Statistical Methods for Preprocessing Microarray Gene Expression Data

Md. Mizanur Rahman Khondoker

Doctor of Philosophy
University of Edinburgh
2006

Abstract

DNA microarrays facilitate the simultaneous monitoring of expression levels of thousands of genes in cell samples. Preprocessing is an important first step in the analysis of microarray data, to correct for effects arising from imperfections in the technology rather than real biological differences. This thesis deals with developing statistical methods to resolve problems in the data preprocessing step of microarray analysis.

Following a brief overview of the microarray technology and statistical issues in the design, image processing and analysis of microarray experiments, a novel method is developed for combining multiple laser scans of microarrays to correct for "signal saturation" and "signal deterioration" effects in the gene expression measurement. After initial exploratory analysis, a multivariate nonlinear functional regression model with censored Cauchy distributed errors having additive plus multiplicative scale is proposed as a model for combining multiple scan data. The nonlinear relationship in the functional model is realistically defined as the expected value of a pixel accounting for the possibility of being censored at $2^{16} - 1 = 65535$, which is the upper threshold for 16-bit image converting software available in most microarray scanners. The model has been found to flexibly describe the nonlinear relationship in multiple scan data. The censored Cauchy distribution with additive plus multiplicative scale provides a basis for objective and robust estimation of gene expression from multiple scan data adjusting for censoring and deterioration bias in the observed intensity. Through combining multiple scans, the model reduces sampling variability in the gene expression estimates.

A unified approach for nonparametric location and scale normalisation of log-ratio data is considered. A Generalised Additive Model for Location, Scale and Shape (GAMLSS, Rigby and Stasinopoulos, 2005, *Applied Statistics* **54**, 507–554) is proposed for nonparametric location and scale normalisation of log-ratio data. The nonparametric location and scale normalisation based on GAMLSS is attractive mainly for two reasons. First, unlike usual practice where location and scale normalisation are treated as separate problems, GAMLSS incorporates both of them in a single framework. Second, GAMLSS uses a nonparametric approach for modelling both location and scale of log-ratio data, in contrast to the general tendency of using a parametric transformation, such as arcsinh, for variance stabilisation. We compare the performance of GAMLSS with that of Huber *et al.*'s (2002, *Bioinformatics* **18**, S96–S104) arcsinh variance stabilising transformation in detecting differential expression. Simulation studies demonstrate GAMLSS to be more powerful than the parametric method when a GAMLSS location and scale model, fitted to real data, is assumed correct. GAMLSS has been found to be as powerful as the parametric approach even when the parametric model is appropriate. Another advantage of the GAMLSS method is that, within slide GAMLSS normalised data automatically achieves between slides comparability.

Finally, we investigate the optimality of different estimation methods for analysing functional regression models. Parameters of functional models are often not identifiable by direct application of maximum likelihood estimation, and sometimes lead to inconsistent estimators when they are estimable. Alternative estimators are available in the literature to deal with the problems of identifiability and consistency. However, questions still remain concerning the *efficiency* of such estimators. We investigated these estimators in terms of unbiasedness and efficiency for a specific case involving multiple laser scans of microarrays, and found that, in addition to being consistent, methods of Morton (1981, *Biometrika* **68**, 735–737) and Chan and Mak (1983, *Biometrika* **70**, 263–267) are highly efficient and unbiased.

Acknowledgements

I would like to thank my supervisors, Professor Chris A. Glasbey and Dr. Bruce J. Worton, for all their valuable guidance, advice and encouragement over the course of this work. I owe them lots of gratitude for giving me the ideas, guiding me throughout to implement them, and most of all, showing me the way of doing research. I am thankful to Bruce Worton for taking the tedious effort in reading and providing me with valuable comments on earlier versions of this thesis. I would like to thank my mentor Glenn Marion who kept an eye on the progress of my work.

I am grateful to Biomathematics & Statistics Scotland (BioSS) and Scottish Executive Environment and Rural Affairs Department (SEERAD) for funding my Ph.D. studentship. I would like to thank Scottish Centre for Genomic Technology & Informatics (GTI), Harry McArdle and Lorraine Gambling of Rowett Research Institute for kindly providing the data used in this research. I also gratefully acknowledge the helpful comments by Claus-Dieter of BioSS and Thorsten Forster of GTI on certain parts of this thesis.

I feel very fortunate for being able to work at BioSS, and thanks to all staffs and students for being extremely helpful and providing a very friendly atmosphere.

I am grateful to my wife Farhana, for her support and patience during the final year of my Ph.D. Finally, I feel a deep sense of gratitude for my mother and extended family who always encouraged and supported me in my endeavour of achieving the best.

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text.

(Md. Mizanur Rahman Khondoker)

Table of Contents

Chapter 1

Introduction to microarrays

1.1 Introduction

The mystery of life is widely believed to be the result of complicated and organised functionality of thousands of genes and their products, *i.e.*, RNA and proteins, in a living organism. Understanding the function of each gene and the pathways and networks it influences is an enormous challenge, and high throughput technologies such as microarrays have emerged as tools for providing insights into which genes are important in different conditions. Traditional methods in molecular biology generally work on a "one gene in one experiment" basis with which it is difficult to monitor the whole picture of gene function. DNA microarrays can track tens of thousands of molecular reactions in parallel and provide a basis for comparing gene activities of thousands of genes simultaneously in different biological samples. A detail review of the technical aspects of current microarray technologies can be found in Schena (2000). A brief overview can be found in Southern (2001) or Hardiman (2002).

1.2 Quantification of gene expression

In order to understand the role and function of the genes one needs the complete information about their messenger RNA (mRNA) transcripts and proteins. Unfortunately, exploring the protein functions is very difficult due to their unique 3-dimensional structure and a shortage of efficient technologies. To overcome this difficulty one may quantify the amount of mRNA transcripts produced by the genes of interest to measure gene expression. This idea was a motivation for the development of microarray technique as a method allowing for studying the interaction between thousands of genes based on their mRNA transcript level. In eukaryotes, the vast majority of genes are encoded in the double stranded DNA found in the nucleus of most cells. According to the central dogma of molecu-

lar biology (Figure 1.1), first enunciated by Crick (1958) and re-stated in Crick (1970), genes are transcribed into single stranded mRNA before exiting the nucleus where they are used as a template for protein synthesis. The process of a selected target sample binding to matching gene probes on the array is called hybridisation.

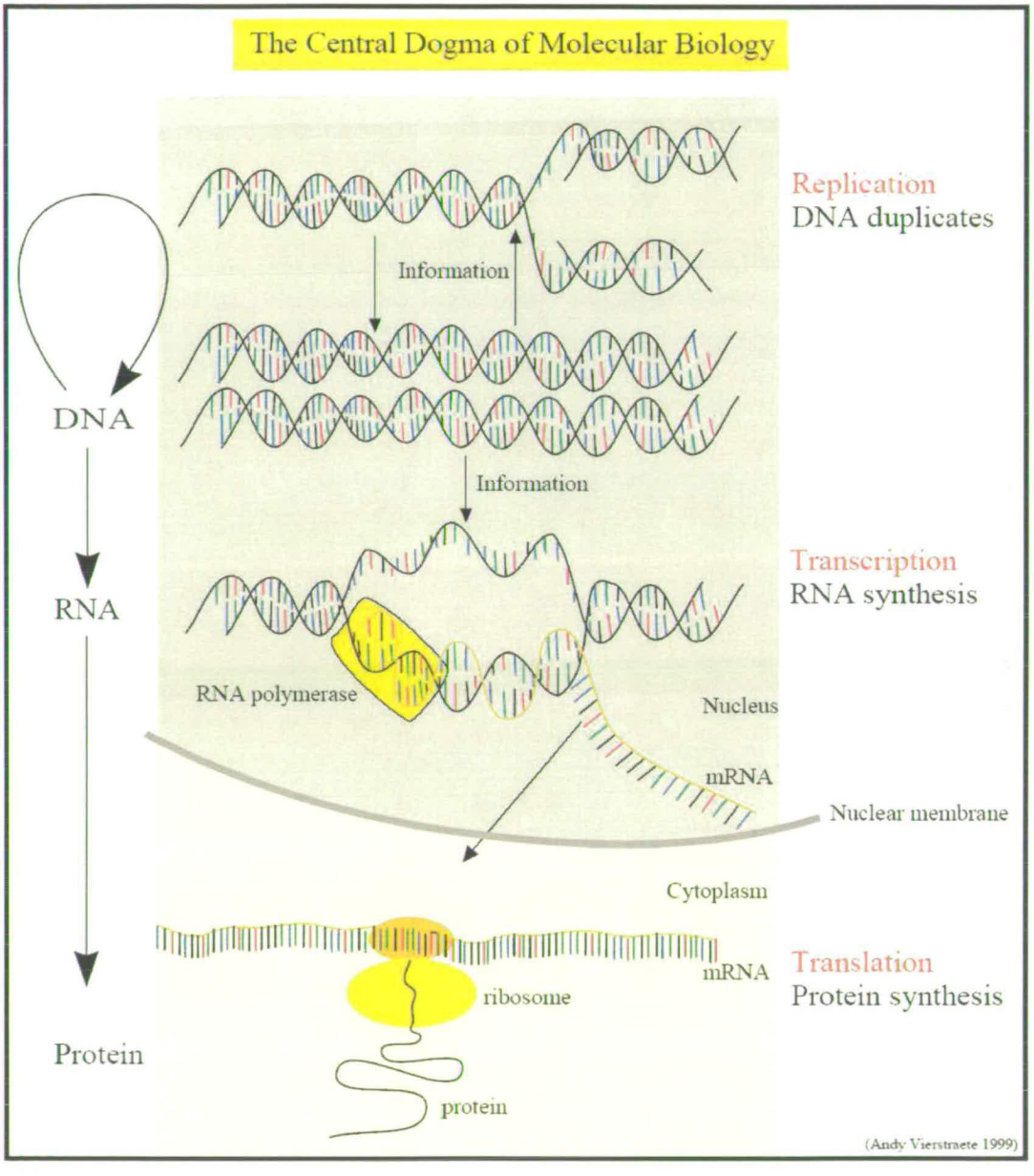


Figure 1.1: Central Dogma of Molecular Biology; DNA is transcribed into RNA which is translated into protein outside the nucleus. Image reproduced from `http://users.ugent.be/~avierstr/principles/centraldogma.html`

Microarray technology is an effort to measure gene expression by quantifying the amount of mRNA produced during transcription by exploiting the complementary base-pairing rule of DNA (A pairs with T and G pairs with C, or in the case of RNA, A pairs with U, uracil).

1.3 DNA microarrays

A DNA microarray, variously known as DNA chip, gene chip, gene array or biochip, is a densely packed array of identified DNA sequences attached to a solid surface, such as glass, plastic or silicon chip. Microarray technology evolved from Southern Blotting, where fragmented DNA is attached to a substrate and then probed with a known gene or fragment.

DNA sequences representing tens of thousands of genes are spotted or *in situ* synthesised on a very small slide like the one in Figure 1.2. The microarray in the picture is comprised of more than 54,000 probe sets capable of analysing expression level of over 47,000 transcripts and variants, including 38,500 well-characterised human genes. In terms of the property of the arrayed sequence, there are two major variants of the DNA microarray technology:

(i) Oligonucleotide arrays and

(ii) Complementary DNA (cDNA) arrays.

Oligonucleotide expression array technology (Lockhart *et al.*, 1996) has recently been adopted in many areas of biomedical research.

Figure 1.2: GeneChip® Human Genome U133 Plus 2.0 Array. Image reproduced from `www.servicexs.com`

On oligonucleotide arrays produced by Affymetrix, as reviewed in Lipshutz *et al.* (1999), each gene is represented by a set of 11 to 20 short sequences of DNA, termed oligonucleotides (Figure 1.3). These oligonucleotides are referred to as the

perfect match (PM) and each of them is paired with a corresponding mismatch (MM), which is identical to the PM-probe except for one nucleotide in the centre of the sequence. These types of chips are produced by GeneChip® technology of Affymetrix Inc.

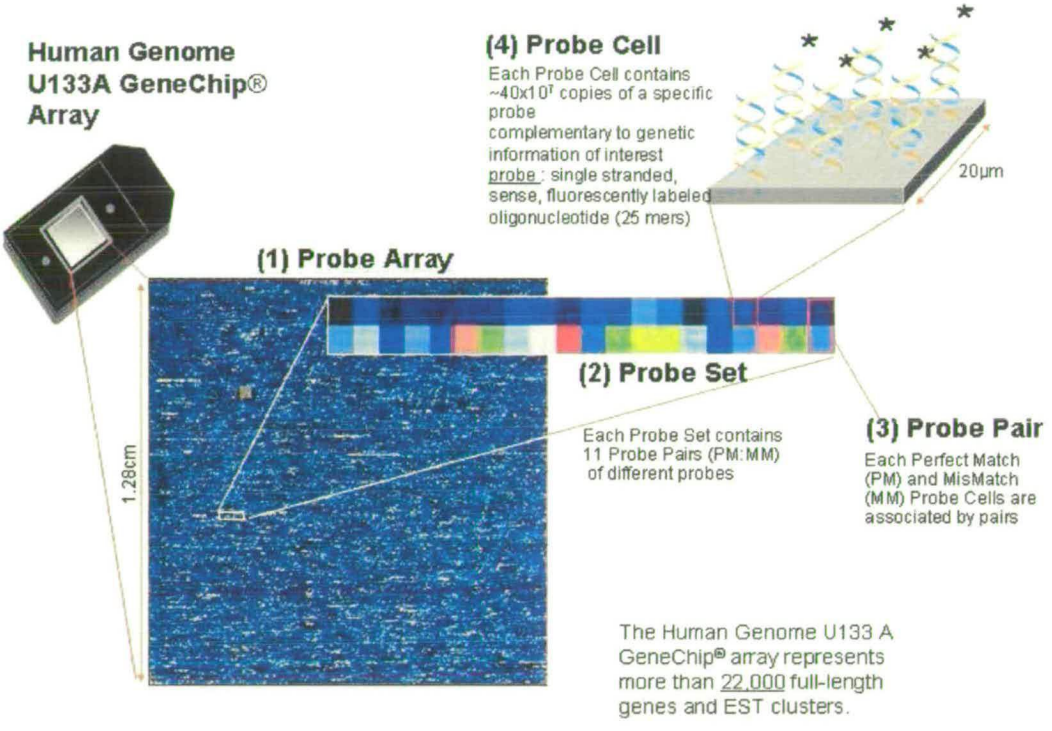


Figure 1.3: Oligonucleotide expression array (HGU133A). Image reproduced from `http://www.weizmann.ac.il/home/ligivol/pictures/system.jpg`

In cDNA technology (Schena *et al.*, 1995; DeRisi *et al.*, 1997), the other widely used type of microarrays, probes of purified DNA are spotted onto a glass slide by specialised robotics. These are usually much longer sequences than in the case of oligonucleotide arrays, often consisting of tens or hundreds of bases. On cDNA arrays typically two samples are analysed in a comparative fashion.

In general, two major differences exist between the cDNA microarray platform and oligonucleotide platforms. First, on cDNA microarray the length of the DNA fragment is generally 500–1500 mer versus 25–60 mer on the oligonucleotide array. Second, in cDNA microarray experiments two RNA samples, control and experimental, are labeled with different fluorophores (Cy5 and Cy3) and competitively hybridised to the same microarray slide. This distinguishes the cDNA microarray platform from most oligonucleotide based platforms where one sample is hybridised to one slide.

There are two major application forms for the DNA microarray technology:

(i) identification of sequence (gene/gene mutation) and

(ii) determination of activity (expression) level of genes.

DNA microarrays are proving immensely valuable to cell biologists, to scientists who study the roots of cancer and other complex diseases and to drug researchers. Microarrays may also be useful as quick diagnostic and prognostic tools. The research and diagnostic information provided by DNA chips may eventually help physicians provide highly individualised therapies.

1.4 Comparative cDNA hybridisation experiment

There are numerous paper-based and electronic references on good descriptions of cDNA hybridisation experiments. An animated description of a comparative hybridisation can be found at `http://www.cs.wustl.edu/~jbuhler/research/array/`. The following descriptions are adapted from the above website, Amaratunga and Cabrera (2004) and the Microarray Facility link of the website of Brunel University (`http://www.brunel.ac.uk/`). The goal of a comparative cDNA hybridisation is to compare gene expression in two or more different samples. The major steps of a comparative cDNA hybridisation experiment are:

(i) selection of cell populations,

(ii) mRNA extraction and reverse transcriptions,

(iii) fluorescent labelling of cDNAs,

(iv) hybridisation to a DNA microarray, and

(v) scanning the hybridised array.

Selection of cell population

Selection of cell populations depends on the biological question under investigation. For example, genetic diseases are result of mutations in a gene or set of genes. Consequence of thus altered mutant genes can be a disease as these genes express inappropriately or do not express at all. Genetic disease, cancer for example, could occur when certain regulatory genes such as p53 tumor suppressor gene, are deleted, inactivated or become constitutively active, *i.e.*, become always transcribed regardless of any regulatory factors. Microarray experiment can be

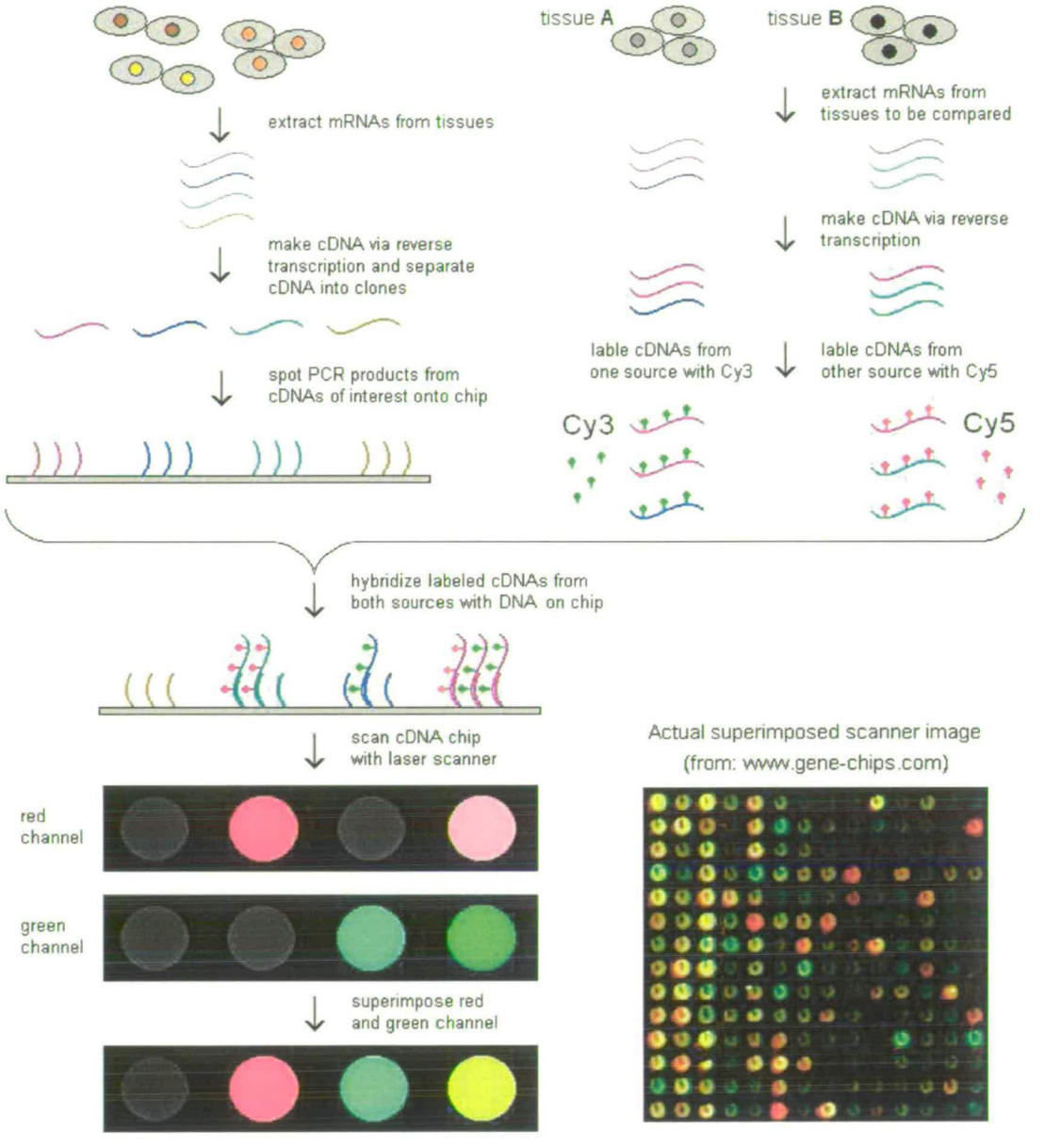


Figure 1.4: Two-channel cDNA experiment. Image courtesy of Helge Roider, Chalmers University, Sweden.

used to identify which genes are differentially expressed in cancerous cells versus normal cells. The choice of cell populations in this case is clear: the healthy cells become a *de facto* reference population and the cancerous cells will be the experimental population. Amaratunga and Cabrera (2004, Chapter 3) described seven common types of microarray experiments used respectively in tissue-specific gene expression studies, developmental genetics, genetic disease studies, complex disease studies, study of pharmacological agents, plant breeding studies and environmental monitoring. Wit and McClure (2004, pp. 3–12) described several microarray experiments of very different nature to illustrate the scope of the technology and to demonstrate different statistical concepts to analyse such experiments.

mRNA extraction and reverse transcriptions

After selection of cell populations, sample messenger RNA's (mRNA's), the mobile copies of genes and the templates for protein synthesis in cells, are extracted from both experimental and reference cell populations. As the extracted mRNA's are prone to being destroyed, they are reverse-transcribed back into more stable DNA form. The products of this reaction are called complementary DNA's (cDNA's) because their sequences are the complements of the original mRNA sequences. Impurities in RNA preparations can have an adverse effect on both the labelling and the stability of the fluorescent dyes used to label the RNA and impurities such as cellular protein, lipids and carbohydrates can cause significant non-specific binding to a cDNA spot on the array, resulting in false positive expression values, therefore purified RNA samples are crucial in a microarray experiment.

Fluorescent labelling of cDNAs

In order to detect cDNA's bound to the microarray, they are labelled with a reporter molecule that identifies their presence. The reporters most commonly used in comparative hybridisation to microarrays are fluorescent dyes (fluors). A differently coloured fluor, usually rhodamine or Cy3 (green) for one sample and fluorescein or Cy5 (red) for the other sample, is used for each sample.

Hybridisation to a DNA microarray

The samples are then pooled and applied to the microarray to allow hybridisation to the cDNA on the array via complementary interaction. Hybridisation is carried out in a hybridisation chamber containing the microarray slide. The chamber is placed in a 42 degree Celcius water bath and incubated for 16–20 hours. The

array holds tens/hundreds of thousands of spots, each of which contains a different DNA sequence. If the sample contains a cDNA whose sequence is complementary to the DNA on a given spot, that cDNA will hybridise to the spot, where it will be detectable by its fluorescence. In this way, every spot on an array is an independent assay for the presence of a different cDNA. There is generally enough DNA on each spot so that both probes can hybridise to it at once without interference.

Scanning the hybridised array

Once the pooled sample has been hybridised to the array and any loose cDNA sequence has been washed off, the array must be scanned, at the wavelength of each fluor, to determine the amount of labelled sample bound to each spot. The reporter molecules emit detectable light when stimulated by a laser. The emitted light is captured by a detector, either charged-coupled device (CCD) camera or a confocal microscope, which records its intensity. Spots with more bound sample will have more reporters and will therefore fluoresce more intensely. Although it is supposed to pick up light emitted by the target cDNAs bound to their complementary spots, the scanner will inevitably also pick up light from various other sources, including the labelled sample hybridising non-specifically to the slide, residual (unwashed) labelled sample adhering to the slide, various chemicals used in processing the slide, and even the slide itself. This extra light is called background. Scanner settings can affect both the precision of the intensity measurements as well as the lower and upper threshold intensity levels that can be measured. Intensities outside this range, called the dynamic range, can not be properly quantified and are often set to the corresponding threshold level. When intensities exceed the upper threshold, saturation is said to have occurred. There is a trade-off between the precision and the dynamic range, and a reasonable balance is important. Arrays are often scanned more than once but common practice is to use data from a single scan, based on some arbitrary exploratory checks, in the subsequent analysis. Use of combined data from multiple scans of hybridised microarray may be useful to get improved gene expression measurements. One part of this thesis concerns combining data from multiple scans at different scanner settings to adjust for the saturation problem in the image analysis process.

1.5 Summary

Microarray technology has been rapidly advancing since its introduction in 1995. New and advanced platforms are being developed to overcome the imperfections in the technology. In this chapter, we described the basic technology and the underlying biology behind it. Emphasis was given to the cDNA platforms, because all the data used in this thesis concern cDNA experiments. In the next chapter, we provide a brief overview of the statistical issues in different steps of the analysis process, such as, the design, image processing, data preprocessing and statistical inference of microarray experiments.

Chapter 2

Statistical design and analysis

2.1 Introduction

Microarrays are powerful for studying gene expressions but are based on technologies that still require improvements. The sources of technological variability arising throughout the measurement process can obscure the biological signals of interest. Properly designed statistical experiments, reliable image processing and appropriately chosen, possibly modified or newly developed, statistical techniques are required to maximise the gain from microarray experiments by detecting and removing the numerous sources of variability. In this chapter, we review the available resources in the design, image processing, data preprocessing and analysis of microarray experiments.

2.2 Experimental design

As in any other area of statistics, proper statistical design of microarray experiments is essential to ensure that the effects of interest are accurately and precisely measured. Although this may not be always feasible in the microarray context, Amaratunga and Cabrera (2004, p. 146) advised to adhere, as much as possible, to the fundamental principles of the theory of design of experiments: randomization, replication and balance, while designing a microarray experiments. Apart from the basic principles of design of experiments, there are additional issues to be considered for designing microarray experiments. In an experiment, 'crossing' means considering all possible factor combinations as the potential set of conditions. In addition to blocking and randomisation, Wit and McClure (2004, pp. 37–38) described 'crossing' as a basic microarray design criterion to compare the effects of two or more sets of conditions in the same experiment. As discussed by Speed (2003, pp. 35–37), any microarray experiment involves two main design aspects:

(i) design of the array itself, and

(ii) allocation of mRNA samples to the microarrays.

The design of the array itself involves deciding which DNA probes are to be printed on the solid substrate, be it a membrane, glass slide or silicon chip, and where they are to be printed. Allocation of mRNA samples to the microarrays involves deciding how mRNA samples should be prepared for the hybridisations, how they should be labelled, and the nature and number of the replicates to be done. A review of design issues for cDNA microarray experiments can be found in Yang and Speed (2002).

2.2.1 Replication

Replication is a key aspect of any comparative experimentation to increase precision and more importantly to provide a basis for formal statistical inference. It is now becoming widely accepted that microarray experiments need to be replicated (Parmigiani *et al.*, 2003, p. 7). The presence of internal control is helpful but not sufficient to eliminate the numerous sources of experimental error. In microarray context, replication can have a number of different forms: duplicate spots, technical replicates and biological replicates. It is important to realise that any type of replication offers information only regrading the particular source of variability associated with that type of replication and no other (Amaratunga and Cabrera, 2004, pp. 82–83). Depending on the experimental setting, it may therefore be important to consider one, two or all these types of replicates. The type of replication to be used in a given experiment depends on the precision and generalisability of the experimental results sought by the experimenter (Speed, 2003, p. 42). In general, biological replicates are used to support generalisations of conclusions and technical replicates to reduce the variability of the measurements themselves. Given that several possible forms of technical and biological replication exist, judgment will need to be exercised on the question of how much replication of a given kind is desirable, subject to experimental and cost constraints.

Duplicate spots

As discussed by Speed (2003, p. 41), duplicate spots provide valuable quality information, as the degree of concordance between duplicate spot intensities is an excellent quality indicator. However, data from the pairs can not be regarded as independent because replicate on the same slide, particularly adjacent spots,

will share most of their experimental conditions. Nevertheless, averaging log-ratios from duplicate spots is appropriate. Their close association means that the information is less than that from pairs of truly independent duplicate spot measurements.

Technical replicates

Technical replicates refer to multiple-array hybridisation where the target mRNA is collected from the same pool, *i.e.*, from the same biological extraction. Technical replication is useful for controlling technical variation, which arises from the handling steps, such as mRNA extraction, amplification, labelling, hybridisation, and scanning. This variation introduces uncertainty to the intensity measurements associated with a gene. Using technical replicates and averaging across them allows gene expression levels to be estimated with greater precision. The higher the number of replicates, the greater is the precision.

Biological replicates

Biological replicates refer to analysis of mRNA of the same type from different subjects, for example muscle tissue treated with the same drug in different mice in the same species or inbred strain. This type of replication was termed as Biological replicates–type II by Speed (2003, p. 42). Biological replicates–type I was used to refer to hybridisations involving mRNA from different extractions, for example, different sample of cells from a particular cell line or from the same tissue.

Biological replicates are used to deal with biological variation, which is the natural variability among subjects due to genetic diversity, environmental effects and other causes. This variation also contributes uncertainty to the intensity measurement associated with a gene. Averaging across biological replicates allows gene expression levels to be estimated with greater biological precision.

2.2.2 Optimal design

Use of indirect designs, also known as common reference designs was common in early microarray studies (DeRisi *et al.*, 1996; Spellman *et al.*, 1998; Perou *et al.*, 1999). Common references are frequently used to provide easy means of comparing many samples against one another. Designs that provide direct estimates of log-ratios were considered by several studies, *e.g.*, Jin *et al.* (2001); Kerr *et al.* (2001). Fixed or random effect linear models and analysis of variance can be used to combine data from such designs. Until recently, the main work

on design of two-channel microarray experiments is due to Kerr and Churchill (2001), and Glonek and Solomon (2004), who have applied ideas from optimal experimental design to suggest efficient designs for some of the common cDNA microarray experiments. Kerr and Churchill (2001) based their comparisons of different designs on the A-optimality criterion. In addition, they introduced a novel class of design, called loop designs, and found that under A-optimality, loop designs were more efficient than common reference designs. However, loop designs are efficient for comparing small to moderate number of conditions. Kerr and Churchill (2001) noticed, through an exhaustive search of all possible designs of limited number of slides and conditions, that the optimalilty of loop designs does not hold when there are more than eight conditions. More efficient way of searching for optimal designs, termed *simulated annealing*, was proposed by Wit *et al.* (2004). Glonek and Solomon (2004) studied optimal design for time course and factorial experiments. Determination of optimal sample size for multiple testing in the case of gene expression microarray data has been discussed by Müller *et al.* (2004).

2.3 Image processing

Once the target cDNAs have been hybridised to the microarray and any loose sequence has been washed off, the array must be scanned to determine how much of each target is bound to each spot. The result is a series of images, one per channel. Oligonucleotide array (one-channel array) gives one image per array, whereas a two-channel microarray yields two images, one for each channel, per array. Scanner reads a microarray by dividing it up into a large number of pixels and recording the intensity level of the fluorescence at each pixel. The resulting rectangular array of pixels and their associated intensities constitute the raw image of the microarray. Image analysis methods are then applied to extract spot intensity from these raw image.

Yang *et al.* (2002) and Glasbey and Ghazal (2003) reviewed a number of existing methods and proposed their own methods of image analysis for microarrays. Processing of the raw image generally involves three major steps: addressing or gridding, segmentation and intensity extraction.

2.3.1 Addressing or gridding

Addressing or gridding is the process of assigning coordinates to the center of each spot. As explained by Yang *et al.* (2002), a number of parameters need to be estimated to address the spot on the image, *i.e.*, to match an idealised model

of the array with the scanned image data. In practice, the arraying process is not perfect, so that the grid that is actually arrayed tends to be slightly deformed version of the target regular rectangular grid. As a result the overlaid grid needs some fine-tuning, which can be done by manipulating the rows and columns of the overlaid grid until it is satisfactorily aligned. Reliability of the addressing stage can be increased by allowing user intervention. However this can make the process potentially slow. Most software systems now provide both manual and automatic gridding procedures. Ideally, one seeks reliability while attempting to minimise user intervention to maximise efficiency.

2.3.2 Segmentation and intensity extraction

Segmentation of an image generally refers to the process of partitioning the image into different regions, each having certain properties. Once the locations of the centers of the spots have been determined, the next step is to separate the spot, *i.e.*, the region of the slide on which cDNA was actually arrayed. In a microarray experiment, pixels belonging to a spot of interest is called foreground and all other pixels surrounding the arrayed spot constitute background. Segmentation in microarray experiment therefore refers to the process of classification of pixels as foreground or background, so that fluorescence intensities can be calculated for each cDNA sequence as measures of gene expression. Depending on the geometry of the spots produced, Yang *et al.* (2002) categorised the existing microarray image segmentation methods into four groups as listed below. In general, most software packages implement a number of segmentation methods. Examples are shown within parentheses.

(i) Fixed circle segmentation (*ScanAlyze*, Eisen, 1999; *GenePix*, Axon Instruments Inc., 1999; *QuantArray*, GSI Luminomics, 1999),

(ii) Adaptive circle segmentation (*GenePix*),

(iii) Adaptive shape segmentation (*Spot*, Buckley, 2000), and

(iv) Histogram segmentation (*QuantArray*).

Fixed circle segmentation fits a circle with a constant diameter to all the spots in the image. This method is easy to implement and works nicely when all the spots are circular and of the same size. However, this method is not satisfactory as the spots tend to vary in size and shape due to a number of reasons.

Adaptive circle segmentation estimates the circle's diameter for each spot. This method may also give poor fit as the spots are rarely circular and can exhibit

oval or donut shapes. Segmentation algorithm that do not place restrictions on the shape of the spots are therefore desirable.

Adaptive shape segmentation methods are beginning to be applied in microarray analysis although not yet available in the most widely used software packages. Two commonly used methods for adaptive segmentation in image analysis are the watershed (Beucher and Meyer, 1993) and seeded region growing (SRG), (Adams and Bischof, 1994). Both watershed and SRG segmentation require the specification of starting points, or seeds. The weak point of segmentation procedures using these methods can be the selection of the number and location of the seed points. In microarray image analysis the number of features (spots) is known exactly a priori and the approximate locations of the spot centres are determined at the addressing stage. Microarray images are therefore well-suited to such methods.

Histogram segmentation places a mask, circular or square, over each spot. This mask should be larger than the spot. The histogram of pixel values within the mask is examined to determine a threshold value. Each pixel within the mask is then classified as foreground or background depending on whether its intensity is above or below this threshold. The histogram method that is implemented in *QuantArray* uses a square target mask and defines foreground and background as the mean intensities between some predefined percentile values. By default, these are the 5th and 20th percentiles for the background and the 80th and 95th percentiles for the foreground.

Intensity extraction refers to computation of average foreground and background intensities for each spot and possibly computation of some quality measures. Most microarray analysis packages define the foreground intensity as the mean pixel values within the segmented spot mask. However, because the distribution of pixel intensities might be irregular, other measures of location, such as the median or trimmed mean, biweight, and mode, are also sometimes used. Mean or median are also commonly used for estimating spot background. More variations however exist for estimating spot background. Background is generally removed from the foreground intensities prior to formal analysis.

2.3.3 Correction for saturated pixels

One of the sources of systematic bias in the gene expression measurements is signal/pixel censoring. The scanned gray scale images of microarrays are usually stored in 16-bit tagged image file format (TIFF). Image processing software converts the image into intensity measurements. For 16-bit image converters these measurements range between 0 and $2^{16} - 1$. Therefore, any pixel having fluorescence intensity greater than $2^{16} - 1$ is censored at this upper limit.

Recent image analysis methods (*e.g.*, Ekstrøm *et al.*, 2004; Glasbey *et al.*, 2006) also deal with the saturation problem which occurs because image converting software can not record intensities beyond a certain threshold. Spatial statistical model was considered by Ekstrøm *et al.* (2004) to predict signal intensities of the censored pixels. Censored pixel values can be estimated optimally for models using transformed data having approximately Gaussian distribution with a mean value function determined by gene intensities and spot shapes and a corresponding covariance function. The authors investigate several types of transformations on the pixel level such as the logarithmic transformation, the Box-Cox family (Box and Cox, 1964) and the inverse hyperbolic sine (arcsinh) transformation (Huber *et al.*, 2002; Durbin *et al.*, 2002), also called the generalised logarithm (Rocke and Durbin, 2003). The paper compares these transformations in combination with four spot shape models: (i) a cylindric plateau spot distribution, (ii) an isotropic two-dimensional Gaussian distribution, (iii) a crater spot distribution consisting of a difference between two scaled isotropic two-dimensional Gaussian distribution and (iv) polynomial-hyperbolic spot shape model. The first three models were suggested by Wierling *et al.* (2002), and according to the paper, do not seem to provide satisfactory description for the data set considered. The proposed polynomial-hyperbolic spot shape model with a second degree polynomial gives considerable improvement in performance.

Glasbey *et al.* (2006) proposed a linear model to impute censored pixels based on the principal components of the uncensored spots on the same array. Arrays with censored spots generally also has many uncensored spots. The idea, as being used in other domain of image analysis, is to use the principal components or eigenvectors computed from these uncensored spots as a basis for a model. The method is sufficiently flexible for modelling non-circular spot shapes and profiles that do not conform to parametric models and has been shown to be more effective than the polynomial-hyperbolic model of Ekstrøm *et al.* (2004) in correcting for the censoring bias.

2.4 Combining multiple scans

Analysis of microarray experiments is commonly based on data from a single scan of hybridised microarrays, although it is standard practice to scan a single microarray several times. Use of multiple scan can be useful as illustrated by Romualdi *et al.* (2003) who used multiple scan data to improve detection of differentially expressed genes through image integration.

One way of minimising pixel censoring is to reduce the amplification setting

(gain) in the scanner. This is not a good solution as it causes another problem, signal deterioration, for weakly expressed genes. In single scan analysis a compromise scanner setting is chosen, generally based on some informal exploratory checks, to make a balance between the amount of censoring and deterioration. It is therefore not usual for gene expressions from a single scan to be approximately proportional to the underlying expression levels over the entire range of gene expressions. Since the problems at the two ends, signal censoring and signal deterioration, are in conflict, no single scan is optimal. Noting that low expression levels are better measured at high setting, while high expression levels are better measured at low setting, combining multiple scans may be a good idea for getting improved gene expression measures across the whole range of data.

Motivated by the fact that spot intensity reported by a scanner is linear only within a certain range of intensities, being dominated by noise below and subject to signal saturation above that range, Dudley *et al.* (2002) suggested a linear regression algorithm to combine the linear ranges of multiple scans taken at different scanner sensitivity settings on to an extended linear range.

Bell (2003) developed an algorithm, implemented in an image analysis software called MAVI Pro, for dealing with signal saturation and deterioration. The algorithm reads in image analysis data from arrays scanned with a range of different amplification settings, eliminates the saturated and deteriorated values, and then computes the intensity for a specific amplification using linear regression of intensity on amplification. In Figure 2.1, the intensity of the spot derived from the image analysis software is plotted against the photomultiplier amplification gain for one channel. Every gray line in the plot connects the intensity values belonging to one spot. MAVI first calculates an amplification value for which it is going to give out the result. It chooses the amplification, which is closest to 40% of the amplification span. In Figure 2.1 this value is 45. Then a straight line is fitted to intensities of all spots that very probably do not have saturation or signal deterioration. The assessment about the presence of saturation or deterioration is made informally from the range of the data. Spots are chosen for this if their highest value is below 40,000 and their lowest value is higher than 100. This might need to be adapted if MAVI should be used for a different scanner. From these fits an average slope is calculated. Then MAVI goes through the intensities of all spots separately. To check for saturation the slope between the intensities of the two highest amplification settings is compared with the average slope. If the discrepancy is larger than 30% the intensity to the higher of the two amplifications is removed and the slope of the next two lower amplifications is checked the same way. This is done consecutively until one slope fulfills the criterion of

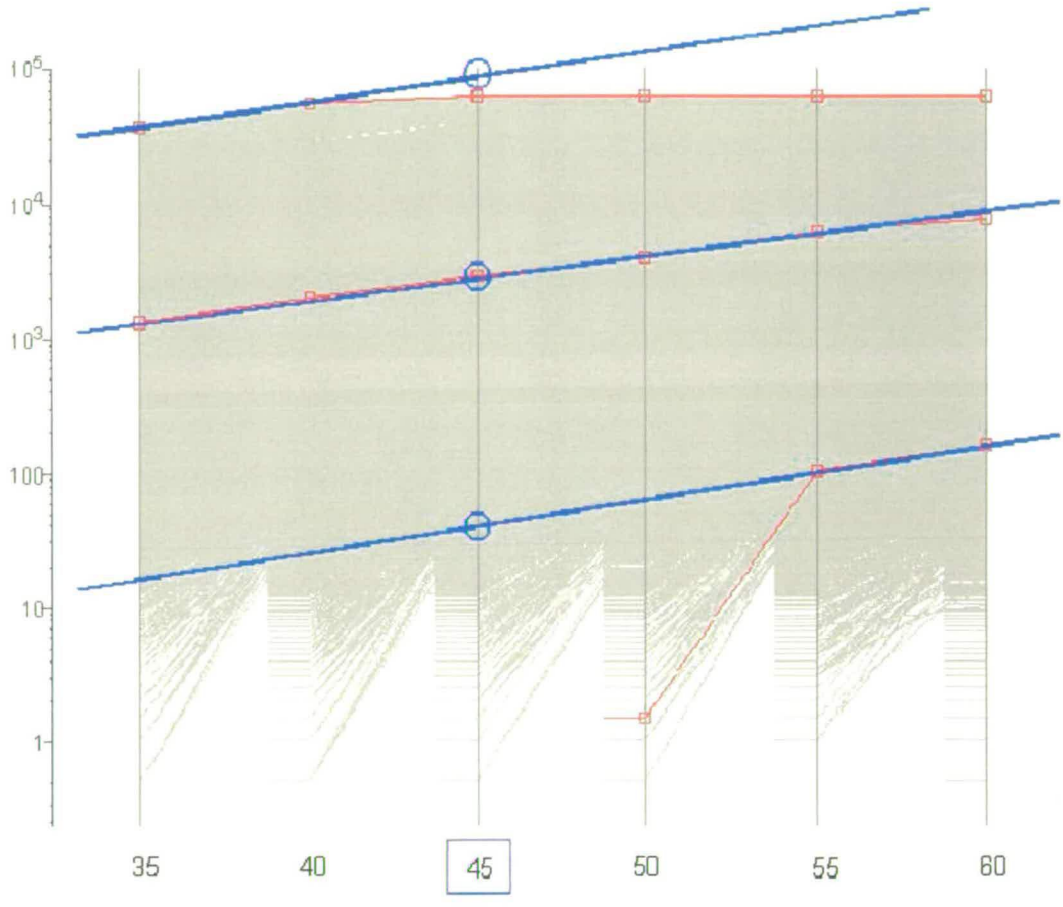


Figure 2.1: Intensities against amplification. Reproduced from Bell (2003).

less than 30% similarity to the average slope. The same is done consecutively from lower amplifications upwards to eliminate intensity with signal deterioration. MAVI then takes the surviving intensity values and again does a linear regression on them (blue lines in Figure 2.1). From the fitted line, the middle one in Figure 2.1, it then calculates the intensity at the previously determined amplification for calculation (blue circles in Figure 2.1).

