\{ \exp(-az)\Phi(z/\sigma_\epsilon - a\sigma_\epsilon) + \exp(az)\tilde{\Phi}(z/\sigma_\epsilon + a\sigma_\epsilon) \right\}.
\]  

(3.7)
The marginal density of the observations $Z_i$ can then be written as

$$\frac{(1 - \omega)}{\sigma_e} \phi(z/\sigma_e) + \omega g_a(z).$$

The marginal maximum likelihood estimators of $\omega$ and $a$, denoted $\hat{\omega}$ and $\hat{a}$ respectively, are estimated by finding the maximum over both parameters of the marginal log-likelihood

$$\ell(\omega, a) = \sum_{i=1}^{n} \log \left( \frac{(1 - \omega)}{\sigma_e} \phi(Z_i/\sigma_e) + \omega g_a(Z_i) \right),$$

subject to the constraint on $\omega$ that the threshold satisfies $t(\omega) \leq \sigma_e \sqrt{2 \log n}$, where $\sigma_e \sqrt{2 \log n}$ is commonly known as the universal threshold of Donoho and Johnstone (1994). We apply this constraint since Donoho and Johnstone showed that in cases where the universal threshold is employed, there will be a high probability of every zero signal value being correctly estimated when $n$ is large. Thus in the case of the empirical Bayesian procedure, there is no need to consider those thresholds any greater than the universal threshold.

Having used the data to estimate $\hat{\omega}$ and $\hat{a}$, these estimated values are substituted back into the prior distribution (3.2) and the means $\mu_i$ estimated.

In the absence of replicates, Johnstone and Silverman (2004) chose to estimate $\sigma_e$ by the median absolute deviation of the $Z_i$ from zero. This can be defined as

$$\sigma_e = \text{mad}(z) = k \times \text{median}(|z|), \quad \text{for } z = (Z_1, \ldots Z_n)$$

(3.8)

where $k = 1.4826$ is chosen to ensure that $\text{mad}(z)$ is an unbiased estimate of $\sigma_e$ when $\omega = 0$. Providing that the sequence $\mu_i$ is reasonably sparse (i.e. $\omega$ is not too far from zero), the median of the absolute deviations should not be very affected by those outlying
observations with non-zero means. How the behaviour of the $\text{mad}(z)$ estimator changes with the sparseness of the sequence is examined more closely in Section 3.4.1.

A distinct advantage of microarray data over the wavelet context, for which the method was originally described, is the presence of replicate information in many examples. Even a limited amount of replication should allow a more reliable estimate of $\sigma^2$. We propose a linear additive model and use an approach based on the residual sum of squares to estimate the components of variance common to all genes. This additive model is described in greater detail in Section 3.4.

The majority of the described empirical Bayesian methodology has been implemented in the package `EbayesThresh` (Johnstone and Silverman, 2005) for the R project, which can be found at http://CRAN.R-project.org/. However, this package only allows for estimation of $\sigma$ by $\text{mad}(z)$; other methods preferred by the user must be implemented separately from the existing software.

### 3.3 Applying the empirical Bayesian method to microarray data

An empirical Bayesian approach is very attractive in the microarray context as we do have some genuine prior information, which is captured intuitively through the use of the 2-component mixture model. It is expected that the majority of genes on the chip will not be differentially expressed between the two samples and will thus have a contrast value of zero; those genes which are substantially differentially expressed will have large (positive or negative) contrast values and be randomly interspersed amongst the near zero values. Given this, it is plausible to interpret this type of dataset as a sparse sequence containing a few large signal values interspersed amongst many zero values, all corrupted by noise. The high dimensionality of microarray data and the limited number of replicates
means that the information per gene is relatively low. The empirical Bayesian method allows the evaluation of the measurement for a particular gene to be influenced by the overall measurements by “borrowing” information across genes through the combination of related inference problems.

Johnstone and Silverman (2004) suggested that some mild correlation between the observations is acceptable; there will be some loss of information in the estimation procedure though reasonable results will be achieved. We investigate further to what extent the correlation can still be considered “mild” in Section 3.4.1, however it should diminish somewhat the problem of gene coregulation and correlated test statistics. Furthermore, this type of modelling approach neatly circumvents any issues regarding multiple testing.

Consider the sequence of observed $X_{ijk}$ values from equation (2.5). In order to reach a conclusion on each gene, some combination of the replicated information has to occur, either before or after thresholding. Whilst thresholding all of the replicated data is possible, we do not recommend this approach. Suppose that as many as $mr - 1$ of the replicates for a particular gene $i$ were thresholded to zero. The inclusion of the one non-zero measure would result in a non-zero mean for the gene, thus defeating the purpose of adopting the posterior median thresholding rule. Thus, in this context, combining the data before thresholding is preferred. Averaging over all the $mr$ replicates, we obtain a mean contrast $\bar{X}_i$ for each gene $i$. Taking these $\bar{X}_i$ as the $Z_i$ values described in Section 3.2, we can apply the empirical Bayesian methodology of Sections 3.2.1 and 3.2.2 to obtain an estimate of the true contrast for each gene $i$.

3.4 Linear additive models for the estimation of $\sigma^2_e$

Here we construct a model for the normalised data for the purpose of estimating $\sigma^2_e$. The model is built on the assumption that whilst the normalisation procedures have
standardised the chips and made them relatively comparable, some local effects and anomalies from chip to chip may have not been removed. These could be due to various sources, such as defects on the chip, (for example, scratches, hair or fingerprints), saturated spots, or simply very small or very large observations with no explanation. We account for two sources of variation: variance between the spots for a particular gene within a chip and variance between the chips.

When there are multiple levels of replication present, it is natural to consider a hierarchical model to provide the necessary structure. Thus, our proposed model is a relatively simple hierarchical linear additive model, similar to ideas discussed by Wernisch et al. (2003). In any analysis of variance, a key assumption is that the data have some constant variance across all observations. Therefore, a variance stabilising transformation, such as that incorporated into the \texttt{vsn} normalisation, is highly useful here.

Consider $X_{ijk}$, the contrast for the $i$th gene ($i = 1, \ldots, n$) and $k$th spot or replicate ($k = 1, \ldots, r$) nested within the $j$th chip ($j = 1, \ldots, m$). Then, conditional on the $\mu_i$,

$$X_{ijk} = \mu_i + B_j + \epsilon_{k(ij)}, \quad (3.9)$$

where $\mu_i$ is the gene effect for gene $i$ from the density (3.2); the term $B_j \sim N(0, \sigma^2_B)$ is a random effect and represents the “between” chip variability; the term $\epsilon_{k(ij)} \sim N(0, \sigma^2_W)$ represents the “within” chip variability. Clearly, if the normalisation scheme employed has worked successfully, the “between” chip variation should not be significant. Further to this, the use of the contrast variable, which is essentially a difference, will eliminate this $B_j$ term. However, we keep the the term in the model to be cautious and for illustrative purposes. Thus the variance associated with each $X_{ijk}$ results partly from the variation between the spot replicates and partly from the additional variation between the chips. Adopting a suitable normalisation practice allows global estimates of $\sigma^2_B$ and $\sigma^2_W$ to be made from the whole dataset leading to an increase in robustness. Given the components
of variance for the model (3.9), the variance $\sigma^2_\epsilon$ associated with each $\bar{X}_i$. can be estimated by

$$\hat{\sigma}^2_\epsilon = \frac{\hat{\sigma}^2_B}{m} + \frac{\hat{\sigma}^2_W}{mr}. \quad (3.10)$$

Clearly, unless each gene has an identical number of replicates, unequal estimates of $\hat{\sigma}^2_\epsilon$ will occur. Hence to ensure a common distribution for each observation, some standardisation of the data will be required before applying the empirical Bayesian thresholding procedure. Quite simply, each observation can be standardised by dividing by its corresponding $\hat{\sigma}_\epsilon$ and assuming a common variance of 1 from there onwards. However, one should note that it will be necessary to multiply the resulting estimates of the $\mu_i$ by $\hat{\sigma}_\epsilon$ after the procedure.

### 3.4.1 Some simulation studies

The first simulation presented here was designed to investigate the claim made by Johnstone and Silverman (2004) that some mild correlation between the observations does not affect the efficiency of the empirical Bayesian procedure. We created a sequence $\mu$ of length $n$, which has $\mu_i = 0$ except in $K$ randomly chosen positions, where it instead takes a non-zero signal value $\mu_i$. The non-zero values reported were 3, 4 and 5. We fixed $n$ to be equal to 1000 and investigated $K = 50, 100, 250$. We also generated a vector of 1000 noise values from an AR(1) process, varying the value of the correlation coefficient $\rho$. A sequence of “observations” was generated by combining the signal and noise vectors. The empirical Bayesian procedure was used to threshold the observations to estimate the true underlying sequence $\mu$. One hundred replications were performed for each of the non-zero signal values and $K$ and for each replication, the total squared error of the estimation $\sum(\hat{\mu}_i - \mu_i)^2$, and the numbers of misclassified signal and noise entries were recorded. Misclassified signal describes true non-zero $\mu_i$ that are incorrectly
concluded to be zero by the thresholding procedure; misclassified noise describes truly zero $\mu_i$ that are incorrectly concluded to be non-zero by the thresholding procedure. The results given in Tables 3.1 to 3.3 are an average over these 100 replications. The same 100,000 noise variables were used for each set of the replications.

The numbers presented in Tables 3.1 to 3.3 suggest that the empirical Bayesian method is resilient to mild correlation between observations, particularly when the signal is strong and sparse. It is clear to see that increasing the value of the correlation coefficient results in an increased proportion of misclassified noise. In a sweeping generalisation, we suggest that the method will tolerate observed correlation values of up to about 0.25 without impacting substantially on the results.

The relative benefits of using the linear additive model introduced in Section 3.4 over the median absolute deviation of $\{\bar{X}_i\}$ for estimating the noise level are illustrated in the second simulation study. Data were generated from the model (3.6) to emulate a simplistic version of noisy microarray data. Motivated by the example we shall study in Section 3.5.2, we consider a small simulated microarray experiment examining $n = 76$ genes with $r = 6$ spot replicates of each gene per chip, and $m = 3$ chips. Gene effects were generated from prior distribution (3.2) with the non-zero part of the prior being the Laplace density (3.3) with a scaling parameter $a = 0.5$. Parameters were $\sigma_B = 0.09$ and $\sigma_W = 0.9$. Fifty simulated datasets were generated for each of a range of values of $\omega$ and for each dataset, the standard deviation $\sigma_e$ was estimated by two different methods: one approach was to use equation (3.10) with the estimated variance components obtained from the linear additive model (3.9); the other was to calculate $\text{mad}(\bar{X}_i)$ (we found that the inclusion of any replicate information had very little or no effect on the estimating power of the $\text{mad}$ function (3.8)). Average values of $\hat{\sigma}_e - \sigma_e$ are shown in Figure 3.2 for both estimators. The results suggest that using the approach based on the residual sum of squares from the linear additive model consistently gives rise to a good estimate of $\sigma_e$, regardless of the proportion of differentially expressed genes. Using
Table 3.1: Results from a simulation study to test the efficiency of the empirical Bayesian thresholding procedure in the presence of correlated observations.

<table>
<thead>
<tr>
<th>Signal sparsity (out of 1000)</th>
<th>Spiked-in values</th>
<th>Correlation coefficient ρ</th>
<th>Total squared error</th>
<th>Misclassified signal</th>
<th>Misclassified noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>3</td>
<td>0.0</td>
<td>216.86</td>
<td>14.96</td>
<td>14.93</td>
</tr>
<tr>
<td>50</td>
<td>3</td>
<td>0.05</td>
<td>220.47</td>
<td>15.43</td>
<td>15.25</td>
</tr>
<tr>
<td>50</td>
<td>3</td>
<td>0.1</td>
<td>215.20</td>
<td>14.36</td>
<td>17.04</td>
</tr>
<tr>
<td>50</td>
<td>3</td>
<td>0.15</td>
<td>216.71</td>
<td>14.41</td>
<td>18.33</td>
</tr>
<tr>
<td>50</td>
<td>3</td>
<td>0.2</td>
<td>213.80</td>
<td>13.34</td>
<td>22.42</td>
</tr>
<tr>
<td>50</td>
<td>3</td>
<td>0.25</td>
<td>218.49</td>
<td>12.97</td>
<td>29.81</td>
</tr>
<tr>
<td>50</td>
<td>3</td>
<td>0.5</td>
<td>275.87</td>
<td>4.38</td>
<td>305.91</td>
</tr>
<tr>
<td>50</td>
<td>3</td>
<td>0.75</td>
<td>852.42</td>
<td>0.01</td>
<td>947.35</td>
</tr>
<tr>
<td>50</td>
<td>4</td>
<td>0.0</td>
<td>157.72</td>
<td>3.61</td>
<td>11.24</td>
</tr>
<tr>
<td>50</td>
<td>4</td>
<td>0.05</td>
<td>162.77</td>
<td>4.05</td>
<td>11.06</td>
</tr>
<tr>
<td>50</td>
<td>4</td>
<td>0.1</td>
<td>162.98</td>
<td>4.03</td>
<td>11.72</td>
</tr>
<tr>
<td>50</td>
<td>4</td>
<td>0.15</td>
<td>159.99</td>
<td>3.66</td>
<td>13.06</td>
</tr>
<tr>
<td>50</td>
<td>4</td>
<td>0.2</td>
<td>163.67</td>
<td>3.6</td>
<td>15.06</td>
</tr>
<tr>
<td>50</td>
<td>4</td>
<td>0.25</td>
<td>173.81</td>
<td>3.65</td>
<td>17.64</td>
</tr>
<tr>
<td>50</td>
<td>4</td>
<td>0.5</td>
<td>269.18</td>
<td>2.41</td>
<td>91.15</td>
</tr>
<tr>
<td>50</td>
<td>4</td>
<td>0.75</td>
<td>930.05</td>
<td>0.1</td>
<td>886</td>
</tr>
<tr>
<td>50</td>
<td>5</td>
<td>0.0</td>
<td>101.98</td>
<td>0.48</td>
<td>7.73</td>
</tr>
<tr>
<td>50</td>
<td>5</td>
<td>0.05</td>
<td>103.00</td>
<td>0.46</td>
<td>7.69</td>
</tr>
<tr>
<td>50</td>
<td>5</td>
<td>0.1</td>
<td>104.95</td>
<td>0.44</td>
<td>8.1</td>
</tr>
<tr>
<td>50</td>
<td>5</td>
<td>0.15</td>
<td>106.17</td>
<td>0.43</td>
<td>8.9</td>
</tr>
<tr>
<td>50</td>
<td>5</td>
<td>0.2</td>
<td>110.77</td>
<td>0.41</td>
<td>9.99</td>
</tr>
<tr>
<td>50</td>
<td>5</td>
<td>0.25</td>
<td>119.15</td>
<td>0.46</td>
<td>11.54</td>
</tr>
<tr>
<td>50</td>
<td>5</td>
<td>0.5</td>
<td>227.33</td>
<td>0.49</td>
<td>48.39</td>
</tr>
<tr>
<td>50</td>
<td>5</td>
<td>0.75</td>
<td>946.33</td>
<td>0.09</td>
<td>645.84</td>
</tr>
</tbody>
</table>
Table 3.2: Results from a simulation study to test the efficiency of the empirical Bayesian thresholding procedure in the presence of correlated observations.

<table>
<thead>
<tr>
<th>Signal sparsity (out of 1000)</th>
<th>Spiked-in values</th>
<th>Correlation coefficient $\rho$</th>
<th>Total squared error</th>
<th>Misclassified signal</th>
<th>Misclassified noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>3</td>
<td>0.0</td>
<td>349.81</td>
<td>17.02</td>
<td>54.14</td>
</tr>
<tr>
<td>100</td>
<td>3</td>
<td>0.05</td>
<td>372.86</td>
<td>20.94</td>
<td>46.46</td>
</tr>
<tr>
<td>100</td>
<td>3</td>
<td>0.1</td>
<td>356.84</td>
<td>17.99</td>
<td>51.06</td>
</tr>
<tr>
<td>100</td>
<td>3</td>
<td>0.15</td>
<td>354.05</td>
<td>16.6</td>
<td>60.41</td>
</tr>
<tr>
<td>100</td>
<td>3</td>
<td>0.2</td>
<td>372.62</td>
<td>18.55</td>
<td>66.67</td>
</tr>
<tr>
<td>100</td>
<td>3</td>
<td>0.25</td>
<td>342.38</td>
<td>12.49</td>
<td>78.15</td>
</tr>
<tr>
<td>100</td>
<td>3</td>
<td>0.5</td>
<td>431.68</td>
<td>6.85</td>
<td>308.64</td>
</tr>
<tr>
<td>100</td>
<td>3</td>
<td>0.75</td>
<td>1021.60</td>
<td>0.07</td>
<td>895.57</td>
</tr>
<tr>
<td>100</td>
<td>4</td>
<td>0.0</td>
<td>246.33</td>
<td>2.94</td>
<td>30.09</td>
</tr>
<tr>
<td>100</td>
<td>4</td>
<td>0.05</td>
<td>250.99</td>
<td>3.36</td>
<td>28.55</td>
</tr>
<tr>
<td>100</td>
<td>4</td>
<td>0.1</td>
<td>253.86</td>
<td>3.4</td>
<td>29.59</td>
</tr>
<tr>
<td>100</td>
<td>4</td>
<td>0.15</td>
<td>255.31</td>
<td>2.91</td>
<td>33.09</td>
</tr>
<tr>
<td>100</td>
<td>4</td>
<td>0.2</td>
<td>262.28</td>
<td>2.81</td>
<td>36.17</td>
</tr>
<tr>
<td>100</td>
<td>4</td>
<td>0.25</td>
<td>274.21</td>
<td>3.36</td>
<td>39.04</td>
</tr>
<tr>
<td>100</td>
<td>4</td>
<td>0.5</td>
<td>409.59</td>
<td>2.91</td>
<td>118.41</td>
</tr>
<tr>
<td>100</td>
<td>4</td>
<td>0.75</td>
<td>1117.49</td>
<td>0.5</td>
<td>756.34</td>
</tr>
<tr>
<td>100</td>
<td>5</td>
<td>0.0</td>
<td>176.56</td>
<td>0.4</td>
<td>18.69</td>
</tr>
<tr>
<td>100</td>
<td>5</td>
<td>0.05</td>
<td>179.42</td>
<td>0.4</td>
<td>17.85</td>
</tr>
<tr>
<td>100</td>
<td>5</td>
<td>0.1</td>
<td>182.43</td>
<td>0.39</td>
<td>18.34</td>
</tr>
<tr>
<td>100</td>
<td>5</td>
<td>0.15</td>
<td>186.87</td>
<td>0.37</td>
<td>19.86</td>
</tr>
<tr>
<td>100</td>
<td>5</td>
<td>0.2</td>
<td>194.29</td>
<td>0.43</td>
<td>21.66</td>
</tr>
<tr>
<td>100</td>
<td>5</td>
<td>0.25</td>
<td>205.81</td>
<td>0.55</td>
<td>23.58</td>
</tr>
<tr>
<td>100</td>
<td>5</td>
<td>0.5</td>
<td>346.90</td>
<td>0.57</td>
<td>66.2</td>
</tr>
<tr>
<td>100</td>
<td>5</td>
<td>0.75</td>
<td>1110.29</td>
<td>0.37</td>
<td>481.31</td>
</tr>
</tbody>
</table>
Table 3.3: Results from a simulation study to test the efficiency of the empirical Bayesian thresholding procedure in the presence of correlated observations.

<table>
<thead>
<tr>
<th>Signal sparsity (out of 1000)</th>
<th>Spiked-in values</th>
<th>Correlation coefficient $\rho$</th>
<th>Total squared error</th>
<th>Misclassified signal</th>
<th>Misclassified noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>250</td>
<td>3</td>
<td>0.0</td>
<td>582.14</td>
<td>2.29</td>
<td>404.29</td>
</tr>
<tr>
<td>250</td>
<td>3</td>
<td>0.05</td>
<td>588.55</td>
<td>2.37</td>
<td>424.95</td>
</tr>
<tr>
<td>250</td>
<td>3</td>
<td>0.1</td>
<td>592.84</td>
<td>2.35</td>
<td>432.58</td>
</tr>
<tr>
<td>250</td>
<td>3</td>
<td>0.15</td>
<td>588.31</td>
<td>2.27</td>
<td>427.25</td>
</tr>
<tr>
<td>250</td>
<td>3</td>
<td>0.2</td>
<td>598.64</td>
<td>2.19</td>
<td>444.72</td>
</tr>
<tr>
<td>250</td>
<td>3</td>
<td>0.25</td>
<td>615.53</td>
<td>2.29</td>
<td>465.26</td>
</tr>
<tr>
<td>250</td>
<td>3</td>
<td>0.5</td>
<td>787.60</td>
<td>1.44</td>
<td>657.71</td>
</tr>
<tr>
<td>250</td>
<td>3</td>
<td>0.75</td>
<td>1409.17</td>
<td>0</td>
<td>750</td>
</tr>
<tr>
<td>250</td>
<td>4</td>
<td>0.0</td>
<td>488.13</td>
<td>0.97</td>
<td>135.94</td>
</tr>
<tr>
<td>250</td>
<td>4</td>
<td>0.05</td>
<td>494.20</td>
<td>1.06</td>
<td>141.95</td>
</tr>
<tr>
<td>250</td>
<td>4</td>
<td>0.1</td>
<td>499.62</td>
<td>1.06</td>
<td>145.09</td>
</tr>
<tr>
<td>250</td>
<td>4</td>
<td>0.15</td>
<td>498.56</td>
<td>1.02</td>
<td>145.14</td>
</tr>
<tr>
<td>250</td>
<td>4</td>
<td>0.2</td>
<td>511.61</td>
<td>1.03</td>
<td>153.56</td>
</tr>
<tr>
<td>250</td>
<td>4</td>
<td>0.25</td>
<td>533.14</td>
<td>1.03</td>
<td>163.77</td>
</tr>
<tr>
<td>250</td>
<td>4</td>
<td>0.5</td>
<td>753.34</td>
<td>1.3</td>
<td>303.1</td>
</tr>
<tr>
<td>250</td>
<td>4</td>
<td>0.75</td>
<td>1558.24</td>
<td>0.34</td>
<td>716.76</td>
</tr>
<tr>
<td>250</td>
<td>5</td>
<td>0.0</td>
<td>399.76</td>
<td>0.2</td>
<td>62.98</td>
</tr>
<tr>
<td>250</td>
<td>5</td>
<td>0.05</td>
<td>403.24</td>
<td>0.12</td>
<td>66.55</td>
</tr>
<tr>
<td>250</td>
<td>5</td>
<td>0.1</td>
<td>409.23</td>
<td>0.14</td>
<td>69.53</td>
</tr>
<tr>
<td>250</td>
<td>5</td>
<td>0.15</td>
<td>411.03</td>
<td>0.12</td>
<td>69.79</td>
</tr>
<tr>
<td>250</td>
<td>5</td>
<td>0.2</td>
<td>424.06</td>
<td>0.11</td>
<td>74.08</td>
</tr>
<tr>
<td>250</td>
<td>5</td>
<td>0.25</td>
<td>443.54</td>
<td>0.19</td>
<td>78.86</td>
</tr>
<tr>
<td>250</td>
<td>5</td>
<td>0.5</td>
<td>660.34</td>
<td>0.39</td>
<td>151.26</td>
</tr>
<tr>
<td>250</td>
<td>5</td>
<td>0.75</td>
<td>1545.29</td>
<td>0.49</td>
<td>502.76</td>
</tr>
</tbody>
</table>
\textit{mad}(\bar{X}_i)\) is a relatively good estimator when the proportion of differentially expressed genes is small. As \(\omega\) increases, the estimate increasingly worsens and the mixed effects estimate consistently outperforms it. The limitation of the \textit{mad} estimator could be highly influential when experiments are conducted with “homemade” chips, common practice in academic circles due to much lower associated costs, or those especially designed for particular experiments, such as the Lymphochip example featured in Section 3.5.3. These types of chip are designed to answer a specific question and usually contain genes chosen using prior information, i.e. those known to be of some particular importance. Consequently they are more likely to contain a high proportion of genes of interest. Furthermore, “homemade” chips tend to contain far fewer gene probes, often in the range of hundreds rather than thousands, and even a small number of differentially expressed genes could be a relatively large overall proportion of the total.

With regard to the empirical Bayesian procedure, an overinflated estimate of the standard deviation \(\sigma_e\) will result in a smaller estimate of \(\omega\) and a larger corresponding threshold value. Thus, a greater number of genes will be thresholded to zero than necessarily should be. Similarly, an underestimate of the variance will result in too few observations being set to zero. Clearly, the estimation of \(\sigma_e\) is crucial to the success of the procedure.

\subsection*{3.4.2 Marginal maximum likelihood estimation for \(\sigma_e\)}

Whilst Johnstone and Silverman (2004) claimed that estimating the \(\sigma_e\) parameter via the marginal maximum likelihood approach could lead to instability in the optimisation procedure and a lack of convergence (by personal communication with Bernard Silverman), for completeness, we investigated and found this approach to work efficiently for our datasets. The resulting estimates are given in Table 3.4 and are similar to those obtained by the \textit{mad} estimator and through the linear additive modelling approach of Section 3.4.
Figure 3.2: Simulation to compare the estimation power of the default $\text{mad}(z)$ compared with the approach described in Section 3.4 over an increasing range of $\omega$. Here we give $\hat{\sigma}_e - \sigma_e$, where $\hat{\sigma}_e$ is estimated by $\text{mad}(\bar{X}_i)$ (dashed line) and equation (3.10) with variance components estimated by the residual sum of squares from model (3.9) (dotted line).

### 3.5 Examples

Each of the examples presented below are analogous to the situation described in Section 2.3.1, despite certain idiosyncrasies specific to each dataset. For each example, the data are manipulated to obtain a sequence of $Z_i$ values (a mean contrast value for each gene, averaged over spot and chip replicates as appropriate) as described in Section 3.3, to which the empirical Bayesian thresholding procedure is applied. To briefly reiterate, we make the prior assumption that the observations $Z_i$ come from a mixture distribution with mixing weight $\omega$, consisting of a point mass at zero and a Laplace distribution with a scaling parameter $a$. Furthermore, we assume that the observations have been corrupted by normally distributed noise with standard deviation $\sigma_e$. The standard deviation of the noise $\sigma_e$ was estimated using the residual sum of squares approach described in
Table 3.4: Comparison of $\hat{\sigma}_\epsilon$ values, obtained by various estimation methods.

<table>
<thead>
<tr>
<th>Method employed</th>
<th>HIV data</th>
<th>E.Coli data</th>
</tr>
</thead>
<tbody>
<tr>
<td>mad($x$)</td>
<td>0.140</td>
<td>0.342</td>
</tr>
<tr>
<td>linear additive model 3.4</td>
<td>0.121</td>
<td>0.249</td>
</tr>
<tr>
<td>MMLE over means</td>
<td>0.158</td>
<td>0.235</td>
</tr>
</tbody>
</table>

Section 3.4 with equation (3.10), whilst $\omega$ and $a$ are estimated by a marginal maximum likelihood approach. Given estimates for the parameters, we estimate the true $\mu_i$ by the median of the posterior distribution. This approach is attractive as it essentially “thresholds” the data: there exists some value $t(\omega)$ below which the posterior median is definitively zero.

### 3.5.1 HIV data

In this case, the sum of squares approach gave a standard deviation estimate $\hat{\sigma}_\epsilon = 0.121$. This is comparable to the mad estimator which gives $\hat{\sigma}_\epsilon = 0.140$. Using either of these estimates for the standard deviation results in an infinite value for $\beta(z)$ and the method breaks down. This problem could occur frequently when we recall that an aim of normalisation is to make the replicate observations comparable across multiple arrays. Hence, it is possible that when using this normalised information to estimate the variability, very small estimates for $\sigma_\epsilon$ may result, due to the normalisation procedure eliminating any substantial variation.

Further investigation suggests that the error distribution in this case is actually heavier-tailed than the normal distribution that the empirical Bayesian method assumes. The heavy tails are clearly apparent on the normal probability plot in the right panel of Figure 3.3, whilst the left panel shows a histogram of the error distribution with the fitted
normal distribution. Gottardo et al. (2006) reported the same feature for this HIV dataset.

Figure 3.3: Left panel: Histogram of error distribution of HIV data with a fitted normal curve, where the standard deviation is estimated by the residual sum of squares approach described in Section 3.4 with equation (3.10). Right panel: Normal probability plot for error distribution of HIV data.

To accommodate this observed heavy tailed behaviour, we artificially inflate the standard deviation using an offset, in a manner similar to the widely used method of Tusher et al. (2001). We choose to estimate the total standard deviation $\hat{\sigma}_e$ by an additive combination of $s_0$ from the modified one-sample $t$-test of Tusher et al. (2001) and the standard deviation by (3.10). For this dataset, $s_0 = 0.232$. Thus we obtain the parameters $\hat{\sigma}_e = 0.353$, $\hat{a} = 0.09$ and $\hat{w} = 0.003$. It is interesting to note that, despite using different mixture models, our estimate of the mixing weight is similarly small to that of Gottardo et al. (2006) and McLachlan et al. (2006), who estimated the mixing weight for these data to be 0.007 and 0.07 respectively.

From the 7680 genes, 20 (0.26%) were identified as being differentially expressed.
between the conditions, including the 12 spiked-in genes. The results are illustrated in Figure 3.4, where each “dot” is an observation $Z_i$. There are $n = 7680$ dots in each plot, however in the “After” panel on the right, 7660 of them lie on the $x$-axis. McLachlan et al. (2006) concluded that a gene was differentially expressed if its posterior probability of non-differential expression was less than a chosen cut-off, $c_0$. Using $c_0 = 0.01$, 15 genes were concluded to be differentially expressed whilst with $c_0 = 0.1$, 37 genes were concluded to be differentially expressed.

![Figure 3.4: Contrast values for 7680 genes in HIV data example before and after thresholding.](image)

As a consequence of our findings in this example, we are left now with an identifiability problem: whether the error distribution is commonly heavier-tailed than normal and should be adjusted for in most instances or whether the HIV dataset is unique in this regard. Data we have obtained from control-control hybridisations (i.e. all error - no differential expression should be observed) illustrates a similar phenomenon, whilst Gottardo et al. (2006) suggested this was the case in the breast cancer data of Hedenfalk et al. (2001). These observations question the success of normalisation schemes, which
aim to transform data into a form that can be treated as normally distributed. Thus for the next two examples we present both analyses: firstly, assuming normal noise and secondly, artificially inflating the standard deviation as in Example 3.5.1, to accommodate for potentially heavy tailed noise. We consider alternative possibilities for the noise distribution later in Chapter 4.

### 3.5.2 E.Coli data

After normalisation, the residual sum of squares approach gave an estimated standard deviation $\hat{\sigma}_e = 0.249$. Without an adjustment to the standard deviation, the empirically estimated prior parameters are $\hat{\omega} = 0.371$ and $\hat{a} = 0.234$. Of the 76 genes, 19 (25%) were identified as being differentially expressed. In this example, $s_0$ is estimated to be 0.197, resulting in a total adjusted standard deviation of 0.447. Using the adjustment resulted in parameter estimates $\hat{\omega} = 0.252$ and $\hat{a} = 0.354$. Of the 76 genes, 13 (17.11%) were identified as being differentially expressed.

Figure 3.5 shows the E.Coli data before and after the empirical Bayesian thresholding procedure, both with and without the adjustment.

### 3.5.3 Lymphoma data

Figure 3.6 shows the lymphoma data before and after the empirical Bayesian thresholding procedure, both with and without the adjustment. Using no adjustment, the estimated parameters are $\hat{\sigma}_e = 0.334$, $\hat{\omega} = 0.592$ and $\hat{a} = 0.697$. Of the 9216, one would conclude that 3182 (34.53%) genes were differentially expressed. In this case, $s_0 = 0.382$, which gives rise to the adjusted parameters, $\hat{\sigma}_e = 0.716$, $\hat{\omega} = 0.01$ and $\hat{a} = 0.568$. Using the inflated standard deviation, the empirical Bayesian method identified 25 (0.27%) differentially expressed genes.
Figure 3.5: Contrast values before and after the empirical Bayesian thresholding method for E.Coli example.
Figure 3.6: Contrast values before and after the empirical Bayesian thresholding method for lymphoma example.
3.6 Discussion

This chapter merely serves to illustrate the potential of empirical Bayesian frameworks within a microarray context. A distinct advantage of the empirical thresholding approach presented here over many existing methods is that the method is completely automatic. The choice of threshold is completely driven by the data and no cut-off point has to be selected arbitrarily by the user. Table 3.5 compares the results of various different procedures for the identification of differential expression. Whilst the examples used here are all from two-channel experiments, the empirical Bayesian thresholding method is also suitable for analysing single-channel data.

In Section 3.4.1, we investigated how varying degrees of correlation between the observations may affect the efficiency of the thresholding procedure. Permutation methods were employed to estimate the degree of correlation present in the datasets considered in this thesis. One should note that these estimates are essentially upper bounds on the true correlation coefficients: the approach involved in the estimation assumes that all the observations are correlated with at least one other; in reality, only a small subset of genes are expected to contribute to coregulation. The correlation coefficients for the E.Coli, HIV and lymphoma datasets were estimated to be 0.0147, 0.0003 and $6.38 \times 10^{-5}$ respectively. Comparing these to the results from the simulation study in Section 3.4.1, we anticipate that such mild correlation will not compromise the usefulness of the empirical Bayesian thresholding procedure within this application.

Independently, Bhowmick et al. (2006) recently proposed this same underlying mixture model with an application to microarray analyses, yet their treatment of the model and emphasis of the method was quite different. Bhowmick et al. (2006) formulated an empirical Bayesian test statistic in terms of the posterior odds of differential expression and offered this as an improvement to similar test statistics in papers by Efron et al. (2001) and Smyth (2004). Regardless of the sophistication of the test statistic, this approach still
Table 3.5: A table of results to compare the numbers of differentially expressed genes identified by various common existing methods. The entries presented for McLachlan et al. (2006) are the numbers of genes that have a posterior probability of non-differential expression less than or equal to 0.1. This cut-off value was chosen to be consistent with their original work. For the other methods that involve arbitrarily selecting some threshold, we have chosen the threshold to maintain a FDR as close as possible to 5%.

<table>
<thead>
<tr>
<th>Method employed</th>
<th>HIV data</th>
<th>E.Coli data</th>
<th>Lymphoma data</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAM (Tusher et al., 2001)</td>
<td>48 (5% FDR)</td>
<td>9 (5.2% FDR)</td>
<td>888 (5% FDR)</td>
</tr>
<tr>
<td>Limma (Smyth, 2004)</td>
<td>353</td>
<td>30</td>
<td>891</td>
</tr>
<tr>
<td>Unnamed (McLachlan et al., 2006)</td>
<td>37</td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td>Empirical Bayesian thresholding (no adjustment)</td>
<td>-</td>
<td>19</td>
<td>3182</td>
</tr>
<tr>
<td>Empirical Bayesian thresholding (with adjustment)</td>
<td>20</td>
<td>13</td>
<td>25</td>
</tr>
</tbody>
</table>
has a distinct drawback as it involves the selection of some arbitrary cut-off criterion for the test statistic.

In the past chapter, we have advocated the use of a thresholding approach, which is completely different to the concept of a test statistic. The crux or novelty of our approach is that the threshold is chosen completely automatically, driven by the data, and no arbitrary selection by the user has to be made in order to conclude that a gene is “interesting”. This is the key idea that distinguishes our method from many of the other empirical Bayesian approaches for microarrays. Whilst Bhowmick et al. (2006) mentioned the thresholding property of the posterior median as a brief passing comment, they did not derive any associated equations or pursue the implementation of this approach anywhere later in their paper.
Chapter 4

Alternative posterior median estimators for sparse sequences

4.1 Introduction

In Section 3.5.1 we observed evidence that the gene expression error distribution was frequently more heavy-tailed than the normal distribution being used to model it. Here, we expand on the method of the last chapter to include longer-tailed error distributions. Firstly, we give expressions for the threshold and posterior median function in the general mixture model case (Bochkina and Sapatinas, 2005) (of which the methodology of Chapter 3 is a special case), before discussing the plausibility and limitations of more specific models. We introduce a new model which incorporates an asymmetric Laplace (ASL) distribution to describe the noise and derive expressions for the positive and negative thresholds, posterior median function and posterior probability of differential expression. Finally, we consider parameter estimation and show some results for the example datasets.
Chapter 4. Alternative posterior median estimators for sparse sequences

4.2 General framework

4.2.1 The single observation case

Consider a single observation $Z$ which can be modelled by the form

$$Z = \mu + \epsilon,$$  \hfill (4.1)

where $\mu$ is an unknown mean and $\epsilon$ is a random noise variable. We assume that $\mu$ follows some prior density and $\epsilon$ comes from some symmetric unimodal distribution with density $\tau$.

As previously, we use a mixture prior on $\mu$

$$\mu \sim (1 - \omega)\delta_0(\mu) + \omega\gamma(\mu),$$  \hfill (4.2)

where $\delta_0$ is a point mass at zero, $\gamma$ is a symmetric unimodal probability density function on $\mathbb{R}$, and $\omega \in [0, 1]$ is the mixing weight. We also assume that that both $\gamma$ and $\tau$ are positive for all $x \in \mathbb{R}$ and finite at zero. We assume that all zero mass is accounted for in the $\delta_0$ component otherwise the mixture is not identifiable. Further, we assume that there is no zero mass in $\tau$ to eliminate the possibility of observing signal without errors.

We define the following functions

$$g(z) = \int_{\mathbb{R}} \tau(z - u)\gamma(u)du \quad \text{(that being the convolution $\gamma \ast \tau$)} \hfill (4.3)$$

and

$$\tilde{F}_1(\mu|z) = \frac{1}{g(z)} \int_{-\infty}^{\mu} \tau(z - u)\gamma(u)du. \hfill (4.4)$$

The posterior density for $\mu$ conditional on $Z = z$ takes the same form as the prior, i.e.
\[ f_{\text{post}}(\mu|Z = z) = (1 - \omega_{\text{post}}) \delta_0(\mu) + \omega_{\text{post}} f_1(\mu|z), \]

where

\[ f_1(\mu|Z = z) = f(\mu|Z = z, \mu \neq 0) \]

and

\[ \omega_{\text{post}}(z) = \omega g(z) \frac{(1 - \omega) \tau(z) + \omega g(z)}{(1 - \omega) \tau(z) + \omega g(z)}. \] (4.5)

The probability that \( \mu \) is non-zero has changed from \( \omega \) to \( \omega_{\text{post}} \), the value of which is clearly dependent on the prior value of the mixing weight and the densities \( \gamma \) and \( \tau \).

Thus the cumulative distribution function of the posterior distribution for \( \mu \) given \( Z = z \) is given by

\[ F_{\text{post}}(\mu|Z = z) = (1 - \omega_{\text{post}}) \delta_0(\mu) + \omega_{\text{post}} \tilde{F}_1(\mu|z), \] (4.6)

where \( \tilde{F}_1(\mu|z) \) can be interpreted as the posterior cumulative distribution function for \( \mu \), given that \( Z = z \) and \( \mu \neq 0 \). Providing that the function \( \tilde{F}_1(\mu|z) \) is invertible, Bochkina and Sapatinas (2005) show that the median of the posterior distribution is given by

\[ m_\mu(z) = \max(0, \text{sign}(|\eta(z)| - 1)) \text{sign}(z) \tilde{F}_1^{-1} \left( \frac{1 - \min\{1, \varsigma(z)\}}{2} \right), \]

where \( \text{sign}(\cdot) \) is the signum function, \( \varsigma(z) = \frac{(1 - \omega) \tau(z)}{\omega g(z)} \), the posterior odds ratio that an observation \( \mu \) comes from the point mass component of the mixture, and \( \eta(z) = \frac{2 \tilde{F}_1(0|z) - 1}{\varsigma(z)} \).
It is worth noting that if $|\eta(z)| \leq 1$ for all $z \in \mathbb{R}$, the posterior median is identically zero. In particular, this is the case when the posterior odds ratio $\varsigma(z) \geq 1$ for all $z \in \mathbb{R}$.