García de la Nava *et al.* (2004) suggested a simple method for saturation reduction based on two scans at different sensitivities, one at a low sensitivity level (L) and the other at a high sensitivity level (H). Two simple mathematical models, based on linear and "gamma" correction curves, which are power functions used to code and decode luminance values in a video or still image system, are presented for relating the two measurements to each other and producing a coherent and extended range of values. Suppose that I_i is the light flux intensity of spot i, L_i stands for the low sensitivity electrical current at photodetector and H_i for the high sensitivity one. Saturation is assumed to be negligible in the low

sensitivity scan, so the value of L_i is given in linear terms

$$L_i = kI_i,$$

where k is some constant. On the other hand, H_i is described by either a clipped linear curve or a power function. The clipped linear curve is described by

$$H_i = \begin{cases} pI_i, & \text{if } I_i < T/p \\ T, & \text{otherwise,} \end{cases}$$

where p is the proportionality constant between read value and spot intensity. The previous equation can be reduced to

$$H_i = \begin{cases} mL_i, & \text{if } I_i < T/m \\ T, & \text{otherwise,} \end{cases}$$

where T is the saturation (clipping) level and $m = p/k$, the proportionality constant between the low sensitivity and high sensitivity scans.

The correction curve based on power function is defined by

$$H_i = cI_i^{\gamma},$$

where c is a constant. The relation between H_i and L_i can therefore be specified by

$$H_i = bL_i^{\gamma},$$

where $b = c/k^{\gamma}$. Least Trimmed Squares (LTS, Rousseeuw and Leroy, 1987) is a robust version of least squares regression obtained by minimising the sum of squares of certain proportion of smallest residuals. García de la Nava *et al.* (2004) used LTS for estimating the parameters of the proposed models.

Wit (2004, personal communication) has considered a generative model for combining multiple scans to improve gene expression estimates. If

$$\mu = (\mu_1, \cdots, \mu_G)$$

represents the underlying expression levels of G genes under a particular condition in a particular fixed RNA sample, then the average observed intensities measured with some ideal scanner having linear infinite dynamic range under a particular set of scanning settings s would be

$$C_s\mu = (C_s\mu_1, \cdots, C_s\mu_G),$$

where C_s is some array-wide constant associated with the particular settings s. In this ideal case the author assumed that the observed intensities ξ_i $(i = 1, \cdots, G)$ follow a log-normal distribution such that

$$E(\xi) = C_s\mu.$$

In particular, the individual spot intensities are assumed to be distributed as

$$\xi \sim LN(m_i, \sigma^2),$$

such that

$$C_s\mu_i = e^{m_i+\sigma^2/2}.$$

However, due to saturation, the actual average is distorted and capped at some maximum intensity level T. To model this distortion, Wit (2004, personal communication) suggested a distortion function

$$f_\delta(\xi) = \frac{T\xi}{\delta + \xi}.$$

The observed intensity I_i for gene i on a particular array is therefore taken to be

$$I_i = f_\delta(\xi_i),$$

where ξ is the log-normally distributed signal for the ideal microarray with mean $C_s\mu_i$. Considering S separate scannings of the same microarray with corresponding distortion functions $f_{\delta_1}, \cdots, f_{\delta_S}$, the author suggested maximum likelihood method for estimating the expression level μ.

2.5 Functional regression for combining multiple laser scans

Standard regression methods are commonly used for modelling a set of responses as functions of a set of predictor variables, where only response variables are allowed to have measurement errors. Application of standard regression, where both response and predictor variables are subject to measurement errors, may often be misleading. For example, agricultural variables such as rainfall, soil nitrogen content, degree of pest infestation etc., which are commonly used for predicting yield, can not be measured precisely. In management sciences, social sciences, and nearly every other field many other variables can only be measured with error. Although it is very unlikely to have a situation in practice where the predictor variables can be measured accurately, analysts commonly prefer to use standard regression method because of its familiarity and ease of application. The method has found applications in dealing with multiple laser scans as well, *e.g.*, Dudley *et al.* (2002) used linear regreesion to relate intensity data obtained from multiple laser scannings of microarrays. The authors used intensity data from one of the scans as response, and the data from other scans as predictors. However, because microarray data are generally very noisy, and each individual laser scan

is subject to similar level of measurement errors, standard regression methods are not appropriate for calibrating such data. Functional regression models, a type of *measurement error* models (Cheng and Van Ness, 1999, Chapter 1), where both response and predictor variables are allowed to have measurement errors, are more realistic.

The basic functional model postulates a linear relationship

$$\eta = \alpha + \beta\mu$$

between two unobservable nonstochastic variables η and μ. The variables η and μ can only be observed with additive errors, *i.e.*, instead of observing η and μ directly, one observes the variables

$$x = \mu + \epsilon_1 \quad \text{and} \quad y = \eta + \epsilon_2,$$

where (ϵ_1, ϵ_2) is normally distributed with zero mean vector and covariance matrix

$$\text{var}\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}.$$

Functional regression models have both advantages and disadvantages over the standard regression models. One major limitation is that the model generally has more parameters than the number of observations, and the parameters are not estimable by direct application of maximum likelihood method. Although likelihood solution exists when restrictions are imposed on certain parameters, maximum likelihood method often leads to inconsistent estimators.

Alternative estimators have been suggested in the literature, *e.g.*, Sprent (1976), Morton (1981), Chan and Mak (1983), to deal with the problems of estimability and consistency. Questions still remain about how good these estimators are with respect to efficiency. We have investigated the efficiency of such estimators through simulation studies (see Chapter 6).

However, these alternatives are based on linear functional relationship with Gaussian distributed errors, and do not directly apply to our study on combining multiple laser scans, in Chapter 3 and 4, which mainly concerns nonlinear functional models with non-Gaussian errors.

2.6 Normalisation

Two important topics in the analysis of microarray data are the calibration of data from different samples and the problem of variance inhomogeneity, in the sense that the variance of the measured intensity depends on their mean. Due

to variations in sample treatment, labelling, dye efficiency and detection, the fluorescence intensities can in general not be compared directly, but only after appropriate calibration, called "normalisation". Many commonly used statistical methodologies, such as regression or the analysis of variance, are based on the assumption that the data are normally or, at least symmetrically distributed with constant variance, not depending on the mean of the data. If these assumptions are violated, the statistician may choose either to develop some new statistical technique which accounts for the specific ways in which the data fail to comply with the assumptions, or to transform the data. Where possible, data transformation is generally the easier of the two options. A considerable number of techniques have been suggested to date to address the issues of calibration and variance-stabilising transformation of gene-expression microarray data. A brief review of some common methods for location and scale normalisation is given in the next two sections.

2.6.1 Location normalisation

Loess normalisation

The purpose of normalisation is to identify and remove sources of systematic variation in the measured fluorescence intensities and bring the data from different microarrays onto a common scale. The systematic bias arises due to different labelling efficiencies and scanning properties of the Cy3 and Cy5 dyes, different scanning parameters such as PMT settings, print-tip, spatial or plate effects etc. The simplest approach to within-slide normalisation is to subtract a constant from all intensity log-ratios, typically their mean or median. The affine-linear calibration technique of Huber *et al.* (2002, 2003) is also global in nature which normalise the data for the differential behaviour of samples and arrays. Such global normalisation methods can not normalise the intensity data for some locally active artefacts, *e.g.*, print-tip effects, spatial or intensity dependent dye biases. Dudoit *et al.* (2002) proposed more flexible normalisation methods which allow the normalisation function to depend on a number of predictor variables, such as average spot intensity (x), location and plate origin. They used *loess*, a *robust locally weighted regression* (Cleveland, 1979; Cleveland and Devlin, 1988; Cleveland *et al.*, 1993) of the log-ratio (y) on the predictor variables. Suppose that R and G represent the intensity values corresponding to red (Cy5) and green (Cy3) dyes. Within print-tip group intensity dependent normalisation can be performed using the mappings: $\log_2(R/G) \leftarrow \log_2(R/G) - l(x, j)$, where $l(x, j)$ is the loess fit to the scatter plot of y *vs.* x for the spots printed using the jth print-tip,

i.e., data from the jth grid only.

Bayesian and semiparametric approaches

Reilly *et al.* (2003) proposed a Bayesian approach for normalising microarray data. The basic idea is to use genes that are not differentially expressed to conduct the normalisation. The problem is a nontrivial one as one can not determine which genes are differentially expressed until the normalisation is done. However, in this paper, a general framework and computational method using the Gibbs sampler is devised to allow for such a normalisation.

Semiparametric approaches have also been suggested for correcting for trends in log-ratio data which are claimed to relax some of the underlying assumptions in loess normalisation. Fan *et al.* (2005) suggested a semilinear high-dimensional model for within-slide normalisation of microarray data having replicated spots. If there is no within slide replication, within slide replications are artificially created by constructing a super array from the replicate arrays. This way, the model can be used for across-array normalisation as well. Unlike the non-parametric (loess) normalisation, the model is not restricted to the assumption that up-regulated and down-regulated genes at each intensity level are about the same in each print-tip block.

Huang *et al.* (2005) proposed a two-way semilinear model for normalisation and analysis of cDNA microarray data. The semiparametric approach uses polynomial splines to estimate the normalisation curves and the normalised expression values. The method also naturally incorporates uncertainty due to normalisation into significance analysis of microarrays. This method also does not make the usual assumptions underlying some of the existing methods. For example, it does not assume that the percentage of differentially expressed genes is small or that there is symmetry in the expression levels of up-regulated and down-regulated genes as required by the loess normalisation.

Ma *et al.* (2006) proposed a robust semiparametric location and scale model for normalisation and significance analysis purposes. Weighted least absolute deviation regression was used as a robust estimation method. The proposed method naturally combines normalisation and significance analysis, and incorporates the variations due to normalisation into the significance analysis properly.

2.6.2 Variance stabilisation

Generalised logarithm (glog) or inverse hyperbolic sine (arcsinh) transformation

Huber *et al.* (2002, 2003) proposed a statistical model which can address the problem of calibration and variance-stabilising transformation together. In particular, they derived a transformation h for intensity measurements and a difference statistic d (as alternative to log-ratio) whose variance is constant along the whole intensity range. For the transformation h, the parametric form $h(x) = \text{arcsinh}(a+bx)$, a and b being the calibration parameters, was derived from a model of the variance-versus-mean dependence for measuring intensity data using the method of variance stabilising transformations. For large intensities, h coincides with the logarithmic transformation and d with the log-ratio. The difference statistic d as a measure of differential expression having constant variance throughout the entire range of the intensity data seems to be an improvement over log-ratio, variability of which depends on intensity. Another advantage of the proposed technique is that it takes into account the problem of calibration—differential behaviour of dyes, arrays and samples—to bring the measurements on a common scale before the difference statistic is computed. The variance stabilising transformation introduced by Huber *et al.* (2002) was independently derived by Durbin *et al.* (2002) and Munson (2001) and sometimes referred to as *generalised logarithm* or *glog* transformation.

Started logarithm and log-linear hybrid transformation

Rocke and Durbin (2003) suggested two alternative variance stabilising transformations (*started logarithm* and *log-linear hybrid* transformation) that may be easier to use in some applications. In Durbin *et al.* (2002), Huber *et al.* (2002) and Munson (2001), it was shown that for a random variable z satisfying $\text{var}(Z) = a^2 + b^2\mu^2$ with $E(Y) = \mu$ where $Z = Y - \alpha$, there is a transformation (glog transformation) that stabilises the variance to the first order, meaning that the variance is almost constant no matter what the mean might be. One of several equivalent ways of writing this transformation is,

$$f_c(z) = \log\left(\frac{z + \sqrt{z^2 + c^2}}{2}\right),$$

where $c = \frac{a}{b}$. This transformation converges to $\log(z)$ for large z, and is approximately linear at 0 (Durbin *et al.*, 2002). This is exactly $\log(z)$ when $c = 0$ and for that reason it was termed as generalised logarithm or glog by Munson (2001).

While proposing the started logarithm as a variance stabilising transformation Rocke and Durbin (2003) pointed out some limitations of the logarithmic transformation. For a random variable z satisfying $E(Z) = \mu$ and

$$\text{var}(Z) = a^2 + b^2\mu^2,$$

the logarithmic transformation has approximate variance

$$\text{var}\left[\log(Z)\right] \approx b^2 + a^2/\mu^2,$$

which goes to infinity as $\mu \longrightarrow 0$. Furthermore, when $\mu = 0$, z will be frequently non-positive for which the transformation is not defined. A common modification of logarithmic transformation to avoid negative arguments is to add a constant to all of the values before taking logarithm, called the started logarithm, given by

$$g_c(z) = \log(z + c) \quad (c > 0),$$

with approximate variance

$$\text{var}\left[\log(Z)\right] \approx \frac{a^2 + b^2\mu^2}{(\mu + c)^2}. \tag{2.1}$$

This transformation does not completely stabilises the variance when the variance of z is additive and multiplicative. It is however possible to find the value of c which minimises the maximum deviance from constancy. It follows from equation (2.1) that it takes the value a^2/c^2 at $\mu = 0$ and has an asymptote at b^2 as $\mu \longrightarrow \infty$. Rocke and Durbin (2003) focused on the deviation of the variance from the limiting value b^2.

The derivative of (2.1) with respect to μ is

$$\frac{2b^2\mu(\mu + c)^2 - 2(a^2 + b^2\mu^2)(\mu + c)}{(\mu + c)^4}. \tag{2.2}$$

The denominator of (3.2) is never zero for $\mu \geq 0$, so any change in the sign of the derivative will occur where

$$2b^2\mu(\mu + c)^2 - 2(a^2 + b^2\mu^2)(\mu + c) = 0,$$

or,

$$\mu = \frac{a^2}{b^2 c}.$$

It may be noted that the derivative of the variance function at $\mu = 0$ is

$$-2a^2/c^3 < 0,$$

indicating that the variance decreases initially, before increasing again at $\mu = a^2/(b^2c)$. It is clear that the value of c that minimises the maximum deviation of (2.1) from b^2 is where the variance at zero, a^2/c^2, is as much above b^2 as the variance at the minimum is below b^2. Since the minimum is at $\mu = a^2/(b^2c)$, the variance at the minimum is

$$\frac{a^2 + b^2a^4/(b^4c^2)}{(a^2/b^2c + c)^2} = \frac{a^2b^2}{a^2 + b^2c^2}.$$

The condition to minimise the maximum deviation from constant variance is

$$\frac{a^2}{c^2} - b^2 = b^2 - \frac{a^2b^2}{a^2 + b^2c^2},$$

or,

$$c = a/(2^{\frac{1}{4}}b).$$

The achieved minimum deviation is $b^2\sqrt{2} - b^2$ and the ratio of the standard deviation at 0 to the asymptotic standard deviation b is about 1.2.

Rocke and Durbin (2003) considered another variant of logarithmic transformation that may be appropriate for microarray data, called the log-linear hybrid transformation, originally suggested by Holder *et al.* (2001). In this approach, the transformation is taken to be $\log(Z)$ for Z greater than some cutoff k and a linear function $cZ + d$ below that cutoff. This eliminates the singularity at zero. The constants c and d are chosen such that the transformation is continuous with continuous derivative at k. This requires

$$ck + d = \log(k),$$

and

$$c = 1/k,$$

which gives

$$d = \log(k) - 1.$$

The transformation family therefore is

$$f_k(z) = \begin{cases} z/k - \log(k) - 1, & z \leq k \\ \log(z), & z > k. \end{cases}$$

The asymptotic delta-method variance function is given by

$$\text{var}(f_k(Z)) = \begin{cases} (a^2 + b^2\mu^2)/k^2, & Z \leq k \\ b^2 + a^2/\mu^2, & Z > k. \end{cases}$$

The value of k that leads to the minimum deviation form the constant variance is the one for which the variance at $\mu = 0$ (a^2/k^2) is as much below b^2 as the variance at the splice point $(\mu = k)$ is above b^2. Thus,

$$b^2 - a^2/k^2 = (b^2 + a^2/k^2) - b^2,$$

which gives $k = \sqrt{2}a/b$. The motivation behind this transformation is the additive multiplicative variance model of Rocke and Durbin (2001)

$$y = \alpha + \mu e^{\eta} + \varepsilon,$$

which has approximately constant variance for μ close to zero and approximately constant coefficient of variation for μ large.

Mixture model for variance modelling of gene expression

A mixture model approach for the variance of gene expression data was considered by Delmar *et al.* (2005). Their approach can be considered as intermediate between the too stringent homoscedastic models and the over parameterised models assuming specific variance for each gene. The proposed method assumes groups of genes with equal variance and uses a mixture model based on the gene variance distribution.

Let observation y_{gcr} representing expression level of gene g ($g = 1, \cdots, G$) in condition c ($c = 1, 2$) and replicate r ($r = 1, \cdots, n_1 + n_2$), is modelled according to a simple linear model

$$y_{gcr} = \mu_{gc} + \epsilon_{gcr}, \tag{2.3}$$

where ϵ_{gcr} is normally distributed with mean zero. Instead of fitting a separate variance (σ_g^2) for each gene, the paper proposed fitting fewer, say K, where $K < G$, variance parameters assuming that there are groups of genes with equal variance. The authors suggested a mixture model to the distribution of sum of squares of errors to fit these variances.

2.7 Analysis of gene expression data

After the image processing and subsequent normalisation steps, data should reflect only the biological signal of interest plus random noise. Appropriate statistical tools are then applied to answer the biological question under investigation. Statistical analysis of microarray data can be categorised into two broad classes:

(i) Identification of differential expression, and

(ii) Pattern discovery and class prediction.

2.7.1 Identification of differential expression

Many microarray experiments are comparative in nature. Their objective therefore is to compare the expression levels of a set of genes across two or more conditions. This comparison usually involves identifying genes that are significantly

differentially expressed across these conditions. The simplest way to analyse comparative experiments is to consider each gene separately and compare its expression levels across the groups. More complex analysis may involve comparing clusters of genes across conditions. Both formal and informal ways of identifying differential expression are found in the literature. Use of fold change is one of the common informal method of identifying differential expression. Statistical hypothesis testing and regression and analysis of variance types models are also being used for comparing gene expressions across conditions.

Fold changes have been used in early microarray studies (*e.g.*, Schena *et al.*, 1995) to compare differential expressions. A gene is declared differentially expressed if its fold increase or fold decrease exceeds a specified cutoff. For example, in their seminal paper, Schena *et al.* (1995) declared a gene differentially expressed if its expression level showed a fivefold difference between the two mRNA samples. The decision rule that declares, on a logarithmic scale, that changes of h-fold or greater are significant means that a gene should be declared differentially expressed if $|\overline{I}_2 - \overline{I}_1| > \log(h)$, where $\overline{I}_1$ and $\overline{I}_2$ represent the replicate means of the gene expressions of a particular gene in the two samples. Reliance on fold change alone to designate significance has however been criticised as the means estimating true gene expressions are subject to variability. The variability of the estimates can be assessed and should be used to adjust the threshold. This is the idea behind using formal hypothesis testing procedures, *e.g.*, t-test and its modified versions.

Statistical tests are commonly used for inferring differential expressions in comparative microarray studies. The most basic statistical test for comparing two groups is the two-sample t-test. With small samples, t-test statistic tends to be highly correlated with the standard error term that appears in its denominator. As a result the test has a propensity for picking up significant findings at a higher rate from among those genes with low sample variance. Since the sample sizes in the microarray experiments are typically very small, some adjustments of t-test have been suggested. One adjustment was suggested by Tusher *et al.* (2001) by adding a carefully chosen constant, a so-called fudge factor, to the denominator of the t-statistic. This statistic is often called SAM t-statistic where SAM stands for "significance analysis of microarrays". The Mann-Whitney-Wilcoxon rank sum test can be used as an alternative of t-test when the underlying distribution is far from normal (Chambers *et al.*, 1999).

Combining information across genes in the statistical analysis of microarray data is desirable because of the relatively small number of replications for each gene. Cui *et al.* (2005) proposed improved statistical tests for differential gene

expression by shrinking variance components estimates. They suggested an estimator of the error variance that can borrow information across genes using the James-Stein shrinkage concept. The test statistic is constructed using this estimator, and the statistic showed best or nearly best power compared with other statistics, such as, gene-specific F-test, the pooled-variance F-test, a hybrid F-test, the generalised t-statistic, the posterior odds statistic B, and the SAM t-test.

Wei (2006) proposed incorporating existing biological knowledge, such as gene functional annotations, in detecting differential gene expression using stratified mixture models allowing genes with different annotations to have different distributions, such as prior probabilities. Rather than treating parameters in stratified mixture models independently, the author proposed a hierarchical model to take advantage of the hierarchical structure of most gene annotation systems, such as gene ontology. An application to a mouse microarray data set and a simulation study demonstrate the improvement of the new approaches over the standard mixture model.

Bayesian methods for detecting differential expressions have been suggested in some recent papers. Lewin *et al.* (2006) proposed a Bayesian hierarchical model for detecting differentially expressed genes. The method includes simultaneous estimation of array effects, and the authors show how to use the output for choosing list of genes for further investigation. By modelling the array effects (normalisation) simultaneously with differential expression the method reduces the false positive rates.

Another robust Bayesian hierarchical model for testing for differential expression was proposed by Gottardo *et al.* (2006). The model takes account of outliers by explicitly using a t-distribution for the errors, and includes an exchangeable prior for the variances. The model can be used for testing for differentially expressed genes among multiple samples, and it can distinguish between the different possible patterns of differential expression when there are three or more samples. Parameter estimation is carried out using a novel version of Markov chain Monte Carlo. The method performed better than the commonly used techniques, namely, the t-test, the Bonferroni-adjusted t-test, significance analysis of microarrays (SAM) and Efron's empirical Bayes method in an experiment with HIV data.

Hong and Li (2006) proposed a Bayesian approach for detecting differential expression in time-course experiment. A functional hierarchical model was suggested for detecting temporally differentially expressed genes between two experimental conditions for cross-sectional designs, treating gene expression profiles as functional data by basis function expansions. A Monte Carlo EM algorithm

was developed for estimating both the gene-specific and hyperparameters in the second level of modelling. Simulation results suggested that the procedure performs better than the two-way ANOVA in identifying temporally differentially expressed genes.

Adjustment for multiplicity in microarray hypothesis testing is important. Hypothesis testing in microarrays involves performing a large number of tests, one for each gene, and one problem of doing so many tests is that the more the number of tests performed, the higher the overall false positive rate and the higher the expected number of false positives. Several adjustments of p-values, for example, false discovery rate (FDR) (Benjamini and Hochberg, 1995; Yuketieli and Benjamini, 1999) have been suggested to combat the problem. FDR is defined as the expected proportion of false positives among the positive findings. Storey and Tibshirani (2001) proposed a modified version of FDR, called positive false discovery rate (pFDR). The pFDR emphasises the fact that an adjustment is only necessary when there are positive findings.

2.7.2 Pattern discovery and class prediction

In microarray experiments, interest sometimes concerns finding group of genes performing similar functions or genes operating along a genetic pathway. One of the limitations of comparing differential expression on a gene-by-gene basis is that this analysis does not expose or exploit the correlated patterns of gene expression. Performing only gene-by-gene analysis is therefore not sufficient to make use of what should ideally be the full potential of multi-gene experiments. Multivariate methods can be used both for finding multivariate patterns in data, called pattern discovery or unsupervised classification or cluster analysis, and for predicting classes, called class prediction or supervised classification or discriminant analysis. An overview of different types of supervised and unsupervised classification techniques in microarray applications can be found in several recent books on microarray analysis, *e.g.*, Speed (2003), Wit and McClure (2004), Amaratunga and Cabrera (2004), Gentleman *et al.* (2005). Parmigiani *et al.* (2002) proposed a statistical modelling framework for expression-based molecular classification in cancer. The modelling framework can be used to inform and organise the development of exploratory tools for classification. The framework uses latent categories to provide both a statistical definition of differential expression and a precise experiment-independent, definition of a molecular profile. It also generates natural similarity measures for traditional clustering and gives probabilistic statements about the assignment of tumors to molecular profiles. Dudoit *et al.* (2002) compared the performance of different discrimination methods for

the classification of tumors based on gene expression data. The methods include nearest-neighbour classifiers, linear discriminant analysis, and classification trees. Machine learning approaches, such as bagging and boosting, are also considered. Bayesian classification of tumors by using gene expression data was considered by Mallick *et al.* (2005). The paper considers several Bayesian classification methods based on reproducing kernel Hilbert spaces for the analysis of microarray data. A Bayesian mixture model for partitioning gene expression data collected over time was proposed by Zhou and Wakefield (2006). The method assumes a nonparametric random walk model, and partition on the basis of the parameters of the model. The model is flexible and can be tuned to the specific context, respects the order of observations within each curve, acknowledges measurement error, and allows prior knowledge on parameters to be incorporated. The number of partitions may also be treated as unknown, and inferred from the data. Qin and Self (2006) proposed a regression model-based clustering method, which groups genes that share a similar relationship to the covariate(s). The method provides a unified approach for a family of clustering procedures and can be applied for data collected with various experimental designs.

Dynamic modelling of microarray time course data are also suggested (Garcia and Wolkenhauer, 2001) to identify genes with similar dynamic response profiles. Dynamic Bayesian networks (DBNs) are being considered (Kim *et al.*, 2003) as a promising model for inferring gene networks from time series microarray data.

2.8 Scope of thesis

This thesis is concerned with developing statistical methods for the data preprocessing step of microarray analysis. Chapter 1 has described the microarray technology and the underlying biology behind the technology. Chapter 2 gives a brief overview of the statistical issues in the design, image processing, data preprocessing and analysis of microarray experiments. Chapters 3 and 4 are concerned with our proposed statistical models for estimating gene expression using multiple laser scans of hybridised microarrays. In Chapter 3, experimental results on finite mixture modelling and hyperbolic and censored mean functional regression approaches of combining multiple laser scans of microarray data are described. While in Chapter 4, we present a refined statistical model for this problem based on a censored Cauchy model to account for the outlying observations and also for the fact that spot averages cannot exceed the censoring threshold T. The model is capable of estimating gene expression adjusting for signal censoring and random outliers in the intensity measurements.

In Chapter 5, we suggest a new nonparametric normalisation method of microarray data. The method incorporates location and scale normalisation simultaneously using Generalised Additive Models for Location, Scale and Shape (GAMLSS, Rigby and Stasinopoulos, 2005) as alternative to the parametric approaches of variance stabilisation.

In Chapter 6, we compare the efficiency of different estimation methods for functional regression models through simulation studies. The motivation behind this investigation is the fact that the parameters of functional regression models, as seen in Chapters 3 and 4, are often not estimable by direct application of maximum likelihood estimation, and sometimes lead to inconsistent estimators when they are estimable. Alternative methods of estimation are suggested in the literature to address the estimability and consistency problems, but questions still remain about how good these estimators are in terms of efficiency.

Finally, Chapter 7 concludes the thesis, providing an overall discussion and indication for possible further research on the work presented in this thesis.

Chapter 3

Combining multiple laser scans: exploratory analysis

3.1 Introduction

Although microarrays are often scanned more than once, common practice is to use data from a single scan, based on some arbitrary exploratory checks, in the subsequent analysis. Use of combined data from multiple scans of hybridised microarray may be useful to get improved gene expression measurements.

In this thesis, one of our aims is to use data from multiple scans at different scanner settings to handle pixel censoring, also known as signal saturation, in the gene expression measurements.

Only handful of methods are available (see Section 2.4) for combining multiple scans, and none of them are fully adequate to address the problem. For example, algorithmic approaches of Dudley *et al.* (2002) and Bell (2003) are not based on full information of multiple scan data. The methods discard information outside the linear range, and there is arbitrariness and subjectivity involved in choosing such a range. Furthermore, the methods consider standard linear regressions to combine multiple laser scans, which may be misleading because each individual scan of data are subject to measurement errors.

Method of García de la Nava *et al.* (2004) is limited to only two scans at different sensitivities. Also, the use of either linear function or power function to relate the measurements of two scans does not seem very realistic.

Our aim is to use functional regression models, which are more appropriate for multi-scan microarray data as they allow measurement errors in both response and explanatory variables. We also plan to use the full information in multi-scan data to suggest a more elegant and objective way of combining multiple laser scans.

In this chapter we explore the patterns of multiple scan data, and investigate

and evaluate three different approaches:

(i) Functional regression model with Gaussian mixture distribution (Section 3.4)

(ii) Hyperbolic functional regression model based on Cauchy likelihood (Section 3.5), and

(iii) Censored mean functional regression model based on M-estimation and t-likelihood (Section 3.6)

for estimating gene expression from multiple laser scans of microarrays. Based on our exploratory analysis in this chapter, we then propose our final model, in Chapter 4, for estimating gene expression using multiple laser scans of microarrays.

3.2 Motivation

The sensitivity level of microarray scanners is adjustable and plays a crucial role in getting reliable measurement of the fluorescence intensity. In an ideal situation, when there is no censoring or deterioration, a change in scanner setting should transform the intensity measurements by a multiplicative constant. That is, the average relation between the data from any two scans should be a straight line through the origin. A scanner's sensitivity has to be raised to a certain level to ensure that the intensity levels of weakly expressed genes exceed the intrinsic noise level of the scanner and so become measurable. This may, however, cause another problem: signal censoring for highly expressed genes. Scanners cannot record pixel intensities above some software dependent threshold, which is $2^{16} - 1 = 65535$, for a 16-bit computer storage system. So highly expressed genes can have pixel values which are right censored at the largest possible value that the scanner software allows, and the linear relation is distorted. As the problem at the two ends are in conflict, no unique scanner setting is optimal. Moreover, there is no objective guideline to date for choosing optimum scanner setting to address these issues. It therefore seems reasonable to consider multiple scanning, some at relatively lower settings ensuring that there is no censoring at the upper end, and the others at higher settings ensuring the visibility of the weakly expressed genes over the scanner's intrinsic noise level, and combine the information together to get final gene expression measures. Combining the data through simple or weighted average over the scans is likely to give biased result as the data are not generally proportional to the true expression levels over the entire range.

Our attempt therefore is to suggest a statistical model based on multiple scan data to improve gene expression estimates over the entire range of intensity level, adjusting for signal censoring at the upper end.

3.3 Murine macrophage data

Before introducing the models it seems relevant to explain some patterns of multiple-scan microarray data. For this explanation and investigation of the prospective models, we use a data set kindly provided by the Scottish Centre for Genomic Technology and Informatics (GTI), which is a post-genomic research centre located within the University of Edinburgh Medical School. The experiment was designed to examine the effects of ingestion of apoptotic cells on murine macrophage gene expressions 24 hours after administration and compare this expression profile against control untreated cells. We shall call this data as 'murine macrophage data' throughout this thesis. There are two arrays of data, one of which is a dye-swap of the other. Each array represents 4624 genes, each of which has been replicated twice within the same array. Total number of spots on each of the arrays is therefore 9248. The arrays were scanned with an Affymetrix 428 scanner at four different sensitivity levels for each of the Cy3 and Cy5 dyes. We have used the data of channel 1 from both arrays (arrays 1 and 2) where the control and treated samples were labeled with (Cy3 and Cy5) and (Cy5 and Cy3) respectively. For both arrays, the observed spot intensities for the ith spot at the jth setting are denoted by y_{ij}.

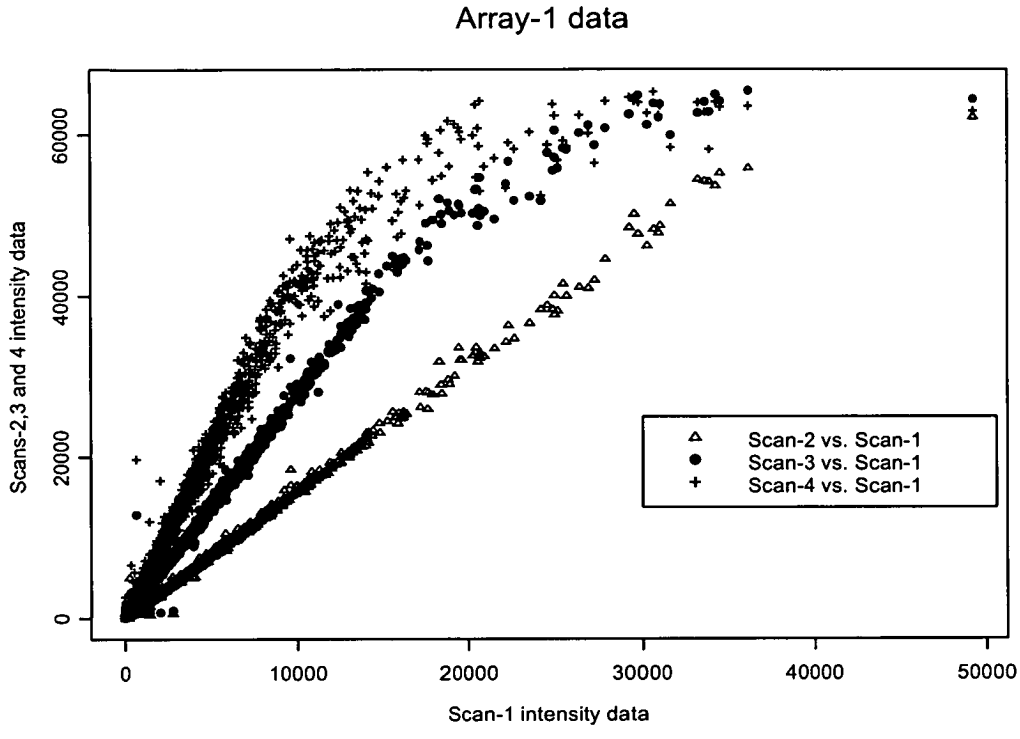
In the absence of any pixel censoring one scan of data, on the average, should just be a multiple of the other. That is, the average relation should be a straight line through the origin. The scatterplots of scans 2, 3 and 4 against scan 1 intensity data in Figure 3.1 show how pixel censoring affects this relationship. The relationships appear linear within the lower range of the intensity data. The limit of the range within which data are linear varies with the sensitivity level of the scanner. The higher the difference between the scanner sensitivity of scan j ($j = 2, 3, 4$) and scan 1, the narrower is the range of linear relationship of scan j vs. scan 1 data. For example, relationship of scan 2 vs. scan 1 data are linear within a broader range than that of scan 3 vs. scan 1 or scan 4 vs. scan 1. As the data of scan 1 are least likely to be affected by pixel saturation, and hence should be approximately proportional to the true expression levels over the entire range, the same conclusion holds about the relationship of any observed intensity data and the true gene expression levels. It can be seen from Figure 3.1 that departure from linearity starts well before the threshold value of 65535. This is

because of the fact that the observed intensity measurement for a particular spot is obtained by taking the average of a set of pixels belonging to that spot, and spots representing highly expressed genes generally contain both censored and uncensored pixels.

Suppose that $y_{.j}$ represents the vector of intensity data across all genes from scan j for $j = 1, 2, \cdots, m$. The *Least Trimmed Squares* (LTS, Rousseeuw and Leroy, 1987) residuals from the simple regression

$$y_{.4} = \beta y_{.1} + e,$$

of scan 4 ($y_{.4}$) on scan 1 ($y_{.1}$) data for array 1 are plotted against scan 1 data in Figure 3.2. It is noted that the majority of the data points belong to the lower range of the intensity and the downward tendency of the residuals roughly after $y_{.1} = 10000$ is clearly due to the bias effect of the pixel censoring. To have a closer look at the residual mean and variability for the main body of the data, a portion, indicated by the rectangle, of the top panel of Figure 3.2 has been magnified in the bottom panel. It is observed that residuals have non-zero variability near the origin ($y_{.1} = 0$) around the horizontal reference (zero) line. Residual variability then increases with the level of intensity ($y_{.1}$) giving rise to a funnel-like shape to the plot. It is therefore reasonable to model error variance as having both additive and multiplicative components. This additive-multiplicative nature of error variability for microarray data has been noted previously by Ideker *et al.* (2000), Rocke and Durbin (2001), Huber *et al.* (2002, 2003) and Rocke and Durbin (2003).



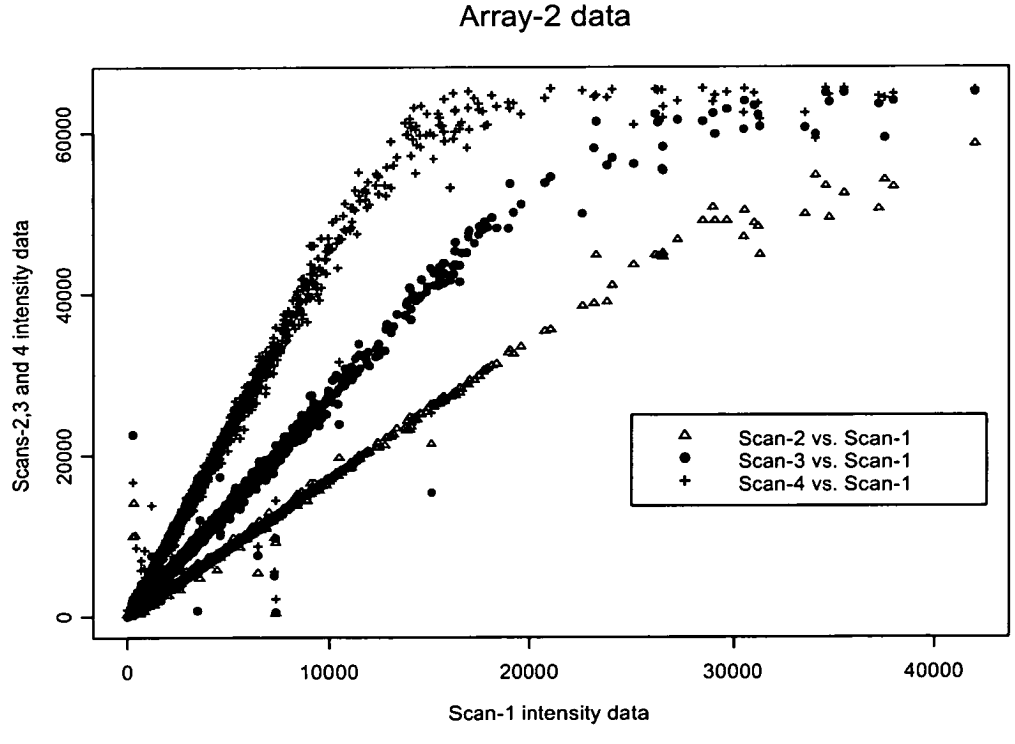


Figure 3.1: Scatterplots of scans-2, 3 and 4 vs. scan-1 intensity data.

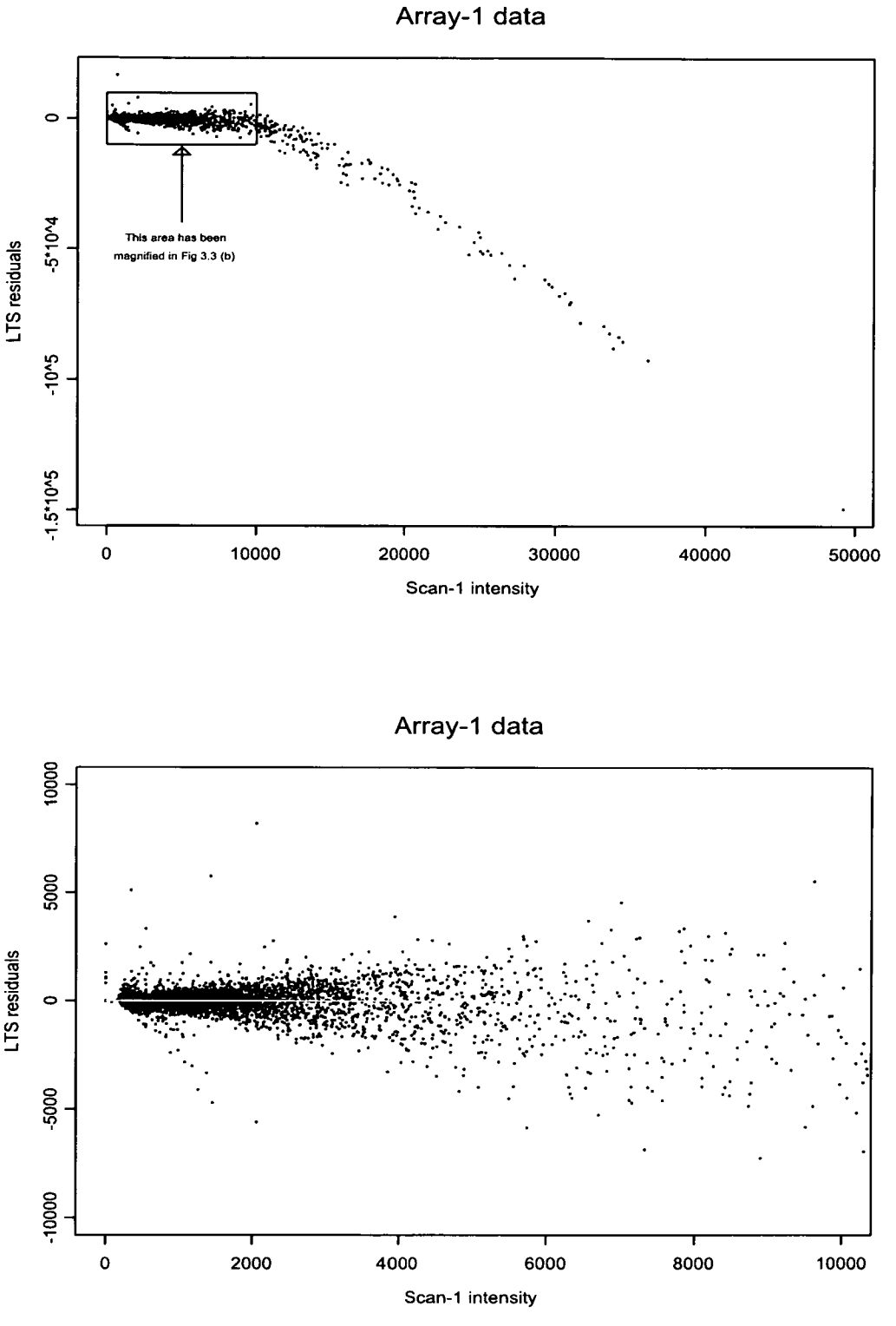


Figure 3.2: LTS residuals vs. $y_{.1}$ for the model $y_{.4} = \beta y_{.1} + e$ (top); magnified portion of the top panel (bottom) for array 1 data.

Normal Q-Q plot of the standardised LTS residuals for the main body of the array 1 data from the simple model $y_{.4} = \beta y_{.1} + e$ is shown in Figure 3.3. This plot suggests a distribution for errors with heavier tails than that of Gaussian. The patterns of multiple scan data as described above suggest

(i) a nonlinear relationship of the observed intensity with the true expression levels,

(ii) additive plus multiplicative variance model for the errors, and

(iii) heavy-tailed distributions for the errors.

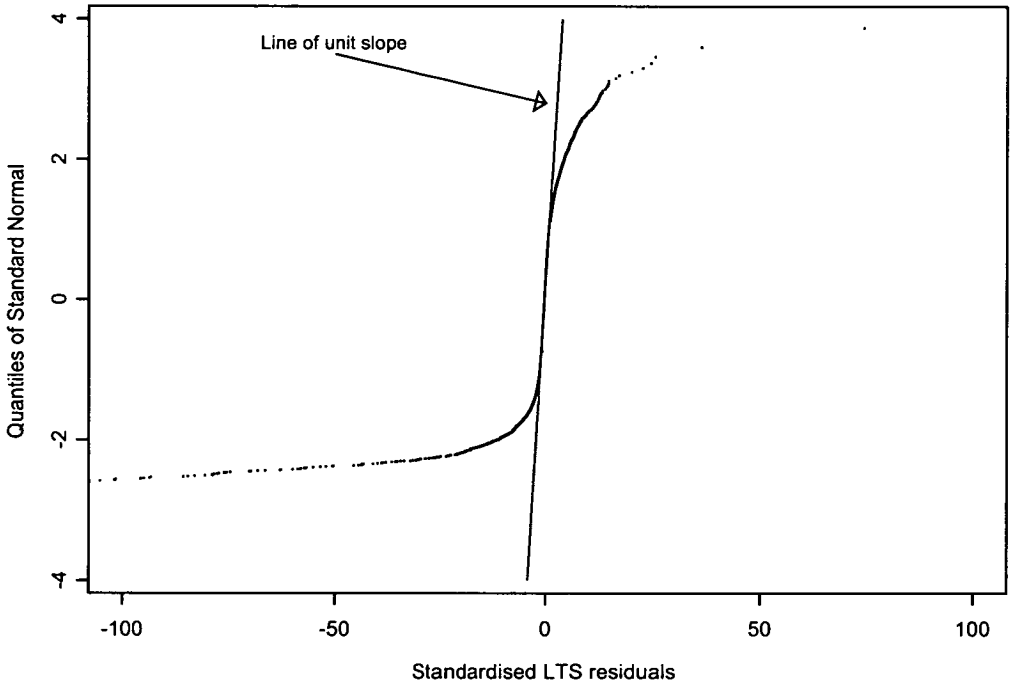


Figure 3.3: Normal Q-Q plot of standardised LTS residuals from the regression model $y_{.4} = \beta y_{.1} + e$ for array 1 data.

3.4 Linear functional regression with Gaussian mixture distribution

Although the relationship of the observed spot intensity with the expression levels appears nonlinear, we start with a linear functional regression model, detailed in

Cheng and Van Ness (1999, Chapter 1), assuming Gaussian mixture distribution for the errors. We considered same linear location but different variance models for the two components of Gaussian mixture. The additive plus multiplicative variance model, in the first component of the mixture, is intended to represent the main body of the data showing approximately a linear relationship with the gene expression levels. The variance of the second component of the mixture is assumed to be a constant, probably very large, to model the nonlinearity in the contaminated region as noise. The idea therefore is to fit a linear model, which appears reasonable for the main body of the data, capturing the nonlinearity at the upper end by a high dispersion parameter of the second component of Gaussian mixture.

First we explore the pairwise ordinary regressions with Gaussian mixture distribution of errors as described above to investigate the suitability of the functional model.

3.4.1 Pairwise regression models

We describe the pairwise Gaussian mixture regression relationships in terms of five parameters: slope parameter (β), additive and multiplicative scales (σ_1 and σ_2) of the first component of the mixture, scale parameter (τ) of the second component of the mixture and the mixing proportion parameter (π).

Suppose that the same microarray has been scanned several (say, m) times at different scanner sensitivity levels. Let (y_{ij}, y_{ih}); $j, h = 1, 2, 3, 4$ $(j > h)$, $i = 1, \cdots, n$ be the n pairs of observations corresponding to any two scans where n is the total number of spots on the array. The linear regression relation of $Y_{.j}$ on $Y_{.h}$ can be described as

$$y_{ij} = \beta y_{ih} + \epsilon_i, \tag{3.1}$$

where ϵ_i is distributed according to

$$f(\epsilon_i) = \frac{(1-\pi)}{\sqrt{\sigma_1^2 + \sigma_2^2(\beta y_{ih})^2}} \phi\left(\frac{y_{ij} - \beta y_{ih}}{\sqrt{\sigma_1^2 + \sigma_2^2(\beta y_{ih})^2}}\right) + \frac{\pi}{\tau}\phi\left(\frac{y_{ij} - \beta y_{ih}}{\tau}\right), \tag{3.2}$$

with $\phi(.)$ being the density function of a standard normal variable. Assuming independence of the ϵ_i's the log-likelihood function for estimation of the parameters of model (3.1) can be expressed as

$$L(\beta, \sigma_1, \sigma_2, \tau, \pi) = \sum_{i=1}^{n} \log\left\{\frac{(1-\pi)}{\sqrt{\sigma_1^2 + \sigma_2^2(\beta y_{ih})^2}} \phi\left(\frac{y_{ij} - \beta y_{ih}}{\sqrt{\sigma_1^2 + \sigma_2^2(\beta y_{ih})^2}}\right) + \frac{\pi}{\tau}\phi\left(\frac{y_{ij} - \beta y_{ih}}{\tau}\right)\right\} \tag{3.3}$$

Table 3.1: Maximum likelihood estimates of the parameters of model (3.1) applied to murine macrophage data.

Pair $(Y_{.j}, Y_{.h})$	Estimates (standard errors) of the parameters				
	β	σ_1	σ_2	τ	π
$(Y_{.2}, Y_{.1})$	1.56 (0.0007)	43 (0.57)	0.024 (0.0005)	751 (53)	0.017 (0.0019)
$(Y_{.3}, Y_{.1})$	2.75 (0.0015)	82 (1.05)	0.029 (0.0005)	12100 (814)	0.012 (0.0012)
$(Y_{.4}, Y_{.1})$	4.26 (0.0034)	111 (2.38)	0.059 (0.0008)	32100 (1950)	0.016 (0.0014)
$(Y_{.3}, Y_{.2})$	1.76 (0.0051)	68 (0.74)	0.013 (0.0003)	9990 (574)	0.017 (0.0015)
$(Y_{.4}, Y_{.2})$	2.74 (0.0023)	71 (2.85)	0.071 (0.0007)	11300 (835)	0.015 (0.0014)
$(Y_{.4}, Y_{.3})$	1.56 (0.0010)	105 (1.84)	0.046 (0.0062)	16300 (852)	0.022 (0.0017)

Maximum-likelihood estimates of the parameters was obtained through numerical maximisation of the log-likelihood function (3.3). The optimisation algorithm of Nelder and Mead (1965) has been implemented using FORTRAN 90 and IMSL routine `DUMPOL` for this purpose. The algorithm minimises a function $g(\theta)$ over p parameters using a direct search polytope algorithm. The polytope method is based on function comparison. It starts with $p+1$ points $\theta_1, \cdots, \theta_{p+1}$. At each iteration, a new point is generated to replace the worst point θ_j, which has the largest function value among the $p+1$ points. We have chosen this algorithm because it does not require the expressions for the score and information functions, which are quite tedious to derive for the mixture distribution we considered. Standard errors of the parameter estimates can be approximated from the diagonal elements of the inverse of observed information matrix. We have evaluated the information matrix through numerical differentiation. Results of applying this technique to murine macrophage data are summarised in Table 3.1. The data set has four columns labeled $Y_{.1}$, $Y_{.2}$, $Y_{.3}$ and $Y_{.4}$, in ascending order of scanner sensitivity, corresponding to four scans. We have considered six regression models corresponding to the pairs $(Y_{.j}, Y_{.h})$; $j, h = 1, 2, 3, 4$ $(j > h)$. Sufficiently small, relative to the estimates, standard errors indicate clear evidence of statistical significance of the parameters. It may be noted that there is a systematic pattern (Table 3.2) in the regression and scale estimates with respect to the label of the variables, *i.e.*, order of scanner sensitivity. For a particular $Y_{.h}$, slope and all scale parameters increase with the increase in the scanner sensitivity associated with the $Y_{.j}$ variable.

3.4.2 Multivariate functional regression model

The results of pairwise linear regression model (3.1), provide convincing evidence of suitability of the assumed functional model. Replacing x by the the gene expression parameters (μ) and generalising the model to incorporate information

Table 3.2: Systematic pattern in estimates.

	β			σ_1			σ_2			τ		
	$Y_{.h}$			$Y_{.h}$			$Y_{.h}$			$Y_{.h}$		
$Y_{.j}$	$Y_{.1}$	$Y_{.2}$	$Y_{.3}$	$Y_{.1}$	$Y_{.2}$	$Y_{.3}$	$Y_{.1}$	$Y_{.2}$	$Y_{.3}$	$Y_{.1}$	$Y_{.2}$	$Y_{.3}$
$Y_{.2}$	1.56			43			0.024			751		
$Y_{.3}$	2.75	1.76		82	68		0.029	0.013		12100	9990	
$Y_{.4}$	4.26	2.74	1.56	111	71	105	0.059	0.071	0.046	32100	11300	16300

from all m scans, the linear functional model can be defined as

$$y_{ij} = \mu_i \beta_j + \epsilon_{ij}, \tag{3.4}$$

where y_{ij} is the intensity measure of the ith gene at the jth scan, μ_i is the true gene expression parameter of the ith gene, β_j ($\beta_1 = 1$ being the identifiability constraint) is the jth scanner's effect and ϵ_{ij} is the random error term distributed according to

$$f(\epsilon_{ij}) = \frac{(1-\pi_j)}{\sqrt{\sigma_{1j}^2 + \mu_i^2 \sigma_{2j}^2}} \phi\left(\frac{y_{ij} - \mu_i \beta_j}{\sqrt{\sigma_{1j}^2 + \mu_i^2 \sigma_{2j}^2}}\right) + \frac{\pi_j}{\tau_j} \phi\left(\frac{y_{ij} - \mu_i \beta_j}{\tau_j}\right). \tag{3.5}$$

The parameters σ_{1j} and σ_{2j} are the additive and multiplicative scales respectively for the first component of the mixture distribution, τ_j is the scale parameter of the second component of the mixture and π_j is the mixing proportion.

A model of the form (3.4)–(3.5) can be regarded as a mixture version of the class of multivariate measurement error (ME) models. Depending on the assumption about μ, ME model has three different subclasses (Cheng and Van Ness, 1999, Chapter 1). If the μ_i's are unknown constants, the model is known as a *functional model*; whereas, if the μ_i's are i.i.d. random variables and independent of the errors, the model is known as a *structural model*. A third type of model, known as *ultrastructural model*, assumes that μ_i's are independent random variables, as in the structural model, but not identically distributed.

With μ_i as a latent Gaussian variables, model (3.4)–(3.5) is a mixture version of one dimensional factor analysis model (Mardia, Kent and Bibby, 1979, Exercise 9.2.7, p. 277). We are considering μ_i as unknown constant. Model (3.4)–(3.5) is then a mixture version of multivariate functional model.

3.4.3 Maximum likelihood estimation

The main challenge of working with this model is the estimation of large ($n + 5m - 1$) number of parameters which increases with the number of spots (n) and

the number of scans (m). We consider an alternating algorithm for estimating the parameters of the model (3.4)–(3.5) through maximum likelihood method as described below.

The log-likelihood function (log(LF)) under independence assumption can be expressed as

$$L(\mu, \beta, \sigma_1, \sigma_2, \tau, p) = \sum_{i=1}^{n} \sum_{j=1}^{m} \log\{l(\mu_i, \beta_j, \sigma_{1j}, \sigma_{2j}, \tau_j, \pi_j)\}, \tag{3.6}$$

where the function $l(.)$ has the same form as $f(.)$ in equation (3.5). The log-likelihood function (3.6) can be maximised numerically with respect to the parameters using the following algorithm:

1. Choose initial values of all the parameters, $(\mu^{(0)}, \beta^{(0)}, \sigma_1^{(0)}, \sigma_2^{(0)}, \tau^{(0)}, \pi^{(0)})$ where $\mu^{(0)}$ is a vector of dimension n, $\beta^{(0)}$ is a $(m-1)$-vector and $\sigma_1^{(0)}, \sigma_2^{(0)}, \tau^{(0)}, \pi^{(0)}$are all m-vectors.

2. Update all μ_i, $(i = 1, 2, \cdots, n)$ individually by maximising $\sum_{j=1}^{m} l(\mu_i, \beta_j^{(0)}, \sigma_{1j}^{(0)}, \sigma_{2j}^{(0)}, \tau_j^{(0)}, \pi_j^{(0)})$. Denote the updated μ_i's by $\mu^{(1)} = (\mu_1^{(1)}, \mu_2^{(1)}, \cdots, \mu_n^{(1)})$.

3. Update $(\beta_j, \sigma_{1j}, \sigma_{2j}, \tau_j, \pi_j)$ for every j separately by maximising $\sum_{i=1}^{n} l(\beta_j, \sigma_{1j}, \sigma_{2j}, \tau_j, \pi_j, \mu^{(1)})$. Denote the updated estimates as $(\beta_j^{(1)}, \sigma_{1j}^{(1)}, \sigma_{2j}^{(1)}, \tau_j^{(1)}, \pi_j^{(1)})$.

Repeat steps (2) and (3) replacing previous estimates by the updated ones until convergence. We have tried to estimate the parameters of the model (3.4)–(3.5) using the above algorithm for the murine macrophage data . The effort has, however, not been successful. As the iteration proceeds the log-likelihood function continues to increase and the scale parameters for one of the scans is driven down to zero. Some iterations of the algorithm applied to murine macrophage data is displayed in Table 3.3. The simplex method of Nelder and Mead (1965) has been used as the optimisation tool. It is seen that by iteration 13 estimates of σ_1 and σ_2 for the second scan are approximately zero. We have encountered the same kind of problem for the simulated data as well. Cheng and Van Ness (1999, Chapter 1) argued that the likelihood function is actually unbounded, which was first shown by Anderson and Rubin (1956). To show the unboundedness of the likelihood function Cheng and Van Ness (1999, Chapter 1) considered a simpler version of the model (3.4)–(3.5), namely $y_{ij} \sim N(\mu_i \beta_j, \sigma_j^2)$. The log-likelihood function is

$$L(\mu_1, \cdots, \mu_n, \beta_2, \cdots, \beta_m, \sigma_1^2, \cdots, \sigma_m^2) \propto -\frac{n}{2} \sum_{j=1}^{m} \left(\log \sigma_j^2 + \frac{S_j}{\sigma_j^2} \right) \tag{3.7}$$

Table 3.3: Some iterations of maximum likelihood algorithm for estimating the parameters of model (3.4)–(3.5) applied to the murine macrophage data.

Iteration	j	Parameter (other than μ) estimates β_j	σ_{1j}	σ_{2j}	τ_j	$\pi_j(\%)$	log(LF)
	1	1.00	40.0	.010	1000	1.00	
Initial	2	1.56	50.0	.020	2000	1.00	–182295
	3	2.75	60.0	.030	3000	1.00	
	4	4.24	70.0	.040	4000	1.00	
	1	1.00	25.1	.010	2580	1.57	
1	2	1.56	27.7	.012	4570	1.26	–172526
	3	2.74	41.9	.016	6890	0.80	
	4	4.28	27.3	.041	20000	2.11	
	1	1.00	23.7	.012	2360	1.37	
2	2	1.56	23.9	.008	4360	1.19	–170648
	3	2.75	39.9	.013	7000	0.84	
	4	4.29	24.2	.044	19800	2.23	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	1	1.00	28.3	.023	2850	1.18	
9	2	1.56	0.9	.000	4280	1.51	–159911
	3	2.75	65.0	.013	6550	0.87	
	4	4.30	96.4	.048	21700	1.92	
	1	1.00	28.8	.023	2840	1.18	
10	2	1.56	0.03	.000	4290	1.51	–147073
	3	2.75	65.0	.013	6550	0.87	
	4	4.30	96.4	.048	21700	1.92	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	1	1.00	28.8	.023	2840	1.18	
13	2	1.56	0.0	.000	4290	1.51	–59589
	3	2.75	65.0	.013	6550	0.87	
	4	4.30	96.4	.048	21700	1.92	

where $S_j = \sum_{i=1}^{n}(y_{ij} - \mu_i\beta_j)^2/n$. Defining $h(a,b) = \log a + b/a$, (3.7) can be written as

$$-\frac{2}{n}L \propto h(\sigma_1^2, S_1) + \cdots + h(\sigma_m^2, S_m). \tag{3.8}$$

Now $h(a,0) = \log a$, so that $h(a,0) \to -\infty$ as $a \to 0$. Thus (3.8), the sum of m such functions, will tend to minus infinity if any of the functions on the right hand side does so. Now consider the values $\mu_i = y_{i1}$, $i = 1, \cdots, n$; $\sigma_1^2 \to 0$. This makes $S_1 = 0$ and $2L/n \to \infty$, and L itself tend to infinity irrespective of the values of other parameters.

Copas (1972) however showed that the likelihood function is bounded when account is taken of the rounding-off errors in the observations. The author showed that a solution to the appropriate likelihood can be found which is approximately the maximum likelihood estimate. Copas' (1972) argument is based on the fact that the likelihood functions such as (3.7) are only approximations to the likelihood functions based on observed data. The approximation is good enough provided that the grouping error is small compared with the underlying variability, but is invalid in the neighbourhood of parameter points which give zero values to some or all of the standard deviations in the model. To take account of the rounding-off errors, Copas (1972) assumed that the observations have infact been recorded to within an accuracy of $h/2$, *i.e.*, a grouping interval of length h. The likelihood function can be represented as

$$L = \prod_{i=1}^{n}\prod_{j=1}^{m} P_{ij}(y_{ij}),$$

where

$$P_{ij}(y) = P(y - h/2 \leq y_{ij} < y + h/2).$$

When $\sigma_j > 0$, $P_{ij}(y)$ can be written as

$$P_{ij}(y) = \Phi\left(\frac{y - \mu_i\beta_j + h/2}{\sigma_j}\right) - \Phi\left(\frac{y - \mu_i\beta_j - h/2}{\sigma_j}\right),$$

where Φ denotes the standard normal distribution function. When $\sigma_j = 0$, $P_{ij}(y)$ takes the value 1 if $y - h/2 \leq \mu_i\beta_j < y + h/2$ and 0 otherwise. The likelihood is therefore bounded and uniquely defined at all points in the parameter space. However, the solutions to the appropriate likelihood defined above are not the exact maximum likelihood estimates, and are not consistent estimators. These illustrates the problems with application of maximum likelihood estimation in this context.

3.4.4 Alternative estimation

As it does not appear to be possible to estimate the parameters of the proposed model directly through maximum likelihood, we tried an alternative algorithm using a combination of likelihood and quasi-likelihood. According to the model (3.4)–(3.5), for any $j \neq h$ $(j, h = 1, 2, \cdots, m)$ the random variable

$$d_{ijh} = \left(y_{ij} - \frac{\beta_j}{\beta_h} y_{ih} \right) \tag{3.9}$$

is distributed according to a mixture of four normal distributions with common mean zero and different variance parameters. The p.d.f $g(.)$ of d_{ijh} can be formulated as

$$\begin{aligned} g(d_{ijh}) = & \frac{(1-\pi_j)(1-\pi_h)}{\sqrt{\sigma_{1j}^2 + (\beta_j/\beta_h)^2\sigma_{1h}^2 + \mu_i^2(\sigma_{2j}^2 + (\beta_j/\beta_h)^2\sigma_{2h}^2)}} \\ & \phi\left(\frac{y_{ij} - (\beta_j/\beta_h)y_{ih}}{\sqrt{\sigma_{1j}^2 + (\beta_j/\beta_h)^2\sigma_{1h}^2 + \mu_i^2(\sigma_{2j}^2 + (\beta_j/\beta_h)^2\sigma_{2h}^2)}} \right) + \\ & \frac{(1-\pi_j)\pi_h}{\sqrt{\sigma_{1j}^2 + (\beta_j/\beta_h)^2\tau_h^2 + \mu_i^2\sigma_{2j}^2}} \phi\left(\frac{y_{ij} - (\beta_j/\beta_h)y_{ih}}{\sqrt{\sigma_{1j}^2 + (\beta_j/\beta_h)^2\tau_h^2 + \mu_i^2\sigma_{2j}^2}} \right) + \\ & \frac{\pi_j(1-\pi_h)}{\sqrt{\tau_j^2 + (\beta_j/\beta_h)^2\sigma_{1h}^2 + \mu_i^2(\beta_j/\beta_h)^2\sigma_{2h}^2}} \phi\left(\frac{y_{ij} - (\beta_j/\beta_h)y_{ih}}{\sqrt{\tau_j^2 + (\beta_j/\beta_h)^2\sigma_{1h}^2 + \mu_i^2(\beta_j/\beta_h)^2\sigma_{2h}^2}} \right) + \\ & \frac{\pi_j\pi_h}{\sqrt{\tau_j^2 + (\beta_j/\beta_h)^2\tau_h^2}} \phi\left(\frac{y_{ij} - (\beta_j/\beta_h)y_{ih}}{\sqrt{\tau_j^2 + (\beta_j/\beta_h)^2\tau_h^2}} \right). \end{aligned} \tag{3.10}$$

The log-quasi-likelihood function (log(Q-LF)) for the parameters of the model (3.4)–(3.5) can be defined, under independence assumption, as

$$QL(\mu, \beta, \sigma_1, \sigma_2, \tau, \pi) = \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{h \neq j=1}^{m} \log\{g(d_{ijh})\} \tag{3.11}$$

The alternative algorithm using the combination of log-likelihood function (3.6) and the log-quasi-likelihood function (3.11) is described below:

1. Give starting values, $(\mu^{(0)}, \beta^{(0)}, \sigma_1^{(0)}, \sigma_2^{(0)}, \tau^{(0)}, \pi^{(0)})$.

2. Update all μ_i $(i = 1, 2, \cdots, n)$ individually by maximising $\sum_{j=1}^{m} l(\mu_i, \beta_j^{(0)}, \sigma_{1j}^{(0)}, \sigma_{2j}^{(0)}, \tau_j^{(0)}, \pi_j^{(0)})$. Denote the updated μ_i's by $\mu^{(1)} = (\mu_1^{(1)}, \mu_2^{(1)}, \cdots, \mu_n^{(1)})$.

3. Update the ($5m$-1) parameters in $(\beta, \sigma_1, \sigma_2, \tau, \pi)$ by maximising $QL(\mu^{(1)}, \beta, \sigma_1, \sigma_2, \tau, \pi)$.

Continue repeating steps (2) and (3) replacing the previous estimates by the updated ones until convergence, *i.e.*, the gain in the L and QL is negligible, is achieved. When applied to murine macrophage data, the algorithm terminates successfully but with unrealistic estimates for some of the parameters. The estimates also appear to depend on the initial choice of the parameters, particularly the values of mixing proportions, π_j. Results for two different sets of initial parameters are tabulated in Table 3.4.

Table 3.4: Results of alternative algorithm applied to murine macrophage data for two sets of initial values.

Set of initial	Estimates	j	Parameter (other than μ)estimates β_j	σ_{1j}	σ_{2j}	τ_j	$\pi_j(\%)$	log(LF) log(Q-LF)
1	Initial values	1	1.00	40.0	.020	500	1.00	
		2	1.56	50.0	.030	1000	1.00	−183703
		3	2.75	60.0	.080	2000	1.00	−597423
		4	4.26	70.0	.250	3000	1.00	
	Final estimates (8th iteration)	1	1.00	22.4	.024	394	1.42	
		2	1.55	6.4	.000	100	12.15	−160765
		3	2.74	52.2	.034	14000	0.71	−573488
		4	4.28	65.7	.203	32900	1.17	
2	Initial values	1	1.00	40.0	.020	500	1.00	
		2	1.56	50.0	.030	1000	2.00	−183980
		3	2.75	60.0	.080	2000	3.00	−599359
		4	4.26	70.0	.250	3000	4.00	
	Final estimates (10th iteration)	1	1.00	23.4	.024	7390	0.68	
		2	1.55	26.2	.016	10700	0.78	−169762
		3	2.74	41.2	.034	615	2.35	−575109
		4	4.31	91.9	.316	29	40.81	

The only difference between the two sets of initial values is in terms of π_j's. For the first set all π_j's have been initialised at 1 percent; whereas π_1, π_2, π_3 and π_4 in the second set of initial values are chosen to be 1, 2, 3 and 4 percent respectively. This produces dramatic changes in the final estimates of τ's.

3.4.5 Simulation study

As the alternative algorithm does not give satisfactory results when applied to real data, we consider doing some simulation studies to investigate the problem in more detail. We generate four data sets according to model (3.4)–(3.5) for

different combinations of the parameters as given in Table 3.5. Each of the data set corresponds to different starting seed of the random number generator. The purpose of this simulation is to check if the problem is with

(i) the estimation algorithm,

(ii) the program code, or

(iii) the model itself.

The results of this simulation are summarised in Table 3.5. It is seen that for all four sets of simulated data, the estimated parameters ($\hat{\mu}_i$ not shown here) by the alternative algorithm are very close to the true values. Reproducibility of the model parameters from the the simulated data generally suggests the validity of the program code and estimation algorithm. We therefore suspect that either the model is not adequate to address the problem or the algorithm is not robust to the lack of fit of the model.

3.5 Hyperbolic functional regression model

The linear functional regression model with Gaussian mixture distribution has been found to be inadequate (previous section) for combining multiple laser scans, and we focus our search to nonlinear functional models. Plots of data in Figure 3.1 motivated us to consider a hyperbolic function to describe the location of the data. We derive a hyperbolic function as a function of the scanning effect and gene expression parameters as described in the following subsection.

Table 3.5: Results of alternative algorithm applied to simulated data.

Data Set	Estimates	j	Parameter (other than μ) estimates β_j	σ_{1j}	σ_{2j}	τ_j	$\pi_j(\%)$	log(LF) log(Q-LF)
1	True values	1	1.00	40.0	.020	500	1.00	
		2	1.56	50.0	.030	1000	1.00	−178203
		3	2.75	60.0	.040	2000	1.00	−585197
		4	4.26	70.0	.050	4000	1.00	
	Final estimates (5th iteration)	1	1.00	30.3	.020	493	0.95	
		2	1.56	40.6	.029	1010	0.79	−173694
		3	2.75	49.2	.041	2080	1.00	−585179
		4	4.26	60.3	.051	2850	1.07	
2	True values	1	1.00	60.0	.050	500	1.00	
		2	1.56	50.0	.040	1000	1.00	−172840
		3	2.75	40.0	.030	2000	1.00	−598713
		4	4.26	30.0	.020	3000	1.00	
	Final estimates (10th iteration)	1	1.00	59.9	.049	555	0.94	
		2	1.56	49.8	.040	939	0.91	−168814
		3	2.74	39.7	.028	1710	1.14	−598672
		4	4.25	29.9	.026	3340	1.12	
3	True values	1	1.00	60.0	.050	3000	1.00	
		2	1.56	50.0	.040	2000	1.00	−172698
		3	2.75	40.0	.030	1000	1.00	−597695
		4	4.26	30.0	.020	500	1.00	
	Final estimates (8th iteration)	1	1.00	60.1	.049	3200	0.99	
		2	1.56	49.9	.039	1940	0.91	−167862
		3	2.74	40.2	.031	1220	0.81	−597590
		4	2.74	40.2	.031	1220	0.81	
4	True values	1	1.00	60.0	.050	3000	5.00	
		2	1.56	50.0	.040	2000	6.00	−172840
		3	2.75	40.0	.030	1000	7.00	−646643
		4	4.26	30.0	.020	500	8.00	
	Final estimates (7th iteration)	1	1.00	59.8	.051	3110	5.04	
		2	1.56	50.1	.039	2170	5.73	−177327
		3	2.74	43.1	.030	999	6.63	−646576
		4	4.24	22.0	.016	508	8.20	

3.5.1 The model

Suppose that the same microarray has been scanned m times at different sensitivity levels of the scanner. Let y_{ij} denote the observed intensity of the ith of n spots in the jth scan. In the absence of censoring, we assume that the expectation of y_{ij} would be $\mu_i\beta_j$, where μ_i is the expression level of gene i and β_j is the multiplicative scaling effect due to scanner setting j. The observed intensity is the average of pixel values. For example, the data plotted in Figure 3.1 were produced by Quantarray, using the average of pixels between the 80th and 95th percentiles of the pixel distribution contained in a 25 by 25 square centred on each spot. If some of these pixels are censored at T then the expectation of y_{ij} will be less than $\mu_i\beta_j$. Figure 3.1 suggests that hyperbolic function (see Figure 3.4) may be appropriate to model the behaviour of the data. We explored the possibility of using a hyperbolic function with the asymptotes $E(y_{ij}) = \mu_i\beta_j$ and $E(y_{ij}) = T = 2^{16} - 1 = 65535$ as the location of the model. The expression of

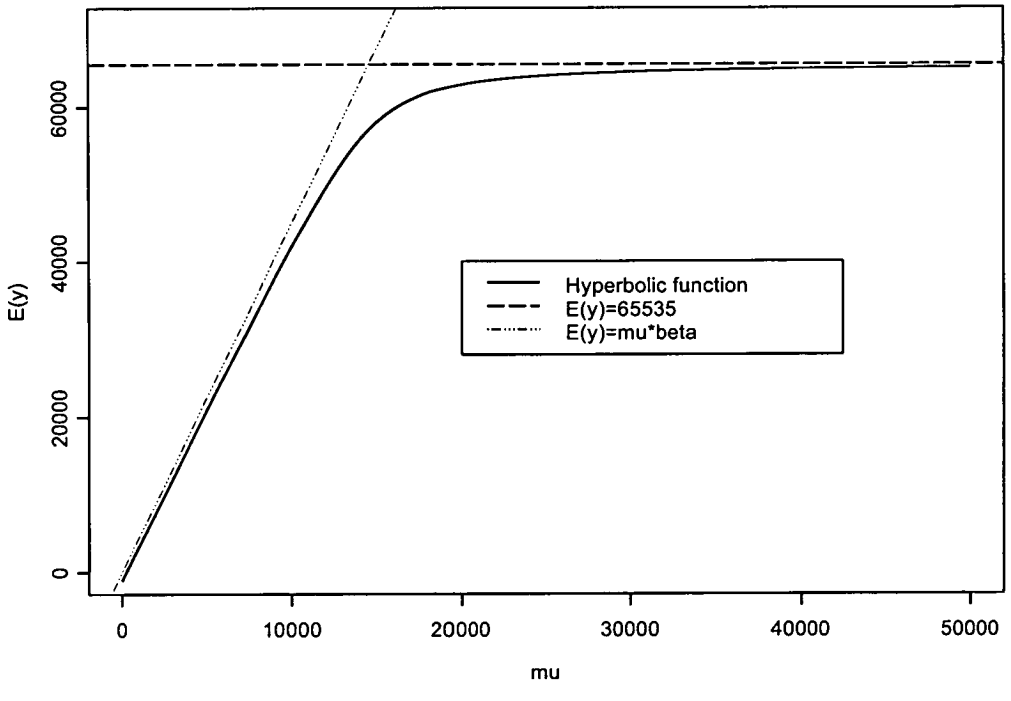


Figure 3.4: Typical curve ($\beta = 4.5, \alpha = 4000$) of the hyperbolic function defined in equation (3.12).

the function can be obtained as a solution of the quadratic equation

$$\left\{\frac{E(y_{ij})}{\beta_j} - \mu_i\right\}\{E(y_{ij}) - T\} = \alpha^2,$$

giving

$$E(y_{ij}) = \frac{T + \mu_i\beta_j}{2} - \sqrt{\left(\frac{T - \mu_i\beta_j}{2}\right)^2 + \alpha^2\beta_j}. \tag{3.12}$$

A typical curve of the hyperbolic function (3.12) has been depicted in Figure 3.4. The hyperbolic functional regression model relating y_{ij} to the true gene expression (μ_i) can be expressed as

$$y_{ij} = \frac{T + \mu_i\beta_j}{2} - \sqrt{\left(\frac{T - \mu_i\beta_j}{2}\right)^2 + \alpha^2\beta_j} + e_{ij}, \tag{3.13}$$

where $\beta_1 = 1$ (the identifiability condition), $\beta_j, (j = 2, \cdots, m)$ is the scanning effect of the jth setting and $T(= 65535)$ is the maximum detectable intensity by the scanning software.

The random error terms e_{ij} are assumed to follow independent Cauchy distributions with location zero and dispersion parameters $\sigma_{ij}^2 = \beta_j^2\sigma_1^2 + \mu_i^2\beta_j^2\sigma_2^2$. The Cauchy distribution is chosen to take account of the outlying observations as evident from Figures 3.1 and 3.3. We investigated the use of other robust methods, *e.g.*, M-estimation using a Gaussian likelihood like objective function and maximum likelihood method based on t-distribution. The Cauchy model have been found to perform better than the other robust methods we have investigated. Performance of M-estimation and maximum likelihood method based on t-distribution using censored mean functional model is described in Section 3.6. The scale parameters have been scaled by the corresponding scanning effects (β) to allow for increasing variance, as evident from the data (Figure 3.1), across the scans of increasing sensitivity. For functional regression models it is not possible to estimate separate variance terms for individual scan of data using maximum likelihood estimation, because the likelihood is unbounded unless taken account of rounding-off errors (Copas, 1972). Even when appropriately defined by taking account of grouping errors, the maximum likelihood method lead to inconsistent estimates.

3.5.2 Maximum likelihood estimation

We have seen that the use of maximum likelihood method in estimating functional models has some limitations. The likelihood function is unbounded unless account

is taken of the rounding-off errors in the observations (Copas, 1972). Even when appropriately defined, assuming that observations belong to certain intervals, the solution is only an approximation to the maximum likelihood estimate, and does not lead to consistent estimator. Alternative methods, *e.g.*, Morton (1981), Chan and Mak (1983), exist to deal with the problems of unboundedness and inconsistency, but these apply to models with Gaussian distributed errors, and are not straight forward to modify to accommodate Cauchy errors. However, parameters of the hyperbolic model (3.13) are estimable through maximum likelihood, because we are not using separate scale parameters for each scan.

The probability density function of y_{ij} assuming a Cauchy distribution can be written as

$$f(y_{ij}) = \left[\pi\sigma_{ij}\left\{1 + \left(y_{ij} - (T - \mu_i\beta_j)/2 + \sqrt{((T - \mu_i\beta_j)/2)^2 + \alpha^2\beta_j}\right)^2/\sigma_{ij}^2\right\}\right]^{-1} \quad (3.14)$$

Under independence assumption of the errors, the log-likelihood function of the parameters of model (3.14) is equivalent to

$$L(\mu, \beta, \sigma_1, \sigma_2, \alpha) = \sum_{i=1}^{n}\sum_{j=1}^{m} l(\mu, \beta, \sigma_1, \sigma_2, \alpha), \quad (3.15)$$

where

$$l(\mu, \beta, \sigma_1, \sigma_2, \alpha) = -0.5\ \log(\sigma_{ij}^2) - \log\left\{1 + \frac{\left(y_{ij} - (T - \mu_i\beta_j)/2 + \sqrt{((T - \mu_i\beta_j)/2)^2 + \alpha^2\beta_j}\right)^2}{\sigma_{ij}^2}\right\}. \quad (3.16)$$

The main challenge of working with this model is the estimation of the large number $(n+m+2)$ of parameters which increases with the number of spots (n) on the array and number of scans (m). According to the literature on *measurement error* (ME) models (Cheng and Van Ness, 1999, Chapter 1) the maximum likelihood estimate of the functional model with Gaussian errors does not exist when the variance parameters are left free to vary. We therefore scale the scale parameters by the corresponding scanning effects (β) to allow for increasing variance across the scans of increasing sensitivity. Leaving α as a free parameter makes the optimisation algorithm extremely slow. We propose an algorithm for the maximum likelihood estimates of the parameters keeping α as fixed. Optimum value of α can be determined by investigating its profile likelihood.

The alternating maximum-likelihood algorithm for simultaneous estimation of the parameters of the proposed model for fixed $\alpha = \alpha^*$ consists of the following steps:

1. Give starting values of all parameters $(\mu^{(0)}, \beta^{(0)}, \sigma_1^{(0)}, \sigma_2^{(0)}, \alpha^*)$ where $\mu^{(0)}$ is a vector of dimension n, $\beta^{(0)}$ is a $(m-1)$ vector and $\sigma_1^{(0)}, \sigma_2^{(0)}$ are scalars.

2. Update all $\mu_i, (i = 1, \cdots, n)$ individually by maximizing $\sum_{j=1}^{m} l(\mu_i, \beta^{(0)}, \sigma_1^{(0)}, \sigma_2^{(0)}, \alpha^*)$ with respect to μ_i alone. Denote the updated vector by $\mu^{(1)} = (\mu_1^{(1)}, \cdots, \mu_n^{(1)})$.

3. Update the $(m+1)$ parameters in $(\beta, \sigma_1, \sigma_2)$ by maximising $L(\beta, \sigma_1, \sigma_2, \alpha^*, \mu^{(1)})$.

Continue repeating steps (2) and (3) replacing the previous estimates by the updated ones until convergence, *i.e.*, gain in the log-likelihood function is negligible, is achieved. The simplex method of Nelder and Mead (1965) using FORTRAN 90 and IMSL routine `DUMPOL` can be used as optimisation tool. Because of the multimodal nature of the Cauchy Likelihood, choice of the initial values, particularly of the parameters in μ, is crucial. With reasonable given values of the other parameters $(\beta, \sigma_1, \sigma_2)$ we choose the value of $\mu_i^{(0)}$ from m possible candidates $y_{ij}/\beta_j, (j = 1, \cdots, m)$ as the one for which $\sum_{j=1}^{m} l(\mu_i, \beta^{(0)}, \sigma_1^{(0)}, \sigma_2^{(0)}, \alpha^*)$ is maximum. Further discussion of Cauchy model properties and estimation techniques of the parameters are given in Chapter 4.

3.5.3 Application

We provide an application of the proposed method to the murine macrophage data described in Section 3.3. As mentioned before, leaving α as a free parameter in model (3.13) seriously reduces the efficiency of the computing algorithm with respect to the convergence time. This is possibly because the relative change in the likelihood function with respect to change in α is very small. Investigating the profile likelihood of α it is seen that, for both arrays of data, the optimum values of α^2 lie between $10^{6.2}$ and $10^{6.4}$ and the approximate relation between $\log\alpha^2$ and the profile log-likelihood function within the range is quadratic. Using quadratic interpolations the optimum values of α^2 have been found to be $10^{6.31}$ and $10^{6.27}$ for the data of arrays 1 and 2 respectively. We therefore run the estimation algorithm fixing α^2 at these levels. The maximum likelihood estimates of the parameters, other than μ, for both sets of data are tabulated in Table 3.6.

Observed intensity divided by the corresponding scanning effects (β) for both sets of data are plotted against the corresponding estimated gene expressions μ

Table 3.6: Maximum likelihood estimates of the scanning effects, scale parameters of the hyperbolic functional model applied to murine macrophage data.

	Scanning effects			Scale		
Data set	β_2	β_3	β_4	σ_1	σ_2	$\log_{10}(\alpha^2)$
Array-1	1.56	2.75	4.33	5.35	0.0065	6.31
Array-2	1.71	2.71	4.53	5.38	0.0051	6.27

in Figure 3.5. It is seen that for the highly expressed genes the estimated gene expressions are, as desired, consistent with the data of scan-1, which are not likely to be affected by pixel censoring. However, the hyperbolic function has only one degree of freedom, and because of insufficient degrees of freedom, the model fails to comply with the trend of the data properly, for example, the model is not fitting the data well between 10,000 and 15,000 of scan-4 of array 1 data.

3.6 Censored mean functional regression model

As the hyperbolic function, because of not being flexible enough due to lack of degrees of freedom, does not fit the data well, we consider an alternative nonlinear function to describe the relationship of the multiple scan data. We call it *censored mean function*, which is derived as the expectation of the minimum of an individual pixel value and T, the censoring threshold. This function overcomes the 'symmetry' problem of the hyperbolic function and the derivation of the function is consistent with the data generation mechanism of the scanner.