When considering the usefulness of potential estimation rules (thresholding or shrinkage) for an application to microarray data, there are key desirable features that we would wish to observe. In particular, we are interested in estimation rules that possess the bounded shrinkage property. Suppose we have a shrinkage rule $0 \leq \Omega(z,t) \leq z$, for all $z \geq 0$. Mathematically, and as noted by Bochkina and Sapatinas (2005), the term “bounded shrinkage” (relative to some threshold $t > 0$) refers to a shrinkage rule that, for some constant $s$, satisfies

$$|z - \Omega(z,t)| \leq s + t \quad \text{for all } z \text{ and } t. \quad (4.7)$$

Intuitively, bounded shrinkage describes the behaviour of an estimation rule when rare large observations are more or less reliably assigned to sparse signals rather than the noise in the Bayesian model. Therefore an estimator satisfying this property ensures that when $\mu$ is large, $Z$ is not shrunk severely in the estimation of $\mu$. Furthermore, bounded shrinkage implies that the estimated $\mu$ does not diverge in its asymptotic behaviour from the observed value $Z$, as $|Z| \to \infty$. Such performance is important in the context of thresholding microarray data, particularly as a secondary aim of the procedure is to quantify the level of differential expression.

### 4.2.2 Parameter estimation

Now suppose that we have a sequence of $n$ i.i.d. observations $Z_i$ ($i = 1, \ldots, n$), each from a distribution with a unique mean parameter $\mu_i$ and subject to random stochastic noise. The model in (4.1) can be extended accordingly and written as
Chapter 4. Alternative posterior median estimators for sparse sequences

\[ Z_i = \mu_i + \epsilon_i, \]  

(4.8)

where the \( \mu_i \) are independent means from the prior distribution (4.2) and the \( \epsilon_i \) are independent random variables that follow some distribution \( T \).

The marginal density of the observations \( Z_i \) can be written as

\[ (1 - \omega)\tau(z) + \omega g(z). \]

Estimates for the model parameters can be found by maximising the marginal log-likelihood:

\[ \ell(\omega, \theta) = \sum_{i=1}^{n} \log \left\{ (1 - \omega)\tau(Z_i|\theta) + \omega g(Z_i|\theta) \right\}, \]

where \( \theta \) denotes the vector of model parameters, with the exclusion of \( \omega \).

### 4.3 Specific cases

Here we describe three special cases formulated by Bochkina and Sapatinas (2005) and discuss their limitations for use within a microarray context. The models are denoted in the form “non-zero component of the prior distribution”–“noise distribution”. Hence, in the form of this notation, the model in Chapter 3 would be considered to be the Laplace-Gaussian case.

#### 4.3.1 Gaussian-Laplace model

Suppose that \( \gamma \) is a standard Gaussian density (the signal) and that \( \tau \) is a Laplace density with a scaling parameter \( b \) (the noise).
Chapter 4. Alternative posterior median estimators for sparse sequences

The median of the posterior distribution has been shown (Bochkina and Sapatinas, 2005) to be given by

\[ m_\mu(z) = \begin{cases} 
0 & \text{if } |z| \leq t \\
\text{sign}(z) \left[ b\sigma^2 + \sigma^{-1} \left( \frac{(1-\text{sign}(z)\varsigma(z))\nu(z) \exp(b|z|)}{2} \right) \right] & \text{if } |z| > t.
\end{cases} \]

where

\[ \varsigma(z) = \frac{(1 - \omega) e^{-b|z|}}{\omega e^{(b\sigma)^2/2} \nu(z)}, \]
\[ \nu(z) = e^{bz} \Phi(-z/\sigma - b\sigma) + e^{-bz} \Phi(z/\sigma - b\sigma), \]

and the threshold \( t > 0 \) is defined as a solution of the equation

\[ \Phi(t/\sigma - b\sigma) + \Phi(t/\sigma + b\sigma)e^{2bt} = 2\Phi(-b\sigma) + \frac{1 - \omega}{\omega} e^{-(b\sigma)^2/2}. \]

The above equation has a finite solution if and only if

\[ \omega > \frac{1}{1 + (2\Phi(b\sigma) - 1)e^{(b\sigma)^2/2}}. \] (4.9)

If the constraint on \( \omega \) does not hold, \( t = +\infty \), which implies that the median \( m_\mu(z) = 0 \) for all \( z \in \mathbb{R} \). This may prove restrictive for sparse sequences where \( \omega \) is expected to be small, particularly if \( b \) and \( \sigma \) are both small.

A further limitation of this particular model is that, despite the posterior median function being antisymmetric \( (f(-z) = -f(z)) \) and a thresholding rule, it does not possess the bounded shrinkage property. This means that large observations will be shrunk dramatically towards zero, which is not desirable within the context of our problem.

Furthermore, Bochkina and Sapatinas (2005) provided an explicit expression for the value of the asymptotes as \( z \to \pm\infty \) and showed that in its asymptotic behaviour for \( |Z| \to \infty \), the posterior median will in fact tend to a constant rather than the \( y = x \) line.
This feature is illustrated in Figure 4.1, which shows the posterior median function for parameters estimated from the E.Coli data by marginal maximum likelihood. In this case, $\hat{b} = 3.59$, $\hat{\sigma} = 1.69$ and $\hat{\omega} = 0.22$. The range of values to be affected by the non-bounded shrinkage and the constant asymptotic behaviour is clearly dependent on the values of the parameters.

![Figures 4.1 and 4.2](image-url)

Figure 4.1: Posterior median function for Gaussian-Laplace model with estimated parameters $\hat{b} = 3.59$, $\hat{\sigma} = 1.69$, $\hat{\omega} = 0.22$. The left panel shows a close-up of the thresholded region, whilst the right panel shows the function on a larger scale, in order to illustrate the effects of the non-bounded shrinkage.

Figure 4.2 shows the 76 contrast values for the E.Coli data before and after thresholding using the estimated parameters. Thirteen genes are retained as being differentially expressed whilst the remainder are thresholded to zero. This is the same number of differentially expressed genes as identified by the adjusted Laplace-Gaussian model in Chapter 3.

Suppose we have distributions $X$ and $Y$ with $E(X) = E(Y) = 0$ and $\text{Var}(X) = \text{Var}(Y)$. We say that distribution $Y$ is heavier-tailed than $X$ if there exists some $k > 0$ such
that for all $|z| > k$, $f_Y(z) > f_X(z)$, with $f(\cdot)$ being the probability density function of the respective distribution. Bochkina and Sapatinas (2005) generalised the findings of the Gaussian-Laplace case to conclude that for any model where the tails of the noise distribution are heavier than the tails of the signal distribution, the posterior median will tend to a constant for large observations, thus severely underestimating the corresponding values of $\mu$. Thus in order to develop a more useful model for thresholding microarray data, we restrict our future attention to models where the tails of the non-zero distribution are at least as heavy as those of the noise distribution.

### 4.3.2 Scaled $t_1$-$t_1$ model

Assume now that both the signal and noise are scaled $t_1$-distributions, with scaling parameters $\xi$ and $\sigma$ respectively.
Bochkina and Sapatinas (2005) derived implicit equations for the median \( m_\mu(z) \) of the posterior distribution of \( \mu \) given \( z \) and the threshold \( t \). However, they also demonstrated that the posterior median is an antisymmetric function, which is a thresholding and shrinkage rule if and only if \( \sigma < \xi \) and \( \omega \geq \frac{\sigma}{\xi} \). Further, the bounded shrinkage property only exists in the case where \( \omega > \frac{\sigma}{\xi} \). If \( \omega < \frac{\sigma}{\xi} \) the posterior median is not a thresholding function and moreover, if \( \omega < \frac{\sigma}{\sigma + \xi} \) then the posterior median is identically zero for all values of \( z \in \mathbb{R} \).

In the context of microarrays, only a few genes amongst thousands may be differentially expressed. Thus it is not uncommon to expect very small values of \( \omega \). Consequently, the constraint on \( \omega \) will be too stringent unless we allow implausible values for \( \xi \) and \( \sigma \). However, compromising with “incorrect” values for the scaling parameters will have dramatic repercussions on the threshold and median values, hence should be avoided.

### 4.3.3 Laplace-Laplace model

Assume now that \( \gamma \) comes from a Laplace distribution with scaling parameter \( a \), with density

\[
\gamma_a(\mu) = \frac{a}{2} \exp(-a|\mu|).
\]

(4.10)

For future reference, we take this opportunity to state that for a Laplace distribution with density (4.10), the variance of the distribution is known to be \( 2/a^2 \).

We take \( \tau \) to be the probability density function of the Laplace distribution also, but with scaling parameter \( b \), then

\[
\tau_b(\epsilon) = \frac{b}{2} \exp(-b|\epsilon|).
\]

(4.11)
The median of the posterior distribution can be shown to be given by

\[
m_{\mu}(z) = \begin{cases} 
0 & \text{for } |z| \leq t \\
\text{sign}(z) \left[ \frac{1}{b-a} \log \left( \frac{a+b}{a+b} + \frac{(1-\omega)(b-a)}{\omega} \right) \right] & \text{for } |z| > t.
\end{cases}
\]

The threshold \( t > 0 \) is defined by

\[
t = \frac{1}{b-a} \log \left[ \max \left( 0, 1 + \frac{(1-\omega)(b^2-a^2)}{\omega ab} \right) \right].
\] (4.12)

Bochkina and Sapatinas (2005) show that when \( a > b \), the posterior median is an antisymmetric, thresholding rule only when \( \omega > \frac{a^2-b^2}{ab+a^2-b^2} \). If this condition does not hold, then the posterior median is zero for all \( z \in \mathbb{R} \), which proves problematic for sparse sequences.

Provided that \( a < b \), the condition on \( \omega \) is redundant and bounded shrinkage follows. It is this particular fact that makes the Laplace-Laplace choice of model the most suitable for modelling microarray data. With no condition on \( \omega \) to be satisfied, varying degrees of sparsity can be accommodated. Very small values of \( \omega \) should not present a problem and can be reached without having to compromise on the values of the other parameters, as might have been necessary with the scaled \( t_1 - t_1 \) model previously discussed. In terms of its practical implications, enforcing the condition \( a < b \) is not unreasonable. The constraint implies that the variance of the signal distribution should be greater than that of the noise distribution; in cases where this is thought not to be justified, one would expect that the quality of the data should be closely questioned.

The convolution of \( \gamma_a(\cdot) \) with \( \tau_b(\cdot) \) takes the explicit form

\[
g(z) = \frac{ab}{2} \frac{ae^{-b|z|} - be^{-a|z|}}{a^2 - b^2}.
\] (4.13)
The marginal maximum likelihood estimators of $\omega$, $a$ and $b$, denoted $\hat{\omega}$, $\hat{a}$ and $\hat{b}$ respectively, can be estimated by finding the maximum over all three parameters of the marginal log-likelihood

$$
\ell(\omega, a, b) = \sum_{i=1}^{n} \log \left( \frac{(1 - \omega)b}{2} \exp(-b|Z_i|) + \omega g(Z_i) \right).
$$

We demonstrate the potential of this model with some results in Section 4.6.

### 4.4 The Laplace-asymmetric Laplace model

#### 4.4.1 Motivation

The control-control data form perhaps one of the more interesting datasets that we have to explore: what we observe is the pure noise realised in the experiment and this is useful for investigating the error distribution most appropriate for gene expression. The control-control data display substantial heavy-tailed behaviour, similar to that seen in the HIV data; however, visual inspections suggest that no symmetric distribution seems to fully encompass the flavour of the true distribution. Thus, it is a natural extension of this work to consider the asymmetric Laplace distribution as a potential error distribution.

#### 4.4.2 The asymmetric Laplace distribution

Various forms of a skewed Laplace distribution have appeared in the literature over the years (McGill, 1962; Holla and Bhattacharya, 1968; Poiraud-Casanova and Thomas-Agnan, 2000; Bhowmick et al., 2006). We use a formulation introduced by Hinkley and Revankar (1977) which generalises the heavy-tailed Laplace distribution (4.11) to allow for asymmetry. The probability density function can be written as below
Chapter 4. Alternative posterior median estimators for sparse sequences

$$f(x) = \begin{cases} 
\frac{b\kappa}{1+\kappa} \exp(-b\kappa|x-\theta|) & \text{if } x \geq \theta \\
\frac{b\kappa}{1+\kappa^2} \exp(-\frac{b}{\kappa}|x-\theta|) & \text{if } x < \theta,
\end{cases}$$

(4.14)

where $\kappa > 0$ and controls the skewness of the distribution and the differing tail behaviours. The distribution is constructed in such a way as to allow the probability density function to be continuous at $x = 0$.

The asymmetric Laplace distribution offers great flexibility as a model: when $\kappa = 1$, density (4.14) reduces to that of the symmetrical Laplace case. Thus, assuming that $\kappa$ is to be estimated from the data, the user does not have to make a prior choice of asymmetric versus symmetric; instead, the choice of model will be driven by the data itself.

Purdon and Holmes (2005) explored the use of an asymmetric Laplace distribution within the microarray context. The authors concurred with our observation that the distribution of normalised gene expression ratios within arrays often possesses heavy tails and has varying degrees of asymmetry. They investigated several datasets and found this to be the case, regardless of the normalisation procedure employed. Purdon and Holmes worked under the assumption that a large majority of genes were not differentially expressed and that on the whole, the observed gene expression should represent observations from some error distribution. They proposed that the asymmetric Laplace distribution could effectively model the error distribution and showed evidence of the superior fit for their various sets of data. Further to this, Purdon and Holmes attempted to interpret the asymmetric Laplace in the microarray context and provided intuitive discussion as to why this distribution gives a good fit to cDNA data.

Figure 4.3 illustrates the suitability of the asymmetric Laplace distribution over the Laplace and Gaussian distributions for the control-control data and HIV data (with the spike-in genes removed). All of the observations here are known to be truly zero as no genes are differentially expressed. Hence we assume that the location parameter $\theta$ in representation (4.14) is zero. Maximum likelihood estimates of the parameters for all
We can use Akaike’s Information Criterion (AIC) (Akaike, 1973; Burnham and Anderson, 1998) as an objective criterion by which to compare the appropriateness of the different models. It should be noted that the choice of the AIC over any other information criterion is an arbitrary one, simply motivated by its prevalence amongst existing bioinformatics literature. The AIC for a model $h(\theta)$ is defined as

$$AIC = -2 \log \left( \mathcal{L}(\hat{\theta}|y_1, \ldots, y_n) \right) + 2K,$$

(4.15)

where $\mathcal{L}$ is the likelihood function of the model $h$, $\hat{\theta}$ is a vector consisting of the maximum likelihood estimates of the parameters of $h$ and $K$ is the number of model parameters.
Chapter 4. Alternative posterior median estimators for sparse sequences

The AIC is an estimate of the relative Kullback-Liebler distance between the true model \( f \) and the model under consideration \( h \), adjusted for the number of parameters. Thus, the smaller the AIC, the better the fit of model \( h \). The AIC values for different datasets will not be on equivalent scales, therefore the AIC should not be compared across different datasets.

Specifically for the control-control and HIV datasets explored in this thesis, the AIC values given in Table 4.1 suggest that the asymmetric Laplace distribution has the better fit to the data; in both cases the Gaussian distribution yields the highest AIC values.

Table 4.1: Parameter estimates and Akaike’s Information Criterion for control-control and HIV data.

<table>
<thead>
<tr>
<th>Model fitted</th>
<th>Parameter estimates</th>
<th>AIC</th>
<th>Model fitted</th>
<th>Parameter estimates</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control-control data</td>
<td>HIV data</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal</td>
<td>( \hat{\sigma} = 0.328 )</td>
<td>11357.2</td>
<td>Normal</td>
<td>( \hat{\sigma} = 0.194 )</td>
<td>-3358.8</td>
</tr>
<tr>
<td>Laplace</td>
<td>( \hat{b} = 4.240 )</td>
<td>9223.5</td>
<td>Laplace</td>
<td>( \hat{b} = 7.493 )</td>
<td>-4919.1</td>
</tr>
<tr>
<td>Asymmetric Laplace</td>
<td>( \hat{b} = 4.485 ) ( \hat{\kappa} = 1.170 )</td>
<td>8223.9</td>
<td>Asymmetric Laplace</td>
<td>( \hat{b} = 7.550 ) ( \hat{\kappa} = 1.073 )</td>
<td>-4994.7</td>
</tr>
</tbody>
</table>

The Laplace distribution can be considered to be a submodel of the asymmetric Laplace distribution, with a restricted number of parameters, based on the assumption that \( \kappa = 1 \). The likelihood ratio test is appropriate for the comparison of hierarchically nested models, with the aim being to establish whether the more complex model fits a particular dataset significantly better than the simple model. Including additional parameters in a model will always result in a higher likelihood and a better fit to the data, however there inevitably will come a point when the improvement gained is not worth the additional complexity.
incurred. Thus, the likelihood ratio test provides a further criterion on which to base model selection.

Suppose that we have two models of interest $M_1$ and $M_2$ with $r_1$ and $r_2$ fitted parameters respectively. Further suppose that $M_1 \subset M_2$. We can define $\ell_1$ to be the log-likelihood obtained from fitting model $M_1$; similarly $\ell_2$ is the log-likelihood from model $M_2$. We let $k$ describe the number of parameters lost in the restricted model; in other words, model $M_1$ has $k = r_2 - r_1$ less parameters than model $M_2$.

For large amounts of information, the statistic

$$-2(\ell_1 - \ell_2) \sim \chi^2_k$$

(4.16)

can be used to test the importance of the extra parameters in $M_2$. The assumption made in the restricted model is rejected at the $100\alpha\%$ level if the value of the test statistic is greater than $\chi^2_k(\alpha)$.

In the case of the control-control data, we can use the likelihood ratio test to assess the importance of the extra parameter $\kappa$. For these data, we obtain $\ell_1 = 2460.554$, $\ell_2 = 2499.367$ and find the value of the test statistic to be equal to 77.626. Comparing this to $\chi^2_1(0.05) = 3.841$, we strongly reject the assumption that $\kappa = 1$ at the 5% significance level, which indicates that the inclusion of $\kappa$ in the model is significant. The same conclusion is reached for the HIV dataset, indeed the result is even more extreme.

### 4.4.3 Methodology for Laplace-asymmetric Laplace model

In this section we incorporate the asymmetric Laplace distribution into the general framework of Section 4.2 and introduce the Laplace-asymmetric Laplace model. Here, we present explicit expressions for the threshold and posterior median function. Further details of the algebra for this model can be found in Appendix B.
Chapter 4. Alternative posterior median estimators for sparse sequences

The single observation case

Suppose we have a single observation that can be modelled in the form of Section 4.2. Once again, we assume that $\gamma$ is a Laplace density with scaling parameter $a$. To allow for any asymmetry of the noise distribution we take $\tau$ to be the asymmetric Laplace density with parameters $b$ and $\kappa$.

Let $g$ be the convolution of $\gamma(\cdot)$ with $\tau(\cdot)$. The convolution is not symmetric due to the asymmetry of $\tau$ and thus we define the function for positive and negative $Z$ separately:

$$g^+(z) = \frac{aB^+}{2(1 + \kappa^2)} \left( \frac{\exp(-B^+z)}{a + B^+} + \frac{\exp(-B^+z)}{-a + B^+} \left( \exp((-a + B^+)z) - 1 \right) + \frac{\exp(-az)}{a + B^-} \right)$$  
for $z \geq 0$,  

$$g^-(z) = \frac{aB^+}{2(1 + \kappa^2)} \left( \frac{\exp(az)}{a + B^+} + \frac{\exp(B^-z)}{a - B^-} \left( 1 - \exp((a - B^-)z) \right) + \frac{\exp(B^-z)}{a + B^-} \right)$$  
for $z < 0$,  

where $B^+ = b\kappa$ and $B^- = b/\kappa$.

The cumulative distribution function of the posterior distribution for $\mu$ given $Z = z$ can be written as

$$F_{\text{post}}(\mu|Z = z) = (1 - \omega_{\text{post}})\delta_0(\mu) + \omega_{\text{post}}\tilde{F}_1(\mu|z), \quad (4.17)$$

where for this particular model
\[ \tilde{F}_1(\mu|z) = \frac{a\mathcal{B}^+}{2(1 + \kappa^2)g^{\text{sign}(z)}(z)} \left[ \exp(-\mathcal{B}^+ z) \exp \left( (a + \mathcal{B}^+) \min(\mu, p_0) \right) \right] \]

\[ + \mathbb{I}(\mu \geq p_1) \exp(\mathcal{B}^- z) \left\{ \exp \left( - (a + \mathcal{B}^-) p_1 \right) - \exp \left( - (a + \mathcal{B}^-) \mu \right) \right\} \]

\[ + \mathbb{I}(\mu \geq p_0) \frac{\exp(-\text{sign}(z)\mathcal{B}^{\text{sign}(z)} z)}{-\text{sign}(x)(a - \mathcal{B}^{\text{sign}(z)})} \]

\[ \times \left\{ \exp \left( - \text{sign}(z)(a - \mathcal{B}^{\text{sign}(z)}) \min(\mu, p_1) \right) - \exp \left( - \text{sign}(z)(a - \mathcal{B}^{\text{sign}(z)}) p_0 \right) \right\} \] ,

(4.18)

with \( p_0 = \min(0, z) \) and \( p_1 = \max(0, z) \).

Given an observation \( z \), we denote the median of the posterior distribution by \( m_\mu(z) \), i.e. \( m_\mu(z) \) is such that \( F_{\text{post}}(m_\mu(z)|z) = 0.5 \). Alternatively, we can express this mathematically as

\[ (1 - \omega_{\text{post}})\delta_0(m_\mu(z)) + \omega_{\text{post}}\tilde{F}_1(m_\mu(z)|z) = 0.5. \quad (4.19) \]

If \( \omega_{\text{post}} \leq 0.5 \) then the posterior median will be zero. Furthermore, the posterior median is zero if and only if \( \omega_{\text{post}}\tilde{F}_1(0|z) \leq 0.5 \) and \( 1 - \omega_{\text{post}} + \omega_{\text{post}}\tilde{F}_1(0|z) \geq 0.5 \). The term \( \omega_{\text{post}} \) can be written in terms of the posterior odds ratio, \( \varsigma(z) \), and the previous condition thus becomes \( 0.5(1 - \varsigma(z)) \leq \tilde{F}_1(0|z) \leq 0.5(1 + \varsigma(z)) \). Let \( \eta(z) = \frac{2\tilde{F}_1(0|z) - 1}{\varsigma(z)} \) and then the latter condition is equivalent to \(-1 \leq \eta(z) \leq 1 \).

Thus the threshold \( t \) for positive and negative \( z \) can be found as the solution of the equation \( \eta(t) = -1 \) and \( \eta(t) = 1 \) respectively and can be written explicitly as
Chapter 4. Alternative posterior median estimators for sparse sequences

\[ t^+ = \max \left[ 0, \log \left( \max \left( 0, \frac{2(\omega a B^+ + \omega a^2 - a^2 - \omega B^2 + B^+)}{\omega a (a + B^+)(B^+ + B^-)} \right) \right) \right] / (-a + B^+) \] for \( z \geq 0, \)

\[ t^- = \min \left[ 0, \log \left( \max \left( 0, \frac{2(\omega a B^- + \omega a^2 - a^2 - \omega B^2 - B^-)}{\omega a (a + B^-)(B^+ + B^-)} \right) \right) \right] / (a - B^-) \] for \( z < 0. \)

\begin{equation}
(4.20)
\end{equation}

The threshold \( t^+ \) is finite only if

\[ \omega > \frac{a^2 - B^{+2}}{a B^+ + a^2 - B^{+2}}. \]

\begin{equation}
(4.21)
\end{equation}

When \( a \leq B^+ \), the condition becomes redundant as the value of \((a^2 - B^{+2})/(a B^+ + a^2 - B^{+2})\) will always be less than or equal to zero. However, if \( a > B^+ \) and the constraint on \( \omega \) is not satisfied, \( t^+ = \infty \) which implies that the posterior median will be identically zero for all \( z \in \mathbb{R}^+ \).

Similarly, \( t^- \) is finite only if

\[ \omega > \frac{a^2 - B^{-2}}{a B^- + a^2 - B^{-2}}. \]

\begin{equation}
(4.22)
\end{equation}

When \( a \leq B^- \), \( \omega \) can take any value. However, if \( a > B^- \) and the condition on \( \omega \) is not satisfied, \( t^- = -\infty \) which implies that the posterior median will be identically zero for all \( z \in \mathbb{R}^- \). These conditions are restrictive for the purposes of estimating sparse sequences as \( \omega \) is never allowed to be very small. Thus we constrain \( a < B^+ \) and \( a < B^- \).

Suppose that \( a < B^+ \) and \( a < B^- \) holds. Now, \( t^+ \) will be definitively zero unless

\[ \omega < \frac{a^2 - B^{+2}}{a^2 - B^{+2} + a B^+ - \frac{a B^+ + B^-}{2(a + B^-)}}. \]

\begin{equation}
(4.23)
\end{equation}

Similarly, \( t^- \) will be definitively zero unless
Chapter 4. Alternative posterior median estimators for sparse sequences

\[ \omega < \frac{a^2 - B^2}{a^2 - B^2 + aB - \frac{a(B^+ + B^-)(a + B^-)}{2(a + B^+)}}. \]  

(4.24)

If \( \kappa = 1 \), both (4.23) and (4.24) constrain \( \omega \) to be less than 1. Given that \( \omega \in [0, 1] \), this condition ceases to be problematic.

In the case that \( \kappa < 1 \) i.e. \( a < B^+ < B^- \), equation (4.23) constrains \( \omega \) to be less than some value which is always greater than 1. Hence, a non-zero \( t^- \) will exist unconditional on \( \omega \). However, the value of the constraint in (4.24) will be less than 1 and \( t^- \) will be zero unless the condition on \( \omega \) is met. If \( t^- \) is zero, the posterior median function will not be a thresholding and shrinkage rule for \( z \in \mathbb{R}^- \). Conversely, if \( \kappa > 1 \) i.e. \( a < B^- < B^+ \), the value of the constraint in (4.23) will be less than 1 and \( t^+ \) will be zero unless the condition on \( \omega \) is met. If \( t^+ \) is zero, the posterior median function will not be a thresholding and shrinkage rule for \( z \in \mathbb{R}^+ \). Equation (4.24) constrains \( \omega \) to be less than some value which is always greater than 1. Hence, a non-zero \( t^- \) will exist unconditional on \( \omega \). In practice, even after the investigation of some extreme cases, we found \( \omega \) was generally constrained to be less than values ranging from around 0.85 to 0.99. In the microarray context, where very small values of \( \omega \) are expected, and in other sparse situations, it is anticipated that these restrictions should not pose a problem.

The posterior median \( m_\mu(z) \) is defined as

\[
m_\mu(z) = \begin{cases} 
\frac{1}{a - B^-} \log \left( \frac{(1 - \omega)(a - B^-)}{\omega a} \right) + \frac{a}{a + B^-} + \frac{(B^+ + B^-) \exp(az - 2 - z)}{2(a + B^+)} & \text{if } z < t^- , \\
0 & \text{if } t^- \leq z \leq t^+ , \\
\frac{1}{a - B^+} \log \left( \frac{(1 - \omega)(a - B^+)}{\omega a} \right) + \frac{a}{a + B^+} + \frac{(B^+ + B^-) \exp(-az + B^+ + 2 + z)}{2(a - B^-)} & \text{if } z > t^+ . 
\end{cases}
\]

Figure 4.4 shows the posterior median function for chosen parameters \( a = 0.5, b = 3, \omega = 0.0625 \) and \( \kappa = 1 \) and 0.6 respectively. In the left panel \( \kappa = 1 \), hence the Laplace-ASL model reduces to the symmetric Laplace-Laplace model discussed in Section 4.3.3 and the posterior median function is antisymmetric.
Figure 4.4: Posterior median functions for the Laplace-ASL model with some asymmetry parameter $\kappa$. Left panel: $\kappa = 1$ thus is the restricted symmetric Laplace-Laplace model. Right panel: Asymmetry parameter $\kappa = 0.6$. 
If $\kappa > \frac{a + \sqrt{a^2 + b^2}}{b}$ or $\kappa < \frac{-a + \sqrt{a^2 + b^2}}{b}$ strange behaviour will be observed for large positive and negative $z$ respectively: “large” is defined in this instance as

$$z > \log \left( \frac{2(a^2 - B^2 + \omega B^2)(a + B^-)}{\omega a(a + B^-)(2a + B^- - B^+)} \right) - a + B^+ \quad \text{in the positive case} \quad (4.25)$$

and

$$z < \log \left( \frac{2(a^2 - B^2 - \omega B^2)(a + B^+)}{\omega a(a + B^-)(2a + B^- - B^+)} \right) - a - B^- \quad \text{in the negative case.} \quad (4.26)$$

This strange behaviour manifests itself in that the posterior median function does not shrink the large $z$ values as expected, but in fact inflates them slightly i.e. in such instances the posterior median function is not a shrinkage rule for all $z \in \mathbb{R}$. Based on the range of $\kappa$ values observed under our investigation, and those reported by Purdon and Holmes (2005), it is anticipated that this irregularity may occasionally prove to be more than a hypothetical occurrence.

The posterior probability that an observation $z$ originates from the non-zero component of the mixture prior can be written in terms of

$$\omega_{\text{pos}}(z) = \begin{cases} \frac{\omega g^+(z)(1 + \kappa^2)}{1 - \omega} & \text{for } z \geq 0 \\ \frac{\omega g^-(z)(1 + \kappa^2)}{1 - \omega} & \text{for } z < 0. \end{cases}$$

This is analogous to the posterior probability that the a gene corresponding to observation $z$ is differentially expressed.

Figure 4.5 shows the effect that the asymmetry parameter has on the function $w_{\text{pos}}(z)$. For these plots the parameters are $a = 0.5$, $b = 3$, $\omega = 0.0625$ and $\kappa = 1$ (left panel) and $0.6$ (right panel) respectively.
Figure 4.5: Posterior probability of differential expression $\omega_{\text{post}}$ across a range of observed values with chosen parameters $a = 0.5$, $b = 3$, $\omega = 0.0625$ and some $\kappa$. Left panel: $\kappa = 1$ thus is the restricted symmetric Laplace-Laplace model. Right panel: Asymmetry parameter $\kappa = 0.6$. 
Parameter estimation

Suppose we have a sequence of \( n \) i.i.d. observations \( Z_i (i = 1, \ldots, n) \), each from a distribution with a unique mean parameter \( \mu_i \) and subject to asymmetric Laplace noise.

The marginal density of the observations \( Z_i \) can be written as

\[
\frac{(1 - \omega)b\kappa}{1 + \kappa^2} \exp(-B^{\text{sign}(z)}|z|) + \omega g(z).
\]

The marginal maximum likelihood estimators of \( \omega, \ a, \ b \) and \( \kappa \), denoted \( \hat{\omega}, \ \hat{a}, \ \hat{b} \) and \( \hat{\kappa} \) respectively, are estimated by finding the maximum over all four parameters of the marginal log-likelihood

\[
\ell(\omega, a, b, \kappa) = \sum_{i=1}^{n} \log \left( \frac{(1 - \omega)b\kappa}{1 + \kappa^2} \exp(-B^{\text{sign}(Z_i)}|Z_i|) + \omega g(Z_i) \right).
\]

Initially we used an optimisation routine that permitted limits to be defined for each of the parameters, i.e. \( \omega \in (0, 1), \ a, \ b, \kappa > 0 \). In practice, this ran into computational difficulties for some of the datasets, consistently evaluating the parameters to be the set limits. In response to this problem, some reparameterisation was necessary in order to allow the parameters being estimated to be unbounded. Consequently, \( a, \ b \) and \( \kappa \) were reparameterised as \( \log(a), \ \log(b) \) and \( \log(\kappa) \) respectively; \( \omega \) was reparameterised as \( \log(\omega/(1 - \omega)) \).

Further, when tested on the control-control data, the maximum likelihood estimation procedure consistently estimated \( \omega = 1 \), instead of the expected zero. From this, one would conclude that the data observed are all signal rather than noise. Additionally, the procedure estimated \( a >>> B^+ \) and \( a >>> B^- \), which violates a condition previously set in response to the constraints (4.21) and (4.22). This result suggests that there is an identifiability problem with the scaling parameters \( a \) and \( b \) when \( \omega \) takes its most
extreme values. Intuitively, there is the further suggestion that the estimation procedure has found the observed data more likely to have come from a Laplace-asymmetric Laplace convolution, than an unconvolved asymmetric Laplace distribution. This perhaps indicates that the noise distribution is even more heavy-tailed than we have allowed for with the asymmetric Laplace assumption. To ensure that $a < B^+$ and $a < B^-$ hold for datasets that we may wish to analyse in the future, we incorporated these constraints into the likelihood function, which simultaneously resolved the problem of falsely estimating $\omega$.

4.5 Incorporating the false discovery rate

As an alternative to the thresholding procedure, one could instead use the posterior probability of non-differential expression $1 - \omega_{\text{post}}(Z_i)$ to make a conclusion about the status of $\mu_i$. This approach is analogous to the idea of Bayesian hypothesis testing, with the null hypothesis being that a parameter $\mu_i$ is from the point mass at zero compared with the alternative hypothesis that $\mu_i$ comes from the heavy-tailed density. For each gene with $1 - \omega_{\text{post}}i < t_0$, where $t_0$ is some cut-off criterion, we would reject the null hypothesis in favour of the alternative, concluding that gene $i$ is probably truly differentially expressed.

In the results presented so far, we have chosen not to adopt this practice as it involves the arbitrary selection of $t_0$. However, if for each cut-off value $t_0$, an estimate of the implied false discovery rate could be calculated, the choice of $t_0$, whilst still rather arbitrary, at least becomes a more informed decision. Furthermore, having some estimate of the FDR allows the choice of cut-off $t_0$ to be driven by the aim of the experiment: in cases where the experimenter has some prior knowledge of which genes are differentially expressed and is looking to confirm this for a small number of specific genes, only a low FDR is likely to be tolerated; for those experiments where there is no prior knowledge and the purpose is simply to “fish” for potentially interesting genes, the experimenter is more
likely to accept an increased FDR.

We can compile a shortlist of “interesting genes” by ranking genes according to increasing values of $1 - \omega_{\text{post}_i}$ and including all those genes that satisfy $1 - \omega_{\text{post}_i} < t_0$. Newton et al. (2004) derived the direct posterior probability approach for estimating the false discovery rate, which we apply here. Suppose we select all the genes with

$$1 - \omega_{\text{post}_i} \leq t_0. \tag{4.27}$$

The expected number of false discoveries is then

$$C(t_0) = \sum_{i=1}^{n} (1 - \omega_{\text{post}_i}) \cdot \mathbb{I}[1 - \omega_{\text{post}_i} < t_0], \tag{4.28}$$

since $1 - \omega_{\text{post}_i}$ is the conditional probability that putting gene $i$ on the shortlist will result in a Type I error. Thus, the expected false discovery rate is

$$\widehat{FDR} = \frac{C(t_0)}{N_r}, \tag{4.29}$$

where $N_r$ is the number of selected genes on the shortlist.

Therefore, we can choose a data-dependent $t_0 \leq 1$ as large as possible such that

$$\widehat{FDR} \leq \alpha, \tag{4.30}$$

where $\alpha$ is the value by which one wishes to bound the false discovery rate. The existence of $t_0$ assumes that there are some genes with $1 - \omega_{\text{post}} < \alpha$, which is plausible in practice. However, it should be noted that the estimate of the FDR is only approximate due to the use of estimates in calculating the posterior probabilities $1 - \omega_{\text{post}}$ and that the accuracy of the approximation depends on the fit of the densities assumed in the model.
Having established an expression to estimate the false discovery rate, we can use this in collaboration with the empirical Bayesian threshold calculated as part of the standard routine. Suppose that we conclude that $N_r$ genes are differentially expressed. The implied FDR associated with that particular threshold value can be found by summing the $1 - \omega_{\text{post}}$ values for those $N_r$ genes, and then dividing by $N_r$.

### 4.6 Examples

#### 4.6.1 HIV data

**Laplace-Laplace model**

The model parameters were estimated to be $\hat{a} = 0.343$, $\hat{b} = 7.681$ and $\hat{\omega} = 0.0047$. Thus the threshold was found to 1.153. Twenty-six (0.34%) genes were found to be differentially expressed with an implied FDR of 0.073.

![HIV data before and after thresholding using the Laplace-Laplace model.](image)

Figure 4.6: HIV data before and after thresholding using the Laplace-Laplace model.
Chapter 4. Alternative posterior median estimators for sparse sequences

One may prefer to select genes based on the posterior probabilities of differential expression. Using a chosen cut-off $t_0 = 0.01$, 18 (0.23\%) genes are concluded to be differentially expressed, which has an implied FDR of 0.001. These selected genes and their corresponding $1 - \omega_{\text{post}}$ values are shown in Table 4.2. The 12 spiked-in genes are the top 12 genes that appear in the shortlist of 18 above. All have a posterior probability of non-differential expression equal to 0 to seven decimal places (or equivalently, a posterior probability of differential expression equal to 1). Using $t_0 = 0.1$ resulted in 20 (0.26\%) genes being identified as differentially expressed and an implied FDR of 0.004. The two extra genes are shown in Table 4.3.

Alternative choices for the cut-off $t_0$ can be made to make the procedure less conservative. Possible values for $t_0$ are shown in Table 4.4, along with the number of genes found to be differentially expressed by that choice of cut-off and the corresponding implied FDR.

**Laplace-asymmetric Laplace model**

The model parameters were estimated by the marginal maximum likelihood approach to be $\hat{a} = 0.321$, $\hat{b} = 7.707$, $\hat{\kappa} = 1.063$ and $\hat{\omega} = 0.0042$. Using equation (4.20), $t^+ = 1.097$ and $t^- = -1.246$. Thresholding by $t^+$ and $t^-$ gave rise to 24 (0.31\%) genes identified as differentially expressed, with an implied FDR of 0.062. The contrast values before and after thresholding are shown in Figure 4.7.

Suppose we adopt the alternative procedure of thresholding based on the $1 - \omega_{\text{post}}$ values. Using a chosen cut-off $t_0 = 0.01$, 16 (0.21\%) genes are concluded to be differentially expressed, which has an implied FDR of 0.001. These selected genes and their corresponding $1 - \omega_{\text{post}}$ values are shown in Table 4.5. Using $t_0 = 0.1$ found an additional 4 differentially expressed genes, which are given in Table 4.6. Overall, the 20 (0.26\%) identified genes had an implied FDR of 0.0092.

Table 4.7 gives alternatives for $t_0$, along with the number of genes found to be
Table 4.2: Laplace-Laplace model: Genes identified as differentially expressed using a cut-off $t_0 = 0.01$ and their corresponding $1 - \omega_{\text{post}}$ values.

<table>
<thead>
<tr>
<th>Gene reference</th>
<th>Posterior probability of non-differential expression $(1 - \omega_{\text{post}})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.000</td>
</tr>
<tr>
<td>1285</td>
<td>0.000</td>
</tr>
<tr>
<td>6419</td>
<td>0.000</td>
</tr>
<tr>
<td>2565</td>
<td>0.000</td>
</tr>
<tr>
<td>1287</td>
<td>0.000</td>
</tr>
<tr>
<td>2563</td>
<td>0.000</td>
</tr>
<tr>
<td>1283</td>
<td>0.000</td>
</tr>
<tr>
<td>3</td>
<td>0.000</td>
</tr>
<tr>
<td>3845</td>
<td>0.000</td>
</tr>
<tr>
<td>3847</td>
<td>0.000</td>
</tr>
<tr>
<td>3843</td>
<td>0.000</td>
</tr>
<tr>
<td>2567</td>
<td>0.000</td>
</tr>
<tr>
<td>6</td>
<td>0.000</td>
</tr>
<tr>
<td>2566</td>
<td>0.000</td>
</tr>
<tr>
<td>7059</td>
<td>0.003</td>
</tr>
<tr>
<td>1284</td>
<td>0.004</td>
</tr>
<tr>
<td>3082</td>
<td>0.004</td>
</tr>
<tr>
<td>1286</td>
<td>0.007</td>
</tr>
</tbody>
</table>
Table 4.3: Laplace-Laplace model: Two further genes identified as differentially expressed when using a cut-off $t_0 = 0.1$ and their corresponding $1 - \omega_{\text{post}}$ values.

<table>
<thead>
<tr>
<th>Gene reference</th>
<th>Posterior probability of non-differential expression $(1 - \omega_{\text{post}})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1923</td>
<td>0.028</td>
</tr>
<tr>
<td>3842</td>
<td>0.034</td>
</tr>
</tbody>
</table>

Table 4.4: Laplace-Laplace model: Alternative choices for the cut-off $t_0$, the number of differentially expressed genes found using this threshold and the corresponding implied FDR.

<table>
<thead>
<tr>
<th>Cut-off $t_0$</th>
<th>Number of differentially expressed genes</th>
<th>Associated FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>20</td>
<td>0.004</td>
</tr>
<tr>
<td>0.10</td>
<td>20</td>
<td>0.004</td>
</tr>
<tr>
<td>0.15</td>
<td>21</td>
<td>0.009</td>
</tr>
<tr>
<td>0.20</td>
<td>22</td>
<td>0.017</td>
</tr>
<tr>
<td>0.25</td>
<td>22</td>
<td>0.017</td>
</tr>
<tr>
<td>0.30</td>
<td>22</td>
<td>0.017</td>
</tr>
<tr>
<td>0.35</td>
<td>23</td>
<td>0.030</td>
</tr>
<tr>
<td>0.40</td>
<td>25</td>
<td>0.058</td>
</tr>
<tr>
<td>0.45</td>
<td>26</td>
<td>0.073</td>
</tr>
<tr>
<td>0.50</td>
<td>26</td>
<td>0.073</td>
</tr>
</tbody>
</table>
Table 4.5: Laplace-asymmetric Laplace model: Genes identified as differentially expressed using a cut-off \( t_0 = 0.01 \) and their associated \( 1 - \omega_{\text{post}} \) values.

<table>
<thead>
<tr>
<th>Gene reference</th>
<th>Posterior probability of non-differential expression ( (1 - \omega_{\text{post}}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.000</td>
</tr>
<tr>
<td>1285</td>
<td>0.000</td>
</tr>
<tr>
<td>6419</td>
<td>0.000</td>
</tr>
<tr>
<td>2565</td>
<td>0.000</td>
</tr>
<tr>
<td>1287</td>
<td>0.000</td>
</tr>
<tr>
<td>2563</td>
<td>0.000</td>
</tr>
<tr>
<td>1283</td>
<td>0.000</td>
</tr>
<tr>
<td>3</td>
<td>0.000</td>
</tr>
<tr>
<td>3845</td>
<td>0.000</td>
</tr>
<tr>
<td>3847</td>
<td>0.000</td>
</tr>
<tr>
<td>3843</td>
<td>0.000</td>
</tr>
<tr>
<td>2567</td>
<td>0.000</td>
</tr>
<tr>
<td>6</td>
<td>0.000</td>
</tr>
<tr>
<td>2566</td>
<td>0.001</td>
</tr>
<tr>
<td>7059</td>
<td>0.007</td>
</tr>
<tr>
<td>1284</td>
<td>0.009</td>
</tr>
</tbody>
</table>
Figure 4.7: HIV data before and after thresholding using the Laplace-ASL model.

Table 4.6: Laplace-asymmetric Laplace model: Four further genes identified as differentially expressed when using a cut-off $t_0 = 0.1$ and their corresponding $1 - \omega_{post}$ values.

<table>
<thead>
<tr>
<th>Gene reference</th>
<th>Posterior probability of non-differential expression $(1 - \omega_{post})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3082</td>
<td>0.010</td>
</tr>
<tr>
<td>1286</td>
<td>0.017</td>
</tr>
<tr>
<td>1923</td>
<td>0.063</td>
</tr>
<tr>
<td>3842</td>
<td>0.076</td>
</tr>
</tbody>
</table>
differentially expressed if that choice of cut-off was to be selected and the corresponding implied FDR.

Table 4.7: Laplace-asymmetric Laplace model: Alternative choices for the cut-off $t_0$, the number of differentially expressed genes found using this $t_0$ and the corresponding implied FDR.

<table>
<thead>
<tr>
<th>Cut-off $t_0$</th>
<th>Number of differentially expressed genes</th>
<th>Associated FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>18</td>
<td>0.002</td>
</tr>
<tr>
<td>0.10</td>
<td>20</td>
<td>0.009</td>
</tr>
<tr>
<td>0.15</td>
<td>20</td>
<td>0.009</td>
</tr>
<tr>
<td>0.20</td>
<td>20</td>
<td>0.009</td>
</tr>
<tr>
<td>0.25</td>
<td>21</td>
<td>0.019</td>
</tr>
<tr>
<td>0.30</td>
<td>22</td>
<td>0.031</td>
</tr>
<tr>
<td>0.35</td>
<td>23</td>
<td>0.043</td>
</tr>
<tr>
<td>0.40</td>
<td>23</td>
<td>0.043</td>
</tr>
<tr>
<td>0.45</td>
<td>23</td>
<td>0.043</td>
</tr>
<tr>
<td>0.50</td>
<td>24</td>
<td>0.062</td>
</tr>
</tbody>
</table>

4.6.2 E.Coli data

For the Laplace-asymmetric Laplace model, the parameters were estimated to be $\hat{a} = 0.925$, $\hat{b} = 4.566$, $\hat{\kappa} = 1.000$ and $\hat{\omega} = 0.324$. Thus in this example, the more complex model is equivalent to the reduced symmetric Laplace case. The posterior median function is therefore antisymmetric and $t^{\pm} = \pm0.712$. Using these parameter estimates and thresholds, 16 (21.05%) genes were found to be differentially expressed with an implied FDR of 0.122. The contrast values before and after thresholding are illustrated in Figure 4.8.
Alternatively, using a cut-off $t_0 = 0.01$ for the $1 - \omega_{\text{post}}$ values, 4 (5.26%) genes were found to be differentially expressed with an implied FDR of 0.001. Using $t_0 = 0.1$ as the cut-off, 10 (13.16%) genes were differentially expressed with an implied FDR of 0.028. The genes identified as being differentially expressed, along with their corresponding values of $1 - \omega_{\text{post}}$ are shown in Tables 4.8 and 4.9 respectively.

### 4.6.3 Lymphoma data

The formulation of the lymphoma data is different to the other datasets, which brings into question the justification of applying the above described model. The lymphoma data were acquired from an experiment that employed a common reference design. Thus, we can define the contrast measurements for the two samples of interest, $T$ and $C$, compared with a reference sample, $R$, to be:
Table 4.8: Laplace-Laplace model: Genes identified as differentially expressed using a cut-off $t_0 = 0.01$ and their corresponding $1 - \omega_{\text{post}}$ values.

<table>
<thead>
<tr>
<th>Gene reference</th>
<th>Posterior probability of non-differential expression $(1 - \omega_{\text{post}})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>metE</td>
<td>0.000</td>
</tr>
<tr>
<td>GFP</td>
<td>0.000</td>
</tr>
<tr>
<td>metF</td>
<td>0.002</td>
</tr>
<tr>
<td>folE</td>
<td>0.003</td>
</tr>
</tbody>
</table>

Table 4.9: Laplace-Laplace model: Six further genes identified as differentially expressed when using a cut-off $t_0 = 0.1$ and their corresponding $1 - \omega_{\text{post}}$ values.

<table>
<thead>
<tr>
<th>Gene reference</th>
<th>Posterior probability of non-differential expression $(1 - \omega_{\text{post}})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>metL</td>
<td>0.024</td>
</tr>
<tr>
<td>cpsA</td>
<td>0.025</td>
</tr>
<tr>
<td>metK</td>
<td>0.028</td>
</tr>
<tr>
<td>metA</td>
<td>0.039</td>
</tr>
<tr>
<td>c23pUC</td>
<td>0.059</td>
</tr>
<tr>
<td>pUC</td>
<td>0.0981502</td>
</tr>
</tbody>
</table>
Chapter 4. Alternative posterior median estimators for sparse sequences

\[ X_{ij}^T \approx \log \left( \frac{T_{ij}}{R_{ij}} \right) \quad \text{and} \quad X_{ij}^C \approx \log \left( \frac{C_{ij}}{R_{ij}} \right), \]

where \( T_{ij}, C_{ij} \) and \( R_{ij} \) are the raw intensity measurements for the treatment, control and reference samples respectively. We use “\( \approx \)” rather than “=” as in practice, the raw measurements would have been background-corrected and normalised.

An indirect comparison of the level of differential expression between samples \( T \) and \( C \) for gene \( i \) can be found through the measurement

\[ \bar{X}_i^T - \bar{X}_i^C. \]

Whilst we have observed that it might be reasonable to model both \( \bar{X}_i^T \) and \( \bar{X}_i^C \) with asymmetric Laplace distributions, this does not necessarily hold for their difference.

Indeed, Kotz et al. (2005) show that for two independent Laplace variates \( X_1 \) and \( X_2 \) with densities of the form

\[ f_i(x) = \frac{a_i}{2} \exp(-a_i|x|) \quad i = 1, 2, \quad x \in \mathbb{R}, \]

the distribution of the difference \( X_1 - X_2 \) (or the summation \( X_1 + X_2 \) by symmetry) is

\[ f_{X_1+X_2}(x) = \begin{cases} \frac{1}{4a}(1 + |x|/a) \exp(-|x|/a) & \text{for } a_1 = a_2 = a \\ \frac{1}{2a_2} \frac{1}{a_1-\{a_1/a_2\}^2}(a_2 \exp(-|x|/a_2) - a_1 \exp(-|x|/a_1)) & \text{for } a_1 \neq a_2. \end{cases} \]

Knowing this, it can be seen that the Laplace-asymmetric Laplace model has potential to be useful for experiments of a common reference design, however the algebra would require some reworking. This also is applicable to single-channel experiments.

In spite of this, we apply the model in order to draw out any inference. The parameter estimates were as follows: \( \hat{\alpha} = 0.988, \hat{b} = 2.28, \hat{\kappa} = 1.00 \) and \( \hat{\omega} = 0.0012. \) Once
again, the model has reduced to the symmetric Laplace case. The parameters gave rise to thresholds $t^\pm = \pm 5.630$. No gene had an absolute contrast value greater than $t^+$, hence we conclude that no genes are differentially expressed. Referring back to Table 3.5, one can observe that the method of McLachlan et al. (2006) gave rise to the same conclusion. Knowing that our results are unjustified and highly questionable casts doubt on the success of the McLachlan et al. (2006) method for these data.

4.7 Simulations

Given that in most experiments we have no prior knowledge of the true noise distribution, a major identifiability problem is posed regarding the choice of model. It would be useful to ascertain whether any particular model consistently outperforms the others when the noise distribution is incorrectly specified. If so, in situations where the noise distribution is unknown, this prevailing model should be adopted by default as the “best”. Thus, we provide a variety of simulation studies to drive this conclusion. Amongst others, we assess the behaviour of the Laplace-Gaussian model in the presence of noise from an asymmetric Laplace distribution and the performance of the Laplace-asymmetric Laplace model in the presence of normally distributed noise. In each case, we consider sequences with a wide range of sparse behaviour and take into account the intensity of the signal relative to the noise.

We present a short summary after the results of each study, followed by an overall comparative conclusion at the end of the section.

4.7.1 Simulation 1: Efficiency of Laplace-Laplace model

This first simulation is designed to assess the accuracy of parameter estimation and the efficiency of the symmetric Laplace-Laplace model on data of such a form. We were
interested in the estimation of a sequence $\mu$ where $\mu_i = 0$ except for in $K$ randomly chosen positions, where it takes some value sampled from a Laplace distribution with a scaling parameter $a$ equal to 1. For each $i$, an “observation”, $Z_i$, is generated by corrupting the $\mu_i$ with noise sampled from a Laplace distribution with scaling parameter $b$. We used the Laplace-Laplace thresholding model to extract the sequence $\mu$ from the sequence of $Z_i$. We fixed the sample size $n = 1000$ and reported results for a wide range of sparse behaviour, that being $K = 50, 100, 250$ and $500$. We only considered situations where the signal had greater variance than the noise i.e. $a < b$, however this ranged from being only marginally so ($b = 1.1$), to substantially so ($b = 5$). We performed one hundred replications for each of the values of $K$ and $b$ and the results given are an average over these 100 replications. We report mean values of $\hat{a}$, $\hat{b}$ and $\hat{\omega}$ and the number of “interesting” non-zero values identified by the method. These results are given in Table 4.10. One should bear in mind that the signal identified only consists of the “large” values sampled from the Laplace distribution: some of the sampled signal may be relatively close to zero thus will too be thresholded along with the noisy values. Therefore, the values in the “Signal identified” column are expected to be less than the true signal sparsity.

The Laplace-Laplace model proves over the 100 simulations to be particularly efficient, regardless of the true signal sparsity, in cases where the signal is strong, relative to the noise. When the signal is weak (relative to the noise) and sparse, the parameter estimates are poor and many more observations are identified as “interesting” than we would have hoped. In cases where this is most extreme, one should note that the signal and noise have similar scaling parameters, again highlighting the need for the signal to be of substantially greater variance than the noise.
Table 4.10: Results from a simulation study designed to assess the performance of the symmetric Laplace-Laplace model.

<table>
<thead>
<tr>
<th>Noise scaling parameter</th>
<th>Signal sparsity (out of 1000)</th>
<th>Mean ( \hat{a} )</th>
<th>Mean ( \hat{b} )</th>
<th>Mean ( \hat{\omega} )</th>
<th>Signal identified</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>50</td>
<td>1.070</td>
<td>3.122</td>
<td>0.699</td>
<td>588.7</td>
</tr>
<tr>
<td>1.1</td>
<td>100</td>
<td>1.060</td>
<td>2.319</td>
<td>0.538</td>
<td>82.1</td>
</tr>
<tr>
<td>1.1</td>
<td>250</td>
<td>0.943</td>
<td>1.414</td>
<td>0.473</td>
<td>296.9</td>
</tr>
<tr>
<td>1.1</td>
<td>500</td>
<td>0.983</td>
<td>1.475</td>
<td>0.677</td>
<td>494.6</td>
</tr>
<tr>
<td>1.5</td>
<td>50</td>
<td>1.273</td>
<td>3.274</td>
<td>0.491</td>
<td>257.3</td>
</tr>
<tr>
<td>1.5</td>
<td>100</td>
<td>1.066</td>
<td>2.576</td>
<td>0.294</td>
<td>61.1</td>
</tr>
<tr>
<td>1.5</td>
<td>250</td>
<td>1.910</td>
<td>2.075</td>
<td>0.566</td>
<td>442.7</td>
</tr>
<tr>
<td>1.5</td>
<td>500</td>
<td>1.009</td>
<td>2.913</td>
<td>0.650</td>
<td>523.7</td>
</tr>
<tr>
<td>2</td>
<td>50</td>
<td>1.135</td>
<td>2.966</td>
<td>0.271</td>
<td>134.4</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>1.404</td>
<td>3.028</td>
<td>0.424</td>
<td>56.1</td>
</tr>
<tr>
<td>2</td>
<td>250</td>
<td>0.968</td>
<td>2.320</td>
<td>0.300</td>
<td>171</td>
</tr>
<tr>
<td>2</td>
<td>500</td>
<td>0.953</td>
<td>4.787</td>
<td>0.551</td>
<td>315.8</td>
</tr>
<tr>
<td>3</td>
<td>50</td>
<td>1.284</td>
<td>3.156</td>
<td>0.100</td>
<td>11.1</td>
</tr>
<tr>
<td>3</td>
<td>100</td>
<td>1.082</td>
<td>3.081</td>
<td>0.126</td>
<td>17.2</td>
</tr>
<tr>
<td>3</td>
<td>250</td>
<td>0.992</td>
<td>3.074</td>
<td>0.263</td>
<td>129.5</td>
</tr>
<tr>
<td>3</td>
<td>500</td>
<td>1.053</td>
<td>3.327</td>
<td>0.548</td>
<td>344.6</td>
</tr>
<tr>
<td>4</td>
<td>50</td>
<td>1.097</td>
<td>4.099</td>
<td>0.070</td>
<td>18.8</td>
</tr>
<tr>
<td>4</td>
<td>100</td>
<td>1.051</td>
<td>4.003</td>
<td>0.098</td>
<td>38.5</td>
</tr>
<tr>
<td>4</td>
<td>250</td>
<td>1.035</td>
<td>4.263</td>
<td>0.268</td>
<td>154.6</td>
</tr>
<tr>
<td>4</td>
<td>500</td>
<td>0.950</td>
<td>3.799</td>
<td>0.454</td>
<td>308.3</td>
</tr>
<tr>
<td>5</td>
<td>50</td>
<td>1.202</td>
<td>5.096</td>
<td>0.063</td>
<td>20.8</td>
</tr>
<tr>
<td>5</td>
<td>100</td>
<td>0.882</td>
<td>4.864</td>
<td>0.091</td>
<td>41.8</td>
</tr>
<tr>
<td>5</td>
<td>250</td>
<td>0.988</td>
<td>5.117</td>
<td>0.251</td>
<td>149.8</td>
</tr>
<tr>
<td>5</td>
<td>500</td>
<td>1.019</td>
<td>5.098</td>
<td>0.512</td>
<td>400.3</td>
</tr>
</tbody>
</table>
4.7.2 Simulation 2: Laplace-asymmetric Laplace model in the presence of asymmetric Laplace noise

This simulation, inspired by Johnstone and Silverman (2004), was designed to assess the capability of the Laplace-asymmetric Laplace model to adapt to sparsity in the true signal, in the presence of asymmetric Laplace noise. We created a sequence $\mu$ of length $n$, which has $\mu_i = 0$ except in $K$ randomly chosen positions, where it instead takes a non-zero signal value $\mu_i$. The non-zero values reported were 2, 2.5, 3, 4 and 5. We fixed $n$ to be equal to 1000 and investigated $K = 50, 100, 250, 500$. We also generated a sequence of 1000 noise values from a random sample of the control-control data, which we hoped would reflect the true nature of microarray noise. The true range of the control-control data is (-3.12, 2.93), thus one can see that non-zero values under investigation ranged from being embedded heavily in the noise to being sufficiently distinct. A sequence of “observations” was generated by combining the signal and noise vectors. We then used the Laplace-asymmetric Laplace model to threshold the observations to estimate the true underlying sequence $\mu$. One hundred replications were performed for each of the non-zero signal values and $K$. For each replication, the total squared error of the estimation $\sum(\hat{\mu}_i - \mu_i)^2$, and the numbers of misclassified signal and noise entries were recorded. Misclassified signal describes true non-zero $\mu_i$ that are incorrectly concluded to be zero by the thresholding procedure; misclassified noise describes truly zero $\mu_i$ that are incorrectly concluded to be non-zero by the thresholding procedure. The results given in Tables 4.11 and 4.12 are an average over these 100 replications. The same 100,000 noise variables were used for each set of the replications and the true mean values of $b$ and $\kappa$ were found to be 4.407 and 1.165 respectively.
Table 4.11: Results from a simulation study to test the efficiency of the Laplace-asymmetric Laplace model in presence of asymmetric Laplace noise.

<table>
<thead>
<tr>
<th>Signal sparsity</th>
<th>Spiked-in values</th>
<th>$\hat{w}$</th>
<th>$\hat{b}$</th>
<th>$\hat{\kappa}$</th>
<th>Total squared error</th>
<th>Misclassified signal</th>
<th>Misclassified noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>2</td>
<td>0.016</td>
<td>4.461</td>
<td>1.171</td>
<td>11.083</td>
<td>1.57</td>
<td>1.17</td>
</tr>
<tr>
<td>5</td>
<td>2.5</td>
<td>0.016</td>
<td>4.492</td>
<td>1.166</td>
<td>5.949</td>
<td>0.06</td>
<td>1.39</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>0.014</td>
<td>4.493</td>
<td>1.164</td>
<td>5.467</td>
<td>0.02</td>
<td>1.26</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>0.011</td>
<td>4.478</td>
<td>1.163</td>
<td>4.954</td>
<td>0.00</td>
<td>0.98</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>0.009</td>
<td>4.467</td>
<td>1.164</td>
<td>4.643</td>
<td>0.00</td>
<td>0.83</td>
</tr>
<tr>
<td>50</td>
<td>2</td>
<td>0.123</td>
<td>4.788</td>
<td>1.162</td>
<td>20.757</td>
<td>0.16</td>
<td>10.86</td>
</tr>
<tr>
<td>50</td>
<td>2.5</td>
<td>0.102</td>
<td>4.760</td>
<td>1.153</td>
<td>18.781</td>
<td>0.03</td>
<td>8.24</td>
</tr>
<tr>
<td>50</td>
<td>3</td>
<td>0.089</td>
<td>4.717</td>
<td>1.156</td>
<td>17.353</td>
<td>0.00</td>
<td>6.29</td>
</tr>
<tr>
<td>50</td>
<td>4</td>
<td>0.076</td>
<td>4.655</td>
<td>1.157</td>
<td>15.627</td>
<td>0.00</td>
<td>4.51</td>
</tr>
<tr>
<td>50</td>
<td>5</td>
<td>0.070</td>
<td>4.616</td>
<td>1.158</td>
<td>14.478</td>
<td>0.00</td>
<td>3.62</td>
</tr>
<tr>
<td>100</td>
<td>2</td>
<td>0.221</td>
<td>5.075</td>
<td>1.150</td>
<td>32.916</td>
<td>0.16</td>
<td>26.5</td>
</tr>
<tr>
<td>100</td>
<td>2.5</td>
<td>0.186</td>
<td>4.974</td>
<td>1.148</td>
<td>29.672</td>
<td>0.05</td>
<td>17.11</td>
</tr>
<tr>
<td>100</td>
<td>3</td>
<td>0.165</td>
<td>4.893</td>
<td>1.149</td>
<td>27.379</td>
<td>0.01</td>
<td>12.21</td>
</tr>
<tr>
<td>100</td>
<td>4</td>
<td>0.144</td>
<td>4.790</td>
<td>1.152</td>
<td>24.527</td>
<td>0.00</td>
<td>8.28</td>
</tr>
<tr>
<td>100</td>
<td>5</td>
<td>0.134</td>
<td>4.741</td>
<td>1.153</td>
<td>22.935</td>
<td>0.00</td>
<td>6.28</td>
</tr>
</tbody>
</table>
Table 4.12: Results from a simulation study to test the efficiency of the Laplace-asymmetric Laplace model in the presence of asymmetric Laplace noise.

<table>
<thead>
<tr>
<th>Signal sparsity (out of 1000)</th>
<th>spiked-in values</th>
<th>$\hat{w}$</th>
<th>$\hat{b}$</th>
<th>$\hat{\kappa}$</th>
<th>Total squared error</th>
<th>Misclassified signal</th>
<th>Misclassified noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>250</td>
<td>2</td>
<td>0.464</td>
<td>5.837</td>
<td>1.132</td>
<td>61.277</td>
<td>0.31</td>
<td>91.43</td>
</tr>
<tr>
<td>250</td>
<td>2.5</td>
<td>0.408</td>
<td>5.553</td>
<td>1.133</td>
<td>55.215</td>
<td>0.07</td>
<td>54.37</td>
</tr>
<tr>
<td>250</td>
<td>3</td>
<td>0.374</td>
<td>5.359</td>
<td>1.135</td>
<td>51.013</td>
<td>0.01</td>
<td>37.69</td>
</tr>
<tr>
<td>250</td>
<td>4</td>
<td>0.337</td>
<td>5.152</td>
<td>1.140</td>
<td>45.987</td>
<td>0.00</td>
<td>22.44</td>
</tr>
<tr>
<td>250</td>
<td>5</td>
<td>0.317</td>
<td>4.999</td>
<td>1.143</td>
<td>43.063</td>
<td>0.00</td>
<td>15.84</td>
</tr>
<tr>
<td>500</td>
<td>2</td>
<td>0.557</td>
<td>1.815</td>
<td>1.019</td>
<td>1150.845</td>
<td>62.61</td>
<td>23.80</td>
</tr>
<tr>
<td>500</td>
<td>2.5</td>
<td>0.514</td>
<td>1.666</td>
<td>1.006</td>
<td>2009.636</td>
<td>147.87</td>
<td>19.23</td>
</tr>
<tr>
<td>500</td>
<td>3</td>
<td>0.538</td>
<td>3.350</td>
<td>1.038</td>
<td>1969.789</td>
<td>179.48</td>
<td>47.59</td>
</tr>
<tr>
<td>500</td>
<td>4</td>
<td>0.635</td>
<td>5.781</td>
<td>1.097</td>
<td>77.857</td>
<td>0.00</td>
<td>60.84</td>
</tr>
<tr>
<td>500</td>
<td>5</td>
<td>0.607</td>
<td>5.412</td>
<td>1.086</td>
<td>76.360</td>
<td>0.00</td>
<td>40.68</td>
</tr>
</tbody>
</table>
The results suggest that the model is at its most efficient, in terms of the total squared error and parameter estimation, when the signal is sparse, yet strong. Within each “sparsity level”, it is clear that the performance of the model generally improves as the strength of the signal increases. However, even when the signal is weak, relatively good performance is observed, providing the sequence is not too sparse. As the sequence becomes less sparse, the misclassified signal values remain largely unaffected (except when \( K = 500 \)), yet the misclassified noise values increase greatly, despite the presence of fewer true zero values in the sequence.

### 4.7.3 Simulation 3: Laplace-asymmetric Laplace model in the presence of normally distributed noise

Here we consider the flexibility of the Laplace-asymmetric Laplace model. We assess how well it performs when in fact the model is misspecified and the noise actually follows a normal distribution. As previously, we created a sequence \( \mu \) and fixed \( n = 1000 \) and \( K = 50, 100, 250, 500 \). The non-zero values reported were 2, 2.5, 3, 4 and 5. A sequence of observations was generated as in the previous simulation. We then used the Laplace-asymmetric Laplace model to threshold the vector of observations, to obtain the sequence \( \mu \). One hundred replications were performed for each of the non-zero signal values and \( K \). For each replication, the total squared error of the estimation and the numbers of misclassified signal and noise entries were recorded. The results given in Tables 4.13 are an average over these 100 replications. As before, the same 100,000 noise variables were used for each set of the replications.

Applying the Laplace-asymmetric Laplace model when the noise is truly normally distributed does not yield desirable results. The mixing weight is consistently overestimated. When the sequence is sparse, the majority of the true signal is misclassified as noise; when the sequence is less sparse, the majority of noise is misclassified as signal.
Table 4.13: Results from a simulation study to test the efficiency of the Laplace-asymmetric Laplace model in the presence of normally distributed noise.

<table>
<thead>
<tr>
<th>Signal sparsity (out of 1000)</th>
<th>Spiked-in values</th>
<th>Mean $\hat{w}$</th>
<th>Total squared error</th>
<th>Misclassified signal</th>
<th>Misclassified noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>2</td>
<td>0.035</td>
<td>19.978</td>
<td>4.89</td>
<td>0.49</td>
</tr>
<tr>
<td>5</td>
<td>2.5</td>
<td>0.036</td>
<td>30.952</td>
<td>4.84</td>
<td>0.55</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>0.032</td>
<td>43.716</td>
<td>4.71</td>
<td>0.45</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>0.039</td>
<td>73.296</td>
<td>4.39</td>
<td>0.61</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>0.047</td>
<td>108.019</td>
<td>4.09</td>
<td>0.79</td>
</tr>
<tr>
<td>50</td>
<td>2</td>
<td>0.065</td>
<td>196.104</td>
<td>48.18</td>
<td>0.87</td>
</tr>
<tr>
<td>50</td>
<td>2.5</td>
<td>0.110</td>
<td>297.552</td>
<td>45.09</td>
<td>0.73</td>
</tr>
<tr>
<td>50</td>
<td>3</td>
<td>0.159</td>
<td>396.939</td>
<td>37.92</td>
<td>1.77</td>
</tr>
<tr>
<td>50</td>
<td>4</td>
<td>0.245</td>
<td>524.570</td>
<td>18.83</td>
<td>4.31</td>
</tr>
<tr>
<td>50</td>
<td>5</td>
<td>0.131</td>
<td>356.993</td>
<td>8.92</td>
<td>3.70</td>
</tr>
<tr>
<td>100</td>
<td>2</td>
<td>0.173</td>
<td>385.078</td>
<td>91.81</td>
<td>2.02</td>
</tr>
<tr>
<td>100</td>
<td>2.5</td>
<td>0.352</td>
<td>582.242</td>
<td>65.41</td>
<td>101.06</td>
</tr>
<tr>
<td>100</td>
<td>3</td>
<td>0.458</td>
<td>722.226</td>
<td>41.41</td>
<td>192.41</td>
</tr>
<tr>
<td>100</td>
<td>4</td>
<td>0.352</td>
<td>535.817</td>
<td>11.00</td>
<td>46.40</td>
</tr>
<tr>
<td>100</td>
<td>5</td>
<td>0.267</td>
<td>253.241</td>
<td>1.79</td>
<td>14.04</td>
</tr>
<tr>
<td>250</td>
<td>2</td>
<td>0.758</td>
<td>786.536</td>
<td>75.58</td>
<td>471.34</td>
</tr>
<tr>
<td>250</td>
<td>2.5</td>
<td>0.544</td>
<td>1126.228</td>
<td>102.96</td>
<td>198.91</td>
</tr>
<tr>
<td>250</td>
<td>3</td>
<td>0.610</td>
<td>1273.602</td>
<td>36.10</td>
<td>190.66</td>
</tr>
<tr>
<td>250</td>
<td>4</td>
<td>0.627</td>
<td>1089.418</td>
<td>27.14</td>
<td>269.42</td>
</tr>
<tr>
<td>250</td>
<td>5</td>
<td>0.616</td>
<td>574.353</td>
<td>0.03</td>
<td>243.12</td>
</tr>
<tr>
<td>500</td>
<td>2</td>
<td>0.870</td>
<td>933.542</td>
<td>57.48</td>
<td>377.49</td>
</tr>
<tr>
<td>500</td>
<td>2.5</td>
<td>0.975</td>
<td>1091.948</td>
<td>9.37</td>
<td>472.58</td>
</tr>
<tr>
<td>500</td>
<td>3</td>
<td>0.989</td>
<td>1371.184</td>
<td>0.25</td>
<td>484.78</td>
</tr>
<tr>
<td>500</td>
<td>4</td>
<td>0.947</td>
<td>2423.217</td>
<td>5.33</td>
<td>429.72</td>
</tr>
<tr>
<td>500</td>
<td>5</td>
<td>0.857</td>
<td>2745.316</td>
<td>0.04</td>
<td>281.26</td>
</tr>
</tbody>
</table>
Neither of these of scenarios would be useful for real-life analysis purposes.

### 4.7.4 Simulation 4: Laplace-Gaussian model in the presence of asymmetric Laplace noise

Here we consider the flexibility of the Laplace-Gaussian model and assess how well it performs when in fact the model is misspecified and the noise actually follows an asymmetric Laplace distribution. As previously, we fixed $n = 1000$, $K = 50, 100, 250, 500$ and the non-zero values reported were 2, 2.5, 3, 4 and 5. Three distinct noise vectors were generated from random samples of the control-control data. Each of these were combined with the signal vector: the aim being to represent observations from 3 replicate chips. We estimated the standard deviation using the linear additive model approach in Section 3.4 and went on to apply the Laplace-Gaussian model, in conjunction with the value of $\sigma_\varepsilon$. Further to this, we applied the adjusted Laplace-Gaussian model, with $s_0$ estimated from the replicates. One hundred replications were performed for each of the non-zero signal values and $K$. For each replication, the total squared error of the estimation and the numbers of misclassified signal and noise entries were recorded for both the unadjusted and adjusted model. The results given in Tables 4.14 and 4.15 are an average over these 100 replications. The same 300,000 noise variables were used for each set of the replications.

For the unadjusted model, the mixing weight is consistently overestimated however generally, the estimate improves as the signal strength increases. As expected from previous discussion, the unadjusted Laplace-Gaussian thresholding procedure is too conservative for use on asymmetric Laplace noise and the noise is not completely removed. Generally, the performance is improved for stronger signals.

In the adjusted case, the model is efficient at removing noise, particularly when the signal is very sparse. However, when the signal is also weak, this is at the expense of
Table 4.14: Results from a simulation study to test the efficiency of the unadjusted Laplace-Gaussian model in the presence of asymmetric Laplace noise.

<table>
<thead>
<tr>
<th>Signal sparsity (out of 1000)</th>
<th>Spiked-in values</th>
<th>Mean $\hat{w}$</th>
<th>Total squared error</th>
<th>Misclassified signal</th>
<th>Misclassified noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>2</td>
<td>0.089</td>
<td>5.105547</td>
<td>0</td>
<td>42.43</td>
</tr>
<tr>
<td>5</td>
<td>2.5</td>
<td>0.039</td>
<td>3.915680</td>
<td>0</td>
<td>17.36</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>0.126</td>
<td>5.543318</td>
<td>0</td>
<td>49.79</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>0.055</td>
<td>3.697996</td>
<td>0</td>
<td>6.95</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>0.019</td>
<td>3.041968</td>
<td>0</td>
<td>5.14</td>
</tr>
<tr>
<td>50</td>
<td>2</td>
<td>0.154</td>
<td>6.587</td>
<td>0</td>
<td>12.56</td>
</tr>
<tr>
<td>50</td>
<td>2.5</td>
<td>0.081</td>
<td>6.039</td>
<td>0</td>
<td>10.03</td>
</tr>
<tr>
<td>50</td>
<td>3</td>
<td>0.130</td>
<td>6.122</td>
<td>0</td>
<td>10.16</td>
</tr>
<tr>
<td>50</td>
<td>4</td>
<td>0.069</td>
<td>5.478</td>
<td>0</td>
<td>7.66</td>
</tr>
<tr>
<td>50</td>
<td>5</td>
<td>0.066</td>
<td>5.222</td>
<td>0</td>
<td>6.78</td>
</tr>
<tr>
<td>100</td>
<td>2</td>
<td>0.157</td>
<td>9.285</td>
<td>0</td>
<td>16.07</td>
</tr>
<tr>
<td>100</td>
<td>2.5</td>
<td>0.145</td>
<td>8.896</td>
<td>0</td>
<td>13.52</td>
</tr>
<tr>
<td>100</td>
<td>3</td>
<td>0.137</td>
<td>8.608</td>
<td>0</td>
<td>11.84</td>
</tr>
<tr>
<td>100</td>
<td>4</td>
<td>0.128</td>
<td>8.183</td>
<td>0</td>
<td>9.89</td>
</tr>
<tr>
<td>100</td>
<td>5</td>
<td>0.123</td>
<td>7.851</td>
<td>0</td>
<td>8.59</td>
</tr>
<tr>
<td>250</td>
<td>2</td>
<td>0.518</td>
<td>17.959</td>
<td>0</td>
<td>52.64</td>
</tr>
<tr>
<td>250</td>
<td>2.5</td>
<td>0.886</td>
<td>22.514</td>
<td>0</td>
<td>123.87</td>
</tr>
<tr>
<td>250</td>
<td>3</td>
<td>0.332</td>
<td>15.192</td>
<td>0</td>
<td>19.45</td>
</tr>
<tr>
<td>250</td>
<td>4</td>
<td>0.299</td>
<td>14.423</td>
<td>0</td>
<td>14.20</td>
</tr>
<tr>
<td>250</td>
<td>5</td>
<td>0.290</td>
<td>13.961</td>
<td>0</td>
<td>12.01</td>
</tr>
<tr>
<td>500</td>
<td>2</td>
<td>0.646</td>
<td>25.945</td>
<td>0</td>
<td>45.23</td>
</tr>
<tr>
<td>500</td>
<td>2.5</td>
<td>0.617</td>
<td>25.166</td>
<td>0</td>
<td>31.40</td>
</tr>
<tr>
<td>500</td>
<td>3</td>
<td>0.598</td>
<td>24.647</td>
<td>0</td>
<td>25.21</td>
</tr>
<tr>
<td>500</td>
<td>4</td>
<td>0.574</td>
<td>23.910</td>
<td>0</td>
<td>18.43</td>
</tr>
<tr>
<td>500</td>
<td>5</td>
<td>0.688</td>
<td>26.985</td>
<td>0</td>
<td>125.23</td>
</tr>
</tbody>
</table>
Table 4.15: Results from a simulation study to test the efficiency of the adjusted Laplace-Gaussian model in the presence of asymmetric Laplace noise.

<table>
<thead>
<tr>
<th>Signal sparsity (out of 1000)</th>
<th>Spiked-in values</th>
<th>Mean $\hat{w}$</th>
<th>Total squared error</th>
<th>Misclassified signal</th>
<th>Misclassified noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>2</td>
<td>0.009</td>
<td>14.466</td>
<td>3.51</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>2.5</td>
<td>0.008</td>
<td>15.310</td>
<td>2.41</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>0.011</td>
<td>20.149</td>
<td>2.29</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>0.009</td>
<td>14.138</td>
<td>0.80</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>0.025</td>
<td>1.515</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td>50</td>
<td>2</td>
<td>0.057</td>
<td>100.505</td>
<td>23.84</td>
<td>9.50</td>
</tr>
<tr>
<td>50</td>
<td>2.5</td>
<td>0.068</td>
<td>143.670</td>
<td>22.43</td>
<td>9.50</td>
</tr>
<tr>
<td>50</td>
<td>3</td>
<td>0.128</td>
<td>146.378</td>
<td>14.98</td>
<td>29.30</td>
</tr>
<tr>
<td>50</td>
<td>4</td>
<td>0.110</td>
<td>52.012</td>
<td>2.23</td>
<td>4.21</td>
</tr>
<tr>
<td>50</td>
<td>5</td>
<td>0.079</td>
<td>5.798</td>
<td>0.00</td>
<td>9.50</td>
</tr>
<tr>
<td>100</td>
<td>2</td>
<td>0.091</td>
<td>192.917</td>
<td>46.55</td>
<td>0.08</td>
</tr>
<tr>
<td>100</td>
<td>2.5</td>
<td>0.093</td>
<td>256.677</td>
<td>38.60</td>
<td>0.14</td>
</tr>
<tr>
<td>100</td>
<td>3</td>
<td>0.098</td>
<td>241.592</td>
<td>22.56</td>
<td>0.14</td>
</tr>
<tr>
<td>100</td>
<td>4</td>
<td>0.143</td>
<td>43.725</td>
<td>0.30</td>
<td>0.14</td>
</tr>
<tr>
<td>100</td>
<td>5</td>
<td>0.121</td>
<td>9.600</td>
<td>0.00</td>
<td>0.18</td>
</tr>
<tr>
<td>250</td>
<td>2</td>
<td>0.327</td>
<td>625.312</td>
<td>122.04</td>
<td>4.98</td>
</tr>
<tr>
<td>250</td>
<td>2.5</td>
<td>0.446</td>
<td>584.716</td>
<td>52.38</td>
<td>7.09</td>
</tr>
<tr>
<td>250</td>
<td>3</td>
<td>0.483</td>
<td>332.353</td>
<td>2.31</td>
<td>5.16</td>
</tr>
<tr>
<td>250</td>
<td>4</td>
<td>0.404</td>
<td>75.174</td>
<td>0.00</td>
<td>2.25</td>
</tr>
<tr>
<td>250</td>
<td>5</td>
<td>0.272</td>
<td>31.218</td>
<td>0.00</td>
<td>1.80</td>
</tr>
<tr>
<td>500</td>
<td>2</td>
<td>0.990</td>
<td>842.597</td>
<td>5.80</td>
<td>485.73</td>
</tr>
<tr>
<td>500</td>
<td>2.5</td>
<td>0.997</td>
<td>542.648</td>
<td>0.00</td>
<td>491.75</td>
</tr>
<tr>
<td>500</td>
<td>3</td>
<td>0.975</td>
<td>345.915</td>
<td>0.00</td>
<td>401.07</td>
</tr>
<tr>
<td>500</td>
<td>4</td>
<td>0.842</td>
<td>156.695</td>
<td>0.00</td>
<td>85.66</td>
</tr>
<tr>
<td>500</td>
<td>5</td>
<td>0.726</td>
<td>82.360</td>
<td>0.00</td>
<td>0.09</td>
</tr>
</tbody>
</table>
misclassifying much of the signal. This observation suggests that the adjustment is too extreme as previously hypothesised. When the sequence contains half signal and the true signal is weak, the model performs poorly, retaining much of the noise.

### 4.7.5 Simulation 5: Laplace-Gaussian model in the presence of normal noise

Johnstone and Silverman (2004) considered the capability of the Laplace-Gaussian model when the noise is normally distributed. However here, our primary interest is in assessing how adaptive the adjusted Laplace-Gaussian model is: when the noise is truly normally distributed it would be highly useful if $s_0$ were estimated to be small. As previously, we fixed $n = 1000$ and $K = 50, 100, 250, 500$. The non-zero values reported were 2, 2.5, 3, 4 and 5. Three distinct noise vectors were generated with a mean of zero and standard deviation equal to 1. Once again, observations from replicate chips were simulated by combining each of the 3 noise vectors with the signal vector. For completeness, we estimated the standard deviation using the linear additive model approach in Section 3.4 and went on to apply the Laplace-Gaussian model, in conjunction with the value of $\sigma$. Of more interest, we applied the adjusted Laplace-Gaussian model, with $s_0$ estimated from the replicates. One hundred replications were performed for each of the non-zero signal values and $K$. For each replication, the total squared error of the estimation and the numbers of misclassified signal and noise entries were recorded for both the unadjusted and adjusted model. The results given in Tables 4.16 and 4.17 are an average over these 100 replications. The same 300,000 noise variables were used for each set of the replications.