3.6.1 The model

With the notations defined in Section 3.4, we derive the censored mean function assuming that the pixel values associated with a spot have mean $\mu_i\beta_j$ and variance $\mu_i^2\beta_j^2\nu^2$, where ν is a variance scaling term, and is distributed as Gaussian. That is

$$y_{ijk} \sim N(\mu_i\beta_j,\ \mu_i^2\beta_j^2\nu^2),$$

where y_{ijk} represents the kth pixel value in the jth scanning of the ith spot. It can be shown that the censored mean function, the expectation of the minimum of y_{ijk} and T can be expressed as

$$\begin{aligned} E(y_{ij} \vee T) &= T + (\mu_i\beta_j - T)\Phi\left(\frac{T-\mu_i\beta_j}{\mu_i\beta_j\nu}\right) - \mu_i\beta_j\nu\phi\left(\frac{T-\mu_i\beta_j}{\mu_i\beta_j\nu}\right) \\ &= g(\mu_i\beta_j, \nu), \text{ say,} \end{aligned} \tag{3.17}$$

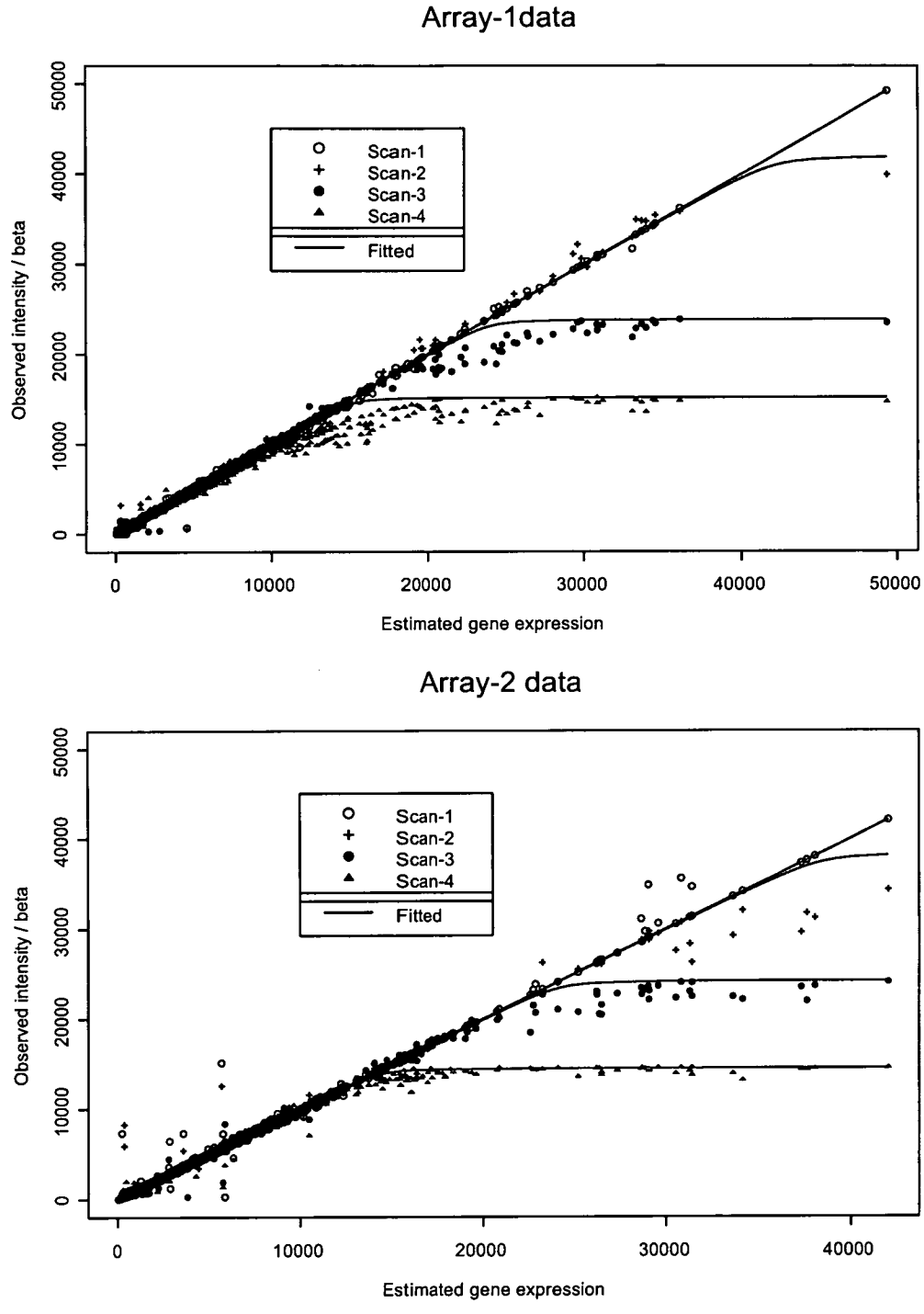


Figure 3.5: Rescaled intensities $(y_{ij}/\hat{\beta}_j)$ plotted against estimated gene expressions $(\hat{\mu}_i)$ for the hyperbolic functional model applied to arrays 1 and 2 of murine macrophage data. The solid lines indicate the corresponding fitted model.

where $\phi(.)$ and $\Phi(.)$ are the density and distribution functions of the standard Gaussian random variable respectively, and $y_{ij} \vee T$ indicates $\min(y_{ij}, T)$. This is derived from expressions for truncated normal distributions (Johnson *et al.*, 1994, p. 156). Typical curves of the function are shown in Figure 3.6.

We now assume that the observed spot intensity y_{ij} is distributed with mean $g(\mu_i\beta_j, \nu)$. Further, we assume that, apart from a few outliers, as are evident in Figure 3.1, the distribution is Gaussian with variance $\sigma_{ij}^2 = \sigma_1^2\beta_j^2 + \sigma_2^2\mu_i^2\beta_j^2$. Some heavy-tailed distribution truncated below 0 and above T with the same location and scale parameters can used to account for the outliers. We therefore assume

$$y_{ij} \sim \begin{cases} N\left(g(\mu_i\beta_j, \nu),\ \sigma_{ij}^2\right) & \text{with probability} \quad (1-\pi) \\ H^{(0,T)}\left(g(\mu_i\beta_j, \nu),\ \sigma_{ij}^2\right) & \text{with probability} \quad \pi \end{cases} \tag{3.18}$$

where $\beta_1 \equiv 1$ for identifiability and the proportion π is very small. The notation $H^{(0,T)}$ represents some heavy-tailed distribution with support $[0, T]$. Model (3.18) belongs to the class of *functional* regression model, a form of *measurement error* model (Madansky, 1959).

3.6.2 M-estimation

To overcome the influence of outliers, we propose a robust method, a form of M-estimation, where the objective function to be minimised is:

$$L(\mu, \beta, \sigma_1, \sigma_2, \nu) = \sum_{i=1}^{n}\sum_{j=1}^{m}\left[\left(\frac{m-1}{m}\right)\log\sigma_{ij}^2 + \rho\left(\frac{y_{ij} - g(\mu_i\beta_j, \nu)}{\sigma_{ij}}\right)\right] \tag{3.19}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m} l(\mu, \beta, \sigma_1, \sigma_2, \nu), \text{ (say)}, \tag{3.20}$$

with

$$\rho(e) = \begin{cases} e^2, & \text{if } |e| < 3 \\ 9, & \text{otherwise.} \end{cases} \tag{3.21}$$

The factor $(m-1)/m$ in (3.19) is used as an adjustment for degrees of freedom for variance estimation. The main challenge of working with this model is the estimation of the large number $(n + m + 2)$ of parameters. We propose an alternating algorithm for simultaneous estimation of all the parameters of model (3.18) as follows:

1. Set $\mu = y_{.1}$ (intensity data of scan-1) as the starting values and minimise L (but with $\rho(e) = e^2$) with respect to all other parameters $(\beta, \sigma_1, \sigma_2, \nu)$, where μ is a vector of dimension n, β is a $(m-1)$ vector and σ_1, σ_2

and ν are scalars. Denote the updated values of other parameters by $(\beta^{(1)}, \sigma_1^{(1)}, \sigma_2^{(1)}, \nu^{(1)})$.

2. Update each μ_i, $(i = 1, \cdots, n)$ individually according to the following sub-steps:

   (a) For each j, set $\mu_i = g^{-1}(y_{ij}, \nu^{(1)})/\beta_j^{(1)}$.

   (b) Minimise $L_i = \sum_{j=1}^{m} l(\mu_i, \beta^{(1)}, \sigma_1^{(1)}, \sigma_2^{(1)}, \nu^{(1)})$ with respect to μ_i alone.

   (c) Repeat (a)–(b) for $j = 1, \cdots, m$.

   (d) From among m updated values of μ_i, choose the one with minimum L_i value. Denote the updated vector by $\mu^{(1)}$.

3. Update the $(m + 2)$ parameters in $(\beta, \sigma_1, \sigma_2, \nu)$ by minimising $L(\beta, \sigma_1, \sigma_2, \nu, \mu^{(1)})$ for given values of the gene expression parameters in $\mu^{(1)}$.

Continue repeating steps (2) and (3), replacing the previous estimates by the updated ones, until gain in the objective function is negligible. The sub-steps under step (2), that update each μ_i starting from m different initial values, are essential. Otherwise, the algorithm may be trapped in a local optimum. The simplex method of Nelder and Mead (1965) using FORTRAN 90 and IMSL Library was used as an optimisation tool. The IMSL routine `DUMPOL` implements the simplex method of function minimisation.

3.6.3 Application

We apply the method to data from a single channel of two microarrays, plotted in Figure 3.1. CPU time, with a single processor Ultra-1 Sun machine, for executing the program codes to apply the method of Section 3.6.2 to each microarray took 11 minutes. Estimates of the parameters, other than μ, for both sets of data are tabulated in Table 3.7.

Table 3.7: M-estimates of the scanning effects, scale parameters and ν applied to murine macrophage data.

	Scanning effects			Scale		
Data set	β_2	β_3	β_4	σ_1	σ_2	ν
Array-1	1.6	2.7	4.3	15.5	0.019	0.43
Array-2	1.7	2.7	4.5	15.5	0.015	0.25

Observed intensity data divided by the corresponding scanning effects (β) for both sets of data are plotted against the corresponding estimated gene expressions (μ) in Figure 3.6. Although the fit of the model to the data appears better than the hyperbolic model, a considerable number of points at the upper end of scan 1 data are not consistent with the fitted model.

Figure 3.7, a plot of the standardised residuals against the rank of the estimated gene expressions from one microarray, does not indicate any obvious model violations.

On each array each gene has been replicated twice in such a way that spot i and $i + n/2$ represent the same gene where $i = 1, \cdots, n/2$. To compare the between replicate variations in the data and fit, we compute

$$BSS(\mu) = \sum_{i=1}^{n/2} \frac{(\mu_i - \mu_{i+n/2})^2}{((\mu_i + \mu_{i+n/2})/2)^2}, \tag{3.22}$$

and

$$BSS(y_{\cdot j}) = \sum_{i=1}^{n/2} \frac{(y_{ij}/\beta_j - y_{i+n/2,j}/\beta_j)^2}{((y_{ij}/\beta_j + y_{i+n/2,j}/\beta_j)/2)^2} = \sum_{i=1}^{n/2} \frac{(y_{ij} - y_{i+n/2,j})^2}{((y_{ij} + y_{i+n/2,j})/2)^2}, \tag{3.23}$$

for $j = 1, \cdots, m$.

The results are summarised in Table 3.8. It is seen that except for one instance (array-2, scan-1 data), between replicate variation in the estimated gene expressions is less than that in any individual scan of data. This suggests that it is possible to reduce the between replicate variation of the gene expression measurements by combining the data according to the proposed model from several scans.

Table 3.8: Comparison of between replicate variation in data and fit (M-estimation).

	Between replicate variation				
Data set	$BSS(\hat{\mu})$	$BSS(y_{\cdot 1})$	$BSS(y_{\cdot 2})$	$BSS(y_{\cdot 3})$	$BSS(y_{\cdot 4})$
Array-1	794.42	816.59	825.07	806.72	817.38
Array-2	839.73	835.80	851.82	878.24	867.67

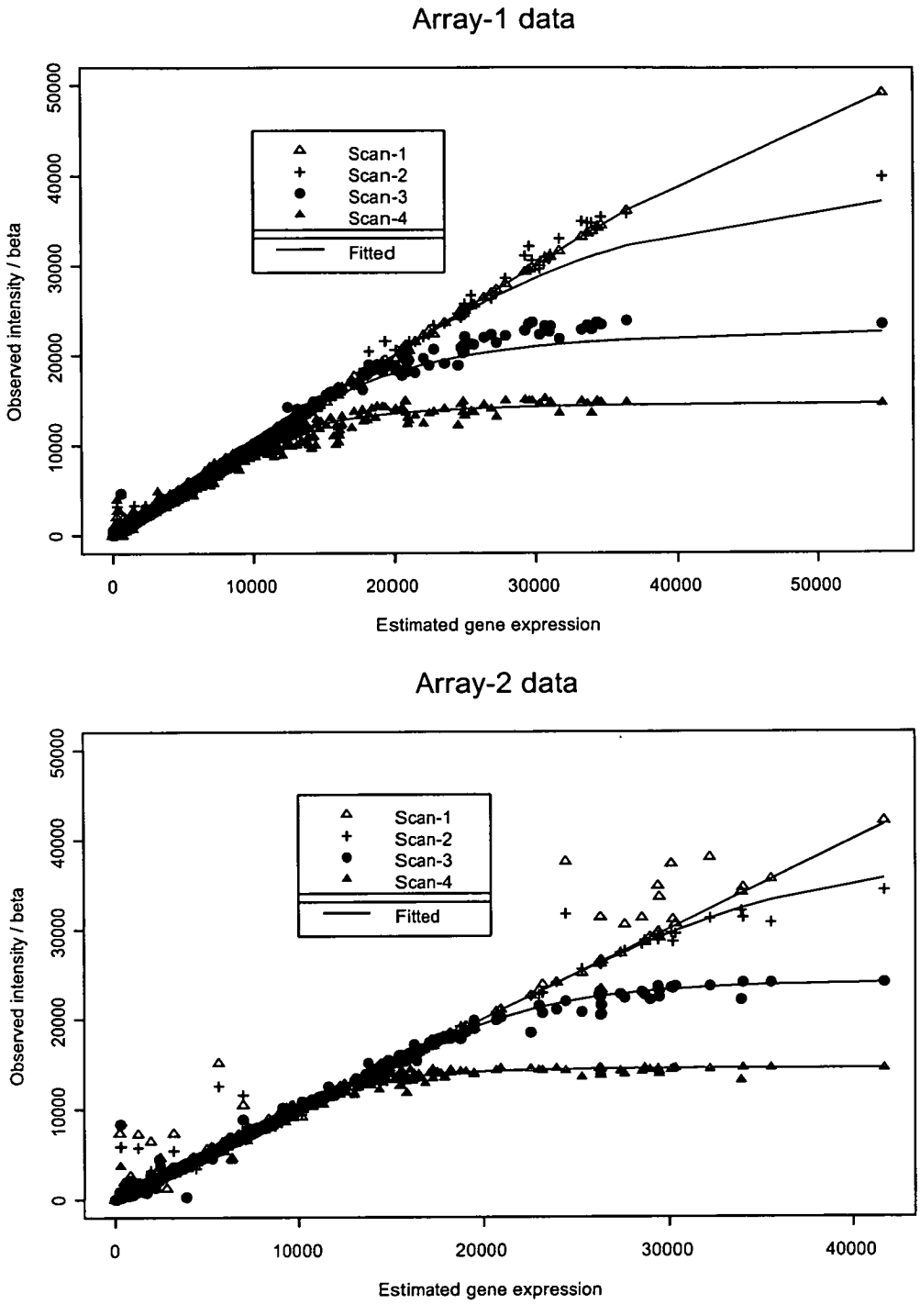


Figure 3.6: Rescaled intensities $(y_{ij}/\hat{\beta}_j)$ plotted against estimated gene expressions $(\hat{\mu}_i)$ for arrays 1 and 2. The solid lines indicate the corresponding fitted model by M-estimation.

Figure 3.7: Standardised residuals against the ranks of estimated gene expressions. The limits $\pm$ 5 (solid lines) and $\pm$ 3 (dashed lines) are also shown.

3.6.4 Simulation study

We performed some simulation experiments to check the validity of the estimation algorithm and the properties of the estimates obtained. We simulated 100 datasets from the model (3.18) for different level of contamination (π = 0.0, 0.01 and 0.05), using the parameter values as estimated for array-2 data (Table 3.7). For the gene expression parameters we used the same set of values for both replicates, obtained as the average of the estimated gene expressions of the two replicates for array-2 data. We used truncated Cauchy distribution with the same location and scale parameters for $H^{(0,T)}$. Empirical biases and standard errors of the parameter estimates, other than μ, are summarised in Table 3.9. It is seen that the parameters are estimated with high precision and negligible bias. For the gene expression parameters (μ), we plot empirical biases, as percentage of true values, against the rank of true values (for $\pi = 0.01$) in Figure 3.8. The bias in estimating gene expression parameters is seen to be in an acceptable range, in most cases less than 0.5%.

Table 3.9: Estimated biases and standard errors for the method of M-estimation. The results are based on 100 simulated data sets.

% Contamination (100π)	True values	Parameters					
		β_2	β_3	β_4	σ_1	σ_2	ν
		1.7	2.7	4.5	15.5	0.015	0.25
0	Bias	0.00035	0.00066	0.0012	−0.43	−0.00036	0.0062
	SE	0.00060	0.0010	0.0017	0.14	0.00018	0.0057
1	Bias	0.00062	0.0010	0.0018	−0.39	−0.00036	0.0080
	SE	0.00074	0.0013	0.0020	0.12	0.00019	0.0049
5	Bias	0.00090	0.0015	0.0026	−0.33	−0.00024	0.0093
	SE	0.00078	0.0012	0.0019	0.13	0.00020	0.0049

Between replicate variations computed according to the formulae (3.22) and (3.23) are summarised in Table 3.10. These results suggest obvious gain in reducing between replicate variations by combining data of multiple scans. These values are considerably less than those in Table 3.8, because the experimental data has other sources of variability in addition to that due to sampling.

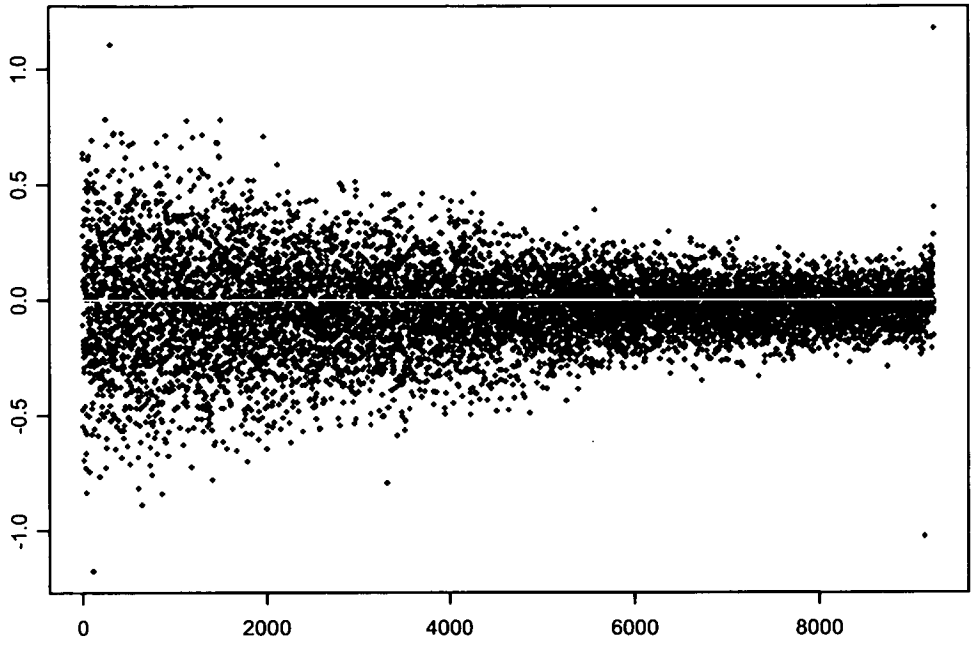


Figure 3.8: Percentage of bias against the ranks of true gene expression values.

Table 3.10: Comparison of between replicate variation for simulated data. Results are averages over 100 data sets.

	Between replicate variation				
100π	$BSS(\hat{\mu})$	$BSS(y_{.1})$	$BSS(y_{.2})$	$BSS(y_{.3})$	$BSS(y_{.4})$
0	2.85	10.71	10.73	10.79	10.77
1	2.91	14.14	14.13	14.27	14.18
5	3.17	27.81	27.67	27.56	26.66

3.6.5 Censored mean functional model based on t-distribution

One problem with the M-estimation is the subjectivity about the proportion of data to be considered as outliers. In our case we considered observations with standardised residuals ≥ 3 as outliers. There is however no rigorous justification for this cut-off point and this may depend on the particular data set being used. One alternative for modeling outliers is to consider a maximum likelihood estimation with distribution having heavier tails than Gaussian but not as heavy as that of a Cauchy distribution. Gaussian and Cauchy distributions are two extreme special cases of t-distribution with sufficiently large and unit degrees of freedom respectively. It is therefore a good idea to model the data with a t-distribution with appropriate degrees of freedom. We considered maximum-likelihood estimation based on t-distribution treating degrees of freedom as a parameter to be estimated from the data .

3.6.6 The model and estimation

With the notations used in previous sections, we assume that the spot intensity data y_{ij} is distributed as a t-distribution with location function $g(\mu_i\beta_j, \nu)$, additive plus multiplicative variance $\sigma_{ij}^2 = \sigma_1^2\beta_j^2 + \sigma_2^2\mu_i^2\beta_j^2$ and degrees of freedom η. The log-likelihood function can be expressed as

$$L(\mu, \beta, \sigma_1, \sigma_2, \nu, \eta) = \sum_{i=1}^{n} L_i(\mu_i, \beta, \sigma_1, \sigma_2, \nu, \eta), \tag{3.24}$$

where

$$\begin{aligned} L_i(\mu_i, \beta, \sigma_1, \sigma_2, \nu, \eta) \;=\; & \textstyle\sum_{j=1}^{m} \Big[-0.5 \log \sigma_{ij}^2 - \log \Gamma(\eta/2) + \log \Gamma(\eta + 1/2) + \\ & (\eta/2)\log\eta - (\eta+1)/2 \log \left\{ \eta + \left(\tfrac{y_{ij} - g(\mu_i\beta_j, \nu)}{\sigma_{ij}} \right)^2 \right\} \Big]. \end{aligned} \tag{3.25}$$

The model has $(n + m + 3)$ parameters. We propose an alternating algorithm for simultaneous estimation of all the parameters of the model as follows:

1. Set $\mu = y_{\cdot 1}$ (intensity data of scan-1) as the starting values and maximise L with respect to all other parameters $(\beta, \sigma_1, \sigma_2, \nu, \eta)$, where μ is a vector of dimension n, β is a $(m-1)$ vector and σ_1, σ_2, ν and η are scalars. Denote the updated values of other parameters by $(\beta^{(1)}, \sigma_1^{(1)}, \sigma_2^{(1)}, \nu^{(1)}, \eta^{(1)})$.

2. Update each μ_i, $(i = 1, \cdots, n)$ individually according to the following substeps:

   (a) For each j, set $\mu_i = g^{-1}(y_{ij}, \nu^{(1)})/\beta_j^{(1)}$.

(b) Maximise L_i with respect to μ_i alone.

(c) Repeat (a)–(b) for $j = 1, \cdots, m$.

(d) From among the m updated values of μ_i, choose the one with maximum L_i value. Denote the updated vector by $\mu^{(1)}$.

3. Update the $(m+3)$ parameters in $(\beta, \sigma_1, \sigma_2, \nu, \eta)$ by maximising $L(\beta, \sigma_1, \sigma_2, \nu, \eta, \mu^{(1)})$ for given values of the gene expression parameters in $\mu^{(1)}$.

Continue repeating steps (2) and (3), replacing the previous estimates by the updated ones, until gain in the log-likelihood function is negligible. The sub-steps under step (2), that update each μ_i starting from m different initial values, are essential. Otherwise, the algorithm may be trapped in a local optimum. We use the polytope search algorithm of Nelder and Mead (1965) for locating the optimum solution. The IMSL routine `DUMPOL` implements the Nelder-Mead algorithm; we use the FORTRAN 90 routine.

3.6.7 Application

We apply the method to the data from a single channel of two microarrays, plotted in Figure 3.1. Estimates of the parameters, other than μ, for both sets of data are tabulated in Table 3.11.

Table 3.11: Estimates of the scanning effects, scale parameters and degrees of freedom.

	Scanning effects			Scales			df
Data set	β_2	β_3	β_4	σ_1	σ_2	ν	η
Array-1	1.6	2.7	4.3	6.15	0.0076	0.42	1.12
Array-2	1.7	2.7	4.5	7.05	0.071	0.26	1.33

Observed intensity data divided by the corresponding scanning effects (β) for both sets of data are plotted against the corresponding estimated gene expressions (μ) in Figure 3.9. If compared with Figure 3.6, corresponding figure for M-estimation, it is seen that scan 1 data are now more consistent, particularly for array-2 data, with the estimated gene expressions. However, in terms of between replicate variation, comparing Table 3.8 with Table 3.12, M-estimation performs better than that of maximum likelihood estimation based on t-distribution.

Table 3.12: Comparison of between replicate variation in data and fit (maximum likelihood estimation based on t-distribution).

	Between replicate variation				
Data set	$BSS(\hat{\mu})$	$BSS(y_{.1})$	$BSS(y_{.2})$	$BSS(y_{.3})$	$BSS(y_{.4})$
Array-1	812.09	816.59	825.07	806.72	817.38
Array-2	857.37	835.80	851.82	878.24	867.67

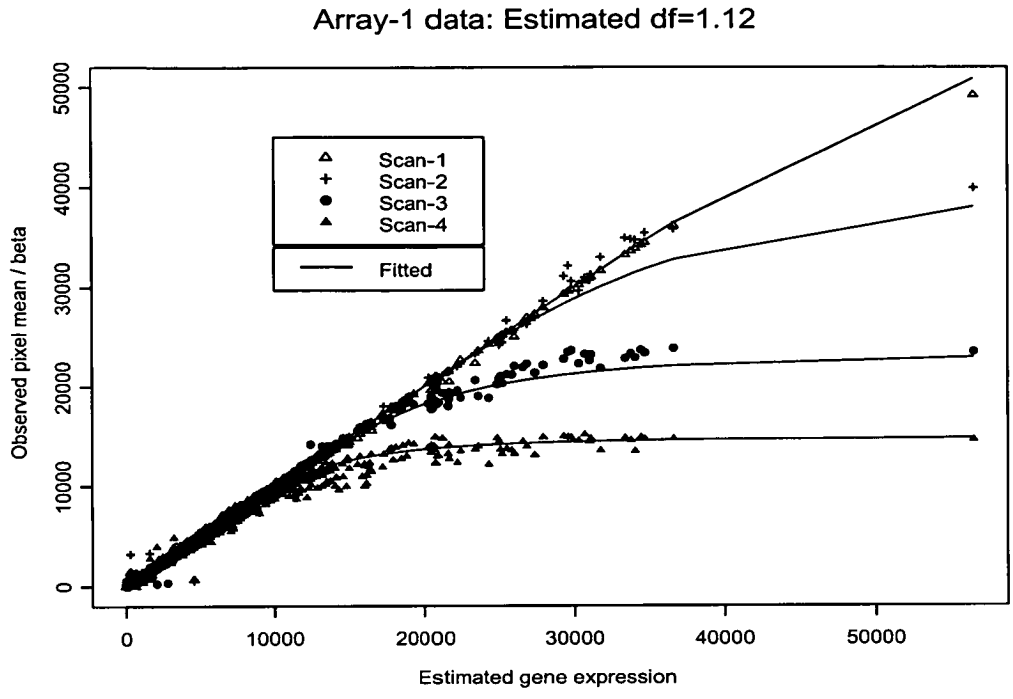
3.6.8 Simulation study

We have performed a small scale simulation, based on only 5 data sets, to investigate bias in the parameter estimates. Data are generated according to the proposed model. True parameter values and the simulation results are tabulated in Table 3.13. It is seen that the scale parameters, σ_1 and σ_2, and degrees of

Table 3.13: Simulation results for maximum likelihood estimation based on t-distribution

		Parameters						
Data		β_2	β_3	β_4	σ_1	σ_2	ν	η
set	True	1.7	2.7	4.5	15.5	0.015	0.25	4.0
1		1.7004	2.7011	4.4991	10.9144	0.0104	0.2581	2.3184
2		1.6982	2.6963	4.4974	11.5165	0.0108	0.2520	2.5952
3	Estimates	1.6993	2.7008	4.5021	11.1061	0.0109	0.2588	2.4683
4		1.6993	2.6989	4.4999	11.4129	0.0106	0.2654	2.4784
5		1.6992	2.6981	4.4952	10.6150	0.0105	0.2592	2.2584

freedom (η) are under estimated. Cheng and Van Ness (1999, pp. 50–51) showed that the maximum likelihood estimates of the variance parameters of a functional relationship are not consistent. There are negligible biases in the estimates of the other parameters. It therefore seems that estimation of appropriate degrees of freedom may not be possible in a straight forward way. This is the main drawback of using t-distribution for modeling these data.



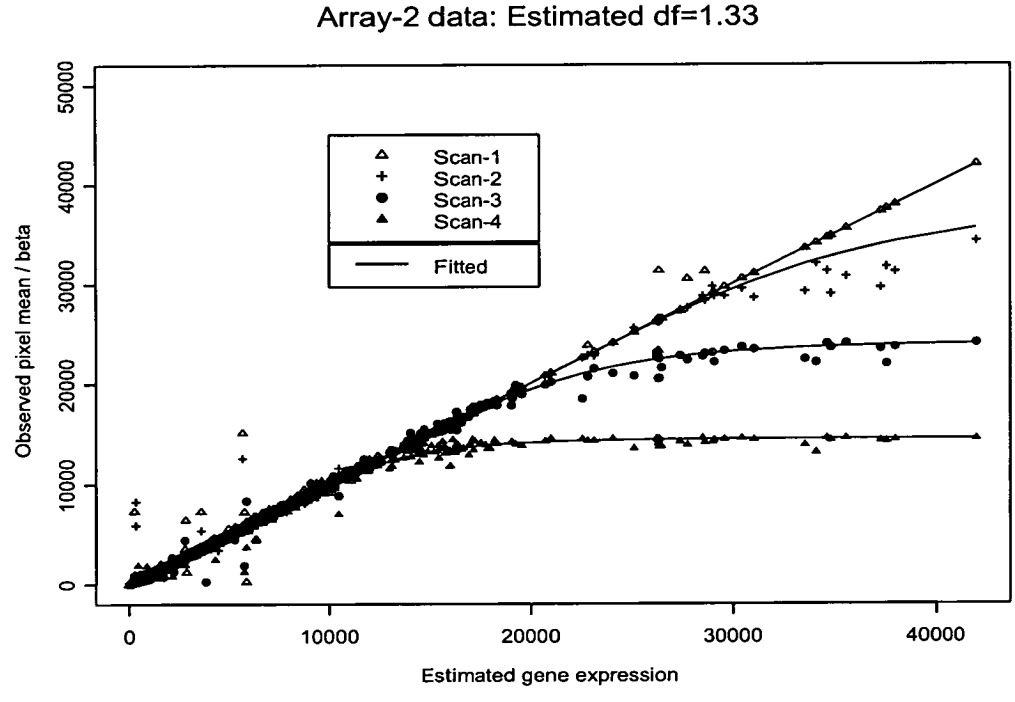


Figure 3.9: Rescaled intensities($y_{ij}/\hat{\beta}_j$) plotted against estimated gene expressions ($\hat{\mu}_i$) for arrays 1 and 2. The solid lines indicate the corresponding fitted model.

3.7 Summary

The exploratory analysis presented in this chapter gives substantial knowledge about the patterns of multiple scan data. The relationship of intensity data, observed at higher settings of the scanner, with the true gene expression levels appears nonlinear at the upper level of gene expression. The variance of the observed intensity seems to depend on the level of the intensity, and an additive plus multiplicative model may be a reasonable choice. In the linear functional model with Gaussian mixture, nonlinearity at the upper level may not be adequately modelled by the variance parameter of the second component of the mixture distribution. An appropriate nonlinear location function is therefore needed to explain the curvature in the relationship. The hyperbolic function has also been found not to be fully adequate to describe the trend of the data. The censored mean function, having an intuitive similarity with the scanner's data generation mechanism, has been found to best describe the nonlinearity in the data. However, choice of an appropriate error distribution still remains a problem. M-estimation with a Gaussian-likelihood type objective function was not fully successful to provide adequate fit to the data. In terms of describing the tail behaviour of the distribution of error, a t-distribution might be a reasonable choice. However, because of the bias problem in the maximum likelihood estimate of the degrees of freedom, we need to find other alternatives. In Chapter 4, we propose our refined model, based on a censored Cauchy distribution, for combining multiple laser scans of microarrays.

Chapter 4

Combining multiple laser scans: refined model

4.1 Introduction

Our investigations in Chapter 3 suggest that the linear or the hyperbolic functional model may not be completely adequate for combining multiple scan data. The idea behind considering a linear model was to model the nonlinearity through the variance parameter in the second component of the Gaussian mixture. However, it was not possible to estimate the parameters of the mixture model through maximum likelihood method, and the alternative algorithm, though terminates successfully, did not give realistic estimates of the parameters, particularly the mixing proportion parameters. It therefore seems that the mixture model we investigated is not adequate to distinguish the contaminated portion from the main body of the data.

The hyperbolic functional model (Section 3.5) using a Cauchy distribution provided considerable improvement in fitting the relationship for combining multiple scan data. However, detail investigation of the model suggests that the function is not flexible enough to model the slightly varying, from array to array, nonlinear patterns found in multiple scan data. The censored mean function in equation (3.17), depicted in Figure 4.1, appears appropriate to describe the trends in multiple scan data. The function has a natural correspondence with the data generation mechanism of the microarray laser scanners. The parameter ν controls the amount of curvature of the function depending on the amount of censoring present in the data. The function is therefore flexible enough to capture the varying degree of nonlinearity trends found in multiple scan data.

However censored mean functional model based on M-estimation, assuming that distributions of errors are Gaussian for majority of the genes, does not provide adequate fit to the data (see Figure 3.6, array 2 data). Although improves

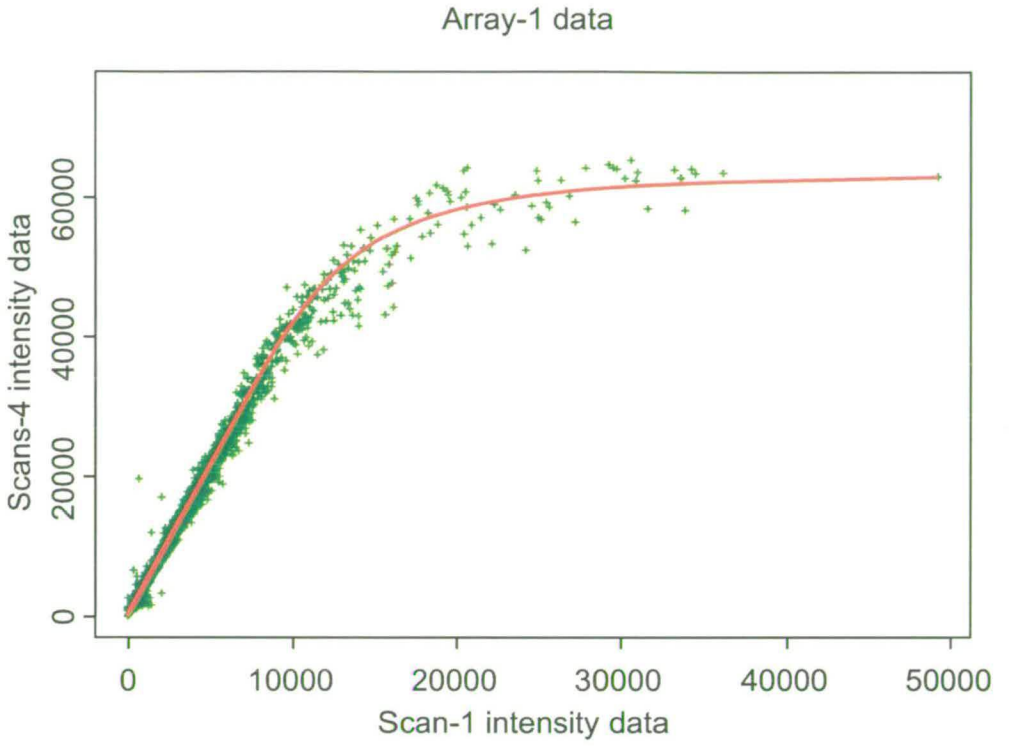


Figure 4.1: A typical curve of the censored mean function (3.17) plotted on scan 4 vs. scan 1 (array 1) data.

the fit compared to M-estimation, use of a t-distribution for describing the error of the censored mean functional regression also appears problematic. The maximum likelihood method underestimates the degrees of freedom as well as the additive and multiplicative components of the scale. The downward bias in the estimated degrees of freedom makes it difficult to use the t-distribution because the fitted model will not match the tail behaviour of the distribution of the data, which was the main point of choosing a t-distribution. However underestimation in the scale parameters does not seem to affect the estimation of gene expression parameters. We therefore investigate a Cauchy model and a censored Cauchy model to describe the distribution of error of the censored mean functional model. A description of the method based on a Cauchy distribution can be found in Khondoker *et al.* (2006a). This chapter studies in detail the method of combining multiple scan data based on a robust likelihood method using a Cauchy distribution and a censored Cauchy distribution.

4.2 Cauchy distribution and its properties

The Cauchy density has been studied in the mathematical world for over three centuries (Johnson *et al.*, 1994, p. 298). An excellent historical account of the distribution has been prepared by Stigler (1974).

A random variable X is said to be distributed as Cauchy with location parameter μ and scale parameter $\sigma > 0$ if its probability density function is given by

$$f(x;\mu,\sigma) = (\pi\sigma)^{-1}\left[1+\left\{\frac{x-\mu}{\sigma}\right\}^2\right]^{-1}, \quad -\infty < x < \infty, \ -\infty < \mu < \infty. \tag{4.1}$$

The cumulative distribution function is

$$\frac{1}{2} + \pi^{-1}\tan^{-1}\left\{\frac{x-\mu}{\sigma}\right\}. \tag{4.2}$$

The distribution is symmetrical about $x = \mu$. The median is μ, the upper and lower quartiles are $\mu \pm \sigma$. The 95% probability limits are $\mu \pm 12.71\sigma$, as compared with the $\mu \pm 1.96\sigma$ of a Gaussian distribution. The distribution does not possess finite moments of order greater than or equal to 1, and so does not possess a finite expected value or standard deviation. However, μ and σ are location and scale parameters respectively and can be regarded as being analogous to mean and standard deviation.

Despite having some peculiar properties, Cauchy distribution may be useful in modelling distribution of data having heavier tails than normal. Maximum likelihood method based on a Cauchy model provides a basis for robust estimation.

In situations where values of X greater than a fixed value (x_0) cannot be observed, X can be regarded as having a censored Cauchy distribution,

$$g(x,\mu,\sigma) = \begin{cases} (\pi\sigma)^{-1}\left[1+\left\{\frac{x-\mu}{\sigma}\right\}^2\right]^{-1}, & -\infty < x < x_0 \\ \frac{1}{2} - \pi^{-1}\tan^{-1}\left\{\frac{x_0-\mu}{\sigma}\right\}, & x \geq x_0. \end{cases} \tag{4.3}$$

It may be noted that the distribution has a point mass at x_0 equal to

$$P(X \geq x_0) = 1 - P(X < x_0) = \frac{1}{2} - \pi^{-1}\tan^{-1}\left\{\frac{x_0-\mu}{\sigma}\right\},$$

because the only information obtained about X when X is censored at x_0 is that $X \geq x_0$.

Estimation of the Cauchy parameters is however somewhat problematic as the likelihood for location parameter for given scale is multimodal and, in general, no explicit solution of the likelihood equation exists. Nevertheless, Ferguson (1978)

derived closed-form expressions for maximum likelihood estimates for small samples. Barnett (1966) has noted that likelihood equation for the location parameter of the Cauchy model often has multiple roots and suggested to examine the likelihood function over an extensive but fine grid for the location. Barnett (1966) has also investigated the distribution of number of local maxima for different sample size through simulation study. Reeds (1985) showed that the number of local maxima of the Cauchy location likelihood function which are not global maxima is asymptotically Poisson distributed with mean parameter $1/\pi$. This agrees with simulation experiment results obtained by Barnett (1966).

However, Copas (1975) showed that under regularity conditions the joint likelihood for both location and scale parameters for independent, identical Cauchy variables has exactly one maximum point, and at most one stationary point. This was studied subsequently by several authors, *e.g.*, Mäkeläinen *et al.* (1981), Gabrielsen (1982), Clarke (1983).

Estimation methods of the parameters of this model have been discussed extensively in the literature. For samples of sizes 3 and 4, Ferguson (1978) obtained the closed-form expressions for the maximum likelihood estimates of the location and scale parameters of the Cauchy model. Haas *et al.* (1970) studied the performance of Newton-Raphson method in finding maximum likelihood estimators through simulation experiments. Koutrouvelis (1980, 1982) discussed the estimation of parameters of a Cauchy model utilising the empirical characteristic function. Brooks and Morgan (1995) studied the estimation of the Cauchy parameters using simulated annealing. Ionides (2005) discussed the use of maximum smoothed likelihood estimation (MSLE) method for multimodal likelihood. These estimators are shown to be asymptotically efficient for models possessing local asymptotic normality.

Bai and Fu (1987) proved that even in the case where the likelihood equation has multiple roots, the maximum likelihood estimator (global maximum) remains as an asymptotically optimal estimator in Bahadur sense.

4.3 The Cauchy model and estimation

In the Cauchy model considered in this chapter, we assume that the observed spot intensity y_{ij} is distributed as a Cauchy distribution with location $g(\mu_i\beta_j, \nu)$, given by equation (3.17) in Section 3.6.1, and with additive plus multiplicative scale model $\sigma_{ij} = \sqrt{(\sigma_1^2 + \sigma_2^2\mu_i^2)\beta_j^2}$. The proposed model therefore is

$$y_{ij} \sim C\left(g(\mu_i\beta_j, \nu),\ \sigma_{ij}^2\right), \tag{4.4}$$

where $\beta_1 \equiv 1$ for identifiability. The notation $C(a, b^2)$ represents a Cauchy distribution (Johnson *et al.*, 1994) with location and scale parameters a and b respectively.

The log-likelihood function for estimating the parameters of model (4.4) can be expressed as:

$$L(\mu, \beta, \sigma_1, \sigma_2, \nu) = \sum_{i=1}^{n} L_i(\mu_i, \beta, \sigma_1, \sigma_2, \nu), \tag{4.5}$$

where

$$L_i(\mu_i, \beta, \sigma_1, \sigma_2, \nu) = -\sum_{j=1}^{m} \left[\log \sigma_{ij} + \log \left\{ 1 + \left(\frac{y_{ij} - g(\mu_i \beta_j, \nu)}{\sigma_{ij}} \right)^2 \right\} \right]. \tag{4.6}$$

We propose an alternating algorithm for simultaneous estimation of all the parameters of model (4.4) as follows:

1. Set $\mu = y_{\cdot 1}$ (intensity data of scan-1) as the starting values and maximise L with respect to all other parameters $(\beta, \sigma_1, \sigma_2, \nu)$, where μ is a vector of dimension n, β is a $(m-1)$ vector and σ_1, σ_2 and ν are scalars. Denote the updated values of other parameters by $(\beta^{(1)}, \sigma_1^{(1)}, \sigma_2^{(1)}, \nu^{(1)})$.

2. Update each μ_i, $(i = 1, \cdots, n)$ individually according to the following sub-steps:

   (a) For each j, set $\mu_i = g^{-1}(y_{ij}, \nu^{(1)})/\beta_j^{(1)}$.

   (b) Maximise L_i with respect to μ_i alone.

   (c) Repeat (a)–(b) for $j = 1, \cdots, m$.

   (d) From among the m updated values of μ_i, choose the one with maximum L_i value. Denote the updated vector by $\mu^{(1)}$.

3. Update the $(m+2)$ parameters in $(\beta, \sigma_1, \sigma_2, \nu)$ by maximising $L(\beta, \sigma_1, \sigma_2, \nu, \mu^{(1)})$ for given values of the gene expression parameters in $\mu^{(1)}$.

Continue repeating steps (2) and (3), replacing the previous estimates by the updated ones, until gain in the log-likelihood function is negligible. The sub-steps under step (2), that update each μ_i starting from m different initial values, are essential. Otherwise, the algorithm may be trapped in a local optimum. Because for the gene expression parameters (μ) the likelihood naturally decomposes into n components, and μ_i can be estimated my maximising the ith component (L_i),

which generally has m peaks, one near to the intensity value for each scan. Multiple starts for each μ_i therefore improves the chance of finding the highest peak. Typical nature of the profile likelihood of μ, particularly for genes affected by pixel censoring, is shown in Figure 4.2.

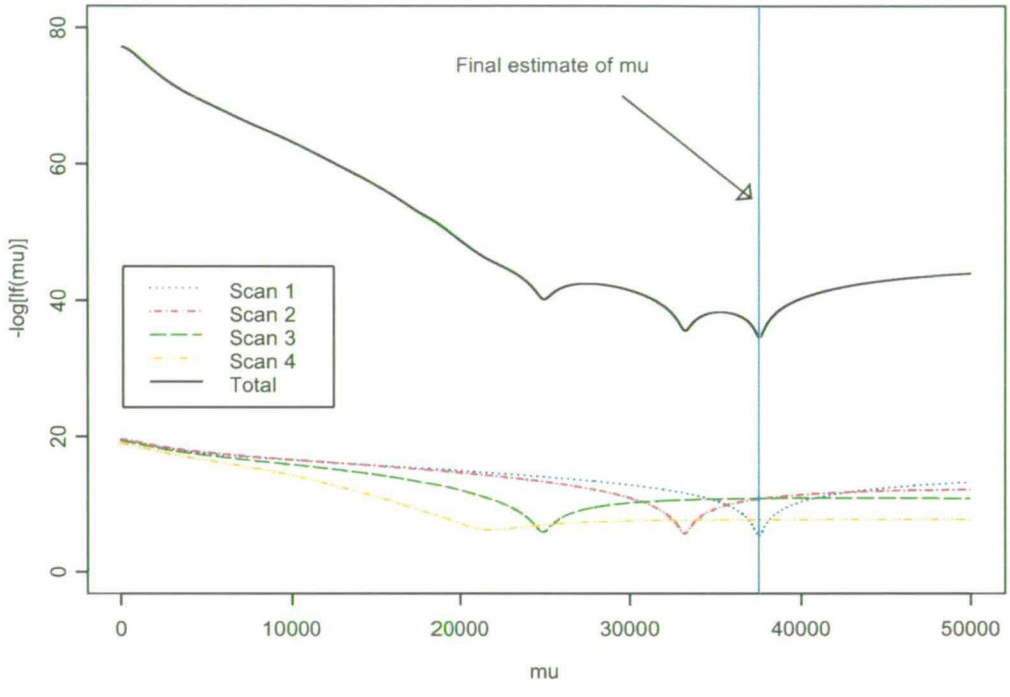


Figure 4.2: Negative profile log-likelihood of μ for the gene 6822 (array 2) at the final estimates of the other parameters.

The figure shows the negative profile log-likelihood of the gene corresponding to the spot 6822. Observed intensity of that spot was affected by pixel censoring in scans 2–4. The figure illustrates that use of multiple start facilitate choosing the right estimate of the gene expression. The algorithm seems to have rightly chosen the final gene expression estimate of that spot near to the observed intensity of scan 1, which has not been affected by pixel censoring.

We have used the numerical optimization routine of Nelder and Mead (1965), which has been implemented in the IMSL routine `DUMPOL`, for implementing the above algorithm.

We have found that L increases at each iteration of our proposed algorithm. Therefore, because L is bounded above with probability 1, the alternating algorithm is guaranteed to terminate at a local stationary point. We did not encounter any convergence problem for the real and simulated data used in this study. How-

ever, as is usually the case with optimization algorithms, there is no guarantee that the global maximum will be found.

4.4 Applications

We apply the method to two different data sets:

1. Murine macrophage data, and
2. Iron-deficiency data.

Murine macrophage data have been described in Section 3.3. A brief description of the iron-deficiency data is given in Section 4.4.2, where we apply the method to the data. We have described and used the iron-deficiency data more extensively in Chapter 5 to illustrate our nonparametric location and scale normalisation method.

4.4.1 Murine macrophage data

To apply the method to murine macrophage data we consider data from a single channel of two microarrays, plotted in Figure 3.1. The best way to choose the initial value of β_j $(j = 2, \cdots, m)$ is to consider the slope of simple no-intercept LTS regression (Rousseeuw and Leroy, 1987) of data of scan j on scan 1. For our chosen initials, the algorithm took 14 and 7 complete iterations for data of arrays 1 and 2 respectively. Results of the iterations are given in Table 4.1. CPU time, with a single processor Ultra-1 Sun machine, for executing the program to apply the method of Section 4.3 to arrays 1 and 2 took 11 and 8 minutes respectively. Number of iterations and CPU time required however depends on the choice of the initial parameters. Final estimates of the parameters (other than μ) for both arrays of data are tabulated in Table 4.2. The estimated scanning effects and scale parameters appears reasonably similar for the two arrays of data. The estimate of ν, the parameter that controls the degree of curvature of the relationship, is however smaller for array 2 data indicating greater degree of curvature in the relationship and hence more censoring. This is also evident from the plots of data in Figure 3.1.

Observed intensity data divided by the corresponding scanning effects (β) for both sets of data are plotted against the corresponding estimated gene expressions (μ) in Figure 4.3. It appears that the model provide satisfactory fit to the data. The outlying points at the upper end of scan 1 data (as seen in Figure 3.6, array 2 data) have now been disappeared (Figure 4.3). Figure 4.4 shows

Table 4.1: Iterations of the algorithm applied to arrays 1 and 2 of murine macrophage data.

		Parameter (other than μ) estimates.						
Data set	Iterations	β_2	β_3	β_4	σ_1	σ_2	ν	$-\log(\text{LF})$
	0 (initial)	1.5000	2.7400	4.3200	6.5000	0.0070	0.4500	181340.96
	1	1.5589	2.7474	4.2981	6.4086	0.0077	0.4677	160239.53
	2	1.5585	2.7471	4.3025	5.6385	0.0068	0.4481	159813.35
	3	1.5581	2.7469	4.3072	5.6111	0.0069	0.4482	159754.63
	4	1.5580	2.7466	4.3091	5.5379	0.0069	0.4492	159729.86
	5	1.5579	2.7465	4.3123	5.4891	0.0070	0.4510	159721.92
1	6	1.5578	2.7464	4.3135	5.4188	0.0070	0.4525	159710.86
	7	1.5579	2.7460	4.3152	5.4732	0.0070	0.4503	159708.95
	8	1.5582	2.7460	4.3166	5.3304	0.0067	0.4445	159698.94
	9	1.5583	2.7464	4.3175	5.3676	0.0067	0.4411	159691.63
	10	1.5583	2.7466	4.3187	5.3868	0.0067	0.4381	159688.87
	11	1.5584	2.7469	4.3195	5.4242	0.0067	0.4347	159686.87
	12	1.5584	2.7468	4.3195	5.3636	0.0068	0.4193	159681.29
	13	1.5584	2.7468	4.3195	5.3636	0.0068	0.4193	159678.42
	14 (final)	1.5584	2.7468	4.3195	5.3636	0.0068	0.4193	159678.42
	0 (initial)	1.7000	2.7000	4.5000	10.0000	0.0500	0.2500	182181.61
	1	1.7102	2.7114	4.5278	6.5479	0.0058	0.2638	156273.30
	2	1.7100	2.7113	4.5276	5.6884	0.0054	0.2796	155731.11
2	3	1.7102	2.7112	4.5272	5.3722	0.0053	0.2761	155683.22
	4	1.7099	2.7111	4.5276	5.4202	0.0052	0.2751	155679.16
	5	1.7098	2.7109	4.5277	5.3566	0.0051	0.2662	155671.26
	6	1.7098	2.7109	4.5277	5.3566	0.0051	0.2662	155667.32
	7(final)	1.7098	2.7109	4.5277	5.3566	0.0051	0.2662	155667.32

Table 4.2: Estimates of the scanning effects, scale parameters and ν for murine macrophage data.

	Scanning effects			Scale		
Data set	β_2	β_3	β_4	σ_1	σ_2	ν
Array-1	1.56	2.75	4.32	5.36	0.0068	0.42
Array-2	1.71	2.71	4.53	5.36	0.0051	0.27

a plot of standardised residuals against the rank of estimated gene expressions from one microarray and does not indicate any obvious model violations. Assessment of model fit is also possible via likelihood-based criteria such as AIC and GAIC. However, more pertinent is whether the use of multiple scans can improve the signal-to-noise ratio in the estimates of gene expression. Between replicate variations in the data and fit are computed according to the formula

$$S(\tilde{\mu}) = \sum_{i=1}^{n/2} \frac{(\tilde{\mu}_i - \tilde{\mu}_{i+n/2})^2}{((\hat{\mu}_i + \hat{\mu}_{i+n/2})/2)^2} \tag{4.7}$$

where $\tilde{\mu}$ is replaced by $\hat{\mu}$ to assess the multi-scan estimate, and by $y_{.j}/\hat{\beta}_j$ to assess the use of scan j alone. Because variability increases approximately as

the square of the expression level, we give equal weight in S to genes at low and high levels by dividing by the square of the estimated expression level for each gene. However, rather than computing this using $\tilde{\mu}$, which is downward biased for censored spots, we use $\hat{\mu}$ in all cases. Results are summarised in Table 4.3. It is seen that between replicate variation in the estimated gene expressions is

Table 4.3: Comparison of between replicate variation in data and fit.

	Between replicate variation				
Data set	$S(\hat{\mu})$	$S(y_{.1}/\hat{\beta}_1)$	$S(y_{.2}/\hat{\beta}_2)$	$S(y_{.3}/\hat{\beta}_3)$	$S(y_{.4}/\hat{\beta}_4)$
Array-1	812	958	913	823	927
Array-2	858	1683	1768	882	863

less than that in any individual scan of data. This suggests that it is possible to reduce the between replicate variation of the gene expression measurements by combining the data according to the proposed model from several scans. Results of Table 4.3 indicate that by combining scans we improve the signal-to-noise ratio in the data, particularly relative to scan 1, which would be the scientists' preferred single scan, as the other ones are affected by censoring bias.

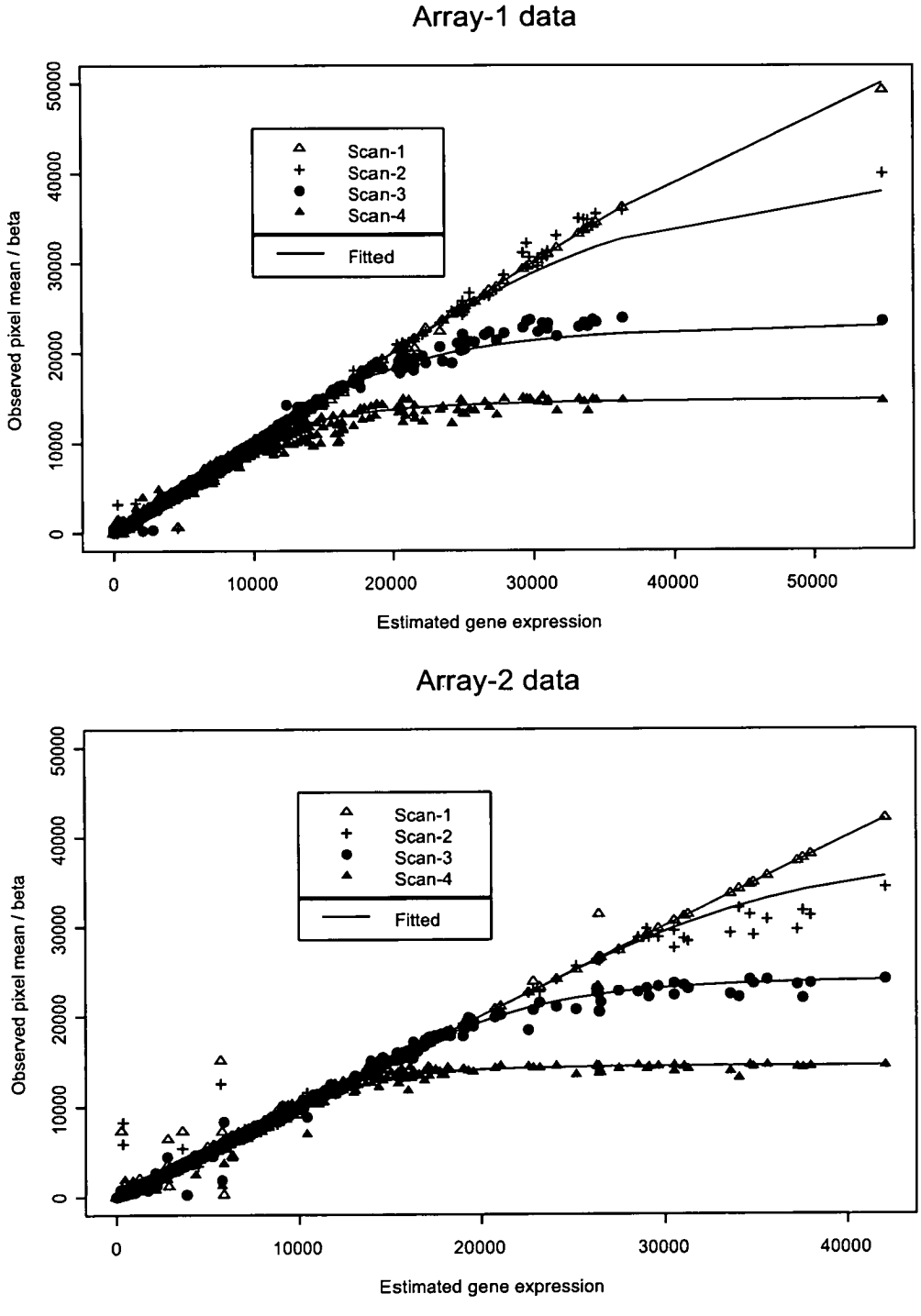


Figure 4.3: Rescaled intensities ($y_{ij}/\hat{\beta}_j$) plotted against estimated gene expressions ($\hat{\mu}_i$). The solid lines indicate the corresponding fitted model.

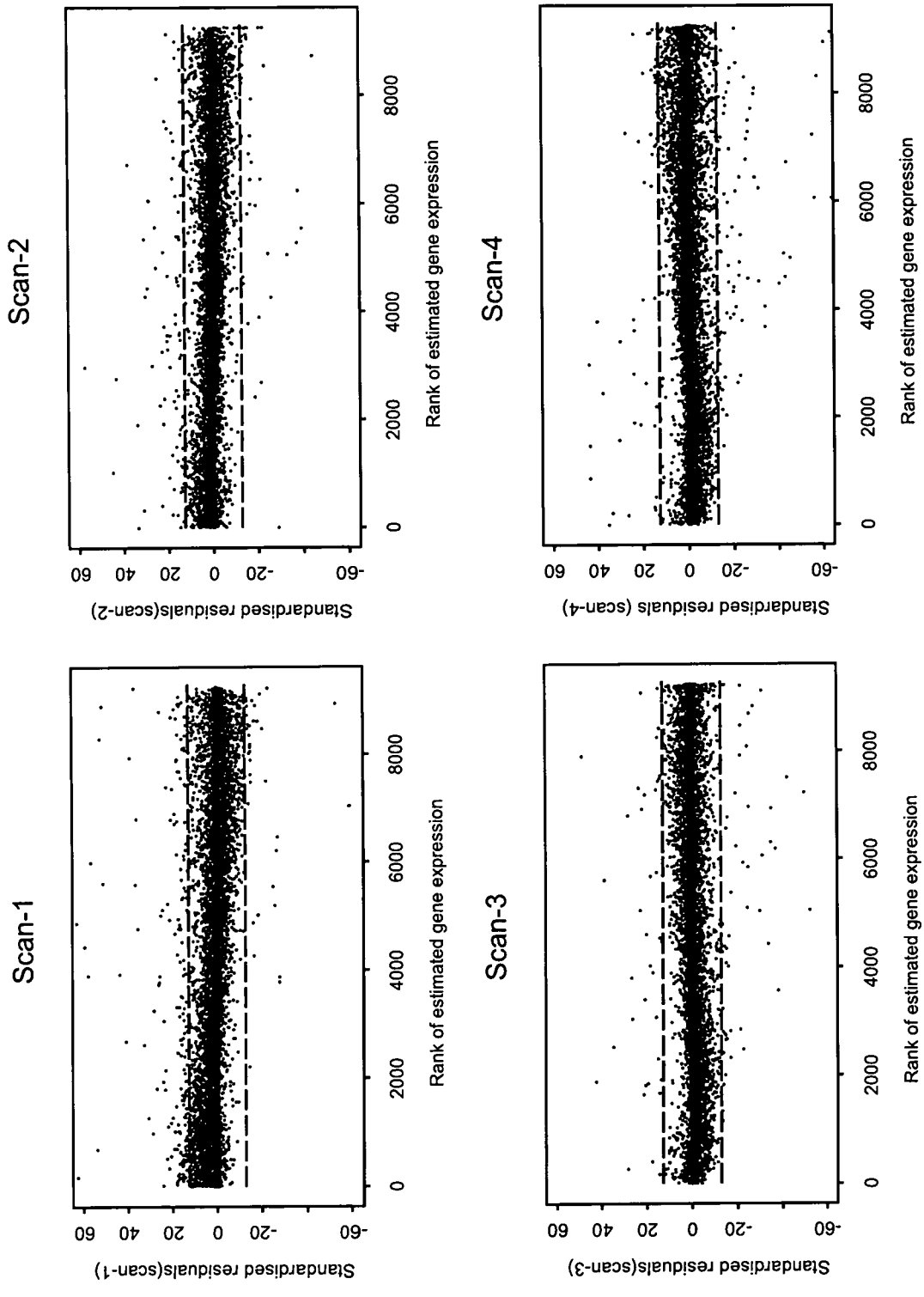


Figure 4.4: Standardised residuals against the ranks of estimated gene expressions. The dashed lines show 95% probability limits (± 12.71). The value 12.71 is the 97.5th percentile point of the standard Cauchy distribution.

4.4.2 Iron-deficiency data

The iron-deficiency experiment, conducted at Rowett Research Institute, Scotland, deals with the impact of iron deficiency on maternal rats and their offspring. The current data set deals with liver (slides 1–8) and kidney (slides 9–16) taken from 4 iron deficient and 4 control mothers. There are intensity measurements for 9984 spots on each of the 16 arrays. The data set gives the observed intensity for both Cy3 and Cy5 channels at 3 different scanner settings as well as the combined intensity measurements using the MVI Pro 2.6.0 (Bell, 2003) software. A description of the algorithm for combining multiple scan data in MVI Pro 2.6.0 is given in Section 2.4. In this section we use the data from both channels of array 3 to illustrate our method of combining multiple scan data. Estimates of the param-

Table 4.4: Estimates of the scanning effects, scale parameters and ν applied to iron-deficiency data.

	Scanning effects		Scale		
Data set	β_2	β_3	σ_1	σ_2	ν
Array-3 (Cy3)	4.97	16.43	0.78	0.0430	0.30
Array-3 (Cy5)	4.36	16.51	2.90	0.0128	0.17

eters other than the gene expression parameters applied to this data are given in Table 4.12. Smaller values, compared to the murine macrophage data, of the scale estimates indicate that the data are less noisy in this case. The multiplicative scanning effects (β_1 and β_2) appear similar for both Cy3 and Cy5 data. Large differences between the scanning effect parameters reflect the substantial differences between the settings of consecutive scanning. Although the scanner settings are the same for both Cy3 and Cy5 data, the smaller value of the estimate of ν indicates higher amount of censoring in Cy5 data. The reason for such difference is probably the differential dye behaviour in response to laser scanning. Plots of the data as well as the corresponding fitted models are shown in Figure 4.5. Plots suggest less noise in the data as compared with the murine macrophage data which was also reflected in the scale estimates. Another difference between the two data sets is the degree of the curvature in the relationship. Curvature in the iron deficiency data, particularly Cy5 data, appears more extreme, almost like linear splines, as compared with the murine macrophage data. This reflects that almost all pixels belonging to a spot representing highly expressed gene are censored at higher settings. It can be seen from the fitted model on the data in Figure 4.5 that the censored mean functional model is flexible enough to represent even this extreme nature of censoring effect.

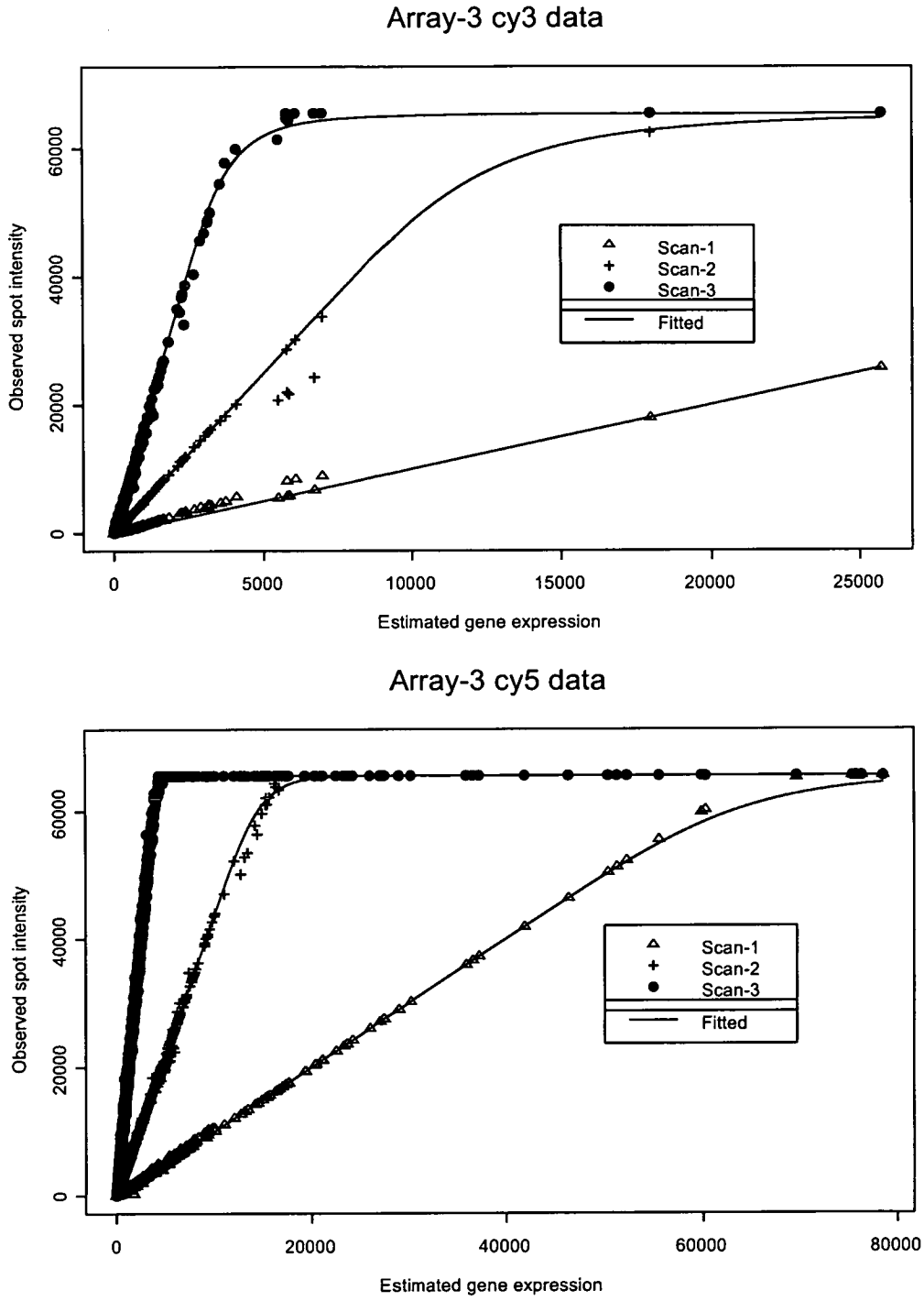


Figure 4.5: Observed spot intensities (y_{ij}) plotted against estimated gene expressions ($\hat{\mu}_i$). The solid lines indicate the corresponding fitted models.

4.5 Simulation study

We performed some simulation experiments to check the validity of the estimation algorithm and the properties of the estimators. The IMSL routine RNCHY has been used for data generation. We simulated 100 datasets from model (4.4) using the parameter values as estimated for array-2 of murine macrophage data (Table 4.2). For the gene expression parameters we used the same set of values for both replicates, obtained as the average of the estimated gene expressions of the two replicates for array-2 data. Empirical biases and standard errors of the parameter estimates, other than μ, are summarised in Table 4.5. It is seen

Table 4.5: Estimated biases and standard errors of the parameters of the Cauchy model for combining multiple scans. The results are based on 100 simulated data sets.

	Parameters					
	β_2	β_3	β_4	σ_1	σ_2	ν
True	1.71	2.71	4.53	5.36	0.0051	0.27
Bias	−0.00005	−0.00007	−0.00015	−2.036	−0.00187	0.00077
SE	0.00038	0.00069	0.00111	0.053	0.00008	0.00235

that the parameters, except for σ_1 and σ_2, are estimated with high precision and negligible bias. There is substantial downward bias in the maximum likelihood estimates of σ_1 and σ_2. The maximum likelihood estimation bias in the scale parameters is studied in more detail in Section 4.6. This bias, however, does not affect the estimation of the other parameters and in particular the gene expression parameters (μ_i). We think that there is little concern as this bias does not affect the estimation of gene expression parameters. Simulation results do not suggest any notable systematic bias in the gene expression estimates. We plot empirical biases, as percentage of the true values, against the rank of the true values in Figure 4.6. The bias in estimating gene expression parameters is seen to be in an acceptable range, in most cases less than 0.5%.

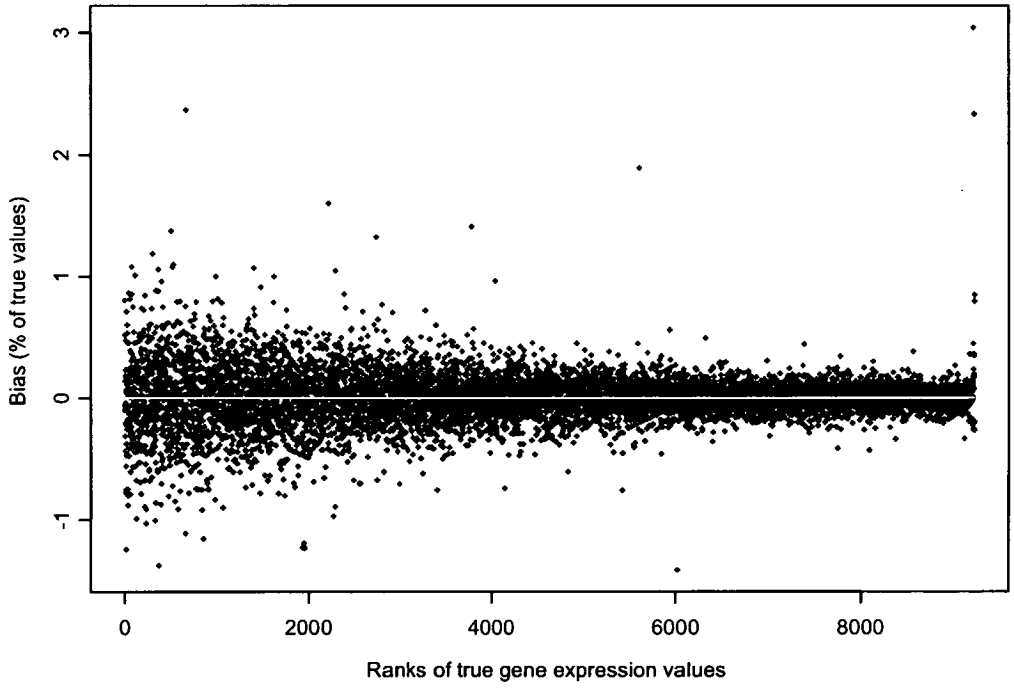


Figure 4.6: Percentage of bias against the ranks of true gene expression values.

4.6 Investigating the bias in the Cauchy scale

Maximum likelihood estimation (MLE) is an elegant and probably the most widely used estimation method because of its many desirable properties. However, maximum likelihood estimators may often be biased. For example, the maximum likelihood estimator of the variance σ^2 of the Gaussian distribution $N(\mu, \sigma^2)$ with unknown μ is biased by the factor $(m-1)/m$, where m is the sample size. This bias can be adjusted easily as the factor $(m-1)/m$, for fixed given sample, is known. Quantification and subsequent adjustment for bias are however difficult for the maximum likelihood estimator of Cauchy scale as it does not have any analytic closed form expression. Mardia *et al.* (1999) proposed a method for estimating first-order bias in the maximum-likelihood estimators using the expressions for score and information of the parameter, and showed that bias in the Cauchy scale σ is linear in σ and the sample size. The expression for first-order expected bias in the maximum likelihood estimate $\hat{\theta}$ of the parameter θ was derived to be

$$E(\hat{\theta} - \theta) = \frac{1}{2mI(\theta)^2}[2E\{U(\theta)U'(\theta)\} + E\{U''(\theta)\}] + o(m^{-1}) \tag{4.8}$$

where $U(\theta)$ is the score for θ, $U'(\theta)$ and $U''(\theta)$ denote the first and second derivatives of score with respect to θ and $I(\theta)$ is the information for θ.

Mardia *et al.* (1999) gave the bias expression for the Cauchy scale (σ) with $\mu = 0$ according to the formula (4.8), which was derived as follows. The Cauchy density (4.1) with $\mu = 0$ has the score function

$$U(\sigma) = \frac{1}{\sigma} - \frac{2\sigma}{\sigma^2 + x^2}. \tag{4.9}$$

Therefore

$$E\{U(\sigma)U'(\sigma)\} = -\frac{1}{2\sigma^3}, \tag{4.10}$$

$$E\{U''(\sigma)\} = \frac{3}{2\sigma^3}, \tag{4.11}$$

and

$$I(\sigma)^2 = \frac{1}{4\sigma^4}. \tag{4.12}$$

Substitution of these equations into (4.8) gives,

$$E(\hat{\sigma} - \sigma) \approx \sigma/m. \tag{4.13}$$

Thus the bias is linear in σ. Therefore bias in the maximum likelihood estimator of Cauchy scale may be negligible for large samples.

Although we are using a Cauchy distribution to define our model (4.4), estimation of the model parameters is not the same as that of the Cauchy density

(4.1). We are estimating each of the gene expression parameters μ_i from a sample of size $m = 4$, the number of scans. On the other hand, the scale parameters σ_1 and σ_2 are being estimated by combining a large number, $n = 9248$, of small samples of size m. We have conducted some additional investigation of the bias in the scale parameter estimation of the Cauchy model through simulation study.

We have noted that the bias in the scale estimates of our proposed model is different from that shown by Mardia *et al.* (1999) in the Cauchy scale of a simple model. We consider two simple models as in equation (4.14) and (4.15) given by

$$y_{ij} \sim C(\mu_i, \sigma_i^2), \tag{4.14}$$

and

$$y_{ij} \sim C(\mu_i, \sigma^2). \tag{4.15}$$

The second model is of a similar nature of our multiple scan model (4.4) where the scale σ is estimated by combining a large number of small samples of size $m = 4$. Whereas in the first model each μ_i and σ_i are estimated from a sample of size m. We have investigated the bias in the scale estimates of models (4.14) and (4.15) through simulation experiments. For model (4.14), we simulate 10000 samples of size $m = 4$ from $C(\mu = i, \sigma_i^2 = 1)$ and estimated the location and scale for each of the samples. Results are summarised in Table 4.6. We see from Table 4.6

Table 4.6: Mean, estimated SE and bias of the scale estimates of the Cauchy model: $y_{ij} \sim C(\mu_i = i, \sigma_i^2 = 1)$. Results are averages over 10000 simulated data sets.

$E(\hat{\sigma}_i)$	SE($\hat{\sigma}_i$)	Bias	$E(\hat{\sigma}_i^2)/\sigma_i^2$
0.9993	1.6481	-0.0007	0.9987

that the average of the scale estimates over the samples is $E(\hat{\sigma_i}) \approx 1.0$, which indicates a very negligible bias, much smaller than the bias according to Mardia *et al.* 's (1999) formula (4.13), in the scale estimates when they are estimated from individual samples.

For model (4.15), we generate n, ($n = 5, 10, 100, 500, 1000$ and 10000), samples of size $m = 4$ from $C(\mu = i, \sigma^2 = 1)$ and then estimate the parameters in model (4.15). Results of simulations are summarised in Table 4.7. Results for each n are averages over 100 replicated data sets. Results in Table 4.7 show that there is substantial downward bias in the scale estimate, and this bias has similar pattern to that in the case of our proposed multi-scan model. The amount of bias depends on the value of n (number of spots) but the changes are negligible when n exceeds some large (say, 100) value. From the simulation results we

Table 4.7: Mean, estimated SE and bias of the scale estimates of simple Cauchy model: $y_{ij} \sim C(\mu_i = i, \sigma^2 = 1)$. Results for each n are averages over 100 simulated data sets.

n	$E(\hat{\sigma}_i)$	SE($\hat{\sigma}_i$)	Bias	$E(\hat{\sigma}_i^2)/\sigma_i^2$
5	0.7132	0.2998	−0.2868	0.5086
10	0.6776	0.1978	−0.3224	0.4591
100	0.6226	0.0540	−0.3774	0.3877
500	0.6231	0.0281	−0.3769	0.3882
1000	0.6260	0.0207	−0.3740	0.3918
10000	0.6271	0.0058	−0.3729	0.3933

found that $E(\hat{\sigma^2}) \approx 0.4\sigma^2$ for $n \geq 100$, $m = 4$ but each μ_i has been found to be approximately unbiased.

4.7 The censored Cauchy model

In this section, we investigate how the Cauchy model for combining multiple laser scans discussed in Section 4.3 can be improved. Although the censored mean function, $g(\mu_i\beta_j, \nu)$, nicely describes the distortion from linearity of observed spot summary data caused by pixel censoring, one major drawback of the Cauchy model is that it defines a density having probability above the censoring threshold T. That is, it ignores the fact that, like the individual pixel values, the spot summary data (y_{ij}) cannot exceed the threshold T. We therefore investigate an alternative model, the censored Cauchy model, to address this issue.

4.7.1 The model

Following equation (4.3), the probability distribution of y_{ij} under a censored Cauchy model with location $g(\mu_i\beta_j, \nu)$ and additive plus multiplicative scale $\sigma_{ij} = \sqrt{(\sigma_1^2 + \sigma_2^2\mu_i^2)\beta_j^2}$ for combining multiple laser scans can be expressed as

$$g(y_{ij}) = \begin{cases} f(y_{ij}), & \text{if } y_{ij} < T \\ S(T), & \text{if } y_{ij} = T, \end{cases} \tag{4.16}$$

where

$$f(y_{ij}) = (\pi\sigma_{ij})^{-1}\left[1 + \left\{\frac{y_{ij} - g(\mu_i\beta_j, \nu)}{\sigma_{ij}}\right\}^2\right]^{-1}, \tag{4.17}$$

is the density function, and

$$S(T) = \frac{1}{2} - \pi^{-1}\tan^{-1}\left\{\frac{T - g(\mu_i\beta_j, \nu)}{\sigma_{ij}}\right\}, \tag{4.18}$$

is the survival function (1 – cumulative distribution function) at T of a $C\left(g(\mu_i\beta_j,\nu),\ \sigma_{ij}^2\right)$ variate. By defining a censoring indicator

$$c_{ij} = \begin{cases} 1, & \text{if } y_{ij} < T \\ 0, & \text{if } y_{ij} = T, \end{cases} \tag{4.19}$$

the likelihood function for estimating the parameters of the censored Cauchy model (4.16) can be expressed as

$$\prod_{i=1}^{n}\prod_{j=1}^{m} f(y_{ij})^{c_{ij}} S(T)^{1-c_{ij}} \tag{4.20}$$

The corresponding log-likelihood function is

$$L_c(\mu,\beta,\sigma_1,\sigma_2,\nu) = \sum_{i=1}^{n} L_{ci}(\mu_i,\beta,\sigma_1,\sigma_2,\nu), \tag{4.21}$$

where

$$L_{ci}(\mu_i,\beta,\sigma_1,\sigma_2,\nu) = \sum_{j=1}^{m} \left[c_{ij}\log f(y_{ij}) + (1-c_{ij})\log S(T)\right]. \tag{4.22}$$

Maximum likelihood estimates of the parameters can be obtained by applying the alternating algorithm described in Section 4.3 for the Cauchy model by replacing L and L_i by L_c and L_{ci} respectively.

4.7.2 Application

We apply the censored Cauchy model (4.16) to both murine macrophase and iron-deficiency data described in Section 4.4. In practice, even in presence of heavy pixel censoring, very few of the spot averages (y_{ij}) are likely to be exactly equal to T, because the spot summary data are averages of pixel values within the segmented spots, which generally contain both censored and uncensored pixels. Table 4.8 shows the number of spot averages in murine macrophase and iron-deficiency data that are exactly equal to T. We see that array 1 of murine macrophage data and array 3 (Cy3) of iron-deficiency data do not have any observation equal to the censoring threshold T. Application of censored Cauchy model (4.16) to this data set would give the same result as that of the Cauchy model (4.4). We therefore apply the model to the other two arrays, array 2 of murine macrophage and array 3 (Cy5) of iron-deficiency data, having 3 and 19 observations equal to T respectively. Similar to the Cauchy model (4.4), we have used the simplex method of Nelder and Mead (1965), implemented in the IMSL routine `DUMPOL`, for estimating the censored Cauchy model (4.16). The results

Table 4.8: Numbers of spot averages equal to T in different laser scans of murine macrophase and iron-deficiency data.

Data set	Array	Scan	Number censored	Total spots
Murine macrophage	Array-1	1	0	9248
		2	0	9248
		3	0	9248
		4	0	9248
	Array-2	1	0	9248
		2	0	9248
		3	0	9248
		4	3	9248
Iron-deficiency	Array-3 (Cy3)	1	0	9984
		2	0	9984
		3	0	9984
	Array-3 (Cy3)	1	0	9984
		2	4	9984
		3	15	9984

of application are summarised in Table 4.9. Comparing these results with those of the Cauchy model (4.4) summarised in Tables 4.2 and 4.12, we see that the censored Cauchy model produces very similar results to the Cauchy model for these particular data sets. The fitted models are superimposed on the scatterplots of data against the estimated gene expressions in Figure 4.7. Observed spot averages and the fitted models for murine macrophage (array 2) data are rescaled by the corresponding scanning effects (β) before plotting for ease of comparison with Figure 4.3. Comparison of these plots with their Cauchy model counterparts in Figures 4.3 and 4.5 does not indicate any notable differences between the fits of the Cauchy and censored Cauchy models. Similar conclusion holds for

Table 4.9: Estimates of the scanning effects, scale parameters and ν of the censored Cauchy model applied to array 2 and array 3 (Cy5) of murine macrophage and iron-deficiency data respectively.

	Scanning effects			Scale		
Data set	β_2	β_3	β_4	σ_1	σ_2	ν
Murine macrophage [Array-2]	1.71	2.71	4.53	5.29	0.0053	0.26
Iron-deficiency [Array-3 (Cy5)]	4.36	16.49	–	2.93	0.0126	0.17

the between replicate variations as calculated according to formula (4.7). Results for array 2 of murine macrophage data obtained from Cauchy and censored Cauchy models can be compared from Tables 4.3 and 4.10 respectively. We see

Table 4.10: Comparison of between replicate variation in the data and fit for the censored Cauchy model applied to array 2 of murine macrophage data.

	Between replicate variation				
Data set	$S(\hat{\mu})$	$S(y_{.1}/\hat{\beta}_1)$	$S(y_{.2}/\hat{\beta}_2)$	$S(y_{.3}/\hat{\beta}_3)$	$S(y_{.4}/\hat{\beta}_4)$
Array-2	858	1684	1769	882	864

that the results are almost identical for the two models, and the between replicate variation in the estimated gene expressions is smaller than that in the per-scan observed spot averages.

4.7.3 Simulation study

In this section, we investigated the properties of the maximum likelihood estimators of the censored Cauchy model for combining multiple laser scans through simulation study. For convenience of comparison, we used the same set of true parameter values as used for the simulation study of the Cauchy model in Section 4.5 for generating data from the censored Cauchy model. Also we used the same starting seed for the two simulation studies. We simulated 100 data sets from the Cauchy model (4.4), and considered any observations greater than or equal to $T(= 65535)$ as censored at T. On an average, we found 0.05% censored observations in 100 replicated data sets. Table 4.11 summarises the true parameter values as well as the estimated biases and standard errors of the estimates of the structural parameters. Again, we see that except for the scale parameters, σ_1 and σ_2, all the structural parameters are estimated with high precision

Table 4.11: Estimated biases and standard errors of the maximum-likelihood estimates of the censored Cauchy model for combining multiple laser scans. The results are based on 100 simulated data sets.

	Parameters					
	β_2	β_3	β_4	σ_1	σ_2	ν
True	1.71	2.71	4.53	5.36	0.0051	0.27
Bias	−0.00004	−0.00007	−0.00011	−2.020	−0.00184	0.00068
SE	0.00037	0.00072	0.00115	0.054	0.00009	0.00218

and negligible bias. Comparison of the simulation results for the Cauchy model (Table 4.5) and the censored Cauchy model (Table 4.11) does not show any notable differences between the properties of the maximum likelihood estimators of the two models. Figure 4.8 shows the estimated % biases and % standard errors in the gene expression estimates. Except for the two points at the upper end, estimated biases in the gene expression estimates obtained for the censored Cauchy model are consistent with the corresponding biases for the Cauchy model (see Figure 4.6). Biases for the censored Cauchy models have also been found to be in an acceptable range, less than 0.5% in most cases, and are symmetrically distributed about the zero reference line. The two gene expression values at the upper end showing relatively high positive biases in the censored Cauchy model estimates also have higher standard errors (see Figure 4.8, bottom).