The unadjusted model performs relatively well when the sequence is sparse and strong. As the sparsity lessens, whilst most signal is classified correctly, the model does not effectively remove all the noise. The adjusted model, as expected, generally
Table 4.16: Results from a simulation study to test the efficiency of the unadjusted Laplace-Gaussian model in the presence of normally distributed noise.

<table>
<thead>
<tr>
<th>Signal sparsity (out of 1000)</th>
<th>Spiked-in values</th>
<th>Mean $\hat{w}$</th>
<th>Total squared error</th>
<th>Misclassified signal</th>
<th>Misclassified noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>2</td>
<td>0.071</td>
<td>11.952</td>
<td>2.38</td>
<td>1.59</td>
</tr>
<tr>
<td>5</td>
<td>2.5</td>
<td>0.037</td>
<td>8.607</td>
<td>0.83</td>
<td>1.16</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>0.026</td>
<td>5.054</td>
<td>0.18</td>
<td>0.88</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>0.021</td>
<td>3.120</td>
<td>0.00</td>
<td>0.51</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>0.018</td>
<td>2.543</td>
<td>0.00</td>
<td>0.35</td>
</tr>
<tr>
<td>50</td>
<td>2</td>
<td>0.204</td>
<td>62.499</td>
<td>8.11</td>
<td>12.93</td>
</tr>
<tr>
<td>50</td>
<td>2.5</td>
<td>0.161</td>
<td>42.585</td>
<td>1.92</td>
<td>9.67</td>
</tr>
<tr>
<td>50</td>
<td>3</td>
<td>0.131</td>
<td>29.636</td>
<td>0.22</td>
<td>6.83</td>
</tr>
<tr>
<td>50</td>
<td>4</td>
<td>0.096</td>
<td>22.500</td>
<td>0.00</td>
<td>3.55</td>
</tr>
<tr>
<td>50</td>
<td>5</td>
<td>0.111</td>
<td>21.038</td>
<td>0.00</td>
<td>2.43</td>
</tr>
<tr>
<td>100</td>
<td>2</td>
<td>0.347</td>
<td>99.520</td>
<td>8.16</td>
<td>35.85</td>
</tr>
<tr>
<td>100</td>
<td>2.5</td>
<td>0.278</td>
<td>72.674</td>
<td>1.85</td>
<td>22.88</td>
</tr>
<tr>
<td>100</td>
<td>3</td>
<td>0.231</td>
<td>55.671</td>
<td>0.21</td>
<td>15.24</td>
</tr>
<tr>
<td>100</td>
<td>4</td>
<td>0.179</td>
<td>46.113</td>
<td>0.00</td>
<td>8.42</td>
</tr>
<tr>
<td>100</td>
<td>5</td>
<td>0.156</td>
<td>42.874</td>
<td>0.00</td>
<td>5.55</td>
</tr>
<tr>
<td>250</td>
<td>2</td>
<td>0.713</td>
<td>183.386</td>
<td>1.92</td>
<td>236.37</td>
</tr>
<tr>
<td>250</td>
<td>2.5</td>
<td>0.579</td>
<td>151.743</td>
<td>0.54</td>
<td>102.87</td>
</tr>
<tr>
<td>250</td>
<td>3</td>
<td>0.502</td>
<td>128.990</td>
<td>0.06</td>
<td>56.40</td>
</tr>
<tr>
<td>250</td>
<td>4</td>
<td>0.918</td>
<td>168.244</td>
<td>0.00</td>
<td>146.48</td>
</tr>
<tr>
<td>250</td>
<td>5</td>
<td>0.918</td>
<td>167.459</td>
<td>0.00</td>
<td>143.72</td>
</tr>
<tr>
<td>500</td>
<td>2</td>
<td>1.000</td>
<td>289.561</td>
<td>0.00</td>
<td>500.00</td>
</tr>
<tr>
<td>500</td>
<td>2.5</td>
<td>0.992</td>
<td>289.324</td>
<td>0.00</td>
<td>485.99</td>
</tr>
<tr>
<td>500</td>
<td>3</td>
<td>0.876</td>
<td>254.633</td>
<td>0.00</td>
<td>276.28</td>
</tr>
<tr>
<td>500</td>
<td>4</td>
<td>0.747</td>
<td>219.290</td>
<td>0.00</td>
<td>100.81</td>
</tr>
<tr>
<td>500</td>
<td>5</td>
<td>0.685</td>
<td>204.399</td>
<td>0.00</td>
<td>56.24</td>
</tr>
</tbody>
</table>
Table 4.17: Results from a simulation study to test the efficiency of the adjusted Laplace-Gaussian model in the presence of normally distributed noise.

<table>
<thead>
<tr>
<th>Signal sparsity (out of 1000)</th>
<th>Spiked-in values</th>
<th>Mean $\hat{w}$</th>
<th>Total squared error</th>
<th>Misclassified signal</th>
<th>Misclassified noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>2</td>
<td>0.009</td>
<td>20</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>2.5</td>
<td>0.009</td>
<td>31.25</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>0.009</td>
<td>45</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>0.009</td>
<td>80</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>0.009</td>
<td>125</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>50</td>
<td>2</td>
<td>0.009</td>
<td>200</td>
<td>50</td>
<td>0</td>
</tr>
<tr>
<td>50</td>
<td>2.5</td>
<td>0.009</td>
<td>312.5</td>
<td>50</td>
<td>0</td>
</tr>
<tr>
<td>50</td>
<td>3</td>
<td>0.009</td>
<td>450</td>
<td>50</td>
<td>0</td>
</tr>
<tr>
<td>50</td>
<td>4</td>
<td>0.009</td>
<td>800</td>
<td>50</td>
<td>0</td>
</tr>
<tr>
<td>50</td>
<td>5</td>
<td>0.010</td>
<td>1250</td>
<td>50</td>
<td>0</td>
</tr>
<tr>
<td>100</td>
<td>2</td>
<td>0.009</td>
<td>400</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>100</td>
<td>2.5</td>
<td>0.009</td>
<td>625</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>100</td>
<td>3</td>
<td>0.010</td>
<td>899.647</td>
<td>99.95</td>
<td>0</td>
</tr>
<tr>
<td>100</td>
<td>4</td>
<td>0.012</td>
<td>1583.458</td>
<td>98.89</td>
<td>0</td>
</tr>
<tr>
<td>100</td>
<td>5</td>
<td>0.023</td>
<td>2405.787</td>
<td>95.91</td>
<td>0.03</td>
</tr>
<tr>
<td>250</td>
<td>2</td>
<td>0.009</td>
<td>1000</td>
<td>250</td>
<td>0</td>
</tr>
<tr>
<td>250</td>
<td>2.5</td>
<td>0.010</td>
<td>1562.5</td>
<td>250</td>
<td>0</td>
</tr>
<tr>
<td>250</td>
<td>3</td>
<td>0.033</td>
<td>2221.213</td>
<td>242.78</td>
<td>0</td>
</tr>
<tr>
<td>250</td>
<td>4</td>
<td>0.200</td>
<td>3424.817</td>
<td>201.17</td>
<td>0.04</td>
</tr>
<tr>
<td>250</td>
<td>5</td>
<td>0.463</td>
<td>2744.84</td>
<td>30.21</td>
<td>0.02</td>
</tr>
<tr>
<td>500</td>
<td>2</td>
<td>0.010</td>
<td>2000</td>
<td>500</td>
<td>0</td>
</tr>
<tr>
<td>500</td>
<td>2.5</td>
<td>0.125</td>
<td>3125</td>
<td>500</td>
<td>0</td>
</tr>
<tr>
<td>500</td>
<td>3</td>
<td>0.446</td>
<td>4266.442</td>
<td>370.12</td>
<td>95.75</td>
</tr>
<tr>
<td>500</td>
<td>4</td>
<td>1.000</td>
<td>3594.93</td>
<td>0</td>
<td>500</td>
</tr>
<tr>
<td>500</td>
<td>5</td>
<td>1.000</td>
<td>2302.02</td>
<td>0</td>
<td>500</td>
</tr>
</tbody>
</table>
overcompensates and most of the true signal is lost in the bid to remove the noise.

4.7.6 Discussion

With regard to resolving the identifiability problem within the microarray context, we can reject the adjusted Laplace-Gaussian model due to its abysmal performance when the noise is truly normally distributed. Consequently, we are left with two remaining options: the too conservative unadjusted Laplace-Gaussian model, which, in the presence of a incorrectly specified noise distribution, effectively classifies the signal but at the expense of retaining some noise, or the Laplace-asymmetric Laplace model, which, for sparse sequences, is more efficient at removing the noise, but in doing so loses much of the interesting signal. (It is worth noting that for less sparse sequences, the Laplace-asymmetric Laplace model actually misclassifies more noise than the Laplace-Gaussian model). Whilst our aim was to remove all of the noise, considering the relatively small numbers of misclassified noise incurred by the Laplace-Gaussian model in the sparse case, we recommend adopting this model as the default. For the sake of so few misclassified noise observations, it is not worth throwing away the majority of “interesting” genes.

Johnstone and Silverman (2004) discussed a possible modification of the threshold estimator for the Laplace-Gaussian model, which would allow a reduction in the numbers of misclassified noise in very sparse cases. If the estimated threshold is near or at the universal threshold, it is replaced by a higher threshold of the form

$$
\sqrt{2(1 + A) \log n}.
$$

(4.31)

The choice of $A$ is not necessarily clear, though Johnstone and Silverman (2004) arbitrarily adopt $A = 1$ for the purposes of demonstration. In actual fact, the merits of such a modification were reported to be limited: clear improvements were only seen when a sequence of length $n = 1000$ consisted of only a singular very large non-zero parameter
amongst the zeros. Particularly for microarray data, where it is usually expected that the sequence is sparse, this extension warrants further investigation. Any improvement gained through such a modification should further support the choice of Laplace-Gaussian model as the default method.
Chapter 5

Conclusions and directions for future work

5.1 Conclusions

In Chapter 2, we gave a brief overview of the need for normalisation and some of the more common approaches to the problem. We demonstrated their effectiveness visually with some of our datasets. We suggested that the log2 transformation usually adopted by scientists within the field is a rather arbitrary choice and showed that this is not the most appropriate transformation in all cases; we achieved superior results from normalisation when alternative transformations are incorporated.

In Chapter 3, we proposed an empirical Bayesian thresholding procedure for the purposes of analysing gene expression data. We showed that when the assumption of a normal error distribution is appropriate, using the sum-of-squares approach presented in Section 3.4 to estimate $\sigma^2_\varepsilon$ is more robust than the $\text{mad}(z)$ estimator used by Johnstone and Silverman (2004) for datasets with large proportions of differentially expressed genes.

After suitable normalisation and variance stabilisation steps, the method presented in
Chapter 3 works and is useful, however there are grounds for improvement. It is clear that inflating the variance can potentially make a large difference to the number of differentially expressed genes that the method identifies. Many variations exist for the derivation of $s_0$; we adopted a practice similar to Tusher et al. (2001) for its computational ease and because it seemed to be a good approximate fit for the spike-in HIV dataset. However, whether this adjustment can be said to be appropriate in all cases is a very difficult question to answer and not the most pertinent to the problem. We would suggest, based on visual evidence, that in fact, the modification overcompensates for the heavy-tailed behaviour of the error distribution; that $s_0$ is slightly too large, and that the true numbers of differentially expressed genes actually lie somewhere in between the two cases. Whilst we could investigate new alternatives for the value of $s_0$ in the future, this approach will never be sufficient compensation for assuming an “incorrect” model. The spike-in HIV data considered here suggested that a heavier-tailed distribution would be a more appropriate assumption for the true nature of the error distribution and investigating this further formed the essence of Chapter 4.

In Chapter 4, the aim was to identify a model that would allow for heavy-tailed behaviour in the error distribution whilst retaining the attractive thresholding and shrinkage properties of the Laplace-Gaussian model in Chapter 3. We reviewed various models which had potential for the purpose of estimating sparse sequences and discussed their limitations within a gene expression context. We stated that the Laplace-Laplace model was the most usable. However, we also observed that not only was the error distribution for the control-control data and HIV data heavier-tailed than normal, it appeared to display varying degrees of asymmetry. Thus we proposed the Laplace-asymmetric Laplace model.

We showed evidence that the asymmetric Laplace distribution provides a good fit to the control-control and HIV data. For the Laplace-asymmetric Laplace model, we derived expressions for the positive and negative threshold values, in terms of the signal scaling
parameter, $a$, the noise scaling parameter, $b$, the asymmetry parameter, $\kappa$, and the mixing weight, $\omega$. We gave constraints on certain parameters that must be satisfied for the thresholds to exist. Further to this, we derived expressions for the posterior median function for positive and negative observations. We gave a condition on $\kappa$ for the posterior median function to be a bounded shrinkage rule. Given all these expressions, we discussed practical aspects of implementing such a model, including parameter estimation and issues with optimisation routines. We discussed how an estimate of the implied FDR may be incorporated into the method. Finally, we explored the usefulness and success of such a model through various simulation studies and by applying the new method to the three datasets discussed in Chapter 3. The Laplace-Gaussian and Laplace-asymmetric Laplace models both have their place for analysing microarray data, yet their relative successes are strongly dependent on selecting the “correct” distribution to model the noise. When no prior information regarding this selection is available and the problem is more of an arbitrary choice, we recommend adopting the Laplace-Gaussian model; in addition to the assured bounded shrinkage property possessed by such a model, simulations suggested that under misspecified assumptions this model offers better performance than its various alternative counterparts.

Table 5.1 compares the four different models investigated in this thesis and gives the numbers of differentially expressed genes identified by each method, along with the estimated implied FDR. However, one should recall that the accuracy of the estimated FDR is dependent on how well the chosen model fits the data. Table 5.2 gives the AIC values for the four models for each of the different datasets. For the HIV dataset, the AIC criterion suggests that the Laplace-asymmetric Laplace model provides the best fit to the data, as expected. In the cases where we have no genuine prior knowledge regarding the distributional form of the noise, the AIC criterion would acknowledge the unadjusted Laplace-Gaussian model as the most appropriate. This strengthens our decision to adopt this model as the default choice of model.
Table 5.1: A table of results to compare the numbers of differentially expressed genes identified by the four models discussed in this thesis and their corresponding FDR.

<table>
<thead>
<tr>
<th>Model</th>
<th>HIV data</th>
<th>E.Coli data</th>
<th>Lymphoma data</th>
</tr>
</thead>
<tbody>
<tr>
<td>L-G (no adj.)</td>
<td>-</td>
<td>19</td>
<td>3182</td>
</tr>
<tr>
<td></td>
<td>(-% FDR)</td>
<td>(6.0% FDR)</td>
<td>(19.3% FDR)</td>
</tr>
<tr>
<td>L-G (with adj.)</td>
<td>20</td>
<td>13</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>(1.1% FDR)</td>
<td>(8.8% FDR)</td>
<td>(24.7% FDR)</td>
</tr>
<tr>
<td>L-L</td>
<td>26</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(7.2% FDR)</td>
<td>(12.2% FDR)</td>
<td>(-% FDR)</td>
</tr>
<tr>
<td>L-ASL</td>
<td>24</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(6.2% FDR)</td>
<td>(12.2% FDR)</td>
<td>(-% FDR)</td>
</tr>
</tbody>
</table>

Table 5.2: A table of results to compare AIC values for the different models discussed in this thesis.

<table>
<thead>
<tr>
<th>Model</th>
<th>HIV data</th>
<th>E.Coli data</th>
<th>Lymphoma data</th>
</tr>
</thead>
<tbody>
<tr>
<td>L-G (no adj.)</td>
<td>-</td>
<td>-810.1</td>
<td>-12106.6</td>
</tr>
<tr>
<td>L-G (with adj.)</td>
<td>-3838.7</td>
<td>-232.7</td>
<td>-77.2</td>
</tr>
<tr>
<td>L-L</td>
<td>-4756.2</td>
<td>149.4</td>
<td>16050.6</td>
</tr>
<tr>
<td>L-ASL</td>
<td>-4810.7</td>
<td>151.4</td>
<td>16052.6</td>
</tr>
</tbody>
</table>
Chapter 5. Conclusions and directions for future work

5.2 Directions for future work

In Chapter 3, we assumed a constant variance $\sigma^2$ term. Whilst a suitably chosen normalisation scheme should make this assumption reasonable, there is the potential to improve the thresholding procedure by incorporating a more robust variance estimation scheme. We suggest therefore combining the variance estimation approach we described in Manda et al. (2007) with the thresholding procedure. Presently, the thresholding procedure classifies genes as interesting based on their absolute log-ratios. Incorporating gene-specific variances into the process would highlight genes with smaller absolute log-ratios but small variances.

In Chapters 3 and 4, we explored potentially useful thresholding models for two distinct cases: firstly, when the noise is normally distributed and secondly, when the noise is heavier-tailed. In practice, choosing which model is the most appropriate to use is a simple task for spike-in experiments as the noise distribution can be inspected and an informed choice made. However, for experiments foraging into the unknown with no prior knowledge available, the choice becomes rather more difficult. Thus, a natural extension to this work is to consider the use of a generalised family of distributions, which has the Gaussian and Laplace distributions as special cases, for the purpose of modelling the noise. This is an attractive development as, given that the generalised distribution is controlled by free parameters to be estimated from the data, the data will essentially drive the choice of model.

The generalised Gaussian distribution $f(x; \alpha, \beta)$ is a family of symmetric distributions of zero mean. The distribution is defined by

$$f(x; \alpha, \beta) = \frac{\alpha}{2\beta \Gamma(1/\alpha)} \exp\left\{ -\left(\frac{|x|}{\beta}\right)^\alpha \right\}, \quad x \in \mathbb{R} \quad (5.1)$$

for $\alpha > 0$, $\beta > 0$ and where $\Gamma(\cdot)$ is the Gamma function. The parameter $\beta$ is a scaling
factor and \( \alpha \) defines the tail lengths. For \( \alpha = 1 \) and \( \alpha = 2 \), one obtains the Laplace and Gaussian distributions respectively.

Nandi and Mampel (1995) extended the distribution (5.1) and developed a distribution, controlled by three parameters, which includes the generalised Gaussian as well as allowing for various degrees of asymmetry and tail lengths. The proposed distribution comprised the negative half of one generalised Gaussian distribution \( f(x; \alpha_n, \beta_n) \) and the positive half of another generalised Gaussian distribution \( f(x; \alpha_p, \beta_p) \). To ensure continuity at \( x = 0 \), the authors constrained \( f(0; \alpha_n, \beta_n) = f(0; \alpha_p, \beta_p) \). Thus, the proposed distribution was defined as

\[
f(x; \alpha_n, \beta_n, \alpha_p) = \begin{cases} \frac{\alpha_n}{2\beta_n \Gamma(1/\alpha_n)} \exp \left\{ -\left( \frac{x}{\beta_n} \right)^{\alpha_p} \right\} & \text{for } x < 0 \\ \frac{\alpha_p}{2\beta_p \Gamma(1/\alpha_p)} \exp \left\{ -\left( \frac{\alpha_n \Gamma(1/\alpha_p) x}{\beta_n \alpha_n \Gamma(1/\alpha_n)} \right)^{\alpha_p} \right\} & \text{for } x > 0 \end{cases}
\]

for \( \alpha_n > 0, \beta_n > 0, \) and \( \alpha_p > 0 \). However, incorporating the above distribution into model (4.1) led to intractable calculations.

An alternative potential noise distribution was described by Kanefsky and Thomas (1965), who achieved asymmetry by having different areas under the curve on the left and right sides of the origin. However, the distribution described was made up of two Gaussian distributions and thus did not possess the necessary flexibility for varying tail lengths.

A further suggestion would be to model the noise as a mixture of two normal distributions controlled by a mixing weight \( \pi \) as below:

\[
\epsilon \sim \pi N(0, \sigma_A^2) + (1 - \pi) N(0, \sigma_B^2) \quad \text{with } \sigma_A^2 << \sigma_B^2.
\]

When \( \pi \) is close to zero or one, the noise distribution will resemble that of a normal. However, estimating \( \pi = 0.5 \) will give rise to a noise distribution with heavier tails. Incorporating this additional aspect would relieve the user of having to specify a noise
distribution explicitly at the outset. In turn, the identifiability problem associated with the choice of noise distribution is alleviated.

5.3 Contributions

The original suggestion of applying the empirical Bayesian thresholding procedure within a microarray context was made by Dr Stuart Barber. However, all subsequent proposals and developments of methodology were initiated and implemented by myself. The majority of work in the first part of this thesis is centred around the theme of empirical Bayesian models for the purposes of thresholding microarray data; the most important of my principal contributions are summarised below:

- The extension of the Laplace-Gaussian model, originally proposed by Johnstone and Silverman (2004), to include a variance parameter \( \sigma^2 \).
- The proposal and implementation of a hierarchical linear additive model for the estimation of \( \sigma^2 \).
- The observation that the error distribution was frequently more heavy-tailed than Gaussian. These findings motivated the investigation of alternative extensions to the empirical Bayesian model.
- In Chapter 4, a review of some of the various empirical Bayesian models described in Bochkina and Sapatinas (2005) was carried out; the implementation of these models and application of them to microarray datasets is my own.
- The proposal and formulation of the Laplace-asymmetric Laplace model. In particular, equations for the threshold value and constraints that should be satisfied are provided for both the positive and negative cases, in addition to expressions for
the posterior median function. These algebraic derivations are my own, as is the realisation of the model in practice.

- The evaluation of the proposed models through a series of simulation studies and the recommendation of a “default” model.

- The inclusion of the existing false discovery rate work of Newton et al. (2004) within this particular empirical Bayesian framework.
Part II

HPLC data
Chapter 6

Proteomics

6.1 Introduction

The analysis of unknown chemical components in complex mixtures arises frequently in many fields including agriculture, medicine, industry and food science. The study of these mixtures is dependent on analytical techniques such as mass spectrometry (MS), high performance liquid chromatography (HPLC), infra red (IR) and nuclear magnetic resonance spectroscopy (NMR), which all give rise to very similar forms of spectral data. Unfortunately, the technologies are highly sensitive to external irregularities and rarely yield robust and consistent data, thus sophisticated statistical algorithms are required for the purpose of analysis, the development of which form the essence of this part of the thesis. Within a medical field, mass spectrometry and high performance liquid chromatography are the predominant types of technology employed. Whilst we suspect that the methods derived here could be applicable to other types of technology, we focus on MS and HPLC from this point forwards. Similarly to Part I, we initially provide an overview of the technology before turning our attention to more statistical issues. The choice to include a discussion of MS technology here is motivated further by the knowledge that the majority of existing literature on spectral analysis focuses on this type
6.2 Mass spectrometry

Whilst analysing changes in gene expression gives us some insight into a cell’s dynamic behaviour, much of the activity occurs at the protein level. Thus it would be highly useful if there existed the protein equivalent of a microarray, allowing one to identify and quantify protein levels in various samples. One popular and promising technology is that of mass spectrometry which can measure the amount of hundreds of proteins present in a sample simultaneously. Mass spectrometry was first evident in the early 1900’s, yet its potential to identify proteomic patterns has only recently been recognised. Since then however, there has been extensive research interest, particularly regarding experiments which aim to discover biomarker patterns that have the potential to discriminate between different pathological conditions from a biological specimen. Comprehensive texts on all aspects of mass spectrometry are provided by Dass (2001), Dass (2007), Boyd (2007), Hoffmann and Stroobant (2007) and Burrell (2008).

Mass spectrometry is designed to separate gas phase ions and works by exploiting the fact that different compounds have different atomic masses. Firstly, the sample of interest must be ionised, or in other words, a charge is added to the substances to be measured. Next, the ions are manipulated with magnetic or electronic fields, which separates them according to their mass to charge \((m/z)\) ratio. Finally, a detector is used to record the relative abundance of ions with a particular \((m/z)\) ratio and a mass spectrum is generated representing the masses of the sample components. From the pattern of abundance values, one can hope to infer which proteins are present and in what quantity. There currently exist various types of mass spectrometers which differ in the types of ions that are produced and the way that the ions are analysed. The choice of ionisation technique
Proteomics is often dependent on the sample state: for gases and vapours, electron and chemical ionisation processes are used; for liquid and solid biological samples, techniques include electrospray ionisation (ESI) and matrix-assisted laser desorption/ionization (MALDI). Within the medical domain, these particular techniques are prevalent as samples of interest are often in a liquid state, for example, human blood serum specimens, urine or saliva. The most widely used mass analysers are the quadrupole and the time-of-flight (TOF), which is often well known through its common coupling with MALDI.

Mass spectrometry experimentation gives rise to complex functional data: on the horizontal axis are mass/charge (m/z) values and on the vertical axis an intensity measurement that indicates the sequentially recorded numbers of ions arriving at the detector. The analysis of these data typically involves inferring from the existence and height of a spike or peak in the spectrum, the existence and (relative) abundance of a peptide of a particular m/z. Unfortunately, and typical of this type of high-throughput technology, many inherent problems have been identified with the data and these need to be confronted before any robust discriminatory analysis can be performed. Meticulous pre-processing is an essential stage and tackles issues including denoising, baseline estimation and correction. Furthermore, the m/z locations and the shape/size of peaks across samples can be highly variable which, given the complexity of the signal, leads to a high risk of falsely finding peaks in the noise. Thus, some normalisation and alignment of multiple spectra is often necessary before one can consider feature extraction, location identification and peak quantification. These ideas are covered in more detail in Chapter 8.

Given successful pre-processing, one can assume that p peaks are identified from n spectra, which yields an \( n \times p \) matrix of “protein peak expression levels” (Morris et al., 2005). From here, many researchers search for proteins that may be differentially expressed between experimental conditions and perform unsupervised clustering or apply unsupervised learning methods to perform discrimination and classification. Essentially the “protein peak expression levels” matrix parallels a matrix of gene expression levels.
obtained from a microarray experiment. Indeed, many of the statistical tools developed for microarray analysis have been found applicable and useful for mass spectrometry analysis.

6.3 High performance liquid chromatography (HPLC)

HPLC is an extremely versatile instrumental technique that can be used to separate, identify and quantify compounds in any sample that can be dissolved in a liquid (Hamilton and Sewell, 1982). Solutions derived from the samples of interest are introduced in small volumes onto an HPLC column that comprises a narrow stainless steel tube, densely packed with fine, chemically modified silica particles, which possess a certain surface chemistry. The sample is forced to flow through the chromatographic column by pumping through a liquid (known as the mobile phase) under high pressure. The unknown compounds in the sample of interest are retarded by specific chemical interactions with the coating of the silica particles as the sample traverses the length of the column. The amount of retardation depends on the nature of the specific compound or analyte, the choice of particles in the column and the composition of the mobile phase. Analytes that “like” the particles in the column will interact and take longer to elute (come out of the end of the column) than analytes where little or no interaction occurs. The time at which a specific analyte elutes is called the retention time, \( t_R \), and is considered a reasonably unique identifying characteristic of a given analyte. A detector is used to quantitatively “see” each separated analyte as it elutes. Various different types of detectors can be used: UV/visible light absorbance, differential refraction index, electrochemical, conductivity and fluorescence are some of the properties that detectors exploit. For example, a UV detector measures the ability of a analyte to absorb UV light and emits a response as a function of the components passing through it. The resulting chromatogram is a representation of the separation that has chemically occurred. Once again, this type of
experimentation yields functional data; an example of such a trace is given in Figure 6.1. Retention times, $t_R$, are on the horizontal axis, where $t_0$ is the time at when the sample was first injected into the column. The values on the vertical axis are essentially “rates” - the count of molecules passing through the detector at a particular time point. Visually, these data yield a tracelike spectrum consisting of a horizontal baseline, interrupted by a series of “burstlike” peaks along the time axis. The baseline represents pure mobile phase passing through the detector. Each peak represents a different compound and can be identified from its location based on the retention time (defined as the apex of the peak). Having identified the different compounds, the concentration of a compound can be quantified by the area under that particular peak. For further reading on the processes involved in HPLC, Hamilton and Sewell (1982), Gilbert (1987) and Meyer (2004) are useful examples.

![Figure 6.1: An example of a raw HPLC trace.](image)

HPLC is sufficiently robust to warrant data analysis on its own however many experimenters combine the separation techniques of HPLC with the analysing power of
mass spectrometry. Incorporating HPLC into the process allows any impurities in the sample that may negatively impact on the ionisation procedure to be eliminated. It is necessary to introduce components eluting from the chromatographic column to the mass spectrometer via a specialised interface, which converts the liquid analytes into a gaseous form. The two most common interfaces used for HPLC/MS are the electrospray ionisation and the atmospheric pressure chemical ionisation interfaces. For more information on these ionisation processes, one should see Cappiello (2006).

### 6.4 Description of data

The data examined in this thesis consist of spectra derived from High Performance Liquid Chromatography (HPLC) experimentation, analysing saliva samples to investigate the effect of caffeine in pregnant women. In this case, a UV light detector was used at a wavelength of 280nm. Whilst the scale of the true measurements runs into the thousands, for presentation purposes, the observations have been linearly transformed internally by the accompanying commercial software to some arbitrarily selected scale, seen here. The dataset consists of spectra from 25 runs. Five of the 25 runs are from standard samples; the remainder are from experimental samples. These data are presently unpublished. Currently, analysis of such data is performed through a combination of manual input (to select peak start- and endpoints) and tools from commercial software (for smoothing and estimating the area under each peak). The aim of this work is to develop a more automatic approach which should reduce the amount of time a user has to spend analysing this data, as well as achieving objective and accurate results.
6.5 Common sources of complications

Here we describe loosely some of the additional technological complications observed by researchers in the area. Some of these problems are not necessarily pertinent to HPLC. However we hope that the methods derived later in this thesis will be applicable to a broad range of data, despite the differences in technology, thus some awareness of these issues is important.

1. Sinusoidal noise: Baggerly et al. (2003) reported visual evidence of systematic distortions, particularly at the high \( m/z \) values: regular sinusoidal noise affected most of the spectra. This noise was periodic on the time scale, rather than the \( m/z \) scale. Due to its regularity, it was hypothesised that this phenomenon was linked to the frequency of the alternating current in the power source, rather than any biological feature. Furthermore, they discovered a periodic dip in the spectrum, occurring at every \( 2^{12} \) clock ticks. Due to its occurrence at powers of 2, it was suggested that this artifact was related to the computer chip inside the instrument recording the data.

2. Chemical noise: In MALDI the sample mixture is combined with a “matrix” compound in the first step of the ionisation procedure. Very simply, the matrix crystallises, holding the proteins in place as the mixture dries. The crystal structure is blasted with a laser which breaks the matrix apart, freeing the protein peptides in a gaseous phase. The matrix compounds readily donate spare protons to molecules in close proximity during this fragmentation process, thus ionising them. However in practice, the laser fire frees not only the peptides of interest, but other unwanted clusters of matrix molecules, which are detected by the mass spectrometer. This surplus information is referred to as “chemical noise”. The addition of chemical noise in the spectra leads to many weak protein/peptide signals being falsely assigned to the chemical noise, as it is impossible to differentiate the
two. Similar forms of chemical noise also hinder electrospray ionisation (ESI) mass spectrometry, due to the buffers and solvents used in preparation.

Many mass spectra are found to contain impurities that have not originated from the biological sample of interest. These pollutants usually materialise from sample preparation or contamination and include keratin (or other proteins from human skin, hair or clothing), and trypsin (used to cleave proteins into peptides).

In HPLC, the noise is due to electronic pulses and the presence of air bubbles dissolved in the solution. Whilst ideally, air should have been removed from the solution by de-gassing, some will inevitably remain and thus the concentration will change with temperature. Furthermore, the system is made more variable because the mobile phase is made up of three components and although homogeneous, changes in the partial pressures of each can cause the equilibrium to vary slightly. Temperature controlled environments will minimize these changes, yet not eliminate them completely, resulting in evident fluctuations.

3. Baseline: Evidence of some sort of drift is frequently observed in these types of spectra. In MALDI spectra, a strong upward drift in the baseline of the mass spectra which falls off rapidly with increasing mass is often evident. This baseline phenomenon is thought to occur due to a cloud of matrix molecules hitting the detector in the early stages of the experiment, or to detector overload (Malyarenko et al., 2005). In electrospray ionisation spectra, chemical noise manifests as a bump in the baseline in the intermediate mass range. In HPLC experimentation, the baseline represents the elution of pure mobile phase and may fluctuate through non-constant pump pressure and the specific composition of the mobile phase (as mentioned previously). With regard to the HPLC data that is analysed in this thesis, the drift appears to be non-decreasing as the retention time increases.

4. Calibration: Technically, a TOF mass analyser will attempt to measure the number of ions hitting the detector as a function of the time of flight. These times are
converted into \( m/z \) values through the application of equations which describe the physics of the ion separation process. The conversion formulae contain various parameters, some of which, for example, the initial velocity and position of the ions, have to be approximated, leading to errors in the \( m/z \) values. Vestal and Juhash (1998) detailed some discussion of these formulae for various TOF configurations. Providing that the flight times of some reference ions of known mass are available, unknown parameters in the conversion equations can be defined by a fitting procedure (Christian et al., 2000). Otherwise some calibration of the spectra will be necessary. Shifts in liquid chromatography retention times are caused by different injection timing (a constant shift), slow and fast temperature fluctuations, and flow rate changes.

5. Peak spread: Electronspray ionisation will often produce peptide ions that have a broad distribution of charge states, especially in the case of large proteins. Consequently, as mass spectrometry “measures” \( m/z \) ratios, that particular protein will be observed at a variety of positions in the spectrum, possibly confounded with signals from other proteins in the sample. Some elements exist naturally as isotopes which have different atomic masses, due to the addition of non-charged neutrons in the nucleus. This can lead to a broadening of the peak for peptides incorporating these isotopes and is most evident at higher \( m/z \) values. Large proteins may fragment during ionisation, which again gives rise to a broader peak than expected. In MALDI experiments, peptides of the same type will break free of the matrix with a distribution of initial velocities, rather than some constant, causing the observed peak to be spread out.

In HPLC experiments, the broadness of the peak will be due to the concentration of the compound as expected, however can also be exaggerated by various technical flaws. These may include how evenly the samples are loaded onto the column, the purity of the column, the temperature of the mobile phase and column, the presence
of leaks in the system and the age of the UV lamp.

6. Ion suppression effects: During a MS ionisation process, if a sample contains multiple analytes that can ionise in the operating conditions, a competition effect may occur. Those analytes producing strong signals, especially those which exist in high concentration, can “drown” the signals of other analytes, which are less amenable to ionisation. This feature has been noted by King et al. (2000) and Tang et al. (2004) for electrospray ionisation experiments and by Kratzer et al. (1998) for MALDI. Thus the strength of the signal for certain analytes is not linearly dependent on its initial concentration, rather it is influenced by the concentration of other analytes in the sample in a complex manner.
Chapter 7

Wavelet methodology

7.1 Wavelets and the discrete wavelet transform

7.1.1 Overview of wavelet analysis

To begin our exploration of wavelets, we first recall Fourier series where a function $f$, defined on $[-\pi, \pi]$, can be represented exactly by a linear combination of elements from the Fourier basis $\{\exp(inx)\}_{n=-\infty}^{\infty}$:

$$f(x) = \sum_{n=-\infty}^{\infty} c_n \exp(inx), \quad (7.1)$$

where the Fourier coefficients are computed by

$$c_n = (2\pi)^{-1} \int_{-\pi}^{\pi} f(x) \exp(-inx) dx. \quad (7.2)$$

The term $\exp(inx)$ can be expanded into $\cos(nx) + i \sin(nx)$; thus the representation of $f$ in (7.1) is essentially a series expansion of the function into sine and cosine functions of differing frequencies, transforming $f$ into a set of coefficients. Such a transformation is simply another way of representing the function and does not change the information.
present in the function. The wavelet transformation, which we are going to briefly outline in this chapter, can be considered similar to the Fourier transform in that both transformations expand a function $f$ in terms of a set of orthogonal basis functions: a Fourier expansion utilises the orthonormal system $(2\pi)^{1/2} \exp(inx)$ on $[-\pi, \pi]$, whereas in a wavelet expansion, the analysing functions are constructed from dilations and translations of a single, well-localised, oscillating function, $\psi$, called the mother wavelet.

Most signals obtained from real-life situations encompass a wide range of frequencies. In addition to this, many tend to be non-stationary in their nature. It is often observed that there is a direct relationship between the characteristic frequency of a given segment of the signal and the time duration of that segment. Low frequency pieces tend to span a long time period, whereas high frequency activity generally occurs in short bursts only. Hence, given that the sines and cosines of the standard Fourier basis are specific only in terms of frequency, the usefulness of the Fourier transform is somewhat limited in such practical applications. Extremal Phase wavelets (Daubechies, 1988), which we will use later in this thesis, differ from the standard Fourier approach in that the wavelet basis elements have compact support, allowing the localisation of a signal in both time and frequency via translation and dilation respectively. Further to this, wavelets have proven superior in many applications to even the windowed Fourier transform, since they adapt to the local frequency in the signal and selects widths of time accordingly.

Briefly, a discrete wavelet decomposition breaks a signal down into a coarse approximation at a given scale and successive residual detail at finer and finer scales. Each of the various contributions represents information of differing frequency contained in the original signal. The decomposition is expressed in terms of a scaling function $\phi$, which resembles a kernel function, and a wavelet function $\psi$, a “localised wiggle”. A linear combination of dyadic shifts of the scaling function provides the coarse scale, whilst linear combinations of dyadic shifts of the wavelet function supply the residual detail in the signal, absent in the coarse approximation. By considering versions of the wavelet at a
number of consecutive dyadic dilations, the detail at each of the corresponding successive scales is recovered. Mathematically, the weights of all these linear combinations are simply given by the inner product of the signal with the appropriate scaling or wavelet function, dilated and translated as necessary.

In this chapter, we only describe those wavelet methods that we later employ for the purposes of analysing the HPLC data; for general reading on wavelets, one is directed to Vidakovic (1999) and Percival and Walden (2000). In the next section, we expand on this brief overview of the wavelet methodology, following the description and notation as detailed in Vidakovic (1999).

### 7.1.2 The discrete wavelet representation

Daubechies (1988), Daubechies (1992) and Meyer (1992) all described how, given a mother wavelet \( \psi \in L_2(\mathbb{R}) \) (the space of square integrable real functions), a family of “self-similar” wavelets can be constructed by translations and rescaling of the function \( \psi \),

\[
\psi_{a,b} = \frac{1}{\sqrt{|a|}} \psi \left( \frac{x - b}{a} \right), \quad a \in \mathbb{R} \setminus \{0\}, b \in \mathbb{R}. \tag{7.3}
\]

For certain choices of \( \psi \), the resulting family of translated and dilated wavelets \( \{ \psi_{a,b} \} \) constitutes an orthonormal basis of \( L_2(\mathbb{R}) \) (Meyer, 1985), which can be used as “building blocks” to describe functions.

The continuous wavelet series representation of a function \( f \) is

\[
f(x) = \int_{\mathbb{R}} f_{a,b} \cdot \psi_{a,b}(x), \tag{7.4}
\]

where the wavelet coefficients are calculated via the continuous wavelet transformation

\[
f_{a,b} = \langle f, \psi_{a,b} \rangle,
= \int_{-\infty}^{\infty} f(x) \psi_{a,b}(x) \, dx, \tag{7.5}
\]
and \(< \cdot , \cdot >\) is the inner product (Grossmann and Morlet, 1984).