4.7.4 Investigating impact of higher level of censoring

In the data sets used in the thesis, it was noted that a large number of spot averages are not exactly equal to T even if they are clearly affected by pixel censoring. Therefore, we investigated a broader definition of censoring to see the impact on the results of censored Cauchy model. In the analysis, in particular, values above 65000 were considered as censored. The parameter estimates applied to array 3 (Cy5) of iron-deficiency data, which have 220 values as censored according to the above definition, are presented in the following table.

Table 4.12: Estimates of the scanning effects, scale parameters and ν applied to array 3 (Cy5) of iron-deficiency data treating values above 65000 as censored.

	Scanning effects		Scale		
Data set	β_2	β_3	σ_1	σ_2	ν
Iron-deficiency [Array 3 (Cy5)] data	4.36	16.47	2.87	0.0135	0.17

Fortunately, these are very similar to the previous estimates. But because of subjectivity involved in this approach we would not advocate it as a final modelling approach. Alternative possible approach, using median, is briefly discussed in Section 4.8.

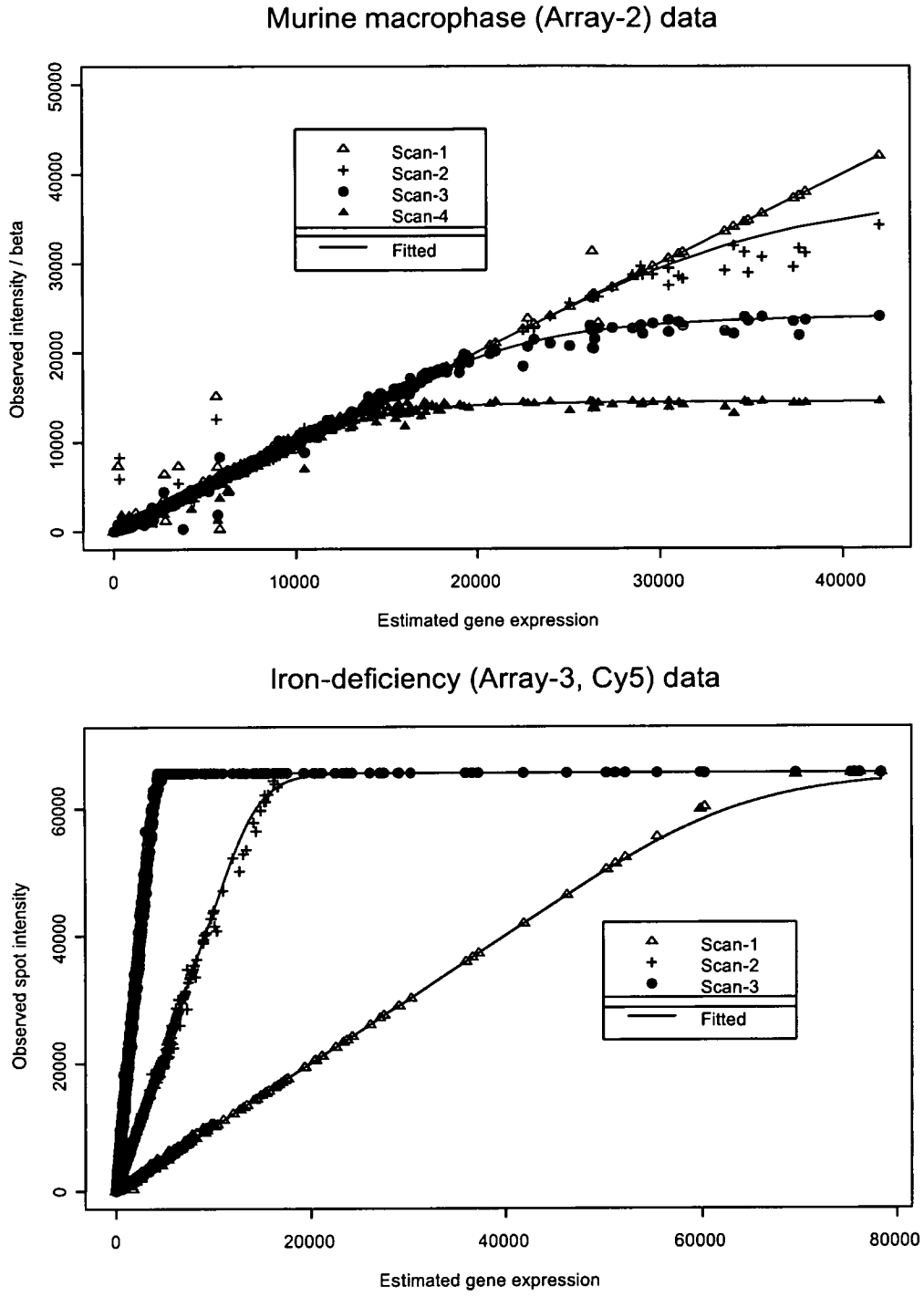


Figure 4.7: Rescaled intensities $(y_{ij}/\hat{\beta}_j)$ plotted against estimated gene expressions $(\hat{\mu}_i)$ for array 2 of murine macrophase data (top), and observed spot intensities (y_{ij}) plotted against estimated gene expressions $(\hat{\mu}_i)$ for array 3 (Cy5) of iron-deficiency data (bottom). The solid lines indicate the corresponding fitted models.

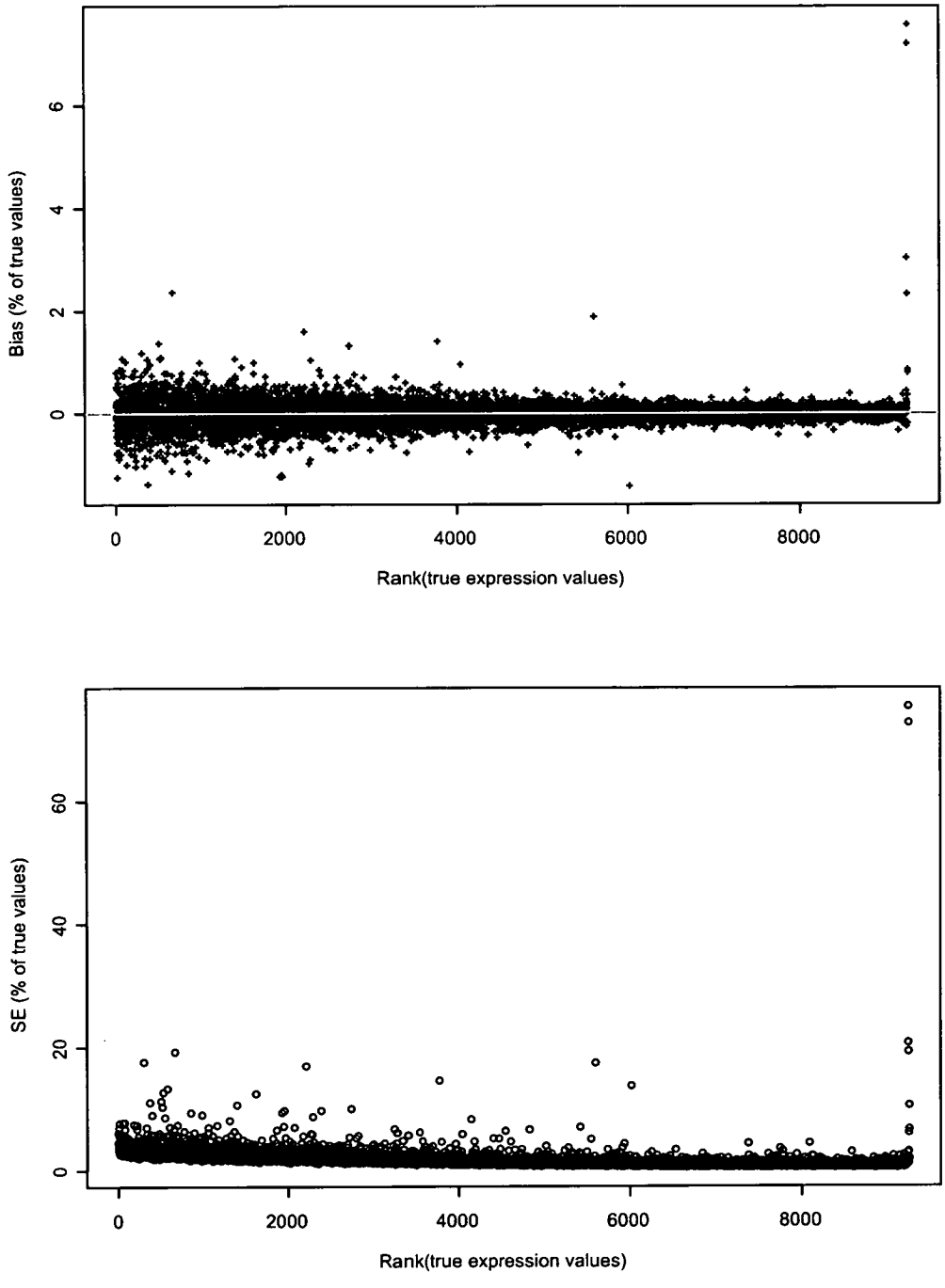


Figure 4.8: Percentage of bias and standard error against the rank of true gene expression values for the maximum likelihood estimates of the censored Cauchy model.

4.8 Discussion and conclusions

Microarray gene expression data obtained as the output of typical image analysis steps are contaminated, in addition to other factors, by the scanner's intrinsic noise level at the lower end, and by the pixel censoring at the upper end. As the problems at the two ends are in conflict, no unique scanner setting is optimal. Moreover, there is no objective guideline to date for choosing optimum scanner setting to address these issues. It therefore seems reasonable to consider multiple scanning, some at relatively lower sensitivity levels, ensuring that there is no censoring at the upper end, and the others at higher sensitivity levels, ensuring the visibility of the weakly expressed genes over the scanner's intrinsic noise level, and combine the information together to get final gene expression measures. The simplest approach of combining the data through simple or weighted average over the scans will give biased result as some individual scans of data are likely to be affected by pixel censoring.

The proposed method can successfully combine the data of multiple scanning to get improved gene expression measures throughout the entire range of intensity data. Although application of the Cauchy and censored Cauchy models to the data sets used in this thesis produces very similar results, the censored Cauchy model is a more realistic choice because it takes account of the fact that spot averages cannot exceed the censoring threshold T, and in case of moderate or heavy censoring, censored Cauchy model can be expected to give better results than the Cauchy model.

The simulation results suggest that the model is capable of estimating gene expressions adjusting for outliers and pixel censoring with reasonable precision and negligible bias. One strength of the model is that the location function specified in Section 3.6.1 explicitly captures the trend of the possibly censored spot summary data. Also, the derivation of the function has a natural correspondence with the data generation mechanism of microarray scanners.

The choice of the censored Cauchy distribution for handling outliers proved to be better than the robust methods with which we have experimented. The censored Cauchy distribution is also a reasonable choice on the grounds of simplicity and objectiveness. Among the few available methods of its kind in the literature, Dudley *et al.*'s (2002) method also considers multiple scan data but loses information discarding data outside the linear range. The method of Wit and McClure (2003) considers single scan data and does not suggest a general pixel distribution. The authors note that their method may produce unstable estimates as it estimates two parameters using only three summary statistics,

mean, median and variance.

We considered how the model may be extended. A natural extension would be to replace the censored Cauchy distribution by a censored t-distribution. This would introduce an additional degrees of freedom parameter which would ideally be estimated from the data, and depend on the tail behaviour. We have conducted some simulation experiments with such a model in the uncensored case. The bias in the estimation of the scale parameter noted in Section 4.5 for the Cauchy model is also present in the estimation of the scale parameter for the t-distribution model but additionally there is a corresponding bias in the estimation of the degrees of freedom parameter (see Table 3.13). However, we found that we get very similar maximum likelihood estimates of the μ_i as with the Cauchy model and therefore there was little advantage in using the slightly more complex model.

For the current data only mean values are available which are censored only if all the pixels belonging to a spot are censored. Alternative approach would be to model the median and this might have the advantage that a spot would be clearly identified as being censored if 50% or more of all the pixels are censored.

Chapter 5

Nonparametric location and scale normalisation

5.1 Introduction

Two of the most discussed issues in microarray literature are normalisation and variance stabilisation of intensity measurements. Due to variations in sample treatment, labelling, dye efficiency and detection, the fluorescence intensities can in general not be compared directly, but only after appropriate calibration, which is termed "normalisation". The purpose of normalisation is to identify and remove sources of systematic variation, *e.g.*, different labelling efficiencies and scanning properties of the Cy3 and Cy5 dyes; different scanning parameters, such as PMT settings; print-tip, spatial or plate effects, in the measured fluorescence intensities. The simplest approach to within-slide normalisation is to subtract a constant from all intensity log-ratios, typically their mean or median, to force the distribution of the intensity log ratios to have a location of zero for each slide. Such global normalisation methods cannot normalise the intensity data for some locally active artefacts, *e.g.*, print-tip effects, spatial or intensity dependent dye biases. Dudoit *et al.* (2002) proposed more flexible normalisation methods which allow the normalisation function to depend on a number of predictor variables, such as the average spot intensity (x), location and plate origin. They used *loess*, a robust locally weighted regression (Cleveland, 1979; Cleveland and Devlin, 1988), of the log-ratio (y) on the predictor variables. Semiparametric approaches have also been suggested for correcting for trends in log-ratio data (Fan *et al.*, 2005; Huang *et al.*, 2005; Ma *et al.*, 2006).

The other common problem is the variance inhomogeneity in the sense that the variance of the measured intensity of a spot depends on that spot's average intensity, which poses complexity in the analysis of microarray data. Many commonly used statistical methodologies, such as regression or the analysis of variance, are

based on the assumption that the data are normally or at least symmetrically distributed with constant variance. If these assumptions are violated, the statistician may choose either to develop some new statistical technique to account for the specific ways in which the data fail to comply with the assumptions, or to transform the data. Where possible, data transformation is generally the easier of the two options. Unlike for the location normalisation, parametric models such as additive plus multiplicative variance model have been preferred for scale normalisation and several transformation methods, *e.g.*, log, started logarithm, arcsinh or glog, log-linear hybrid etc. have emerged as a result. Log transformation is commonly used for variance stabilisation based on the assumption of multiplicative variance model, which is not generally the case. The supposed simplicity of interpretation of log ratios provided a major justification for the use of log transformation on microarray data. There are however two major drawbacks of log transformation in microarray applications. First, background corrected intensity data frequently have non-positive values for which the log-transformation is not defined. This however is not the problem of the log-transformation, but of the crude and in-appropriate nature of background correction methods commonly used by the microarray community. Secondly, log-transformations tend to inflate the variance at the lower gene expression levels, because the multiplicative model is not generally adequate to describe the variance of microarray data.

The "started logarithm" (Rocke and Durbin, 2003) is a modification of logarithmic transformation to avoid negative arguments. The idea is to add a constant to all of the values before taking logarithm. This transformation does not completely stabilise the variance when the variance is additive plus multiplicative. The authors have given the conditions, details described in Section 2.6.2, which minimise the maximum deviation from constancy.

Rocke and Durbin (2003) considered another variant of logarithmic transformation that may be appropriate for microarray data. It is the log-linear hybrid transformation, originally suggested by Holder *et al.* (2001). In this approach, the transformation is taken to be $\log(Z)$ for Z greater than some cutoff k and a linear function $cZ + d$, where c and d are constants, below that cutoff. This eliminates the singularity at zero. The constants c and d are chosen such that the transformation is continuous with continuous derivative at k (Section 2.6.2).

Additive plus multiplicative model has been suggested as a more realistic variance model for microarray data (*e.g.*, Rocke and Durbin, 2001, Huber *et al.*, 2002). Inverse hyperbolic sine (arcsinh), variously known as generalised logarithm (glog), transformation stabilises the variance of additive plus multiplicative structure. This transformation, introduced independently by several research groups

(Munson, 2001; Huber *et al.*, 2002; Durbin *et al.*, 2002), also overcomes the limitations of log transformation and stabilises variance of additive multiplicative structure to the first order, meaning that the variance is almost constant no matter what the mean might be. This transformation converges to $\log(Z)$ for large Z, and is approximately linear at 0 (Durbin, 2002).

However, as can be seen from experimental data, variance of microarray data may be of a more complex nature than can be generally accommodated by any particular parametric model. In this chapter we propose and evaluate a new nonparametric approach that incorporates location and scale normalization simultaneously using Generalised Additive Models for Location, Scale and Shape (GAMLSS, Rigby and Stasinopoulos, 2005). The methods presented in this chapter has been described briefly in Khondoker *et al.* (2006b). GAMLSS is the extended and more flexible version of their Mean and Dispersion Additive Models (MADAM) (Rigby and Stasinopoulos, 1996; Stasinopoulos *et al.*, 2000) and the model for fitting smooth centile curves to skew and kurtotic data (Rigby and Stasinopoulos, 2004).

5.2 Generalised additive models for location, scale and shape

Generalised additive models for location scale and shape (GAMLSS) are a flexible class of statistical models for univariate regressions. GAMLSS generalises a wide variety of statistical models that are considered as separate classes in the statistical literature, *e.g.*, Generalised Linear Mixed Models (GLMM) and Generalised Additive Mixed Models (GAMM), which in turn are more general than the Generalised Linear Models (GLM) and Generalised Additive Models (GAM) respectively.

GAMLSS allows a very general family of distributions for the response. In addition to the location parameter, other parameters of the conditional distribution of the response, such as the scale and shape parameters can be modelled as parametric and/or additive nonparametric smooth functions of explanatory variables and/or as random-effect terms.

Suppose that $f(y|\theta)$ be the probability density/mass function of the response variable y for given p-dimensional parameter vector $\theta^T = (\theta_1, \cdots, \theta_p)$. Let $y^T = (y_1, \cdots, y_n)$ be the vector of the n observations of the response variable y. The model assumes that, for $i = 1, \cdots, n$, observations y_i are independent conditional on θ^i, with probability density/mass function $f(y_i|\theta^i)$, where $\theta^{iT} = (\theta_{i1}, \cdots, \theta_{ip})$ is a vector of p parameters related to explanatory variables and random effects.

For $k = 1, \cdots, p$, let $g_k(.)$ be a known monotonic link function relating θ_k to explanatory variables and random effects through an additive model given by

$$g_k(\theta_k) = \eta_k = X_k\beta_k + \sum_{j=1}^{J_k} Z_{jk}\gamma_{jk} \tag{5.1}$$

where θ_k and η_k are vectors of length n, *e.g.*, $\theta_k^T = (\theta_{1k}, \cdots \theta_{nk})$, $\beta_k^T = (\beta_{1k}, \cdots, \beta_{J_k'k})$ is a parameter vector of length J_k', X_k is a known design matrix of order $n \times J_k'$, Z_{jk} is a fixed known $n \times q_{jk}$ design matrix and γ_{jk} is a q_{jk}-dimensional random variable. The class of models (5.1) is called GAMLSS.

If, for $k = 1, \cdots, p$, $J_k = 0$ then (5.1) reduces to a fully parametric class given by

$$g_k(\theta_k) = \eta_k = X_k\beta_k. \tag{5.2}$$

If $Z_{jk} = I_n$, where I_n is an $n \times n$ identity matrix, and $\gamma_{jk} = s_{jk} = s_{jk}(x_{jk})$ for all combinations of j and k in (5.1), this gives

$$g_k(\theta_k) = \eta_k = X_k\beta_k + \sum_{j=1}^{J_k} s_{jk}(x_{jk}) \tag{5.3}$$

where x_{jk} for $j = 1, \cdots, J_k$ and $k = 1, \cdots, p$ are vectors of length n. The function s_{jk} is an unknown function of the explanatory variable X_{jk} and $s_{jk} = s_{jk}(x_{jk})$ is the vector which evaluates the function s_{jk} at x_{jk}. The explanatory vectors x_{jk} are assumed to be known. The models in equation (5.3) are called the semiparametric GAMLSS. The class of models (5.3) is an important special case of (5.1). If $Z_{jk} = I_n$ and $\gamma_{jk} = s_{jk} = s_{jk}(x_{jk})$ for specific combinations of j and k in (5.1), then the resulting models contain parametric, nonparametric and random-effect terms. The first two population parameters θ_1 and θ_2 in (5.1) are usually characterized as location and scale parameters, denoted here by μ and σ, whereas the remaining parameter(s), if any, are characterized as shape parameters, although the models may be applied more generally to the parameters of any population distribution.

Maximum penalised likelihood estimation is used to fit the nonparametric models. Additive terms in the models are fitted by a backfitting algorithm. The estimation algorithm has been described in Appendix B of Rigby and Stasinopoulos (2005).

It is possible to use different smoothing methods, *e.g.*,

- smoothing splines,
- penalised splines or p-splines (Eilers and Marx, 1996),
- loess (Cleveland, 1979; Cleveland and Devlin, 1988) etc.

to fit the nonparametric models in (5.3). The R package `gamlss` is publicly available at `http://www.londonmet.ac.uk/gamlss/` for implementing the estimation algorithm. Instructions on how to use the package can be found in Stasinopoulos *et al.* (2004).

5.3 Nonparametric location and scale normalisation using GAMLSS

Let $y = \log(I_2/I_1)$ and $x = \log(\sqrt{I_1 I_2})$ be the log ratio and overall log spot intensity respectively for a usual cDNA experiment where I_1 and I_2 represent intensity data corresponding to control and experimental samples respectively. We assume that the conditional distribution of y, for the combined differentially and non-differentially expressed genes, given x can be approximated by a t-distribution with intensity dependent location $\mu(x)$ and scale $\sigma(x)$ parameters respectively and a constant shape parameter ν, the degrees of freedom of the distribution. That is

$$y|x \sim t_\nu(\mu(x),\ \sigma(x)). \tag{5.4}$$

This is a special case of a GAMLSS model (Rigby and Stasinopoulos, 2005). The functions $\mu(x)$ and $\sigma(x)$ are not gene specific and are supposed to capture the intensity-dependent trends in the location and variability of the data, which are induced by the differential behaviour of Cy3 and Cy5 dyes, termed dye bias. Maximum penalised likelihood may be used to obtain estimates of the functions $\mu(x)$ and $\sigma(x)$. In loess smoothing the user-defined parameter f, called span, is the fraction of the data used for smoothing at each point. From among several smoothing options available, we choose p-splines with effective degrees of freedom equivalent to that of loess span of 40%, which gives reasonable amount of smoothing for both location and scale models. We prefer p-splines because it is less computationally expensive than smoothing splines and p-splines smoothing has some desirable properties over loess, *e.g.*, it is free from boundary effects and conserves the moments of the data. The normalised data, from the fitted model (5.4), can be defined as the standardised residuals given by

$$z_G = \frac{(y - \hat{\mu}(x))}{\hat{\sigma}(x)}.$$

GAMLSS normalisation model (5.4) is not necessarily restricted to log-ratio (y) vs. average log intensity (x) data. The model can be applied, for example, to the arsinh transformed data (see Section 5.6). This approach may be useful when log transformation can not be applied to all data points because of non-positive values in the background corrected intensity data. We investigated such

normalisation through simulation study in Section 5.6, where we apply GAMLSS to arcsinh transformed experimental vs. control samples.

Although GAMLSS, in its current form, can only be used for within-slide normalisation, GAMLSS normalised data are comparable across arrays as they are expected to have unit scale for all individual arrays.

5.4 Huber *et al.*'s (2002) parametric normalisation method

We compare the performance of the proposed method with Huber *et al.*'s (2002) arcsinh variance stabilising transformation (AVST) method through application of the methods to two different data sets and through simulation studies. We call the method AVST to distinguish it from `vsn` which is the name of the software package in R that implements the method. The Huber *et al.*'s (2002) parametric normalisation method (AVST) is derived from a model of the variance-versus-mean dependence of the linearly calibrated intensity data using a variance stabilising transformation $h(.)$, called arcsinh. The variance of the calibrated intensity data is assumed to be a quadratic function of the mean. The transformation $h(.)$ is derived such that the variance of the transformed data is approximately independent of the mean. For a two-sample comparative experiment, the difference of the transformed data, called difference statistic (z_A), is given by

$$z_A = h_2(I_2) - h_1(I_1) = \text{arcsinh}(a_2 + b_2 I_2) - \text{arcsinh}(a_1 + b_1 I_1), \tag{5.5}$$

where a_j and b_j, $j = 1, 2$ are transformation parameters to be estimated from the data. The difference statistic represents the changes in expression levels of the genes between samples 1 and 2. For large intensities, the arcsinh transformation becomes equivalent to the logarithmic transformation and therefore z_A becomes equivalent to log-ratio (y). For non-differentially expressed genes, z_A is assumed to be distributed as normal with zero mean and constant variance. That is

$$(z_A)_i \sim N(0, \delta^2), \quad i \in K, \tag{5.6}$$

where K is the set of non-differentially expressed genes. The set K is determined iteratively by Least Trimmed Squares (LTS) regression (Rousseuw and Leroy, 1987). We have used LTS quantile of 90% in our applications. For fixed K, the parameters a_j and b_j, $(j = 1, 2)$ are estimated numerically by maximizing the profile likelihood. The estimated difference statistic is used for inference on differential expression. The software is available as an R package at `http://www.bioconductor.org/`.

5.5 Applications

We provide applications of the proposed GAMLSS normalisation method to two different data sets:

(i) lymphoma data, and

(ii) iron-deficiency data.

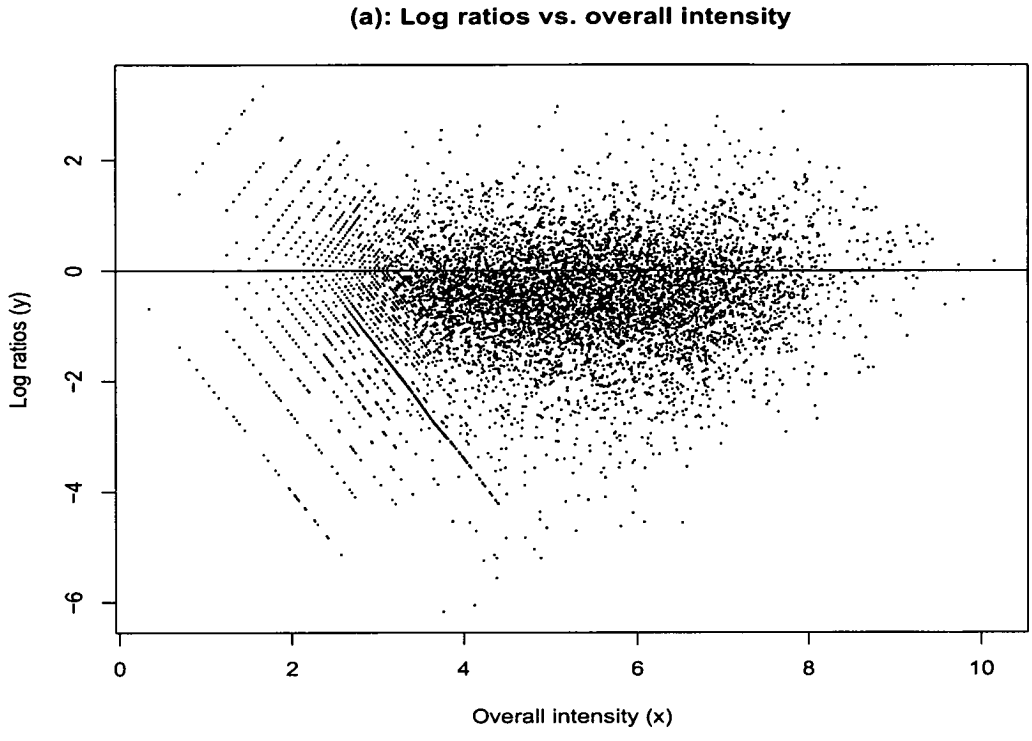
We also compare the results of applying our method to these data sets with that of the parametric AVST method.

5.5.1 Lymphoma data

The lymphoma data is obtained from the Huber *et al.*'s (2002) R package `vsn`, which is available at `http://www.bioconductor.org/`. The data set is based on a series of 8 cDNA arrays on which different lymphoma samples were hybridised together with a reference cDNA (Alizadeh *et al.*, 2000). We use the data on one array, named lc7b047, containing 9216 spots for illustration of the method. Data on treatment and reference samples are background corrected and have some negative values. We replace all negative and zero values by 10 before computing the log ratios. Plots of the log ratio $y = \log(I_2/I_1)$ against the overall log-intensity $x = \log(\sqrt{I_2 I_1})$ and rank of x are shown in Figure 5.1. Systematic patterns in the lower expression region are the effect of artificial replacement of negative and zero values. There is some indication of non-linear trend in the location of the log ratios. Also there is indication of variance inhomogeneity in the data.

Rigby and Stasinopoulos (2005) suggested different criteria such as Akaike Information Criterion (AIC), Generalised Akaike Information Criterion (GAIC) and Schwarz Bayesian Information Criterion (SBC), (Schwarz, 1978), for optimising the amount of smoothing in the fit of the GAMLSS models. For the data set at hand optimal fit using p-splines of the models (5.4) according to GAIC with penalty 2.5 and SBC leads to effective degrees of freedom of about 30 and 21 respectively. Fitted optimal models, as shown in Figure 5.2, according to these criteria appear to give too localised fit.

The span of loess for location smoothing of microarray data are typically suggested to be between 20% and 40% (Dudoit *et al.*, 2002; Yang *et al.*, 2002). The GAMLSS (5.4) fitted to the data at hand with loess ($f = 40\%$) as the smoothing option for nonparametric models results in effective degrees of freedom equal to 13.



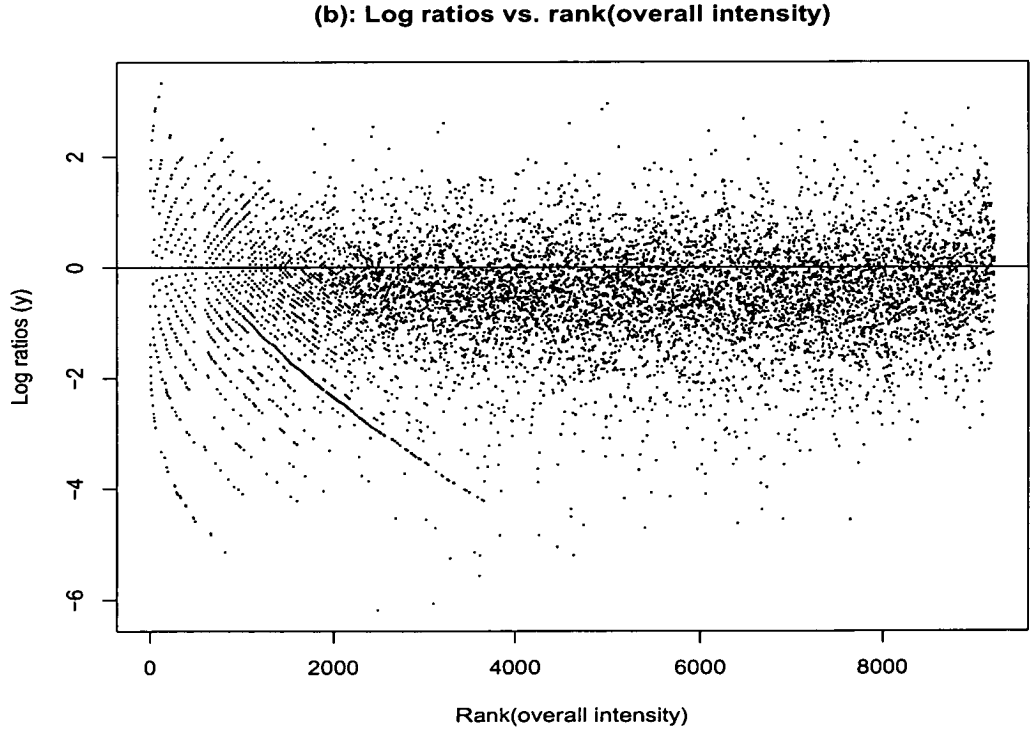


Figure 5.1: Log ratio vs. (a) overall intensity; (b) rank (overall intensity) for one array of lymphoma data.

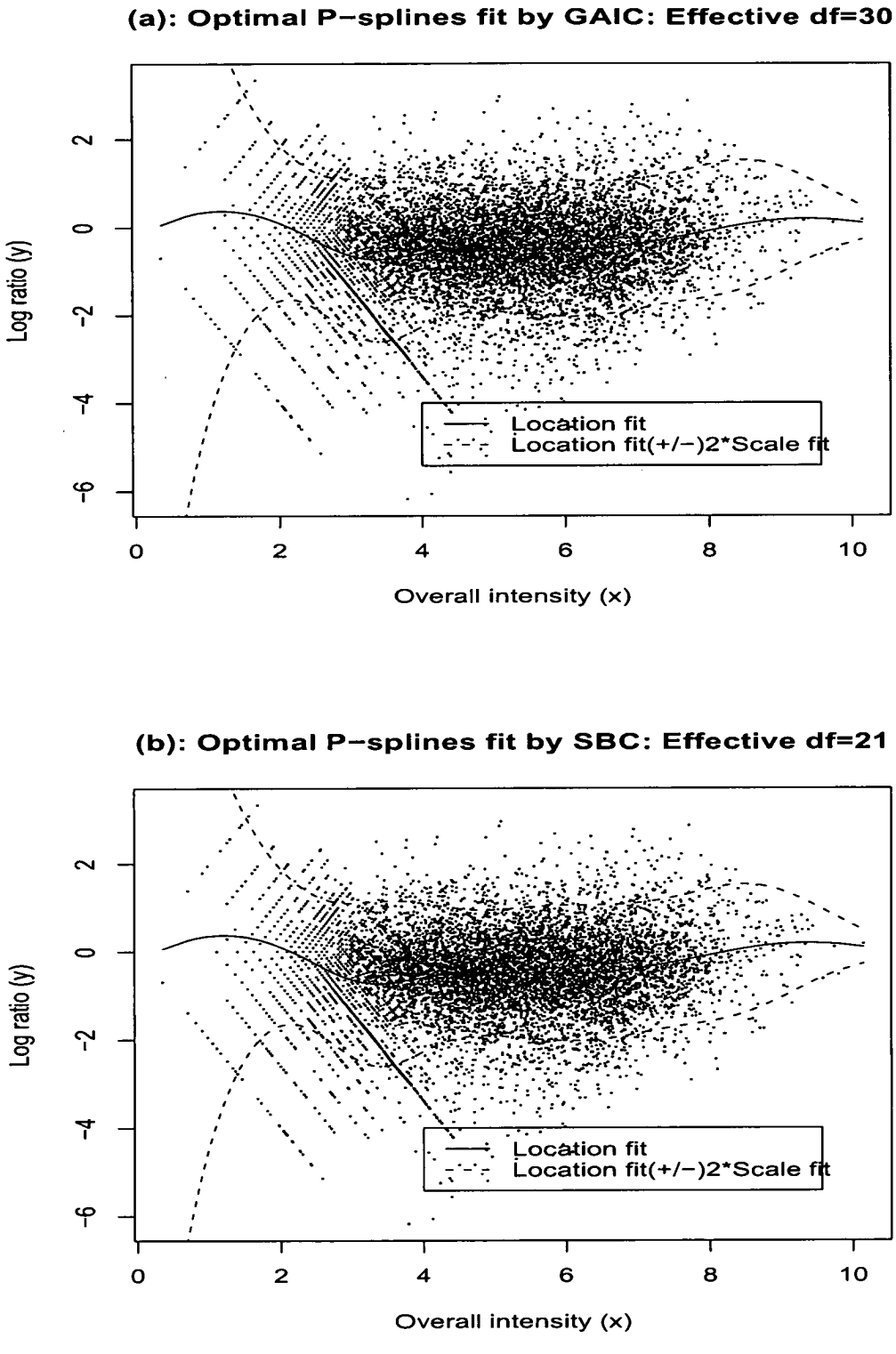


Figure 5.2: Optimal p-splines fit (a) by Generalised Akaike Information Criterion (GAIC) with penalty 2.5; (b) Schwarz Bayesian Information Criterion (SBC). The solid line is fitted location $\hat{\mu}(x)$ and dashed lines are the limits $\hat{\mu}(x) \pm 2\hat{\sigma}(x)$.

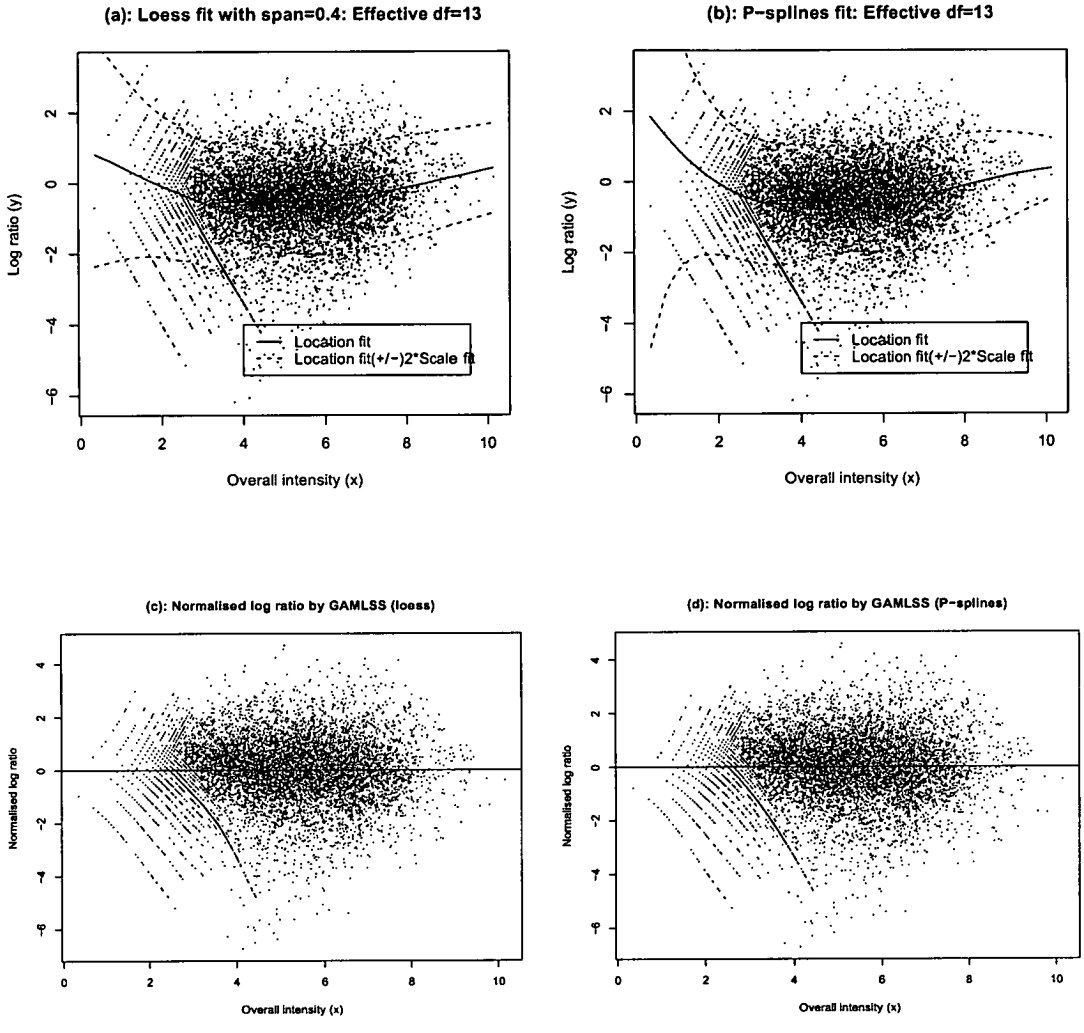


Figure 5.3: GAMLSS fit with effective df = 13: (a) loess smoothing with span = 0.4 (effective df = 13); (b) p-splines smoothing; (c) Normalised log ratios for the fit in (a); (d) Normalised log ratios for the fit in (b). In (a) and (b) the solid lines are the fitted locations $\hat{\mu}(x)$ and the dashed lines are the limits $\hat{\mu}(x) \pm 2\hat{\sigma}(x)$. In (c) and (d) solid lines are the horizontal zero references.

We therefore fit (5.4) using p-splines with effective degrees of freedom equivalent to that of loess, which gives reasonable amount of smoothing for both location and scale models. Fitted models and corresponding normalised data (standardised residuals) for both smoothing options are shown in Figure 5.3. Except at the boundaries where there are a few observations, both smoothing options give similar fit. We compare the results of GAMLSS analysis applied to the lymphoma data to that of AVST analysis. The AVST method is implemented using the R package `vsn` to the same data described in Section 5.5.1. For identifying differential expressions we consider cut-off points of 2 and 3 times the standard deviation of the difference statistic (z_A) of AVST method. AVST method does not give direct estimate of standard deviation of the difference statistic but suggests to estimate robustly from the empirical distribution. We estimate standard deviation robustly using Median Absolute Deviation (MAD) given by

$$\hat{\tau} = \frac{\text{Median}(|z_i - m|)}{\Phi^{-1}(0.75)}, \tag{5.7}$$

where m denotes the median of either z_A or z_G. The estimate is adjusted by the factor $\Phi^{-1}(0.75)$, the value of the inverse standard normal distribution function at the point 0.75, for asymptotically normal consistency (Huber, 1981). For the lymphoma data, we estimate $\hat{\tau}_A = 0.79$. Normalised data using GAMLSS should approximately follow a standard t-distribution with degrees of freedom ν which, using p-splines with effective df 13, is estimated to be 5.42. For GAMLSS normalised data we use cut-off points of 2 and 3 for detecting differential expressions. A summary of the comparison on the inference of differential expression by the two methods are given in Table 5.1 and Figure 5.4. We see that GAMLSS detects slightly more genes as differentially expressed for both 2 and 3 standard deviation cut-offs (Table 5.1). For GAMLSS normalised data there are 966 genes outside 2 standard deviation cut-off whereas the value is 873, 807 of which are common in both methods, for AVST method. Corresponding values for 3 standard deviation cut-off are 250 and 225, 209 of which are common.

Table 5.1: Comparison of differential expression on lymphoma data.

		2SD cut-off			3SD cut-off		
		GAMLSS			GAMLSS		
		Significant	Not significant	Total	Significant	Not significant	Total
AVST	Significant	807	66	873	209	16	225
	Not Significant	159	8184	8343	41	8950	8991
	Total	966	8250	9216	250	8966	9216

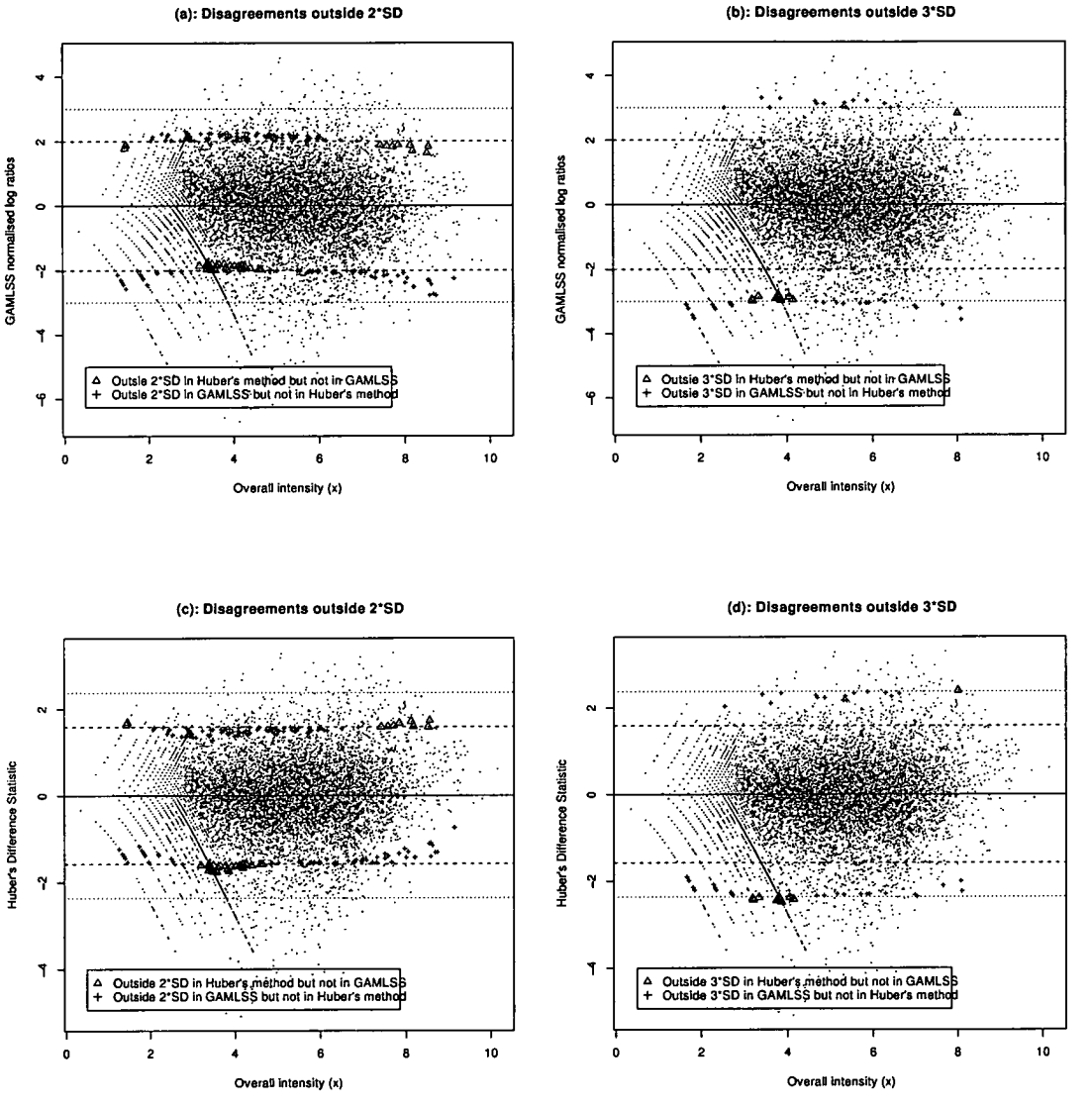


Figure 5.4: Summary of comparisons of differential expressions:(a) disagreements outside 2SD in GAMLSS normalised log ratio vs. x plot; (b) disagreements outside 3SD in GAMLSS normalised log ratio x plot; (c) disagreements outside 2SD in z_A vs. x plot; (d) disagreements outside 3SD in z_A vs. x plot.

Disagreements between the methods are highlighted in Figure 5.4. Of 966 genes detected as differentially expressed by GAMLSS (2SD cut-off), 159 are not expressed according to AVST method. On the other hand, of 873 genes detected as differentially expressed by AVST method (2SD cut-off), 66 genes are not expressed according to GAMLSS. The corresponding disagreement values are 41 and 16 respectively for 3SD cut-off.

5.5.2 Iron-deficiency data

In this section we apply the method to iron-deficiency data briefly described in Section 4.4.2. After deleting the blank spots and the missing (non-detected) values across slides, there are intensity measurements for 9968 spots on each of the arrays. The correspondence between array, organ, dye and sample is given in Table 5.2. The image analysis has been carried out using `MWG MVI PRO 2.6.0`

Table 5.2: Correspondence between array, organ, dye and sample.

Array	Organ	Dye	Sample	Array	Organ	Dye	Sample
1	Liver	Cy3	Control	9	Kidney	Cy3	Treatment
		Cy5	Treatment			Cy5	Control
2	Liver	Cy3	Treatment	10	Kidney	Cy3	Treatment
		Cy5	Control			Cy5	Control
3	Liver	Cy3	Control	11	Kidney	Cy3	Treatment
		Cy5	Treatment			Cy5	Control
4	Liver	Cy3	Treatment	12	Kidney	Cy3	Treatment
		Cy5	Control			Cy5	Control
5	Liver	Cy3	Treatment	13	Kidney	Cy3	Control
		Cy5	Control			Cy5	Treatment
6	Liver	Cy3	Control	14	Kidney	Cy3	Control
		Cy5	Treatment			Cy5	Treatment
7	Liver	Cy3	Treatment	15	Kidney	Cy3	Control
		Cy5	Control			Cy5	Treatment
8	Liver	Cy3	Control	16	Kidney	Cy3	Control
		Cy5	Treatment			Cy5	Treatment

(Bell, 2003). We have compared the results of GAMLSS analyses with that of AVST analyses applied to this data set. We fit model (5.4) individually to all 16 arrays of the data using p-splines as the smoothing option with effective degrees of freedom equivalent to that of loess span of 40%. Fitted models to all liver and kidney samples are shown in Figures 5.5 and 5.6 respectively. The model appears able to capture the trends in location and variability in the data. To see how GAMLSS compares with AVST in normalising gene expression data, we plot GAMLSS normalised log-ratios and AVST difference statistics side by side against the overall log-intensity (x) data for one microarray (array 2) in Figure 5.7. In

terms of both location and variability, GAMLSS normalised data appears much more homogeneous than that of AVST normalised data. We consider cut-off points of 2 and 3 times the estimated standard deviations of the corresponding normalised data to label genes as differentially expressed. We use MAD as the estimate of the standard deviation. For array 2 data, we estimate standard deviations of GAMLSS normalised log-ratio and AVST difference statistic as 1.05 and 1.02 respectively. A cross classification of the summary of the genes identified as differentially expressed by the two methods are given in Table 5.3. The disagreements between the methods are highlighted in Figure 5.7. It is observed

Table 5.3: Comparison of differential expression on array 2 of iron-deficiency data.

		2SD cut-off			3SD cut-off		
		GAMLSS			GAMLSS		
		Significant	Not significant	Total	Significant	Not significant	Total
AVST	Significant	279	49	328	15	1	16
	Not Significant	222	9418	9640	56	9896	9952
	Total	501	9467	9968	71	9897	9968

that GAMLSS identifies comparatively more genes as differentially expressed. It is also seen that most of the genes labelled as differentially expressed by AVST method are also identified as differentially expressed by GAMLSS. Figure 5.7 shows that, for 2SD cutoff, genes identified as differentially expressed after AVST normalisation, but not after GAMLSS normalisation, are the positive ones in the middle of the x-range and negative ones at the two ends of the range, and are likely to be artefacts due to the failure of AVST to fully correct for location and scale trends.

The experiment has both biological and technical replicates: each of the liver and kidney groups has 4 dye-swapped biological replicates. We analysed the normalised liver and kidney data separately, using simple t-tests to identify differentially expressed genes. The number of detected genes by either or both methods, using 5% and 1% thresholds, are shown in Table 5.4. We see that GAMLSS normalisation identifies slightly more genes than AVST for kidney data whereas AVST normalisation picks comparatively more genes for liver data. However, mindful of the observation from Figure 5.7 that differential expression may be falsely identified if normalisation fails to fully correct for location and scale trends, the number of detected genes may not be a good measure of success of a method.

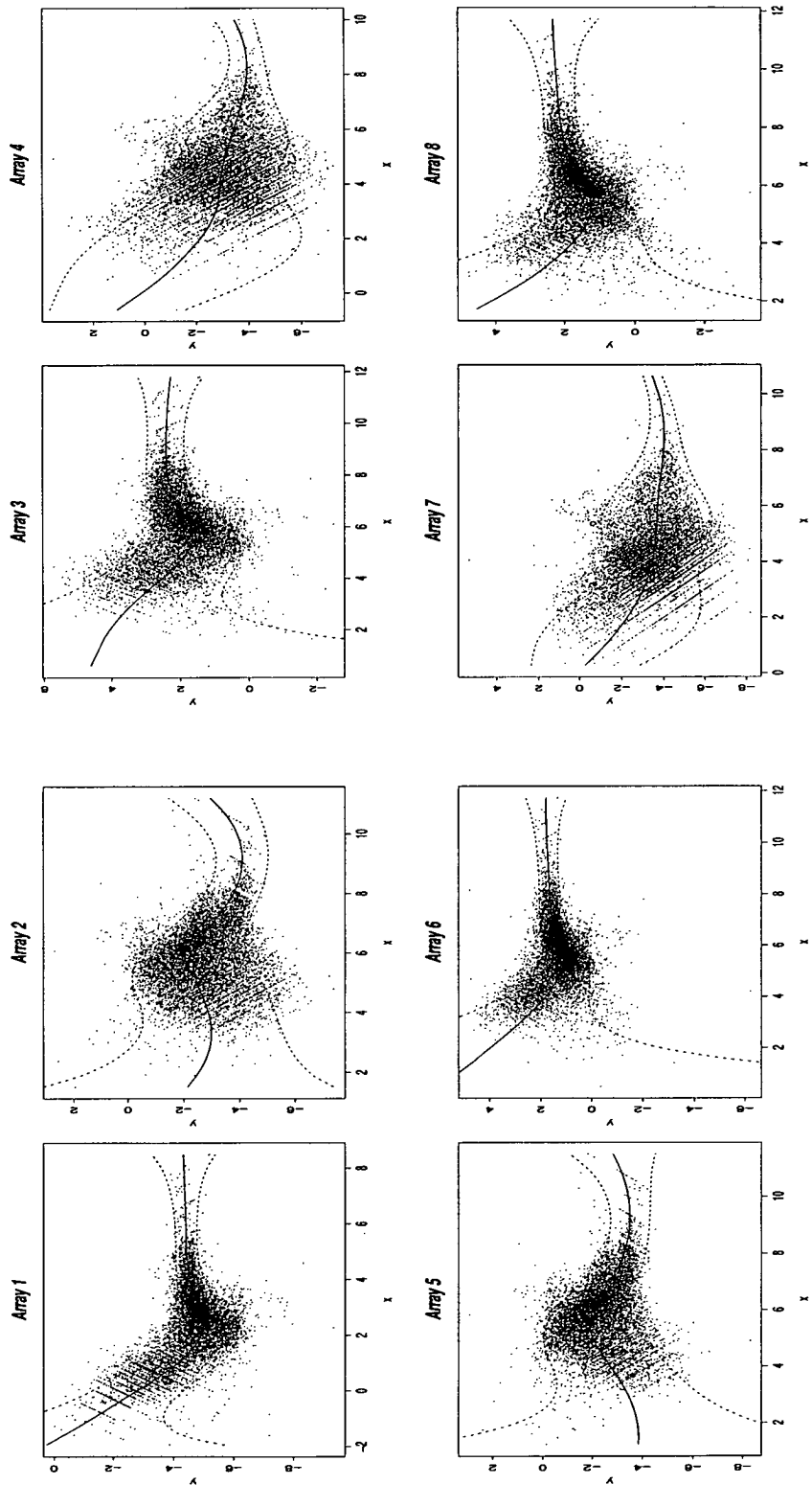


Figure 5.5: GAMLSS fit to 8 arrays of liver data using p-splines of the model $y|x \sim t_\nu(\mu(x), \sigma(x))$. Within each array, the solid line is fitted location $\hat{\mu}(x)$ and dashed lines are the limits $\hat{\mu}(x) \pm 2\hat{\sigma}(x)$.

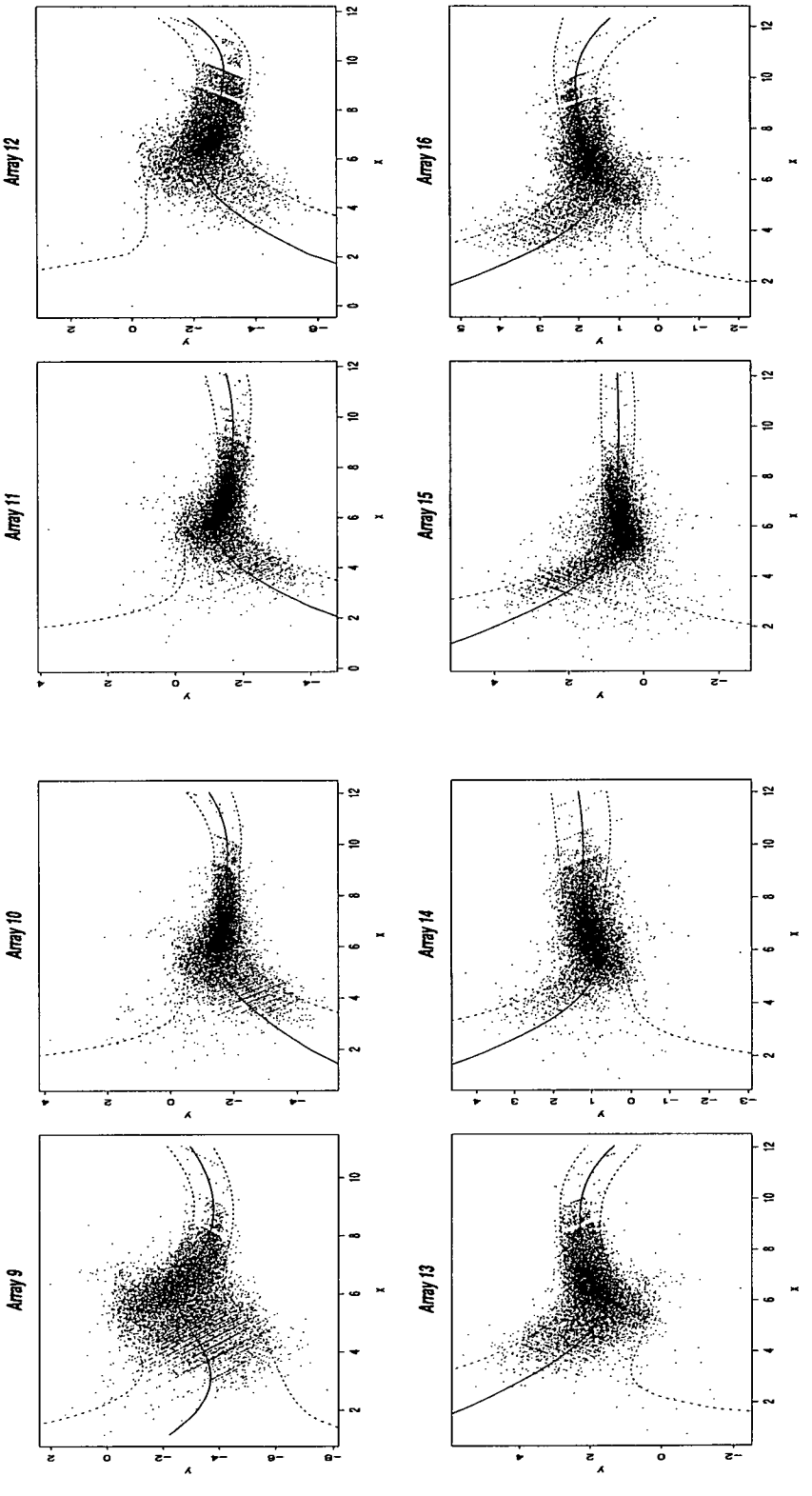


Figure 5.6: GAMLSS fit to 8 arrays of kidney data using p-splines of the model $y|x \sim t_\nu(\mu(x), \sigma(x))$. Within each array, the solid line is fitted location $\hat{\mu}(x)$ and dashed lines are the limits $\hat{\mu}(x) \pm 2\hat{\sigma}(x)$.

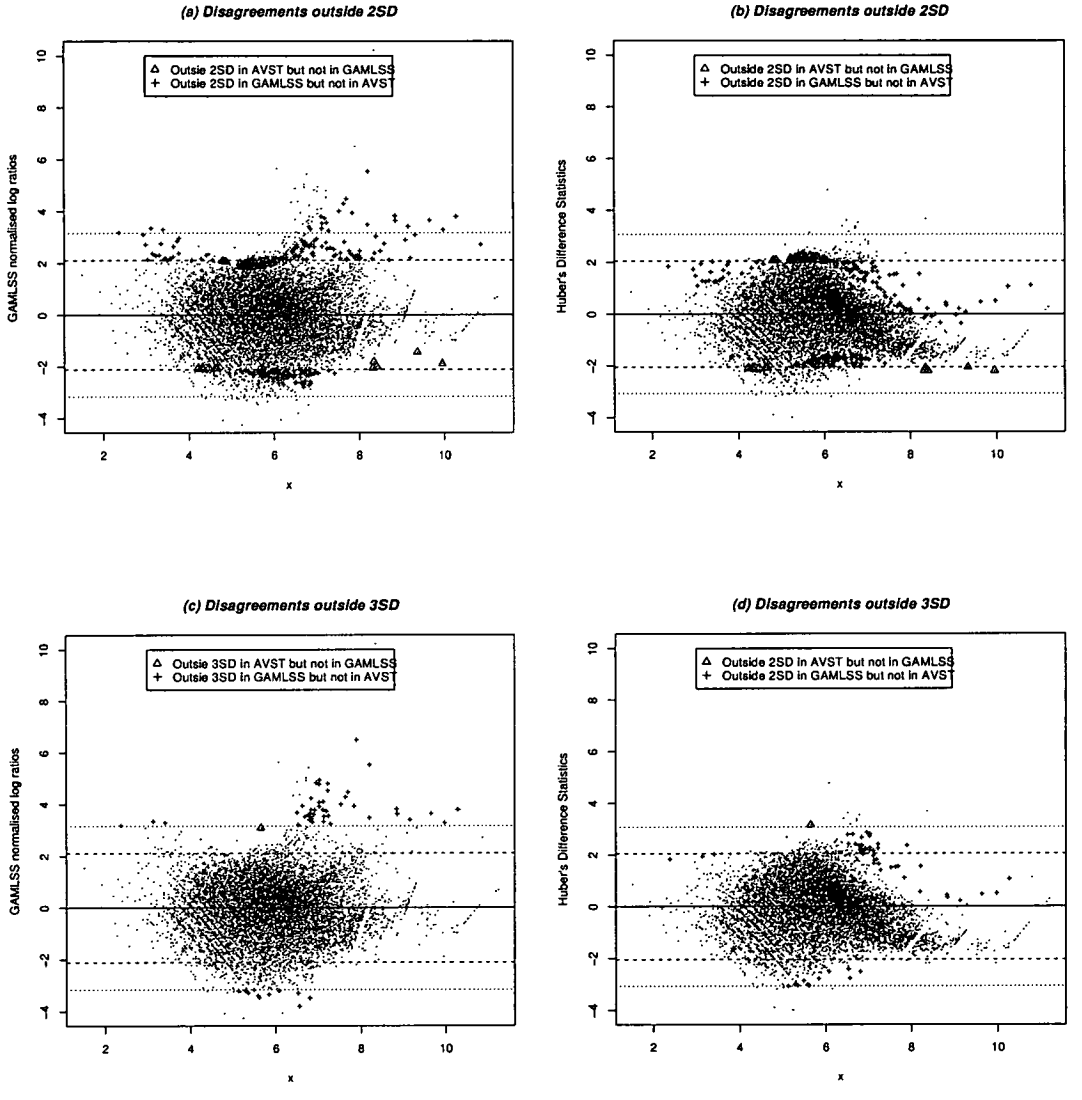


Figure 5.7: Summary of comparisons of differential expressions: (a) disagreements outside 2SD in GAMLSS normalised log-ratio vs. x plot; (b) disagreements outside 2SD in AVST difference statistic vs. x plot; (c) disagreements outside 3SD in GAMLSS normalised log-ratio vs. x plot; (d) disagreements outside 3SD in AVST difference statistic vs. x plot.

Also, we note that the number of detected genes is no more than we would expect by chance in the absence of differential expression. Therefore, simulation, where it is known which genes are differentially expressed, is a more effective way to compare the methods.

Table 5.4: Numbers of genes identified as differentially expressed in iron deficiency experiment, after GAMLSS or AVST normalisation, using t-tests with 5% and 1% thresholds.

		5% level			1% level		
		GAMLSS			GAMLSS		
		Significant	Not significant	Total	Significant	Not significant	Total
Liver data							
AVST	Significant	193	304	497	28	80	108
	Not Significant	186	9285	9471	51	9809	9860
	Total	379	9589	9968	79	9889	9968
Kidney data							
AVST	Significant	309	130	439	40	32	72
	Not Significant	154	9375	9529	49	9847	9896
	Total	463	9505	9968	89	9879	9968

5.6 Simulation study I

We have performed some simulation experiments to compare the performance of the proposed method to that of AVST method in making inference on differential expression. For the simulation in this section, we fitted GAMLSS to arcsinh transformed experimental vs. control samples. In simulation study II (Section 5.7) we consider fitting GAMLSS to the log-transformed data, *e.g.*, to log-ratio(y) vs. overall log-intensity (x).

5.6.1 Data generation

We simulate data according to several parametric and nonparametric location and scale models for the data in arcsinh scale. Suppose that

$$h_1 = \text{arcsinh}(a_1 + b_1 I_1),$$

and

$$h_2 = \text{arcsinh}(a_2 + b_2 I_2)$$

represent the arcsinh-transformed intensity data corresponding to the reference and treatment samples respectively. First we fit the proposed GAMLSS model to one array of lymphoma data transformed in arcsinh scale. We assume that the conditional distribution of h_2 given h_1 can be modelled using a t-distribution with degrees of freedom ν and intensity dependent location and scale functions $\mu(.)$ and $\sigma(.)$ respectively. That is,

$$h_2|h_1 \sim t_\nu(\mu(h_1), \sigma(h_1)). \tag{5.8}$$

We then simulate treatment sample (h_2) in arcsinh scale for several combinations of location (μ) and scale (σ) according to

$$h_{i2} = \mu_i + z_i\sigma_i, \tag{5.9}$$

where i indexes gene, $z_i \sim N(0, \hat{\tau}^2)$, and $\hat{\tau}$ is the estimated standard deviation of the standardised residuals from the model (5.8). We made 10% of the genes differentially expressed by modifying z_i in (5.9) by $z_i \sim N(0, \hat{\tau}^2) \pm U(0, 4\hat{\tau})$. The '+' and '−' are used to generate up- and down-regulated gene expressions respectively. We consider similar number of up- and down-regulations by choosing probability of up-regulation as 50%. Simulated data are then transformed back to the original scale by

$$\tilde{I}_{ij} = \frac{\sinh(h_{ij}) - a_j}{b_j}, \; j = 1, 2, \tag{5.10}$$

where the notation '$\tilde{I}$' is used to denote simulated intensity data. First we simulate according to fitted GAMLSS location and scale models to one array (named lc7b047) of lymphoma data described in Section 5.5.1, *i.e.*, considering $\mu_i = \hat{\mu}(h_{i1})$ and $\sigma_i = \hat{\sigma}(h_{i1})$ in the scatterplot (h_2 vs. h_1) shown in Figure 5.8.

Because the variance of the data is already stabilised to some extent by the AVST transformation, the scale model fitted by GAMLSS appears almost constant. To create non-constant scale functions, we add the difference $(\sigma - \bar{\sigma})$ and some multiple of the difference to the GAMLSS scale σ. This gives a family of scale functions with increasing distortions: σ, $p\sigma - (p-1)\bar{\sigma}$; $p = 2, 3\cdots$ and so on. We have also considered the reflected versions of these distorted scale functions: $p\bar{\sigma} - (p-1)\sigma$; $p = 2, 3\cdots$ and so on. Negative values in the distorted or reflected distorted scale functions are set to zero before using them in simulating data. Furthermore, we have experimented with some parametric scale family, e.g, $a + b\mu$, $a + b\mu^{2.5}$ and $a + b\mu^4$ etc. Similarly, we create alternative location functions by adding the difference of the linear fit from the GAMLSS location (μ) giving a family of alternative locations: μ, $p\mu - (p-1)$(linear fit); $p = 2, 3\cdots$ and so

on. Some of the alternative location functions, distorted and reflected distorted scale functions are displayed in Figures 5.8 (bottom), 5.9 (top) and 5.9 (bottom) respectively.

5.6.2 Results

We have compared the performance of AVST method to that of the proposed GAMLSS model in making inference on differential expressions. Standardised residuals (z_G) and difference statistic (z_A) are used as normalised data for GAMLSS and AVST respectively for identifying differential expressions. We use 2 and 3 times the estimated standard deviations of the respective normalised data as cut-off points for deciding differential expressions. We use median absolute deviation (MAD) for estimating the standard deviations. It is known which genes are differentially expressed in the simulated data and we therefore compute the average false positive rates and power (1 – false negative rates) for both methods for 10 simulated data sets. The results are summarised in Table 5.5. It is seen that

Table 5.5: Estimated false positive rates and power for 10 simulated data sets.

	2 × SD cutoff				3 × SD cutoff			
	False Positive Rate		Power		False Positive Rate		Power	
(Location, Scale)	GAMLSS	AVST	GAMLSS	AVST	GAMLSS	AVST	GAMLSS	AVST
(μ, σ)	0.037	0.032	0.41	0.40	0.003	0.002	0.16	0.15
$(\mu, 2\sigma - \bar{\sigma})$	0.042	0.036	0.39	0.37	0.005	0.005	0.14	0.15
$(\mu, 3\sigma - 2\bar{\sigma})$	0.049	0.046	0.38	0.35	0.011	0.011	0.13	0.11
$(\mu, 4\sigma - 3\bar{\sigma})$	0.059	0.064	0.38	0.35	0.020	0.021	0.15	0.15

both methods lose power as the deviation of the scale model from constancy increases. False positive rate also increases for both methods as the scale function deviates from constancy. However in terms of power difference GAMLSS appears to perform comparatively better as the fluctuation in scale increases.

Since there generally is a trade-off between false positive rate and false negative rate in any statistical decision procedure, an alternative way is to compare one by keeping the other fixed. We therefore compare the power of the methods holding false positive rates fixed at 5% and 1% levels. Standard errors of the power difference at each level are also computed to have an idea about the variability of the estimates in repeated sampling. The results for 10 simulated data sets are tabulated in Table 5.6. A more informative analysis, the Receiver Operating Characteristic (ROC) analysis, which is a graphical plot of the sensitivity vs. 1− specificity for a binary classifier system as its discrimination threshold is varied, has also been performed. In the terminology of statistical hypothesis testing,

ROC is obtained by plotting the power against the level of a test. Power and level are also known as proportions of true positives (TP) and false positives (FP) respectively. The best possible prediction method is the one that has 100% sensitivity (detects all true positives) and 100% specificity (no false positives). ROC curves comparing the performance of GAMLSS and AVST normalisations for each of the 19 different location and scale combinations presented in Table 5.6 are shown in Figures 5.10, 5.11 and 5.12.

The results in Table 5.6, and the ROC analysis in Figures 5.10, 5.11 and 5.12 indicate that, in general, performance of GAMLSS following AVST analysis in detecting differential expressions is better than that of the AVST only analysis. The difference in power increases with the increase in the deviation of the scale model from constancy. For this lymphoma data set, performance of proposed method is considerably better in cases of parametric scales and reflected distorted scale functions. Standard errors of the power difference over repeated sampling appears sufficiently small, and on an average are smaller for 5% level than those for 1% level.

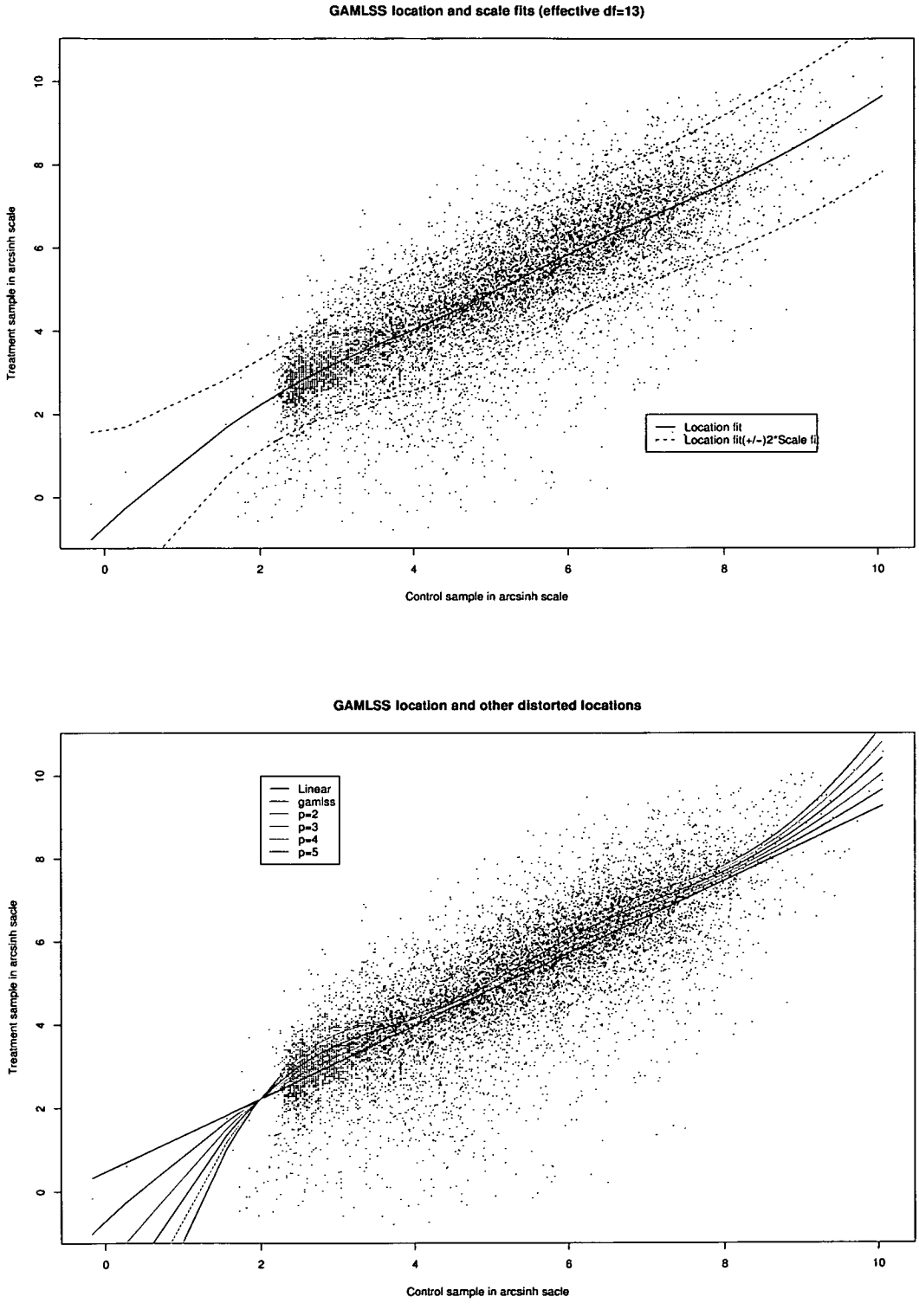


Figure 5.8: Fitted location and scale functions to lymphoma data by GAMLSS (top) and GAMLSS location and other distorted location functions (bottom).

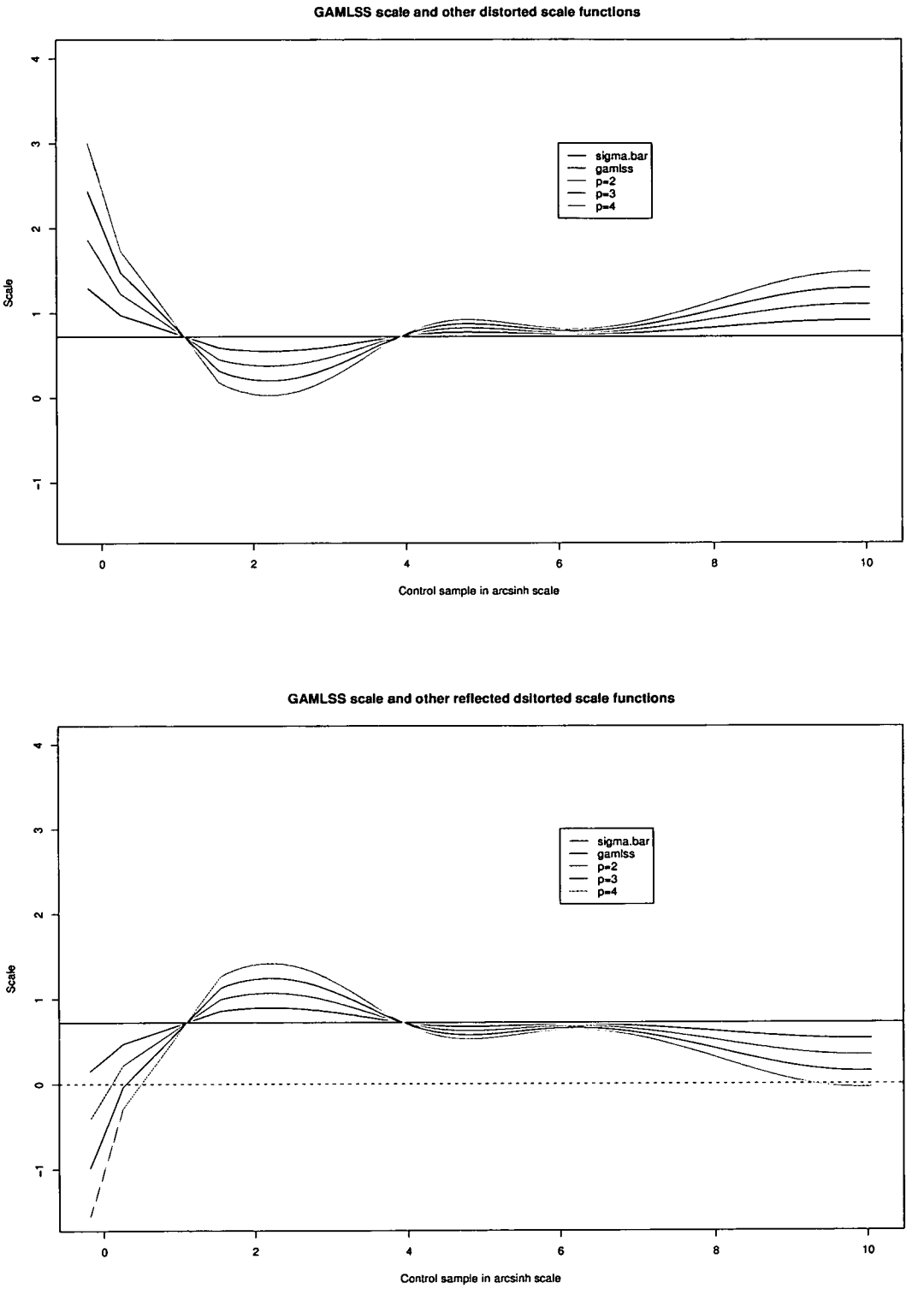
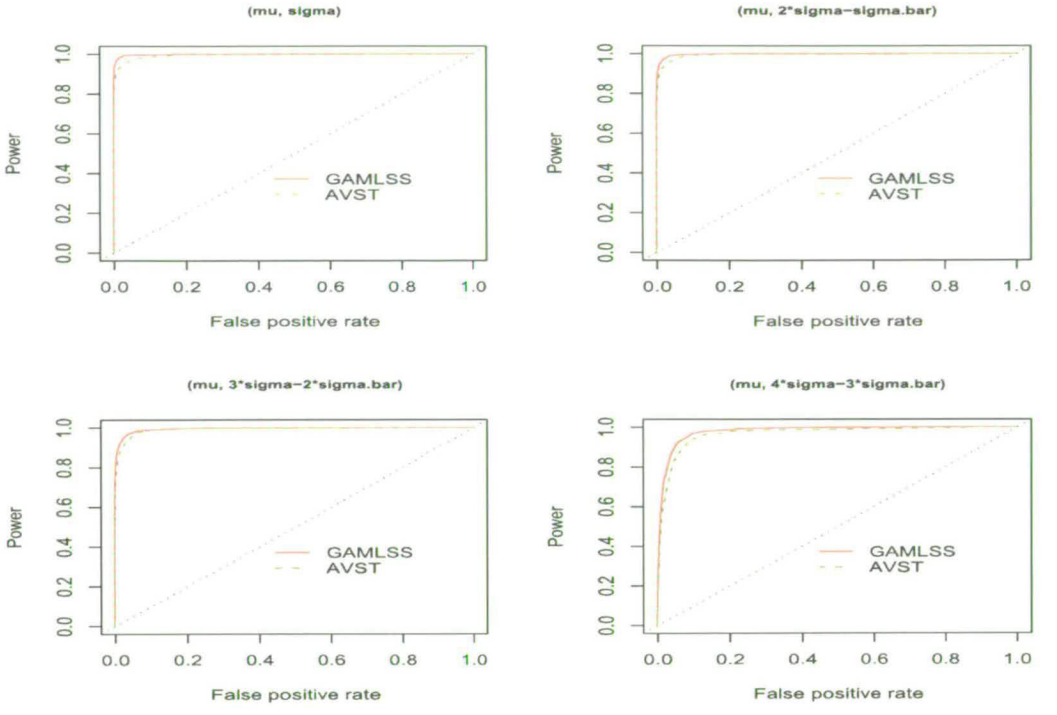


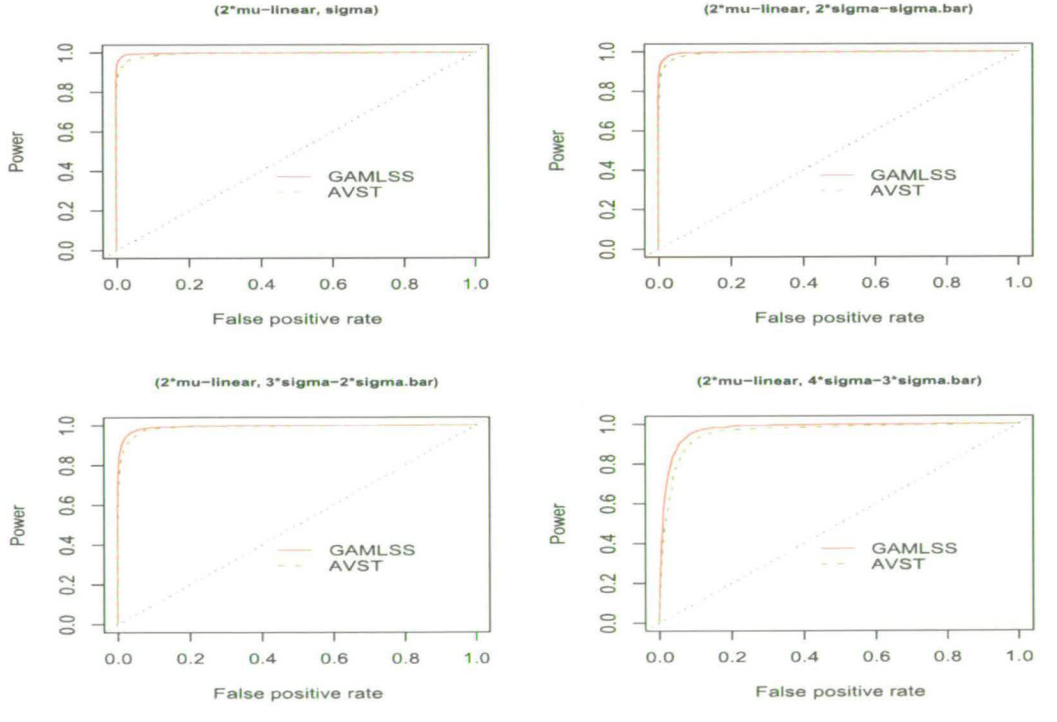
Figure 5.9: GAMLSS scale fitted to lymphoma data and other distorted scale functions (top) and reflected versions of the scale functions (bottom).

Table 5.6: Estimated power (at fixed levels) for 10 simulated data sets.