The continuous wavelet transform operates over every possible dilation and translation value, which has proven to be redundant. Alternatively, it is possible to “minimise” the efforts of the transformation by using a specifically selected subset of discrete dilation and translation values (Mallat, 1989). The most common discrete wavelet transformation (which forms the core of our immediate discussion) incorporates sampling defined by \(a = 2^{-j}\) and \(b = 2^{-j}k\), \(j, k \in \mathbb{Z}\), which yields the general wavelet basis element

\[
\psi_{j,k}(x) = 2^{j/2} \psi(2^j x - k) \quad j, k \in \mathbb{Z}. \tag{7.6}
\]

Dilation, indexed by the parameter \(j\), describes the degree of compression or scale of the wavelet. Translation, indexed by \(k\), determines the location of the wavelet along the horizontal axis. A unit increase in \(j\) packs the oscillations of \(\psi_{j,k}\) into half the width; a unit increase in \(k\) will shift the location of \(\psi_{j,k}\) along the horizontal axis by an amount proportional to its width, \(2^{-j}\). The purpose of the scale factor \(2^{j/2}\) is to normalise \(\psi_{j,k}\), ensuring that \(\|\psi_{j,k}\| = \|\psi\|\). The resulting set of translated and dilated wavelets can be used as “building blocks” to describe a function \(f\) in a corresponding discrete wavelet series as

\[
f(x) = \sum_{j \in \mathbb{Z}} \sum_{k \in \mathbb{Z}} d_{j,k} \cdot \psi_{j,k}(x), \tag{7.7}
\]

where \(d_{j,k} = < f, \psi_{j,k} >\) are the coefficients obtained by the discrete wavelet transform.

In order to cover the entire domain of a function \(f\), infinite \(j\) scales will be required in the representation (7.7). Thus in practice, a scaling function, \(\phi\), is used to provide a coarse approximation at the lowest level of the transform, ensuring that the entire range of \(f\) will be covered. Hence,
Chapter 7. Wavelet methodology

\[ f(x) = \sum_{k \in \mathbb{Z}} c_{j_0,k} \cdot \phi_{j_0,k} + \sum_{j \geq j_0} \sum_{k \in \mathbb{Z}} d_{j,k} \cdot \psi_{j,k}(x), \quad (7.8) \]

where \( c_{j_0,k} = \langle f, \phi_{j_0,k} \rangle \) and \( d_{j,k} = \langle f, \psi_{j,k} \rangle \).

Intuitively, the coefficient \( d_{j,k} \) gives information on the local oscillatory behaviour of \( f(x) \) near time point \( 2^{-j} k \) and near frequency proportional to \( 2^j \). The scaling coefficient \( c_{j_0,0} \) contains information on the global “mean level” of \( f \) at resolution level \( j_0 \).

### 7.1.3 Scaling functions and constructing the mother wavelet

A multiresolution analysis (MRA) is the underlying design of a discrete wavelet transformation (see Hardle et al. (1998), Jawerth and Sweldens (1994) and Mallat (1989), for example) and is defined as a sequence of closed subspaces \( V_m, m \in \mathbb{Z} \) such that

\[ \ldots \subset V_{-2} \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \subset \ldots \quad (7.9) \]

The intersection of these nested spaces is in fact trivial, i.e. \( \bigcap_j V_j = \{0\} \); their union is dense in \( L^2(\mathbb{R}) \), i.e. \( \bigcup_j V_j = L^2(\mathbb{R}) \).

This hierarchy of subspaces is built in such a way as to satisfy some self-similarity criteria: firstly, the \( V \)-spaces should all be scaled versions of the central space \( V_0 \), that is

\[ f(\cdot) \in V_0 \iff f(2^j \cdot) \in V_j. \quad (7.10) \]

Furthermore, the space \( V_0 \) should be invariant under integer translations, that is

\[ f(\cdot) \in V_0 \Rightarrow f(\cdot - k) \in V_0. \quad (7.11) \]
We further define $W_j$ to be the “difference space”, $V_{j+1} \ominus V_j$, and this is orthogonal to $V_j$. Further, for a fixed scale $j$, the set of functions $\{\psi_{j,k}(x), k \in \mathbb{Z}\}$, with $\psi_{j,k}$ as defined in equation (7.6), forms an orthonormal basis of $W_j$.

As the space $L_2(\mathbb{R})$ is represented by the direct sum

$$L_2(\mathbb{R}) = \bigoplus_{j \in \mathbb{Z}} W_j,$$

it is clear to see that

$$\{\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k), j, k \in \mathbb{Z}\},$$

forms an orthonormal basis of $L_2(\mathbb{R})$.

The construction of such a wavelet family exploits the properties of the multiresolution analysis and generates the wavelet function through a companion scaling function, or informally the father wavelet, $\phi \in V_0$. The integer-translates of this scaling function spans the space $V_0$

$$V_0 = \left\{ f \in L_2(\mathbb{R}) | f(x) = \sum_k c_{0,k}\phi(x - k) \right\},$$

and the set $\{\phi(\cdot - k), k \in \mathbb{Z}\}$ forms an orthonormal basis for $V_0$. Furthermore, it is assumed that $\int \phi(x)dx \neq 0$.

Since $V_0 \subset V_1$ by the definition of a multiresolution analysis, the function $\phi(x) \in V_0$ can be written as a linear combination of functions from $V_1$, i.e.,

$$\phi(x) = \sum_{k \in \mathbb{Z}} h_k \sqrt{2}\phi(2x - k),$$
Chapter 7. Wavelet methodology

for some coefficients $h_k, k \in \mathbb{Z}$. This is the key idea underlying wavelet expansions: essentially a function can be represented as a linear combination of functions from the next resolution via a filter, $\{h_l, l = 0 \ldots L - 1\}$, where $L$ is an even integer and describes the length of the filter.

The desired function, $\psi(x) = \psi_{0,0}(x)$, is the aforementioned mother wavelet and this forms an orthonormal basis of $W_0$. Because of the inclusion $W_0 \subset V_1$, it becomes clear that $\psi(x) \in V_1$. Thus, via a suitable chosen wavelet filter, we can also express $\psi(x)$ as a linear combination of functions from $V_1$, or in wavelet terminology, from the next resolution:

$$\psi(x) = \sum_{k \in \mathbb{Z}} g_k \sqrt{2} \phi(2x - k), \quad (7.16)$$

for some coefficients $g_k, k \in \mathbb{Z}$.

The smoothing or low-pass scaling filter $\{h_l\}$ and the high-pass wavelet filter $\{g_l\}$ are known as the quadrature mirror filters (Cohen, 1990). They can be shown to be related through the equation

$$g_l = (-1)^l h_{L-1-l}, \quad l = 0, \ldots, L - 1, \quad (7.17)$$

which is known as the quadrature mirror relation.

In order to form a valid wavelet filter the coefficients $\{g_l\}$ must satisfy the following set
of conditions

\[
\sum_{l=0}^{L-1} g_l = 0, \\
\sum_{l=0}^{L-1} g_l^2 = 1, \\
\sum_{l=0}^{L-1} g_l g_{l+2m} = \sum_{-\infty}^{\infty} g_l g_{l+2m} = 0,
\]

(7.18)

for all non-zero integers \(m\). Numerous filter families have been developed that satisfy these necessary conditions on the filters and the choice of filter family often depends on the application in question. In later chapters, we use the family of Extremal Phase filters proposed by Daubechies (1988). Further to the choice of filter family, there is also a selection to be made regarding the length of the filter. Longer filters are more frequency localised and less likely to introduce spurious features. However, shorter filters are more localised in time and are faster to compute. By definition, shorter filters are also less influenced by boundary effects. In many cases, the choice is motivated by prior knowledge regarding the particular dataset in question or visual exploration. For further discussion of these issues, see Percival and Walden (2000).

Further developments in wavelet methodology include the introduction of multiwavelets (Geronimo et al., 1994; Downie and Silverman, 1998) and complex-valued wavelets (Lina and Mayrand, 1995; Lawton, 1993; Barber and Nason, 2004). These ideas are not covered within this thesis, however one is directed to the given references for further reading.

### 7.1.4 Haar’s Wavelets

The purpose of this section is to illustrate some of the statements given in the previous section through the simplest wavelet family and provide a working example.
Chapter 7. Wavelet methodology

The simplest wavelet basis for $L^2(\mathbb{R})$ is the Haar basis (Haar, 1910), which uses the scaling function given by

$$
\phi(x) = \begin{cases} 
1 & \text{if } 0 \leq x < 1 \\
0 & \text{otherwise.}
\end{cases}
$$

In practice, the general approach to constructing a wavelet family is to define a scaling function $\phi(x)$, from which the scaling function filter coefficients can be deduced. By equation (7.15), the scaling equation $\phi(x)$ can be written as:

$$
\phi(x) = \phi(2x) + \phi(2x - 1) = \frac{1}{\sqrt{2}} \phi(2x) + \frac{1}{\sqrt{2}} \phi(2x - 1)
$$

(7.19)

and so the corresponding filter coefficients are:

$$
h_0 = h_1 = \frac{1}{\sqrt{2}}
$$

(7.20)

The filter for the scaling function is transformed into the Fourier domain; an expression describing the filter for the wavelet function in the Fourier domain can be derived using relationships detailed in Vidakovic (1999). Performing the inverse Fourier transformation will yield the corresponding wavelet function $\psi(x)$. Using this approach, it can be shown that the Haar wavelet function $\psi$ is defined to be

$$
\psi(x) = \phi(2x) - \phi(2x - 1),
$$

and that the high-pass wavelet filter takes the form $g_0 = -g_1 = 1/\sqrt{2}$.

Thus
ψ(x) = \begin{cases} 
1 & \text{if } 0 \leq x < \frac{1}{2} \\
-1 & \text{if } \frac{1}{2} \leq x < 1 \\
0 & \text{otherwise.} 
\end{cases}

Figure 7.1: The Haar wavelet function \( \psi_{j,k} \) for different values of \( j \) and \( k \).

Figure 7.1 illustrates the wavelet functions \( \psi_{0,0} \), \( \psi_{1,1} \) and \( \psi_{-1,0} \).

In order to demonstrate that orthonormality is satisfied, it is necessary to show that

\[
\int_{\mathbb{R}} \psi_{j,k}(x)\psi_{j',k'}(x)dx = \begin{cases} 
0 & \text{if } (j, k) \neq (j', k') \\
1 & \text{if } (j, k) = (j', k'). 
\end{cases}
\]

Clearly in this case, \( \int \psi_{j,k}\psi_{j,k'}dx = 0 \) when \( k \neq k' \) as the supports of wavelets on the same scale but of differing translation will never intersect. Visual inspection of Figure 7.1 suggests that the condition \( \int \psi_{j',k}\psi_{j,k}dx = 0 \) is fully satisfied also when \( j \neq j' \). Wavelets on different scales either have supports which do not intersect, or, if this is not the case,
then one wavelet takes the value \(-k\) and then \(k\) over a set where the other wavelet remains constant (for some \(k\)).

Alternatively, a more rigorous proof can be provided algebraically. Suppose that \(j = j'\) and \(k \neq k'\). Then

\[
\int_{\mathbb{R}} \psi_{j,k}(x) \psi_{j',k'}(x) dx = \int_{\mathbb{R}} 2^j \psi(2^j x - k) 2^j \psi(2^j x - k') dx
= \int_{\mathbb{R}} \psi(y) \psi(y + k - k') dy
= 0
\]

Since \(k\) and \(k'\) are simply shifts of the function up or down the \(x\)-axis, the integrand is identically zero.

Alternatively, assume that \(j < j'\). Then

\[
\int_{\mathbb{R}} \psi(2^j x - k) \psi(2^{j'} x - k') dx = 2^{-j} \int_{\mathbb{R}} \psi(y) \psi(2^{j'-j} y + 2^{j'-j} k - k') dy
= \int_{0}^{\frac{1}{2}} \psi(2^{j'-j} y + k'' dy - \int_{\frac{1}{2}}^{1} \psi(2^{j'-j} y + k'' dy
\]

Both of these integrands are zero since it is clear that \(\int_{0}^{1} \psi(x) = 0\).

Finally, normalisation, in case \((j, k) = (j', k')\), can established by computing

\[
\int_{\mathbb{R}} \psi_{j,k}(x)^2 dx = 2^j \int_{\mathbb{R}} \psi(2^j x - k)^2 dx = \int_{\mathbb{R}} \psi(y)^2 dy = 1.
\]

The Haar wavelets have limited use in practice as clearly the \(\psi_{j,k}\) are discontinuous functions. These are obviously inappropriate for the majority of applications, as most of the time the real-life signals that we wish to approximate are smooth functions, or at least have smooth segments.
7.1.5 Mallat’s cascade algorithm

Given some particular parent wavelets, what remains is to calculate the wavelet coefficients. Each coefficient requires the evaluation of an integral, which will be costly in terms of computational time. However, Mallat (1989) derived an alternative algorithm for discrete datasets, analogous to the Fast Fourier Transform of Cooley and Tukey (1965), which implements the discrete wavelet transformation whilst neatly avoiding the many integral calculations. The algorithm is built on the exploitation of the nested hierarchical structure of the multiresolution analysis and incorporates the use of the wavelet filters described in Section 7.1.3. The discrete wavelet transform (DWT) produces a vector of the same length as the original data, containing the wavelet coefficients of the input data at dyadic scales and locations. The resulting vector constitutes a mapping from the time domain (the original data) to the wavelet domain.

The original data of length $2^J$ are considered to be linked with the space $V_J$ and denoted $c_J$. The spaces $(V_{J-1}, W_{J-1})$, $(V_{J-2}, W_{J-2})$ then denote coarser smooth and corresponding detail spaces. Decreasing the index in $V$-spaces is equivalent to coarsening the approximation to the data. At each stage of the algorithm or decomposition, “smoother signal” $c$ is extracted at a lower resolution level than the current one, in addition to the signal detail $d$ at that resolution.

Suppose there exists the multiresolution analysis $\ldots \subset V_{J-1} \subset V_J \subset V_{J+1} \subset \ldots$ and that $v_J(x) = \sum_k c_{j,k} \phi_{j,k}(x)$. Since $V_J = V_{J-1} \oplus W_{J-1}$, any function $v_j \in V_j$ can be represented uniquely as $v_J(x) = v_{J-1}(x) + w_{J-1}(x)$, where $v_{J-1} \in V_{J-1}$ and $w_{J-1} \in W_{J-1}$.

Mathematically, this is expressed as
\[ v_j(x) = \sum_k c_{j,k} \phi_{j,k}(x) \]
\[ = v_{j-1}(x) + w_{j-1}(x). \]
\[ = \sum_l c_{j-1,l} \phi_{j-1,l}(x) + \sum_l d_{j-1,l} \psi_{j-1,l}(x) \]

where \( c_{j-1,l} = \langle v_j, \phi_{j-1,l} \rangle \) and \( d_{j-1,l} = \langle v_j, \psi_{j-1,l} \rangle \). The \( c_{j,k} \) and \( d_{j,k} \) are the conventional choice to denote coefficients associated with \( \phi_{j,k}(x) \) and \( \psi_{j,k}(x) \) respectively.

Given the orthogonality of \( w_{j-1}(x) \) and \( \phi_{j-1,l}(x) \) for any \( j \) and \( l \), the additivity of inner products and the fact that

\[ \phi_{j-1,k}(x) = \sum_{l \in \mathbb{Z}} h_{l-2k} \phi_{j,k}(x) \quad \text{and} \quad \psi_{j-1,k}(x) = \sum_{l \in \mathbb{Z}} g_{l-2k} \phi_{j,k}(x) \]

by a simple index substitution of (7.15) and (7.16), a single decomposition step can be written as

\[ c_{j-1,l} = \langle v_j, \phi_{j-1,l} \rangle \]
\[ = \langle v_j, \sum_k h_{k-2l} \phi_{j,k} \rangle \]
\[ = \sum_k h_{k-2l} \langle v_j, \phi_{j,k} \rangle \]
\[ = \sum_k h_{k-2l} c_{j,k}. \quad (7.21) \]

Similarly,

\[ d_{j-1,l} = \sum_k g_{k-2l} c_{j,k}. \quad (7.22) \]
In practice, given the data \( c_J \) and relations (7.21) and (7.22), we can apply filters \( \{ h_l \} \) and \( \{ g_l \} \) to give resulting vectors, each of length \( 2^{J-1} \), which contain the scaling coefficients \( \{ c_{J-1,k} \} \) and wavelet coefficients \( \{ d_{J-1,k} \} \) at the next resolution respectively.

The averaging nature of the filter \( \{ h_l \} \) in the frequency domain has lead to the scaling coefficients being referred to as smoothing coefficients; the wavelet coefficients are most commonly known as the detail coefficients, since they fill in the “detail information” that is lost when \( \{ c_{j,k} \} \) is approximated by the “averaged” \( \{ c_{j-1,k} \} \).

Now the \( \{ c_{J-1,k} \} \) are used to yield \( \{ c_{J-2,k} \} \) and \( \{ d_{J-2,k} \} \), both vectors of length \( 2^{J-2} \); similarly, the approach can be applied repeatedly to each scale until one arrives at the coarsest resolution. At this point the vectors will contain just one coefficient, \( c_{0,0} \) and \( d_{0,0} \) respectively.

The DWT of the initial function \( f \) is then described as

\[
\text{DWT}(f) = (c_{0,0}, d_{0,0}, d_{1,0}, d_{1,1}, d_{2,0}, \ldots, d_{2,3}, \ldots, d_{J-1,0}, \ldots, d_{J-1,2^{J-1}-1}).
\]

This representation is intuitively clear if we consider the overall decomposition of the space \( V_J \):

\[
V_J = V_{J-1} \oplus W_{J-1} = V_{J-2} \oplus W_{J-2} \oplus W_{J-1} = V_{J-3} \oplus W_{J-3} \oplus W_{J-2} \oplus W_{J-1} = \ldots = V_0 \oplus W_0 \oplus \cdots \oplus W_{J-2} \oplus W_{J-1}.
\]

Thus, the function \( f \) can be represented as a corresponding wavelet series by
\[ f(x) = c_{0,0}\phi(x) + \sum_{j=0}^{J-1} \sum_{k=0}^{2^j-1} d_{j,k} \psi_{j,k}(x). \]  

(7.24)

One should note that the above algorithm describes the decimated DWT. The non-decimated version is discussed further in Section 7.3.

The cascade algorithm can be applied in the reverse direction as well, known as the inverse DWT, permitting the original signal to be recovered from the coefficients in its wavelet decomposition. Coefficients in the next finer scale corresponding to \( V_j \) can be obtained from the coefficients corresponding to \( V_{j-1} \) and \( W_{j-1} \) through the relation

\[
c_{j,k} = \langle v_j, \phi_{j,k} \rangle = \sum_l c_{j-1,l} \phi_{j-1,l} \langle \phi_{j,k} \rangle + \sum_l d_{j-1,l} \phi_{j,k} \rangle
\]

\[
= \sum_l c_{j-1,l} h_k - 2l + \sum_l d_{j-1,l} g_{k-2l}.
\]

(7.25)

In cases where the selected filter is of length greater than 2, issues occur at the boundaries of the data as the convolving filter extends beyond the range of data-indices. There exist various options to circumvent this problem: the simplest approaches are to assume the data are periodic or symmetric outside of the observed interval. More sophisticated approaches include confining the wavelet transformations to an interval (one should see Cohen et al. (1993)), however for similar reasons as to the choice of padding technique, discussed in the next section, we move to the use of symmetric boundary handling.

### 7.1.6 Padding

The discrete wavelet transform requires the sample size to be of dyadic length. This assumption is very restrictive and a number of modified discrete wavelet transforms have
been developed to overcome it (Ogden, 1997; Vidakovic, 1999; Percival and Walden, 2000). Popular modifications are the partial DWT and the maximal overlap DWT.

The alternative to a modified DWT is to “pad” the observed series at the boundaries to increase the number of observations to the nearest dyadic term. The extra observations are added on both sides and are removed from the series after wavelet analysis.

Suppose we collect \( n \) evenly spaced observations on the function \( f \), denoted \( y_i, i = 0, \ldots, n - 1 \). One possibility is to consider the observations \( \{y_0, y_1, \ldots, y_{n-1}\} \) as a sample from an infinite series with periodic behaviour \( \{\ldots, y_{n-2}, y_{n-1}, y_0, y_1, \ldots, y_{n-2}, y_{n-1}, y_0, y_1 \ldots\} \) and to extend the observed series accordingly. However, in cases such as the HPLC data explored in Chapter 9, where a drift leads to \( y_0 \) and \( y_{n-1} \) being substantially dissimilar, a more appropriate extension approach is to symmetrically reflect the observed data series at each end to the required dyadic length i.e. \( \{\ldots, y_1, y_0, y_0, y_1, \ldots, y_{n-2}, y_{n-1}, y_{n-1}, y_{n-2} \ldots\} \). A discussion of these approaches is provided by Smith and Eddins (1990). We use the symmetric reflection approach when padding is necessary in the remainder of this thesis.

### 7.2 Removing noise from a signal

An observed data series can be considered to contain observations from the true function plus a noise element. This can be described by the simple univariate model

\[
y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n,
\]

where we assume that the \( \epsilon_i \) are independent \( N(0, \sigma^2) \) random variables, that the sample points are equally spaced and that the sample size \( n \) is a power of 2, i.e. \( n = 2^J \) for some \( J > 0 \).
Chapter 7. Wavelet methodology

Given the sample observations \( y_1, \ldots, y_n \), for \( n = 2^J \), one can apply the discrete wavelet transform to yield a sequence of wavelet detail coefficients \( d_{j,k} \), of length \( n - 1 \). It has been noted (Donoho and Johnstone, 1994; Donoho and Johnstone, 1995) that for most spatially inhomogeneous functions, much of the activity in the wavelet domain is concentrated in a small subset of \((j, k)\)-space and manifests itself in only a few, but usually large, wavelet coefficients i.e. it is a sparse representation of the function \( f \). Due to the orthogonality of the discrete wavelet transform, white noise is simply transformed to white noise and conversely, this contaminates all wavelet coefficients equally (Donoho and Johnstone, 1994; Donoho and Johnstone, 1995).

Inducing such sparsity is a remarkable feat of wavelets: what follows is that the representation of a noisy function in the wavelet domain has a greater signal-to-noise ratio than in the time domain. Thus, attempting to identify and eliminate noise from a signal is less of a challenge in the wavelet domain than it would be in the original time domain. Many researchers have mused upon ideas in this area and in this section we describe some of the more important contributions. All are linked through a joint aim, that being to estimate the underlying function \( f \) from the noisy sample observations \( y_i \), without assuming any particular parametric structure for \( f \).

Wavelet shrinkage was a technique first developed by Donoho and Johnstone (1994, 1995) to “denoise” the wavelet transform coefficients, leaving only those representing the true function. The heuristic underlying wavelet shrinkage is the assumption that only the “larger” coefficients contain information regarding the behaviour of function \( f \); the many smaller coefficients are assumed to be associated with the noise, which uniformly contaminates all of the wavelet coefficients. By making a decision about which coefficients are “significantly” large, we can disregard all others by setting them to zero. If we then invert the “shrunk” coefficients, one should be able to recover a selective reconstruction, \( \hat{f} \) of the function \( f \). The resulting estimate can be expressed as
\[ \hat{f}(x) = c_{0,0} \phi(x) + \sum_{j=0}^{J-1} \sum_{k=0}^{2^j-1} d^*_{j,k} \psi_{j,k}(x), \]

where \( d^*_{j,k} \) are the thresholded detail coefficients. Thresholding in the wavelet domain achieves smoothing in the time domain.

### 7.2.1 Thresholding rules

Thresholding is essentially a two-step process. Firstly, some assessment of the relative size of each coefficient is made, which involves comparing the set of coefficients \( \{d_{j,k}\} \) to a selected threshold. Secondly, a thresholding scheme is applied, the particulars of which determine whether coefficients are set to zero, reduced in magnitude or unchanged.

Donoho and Johnstone (1994) pursued the development of two popular thresholding schemes, known as hard and soft. The difference between the two lies in how the coefficients \( d_{j,k} \) are processed when they are larger in absolute value than some threshold \( \lambda > 0 \).

Hard thresholding is often referred to as a “keep or kill” rule. A coefficient \( d_{j,k} \) larger in absolute magnitude than the threshold \( \lambda \) is retained without modification. A coefficient \( d_{j,k} \) smaller in magnitude than the threshold \( \lambda \) is set to zero. Analytically, this policy can be expressed as

\[
\delta^h(d_{j,k}, \lambda) = \begin{cases} 
    d_{j,k} & \text{if } |d_{j,k}| > \lambda \\
    0 & \text{otherwise.}
\end{cases}
\]

Soft thresholding can be interpreted as a “shrink or kill” rule. A coefficient \( d_{j,k} \) larger in absolute magnitude than the threshold \( \lambda \) is reduced in size by an amount \( \lambda \). A coefficient \( d_{j,k} \) smaller in magnitude than the threshold \( \lambda \) is set to zero. Analytically,
\[ \delta^s(d_{j,k}) = \text{sign}(d_{j,k})(|d_{j,k}| - \lambda)_+ \] (7.26)

where

\[ x_+ = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases} \]

It can be shown that hard thresholding gives rise to a larger variance in the function estimate, while soft thresholding produces larger bias, since it shrinks the big coefficients uniformly towards 0 by \( \lambda \). Compromise can be found in the existence of other thresholding schemes. An intermediary trade-off between hard and soft thresholding is obtained with mid- or firm-thresholding, however we do not pursue the use of these rules in this thesis. Discussion of these alternative schemes can be found in Percival and Walden (2000) and Bruce and Gao (1997).

### 7.2.2 Selecting a threshold

The aim of thresholding is to remove noise from the sample in order to try and recover the best approximation to the underlying function. If the threshold \( \lambda \) is too large, there is the possibility that too much of the signal coefficients will be removed by thresholding; key parts of the true function will be missed in the reconstruction and the data will be “over-smoothed”. Conversely, with a threshold that is too small, numerous noise coefficients will be retained and included in the reconstruction leading to an “under-smoothed” estimate. Thus, it is clear that choosing an appropriate threshold is crucial to the effectiveness of the processes discussed above.

**Universal thresholding**

The universal threshold was proposed by Donoho and Johnstone (1994) and is defined as \( \lambda_{UNIV} = \sigma \sqrt{2 \log(n)} \), where \( \sigma \) is the standard deviation of the noise coefficients and \( n \) is
The justification for the form of the universal threshold is to remove all wavelet coefficients smaller in magnitude than the expected maximum of an assumed independent and identically distributed normal noise sequence of the same size. In practice, the value of $\sigma$ is unknown thus should be replaced with a suitable estimate $\hat{\sigma}$ from the sample observations. Typically this estimate is given by the median absolute deviation of the estimated wavelet coefficients from the median at the finest level:

$$
\hat{\sigma} = 1.48258 \times \text{median} \left( \left| \{d_{j-1,k}\} - \text{median} \left( \{d_{j-1,k}\} \right) \right| \right) \quad k \in \mathbb{Z}, \quad (7.27)
$$

where the factor 1.48258 is required to ensure that the estimator is consistent when applied to observations from a normal distribution. Here “consistent” is used to mean that the variance of the estimator tends to zero as the size of the sample from which the estimate is made tends to infinity.

The universal threshold is very efficient at noise removal. However, it has a tendency to remove part of the underlying function also (Percival and Walden, 2000). The impact of these misclassifications is that the reconstructed function can be over-smoothed.

Various other thresholds have been suggested as alternatives to the universal threshold: Donoho and Johnstone (1994) derived the minimax threshold which is similar to the universal threshold in that it is predetermined, however it is not as extreme in its noise removal and thus retains more of the underlying function; in Donoho (1995), the authors developed a SureShrink thresholding approach, which aims to minimise the mean squared error of reconstruction by minimising Stein’s unbiased risk estimate (Stein, 1981) for the mean squared error of soft thresholding. The SURE threshold is data and level dependent; however, it is often the least efficient of the methods presented for removing noise as it generally yields threshold values smaller in magnitude than the universal threshold. Further, Donoho (1995), in a bid to achieve greater adaptivity of the threshold to the
data, suggested that the choice between SURE and universal thresholds should be made according to the result of a test for sparsity; Bruce and Gao (1996) also provide details.

**Thresholding based on the false discovery rate**

The problem of hard thresholding wavelet coefficients may be viewed in a multiple hypothesis test setting (see Abramovich and Benjamini (1996) and Abramovich et al. (2000)), as we are essentially testing each detail coefficient for a significant departure from zero. Under the null hypothesis, a particular wavelet coefficient is considered to be zero; this can be tested against a two-sided alternative hypothesis that it is non-zero. Assuming that the noise effect is normally distributed with a zero mean, an appropriate test statistic for each wavelet noise coefficient is the two-sided *p*-value, given as

\[
p_{j,k} = 2 \left[ 1 - \Phi \left( \frac{d_{j,k}}{\sigma} \right) \right],
\]

(7.28)

where \( \Phi(\cdot) \) denotes the standard normal cumulative distribution function and \( \sigma \) is the standard deviation of the noise coefficients.

Thus a particular detail coefficient \( d_{j,k} \) would be thresholded to zero if its corresponding \( p_{j,k} \) were greater than some pre-chosen significance level, \( \alpha \). Alternatively, if \( p_{j,k} < \alpha \), detail coefficient \( d_{j,k} \) would be retained as non-zero and “interesting”. Applying such an approach to each of the \( n - 1 \) detail coefficients induces severe multiple testing issues: we would expect, on average, \( \alpha(n - 1) \) of the tests to be significant even when the null hypothesis is true. This will give rise to noise coefficients being falsely retained in the model; in turn, this will impede on the success of the signal reconstruction. In response to this problem, Abramovich and Benjamini (1996) incorporated the false discovery rate work of Benjamini and Hochberg (1995) (as described in Section 2.3.2 of this thesis) to yield a threshold that should maximise the number of detail coefficients retained in the

---

*Chapter 7. Wavelet methodology*
model whilst simultaneously maintaining an FDR of less than or equal to $100\alpha\%$. The procedure can be described simply in a few simple steps:

- Calculate a $p$-value for each detail coefficient $d_{j,k}$, according to equation (7.28).
- Order the $p$-values by increasing magnitude i.e. $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(n-1)}$.
- The largest value of $i$ for which $p_{(i)} \leq \alpha i/(n - 1)$ holds true is denoted by $i^\star$.
- The threshold is defined by

$$\lambda_{FDR} = \sigma \Phi^{-1}(1 - p_{(i^\star)}/2),$$

where $\sigma$ is the standard deviation of the detail coefficients. This value can be applied within a hard thresholding capacity to all $d_{j,k}$.

**Block thresholding**

Block thresholding approaches (Cai and Silverman, 2001; Cai, 2002; Hall et al., 1998) aim to borrow information from neighbouring coefficients when setting the threshold for each individual $d_{j,k}$. The method of Cai (2002) partitions the data into distinct blocks and proceeds to threshold all the data contained within each block. The intricacies of the thresholding scheme adopted for a particular block are driven according to the sum of squares of the data in that block. A similar approach is taken in Cai and Silverman (2001), however here the difference is that the blocks are designed to be overlapping. The coefficients in the middle of each block are subjected to hard thresholding i.e. they are either thresholded to zero or retained unaltered, according to the sum of squares over the whole block.
Cross-validation

Cross-validation is a commonly adopted practice in a variety of statistical settings for the purposes of selecting a smoothing parameter (Green and Silverman, 1994; Silverman, 1986). The approach was first applied within a wavelet context by Nason (1996), who described implementations of two-fold cross-validation and leave-one-out cross-validation. Here we give an overview of the former; for discussion of the latter, the reader should consult Nason (1996).

The two-fold cross-validation method can be used to automatically choose a threshold $\lambda$ for a wavelet shrinkage estimator based on $2^J$ data points. As expected, the rationale behind the use of cross-validation in such a context is the minimisation of the mean integrated square error between the wavelet shrinkage estimator $\hat{f}(x)$ and the true function $f(x)$, i.e.

$$E \int \left\{ \hat{f}(x) - f(x) \right\}^2 dx.$$

Very generally, the procedure of Nason (1996) works by first leaving out half of the data points. The remaining $2^{J-1}$ data points are then used to form a wavelet shrinkage estimator using a particular threshold value; the prediction error associated with this threshold value is subsequently estimated by comparing the values of the excluded points with the shrinkage estimator. This quantity can then be numerically minimised over a range of values for the threshold. The algorithm for implementing such an approach in practice is described below.

Suppose that we have observed data $y_1, \ldots, y_n$ with $n = 2^J$. Firstly, the odd-indexed observations are removed from the original set, leaving the $2^{J-1}$ evenly indexed $y_i$, which are then reindexed from $j = 1, \ldots, 2^{J-1}$. A function estimate $\hat{f}^E$ is then constructed from the re-indexed $y_j$, using some threshold $\lambda_{CV} > 0$. In order to make a comparison of the
Chapter 7. Wavelet methodology

function estimator with the left-out noisy data, an interpolated version of $\hat{f}^E$ is necessary, which can be given as

$$\bar{f}_E^j = \frac{1}{2}(\hat{f}_{j+1}^E + \hat{f}_j^E), \quad j = 1, \ldots, n/2.$$  

The approach assumes that the function $f$ is periodic, thus $\hat{f}_{n/2+1}^E = \hat{f}_1^E$ where required. In a similar fashion, $\hat{f}^O$ is formed using the same threshold value and the interpolant $\bar{f}^O$ computed as above. In the penultimate step of the algorithm, the value of the threshold $\lambda_{CV}$ is varied in order to minimise

$$\sum_{j=1}^{n/2} \left\{ (\bar{f}_j^E - y_{2j+1})^2 + (\bar{f}_j^O - y_{2j})^2 \right\}.$$  

It is worth noting that in deriving an estimate of $\lambda_{CV}$, two estimates of $f$ are used, each of which is based upon $n/2$ data points. Thus, the final stage of the algorithm is to make an adjustment to the threshold in order to obtain a cross-validated threshold valid for all $n$ data points. The adjustment proposed by Nason (1996) was motivated by work on the universal threshold by Donoho and Johnstone (1994) and takes the form

$$\lambda_{CV}(n) \approx \left(1 - \frac{\log 2}{\log n}\right)^{-1/2} \lambda_{CV}(n/2).$$

Bayesian thresholding

Various Bayesian approaches have been proposed for thresholding wavelet coefficients; one should see Chipman et al. (1997), Abramovich et al. (1998), Clyde and George (1998), Clyde et al. (1998), Crouse et al. (1998), Johnstone and Silverman (1998), Vidakovic (1998) and Vannuci and Corradi (1999) for examples of this type of work. The general approach is to impose a prior distribution on the wavelet coefficients which
is designed to capture the sparseness of the wavelet expansion. In turn, the shrinkage estimation rule can be derived by applying a suitable Bayesian rule to the resulting posterior distribution of the wavelet coefficients. At this point it is pertinent to mention that the empirical Bayesian thresholding model discussed with regard to microarrays in Chapter 3 was originally developed by Johnstone and Silverman (2004) for this purpose.

### 7.2.3 Correlated errors

The above discussion assumes that the errors are independent and identically distributed. In cases where this assumption is not justified and the errors are autocorrelated, Johnstone and Silverman (1997) showed that the variance of the wavelet coefficients will be dependent on the level of the wavelet decomposition, but will be constant within such a level. Intuitively, this suggests that the coefficients at each level should be treated differently: the natural approach is to compare the \( \{d_{j,k}\} \) to a set of selected thresholds \( \{\lambda_j\} \), the values of which are dependent on the decomposition level \( j \).

### 7.3 The non-decimated wavelet transform

Unfortunately, a feature of the DWT is that it is not invariant under circular shifts. If discontinuities are present in the underlying signal, then wavelet denoising can give rise to oscillating pseudo-Gibbs phenomena near the location of this behaviour, alternatively undershooting and overshooting the desired target level. The artifacts are all connected in some way with the precise alignments between features in the signal and the features of basis elements; signals exhibiting similar features but with slightly different alignment in the time domain might generate less of the undesirable artifacts.

Coifman and Donoho (1995) introduced an approach to correct for this by forcibly shifting signals so that their discontinuities change positions, with the philosophy being
that an analysis of the shifted signal will not exhibit these undesirable artifacts and that this analysis can be later unshifted. It is important to be aware that this idea has also appeared in alternative guises (see the stationary wavelet transform of Nason and Silverman (1995) and the maximal overlap transform referred to by Percival and Walden (2000)).

Suppose we observe a signal \((y_i : 0 \leq i < n - 1)\). Now let \(S_c\) denote the circulant shift by \(c\), \((S_c y)_i = y_{(i+c)\text{mod}(n)}\). Instead of finding the wavelet denoised version of \(y\) as standard, one can find the denoised version of the horizontally shifted data vector, \((S_c y)\) and then unshift the data. However, knowledge of the “correct” choice of shift parameter \(c\) is elusive and where a signal contains many discontinuities, the best shift for one discontinuity may be the worst shift for another discontinuity. Coifman and Donoho (1995) explored an extension of such “cycle-spinning”, where a range of shifts is applied and the results obtained are averaged. Specifically, for a signal with length \(n\), they advocated averaging over all \(n\) circulant shifts, i.e. \(c = 0, \ldots, n - 1\). It was reported that this non-decimated transformation, particularly in conjunction with hard thresholding, offered superiority over the standard transform in terms of visual performance and quantitative similarity.

Whilst it appears that the additional complexity of such a transform will be computationally intensive, Coifman and Donoho (1995) developed an algorithm, parallel to that of Mallat (1989) for the standard DWT, that allows the non-decimated wavelet coefficients for a vector of \(n\) noisy observations to be calculated efficiently in order \(n \log_2(n)\) time. The coefficients are calculated similarly to the standard case, by a series of decimation and filtering operations but with an additional circulant shift incorporated. The intricacies of the implementation of such an algorithm are left for further reading and can be found in Coifman and Donoho (1995).

The thresholding schemes described in Section 7.2 were developed purposefully for the DWT, thus, due to the structure of the non-decimated transform are not applicable without some modification. Most of the literature in this area has been concentrated
around modifying the universal threshold as Coifman and Donoho (1995) prompted that
the universal threshold should be inflated when applied to the non-decimated transform.
In particular, Berkner and Wells (2002) proposed a formal modification to the universal
threshold which takes into account the correlation structure of the wavelet coefficients
induced by the non-decimated transform. However, here we simply use a inflation
factor induced naturally through the increase in the number of coefficients that require
thresholding. Mathematically, this can be expressed as

$$\lambda_{UNIV} = \sigma \sqrt{2 \log(n \log_2 n)}.$$  \hspace{1cm} (7.29)

### 7.4 The vaguelette-wavelet decomposition

In this section, for the purpose of easing the discussion of the vaguelette-wavelet
decomposition we move to matrix notation. We use the vector $f$ to represent the
underlying signal and the matrix $W$ to describe the action of the discrete wavelet
transform. The matrix $W$ is of dimension $n \times n$ and its form is associated with the
orthornormal wavelet basis selected. For further details on the formulation of the discrete
wavelet transformation matrix, one should refer to Vidakovic (1999).

Suppose that we have noisy observations $y$ from the model

$$y = f + \epsilon$$  \hspace{1cm} (7.30)

where $f$ is the signal of interest, contaminated by a vector of Gaussian white noise $\epsilon$.

Previously, we have provided discussion as to how one might extract the signal $f$ from the
noisy observations $y$. However, now suppose instead that we wish to recover the function
Chapter 7. Wavelet methodology

\[ g = g(f) \approx K^{-1} f, \]  

(7.31)

where \( K^{-1} \) is a matrix corresponding to a desired linear transformation.

This linear-inverse problem can be described as “ill-posed” when the matrix \( K^{-1} \) does not possess a well-defined inverse. In such cases, first estimating \( f \) and then applying the matrix \( K^{-1} \) is not a valid approach for estimating \( g \), unless some modifications are made.

Abramovich and Silverman (1998) proposed a “waveletized” method, parallel to a standard singular value decomposition and known as a vaguelette-wavelet decomposition, that could be used to find the function of an observed data series. We provide a description of this approach in Section 7.4.2, adopting the matrix notation as detailed in Baxter (2004), however we must first briefly review the counterpart wavelet-vaguelette approach of Donoho (1995). Further, we describe the necessary modifications to the thresholding procedures described in Section 7.2.

7.4.1 The wavelet-vaguelette transform

An established approach to solving linear-inverse problems presents itself in the form of singular value decomposition. The underlying construction of SVD techniques depends solely on the matrix \( K^{-1} \), with no regard for the specific physical nature of the problem under consideration. Consequently, it is known that such methods can prove restrictive for spatially inhomogeneous signals; see Donoho (1995) for further discussion.

The wavelet-vaguelette transform of Donoho (1995) provides a wavelet-based method which aims to circumvent these limitations, allowing a parsimonious representation of a wide range of functions; the particular spatial properties of the data under study can be reflected in the choice of wavelet filter used.

Donoho (1995) constructed the vaguelette transform, which is described by
\[ v = \mathcal{V} f = \mathcal{W}g, \]  

(7.32)

where \( \mathcal{V} \) is an \( n \times n \) vaguelette transform matrix and \( v \) is an \( n \)-vector of vaguelette coefficients.

The vaguelette transform (equation (7.32)) is assembled in such a way that the expansion of the vector \( f \) into a vaguelette series is equivalent to a wavelet expansion of the unknown function \( g \); thus the standard inverse DWT can be applied to the vaguelette coefficients of \( f \) to obtain \( g \).

The form of the vaguelette transform matrix is derived by considering the relationship in (7.31). Thus

\[ \mathcal{V} f = \mathcal{W}g \approx \mathcal{W}K^{-1}f \]  

(7.33)

and hence

\[ \mathcal{V} \approx \mathcal{W}K^{-1}. \]  

(7.34)

It should be noted that this result is only approximate because the matrix \( K^{-1} \) will only transform to approximate observations on the function \( g \).

Whilst the matrix \( \mathcal{V} \) is not perfectly orthonormal, Donoho (1995) demonstrated that for various transformation matrices \( K^{-1} \) and a number of wavelet filters \( \{h_l\} \), the transform \( \mathcal{V} \) will be close to orthonormal. Furthermore, the vaguelette coefficients will form a multiresolution analysis of \( g \).
Whilst the work of Donoho (1995) is attractive in its theoretical properties, in practice, calculating the vaguelette coefficients, \( v \), is computationally intensive and there is no direct analogy to Mallat’s algorithm.

### 7.4.2 The vaguelette-wavelet transform

In response to the computational difficulties involved in the wavelet-vaguelette decomposition, Abramovich and Silverman (1998) proposed the vaguelette-wavelet transform (VWT), which we consider in this section. Here the function \( f \) is expanded into a wavelet series. A mapping of the wavelet coefficients back into the vaguelette expansion in the original space yields the vaguelette-wavelet decomposition estimator of \( g \). The construction of this inverse mapping is described below.

We denote the DWT of \( g \) as

\[
\mathcal{G} = \mathcal{W}g,
\]

and the DWT of \( f \) as

\[
\mathcal{F} = \mathcal{W}f.
\]

Abramovich and Silverman (1998) equated the inverse vaguelette transform of the coefficients \( \mathcal{F} \) with the inverse DWT of the coefficients \( \mathcal{G} \):

\[
\mathcal{V}^{-1}\mathcal{F} = \mathcal{W}^T\mathcal{G} = g.
\]

The matrix \( \mathcal{V}^{-1} \) is of \( n \times n \) dimension and represents the inverse vaguelette transform. Similarly to the wavelet-vaguelette case, \( \mathcal{V}^{-1} \) will not be exactly orthonormal, however, for a suitably chosen matrices \( K^{-1} \) and \( \mathcal{W} \) it will be near orthonormal.
Substituting equations (7.35) and (7.36) into equation (7.37), it is clear that

\[ \mathcal{W}^T \mathcal{W} g = \mathcal{V}^{-1} \mathcal{W} f \]
\[ \Rightarrow K^{-1} f \approx \mathcal{V}^{-1} \mathcal{W} f \]
\[ \Rightarrow K^{-1} \mathcal{W}^T \approx \mathcal{V}^{-1}. \]  

(7.38)

Thus is it obvious that the form of \( \mathcal{V}^{-1} \) is dependent on the form of the matrix \( K^{-1} \).

Again in this derivation, the application of the inverse vaguelette transform directly to the coefficients \( \mathcal{F} \) can be computationally expensive: it is possible that a full \( O(n^2) \) operations will be required. However, in a key development, Abramovich and Silverman (1998) proposed that the inverse vaguelette transform could be looked upon as a plug-in estimator.

Substituting equation (7.38) into equation (7.37) it can be seen that

\[ g = \mathcal{V}^{-1} \mathcal{F} \]
\[ \approx K^{-1} (\mathcal{W}^T \mathcal{F}). \]  

(7.39)

In other words, taking an inverse vaguelette transform is equivalent to applying the matrix \( K^{-1} \) to the inverse DWT of the coefficients \( \mathcal{F} \). Mallat’s algorithm computes the standard inverse DWT in an efficient \( O(N) \) operations; consequently, the plug-in approach is computationally advantageous.

### 7.4.3 Thresholding the vaguelette-wavelet transform

The situation in the previous section describes a signal uncontaminated by noise. The problem is not quite as simple in practice and indeed we observe a noisy signal \( y \). Having decomposed the noisy signal using the DWT,
\[ w = \mathcal{W}y, \quad (7.40) \]

we can threshold the resulting wavelet coefficients to yield the vector \( \hat{w} \). Thus now

\[
\hat{g} = \mathcal{V}^{-1} \hat{w} \\
\approx K^{-1} (\mathcal{W}^T \hat{w}).
\quad (7.41)
\]

Given that the wavelet coefficients \( w \) are the product of applying the orthonormal DWT, the methods discussed in Section 7.2 are applicable here also. However, considering that the aim of the vaguelette-wavelet procedure is to find \( K^{-1} f \) rather than \( f \), threshold levels should be inflated accordingly, reflecting the ill-posed essence of the problem.

Abramovich and Silverman (1998) proposed inflating the universal threshold by a factor of \( \sqrt{1 + 2\alpha} \), resulting in a threshold of the form

\[
\lambda = \sigma_w \sqrt{2(1 + 2\alpha) \log(n)},
\quad (7.42)
\]

where \( \sigma_w \) denotes the standard deviation of the noise coefficients, \( n \) is the sample size and \( \alpha \) is a multiplier related to the specific form of the matrix \( K^{-1} \). In the case of differentiation, \( \alpha = 1 \). This threshold is deemed appropriate for signals distorted by independently and identically distributed Gaussian noise.
Chapter 8

Existing methods for processing spectral data

The analysis of spectral data poses many interesting, yet challenging, statistical questions. There is a wealth of literature to accompany the explosion of this type of experimentation within the proteomic domain; here we briefly review some of the existing methodology. Many statistical advances in this field tend never to appear in the statistical literature; instead they are confined to chemistry journals in particular.

Spectral data are inherently of a multiscale nature: a variety of effects, most typically localised in different frequency ranges, combine to produce the final signal. Thus, many researchers have considered the potential of wavelets for analysing data of this type. For example, the discrete wavelet transformation has been used for denoising (Coombes et al., 2005) and feature extraction (Qu et al., 2003; Randolph and Yasui, 2006). Lange et al. (2006) recently proposed using the continuous wavelet transform in peak detection and peak parameter estimation.
8.1 Pre-processing of spectral data

Roughly, a spectrum reduces to three components: a smooth baseline, which is typically a slowly varying trend line under the spectrum, a high-frequency noise element and the analytical signal, which we wish to extract. The signal usually will consist of a series of independent, possibly overlapping peaks, each corresponding to a protein or peptide, and should occupy a frequency range in between noise and baseline. The baseline is thought to consist of a very low frequency component of the observed signal; however, there is usually no theoretical model for such. In some cases, the baseline has been reported to depend on the presence of large and intense signals as well as on abundant low-intensity noise. The noise can be a mixture of various consistencies: chemical background (which is usually small), electronic noise, signal intensity fluctuations, warping of the signal shapes and noise in the isotopic clusters. Unfortunately, the level of noise can be quite intense and it is sometimes impossible to differentiate between that and the real signals.

Statistically, one might represent a spectrum by the model

$$y_i(t) = B_i(t) + N_i S_i(t) + \epsilon_{it}$$  \hspace{1cm} (8.1)$$

where $y_i(t)$ is the observed intensity of spectrum $i$ at time $t$, $B_i(t)$ is the baseline, $S_i(t)$ is the true signal of interest (i.e. a set of peaks), $N_i$ is a normalisation factor and $\epsilon_{it}$ is the noise (Morris et al., 2005). In its original proposed form, Gaussian white noise was adopted as the model for the final term in equation (8.1), as the observed noise was thought to originate from electronic noise in the detector. However all the terms can be specifically moulded to the particular dataset in question.

Pre-processing, essentially attempting to disentangle the three components of the model (8.1), forms an essential stage of any spectral analysis. The overall process can be crudely broken down into various subtasks, including spectral denoising, baseline
estimation and subtraction, peak detection and quantification, intensity normalisation and spectra calibration. These tasks are related to one another through complex interactions and possess a natural order and any deviation from its logical sequencing could severely impact the outcome of later classification algorithms. Spectral analysis is clearly very modular in its nature and most researchers strive to find improvements in only one or two of these areas. Consequently, there is a multitude of existing algorithms for all aspects of the analysis, yet very little work has explicitly looked for the optimal combinations of methods.

8.1.1 Denoising

Eliminating the noise from the signal as the primary step reaps the greatest benefit as, post-denoising, it becomes much easier to estimate and separate the baseline from the peaks. Moreover, peak identification reduces in complexity.

It is possible to regard the data described by equation (8.1) as $y_i(t) = f_i(t) + \epsilon_{it}$; this is a standard denoising situation. Various methods for this purpose have been explored over recent years. In order for a smoothing method to be useful, it is crucial that it does not perturb the peak shape, or at the least, its location along the horizontal axis and the width of the peak. Smoothing splines, wavelet smoothing and kernel methods, such as locally weighted linear regression were explored by Hastie et al. (2003). In general, information about true peaks in the time domain is likely represented by a small number of (relatively large) coefficients in the wavelet domain, and the noise should be distributed (at low levels) over most wavelet coefficients. Thus, Coombes et al. (2005) proposed a discrete wavelet transformation (DWT) and applied hard thresholding in the wavelet domain, based on a multiple of the median absolute deviation, before mapping back to the original domain. This method produced good results for low-resolution SELDI spectra. Qu et al. (2003) used an orthogonal (decimated) discrete wavelet transform to study SELDI data. They presented the wavelet transform as a tool for data reduction or feature selection.
and performed discriminant analysis on wavelet coefficients that exceeded a threshold. However, the justifications for doing so are limited as the coefficients in the wavelet domain do not hold any biological interpretation. Coombes et al. (2005) noted that when using the DWT for denoising, significant artifacts tended to be created near the ends of the signal. These artifacts were subject to substantial changes by shifting the starting point of the signal being transformed, a notion previously discussed in Section 7.3. To counteract these problems, Coombes et al. (2005) and Morris et al. (2005) moved to the non-decimated version of the DWT and reported that better visual and qualitative denoising were achieved through its incorporation.

In many experiments, replicate spectra are derived under the same experimental conditions and thus are not independent. Chamrad et al. (2003) considered the correlation between replicate spectra and used these as a basis to smooth the spectra and purge chemical noise or contaminants. They investigated sets of spectra acquired under the same experimental conditions, but with different analytes. It was concluded that the features common to the majority of spectra were likely to be surplus information, incurred due to sample preparation rather than the true content of the sample, thus could be discarded without losing valuable knowledge of the signal of interest.

After an appropriate smoothing method has been applied, an estimate of the noise can be obtained by calculating the expected deviation of the raw data from the smoothed curve in a particular mass or time window.

### 8.1.2 Baseline correction

After denoising the spectra, one should turn attention to estimating and subtracting the baseline. Very few assumptions are made about the baseline and, generally, relatively unsophisticated algorithms are employed for the purpose. However, it is important to note that estimating the baseline from sections of the spectra that contain many intense
peaks is likely to produce an estimate which can be dramatically distorted from the true drift.

The baseline is thought to occupy a considerably lower frequency than the signal of interest and the noise. Thus a technique such as the wavelet transform which decomposes a signal into components of different frequencies can be useful to separate drifting baseline from interesting signal. This idea has been explored in various forms of analytical chemistry (Shao et al., 1998; Shen et al., 1997; Barache et al., 1997; Gunther et al., 2002; Tan and Brown, 2002). In particular, Shao et al. (1999) considered an application to HPLC data. The experimental signal, $c_j$, of length $n = 2^J$ say, was decomposed in the wavelet domain to yield a set of discrete detail coefficients, $\{d_{j,k}\}$, and smooth coefficients, $\{c_{j,k}\}$. However, in a rather dubious twist, at each stage of the decomposition the smooth coefficients from the previous level were “stuffed” with zeros placed in arbitrary positions. The reason for this was to ensure for any resolution level there were $n = 2^J$ smooth coefficients (in other words, at each resolution level there was a vector of smooth coefficients that was the same length as the original signal). The smoothed approximation for some particular resolution level i.e. $\{c_{j,k} : j \text{ fixed}\}$ was chosen to resemble the drifting baseline, a multiple of which was then subtracted from the original trace (hence the need for the vector of $\{c_{j,k} : j \text{ fixed}\}$ to contain as many data points as the original sample). It should be stated that this particular approach is not being advocated here. As far as we are aware, there is no statistical justification for thinking that this approach will yield anything that truly resembles the real underlying baseline. Aside from this, the process of adding zeros in the above manner is highly questionable and the choice of resolution level at which the baseline allegedly manifests itself is far from clear and open to interpretation. Shao et al. (1999) made his selection based on visual examination, however for large-scale experiments involving many spectra, inspecting each one individually is likely to be infeasible.

Shackman et al. (2004) considered one of two approaches for estimating the baseline:
most simply, an \( n \)th-order polynomial was fitted to the entire data set or alternatively, the spectrum was broken up into \( x \) consecutive pieces and \( x \) \( n \)th-order polynomials fitted. The latter approach allowed for extraction of the baseline from peak components in the data. The \( n \)th-order polynomials were generally fitted by a singular value decomposition method. A similar method is evident in Constantinou et al. (2004), who adopted a simple polynomial curve fitting approach to the problem. Other smooth curves have been investigated in the literature: Yasui et al. (2003) and Tibshirani et al. (2004) both considered the use of a “loess” smoother.

Coombes et al. (2003) combined baseline correction and peak detection into a two-step algorithm. A tentative separation of the peaks was made prior to any baseline estimation. Subsequently, the baseline was calculated as a piecewise-constant interpolation of the local minima in a window. Having subtracted the baseline from the original spectrum, the peak detection algorithm was run again. A more simple approach was that of Shao et al. (1997), who estimated the baseline by linking the minimum points of every peak.

Baggerly et al. (2003) explored the use of the local median in a fixed time-scale window as a potential baseline estimator. This method was successful at removing general trends, especially those predominant at lower intensities where matrix elements were the main source of background noise. However, there was a tendency for broad peaks at higher \( m/z \) values to be removed, which could have been valuable signal. One would expect that this problem is not specific only to this type of smoother; when a peak is wider than the bandwidth, the baseline will inevitably rise. Further, with regard to the method of Baggerly et al. (2003), subtracting the local median also made it difficult to reliably quantify the height of a peak and gave negative intensities. To overcome these problems, the local median was replaced by the local minimum and a “semi-monotonic” curve was derived. Enforcing strict monotonicity was found to introduce additional biases. Coombes et al. (2005) estimated the baseline by an imitative “monotone local minimum” curve approach. The curve was established by assuming that the baseline follows the
Chapter 8. Existing methods for processing spectral data

spectrum whenever it decreases, and remains level if the spectrum increases. Further, they emphasised that denoising was critical to the success of baseline estimation: without it, the extremes of the noise at the lower end of the spectra tended to drive the estimated baseline below the actual baseline.

Less conventional approaches to estimating the baseline, such as filters from mathematical morphology (Breen et al., 2000; Soille, 2003) or high pass filters implemented with the fast Fourier transform (Press et al., 1995), have been explored as they are allegedly easy to implement and contain few parameters. However, the results are reported to be undesirably jagged.

All of these non-parametric baseline detection algorithms encounter difficulties when peaks in the spectra overlap due to the coelution of several compounds. A small peak may sit on the shoulder of a large, broad peak in which case the convention is to take the envelope of the large peak as the baseline for the smaller one. Alternatively, several larger peaks may overlap in which case the baseline is assumed to stay on the base level for the width of the convoluted section of signal.

8.1.3 Normalisation and transformation

Even after baseline correction and smoothing, it is possible that large experimental variations remain in the data, since the signal intensities can change between experiments due to different total analyte concentration or ionisation efficiency, for example. In order to even out these experimental variations, signals intensities are usually normalised, that is, the intensity values are transformed to new values, which are less dependent on the experimental conditions. Multiple mass spectra should be normalised individually before any comparisons of peak quantifications can be made across the set of spectra. This is normally done after smoothing and baseline subtraction.

The most commonly adopted approach for MALDI/SELDI data is normalisation by
the total ion count (TIC). In practice, the peak intensities are divided by the sum of all intensities (total ion count) of the spectra. This should result in observations less dependent on variations in laser intensity or matrix crystal formation. However, TIC works on the assumption that there are equal concentrations of proteins present in the sample; this may not have a sound biological basis when comparing two or more classes of samples. Cairns et al. (2007) challenge the suitability of this form of normalisation in all cases and suggest that in fact, biological features can be masked as a result of the process. An alternative strategy is to replace the intensities by their signal-to-noise ratios, where the noise is estimated in a window around a signal (Satten et al., 2004). Tibshirani et al. (2004) normalised each spectrum by a linear transformation that mapped the 10th and 90th percentiles to 0 and 1 respectively.

Many downstream peak detection algorithms assume a typical peak width. In some cases of mass spectrometry, peak widths can vary along the spectrum and may depend on the presence of overlapping peaks. For example, the signal width for time-of-flight data generally increases with the mass. A transformation, such as the logarithmic transform, can reduce this dependency (Tibshirani et al., 2004). Furthermore, it has been noted that the standard deviation of a peak’s intensities depends on the intensity itself. Similarly to microarray data, some form of variance stabilisation is necessary before the application of some statistical tests is justified. Considering SELDI and LC-MS data respectively, Coombes et al. (2005) and Wang et al. (2003) found a linear dependence; if this is the case, a logarithmic transformation of the intensities should theoretically produce constant standard deviation. However, similarly to microarray data, the success of such a transformation is not consistent across experiments: for an alternative SELDI dataset Coombes et al. (2003) reported that a cube root transformation was the most successful at stabilising the variance, whilst Anderle et al. (2004) found that a quadratic variance/intensity relation fitted the data well in his study of noise models for LC-MS experiments.
8.1.4 Peak detection and deconvolution

Analytical spectra usually contain several thousand sampling values, yet often have large regions that do not contain useful information. Therefore an obvious way to reduce the dimension of the dataset to a more manageable size is to extract the relevant features and discard the uninteresting regions. Furthermore, suppose one detects a significant difference between two groups in a noisy region of the spectra, it can be difficult to substantiate an explanation or extract a potential biomarker from the information. It is this lack of interpretability that motivates many researchers to apply a peak detection algorithm before any further analysis is carried out (Fung and Enderwick, 2002; Hilario et al. (2003); Wang et al., 2003; Yasui et al., 2003; Coombes et al., 2005; Tibshirani et al., 2004).

The simplest approach to peak detection is to locate all maxima in the denoised spectrum (Yasui et al., 2003; Tibshirani et al., 2004; Coombes et al., 2003; Coombes et al., 2005). A refinement of this selects those local maxima which have an intensity exceeding a given threshold value, determined by a signal-to-noise ratio (Coombes et al., 2005). The end points of the peaks can be defined as when, moving away from a particular maximum, a minimum is reached or the intensity drops below a pre-defined threshold.

Morris et al. (2005) recommended averaging replicate spectra before peak detection, however this requires the spectra to be roughly aligned. It was suggested that peak features should be at least preserved, if not enhanced, in the mean spectrum. The authors identified all local maxima in the average spectrum and an interval containing each peak. The left and right endpoints of the interval were determined by the \( m/z \) values at the nearest local minima to the left and right respectively of the local maximum. Subsequently, each peak identified in the average spectrum was quantified by the maximum intensity measurement from any of the individual spectra within the interval defining the peak on the average spectrum. An advantage of this approach was that it proved to be forgiving to slight
misalignments across spectra.

Yasui et al. (2003) did not attempt to quantify the peaks; instead they computed a binary indicator for the presence or absence of a peak. They defined a point on the graph to be a peak if it satisfied two properties: firstly, it had to be a local maximum in a fixed width window; secondly, it had to have an intensity value higher than the average intensity in a broad neighbourhood, where this average was computed using a super-smoother method in a window containing 5% of the data points.

Wallace et al. (2004) proposed a different technique to find peaks and troughs: starting with a straight line that connected the first point to the last in the spectrum, the algorithm found the point in the raw spectrum that was furthest from this line. This point was added as a new node in the piece-wise linear interpolation of the raw data and these steps were repeated until no significant peaks were left. Jarman et al. (2003) statistically assessed whether the histogram within a sliding window (ion counts vs. time or m/z bins) resembled a uniform distribution or had a peaked shape. The test considered baseline and noise in the raw data and was performed for varying window widths, in order to cope with different peak widths.

Randolph and Yasui (2006) used a multiscale wavelet decomposition of spectra and focused on local structure occurring at various scales. The philosophy behind this approach was that signal content could be identified and quantified without first smoothing the spectra, estimating signal-to-noise ratio or modelling a baseline. Scale-specific features were defined as local extrema in functions that arose from the multiresolution decomposition of the original spectrum. The location of a scale-based feature was recorded for each spectrum and then a histogram of the locations from all samples was constructed. Subsequently, the histogram served as spectral-density-like function: high-density regions of features in the data, distinguished across all spectra, were indicated by peaks in the histogram. Randolph and Yasui (2006) claimed that identifying features prior to averaging across spectra or smoothing was advantageous: peaks were not “lost” in
performing these processes. To deal with varying peak widths along the spectrum, it was suggested that analysing a combination of scales may be useful. The choice of scales was somewhat arbitrary, however it was hypothesised that the useful information was unlikely to be contained in the most extreme scales: these were generally too wide to pick up localised peptide features, or too fine, leading to something resembling that which might result from a set of pure noise spectra. The authors pointed out that in fact the features manifested themselves across a range of wavelet scales making the identification of the features less sensitive to the specific choice of scale than if this was not the case.

Lange et al. (2006) proposed building a peak picking algorithm around the continuous wavelet transform (CWT), on the basis that the maximum position in the CWT coincides with the maximum position in the data (Wu et al., 2001). The spectrum was initially split into segments. The peaks were identified and parameters estimated from a certain scale in the wavelet domain, before a pre-determined asymmetric peak function was fitted to the raw data. The method proved useful for resolving overlapping peaks however, selecting which CWT scales should be used for different segments of the spectrum is not necessarily obvious.

Du et al. (2006) proposed a novel algorithm which identified the peaks by applying a continuous wavelet transform-based pattern matching scheme. This approach boasted the advantage of making use of the additional information present in the shape of the peaks, greatly enhancing the estimation of the signal-to-noise ratio and robustly identifying peaks across scales and with varying intensities in a single spectrum. The authors demonstrated that using this technique allowed both large and small peaks to be detected whilst maintaining a high sensitivity and low FDR. A further advantage of this method was that the CWT could be applied directly to the raw spectra thus biases introduced through different smoothing algorithms and baseline removal schemes avoided.

A disadvantage of the many existing non-parametric methods is that they do not have the ability to strongly detect overlapping peaks. The deconvolution of peaks is a universal
problem for any analyses involving this form of data. Various approaches have been proposed for curve resolution, including the Fourier self-deconvolution method (Jackson and Griffiths, 1991; Kauppinen et al., 1981) and the Fourier derivation method (Cameron and Moffatt, 1987). Both of these methods worked by multiplying the Fourier transform of the original signal by a weighted function, which decayed more slowly. The multiplied signal was then transformed back to the time domain. Doing so increased the higher frequency part of the time domain signal and made the broad peak narrower. In Vivo-Truyols et al. (2005a) the authors discussed various existing deconvolution algorithms (Press et al., 1995; Vivo-Truyols et al., 2001; Hart, 1994; Vivo-Truyols et al., 2002) and considered an automatic approach for selecting the algorithm most appropriate to the data in question.

Shao et al. (1997, 1998) used a novel application of the wavelet transform to separate overlapping chromatographic peaks. The method is based on the idea that the wavelet transform decomposes the original signal into contributions that represent information of different frequencies. Furthermore, the wavelet transform is a linear operation thus retains the linearity of the decomposed signals. The method assumed that an overlapping chromatographic peak is of relatively low frequency and is essentially the sum of the chromatograms of each separate component, which are in turn signals of relatively high frequency. The higher frequency part was retrieved by applying a discrete wavelet transform and choosing the decomposition level at which any quantitative analysis was made to be the one with the greatest resolution of the separated peaks. In Shao and Cai (1999), exactly the same approach was employed but incorporating wavelet packets. Wavelet packets are outside the scope of our work but their inclusion was reported to lead to better results than the standard wavelet transform.

In a further paper Shao et al. (2000a) returned to the wavelet transform, this time the aim being to increase the resolved information in the overlapping peaks. The signal was decomposed into its wavelet domain counterpart and components representing the
resolved information were selected, somewhat arbitrarily. These selected components are multiplied up by a factor $k$ with $k > 1.0$ and the signal was reconstructed in the normal way using the inverse DWT.

Derivative techniques are often used to identify peaks and separate overlapping peaks as they offer higher resolution of the differential data, compared with the original signal. Vivo-Truyols et al. (2005b) approached peak detection by considering changes in the higher-order derivatives of spectra, in particular, finding negative regions in the second derivative.

Shackman et al. (2004) located peaks by determining the time at which the derivative of the signal crossed a threshold, defined as a multiple of the standard deviation of the derivative in a region known to be baseline. In order for a peak to be noted, the derivative was required to cross the positive threshold twice and similarly for the negative threshold. A minimum peak width was defined so that spikes originating from noise were excluded. The authors incorporated an automatic peak deconvolution procedure using an empirically transformed Gaussian function however this was shown to be only moderately successful unless the components comprising the overlapping peak were known.

Unfortunately, noise levels tend to be increased in calculations of the higher-order derivative. In response to this, Leung et al. (1998) proposed a method which made use of the wavelet transform for calculating approximate higher-order derivatives. Whilst being built on little theoretical justification, the technique boasted enhanced signal-to-noise ratios. The first derivative was expressed approximately as the difference between the scaling coefficients generated from any two Daubechies wavelet functions at different scales. Higher order approximate derivatives could be obtained from the previous order derivative. Shao et al. (2000b) proposed an alternative scheme for estimating an approximate derivative, based on the CWT with the Haar wavelet. Briefly, the $n$th derivative was found by applying the wavelet transform $n$ times to the analytical signal.

A parametric approach to peptide signal detection is to consider matching a given model
Chapter 8. Existing methods for processing spectral data

of a peptide signal against the raw data. Where the match exceeds a certain threshold a signal is assumed to be present. Such a model has the potential to contain various parameters, including an offset from the baseline, signal height and the width of peak (Berndt et al., 1999; Gras et al., 1999). Gras et al. (1999) used a matched filter approach to locate potential peaks and then carried out a linear regression in order to adjust the peak heights and widths. The fitted models were then subtracted from the raw data, and the algorithm was run again in order to find overlapping peaks.

Morris et al. (2006) proposed a novel approach for analysing MALDI-TOF data, using Bayesian wavelet-based functional mixed models in which the mass spectra were modelled as functions, removing any reliance on peak detection methods. This method had a flexible non-parametric representation of fixed and random effects, allowing it to simultaneously model the functional effects of a number of factors, including those of interest in addition to nuisance factors related to the experimental design. The method had potential to produce posterior probabilities to identify regions of interest within the spectra, taking into account both statistical and practical significance, whilst controlling the FDR at a specified level.

8.1.5 Alignment

Considering whether a set of spectra are well aligned is a crucial step in pre-processing, especially when inference is to be made from multiple spectra. Whilst variation in peak position may be indicative of some relevant interesting biological feature worth noting, in comparative studies, small shifts in the retention times or $m/z$ values can blur the distinction between groups of samples. Thus it is necessary to find some balance between an alignment procedure that is too stringent and one that is too conservative.

The simplest way to align multiple spectra is to include known reference analytes in the mixture to be separated. If these can be identified on the spectra, then the spectra
can be shifted accordingly. Provided that two or more reference values are included, errors can be corrected approximately by a simple affine transformation (Egelhofer et al., 2000; Gentzel et al., 2003). If many evenly distributed reference peaks are present, higher order correction polynomials can show substantial improvements over an affine correction (Gobom et al., 2002). Malmquist (1994) aligned the highest peaks first and then refined the calibration by considering those smaller peaks which are highly correlated across spectra.

A more recent paper by Crockford et al. (2005) derived an algorithm to find best-fit transformations between peaks in reference compound spectra and the corresponding peaks in experimental spectra. This method is attractive as it is shift-invariant and makes the alignment problem redundant. Driven by prior knowledge, a “probe” was identified from the pure compound spectrum and a search range defined. Quite simply, the search range was swept over by the probe seeking a match. The quality of a match was described by a form of squared difference statistic and optimal parameters were found by minimising this value.

Heatmaps are often used (Baggerly et al., 2004) due the simplicity of their implementation. All of the spectra are plotted on top of each other and the intensity is represented on a gray-scale. Vertical bands suggest peaks that are present in many spectra. At a finer scale, they show the alignment of the peaks across the multiple spectra and provide hints as to whether the horizontal scales of the spectra are well aligned.

Vogels et al. (1996) described a partial linear fit technique. The method involved the set of spectra being divided into segments, each of which was compared to the mean spectrum for the entire set. Each segment was shifted in order to minimise the squared difference to the mean spectrum and the spectra were reconstructed from these scale-adjusted fragments. Alternatively, Torgrip et al. (2003) proposed an algorithm based on the shifting and matching of peak positions using graph theory.

Tibshirani et al. (2004) considered alignment and peak detection as a 2-step process. To
align the spectra, the authors applied complete hierarchical clustering to the collection of all identified peaks from the individual spectra. This approach was based on the idea that tight clusters should represent the same biological peak that has been horizontally shifted in different spectra. The centroid (mean position) of each cluster was extracted to represent the “consensus” position for that peak across all spectra. Given a list of common peaks from the clustering, the individual spectra were then reconsidered: a peak in an individual spectra was deemed to be present and one of the common peaks if its centre lay within some distance of the estimated centre position of the common peak. An alternative approach is to cluster peaks across spectra that lie within a small fixed window. Fung and Enderwick (2002), Yasui et al. (2003) and Prados et al. (2004) discuss their clustering strategies in more detail.

Dynamic time warping (DTW) is frequently used in signal processing tasks in order to align warped signals (Aach and Church, 2001) and has been used with applications to liquid chromatography spectra (Wang and Isenhour, 1987; Nielsen et al., 1998; Eilers, 2004). The approach is based on a dynamic programming algorithm, which finds a globally optimal solution maximising the similarity between two signals. However, a distinct disadvantage of such a method is that it can be quite time and memory consuming for long spectra.

8.1.6 Peak quantification

For non-parametric peak detection, there are two options to quantify peaks: peak height or peak area above the baseline. Peak height (Coombes et al., 2003) is less sensitive to disturbance by other overlapping signals, but it neglects the width of the signal. Peak area is believed to be a better quantification of the signal than the amplitude of the peaks (Hilario et al., 2006) as it considers the full signal and averages out random noise. However, for this method to be effective the beginning and ends of peaks have to be well defined. Where this approach is adopted, the peak area is generally
calculated by simple numerical integration (Shackman et al., 2004; Lange et al., 2006) i.e. repeated applications of the trapezoidal rule. An alternative approach results from the peak detection algorithm derived by Wallace et al. (2004), which induced the reduction of the dataset to several groups of three strategic points, where generally each group corresponded to the beginning, apex and end of a peak \( p \). Thus each peak was represented by a polygon (in the most straightforward case, a triangle); the area of the true peak was taken to be the area of the corresponding polygon.

With regard to our HPLC data, we anticipate that key issues to be tackled will be denoising, baseline estimation and peak detection, before we can quantify the areas under the peaks. We consider the development of statistical tools for these purposes in the next chapter.

### 8.2 Post-processing analysis

Discriminatory approaches and classification algorithms are outside the scope of this work. However, there are many researchers active in this area (Tibshirani et al., 2004; Qu et al., 2003; Constantinou et al., 2004; Du et al., 2006). Hilario et al. (2006) provides a comprehensive overview of many of these ideas.
Chapter 8. Existing methods for processing spectral data
Chapter 9

Novel approaches to analysing HPLC data

In this chapter we consider the application of the methodology introduced in Chapter 7 to the real-life HPLC data set described in Section 6.4. The key aims of the chapter are to improve on the existing methods for obtaining the derivative of the trace and to investigate approaches for assigning confidence intervals to the estimates of the peaks’ areas. However, as previously mentioned, the process of spectral analysis involves a complicated disentangling of various components, thus some consideration has to be given to the other issues, such as baseline estimation and peak detection.

All of the wavelet methodology described here was implemented through the WaveThresh3 package in R (Nason, 1998).

9.1 Denoising

In the first instance we considered the usefulness of averaging the replicate spectra prior to denoising, in a manner as suggested by Morris et al. (2005). However, due to a lack of
sufficient alignment between the multiple spectra (evident in the top panel of Figure 9.1), this resulted in the creation of unwanted artifacts, such as additional peaks, in the average spectrum. These effects can be observed in the bottom panel of Figure 9.1. For example, around 10-12 minutes, a spurious subsidiary peak has been introduced.

Figure 9.1: Top panel: Plotting the multiple spectra illustrates the lack of alignment across the samples. Bottom panel: Mean trace after averaging replicate standards.
Alternatively, our strategy will be to consider each individual trace separately. We propose decomposing each trace with the non-decimated wavelet transform, in conjunction with the Haar wavelet. The motivation behind the choice of the Haar wavelet is that the resulting wavelet coefficients will have some significance in the original context (Kolaczyk, 1997). Through the nature of their construction, the detail coefficients will be proportional to the difference in counts between adjacent time intervals since

$$d_{j,k} = 2^{-1/2}(c_{j+1,2k+1} - c_{j+1,2k}),$$

(9.1)

where the data are taken to be the finest level, \(c_J\). Similarly, the scaling coefficients will be proportional to the actual number of counts in time intervals of dyadic length. However, due to the block nature of the Haar wavelet, “step-like” phenomena often transpire in the reconstructed trace, particularly in regions of smooth behaviour. This undesirable effect is illustrated in the top panel of Figure 9.2. The non-decimated transform suppresses artifacts such as these, hence is the motivation behind this choice. The improvements are evident in the bottom panel of Figure 9.2. In both of these cases, we combined the universal threshold with the hard policy.

Having verified the stationarity of the residuals obtained from the fitted wavelet model, examination of their autocorrelation and partial autocorrelation functions (Figure 9.3) suggests that the errors are not independent. Moreover, the cyclical decay evident in the ACF and the absence of any infinite geometric decay in the PACF suggest that the noise follows some autoregressive time series model. Hypothesising as to what process might have generated such correlations is more of a biological question than a statistical one, and outside the scope of this work. However, it is most likely to be attenuated to some underlying technological issue; similar observations were made by Baggerly et al. (2003).

The conventional approach to denoising signals in the presence of correlated errors is to
Figure 9.2: Top panel: Implications of denoising with the decimated Haar transform, after which the “step-like” phenomena are very evident. Bottom panel: Raw trace (dot-dashed line) with the reconstructed trace after applying the non-decimated version of the Haar wavelet transform (solid line). In both cases hard thresholding was employed with the universal threshold.
Figure 9.3: Left panel: Autocorrelation function of residuals. Right panel: Partial autocorrelation function of residuals. In real time, each lag equals 0.00667 of a minute.
threshold by level as described in Section 7.2.3. However, having explored the application of this approach to the HPLC dataset in question, it is evident that the results produced are not necessarily superior. In this case, the generation of artificial “lumps and bumps” in the reconstructed trace contributes to less than desirable results. These features are clearly illustrated in Figure 9.4.

An alternative approach is to adopt the empirical Bayesian thresholding procedure of Johnstone and Silverman (2004), which was described extensively in Chapter 3, for the purpose that it was originally proposed. The method is robust to the presence of mild correlation between observations (see Section 3.4.1), therefore is particularly useful in this situation. The results of such an approach are illustrated in the bottom panel of Figure 9.4.

9.2 Baseline estimation

Unlike their well established microarray counterparts, many of the existing pre-processing methods for spectral data, such as those described in Chapter 8, are not yet readily available through common software such as Bioconductor in R. Thus we propose and construct three simple algorithms of our own for the purpose of baseline estimation and inspect their various advantages and limitations.

**Local minimum curve approach** Very briefly, this approach starts from the observed intensity at the maximum time point and progresses backwards in time. The baseline remains constant until it meets the trace, after which it follows the trace when it decreases or otherwise remains constant until it hits the trace again. Mathematically, this can be expressed as
Figure 9.4: Top panel: Raw trace (dot-dashed line) with the reconstructed trace after thresholding by level (solid line). Bottom panel: Raw trace (dot-dashed line) with the reconstructed trace after applying empirical Bayesian thresholding (solid line).
Chapter 9. Novel approaches to analysing HPLC data

\[ B(t) = f(t) \quad \text{for } t = \max(t) \]

\[ B(t) = \min(B(t+1), f(t)) \quad \text{for all other } t, \]

where \( B(t) \) is the baseline at time \( t \) and \( f(t) \) is the value of the observed trace at time \( t \). In practice, this resulted in a rather “step-like” function for the baseline.

**Least squares isotone regression approach**  Least squares isotone regression is a method which will yield a non-decreasing sequence from a set of data points. In the HPLC context under discussion, the non-decreasing baseline is obtained by minimising the loss function which measures the fit between the trace and estimated baseline. For further information, one should consult Kruskal (1964).

**Gradient changepoints approach**  A tentative separation of the peaks is made from the denoised trace. Having done so, we assume that the majority of what remains is baseline, from which we can make a robust estimate of such without any interference or distortion from the peaks.

Such an approach was implemented by firstly calculating the gradient \( g \) at each time increment \( t \), expressed numerically as

\[ g(t) = \frac{\hat{f}(t) - \hat{f}(t-1)}{x(t) - x(t-1)}, \quad (9.2) \]

where \( \hat{f}(t) \) is the value of the denoised trace at time \( t \).

A threshold was defined to be

\[ \bar{g} \pm k \times \sigma, \quad (9.3) \]

where \( \sigma \) is the standard deviation of the gradient function, estimated by the mad estimator, and \( k \) is a user-defined factor. We adopted \( k = 2 \) after exploratory investigations.
Starting from $t = 0$, the startpoints of the peaks were identified as the points on the trace where the gradient function exceeds $\bar{g} + k \times \sigma$ for a period of $c$ consecutive clock ticks. Starting with $t = \max(t)$, a similar process was applied using $\bar{g} - k \times \sigma$ to identify the peak endpoints. Finally, the baseline was constructed by linearly interpolating between consecutive start- and endpoints. We introduced an additional constraint that the interpolated segments should be strictly increasing, to prevent any distortion of the baseline by any overlapping peaks.

Enforcing such an approach gave rise to a rather disjoint baseline; an further undesirable feature was that some in segments, the baseline exceeded the true trace. This can create difficulties further down the line as baseline subtraction will yield negative intensities. Thus to refine the procedure, the baseline was constructed by following the denoised trace in segments between the peaks and linearly interpolating across the peaks.

Results from all three approaches are shown below in Figure 9.5. It seems more natural, given the type of data, to have a smooth baseline rather than one so strictly “step-like”, therefore from this point onwards, we take the baseline to be estimated by the gradient changepoints approach. However, regardless of the method adopted, the resulting estimate of the baseline should be subtracted from the denoised trace. Assuming that the baseline estimation is completed successfully, this should ideally produce a trace with a horizontal baseline at zero, substantially simplifying peak detection. An example of such a baseline-corrected trace can be seen in Figure 9.5. It is now relatively simple to identify the $p$ peaks (in this case $p = 8$) and the corresponding start- and endpoints; by convention, the next step would be to estimate the area under the $p$th peak by simple numerical integration between its start- and endpoint.
Figure 9.5: Top panel: Denoised trace (solid line) with baselines estimated by local minimum curve approach (dotted line), least squares isotone regression approach (dot-dashed line) and gradient changepoints approach (dashed line). Bottom panel: Trace after correcting for increasing baseline by the gradient changepoint method.
9.3 Finding the derivative

Derivative techniques are often incorporated at various stages of spectral analysis, however, it is widely acknowledged that such an approach can be complicated by increased levels of noise, induced by numerically calculating the derivative. A detailed discussion of existing ideas based on derivative techniques and recent proposals to circumvent the issue of increased noise was provided in Section 8.1.4.

The problem of finding the derivative fits neatly into the generalised vaguelette-wavelet framework of Abramovich and Silverman (1998), as described in Section 7.4.2. Here we discuss the details specific to differentiation and explore the application of this approach within the HPLC domain.

9.3.1 Constructing the differentiation matrix

In the case of differentiation, we aim to recover the derivative function

\[ g = \left[ \frac{d}{dx} \right] f \approx K^{-1} f, \tag{9.4} \]

where \( K^{-1} \) is a differentiation matrix.

The differentiation matrix \( K^{-1} \) aims to take observations on the signal \( f \) and transform them to approximate observations on the derivative, \( g \). This transformation can take many forms however, we concentrate on those based on finite-difference coefficients \( \{ r_k \} \) and subject to the constraints

\[ \sum_k r_k = 0 \quad |r_k| \leq 1 \quad \forall k. \tag{9.5} \]
Chapter 9. Novel approaches to analysing HPLC data

210

The signal $f$ is assumed to be periodic outside of its domain of observations i.e. \{\(f_0, f_1, \ldots, f_{n-2}, f_{n-1}\)\} is viewed as a sample from the infinite sequence \{\(\ldots, f_{n-2}, f_{n-1}, f_0, f_1, \ldots, f_{n-2}, f_{n-1}, f_0, f_1, \ldots\)\}, for the purposes of simplicity. We do not anticipate that this will conflict with our earlier assumption of symmetric boundary conditions for two reasons: firstly, our chosen wavelet filter is short in length, thus any undesirable boundary effects will be minimal. Moreover, the features of interest that must be preserved i.e. the peaks, are contained within the middle region of the trace; the ends of the trace where distortions may occur will ultimately be discarded.

Thus, the rows of the \(n \times n\) matrix \(K^{-1}\) are built from circular shifts of the \(1 \times n\) vector:

\[
\mathbf{r} = (r_{-k}, r_{-k+1}, \ldots, r_0, \ldots, r_{k-1}, r_k, 0, 0, 0, \ldots).
\] (9.6)

For example, when \(k = 1\), the matrix \(K^{-1}\) will take the form

\[
K^{-1} = \begin{pmatrix}
  r_0 & r_1 & 0 & \ldots & 0 & 0 & r_{-1} \\
  r_{-1} & r_0 & r_1 & \ldots & 0 & 0 & 0 \\
  0 & r_{-1} & r_0 & \ldots & 0 & 0 & 0 \\
  \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
  0 & 0 & 0 & \ldots & r_0 & r_1 & 0 \\
  0 & 0 & 0 & \ldots & r_{-1} & r_0 & r_1 \\
  r_1 & 0 & 0 & \ldots & 0 & r_{-1} & r_0
\end{pmatrix}.
\]

Given the restrictions in (9.5), the matrix \(K^{-1}\) will not possess a well-defined inverse. The resulting linear-inverse problem will therefore be ill-posed.

The accuracy of the transformation, \(K^{-1}\), will depend on the structure of the signal \(f\) and the choice of filter coefficients, \(\{r_k\}\). The choice regarding the length of the filter can be regarded as a trade-off: longer filters will represent higher-order polynomial behaviour more accurately, however, as they take into account observations further from the point...
Chapter 9. Novel approaches to analysing HPLC data

at which the function is to be estimated, they provide a less instantaneous measure of the function, \( g \).

Given a set of scaling function filter coefficients \( \{h_l\} \), Jameson (1993) provided details on how to construct a set of finite-difference coefficients, \( \{r_k\} \) from the \( \{h_l\} \). It was shown that the coefficients \( \{r_k : k = -L + 2, \ldots, L - 2\} \) may be uniquely determined through their relationship with the filter coefficients \( \{h_l : l = 0, 1, \ldots, L - 1\} \), and are given by the solution to the system of linear algebraic equations:

\[
r_k = 2 \left[ r_{2k} + \frac{1}{2} \sum_{l=1}^{L/2} a_{2l-1}(r_{2k-2l+1} + r_{2k+2l-1}) \right], \tag{9.7}
\]

and

\[
\sum_{k=-L+2}^{L-2} kr_k = -1,
\]

where the coefficients \( a_{2l-1} \) are obtained from the \( \{h_l\} \) via the relationship

\[
a_n = 2 \sum_{i=0}^{L-1-n} h_i h_{i+n}, \quad n = 1, \ldots, L - 1.
\]

In the absence of noise, the resulting finite-difference filter is able to differentiate exact polynomials of degree \( L \). We use the same scaling function filter to produce the coefficients \( \{r_k\} \), as we use to construct the DWT matrix, \( W \). Consequently, the choice of finite-difference filter is coupled to the choice of wavelet filter.

9.3.2 Application to HPLC data

We use the vaguelette-wavelet plug-in estimator approach of Abramovich and Silverman (1998) to estimate the derivative of the HPLC trace. The implications of a less noisy
measure of the derivative could be considerable within this domain. Not only do we anticipate that it will have benefits in terms of our own work, but its incorporation into those existing and future methods that rely on derivative measures (for example, Vivo-Truyols et al. (2005a)) could lead to substantial improvements. Furthermore, higher-order derivatives are obtainable via the same approach; Beylkin (1992) provided a general description on how to construct finite-difference coefficients for the higher-order operator $d^m/dx^m$.

In Section 9.1, we established that the non-decimated transform yielded superior results to the standard decimated transform. Thus we combine the wavelet-vaguelette methodology with that of the non-decimated transform. We are not aware of any other literature which adopts this approach.

Essentially, in combining the two ideas, we could apply the matrix $K^{-1}$ to each of the $n$ unshifted denoised signals which comprise the non-decimated transform, prior to the execution of any averaging; subsequently the $n$ derivative estimates can be combined to yield an “average” measure. However, as the average of the derivatives is equivalent to the derivative of the averages, it is highly advantageous in terms of computational time, to simply apply the matrix $K^{-1}$ once to the estimated signal vector $\hat{f}$, resulting from applying the non-decimated transform in its completeness. Some adjustment will be necessary with regard to thresholding the non-decimated wavelet coefficients: in addition to the compensation made for using the non-decimated transform (see Section 7.3), the universal threshold is inflated by a factor of $\sqrt{1 + 2\alpha}$, as according to Section 7.4.3.

Previously, in the discussion on denoising, the Haar wavelet has been employed with substantial success. However, the relationship described in (9.7) is only sustained when $L \geq 4$, where $L$ is the length of the selected filter, thus eliminating the Haar wavelet as a potential choice. Hence, for the purposes of applying the vaguelette-wavelet methodology, we move to using the Daubechies Extremal Phase wavelet with a filter length of 4, which we will denote $D_4$. This wavelet was chosen above all others simply
for its superior end result, measured by visual inspection. Longer wavelets were less successful at “flattening” the plateaus between the “burstlike” peaks. Similar restrictions exist for obtaining higher-order derivatives: in order that the $m$th derivative may be representable in a wavelet basis, Beylkin (1992) established a necessary yet not sufficient condition in that $L \geq m + 1$.

![Figure 9.6: Comparison of derivative techniques along the length of the trace. Black line: Denoising with Haar wavelet followed by numerical differentiation. Red line: Denoising with Daubechies Extremal Phase wavelet with filter length 4 ($D_4$) followed by numerical differentiation, plus an offset of 50. Green line: Denoising with $D_4$ wavelet followed by vaguelette-wavelet approach (universal threshold), plus an offset value of 100. Blue line: Denoising $D_4$ wavelet followed by vaguelette-wavelet approach (empirical Bayesian threshold), plus an offset value of 150.](image)

Figures 9.6, 9.7 and 9.8 compare the derivative function of the real HPLC trace obtained by a variety of methods. The functions have been offset to aid visual presentation. We consider numerical differentiation of the reconstructed trace after wavelet denoising.
Figure 9.7: Close-up comparison of derivative techniques for sections of the trace, in particular, those containing peaks. Black line: Denoising with Haar wavelet followed by numerical differentiation. Red line: Denoising with $D_4$ wavelet followed by numerical differentiation, plus an offset of 50. Green line: Denoising with $D_4$ wavelet followed by vaguelette-wavelet approach (universal threshold), plus an offset value of 100. Blue line: Denoising $D_4$ wavelet followed by vaguelette-wavelet approach (empirical Bayesian threshold), plus an offset value of 150.
Figure 9.8: Close-up comparison of derivative techniques for sections of the trace, in particular, those containing peaks. Black line: Denoising with Haar wavelet followed by numerical differentiation. Red line: Denoising with $D_4$ wavelet followed by numerical differentiation, plus an offset of 50. Green line: Denoising with $D_4$ wavelet followed by vaguelette-wavelet approach (universal threshold), plus an offset value of 100. Blue line: Denoising $D_4$ wavelet followed by vaguelette-wavelet approach (empirical Bayesian threshold), plus an offset value of 150.
with a Haar wavelet-universal threshold combination, and similarly for a $D_4$ wavelet-universal threshold combination; we also compare the vaguelette-wavelet approach and explore thresholding with an inflated universal threshold and empirical Bayesian threshold. Inspired by the discussion in Section 7.4.3, the empirical Bayesian threshold was increased by the factor $\sqrt{1 + 2\alpha}$. Whilst we have not investigated as to whether this is the most appropriate inflation factor for such a thresholding scheme, the results are comparable to those obtained from using the universal threshold.

In Section 9.1 we illustrated that the non-decimated transform successfully eliminated “step-like” phenomena, induced by the Haar wavelet, from the denoised trace. However, such features are considerably more pronounced in the numerically calculated derivative. Further, this derivative displays high levels of noise, substantially more than the alternative approaches investigated, and this noise is particularly evident around the peak regions. Whilst using the $D_4$ wavelet reduces noise levels in the numerical derivative, both vaguelette-wavelet approaches outperform either numerical tactic. Small peaks towards the end of the trace, which appear to be spurious leftovers from denoising, are evident from the derivative function when it is obtained by the numerical methods, yet not from the vaguelette-wavelet obtained function. This demonstrates the importance of a robust derivative measure in relevant peak detection algorithms: approaches based on the numerical derivative would potentially highlight these artifacts as areas of activity, whereas any approach based on the the vaguelette-wavelet derivative would bypass them completely.

### 9.4 Bootstrap confidence intervals for peak areas

In HPLC experimentation, the concentration of a particular compound is inferred from the area under the corresponding peak. The area is generally found by numerically integrating the denoised trace between the start- and endpoint of the peak, i.e. repeatedly applying the
trapezoidal rule. Here, we attempt to attach some level of confidence to such a statistic, an area in which very little work currently exists. We are only aware of one other paper which describes a similar idea; see Bonate (1998).

### 9.4.1 The general approach

The bootstrap is a general simulation approach to statistical inference, based on building a sampling distribution for a statistic from the data at hand. Using the sample data as a population, the success of the method depends on the notion of a bootstrap sample, randomly drawn from this population.

Suppose that we draw a sample, \( y = \{y_1, y_2, \ldots, y_n\} \), from an infinite population, \( p \).

Given that we are interested in a population parameter, \( \theta = t(p) \), we can obtain an estimate of this statistic, denoted \( T = t(y) \).

Bootstrapping techniques can be categorised into two subsets: parametric and non-parametric varieties. The parametric version relies on some prior knowledge of the distribution of population \( p \) whereas the nonparametric bootstrap allows us to estimate the sampling distribution of a statistic empirically without making assumptions about the form of the population \( p \) and without having to derive the sampling distribution explicitly.

To bootstrap, we proceed in the following manner:

- Draw a sample of size \( n \). In the parametric case, this may consist of random deviates from the supposed distribution. Alternatively, with non-parametric bootstrapping, the sample is drawn from among the elements of \( y \), sampling with replacement. It is necessary to sample with replacement, because we would be otherwise simply reproduce the original sample, \( y \). In effect, we are treating the sample \( y \) as an estimate of the population \( p \); that is, each element \( y_i \) of \( y \) is selected for the bootstrap sample with probability \( 1/n \), mimicking the original selection of the
sample $y$ from the population $p$. The resulting bootstrap sample can be denoted by $y^*_1 = \{y^*_{11}, y^*_{12}, \ldots, y^*_{1n}\}$. The star indicates that the sample is a bootstrap replicate and the subscript on the $y^*$ describes the replicate number.

- The resampling procedure is repeated a large number of times, $B$, generating many bootstrap samples. The $b$th bootstrap is denoted $y^*_b = \{y^*_{b1}, y^*_{b2}, \ldots, y^*_{bn}\}$.

- The statistic $T$ can be calculated for each of the $B$ bootstrap samples; that is $T^*_b = t(y^*_b)$. Providing that $B$ is large enough, the distribution of $T^*_b$ around the original estimate $T$ is assumed to be analogous to the sampling distribution of the estimator $T$ around the population parameter $\theta$.

### 9.4.2 Bootstrapping in regression problems

Suppose we regress a set of observations, $\{y_i\}$, on some predictors, $\{x_i\}$, for $i = 1, \ldots, n$. There are two general methods to tackle regression problems in bootstrapping. The first involves considering the predictors to be random variates, potentially changing from sample to sample. This option is not taken any further in this thesis; one should see Efron and Tibshirani (1993) for more details. Alternatively, as applied later in the work here, the predictors can be treated as fixed and the randomness of the responses is inherited from the error component of the model. Attaching a random error term to each fitted value from the original regression, $\hat{y}_i$, produces a fixed-$x$ bootstrap sample $y^*_b = \{y^*_{bi}\}$. The errors may be generated parametrically from a chosen model e.g. $N(0, s^2)$, where $s^2$ is the estimated error variance in the regression, or nonparametrically, by resampling residuals with replacement from the original regression. The bootstrapped values $y^*_b$ can be subsequently regressed on the fixed $x$ vector to obtain bootstrap replications of the regression coefficients.
9.4.3 Bootstrap confidence intervals

There are various approaches to constructing bootstrap confidence intervals, depending on the question posed. Here we give a brief overview of percentile intervals and bias-corrected accelerated (or $BC_a$) percentile intervals and consider their implementation for HPLC data in Section 9.4.4.

Given the empirical quantiles of the statistic $T^*_b$, a $100(1 - \alpha)%$ confidence interval for $\theta$ can be constructed of the form

$$T^*_{[l]} < \theta < T^*_{[u]}, \quad (9.8)$$

where $T^*_{[1]}, T^*_{[2]}, \ldots, T^*_{[B]}$ are the ordered bootstrap replicates of the statistic; $l = (B + 1)\alpha/2$ rounded to the nearest integer; similarly $u = (B + 1)(1 - \alpha/2)$.

$BC_a$ intervals take the form of (9.8) however, in this case, $l = B a_1$ and $u = B a_2$, with

$$a_1 = \Phi\left[z_0 + \frac{z_0 + z^{(\alpha/2)}}{1 - a(z_0 + z^{(\alpha/2)})}\right],$$

$$a_2 = \Phi\left[z_0 + \frac{z_0 + z^{(1 - \alpha/2)}}{1 - a(z_0 + z^{(1 - \alpha/2)})}\right],$$

where $\Phi$ is the standard normal cumulative distribution function and $z^{\alpha/2}$ is the $100 \cdot \alpha/2$th percentile point of a standard normal distribution. Further, we define

$$z_0 = \Phi^{-1}\left[\frac{\#(T^*_b < T)}{B}\right] \quad (9.9)$$

where $\Phi^{-1}(\cdot)$ is the standard normal quantile function and $\#(T^*_b < T)/B$ is the proportion of bootstrap replicates of the statistic less than the original estimate $T$. If the
bootstrap sampling distribution is symmetric and \( T \) is unbiased, then this proportion will be close to half and the “bias-correction” factor \( z_0 \) will be close to 0.

The term \( a \) is calculated to be

\[
a = \frac{\sum_{i=1}^{n} (T_i - T_{(-i)})^3}{6\{\sum_{i=1}^{n} (T_i - T_{(-i)})^2\}^{3/2}},
\]

where \( T_{(-i)} \) is the value of the statistic \( T \) but with the \( i \)th point deleted from the sample and \( T_{(-i)} = \sum_{i=1}^{n} T_{(-i)}/n \). Further discussion regarding the formulation of \( a \) can be found in Efron (1987).

When the correction factors \( a \) and \( z_0 \) are both 0, \( a_1 = \Phi(z^{(\alpha/2)}) = \alpha/2 \) and \( a_2 = 1 - \alpha/2 \); thus the \( BC_a \) interval reduces to the uncorrected percentile interval in equation (9.8).

A confidence interval should comply with certain theoretical ideals to be considered useful. Firstly, it is advantageous, but not essential, for a confidence interval to be transformation respecting. In other words, it is desirable for the endpoints of the interval to transform correctly if the parameter of interest is mapped from \( \theta \) to some function of \( \theta \). Additionally, a central \( 100(1 - \alpha)\% \) confidence interval \((T_{l}^{*}, T_{u}^{*})\) should have probability \( \alpha/2 \) of not covering the true value of \( \theta \) from above or below i.e.

\[
\text{Prob}\{\theta < T_{l}^{*}\} = \frac{\alpha}{2} \quad \text{and} \quad \text{Prob}\{\theta > T_{u}^{*}\} = \frac{\alpha}{2}.
\]

The term symmetric coverage, in its strictest sense, describes the extent to which this particular criterion is fulfilled. However, the term coverage is used more loosely to mean that an \( 100(1 - \alpha)\% \) confidence interval will include the true value of the statistic \( 100(1 - \alpha) \) times from a 100.

Whilst percentile intervals are advantageous in terms of being simple to calculate and respectful to transformations, they often do not perform well in terms of coverage. Bias-
corrected, accelerated intervals tend to be preferable as, in addition to possessing the transform-respecting property, they are more accurate in their coverage.

With regards to its coverage potential, the BC$_a$ confidence interval can be shown to be second-order accurate (Efron and Tibshirani, 1993). In other words, the errors incurred in matching the conditions in (9.11) go to zero at a rate of $1/n$, where $n$ is the sample size. Mathematically, this can be expressed as

$$\text{Prob}\{\theta < T^*_l\} = \frac{\alpha}{2} + \frac{c_l}{n} \quad \text{and} \quad \text{Prob}\{\theta > T^*_u\} = \frac{\alpha}{2} + \frac{c_u}{n}. \quad (9.12)$$

This should be compared to the standard percentile interval, which is only first-order accurate:

$$\text{Prob}\{\theta < T^*_l\} = \frac{\alpha}{2} + \frac{c_l}{\sqrt{n}} \quad \text{and} \quad \text{Prob}\{\theta > T^*_u\} = \frac{\alpha}{2} + \frac{c_u}{\sqrt{n}}. \quad (9.13)$$

The implication of this result means that the errors in matching the conditions in (9.11) are an order of magnitude larger for percentile intervals than for BC$_a$ intervals.

### 9.4.4 Application to HPLC data

In this application, we use a parametric approach similar to that described in Section 9.4.2. Suppose we have a vector of observations, $y$, on the function $f$, but corrupted by noise. Given the function $\hat{f}$ obtained by wavelet smoothing, we identify $p$ peaks and find corresponding estimates of the areas under the peaks. These areas form a set of statistics of interest, $t = \{T_1, T_2, \ldots, T_p\}$.

We can generate $b = 1, \ldots, B$ bootstrapped samples

$$y^*_b = \hat{f} + \epsilon_b,$$  \hfill (9.14)
where $\hat{f}$ is the wavelet denoised version of $y$ and $\epsilon_b$ is a vector of errors, sampled from an $AR(q)$ model with $q = 36$. The $AR(36)$ model was estimated based on the residuals from the original fitted model and its complexity was determined by the AIC criterion. Whilst this choice seems to be a rather complicated model and indeed simpler ones may suffice, ultimately, our only objective is to sample from such a model. Thus we retain the additional complexity for the purposes of generating noise as similar in its behaviour to that of the original data as is possible. Some examples of such bootstrapped data are given in Figure 9.9. Briefly, for each bootstrapped replication $y_b^*$, we smooth the trace and estimate bootstrapped replicates of the peak areas. Following the completion of the $B$ bootstrapped replications, we calculate 95% percentile and BC$_\alpha$ confidence intervals for each of the $p$ peaks.

### Fixed peak locations

From the original sample, we identified 8 peaks and determined the start- and endpoints to be as described in Table 9.1. We initially proceed with the bootstrapping by assuming the start- and endpoints of the peaks to be fixed in these locations. Furthermore, to circumvent the additional complexity of baseline correction, we adopt the baseline corrected version of the trace as the original smoothed $\hat{f}$. Thus to reiterate, we generate $B = 2000$ bootstrapped replicates, $y_b^* = \hat{f} + \epsilon_b$, with $\epsilon \sim AR(36)$. For each $y_b^*$, we decompose the trace into the wavelet domain, threshold the coefficients and reconstruct the denoised trace, before calculating bootstrapped replicates of the 8 peak areas using the start- and endpoints in Table 9.1. Having completed all 2000 replications, percentile and BC$_\alpha$ confidence intervals can be calculated for each of the peaks. The resulting confidence intervals are given in Table 9.2.
Chapter 9. Novel approaches to analysing HPLC data

Figure 9.9: Some examples of the bootstrapped traces.
Table 9.1: Times of startpoints, endpoints and peak apexes for HPLC trace.

<table>
<thead>
<tr>
<th>Peak number</th>
<th>Time of startpoint (minutes)</th>
<th>Time of peak apex (minutes)</th>
<th>Time of endpoint (minutes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7.120</td>
<td>7.473</td>
<td>7.960</td>
</tr>
<tr>
<td>2</td>
<td>9.760</td>
<td>10.267</td>
<td>11.333</td>
</tr>
<tr>
<td>3</td>
<td>14.033</td>
<td>14.747</td>
<td>15.800</td>
</tr>
<tr>
<td>4</td>
<td>15.853</td>
<td>16.400</td>
<td>17.453</td>
</tr>
<tr>
<td>5</td>
<td>17.453</td>
<td>18.000</td>
<td>19.353</td>
</tr>
<tr>
<td>6</td>
<td>21.993</td>
<td>22.700</td>
<td>23.447</td>
</tr>
<tr>
<td>7</td>
<td>27.660</td>
<td>27.960</td>
<td>28.673</td>
</tr>
<tr>
<td>8</td>
<td>29.587</td>
<td>30.100</td>
<td>32.093</td>
</tr>
</tbody>
</table>

Table 9.2: 95% percentile and BC$_n$ confidence intervals for the areas of the eight identified peaks (using fixed peak locations).

<table>
<thead>
<tr>
<th>Original estimate of area</th>
<th>Percentile interval</th>
<th>BC$_n$ interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lower bound</td>
<td>Upper bound</td>
</tr>
<tr>
<td>1.324</td>
<td>1.296</td>
<td>1.354</td>
</tr>
<tr>
<td>3.857</td>
<td>3.817</td>
<td>3.899</td>
</tr>
<tr>
<td>2.372</td>
<td>2.329</td>
<td>2.414</td>
</tr>
<tr>
<td>3.361</td>
<td>3.314</td>
<td>3.401</td>
</tr>
<tr>
<td>1.169</td>
<td>1.129</td>
<td>1.205</td>
</tr>
<tr>
<td>1.064</td>
<td>1.031</td>
<td>1.098</td>
</tr>
<tr>
<td>3.812</td>
<td>3.754</td>
<td>3.857</td>
</tr>
</tbody>
</table>
Variable peak locations

Here we extend the approach of the previous section to allow for variability in the peak locations. A key difference between the two approaches is that here we revert to using the denoised trace, prior to any baseline correction, as $\hat{f}$. For each bootstrapped replicate, $y_b^*$, we denoise the trace and estimate the baseline by the gradient changepoints approach. This was described in Section 9.2, but now we incorporate the derivative measure obtained via the vaguelette-wavelet approach in Section 9.3. Subsequently, we can subtract the resulting baseline estimate and identify the peaks from this trace. For each peak, we record the start-, endpoint and area under the peak. Having completed $B$ bootstraps, we have sufficient information to yield confidence intervals for the start- and endpoints of the eight identified peaks, in addition to those for the peak areas. This procedure is much more complex and is evident in the computational time required. Consequently we limit the number of bootstrap replications to $B = 1000$. The confidence intervals generated are shown in Tables 9.3, 9.4 and 9.5.

We experimented with obtaining confidence intervals for the location of the peak apex, given that this defines the retention time of a particular compound. However, the variability in position proved minimal and the peak apex mostly only shifted up or down the time axis by one clock tick. These results are not particularly interesting yet, considering the construction of the bootstrapped data, are to be expected; we do not pursue these results any further.

For a single trace, the confidence intervals obtained for the start- and endpoints of peaks are of limited interest. The retention time for a particular compound is defined by the time point at which the apex of the corresponding peak occurs and fluctuations in the start- and endpoint of the peak have little or no impact on the position of the apex. However, given that we have replicate traces and thus can obtain multiple confidence intervals for the location of the same peak, one may consider this a natural source from which to develop algorithms which correct for misalignments. However we leave this for future work.
Table 9.3: 95% percentile and BC$_a$ confidence intervals for the startpoints (in minutes) of the eight identified peaks (using variable peak locations).

<table>
<thead>
<tr>
<th>Original estimate of peak startpoint</th>
<th>Percentile interval</th>
<th>BC$_a$ interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lower bound</td>
<td>Upper bound</td>
</tr>
<tr>
<td>7.120</td>
<td>6.993</td>
<td>7.213</td>
</tr>
<tr>
<td>15.853</td>
<td>15.827</td>
<td>15.887</td>
</tr>
<tr>
<td>17.453</td>
<td>17.407</td>
<td>17.467</td>
</tr>
<tr>
<td>27.660</td>
<td>27.573</td>
<td>27.687</td>
</tr>
<tr>
<td>29.587</td>
<td>29.553</td>
<td>29.607</td>
</tr>
</tbody>
</table>

The confidence intervals for the peak areas obtained from a single trace are useful for describing the extent of variability realised in the experiment. Furthermore, we have seen that spurious small peaks originating from noise can be retained by inefficiencies in the denoising scheme and/or peak detection procedure. The confidence interval for this type of peak should theoretically contain zero: this can be used as a basis on which to declare such peaks insignificant. Consequently, such peaks could be eliminated from further classification/discriminant analysis. Now consider a situation, similar to a 2-channel microarray experiment, in which we wish to compare two biological conditions. Again these might be diseased and non-diseased, or treated versus control for example, and the aim is to determine whether the compound represented by peak $p$ differs significantly in terms of its concentration, between the two types of sample. In the absence of replication, the confidence intervals, or rather the overlap of the confidence intervals, could be used as a substitute for more formal hypothesis testing, which may have important implications.
Table 9.4: 95% percentile and BC$_a$ confidence intervals for the endpoints (in minutes) of the eight identified peaks (using variable peak locations).

<table>
<thead>
<tr>
<th>Original estimate of peak endpoint</th>
<th>Percentile interval</th>
<th>BC$_a$ interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lower bound</td>
<td>Upper bound</td>
</tr>
<tr>
<td>7.960</td>
<td>7.960</td>
<td>8.300</td>
</tr>
<tr>
<td>15.800</td>
<td>15.460</td>
<td>15.833</td>
</tr>
<tr>
<td>17.453</td>
<td>17.407</td>
<td>17.467</td>
</tr>
<tr>
<td>32.093</td>
<td>31.600</td>
<td>32.200</td>
</tr>
</tbody>
</table>

for classification algorithms. A complete separation of the confidence intervals for peak $p$ would imply that the concentration of the compound corresponding to peak $p$ was significantly different in the two types of sample. In the converse however, if the confidence intervals were to overlap, this does not strictly imply that there is no significant difference between the samples. It can be shown (Schenker and Gentleman, 2001) that using the confidence intervals in such a fashion will be much more conservative than the chosen significance level, $\alpha$, making it difficult to detect significant differences even if present. Thus in order to maintain a significance level of $\alpha$ in a testing situation, confidence intervals narrower than the usual 100(1 − $\alpha$)$\%$ will be required. Formal adjustments to the confidence interval widths have been proposed to validate such a procedure (Payton et al. (2003) provided an example of this type of work), however this is not investigated further in this thesis.
Table 9.5: 95% percentile and BCₛ confidence intervals for the areas of the eight identified peaks (using variable peak locations).

<table>
<thead>
<tr>
<th>Original estimate of peak area</th>
<th>Percentile interval</th>
<th>BC₄ interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lower bound</td>
<td>Upper bound</td>
</tr>
<tr>
<td>1.324</td>
<td>1.279</td>
<td>1.424</td>
</tr>
<tr>
<td>3.857</td>
<td>3.784</td>
<td>4.004</td>
</tr>
<tr>
<td>2.372</td>
<td>2.210</td>
<td>2.497</td>
</tr>
<tr>
<td>3.361</td>
<td>3.218</td>
<td>3.519</td>
</tr>
<tr>
<td>1.169</td>
<td>1.102</td>
<td>1.288</td>
</tr>
<tr>
<td>1.064</td>
<td>0.922</td>
<td>1.146</td>
</tr>
</tbody>
</table>

9.4.5 Simulations

A method which has the potential for generating a $100(1 - \alpha)\%$ confidence interval for the area under a peak should ideally fulfill various criteria: the interval obtained should be precise and in terms of physically possible values. Moreover, the interval should contain the true value of the statistic of interest with probability coverage near the nominal $100(1 - \alpha)\%$ level. We investigated how well our procedure performs against these specifications, through the use of some simulation studies.

Firstly, we assume that the real data are adequately modelled by a simple linear combination of signal plus noise. We adopted the bumps function of Donoho and Johnstone (1994) as the true signal, simply because the true peak areas, start- and endpoints are easily obtained from such. However, in order that it might resemble one of the real HPLC traces and be consistent in terms of its root signal to noise ratio (RSNR),
the bumps function was rescaled; this was achieved by ensuring that its standard deviation was equal to the standard deviation of the denoised HPLC trace. Furthermore, it was reduced in length so that only seven “bumps” or peaks remained: the reasons behind such a motion were to mimic a similar number of peaks to that seen in the real trace and minimise computational intensity. The modified bumps function is illustrated in Figure 9.10. Motivated by observations made from the real HPLC dataset, the noise was assumed to come from an AR process with 36 parameters. One hundred realisations of the model were simulated by combining the true signal i.e. the bumps function, with a vector of randomly generated noise. To each of the 100 realisations of data, the bootstrapping procedure was applied (we considered both the fixed and variable peak location approaches) with \(B = 1000\). The resulting 95% percentile and \(BC_\alpha\) confidence intervals were recorded. Considering the percentile and \(BC_\alpha\) cases separately, of the 100 simulated confidence intervals, it would be desirable to have the true value of the statistic contained in close to 95 of them.

The RSNR of the real HPLC trace was estimated to be close to 24; we also investigated the efficiency of the bootstrapping procedure in the presence of increased noise by rescaling the bumps function so that RSNR was equal to 6.

The results of the simulation studies are given in Tables 9.6 to 9.9.

Table 9.6: Results from simulation study designed to test the efficiency of the bootstrapping approach with fixed peak locations. The nominal coverage is 95%.

<table>
<thead>
<tr>
<th>Peak number</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentile interval coverage (%)</td>
<td>62</td>
<td>66</td>
<td>63</td>
<td>74</td>
<td>64</td>
<td>66</td>
<td>58</td>
<td>64.7</td>
</tr>
<tr>
<td>(BC_\alpha) interval coverage (%)</td>
<td>78</td>
<td>84</td>
<td>78</td>
<td>81</td>
<td>73</td>
<td>77</td>
<td>74</td>
<td>77.9</td>
</tr>
</tbody>
</table>

The fixed peak approach is rather crude in terms of obtaining bootstrapped estimates of the peak areas. The greatest problem occurs when the fixed peak start- and endpoints
Figure 9.10: The modified bumps function.

Table 9.7: Results from simulation study designed to test the efficiency of the bootstrapping approach with variable peak locations. The nominal coverage is 95%.

<table>
<thead>
<tr>
<th></th>
<th>Peak number</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peak areas</td>
<td>Perc. int. coverage (%)</td>
<td>75</td>
<td>75</td>
<td>69</td>
<td>67</td>
<td>78</td>
<td>77</td>
<td>79</td>
<td>74.3</td>
</tr>
<tr>
<td></td>
<td>BCₐ int. coverage (%)</td>
<td>90</td>
<td>86</td>
<td>85</td>
<td>80</td>
<td>92</td>
<td>87</td>
<td>89</td>
<td>87.0</td>
</tr>
<tr>
<td>Peak startpoints</td>
<td>Perc. int. coverage (%)</td>
<td>91</td>
<td>71</td>
<td>86</td>
<td>92</td>
<td>77</td>
<td>72</td>
<td>79</td>
<td>81.1</td>
</tr>
<tr>
<td></td>
<td>BCₐ int. coverage (%)</td>
<td>98</td>
<td>95</td>
<td>97</td>
<td>93</td>
<td>89</td>
<td>92</td>
<td>85</td>
<td>92.7</td>
</tr>
<tr>
<td>Peak endpoints</td>
<td>Perc. int. coverage (%)</td>
<td>41</td>
<td>18</td>
<td>22</td>
<td>45</td>
<td>36</td>
<td>37</td>
<td>42</td>
<td>34.4</td>
</tr>
<tr>
<td></td>
<td>BCₐ int. coverage (%)</td>
<td>69</td>
<td>62</td>
<td>52</td>
<td>75</td>
<td>78</td>
<td>67</td>
<td>71</td>
<td>67.7</td>
</tr>
</tbody>
</table>
Table 9.8: Results from simulation study designed to test the efficiency of the bootstrapping approach with fixed peak locations in the presence of increased noise levels. The nominal coverage is 95%.

<table>
<thead>
<tr>
<th>Peak number</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentile interval coverage (%)</td>
<td>61</td>
<td>52</td>
<td>61</td>
<td>66</td>
<td>58</td>
<td>68</td>
<td>56</td>
<td>60.3</td>
</tr>
<tr>
<td>BC$_a$ interval coverage (%)</td>
<td>77</td>
<td>76</td>
<td>78</td>
<td>78</td>
<td>74</td>
<td>76</td>
<td>70</td>
<td>75.6</td>
</tr>
</tbody>
</table>

Table 9.9: Results from simulation study designed to test the efficiency of the bootstrapping approach with variable peak locations in the presence of increased noise levels. The nominal coverage is 95%.

<table>
<thead>
<tr>
<th>Peak number</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peak areas</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perc. int. coverage (%)</td>
<td>70</td>
<td>85</td>
<td>69</td>
<td>75</td>
<td>77</td>
<td>83</td>
<td>72</td>
<td>75.9</td>
</tr>
<tr>
<td>BC$_a$ int. coverage (%)</td>
<td>81</td>
<td>92</td>
<td>82</td>
<td>93</td>
<td>93</td>
<td>83</td>
<td>86</td>
<td>86.9</td>
</tr>
<tr>
<td>Peak startpoints</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perc. int. coverage (%)</td>
<td>27</td>
<td>31</td>
<td>34</td>
<td>21</td>
<td>17</td>
<td>36</td>
<td>25</td>
<td>27.3</td>
</tr>
<tr>
<td>BC$_a$ int. coverage (%)</td>
<td>59</td>
<td>67</td>
<td>65</td>
<td>58</td>
<td>62</td>
<td>76</td>
<td>50</td>
<td>62.4</td>
</tr>
</tbody>
</table>
form a narrower interval than those of the bootstrapped trace; the tails of the peak will be cut from the numerical integral, leading to an underestimate of the area. Conversely, if the fixed peak locations form too wide an interval, it is possible that the peak area will be overestimated. However, as the peaks rise from a zero baseline, any effect incurred from such a problem is expected to be relatively minimal. From Tables 9.2 and 9.5, it is clear that using fixed peak locations yields narrower confidence intervals than its variable peak locations counterpart, however from Tables 9.6 and 9.7, we infer that this is at the expense of the level of coverage obtained.

It would have been optimistic to expect perfect results from the simulations: the complexities involved in the process prior to estimating the peak statistics and the imprecision in the wavelet denoising all give rise to additional uncertainty. However, in the face of this, Table 9.7 implies good levels of coverage are attained by the confidence intervals for the peak startpoints, and relatively so for the peak areas, when generated by the variable peak locations bootstrapping procedure. The results for the peak endpoints are rather disappointing, but it is difficult to distinguish whether it is the bootstrapping procedure or peak detection algorithm which is flawed. Given that coverage levels obtained by bootstrapping for the peaks startpoints are good, one might infer that it is the algorithm for detecting the endpoints that is lacking in robustness. Why the method works so effectively for the startpoints but not the endpoints is not immediately clear. Furthermore, it is important to note that reliable estimation of the peak areas is intimately related to successful identification of the peak start- and endpoints. An improvement to the peak detection algorithm would hopefully yield improved endpoint coverage, as well as improved area estimates, which may in turn improve the corresponding levels of coverage for the bootstrapping of peak areas. The simulations show the definite potential of bootstrapping within such a context, and considering the modularity of spectral analysis, could easily be combined with existing or future peak detection methods. As the start- and endpoints of the peaks have essentially come from a discretely sampled population, a smoothed version of the bootstrap may provide additional benefits. However, given the
frequent sampling rate (150 times per minute) of the real HPLC data, we anticipate that,
in practice, this would actually add little to the procedure.

In all cases, the $BC_\alpha$ intervals achieved much better coverage levels than the percentile
intervals. It was observed that often when the $BC_\alpha$ interval did not contain the true value
of the statistic, it was due to the interval being much too short. The lower and upper
limits, $l$ and $u$, were calculated to be both either very small or very large, which generated
intervals that were barely intervals.

The results in Tables 9.8 and 9.9 strongly imply that the bootstrapping procedure is less
efficient when the root signal to noise ratio is decreased (RSNR of 6 rather than 24). However,
we suspect that in cases of heightened noise, the denoising process, as it is, is
less than sufficient and the noise is not removed as well as it potentially could be with
alternative choices of wavelet filter and thresholding technique. An example of such a
denoised trace is given in Figure 9.11. As a consequence of the poor denoising, in most
cases, the difficulty lay with detecting the peaks: in the presence of heightened noise, the
method, with parameters as described for the HPLC data, identified more peaks than the
signal contained in truth. We circumvented the problem for the purpose of simulation
by constraining the method to detect the seven most significant peaks. If these were
“real” data, in addition to choosing a more efficient denoising scheme, the user-defined
parameters could be tuned to the specific nature of the data. However, it is not simple
to select optimal parameter values. Generally, increasing the value of $k$ in equation (9.3)
will yield less identified peaks. In doing so however, the estimated start- or endpoint of
the peak will move further up the peak; if substantially so, the outcome may be less
than desirable. Increasing the window size should also reduce the number of peaks,
however one should be careful to choose a window size that will eliminate rogue peaks
that originated from the noise, yet have not been removed fully by smoothing, without
eradicating the truly interesting peaks. With regards to the simulation study, the peaks
in the bumps function were generally not as wide as those observed in the HPLC data.
Consequently, combining them with noise sampled from a model observed from the real HPLC data, generated simulated data where the interesting signal peaks were of a similar width to some of the peaks in the noise. Consequently, choosing a window size that would distinguish between the two types of peak was difficult. The paramount way to select the user-defined parameters is to consider each data realisation individually and visual inspect results from a variety of parameters. However, resorting to this does defy the aim of an creating a more automatic approach. Once chosen, the selected parameter values appear to hold sufficiently well for subsequently generated bootstrapped samples.

Figure 9.11: Signal corrupted by increased noise after denoising procedure.
Chapter 10

Conclusions and further work

10.1 Conclusions

In Part II of this thesis, we considered the novel application of statistical methods for the purpose of analysing HPLC data.

Firstly, we explored various wavelet smoothing and baseline estimation methods and developed a simple peak detection algorithm. It is important to note that whilst our methods work for our low noise data and are useful specific to the context described here, these pre-processing components were not the main focus of our work and we acknowledge the existence of more sophisticated, yet not necessarily more efficient, approaches in these areas. Whilst for microarray analysis, pre-processing methods from other authors are easily accessible and applicable through common software such as Bioconductor in R, this is not the case as much for spectral type data. Consequently, we chose to develop our own simple algorithms. We identified autocorrelation in the noise and, after discussion with the scientists concerned, suggest that this is due to some inefficiency of the mobile mixing apparatus.

Derivative techniques have been used for a variety of purposes within the spectral
Chapter 10. Conclusions and further work

analysis domain. Here we described the vaguelette-wavelet method of Abramovich and Silverman (1998) and demonstrated its usefulness in such a context. We combined the vaguelette-wavelet plug-in estimator with the non-decimated transformation of Coifman and Donoho (1995) and an inflated universal threshold. The method showed considerable improvement over simple numerical differentiation and its incorporation increased the efficiency of future downstream analysis, such as peak detection.

We used a variation of parametric bootstrapping to obtain percentile and bias-corrected, accelerated confidence intervals for peak areas. Two approaches were explored: firstly we assumed the start- and endpoints of the peaks to be fixed in the positions determined from the original trace; we further extended the method to allow for variable peak start-and endpoints, which gave rise to confidence intervals for the peak locations too. For a single trace, as considered here, being able to attach a level of confidence to a particular peak area is probably the result that is most informative to scientists. Having confidence intervals for the start- and endpoints of the peaks is of limited use as generally the peak apex is taken as the definitive retention time. However, given that confidence intervals for peak start- and endpoints could be obtained for multiple traces, it seems intuitive that this information is a natural foundation from which to develop alignment algorithms. However, this is left for future work. Being able to calculate a confidence interval for a particular peak area has more than one potential use. Without modification, they are effective at identifying statistically insignificant peaks in a single trace, which can be eliminated from further analysis. In a two-sample scenario, comparing diseased and non-diseased samples say, checking for any overlap between the confidence intervals can yield conclusions regarding hypotheses about differences between the population parameters for the two samples. However, the intricacies of such is a much discussed issue of statistics and its usefulness within this context would require further investigation.

With regard to the analysis of mass spectrometry data, it is anticipated that the methods presented here would be applicable with minimal modifications, if any at all. The
additional complexities involved in analysing mass spectra make it likely that more sophisticated baseline estimation, peak detection and deconvolution methods would be necessary than used here. However, where appropriate, the vaguelette-wavelet derivative technique could be incorporated in its current form into any existing approaches with the potential to offer substantial improvement. Given successful peak detection, it should be possible to apply the bootstrapping procedure to mass spectra as it exists here, providing a more robust measurement than simple peak height.

### 10.2 Directions for further work

Essentially, the data we have is “count” data, which could be considered to be from a Poisson process where each observation is a Poisson random variable, with parameter $\lambda$. The variance of the Poisson random variable changes with the mean, however in the work here and specifically for the purposes of thresholding, we have assumed that the noise component follows a Gaussian distribution where the variance is constant. Many researchers have considered “pre-processing” transformations for Poisson data, which aim to “Gaussian-ise” the errors, making them suitable for standard wavelet denoising techniques. One of the first transformations to appear was that of Anscombe (1948) who suggested the mapping $Y \rightarrow 2\sqrt{Y + 3/8}$. Since the emergence of wavelets, there have been considerable advances in this area, many of which combine such transformations with wavelet decomposition methods. Examples of this type of work include Kolaczyk (1997), Charles and Rasson (2003) and the Haar-Fisz transformation of Fryzlewicz and Nason (2004). In our case, given that the true scale of the data runs into the thousands, we anticipated that the Gaussian approximation was reasonable. However, it would be interesting to explore how the results differed, if at all, when the data were treated as Poisson.
Efromovich et al. (2004) discussed signal denoising and the recovery of its derivative using multiwavelets. This area of wavelets is substantial and warrants separate discussion outside this thesis. However, to briefly summarise, Efromovich et al. (2004) demonstrated further reduced noise levels in the recovered derivative obtained using the multiwavelet method, in comparison to the vaguelette-wavelet approach. The application of such a multiwavelet technique to proteomic data and exploration of potential improvements obtained via this approach are left for future investigation.

### 10.3 Contributions

Part II of this thesis concerns aspects of the analysis of HPLC data. My original contributions are briefly summarised below:

- The derivation and implementation of three simple baseline estimation algorithms and the acknowledgment of their respective limitations. The most successful of these, the gradient changepoints approach, essentially formed the basis of a peak detection algorithm also and was incorporated into further downstream work.

- Abramovich and Silverman (1998) proposed the vaguelette-wavelet approach and gave the estimation of the derivative of a function as an illustration of their method. My contribution is in combining this idea with the non-decimated wavelet transform, developed, amongst others, by Coifman and Donoho (1995). Subsequently, the resulting approach was used to estimate the derivative of an HPLC obtained trace and offered substantial improvements over simple numerical differentiation.

- The application of bootstrapping within an HPLC domain, based on an idea originally suggested by Dr Stuart Barber. The execution of this suggestion is my
own, as is the idea of extending the approach to incorporate variable peak locations, yielding confidence intervals for both the peak area and position.
Appendices

A  Algebra for Laplace-Gaussian model with a variance parameter

A.1 Convolution of normal distribution with a variance parameter and the Laplace distribution

Suppose that we have an observation $z$ which is a linear additive combination as written below:

$$Z = \mu + \epsilon,$$  \hspace{1cm} (A.1)

where $\mu$ comes from a Laplace distribution with scaling parameter $a$ and $\epsilon$ is normally distributed with variance parameter $\sigma$.

$$\gamma(\mu) = \frac{a}{2} e^{-a|\mu|} \quad \text{and} \quad \tau(\epsilon) = \frac{1}{\sigma \sqrt{2\pi}} e^{\frac{\epsilon^2}{2\sigma^2}}$$

Define $g(z) = \gamma(\mu) * \tau(\epsilon)$ to be the convolution of $\mu$ and $\epsilon$.

Rewriting (A.1) as $\epsilon = z - \mu$, we have
\[ g(z) = \int_{\mathbb{R}} \gamma(\mu) \tau(z - \mu) d\mu \\
= \int_{\mathbb{R}} \gamma(\mu) \tau(\mu - z) d\mu \quad \text{as } f_z \text{ is even.} \\
= \int_{-\infty}^{\infty} \frac{a}{2} e^{-a|\mu|} \cdot \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(\mu - z)^2}{2\sigma^2}} d\mu. \quad (A.2) \]

Splitting (A.2) in two parts we get

\[ g(z) = \frac{a}{2\sigma \sqrt{2\pi}} \int_{-\infty}^{0} \exp \left\{ a\mu - \frac{1}{2\sigma^2} (\mu - z)^2 \right\} d\mu + \frac{a}{2\sigma \sqrt{2\pi}} \int_{0}^{\infty} \exp \left\{ -a\mu - \frac{1}{2\sigma^2} (\mu - z)^2 \right\} d\mu. \quad (A.3) \]

Consider the situation when \( z < 0 \). We proceed from the first term of (A.3) as follows:

\[
\frac{a}{2\sigma \sqrt{2\pi}} \int_{-\infty}^{0} \ldots = \frac{a}{2\sigma \sqrt{2\pi}} \int_{-\infty}^{0} \exp \left\{ ay_1 - \frac{1}{2\sigma^2} (\mu - z)(\mu - z) \right\} d\mu \\
= \frac{a}{2\sigma \sqrt{2\pi}} \int_{-\infty}^{0} \exp \left\{ ay_1 - \frac{1}{2\sigma^2} (\mu^2 - 2zM + z^2) \right\} d\mu \\
= \frac{a}{2\sigma \sqrt{2\pi}} \int_{-\infty}^{0} \exp \left\{ -\frac{1}{2\sigma^2} (\mu^2 - 2\mu z + z^2) \right\} d\mu \\
= \frac{a}{2\sigma \sqrt{2\pi}} \int_{-\infty}^{0} \exp \left\{ \frac{1}{2\sigma^2} (\mu^2 - 2\mu (z + a\sigma^2) + z^2) \right\} d\mu \\
= \frac{a}{2\sigma \sqrt{2\pi}} \int_{-\infty}^{0} \exp \left\{ \frac{1}{2\sigma^2} (\mu^2 - (z + a\sigma^2) + z^2) \right\} d\mu \\
= \frac{a}{2\sigma \sqrt{2\pi}} \int_{-\infty}^{0} \exp \left\{ \frac{1}{2\sigma^2} (\mu^2 - 2az\sigma^2 - a^2 \sigma^4) \right\} d\mu \\
= \frac{a}{2\sigma \sqrt{2\pi}} \int_{-\infty}^{0} \exp \left\{ az + \frac{(a\sigma)^2}{2} \right\} \exp \left\{ -\frac{1}{2\sigma^2} [\mu - (z + a\sigma^2)]^2 \right\} d\mu \\
= \frac{a}{2\sigma \sqrt{2\pi}} \int_{-\infty}^{0} \exp \left\{ az + \frac{(a\sigma)^2}{2} \right\} \exp \left\{ -\frac{1}{2\sigma^2} \left[ \frac{\mu - (z + a\sigma^2)}{\sigma} \right]^2 \right\} d\mu. \quad (A.4) \]

Let \( v = \frac{\mu - (z + a\sigma^2)}{\sigma} \). When \( \mu = 0 \), \( v = -(z/\sigma + a\sigma) \). Substituting in the new variable \( v \), equation (A.4) becomes
\[
\frac{a}{2} \exp \left\{ az + \frac{(a\sigma)^2}{2} \right\} \int_{-\infty}^{-\left(\frac{z}{\sigma} + a\sigma\right)} - \frac{1}{2} e^{2v^2} dv = \frac{a}{2\sigma} \exp \left\{ az + \frac{(a\sigma)^2}{2} \right\} \int_{-\infty}^{-\left(\frac{z}{\sigma} + a\sigma\right)} \phi(v) dv \\
= \frac{a}{2} \exp \left\{ az + \frac{(a\sigma)^2}{2} \right\} [1 - \Phi(\frac{z}{\sigma} + a\sigma)].
\]

Similarly for \( z \geq 0 \) we have

\[
\frac{a}{2\sigma\sqrt{2\pi}} \int_{0}^{\infty} \cdots = \frac{a}{2\sigma\sqrt{2\pi}} \int_{0}^{\infty} \exp \left\{ -a\mu - \frac{1}{2\sigma^2} (\mu - z)(\mu - z) \right\} d\mu \\
= \frac{a}{2\sigma\sqrt{2\pi}} \int_{0}^{\infty} \exp \left\{ -a\mu - \frac{1}{2\sigma^2} (\mu^2 - 2z\mu + z^2) \right\} d\mu \\
= \frac{a}{2\sigma\sqrt{2\pi}} \int_{0}^{\infty} \exp \left\{ -\frac{1}{2\sigma^2} (\mu^2 + 2a\mu\sigma^2 - 2z\mu + z^2) \right\} d\mu \\
= \frac{a}{2\sigma\sqrt{2\pi}} \int_{0}^{\infty} \exp \left\{ -\frac{1}{2\sigma^2} (\mu^2 - 2\mu(z - a\sigma^2) + z^2) \right\} d\mu \\
= \frac{a}{2\sigma\sqrt{2\pi}} \int_{0}^{\infty} \exp \left\{ -\frac{1}{2\sigma^2} \left[ (\mu - (z - a\sigma^2))^2 - (z - a\sigma^2)^2 + z^2 \right] \right\} d\mu \\
= \frac{a}{2\sigma\sqrt{2\pi}} \int_{0}^{\infty} \exp \left\{ -\frac{1}{2\sigma^2} \left[ (\mu - (z - a\sigma^2))^2 + 2az\sigma^2 - a^2\sigma^4 \right] \right\} d\mu \\
= \frac{a}{2\sigma\sqrt{2\pi}} \int_{0}^{\infty} \exp \left\{ -az + \frac{(a\sigma)^2}{2} \right\} \exp \left\{ -\frac{1}{2\sigma^2} \left[ \mu - (z - a\sigma^2) \right]^2 \right\} d\mu \\
= \frac{a}{2\sigma\sqrt{2\pi}} \int_{0}^{\infty} \exp \left\{ -az + \frac{(a\sigma)^2}{2} \right\} \exp \left\{ -\frac{1}{2} \left[ \mu - (z - a\sigma^2) \right]^2 \right\} d\mu.
\]

Let \( v = \frac{\mu - (z - a\sigma^2)}{\sigma} \). Therefore, since \( dv d\mu = 1/\sigma, d\mu = \sigma dv \).

When \( \mu = 0, v = -(z/\sigma - a\sigma) \). Substituting in the new variable \( v \), equation (A.6) becomes
Combining equations (A.5) and (A.7), the convolution can be expressed as

$$g(z) = \frac{a}{2} \exp \left\{ -az + \frac{(a\sigma)^2}{2} \right\} \{ e^{-az\Phi(z/\sigma - a\sigma)} + e^{az}(1 - \Phi(z/\sigma + a\sigma)) \}.$$  

(A.8)

Consider the middle term of equation (A.8). This becomes
\[
\frac{\frac{a^2}{2} \Phi(z/\sigma + a\sigma)}{\frac{1}{\sigma \sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left( \frac{z}{\sigma} \right)^2 - az - \frac{1}{2}(a\sigma)^2 \right\}} = \frac{\frac{a^2}{2} \Phi(z/\sigma + a\sigma)}{\frac{1}{\sigma \sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left( \frac{z}{\sigma} \right)^2 + 2az + (a\sigma)^2 \right\}} \\
= \frac{\frac{a^2}{2} \Phi(z/\sigma + a\sigma)}{\frac{1}{\sigma \sqrt{2\pi}} \exp -\frac{1}{2} \left( \frac{z}{\sigma} + a\sigma \right)^2} \\
= \frac{a\sigma \Phi(z/\sigma + a\sigma)}{2\phi(z/\sigma + a\sigma)}
\]

Having worked through the first term of equation (A.8) similarly, we arrive at the expression below for \(\beta(z)\):

\[
\beta(z) = \frac{a\sigma}{2} \left\{ \frac{\Phi(z/\sigma - a\sigma)}{\phi(z/\sigma - a\sigma)} + \frac{\Phi(z/\sigma + a\sigma)}{\phi(z/\sigma + a\sigma)} \right\} - 1.
\]

### A.3 Posterior probability that the parameter is non-zero

The posterior probability \(\omega_{\text{post}}(z) = P(\mu \neq 0|Z = z)\) will satisfy

\[
\omega_{\text{post}}(z) = \frac{\omega g(z)}{\omega g(z) + \frac{(1-\omega)}{\sigma^2} \phi(z/\sigma)}.
\]

Define

\[
\beta(z) = \frac{g(z)}{\sigma \phi(z/\sigma)} - 1,
\]

therefore \(\omega_{\text{post}}(z)\) can be written as

\[
\omega_{\text{post}}(z) = \frac{(1 + \beta(z))}{\omega - 1 + \beta(z)}.
\]
Appendices

A.4 Derivation of the posterior probability distribution of $\mu$ given $Z$

By definition,

$$f(\mu|z) \propto f(z|\mu)\pi(\mu, \mu \neq 0) = \frac{f(z|\mu)\pi(\mu)}{\int f(z|\mu)\pi(\mu)\,d\mu}.$$ 

In this case, we have the following:

$$\pi(\mu) = (1 - \omega)\delta_0 + w \cdot \frac{a}{2} \exp(-a|\mu|)$$

$$\pi(\mu) = \begin{cases} 
  w \cdot \frac{a}{2} \exp(-a\mu) & \text{for } \mu > 0 \\
  w \cdot \frac{a}{2} \exp(a\mu) & \text{for } \mu < 0
\end{cases}$$

$$f_z|\mu(z|\mu) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{(z - \mu)^2}{2\sigma^2} \right\}.$$ 

Thus, when $\mu > 0$,

$$f(z|\mu)\pi(\mu) = \frac{aw}{2} \exp(-a\mu) \cdot \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{(z - \mu)^2}{2\sigma^2} \right\}.$$ 

Therefore,

$$\int_0^\infty f(z|\mu)\pi(\mu) = \frac{aw}{2\sigma\sqrt{2\pi}} \int_0^\infty \exp\left\{ -\frac{1}{2\sigma^2}(z^2 - 2\mu + \mu^2) - a\mu \right\} \,d\mu$$

$$= \frac{aw}{2\sigma\sqrt{2\pi}} \int_0^\infty \exp\left\{ -\frac{1}{2\sigma^2}(z^2 - 2\mu + \mu^2 + 2a\mu\sigma^2) \right\} \,d\mu$$

$$= \frac{aw}{2\sigma\sqrt{2\pi}} \int_0^\infty \exp\left\{ -\frac{1}{2\sigma^2}(\mu^2 - 2\mu(z - a\sigma^2) + z^2) \right\} \,d\mu$$

$$= \frac{aw}{2\sigma\sqrt{2\pi}} \int_0^\infty \exp\left\{ -\frac{1}{2\sigma^2}(\mu^2 - 2\mu(z - a\sigma^2) + z^2) + 2az\sigma^2 - (a\sigma^2)^2 \right\} \,d\mu$$

$$= \frac{aw}{2\sigma\sqrt{2\pi}} \int_0^\infty \exp\left\{ -\frac{1}{2}\left( \frac{\mu - (z - a\sigma^2)}{\sigma} \right)^2 \right\} \exp\left\{ -az + \frac{(a\sigma)^2}{2} \right\} \,d\mu.$$  

(A.9)
Now let \( v = \frac{\mu - (z - a\sigma^2)}{\sigma} \). Then \( d\mu = \sigma dv \).

When \( \mu = 0 \), \( v = -(z/\sigma - a\sigma) \). Therefore integral (A.9) becomes

\[
\frac{aw}{2\sigma} \exp \left\{ -az + \frac{(a\sigma)^2}{2} \right\} \int_{-(z/\sigma - a\sigma)}^{\infty} \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} v^2 \right\} \sigma dv
\]

\[
= \frac{aw}{2} \exp \left\{ -az + \frac{(a\sigma)^2}{2} \right\} \int_{-\infty}^{z/(\sigma - a\sigma)} \phi(v) dv
\]

\[
= \frac{aw}{2} \exp \left\{ -az + \frac{(a\sigma)^2}{2} \right\} \Phi(z/\sigma - a\sigma).
\]

Hence for \( \mu \neq 0 \),

\[
\int_{-\infty}^{\infty} f(z|\mu) \pi(\mu) d\mu = \frac{aw}{2} \exp \left\{ \frac{(a\sigma)^2}{2} \right\} \left\{ e^{-az} \Phi(z/\sigma - a\sigma) + e^{az} \tilde{\Phi}(z/\sigma + a\sigma) \right\}.
\]

Now for \( \mu > 0 \),

\[
f(z|\mu) \pi(\mu) = \frac{aw}{2} \exp(-a\mu) \cdot \frac{1}{\sigma \sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left( \frac{z - \mu}{\sigma} \right)^2 \right\}
\]

\[
= \frac{aw}{2\sigma \sqrt{2\pi}} \exp \left\{ -\frac{1}{2\sigma^2} \left( z^2 - 2z\mu + \mu^2 \right) - a\mu \right\}
\]

\[
= \frac{aw}{2\sigma \sqrt{2\pi}} \exp \left\{ -\frac{1}{2\sigma^2} \left( z^2 - 2z\mu + \mu^2 + 2a\mu \sigma^2 \right) \right\}
\]

\[
= \frac{aw}{2\sigma \sqrt{2\pi}} \exp \left\{ -\frac{1}{2\sigma^2} \left( \mu^2 - 2\mu(z - a\sigma^2) + z^2 \right) \right\}
\]

\[
= \frac{aw}{2\sigma \sqrt{2\pi}} \exp \left\{ -\frac{1}{2\sigma^2} \left( \mu - (z - a\sigma^2) \right)^2 - (z - a\sigma^2)^2 + z^2 \right\}
\]

\[
= \frac{aw}{2\sigma \sqrt{2\pi}} \exp \left\{ -\frac{1}{2\sigma^2} \left( \mu - (z - a\sigma^2) \right)^2 + 2a \sigma^2 \right\}
\]

\[
= \frac{aw}{2\sigma \sqrt{2\pi}} \exp \left\{ -\frac{1}{2\sigma^2} \left( \mu - (z - a\sigma^2) \right)^2 \right\} \exp \left\{ -az + \frac{(a\sigma)^2}{2} \right\}
\]

\[
= \frac{aw}{2\sigma} \exp \left\{ -az + \frac{(a\sigma)^2}{2} \right\} \Phi \left( \frac{\mu - z}{\sigma} + a\sigma \right).
\]
Hence for $\mu > 0$

$$f(\mu \mid z) = \frac{e^{-az\Phi(\frac{\mu - z}{\sigma} + a\sigma)}}{\sigma\{e^{-az\Phi(\frac{\mu}{\sigma} - a\sigma)} + e^{az\Phi(\frac{\mu}{\sigma} + a\sigma)}\}}.$$ 

### A.5 Determination of the posterior median

Let

$$\tilde{F}_1(\mu \mid z) = \int_\mu^\infty f(u \mid z) du.$$ 

Suppose $z > 0$. In the case of $\mu \geq 0$, we have

$$\tilde{F}_1(\mu \mid z) = \frac{e^{-az\tilde{\Phi}(\frac{\mu - z}{\sigma} + a\sigma)}}{e^{-az\Phi(\frac{\mu}{\sigma} - a\sigma)} + e^{az\tilde{\Phi}(\frac{\mu}{\sigma} + a\sigma)}},$$

If $z > 0$, we can find $\hat{\mu}(z, \omega)$ from the properties

$$\hat{\mu}(z; \omega) = 0 \quad \text{if } \omega_{\text{post}} \tilde{F}_1(0 \mid z) \leq \frac{1}{2}$$

$$\tilde{F}_1(\hat{\mu}(z; \omega) \mid z) = \{2\omega_{\text{post}}(z)\}^{-1} \quad \text{otherwise.} \quad (A.10)$$

Hence if the posterior median is greater than zero, using (A.10), we have

$$\frac{e^{-az\tilde{\Phi}(\frac{\mu - z}{\sigma} + a\sigma)}}{e^{-az\Phi(\frac{\mu}{\sigma} - a\sigma)} + e^{az\tilde{\Phi}(\frac{\mu}{\sigma} + a\sigma)}} = \frac{\omega g(z) + \frac{(1-\omega)}{\sigma} \phi(z/\sigma)}{2\omega g(z)}.$$

As

$$\beta(z) = \frac{g(z)}{\frac{1}{\sigma} \phi(z/\sigma)},$$

we have

$$g(z) = \frac{1}{\sigma} \phi(z/\sigma) \cdot \beta(z) + \frac{1}{\sigma} \phi(z/\sigma).$$
Therefore
\[
e^{-az\tilde{\Phi}\left(\frac{\hat{\mu} - z}{\sigma} + a\sigma\right)} e^{-az\Phi\left(\frac{\hat{\mu}}{\sigma} + a\sigma\right)}
= \frac{e^{-az\Phi\left(\frac{\hat{\mu} - z}{\sigma} + a\sigma\right)}}{2\omega g(z)}
= \frac{1}{2\omega g(z)} \phi(z/\sigma)(\omega \beta(z) + \omega + 1 - \omega)
= \frac{\sigma^{-1}\phi(z/\sigma)(1 + \omega \beta(z))}{aw \exp\left(\frac{1}{2}(a\sigma)^2\right)e^{-az\Phi\left(\frac{\hat{\mu}}{\sigma} - a\sigma\right)} + e^{az\tilde{\Phi}\left(\frac{\hat{\mu}}{\sigma} + a\sigma\right)}
= \frac{a^{-1}\omega^{-1}\sigma^{-1}\phi(z/\sigma) \exp \left( -\frac{1}{2}(a\sigma)^2 \right)(1 + \omega \beta(z))}{\phi(z/\sigma) + \omega g(z)}
= \frac{1}{2}\frac{\sigma^{-1}\phi(z/\sigma)(1 + \omega \beta(z))}{aw \exp\left(\frac{1}{2}(a\sigma)^2\right)e^{-az\Phi\left(\frac{\hat{\mu}}{\sigma} - a\sigma\right)} + e^{az\tilde{\Phi}\left(\frac{\hat{\mu}}{\sigma} + a\sigma\right)} \left(1 + \omega \beta(z)\right)}.
\]

Equating both sides, we can simplify to
\[
e^{-az\tilde{\Phi}\left(\frac{\hat{\mu} - z}{\sigma} + a\sigma\right)} = \frac{1}{aw\Phi\sqrt{2\pi}} \exp \left\{ -\frac{1}{2}\left(\frac{z}{\sigma}\right)^2 + (a\sigma)^2 \right\} (1 + \omega \beta(z))
\tilde{\Phi}\left(\frac{\hat{\mu} - z}{\sigma} + a\sigma\right) = a^{-1}\omega^{-1}\sigma^{-1}(\sqrt{2\pi})^{-1} \exp \left\{ -\frac{1}{2}\left(\frac{z}{\sigma}\right)^2 - 2az + (a\sigma)^2 \right\} (1 + \omega \beta(z))
\tilde{\Phi}\left(\frac{\hat{\mu} - z}{\sigma} + a\sigma\right) = a^{-1}\sigma^{-1}\omega^{-1}(\sqrt{2\pi})^{-1} \exp \left\{ -\frac{1}{2}\left(\frac{z}{\sigma} - a\sigma\right)^2 \right\} (1 + \omega \beta(z))
\tilde{\Phi}\left(\frac{\hat{\mu} - z}{\sigma} + a\sigma\right) = a^{-1}\sigma^{-1}\omega^{-1}\phi(z/\sigma - a\sigma)(1 + \omega \beta(z)).
\]

As \(\tilde{\Phi}^{-1}(u) = -\Phi^{-1}(u)\), we have
\[
\frac{\hat{\mu} - z}{\sigma} + a\sigma = -\Phi^{-1}(z_0)
\frac{\hat{\mu}}{\sigma} = \frac{z}{\sigma} - a\sigma - \Phi^{-1}(z_0)
\hat{\mu} = z - a\sigma^2 - \sigma\Phi^{-1}(z_0),
\]
where \(z_0 = a^{-1}\omega^{-1}\phi(z/\sigma - a\sigma)(\omega^{-1} + \beta(z))\).
B Algebra for Laplace-asymmetric Laplace model

B.1 Finding the convolution of a Laplace and asymmetric Laplace distribution

Suppose that we have an observation $z$ which is a linear additive combination as written below:

$$Z = \mu + \epsilon,$$

where $\mu$ follows a Laplace distribution with scaling parameter $a$ and $\epsilon$ follows and asymmetric Laplace distribution with scaling and asymmetry parameters $b$ and $\kappa$ respectively. Thus

$$\gamma(\mu) = \frac{a}{2} e^{-a|\mu|}$$

and

$$\tau(\epsilon) = \begin{cases} \frac{b\kappa}{1+\kappa^2} \exp(-b\kappa|\epsilon|) & \text{if } \epsilon \geq 0 \\ \frac{b\kappa}{1+\kappa^2} \exp(-\frac{b}{\kappa}\epsilon) & \text{if } \epsilon < 0 \end{cases}.$$

We define the convolution as

$$g(z) = \int_{\mathbb{R}} \tau(z-u)\gamma(u)du. \quad (B.1)$$

Consider the case $z > 0$. The integral (B.1) over $\mathbb{R}$ will have 2 points of discontinuity at 0 and $z$. Thus, in order to derive the convolution we consider the integral over three separate intervals, $[-\infty, 0]$, $[0, z]$ and $[z, \infty]$. Thus, the convolution $g(z) = \gamma(\mu) * \tau(\epsilon)$ can be written as
Appendices

\[ g^+(z) = \int_{-\infty}^{0} \frac{ab\kappa}{2(1 + \kappa^2)} \exp \left( au - b\kappa(z-u) \right) du 
+ \int_{0}^{z} \frac{ab\kappa}{2(1 + \kappa^2)} \exp \left( -au - b\kappa z + b\kappa u \right) du 
+ \int_{z}^{\infty} \frac{ab\kappa}{2(1 + \kappa^2)} \exp \left( -au + b/\kappa(z-u) \right) du \]

\[ = \frac{ab\kappa}{2(1 + \kappa^2)} \left[ \exp \left( (a + b\kappa)u \right) \right]_{-\infty}^{0} 
+ \frac{ab\kappa}{2(1 + \kappa^2)} \left[ \exp \left( -(a-b)u \right) \right]_{0}^{z} 
+ \frac{ab\kappa}{2(1 + \kappa^2)} \left[ \exp \left( -u(a+b/\kappa) \right) \right]_{z}^{\infty} \]

\[ = \frac{ab\kappa}{2(1 + \kappa^2)} \left( \frac{\exp(-b\kappa z)}{a + b\kappa} + \frac{\exp(-b\kappa z) \exp((a - b\kappa)z)}{-a + b\kappa} - \frac{\exp(-b\kappa z)}{-a + b\kappa} + \frac{\exp(b\kappa z) \exp((a - b\kappa)z)}{a + b\kappa} \right). \] (B.2)

Equation (B.2) can be simplified to obtain the form given in Section 4.4.3. The process can be repeated in a similar fashion for the \( z < 0 \) case over the intervals \([-\infty, z] \), \([z, 0) \) and \([0, \infty) \).

### B.2 Deriving \( \tilde{F}_1(\mu|x) \)

From hence, we use \( B^+ = b\kappa \) and \( B^- = b/\kappa \) for the purposes of maintaining maximal simplicity. The posterior distribution for \( \mu \) can now be written as

\[ \tilde{F}_1(\mu|z) = \frac{aB^+}{2(1 + \kappa^2)g^{\text{sign}(z)}(z)} \int_{-\infty}^{\mu} \exp(-B^\pm|z-u| - a|u|) du. \]

Now it is more difficult to identify the points of discontinuity: there will clearly be a discontinuity at \( z \) however whether \( z < \mu \) or \( z > \mu \) is unknown and similarly for the discontinuity at 0. Hence we define \( p_0 := \min(0, z) \) and \( p_1 = \max(0, z) \). Thus
\[
\tilde{F}_1(\mu|z) = \int_{-\infty}^{\min(\mu, p_0)} \frac{ab\kappa}{2(1 + \kappa^2)g_{\text{sign}}(z)} \exp \left( au - B^+(z - u) \right) du \\
+ \mathbb{I}(\mu \geq p_1) \int_{p_1}^{\mu} \frac{ab\kappa}{2(1 + \kappa^2)g_{\text{sign}}(z)} \exp \left( -au + B^-(z - u) \right) du \\
+ \mathbb{I}(\mu \geq p_0) \int_{p_0}^{\min(\mu, p_1)} \frac{ab\kappa}{2(1 + \kappa^2)g_{\text{sign}}(z)} \exp \left( -a|u| - B^+|z - u| \right) du.
\]

(B.3)

We now consider these 3 terms in turn. The first term of equation (B.3) becomes

\[
\frac{ab\kappa}{2(1 + \kappa^2)g_{\text{sign}}(z)} \int_{-\infty}^{\min(\mu, p_0)} \exp \left( au - B^+(z - u) \right) du \\
= \frac{ab\kappa}{2(1 + \kappa^2)g_{\text{sign}}(z)} \int_{-\infty}^{\min(\mu, p_0)} \exp \left( (a + B^+)u - B^+z \right) du \\
= \frac{ab\kappa}{2(1 + \kappa^2)g_{\text{sign}}(z)} \exp(-B^+z) \left[ \exp \left( (a + B^+)u \right) \right]_{-\infty}^{\min(\mu, p_0)} \\
= \frac{ab\kappa}{2(1 + \kappa^2)g_{\text{sign}}(z)} \exp(-B^+z) \left( \frac{(a + B^+)\min(\mu, p_0)}{a + B^+} \right).
\]

(B.4)

Similarly, for the second,

\[
\frac{ab\kappa}{2(1 + \kappa^2)g_{\text{sign}}(z)} \int_{p_1}^{\mu} \exp \left( -au + B^-(z - u) \right) du \\
= \frac{ab\kappa}{2(1 + \kappa^2)g_{\text{sign}}(z)} \int_{p_1}^{\mu} \exp \left( B^-z - (a + B^-)u \right) du \\
= \frac{ab\kappa}{2(1 + \kappa^2)g_{\text{sign}}(z)} \exp(B^-z) \left[ \exp \left( -(a + B^-)u \right) \right]_{p_1}^{\mu} \\
= \frac{ab\kappa}{2(1 + \kappa^2)g_{\text{sign}}(z)} \exp(B^-z) \left( -(a + B^-) \right) \\
\times \left( \exp \left( -(a + B^-)\mu \right) - \exp \left( -(a + B^-)p_1 \right) \right).
\]

(B.5)
The third term is slightly more complex and is dependent on the value of the observation $z$. The term can be generalised for all $z$ as

\[
\int_{p_0}^{\min(\mu, p_1)} \frac{ab\kappa}{2(1 + \kappa^2)g(z)} \exp \left( \text{sign}(-z)au + \text{sign}(-z)B^{\text{sign}(z)}(z - u) \right) du
\]

\[
= \frac{ab\kappa}{2(1 + \kappa^2)g(z)} \times \int_{p_0}^{\min(\mu, p_1)} \exp \left( \text{sign}(-z)B^{\text{sign}(z)}z \right) \exp \left( \text{sign}(-z)u(a - B^{\text{sign}(z)}) \right) du
\]

\[
= \frac{ab\kappa}{2(1 + \kappa^2)g(z)} \exp(\text{sign}(-z)B^{\text{sign}(z)}z) \times \left[ \exp \left( \text{sign}(-z)u(a - B^{\text{sign}(z)}) \right) \right]_{p_0}^{\min(\mu, p_1)}
\]

\[
= \frac{ab\kappa}{2(1 + \kappa^2)g(z)} \exp(\text{sign}(-z)B^{\text{sign}(z)}z) \times \left( \exp \left( \text{sign}(-z)(a - B^{\text{sign}(z)}) \min(\mu, p_1) \right) \right.
\]

\[
- \exp \left( \text{sign}(-z)(a - B^{\text{sign}(z)})p_0 \right).
\]  

(B.6)

Combining the three terms, (B.4), (B.5) and (B.6), provides the overall form given in Section 4.4.3. From here, the threshold values were derived as detailed in Section 4.4.3.

**B.3 Identifying parameter constraints**

To find the constraints on $\omega$, we note that the thresholds are given in Section 4.4.3 to be

\[ t^+ = \max \left[ 0, \log \left( \max \left( 0, \frac{2(\omega aB^+ + \omega a^2 - a^2 - B^{-2}B^+ + B^+)(a + B^-)}{\omega a(a + B^-)(B^+ + B^-)} \right) \right) / -a + B^+ \right] \quad \text{for } z \geq 0 \]

\[ t^- = \min \left[ 0, \log \left( \max \left( 0, \frac{2(\omega aB^- + \omega a^2 - a^2 - B^{-2}B^+ + B^+)(a + B^+)}{\omega a(a + B^-)(B^+ + B^-)} \right) \right) / a - B^- \right] \quad \text{for } z < 0. \]
The threshold $t^+$ will only be finite if

\[
\frac{2(\omega aB^+ + \omega a^2 - a^2 - \omega B^+ + B^+^2)(a + B^-)}{\omega a(a + B^+)(B^+ + B^-)} > 0
\]

\[
\omega a B^+ + \omega a^2 - a^2 - \omega B^+^2 + B^+^2 > 0
\]

\[
\omega (aB^+ + a^2 - B^+^2) > a^2 - B^+^2
\]

\[
\omega > \frac{a^2 - B^+^2}{aB^+ + a^2 - B^+^2},
\]

and similarly for the threshold $t^-$.

Furthermore, $t^+$ will be definitively zero unless

\[
\log \left( \frac{2(\omega aB^+ + \omega a^2 - a^2 - \omega B^+^2 + B^+^2)(a + B^-)}{\omega a(a + B^+)(B^+ + B^-)} \right) > 0.
\]

Thus,

\[
\log \left( \frac{2(\omega aB^+ + \omega a^2 - a^2 - \omega B^+^2 + B^+^2)(a + B^-)}{\omega a(a + B^+)(B^+ + B^-)} \right) > 0
\]

\[
\frac{2(\omega aB^+ + \omega a^2 - a^2 - \omega B^+^2 + B^+^2)(a + B^-)}{\omega a(a + B^+)(B^+ + B^-)} > 1
\]

\[
\omega a B^+ + \omega a^2 - a^2 - \omega B^+^2 + B^+^2 > \frac{\omega a(a + B^+)(B^+ + B^-)}{2(a + B^-)}
\]

\[
\omega \left( aB^+ + a^2 - B^+^2 - \frac{a(a + B^+)(B^+ + B^-)}{2(a + B^-)} \right) > a^2 - B^+^2.
\]

The resulting constraint is found to be

\[
\omega < \frac{a^2 - B^+^2}{aB^+ + a^2 - B^+^2 - \frac{a(a + B^+)(B^+ + B^-)}{2(a + B^-)}}.
\]
as $aB^+ + a^2 - B^{-2} - \frac{a(a + B^+)(B^++B^-)}{2(a+B^-)} < 0$.

Similarly, $t^-$ will be definitively zero unless

$$\log \left( \frac{2(\omega aB^- + \omega a^2 - a^2 - \omega B^{-2} + B^{-2})(a + B^+)}{\omega a(a + B^-)(B^+ + B^-)} \right) < 0.$$ 

Thus,

$$\log \left( \frac{2(\omega aB^- + \omega a^2 - a^2 - \omega B^{-2} + B^{-2})(a + B^+)}{\omega a(a + B^-)(B^+ + B^-)} \right) > 0 \quad \text{as } a - B^- < 0$$

$$\frac{2(\omega aB^- + \omega a^2 - a^2 - \omega B^{-2} + B^{-2})(a + B^+)}{\omega a(a + B^-)(B^+ + B^-)} > 1$$

$$\omega aB^- + \omega a^2 - a^2 - \omega B^{-2} + B^{-2} > \frac{\omega a(a + B^-)(B^+ + B^-)}{2(a + B^+)}$$

$$\omega \left( aB^- + a^2 - B^{-2} - \frac{a(a + B^-)(B^+ + B^-)}{2(a + B^+)} \right) > a^2 - B^{-2}.$$ 

Following on, we have

$$\omega \left( aB^- + a^2 - B^{-2} - \frac{a(a + B^-)(B^+ + B^-)}{2(a + B^+)} \right) > a^2 - B^{-2}.$$ 

as $aB^- + a^2 - B^{-2} - \frac{a(a + B^-)(B^+ + B^-)}{2(a + B^+)} < 0$.

### B.4 Finding the posterior median function

For an observation $z > t^+$, the equation for the posterior median $m_\mu(z)$ is given by

$$\tilde{F}_1(m_\mu(z)|z) = 0.5(1 - \varsigma(z)) \quad \text{(Bochkina and Sapatinas, 2005), where } \varsigma(z) = \frac{(1-\omega)\tau(z)}{\omega g^+(z)}\text{ and } \tau(z) \text{ is the asymmetric Laplace density function. We know that } 0 \leq m_\mu(z) \leq z, \text{ hence } p_0 = 0 \text{ and } p_1 = z. \text{ Thus } \tilde{F}_1(m_\mu(z)|z) \text{ reduces substantially and after some algebraic manipulation, the desired result follows.}$$
B.5 Non-shrinkage

To determine which values of $z$ are affected by the strange behaviour, we consider where the value of the median function for $z$ exceeds the observed value of $z$, i.e. for positive $z$, where

$$\frac{1}{-a + B^+} \log \left( \frac{(1 - \omega)(a - B^+)}{\omega a} \right) + \frac{a}{a + B^+} + \frac{(B^+ + B^-) \exp(-az + B^+ z)}{2(a + B^-)} > z,$$

holds true.

Thus,

$$z(-a + B^+) < \log \left( \frac{(1 - \omega)(a - B^+)}{\omega a} \right) + \frac{a}{a + B^+} + \frac{(B^+ + B^-) \exp(-az + B^+ z)}{2(a + B^-)}$$

$$\exp \left( z(-a + B^+) \right) < \left( \frac{1 - \omega}{\omega a} \right) + \frac{a}{a + B^+} + \frac{(B^+ + B^-) \exp(-az + B^+ z)}{2(a + B^-)}$$

$$\exp \left( z(-a + B^+) \right) > \frac{(1 - \omega)(a - B^+)}{\omega a} + \frac{a}{a + B^+}$$

as $1 - \frac{B^+ + B^-}{2(a + B^-)} < 0$ when $\kappa > \frac{a + \sqrt{a^2 + b^2}}{b}$.

$$z > \log \left( \frac{(1 - \omega)(a - B^+)}{\omega a} + \frac{a}{a + B^+} \right)$$

$$-a + B^+$$

and the result follows. A similar derivation can be done for when $z < 0$. 

C List of published papers

Journal papers


Conference proceedings


Bibliography


dynamic range of detection in electrospray ionization mass spectrometry. *Journal of

Tibshirani, R., Hastie, T., Narasimhan, B., Soltys, S., Shi, G., Koong, A., and Le, Q. T.
(2004). Sample classification from protein mass spectrometry by ‘peak probability


microarray analysis: quality filtering, channel normalization, models of variations and

to transcriptional responses to ionizing radiation. *Proceedings of the National Academy
of Science USA*, 98:5116–5121.

and models in a Bayesian perspective. *Journal of the Royal Statistical Society, Series

van’t Wout, A. B., Lehrman, G. K., Mikheeva, S. A., O’Keeffe, G. C., Katze, M. G.,
upon Human Immunodeficiency Virus Type 1 infection of CD4+-T-cell lines. *Journal

laser desorption ionization time-of-flight. *Journal of the American Society for Mass


BIBLIOGRAPHY