	Estimated Power					
	Level fixed at 5%			Level fixed at 1%		
(Location, Scale)	GAMLSS	AVST	SE of difference	GAMLSS	AVST	SE of difference
Distorted scales						
$(\mu,\ \sigma)$	0.44	0.44	0.009	0.26	0.27	0.009
$(\mu,\ 2\sigma - \bar{\sigma})$	0.41	0.41	0.005	0.19	0.19	0.011
$(\mu,\ 3\sigma - 2\bar{\sigma})$	0.40	0.36	0.005	0.10	0.10	0.012
$(\mu,\ 4\sigma - 3\bar{\sigma})$	0.33	0.30	0.006	0.01	0.01	0.005
$(2\mu$−linear fit, $\sigma)$	0.43	0.44	0.008	0.25	0.26	0.008
$(2\mu$−linear fit, $2\sigma - \bar{\sigma})$	0.41	0.41	0.008	0.18	0.18	0.011
$(2\mu$−linear fit, $3\sigma - 2\bar{\sigma})$	0.37	0.35	0.008	0.09	0.08	0.015
$(2\mu$−linear fit, $4\sigma - 3\bar{\sigma})$	0.33	0.28	0.008	0.01	0.01	0.016
Reflected distorted scales						
$(\mu,\ 2\bar{\sigma} - \sigma)$	0.47	0.46	0.011	0.33	0.30	0.009
$(\mu,\ 3\bar{\sigma} - 2\sigma)$	0.48	0.46	0.004	0.36	0.30	0.019
$(\mu,\ 4\bar{\sigma} - 3\sigma)$	0.49	0.44	0.016	0.35	0.25	0.019
$(\mu,\ 5\bar{\sigma} - 4\sigma)$	0.50	0.42	0.015	0.36	0.22	0.016
$(2\mu$−linear fit, $2\bar{\sigma} - \sigma)$	0.47	0.46	0.009	0.33	0.30	0.013
$(2\mu$−linear fit, $3\bar{\sigma} - 2\sigma)$	0.48	0.45	0.009	0.34	0.28	0.015
$(2\mu$−linear fit, $4\bar{\sigma} - 3\sigma)$	0.49	0.44	0.016	0.35	0.24	0.018
$(2\mu$−linear fit, $5\bar{\sigma} - 4\sigma)$	0.50	0.42	0.014	0.36	0.21	0.015
Parametric scales						
$(\mu,\ \sqrt{0.1 + 1.5\mu})$	0.44	0.41	0.010	0.27	0.23	0.009
$(\mu,\ \sqrt{0.5 + 0.05\mu^{2.5}})$	0.40	0.31	0.014	0.20	0.11	0.011
$(\mu,\ \sqrt{0.25 + 0.05\mu^{4}})$	0.50	0.35	0.015	0.35	0.15	0.013

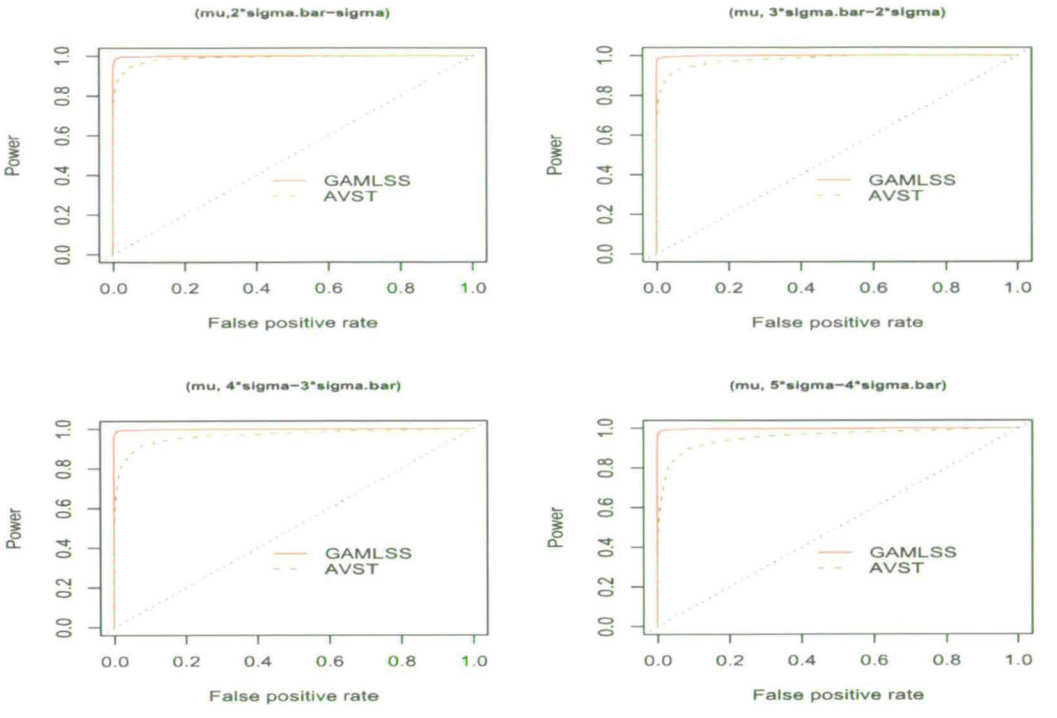


(a) (μ, distorted scales)

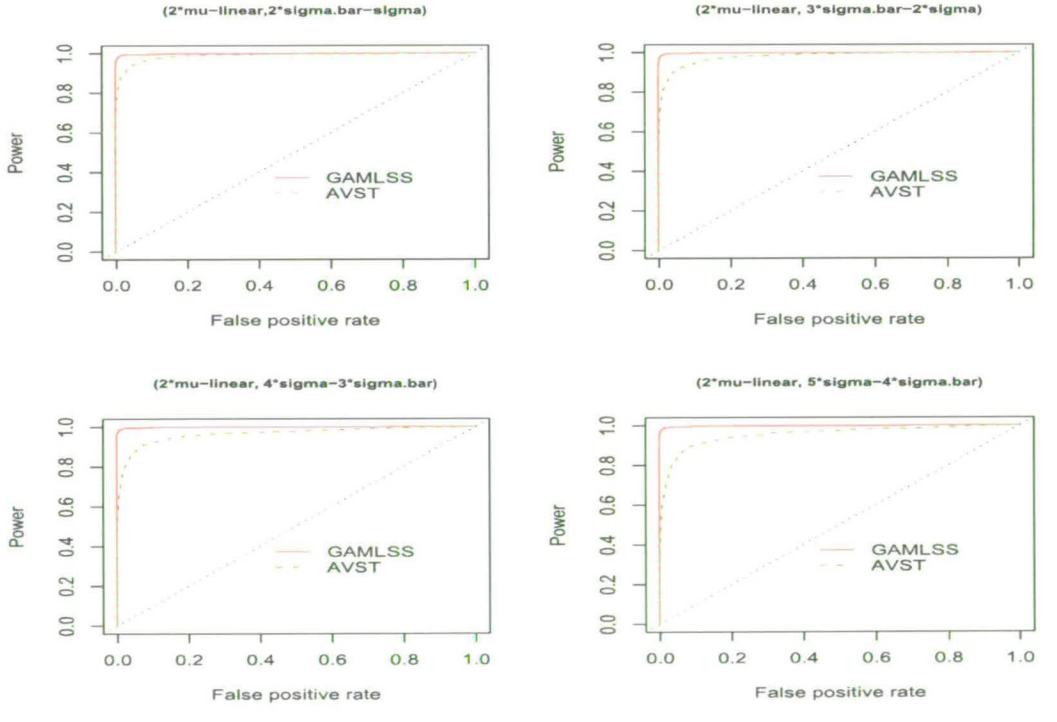


(b) (2μ − linear, distorted scales)

Figure 5.10: ROC curves comparing the performance of GAMLSS and AVST for the simulated data with distorted scales presented in Tables 5.5 and 5.6. The solid red lines and dashed green lines represent the ROC curves for GAMLSS and AVST normalisations respectively.

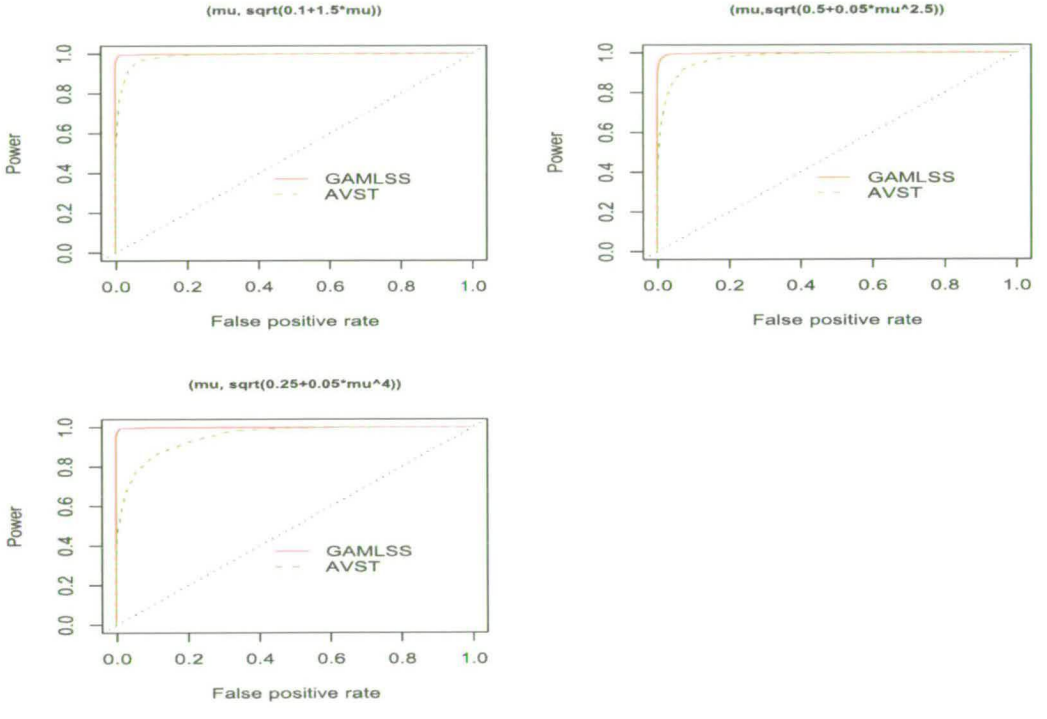


(a) (μ, reflected distorted scales)



(b) (2μ – linear, reflected distorted scales)

Figure 5.11: ROC curves comparing the performance of GAMLSS and AVST for the simulated data with relected distorted scales presented in Tables 5.5 and 5.6. The solid red lines and dashed green lines represent the ROC curves for GAMLSS and AVST normalisations respectively.



(a) (μ, parametric scales)

Figure 5.12: ROC curves comparing the performance of GAMLSS and AVST for the simulated data with parametric scales presented in Tables 5.5 and 5.6. The solid red lines and dashed green lines represent the ROC curves for GAMLSS and AVST normalisation respectively.

5.7 Simulation study II

Simulations in this section correspond to the data of iron-deficiency experiments, and we consider fitting GAMLSS to log-ratio (y) vs. average log intensity (x) data. We simulated data from the GAMLSS model, with 10% of genes differentially expressed, then normalised them by both methods and compared detection rates. Similarly, we simulated data from the AVST model and compared normalisations.

5.7.1 Data generation

For the GAMLSS model, we based simulations on observed log-intensities (x) and estimated location and scale functions $(\hat{\mu}(x), \hat{\sigma}(x))$ from the iron-deficiency experiment, as shown in Figures 5.5 and 5.6. We simulate log-ratio values (y) corresponding to each array of data according to

$$y = \hat{\mu}(x) + z\hat{\sigma}(x) \tag{5.11}$$

where $z \sim N(0, \tau_G^2)$. Every 10th gene is made differentially expressed by modifying z in (5.11) by $z \sim N(0, \tau_G^2) \pm U(0, 4\tau_G)$, of which 50% are made up-regulated; *i.e.*, correspond to $N(0, \tau_G^2) + U(0, 4\tau_G)$. We choose 10% as the proportion of differential expressions to be consistent with the chosen LTS quantile of 90% in fitting AVST model, and because this was one of the cases considered by Huber *et al.* (2003). The choice $U(0, 4\tau_G)$ as the amplitude of differential expression was made simply to achieve reasonable detection rates. The simulated log-ratio data are then transformed back to the original scale by $\tilde{I}_1 = \exp(x - y/2)$ and $\tilde{I}_2 = \exp(x + y/2)$, using '$\tilde{I}$' to denote simulated values. We generate 10 data sets corresponding to each of 16 arrays and compare the performance of GAMLSS with AVST in making inference on differential expression. For each generated data set we fit GAMLSS directly to y vs. x data and AVST to simulated intensity data on the original scale $(\tilde{I}_1, \tilde{I}_2)$.

To simulate data from the AVST model, we generate the z's as above, with τ_G replaced by τ_A, then transform to intensities using estimated values for the a's and b's:

$$\left.\begin{aligned} \tilde{I}_1 &= (\sinh(x' - z/2) - \hat{a}_1)/\hat{b}_1 \\ \tilde{I}_2 &= (\sinh(x' + z/2) - \hat{a}_2)/\hat{b}_2, \end{aligned}\right\} \tag{5.12}$$

where

$$x' = \frac{1}{2}[\operatorname{arcsinh}(\hat{a}_1 + \hat{b}_1 I_1) + \operatorname{arcsinh}(\hat{a}_2 + \hat{b}_2 I_2)].$$

If either $\tilde{I}_1$ or $\tilde{I}_2$ is negative, then the corresponding data points are regenerated. Again, we generate 10 sets of data corresponding to each of the 16 arrays of iron-deficiency data set and apply either GAMLSS or AVST normalisation.

5.7.2 Results

Standardised residuals and difference statistics are used as normalised data for GAMLSS and AVST methods respectively for identifying differential expression. In simulated data, identities of the differentially expressed genes are known. We can therefore compute power of detection and proportion of type I errors for both methods. Tests for differential expression are carried out at 5% and 1% levels using t-tests as with the experimental data. Standard errors of the difference of powers and the difference of type I errors at each cutoff point/level are also computed to indicate the stability of the estimates in repeated sampling. The results summarised in Table 5.7 show that GAMLSS outperforms AVST when

Table 5.7: Estimated power, proportion of type I errors and SE of difference based on t-test for assessing differential expression. Results are based on 10 simulated data sets from either GAMLSS or AVST model.

	5% level						1% level					
	GAMLSS		AVST		SE(difference)		GAMLSS		AVST		SE(difference)	
Organ	Power	Error	Power	Error	Power	Error	Power	Error	Power	Error	Power	Error
Data generated according to GAMLSS												
Liver	0.80	0.049	0.77	0.048	0.0075	0.00240	0.70	0.0100	0.65	0.0098	0.0040	0.00130
Kidney	0.80	0.049	0.77	0.038	0.0075	0.00180	0.69	0.0096	0.66	0.0070	0.0079	0.00080
Data generated according to AVST												
Liver	0.79	0.051	0.79	0.051	0.0029	0.00120	0.68	0.0100	0.68	0.0100	0.0033	0.00039
Kidney	0.80	0.050	0.80	0.050	0.0019	0.00075	0.69	0.0098	0.69	0.0098	0.0020	0.00032

data are generated from a GAMLSS model. Whereas GAMLSS is as powerful as AVST method when the situation is ideal for AVST model. Furthermore, GAMLSS has been found to have good control over the type I errors. Observed levels have consistently been found to be very close the corresponding nominal levels of 5% and 1%.

5.8 Discussion

Normalisation is an important first step in the analysis of cDNA microarray data and will undoubtedly see further development in future. Until recently, separate methods have been suggested for location and scale normalisation with nonparametric methods such as loess preferred for the former and parametric models for the latter. However, as can be seen from experimental data such as the plots in Figures 5.5 and 5.6, the variability in microarray data can be of a more challenging nature than can be handled by any particular parametric transformation.

The approach developed in this chapter is novel in two ways; first, we consider nonparametric models for both location and scale normalisation, and second, we incorporate both location and scale normalisation simultaneously using a flexible model, GAMLSS. As is the case with other flexible models, there is always a risk of overfitting and caution is needed in choosing optimality of the fit of GAMLSS. Rigby and Stasinopoulos (2005) suggested several criteria (Akaike information criterion (AIC), Generalised Akaike information Criterion (GAIC) and Schwarz Bayesian information Criterion (SBC), Schwarz (1978)) for optimising the amount of smoothing in the fit of the GAMLSS model. However for microarray data, optimal fit according to these criteria gives very localised fit. In loess smoothing the user-defined parameter f, called span, is the fraction of the data used for smoothing at each point. The span of loess for location smoothing of microarray data is typically suggested to be between 20% and 40% (Dudoit *et al.*, 2002; Yang *et al.*, 2002). Model (5.4) fitted to the iron-deficiency microarray data with loess ($f = 40\%$) as the smoothing option for nonparametric models results in effective degrees of freedom equal to 13. We therefore fit model (5.4) using p-splines with effective degrees of freedom equivalent to that of loess, which gives a reasonable amount of smoothing for both location and scale models. We prefer p-splines because they are less computationally expensive than smoothing splines, and p-splines smoothing has some desirable properties over loess, *e.g.*, it is free from boundary effects and conserves the moments of the data. Application of the proposed method suggests that it is capable of capturing the trends of whatever shape in both location and scale of the data and therefore may be a suitable normalisation method in microarray applications. Simulations demonstrate GAMLSS to be more powerful than Huber's parametric model in detecting differential expression in a wide variety of realistic situations. GAMLSS has been found to be as powerful as AVST even when the situation is ideal for the parametric AVST model.

GAMLSS normalisation presented in this chapter is applicable for within-slide normalisation. One advantage of GAMLSS normalisation over the other within-array normalisation methods is that it automatically achieves between array standardisation as well. Within-slide GAMLSS normalised log-ratios are comparable across arrays as they are expected to have unit scale for all individual arrays.

Although the GAMLSS method presented in this thesis models only the intensity-based trends in the location and scale of the log-ratio data, the method is flexible enough to accommodate the spatial and print-tip effects in both location and scale models. Therefore spatial and print-tip normalisations can be

incorporated in the same model by including row, column and print-tip factors. Print-tip bias can also be corrected by applying the model to each of the print-tip groups separately. Print-tip group loess normalisation is routinely used to adjust for intensity and print-tip biases in the location. A similar approach is possible with GAMLSS, to adjust for intensity dependent and print-tip trends in both location and scale of log-ratio data.

Fold-change interpretation of gene expression, generally preferred by biologists, does not directly apply to GAMLSS normalised log-ratios. Statements like "this gene is two-fold up-/down-regulated" is therefore not valid for GAMLSS standardised data $(y - \hat{\mu}(x))/\hat{\sigma}(x)$. This type of statement may however be made for GAMLSS residuals $(y - \hat{\mu}(x))$, which does not adjust the data for scale.

Although the within-slide GAMLSS normalised data are comparable across slides, it may be be desirable to generalise the method to multiple-slide normalisation. One possible way of doing this would be to consider multivariate GAMLSS. This will however require extending the existing univariate GAMLSS theory to the multivariate case, which may not be very straight forward. Another possible extension of GAMLSS normalisation would be to consider the scale model as a combination of the smooth component $\sigma(x)$ and a parametric, fixed or random, gene specific component. This kind of formulation, termed semiparametric GAMLSS, is allowed within the GAMLSS framework but would require within-slide replication.

Chapter 6

Functional regression modelling

6.1 Introduction

Our study on the use the functional regression model for combining multiple scan data in Chapter 4 motivated us to investigate some aspects of this type of model in more detail. For example, the functional model presented in Chapter 4 would have been more realistic if we could consider a separate variance term for each individual scan of data. However in this case the model is intractable by maximum likelihood estimation because the parameters are not estimable without further restrictions on the variance parameters. Although maximum likelihood estimation fails, all parameters are estimable by alternative methods. A considerable amount of literature is available on functional, structural and ultrastructural relationship, commonly known as measurement error models, focusing on alternative estimation methods to overcome the limitations of maximum likelihood estimation. These include estimation approaches based on functions of observations rather than the observations themselves (Sprent, 1976; Morton, 1981). Neyman and Scott (1951) defined two types of parameters, 'structural' or 'incidental', in a functional model depending on whether they occur with every observation or with a single observation of the joint distribution. Sprent (1976) suggested using likelihood of certain functions of observations to avoid dependency of the structural parameters on the incidental parameters, which is essentially the cause of the inconsistency in the maximum likelihood estimates of the structural parameters. Morton (1981) approached the problem in a similar way to Sprent (1976) and based the estimation on 'pivots', certain functions of observations whose distributions do not depend on the incidental parameters. Use of pivots therefore eliminate the dependence on the incidental parameters. Estimating equations were derived by taking the expectations of the score functions conditional on the pivots and the sufficient statistics of the parameters. The estimating equations proposed by Chan and Mak (1983) are based on likelihood of the observations, but

modified to give consistent estimates of the parameters. In fact, all the approaches described above for estimating functional relationship are mainly concerned with estimability and consistency of the estimators. A question still remains about how good the estimators are with respect to other criteria. One of the key properties of an optimal estimator is the *efficiency*, which has not received much consideration in the literature of functional models.

In this chapter our main aim is to investigate alternative approaches of estimating functional models and compare them in terms of unbiasedness and efficiency of the estimators. We confine our investigation to linear relationships and restrict attention to a particular 'cleaned' subset of the murine macrophage data described in Section 3.3. A plot of subset of the data on which we base our

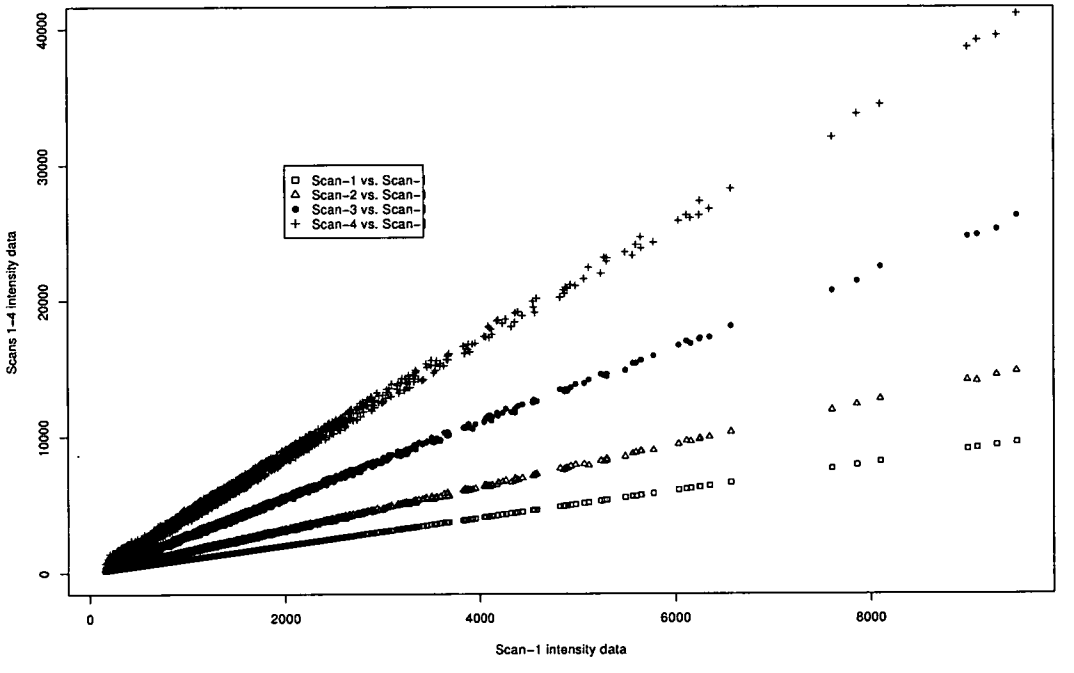


Figure 6.1: Scatterplot of scans 1–4 vs. scan 1 intensity data.

analysis is given in Figure 6.1. The subset contains 7543 observations which were selected by fitting simple no intercept Least Trimmed Squares (LTS, Rousseeuw and Leroy, 1987) regressions ($y_{.j} = \beta_j y_{.1} + e$ for $j = 2, 3, 4$) of scans 2, 3 and 4 data on scan 1 data and choosing observations corresponding to absolute standardised residuals less than or equal to 3. A few observations near zero at the bottom end have also been deleted. The subset of the data is free from outliers and consists of cases where the linearity of the relationship is not affected by censoring of signal at the upper limit of 65535. This nonlinear pattern of the full data set has been

explained and addressed in Chapters 3 and 4.

6.2 The model

Suppose that y_{ij} represents the measure of response of gene i in scan j for $i = 1, \cdots, n;\ j = 1, \cdots, m$. For this investigation we consider the multivariate linear functional model with $y_{\cdot 1}$ as the predictor variable and $(y_{\cdot 2}, \cdots, y_{\cdot m})^T$ as the vector of response variables, and assume

$$y_{ij} \sim N(\mu_i \beta_j, \sigma_j^2), \tag{6.1}$$

where $\mu_i = E(y_{i1})$ denotes the true expression level for gene i, β_j the gain setting in laser scan j, and σ_j^2 the variance of the measured response in scan j. In matrix notation, the model can be represented as

$$Y_i \sim N_m(\mu_i \beta, V), \tag{6.2}$$

where $Y_i = (y_{i1}, \cdots, y_{im})^T$, $\beta = (\beta_1, \cdots, \beta_m)^T$ and V is a $m \times m$ diagonal covariance matrix. We constrain $\beta_1 \equiv 1$ for identifiability. The objective is to estimate the μ's, β's and σ^2's without making distributional assumptions about the μ's.

6.3 Estimation methods

In this section, we describe some estimation methods that can be used to estimate the parameters of model (6.1) avoiding the limitations the of method of maximum likelihood in functional regression estimation. For example, the model can be estimated by minimising the sum of squares of the difference between the observed and expected second moments, which we term *method of second moments* estimation. Sprent (1976) noted that anomalies of the likelihood approach can be avoided by considering likelihood of certain functions of observations, rather than the likelihood of the observations themselves, and making inference on the modified likelihood. Morton (1981) gave estimating equations for the functional model (6.1), and more generally for a multivariate ultrastructural model, derived from functions of pivot-like quantities, which eliminate the dependence on the incidental parameters, and provide basis for consistent estimation. Modified likelihood equations for estimation of the multivariate functional relationship was also suggested by Chan and Mak (1983). Aitkin and Rocci (2002) suggested an EM algorithm for maximum likelihood estimation in generalised linear models with continuous measurement error in the explanatory variables. A brief review of some of these alternative estimating methods are given below.

6.3.1 Method of second moments

The idea is analogous to the method used for solving factor analysis model (Mardia *et al.*, 1979, p. 259) where the coefficient matrix and the variance of specific factors are estimated by equating the sample covariance matrix and population covariance matrix under the factor model. Let S denotes the $m \times m$ matrix of the observed second moments, *i.e.*,

$$S_{jk} = \frac{1}{n}\sum_{i=1}^{n} y_{ij}y_{ik} \qquad \text{for } j,k = 1,\cdots,m. \tag{6.3}$$

We denote its expectation by V, given by

$$V_{jk} = \tau^2\beta_j\beta_k + \sigma_j^2\delta_{jk} \qquad \text{for } j,k = 1,\cdots,m,$$

where δ is the Kronecker delta, and

$$\tau^2 = \frac{1}{n}\sum_{i=1}^{n}\mu_i^2.$$

The parameters β's and σ^2's together with τ^2 can be estimated by numerically minimising the sum of squares:

$$\sum_{j=1}^{m}\sum_{k=1}^{m}(S_{jk} - V_{jk})^2. \tag{6.4}$$

There are $\frac{1}{2}m(m+1)$ distinct terms in S, which equal or exceed the $2m$ parameters in V provided that $m \geq 3$. The maximum likelihood estimates of μ's, conditional on the estimates $\hat{\beta}$'s and $\hat{\sigma^2}$'s, can be obtained as

$$\hat{\mu}_i = \sum_{j=1}^{m}\frac{y_{ij}\hat{\beta}_j}{\hat{\sigma}_j^2}\Big/\sum_{j=1}^{m}\frac{\hat{\beta}_j^2}{\hat{\sigma}_j^2} \qquad \text{for } i = 1,\cdots,n. \tag{6.5}$$

6.3.2 Morton's (1981) estimating equations

The estimating equations in Morton (1981) were derived for a ultrastructural relationship with replicated observations, which also applies to the functional model (6.2) as a special case. The idea in Morton's (1981) method is to start with pivot-like quantities, which eliminate the dependence on the incidental parameters, and derive functions of these which give estimating equations leading to consistent estimators. These equations involve the incidental parameters which are then replaced by estimators. The method overcomes some of the difficulties encountered in likelihood and least squares estimation when there are many incidental parameters.

Morton (1981) gave estimating equations for a more general ultrastructural relationship with replicated observations, which for an m-vector $Y_{iq} = (y_{i1q}, \cdots, y_{imq})^T$ of observations at replicate q is given by

$$Y_{iq} = \alpha + (\mu_i + \delta_{iq})\beta + \epsilon_{iq} \qquad \text{for } i = 1, \cdots, n;\ q = 1, \cdots, r, \tag{6.6}$$

where α, β are m-vectors of parameters with first components $\alpha_1 = 0, \beta_1 = 1$, the μ's are incidental univariate parameters, δ's are independent $N(0, \tau^2)$ errors and the ϵ's are independent m-vectors with multivariate normal distributions of zero mean and covariance matrix U specified by $\theta = (\sigma_1^2, \cdots, \sigma_m^2)^T$.

The sufficient statistics for the parameters are the vector means $\bar{Y}_{i.} = (\bar{y}_{i1.}, \cdots, \bar{y}_{im.})^T$ and the sum of squares and product matrix

$$S_i = \sum (Y_{iq} - \bar{Y}_{i.})(Y_{iq} - \bar{Y}_{i.})^T.$$

To state the estimating equations of Morton (1981), suppose

$$l = V^{-1}\beta, \quad V = U + \tau^2 \beta\beta^T,$$

$$g_i = \bar{Y}_{i.} - \alpha - \beta\bar{y}_{i1.},$$

and define the matrix

$$M = \begin{bmatrix} l_1 & l_2 & \cdots & l_m \\ -\beta_2 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ -\beta_m & 0 & \cdots & 1 \end{bmatrix}. \tag{6.7}$$

Then the augmented matrix S^+ is defined to be

$$S^+ = S_i + r \sum_{i=1}^{n} M^{-1} g_i g_i^T (M^{-1})^T, \tag{6.8}$$

which has expectation

$$E(S^+) = n(r-1)V + nM^{-1}PMVM^TP(M^{-1})^T, \tag{6.9}$$

where P is the $m \times m$ identity matrix except for the first term which is replaced by zero.

Denote the elements of V^{-1} by $\{v^{h,k}\}$ and the vectors of partial derivatives with respect to θ by $\partial v^{hk}/\partial\theta$, where $\theta = (\sigma_1^2, \cdots, \sigma_m^2)^T$.

Morton (1981) showed that, the parameters of model (6.6) can be estimated by solving the system of equations

$$\bar{g}_{.} = 0, \tag{6.10}$$

$$PV^{-1}\left[\sum_i \hat{\mu}_i M^{-1} g_i - \{S^+ - E(S^+)\}U^{-1}\beta\right] = 0, \quad (6.11)$$

$$\sum_{h,k}\{S^+_{h,k} - E(S^+_{h,k})\}\partial v^{h,k}/\partial\theta = 0, \quad (6.12)$$

$$\beta^T U^{-1} S_r - (r-1)VU^{-1}\beta = 0, \quad (6.13)$$

where $\hat{\mu}_i = l^T Y_i / l^T \beta$, and that the estimates obtained by solving the above system of equations are consistent. That is, the method overcomes the inconsistency properties of the likelihood equations in presence of incidental parameters in the functional model.

The functional model (6.2) is a special case of the ultrastructural model for replicated data (6.6) with $r = 1, \alpha = 0$ and $\tau^2 = 0$. The system of the estimating equations for this functional model reduces to

$$PV^{-1}\left[\sum_i \hat{\mu}_i M^{-1} g_i - \{S^+ - E(S^+)\}V^{-1}\beta\right] = 0, \quad (6.14)$$

$$\sum_{h,k}\{S^+_{h,k} - E(S^+_{h,k})\}\partial v^{h,k}/\partial\theta = 0. \quad (6.15)$$

As for the general case, solutions of the above system give consistent estimators of the parameters of model (6.2).

6.3.3 Modified likelihood equations (Chan and Mak, 1983)

Chan and Mak (1983) used the log-likelihood function, which under the notations of model (6.2) is given by

$$L = -(n/2)\log|V| - (1/2)\sum_{i=1}^{n}(Y_i - \mu_i\beta)^T V^{-1}(Y_i - \mu_i\beta) + \text{constant}, \quad (6.16)$$

to derive the usual likelihood equations and then modified them to obtain consistency. For arbitrary given β and $\theta = (\sigma_1^2, \cdots, \sigma_m^2)^T$, the conditional maximum likelihood estimators of μ_i is given by

$$\hat{\mu}_i = (\beta^T V^{-1}\beta)^{-1}\beta^T V^{-1} Y_i \,. \quad (6.17)$$

Substitution of $\hat{\mu}_i$ into the above log-likelihood yields

$$L = -(n/2)\log|V| - (n/2)\text{tr}(V^{-1}S) + (n/2)\text{tr}(C^{-1}\beta^T\Psi\beta) + \text{constant}, \quad (6.18)$$

where

$$S = \sum_{i=1}^{n} Y_i Y_i^T,\ C = \beta^T V^{-1}\beta, \text{and } \Psi = V^{-1}SV^{-1}. \tag{6.19}$$

Thus the maximum likelihood estimators of β and θ are obtained by maximising $2n^{-1}L$ with respect to β and θ. Chan and Mak (1983) gave the following convenient representation of the likelihood equations

$$\text{tr}\{(I - V^{-1}\beta C^{-1}\beta^T)\Psi\beta C^{-1}(\partial\beta^T/\partial\beta_j)\} = 0 \quad (j = 2, \cdots, m), \tag{6.20}$$

$$\text{tr}\{(\Psi - V^{-1} - V^{-1}\beta C^{-1}\beta^T\Psi)(\partial V/\partial\theta_k)\} = 0 \quad (k = 1, \cdots, m). \tag{6.21}$$

Under the assumption that $\sum_{i=1}^{n} \mu_i^2/n$ tends to a finite limit as $n \to \infty$, and if we know V or A of $V = \sigma_1^2 A$, so that $\theta = \sigma_1^2$, the maximum likelihood estimator of β can be obtained by solving (6.20) and the resulting estimator is consistent. It can be verified directly that the left-hand side of the equations in (6.20) converges in probability to zero, but that $\Psi - V^{-1} - V^{-1}\beta C^{-1}\beta^T\Psi$ in (6.21) converges in probability to $-V^{-1}\beta C^{-1}\beta^T V^{-1}$. Since consistency of the maximum likelihood estimators of β and θ implies that left-hand sides of the equations in (6.20) and (6.21) converges in probability to zero when the true values V and β are inserted, it follows that a necessary condition for consistency of the maximum likelihood estimator of θ, and also of β if A is unknown, is

$$V^{-1}\beta C^{-1}\beta^T V^{-1}(\partial V/\partial\theta_k) = 0, \quad \text{for all } k, \tag{6.22}$$

which is rarely satisfied since $V^{-1}\beta C^{-1}\beta^T V^{-1}$ is nonnegative definite. Chan and Mak (1983) therefore modified the likelihood equations in (6.21) by adding an extra term $V^{-1}\beta C^{-1}\beta^T V^{-1}$ to the expression $\Psi - V^{-1} - V^{-1}\beta C^{-1}\beta^T\Psi$, so that the left-hand side now converges in probability to zero when the true values of β and V are inserted. Thus an alternative procedure for estimating β and θ is to solve the equations

$$\text{tr}\{(I - V^{-1}\beta C^{-1}\beta^T)\Psi\beta C^{-1}(\partial\beta^T/\partial\beta_j)\} = 0 \quad (j = 2, \cdots, m), \tag{6.23}$$

$$\text{tr}\{(I - V^{-1}\beta C^{-1}\beta^T(\Psi - V^{-1})(\partial V/\partial\theta_k)\} = 0 \quad (k = 1, \cdots, m). \tag{6.24}$$

This estimation procedure in general yields consistent estimators of β and θ, unless β and θ are not estimable.

For the special case when $m = 2$, Chan and Mak (1983) gave the simplified expressions for the modified estimating equations for estimating β and σ_1^2, assuming

that $\lambda = \sigma_2^2/\sigma_1^2$ is known, and the estimating equations are

$$\beta_2^2 S_{12} + \beta_2(\lambda S_{11} - S_{22}) - \lambda S_{12} = 0, \tag{6.25}$$

$$(\lambda + \beta_2^2)\sigma_1^2 - (\beta_2^2 S_{11} - 2\beta_2 S_{12} + S_{22}) = 0, \tag{6.26}$$

where S is given by (6.3).

6.3.4 Relation between Morton's and Chan and Mak's methods

Let us consider the special case of $m = 2$, and simplify the estimating equations for Morton's (1981) method. Estimating equation for β can be written as

$$\sum_{i=1}^{n} \hat{\mu}_i g_i = 0, \tag{6.27}$$

where

$$\hat{\mu}_i = \beta^T V^{-1} Y_i/(\beta^T V^{-1}\beta) = (\sigma_2^2 y_{i1} + \sigma_1^2 \beta_2 y_{i2})/(\sigma_2^2 + \beta_2^2 \sigma_1^2), \tag{6.28}$$

and $g_i = (0, y_{i2} - \beta_2 y_{i1})^T$.

Therefore, the estimating equation for β simplifies to

$$\sum_{i=1}^{n} (y_{i2} - \beta_2 y_{i1})(\sigma_2^2 y_{i1} + \sigma_1^2 \beta_2 y_{i2}) = 0,$$

$$\text{or, } \sigma_2^2 S_{12} + \beta_2 \sigma_1^2 S_{22} - \beta_2 \sigma_2^2 S_{11} - \beta_2^2 \sigma_1^2 S_{12} = 0,$$

$$\text{or, } \beta_2^2 S_{12} + \beta_2(\lambda S_{11} - S_{22}) - \lambda S_{12} = 0, \tag{6.29}$$

which is identical to that of Chan and Mak (1983).

Now, to simplify the estimating equation for θ, which has a single parameter σ_1^2 as λ is assumed known, we have

$$l = V^{-1}\beta = \begin{bmatrix} 1/\sigma_1^2 \\ \beta_2/\sigma_2^2 \end{bmatrix}, \tag{6.30}$$

and

$$M = \begin{bmatrix} 1/\sigma_1^2 & \beta_2/\sigma_2^2 \\ -\beta_2 & 1 \end{bmatrix}. \tag{6.31}$$

Let $|M| = D$, and therefore

$$M^{-1} = (1/D) \begin{bmatrix} 1 & \beta_2 \\ -\beta_2/\sigma_2^2 & 1/\sigma_1^2 \end{bmatrix}. \tag{6.32}$$

To evaluate S^+, first we need to expand $M^{-1}g_i g_i^T (M^{-1})^T$ which is

$$M^{-1}g_i g_i^T (M^{-1})^T = (1/D^2) \begin{bmatrix} \beta_2^2(y_{i2} - \beta_2 y_{i1})^2 & (\beta_2/\sigma_1^2)(y_{i2} - \beta_2 y_{i1})^2 \\ (\beta_2/\sigma_1^2)(y_{i2} - \beta_2 y_{i1})^2 & (1/\sigma_1^2)^2(y_{i2} - \beta_2 y_{i1})^2 \end{bmatrix}. \tag{6.33}$$

Summing (6.33) over i from 1 to n, we get the expression for S^+ given by

$$S^+ = \sum_{i=1}^{n} M^{-1}g_i g_i^T (M^{-1})^T = (1/D^2) \sum_{i=1}^{n} (y_{i2} - \beta_2 y_{i1})^2 \begin{bmatrix} \beta_2^2 & (\beta_2/\sigma_1^2) \\ (\beta_2/\sigma_1^2) & (1/\sigma_1^2)^2 \end{bmatrix}. \tag{6.34}$$

Similarly, matrix multiplications lead to expression for $E(S^+)$ as

$$\begin{aligned}
E(S^+) &= nM^{-1}PMVM^TP(M^{-1})^T \\
&= n(1/D) \begin{bmatrix} 1 & \beta_2 \\ -\beta_2/\sigma_2^2 & 1/\sigma_1^2 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} MVM^TP(M^{-1})^T \\
&= n(1/D) \begin{bmatrix} 0 & \beta_2 \\ 0 & 1/\sigma_1^2 \end{bmatrix} \begin{bmatrix} 1/\sigma_1^2 & \beta_2/\sigma_2^2 \\ -\beta_2/\sigma_1^2 & 1 \end{bmatrix} VM^TP(M^{-1})^T \\
&= n(1/D) \begin{bmatrix} -\beta_2^2 & \beta_2 \\ -\beta_2/\sigma_1^2 & 1/\sigma_1^2 \end{bmatrix} VM^TP(M^{-1})^T \\
&= n(1/D) \begin{bmatrix} -\beta_2^2 & \beta_2 \\ -\beta_2/\sigma_1^2 & 1/\sigma_1^2 \end{bmatrix} V(1/D) \begin{bmatrix} -\beta_2^2 & -\beta_2/\sigma_1^2 \\ \beta_2 & 1/\sigma_1^2 \end{bmatrix} \\
&\qquad [\text{since } M^TP(M^{-1})^T = (M^{-1}PM)^T] \\
&= n(1/D^2) \begin{bmatrix} -\beta_2^2 & \beta_2 \\ -\beta_2/\sigma_1^2 & 1/\sigma_1^2 \end{bmatrix} \begin{bmatrix} 1/\sigma_1^2 & 0 \\ 0 & 1/\sigma_2^2 \end{bmatrix} \begin{bmatrix} -\beta_2^2 & -\beta_2/\sigma_1^2 \\ \beta_2 & 1/\sigma_1^2 \end{bmatrix} \\
&= (1/D^2)n\sigma_1^2(\beta_2^2 + \lambda) \begin{bmatrix} \beta_2^2 & (\beta_2/\sigma_1^2) \\ (\beta_2/\sigma_1^2) & (1/\sigma_1^2)^2 \end{bmatrix}.
\end{aligned} \tag{6.35}$$

Noting that $\partial v^{1,1}/\partial\theta = \partial(1/\sigma_1^2)/\partial\sigma_1^2 = -(1/\sigma_1^2)^2 \neq 0$, estimating equation for $\theta = \sigma_1^2$ according to equation (6.15) will be

$$\sum_{i=1}^{n} (y_{i2} - \beta_2 y_{i1})^2 - n\sigma_1^2(\beta_2^2 + \lambda) = 0,$$

$$\text{or, } \sigma_1^2(\beta_2^2+\lambda) - (1/n)\sum_{i=1}^{n}(y_{i2}-\beta_2 y_{i1})^2 = 0,$$

$$\text{or, } (\lambda+\beta_2^2)\sigma_1^2 - (\beta_2^2 S_{11} - 2\beta_2 S_{12} + S_{22}) = 0, \tag{6.36}$$

which is identical to the corresponding estimating equation for θ of Chan and Mak's (1983) method. Therefore, the methods are equivalent for unreplicated functional model (6.1) with $m = 2$. We have not investigated the equivalence of the methods theoretically for $m > 2$, because the algebraic expressions for the estimating equations become messier in higher dimensions, and seems analytically intractable. We have however seen numerically that the methods give identical results (see Table 6.1) for the model (6.2) with $m = 4$ when applied to the data set plotted in Figure 6.1. It therefore can be expected that the two methods are equivalent for the special case model (6.2).

6.3.5 Maximum likelihood estimators of structural relationship

Although Morton (1981) and Chan and Mak (1983) have shown their respective approaches produce consistent estimators, the question still remains concerning the existence of *efficient* estimators, or generally optimal estimators with respect to a given criterion.

Structural version of model (6.1), assuming that μ's are random variables and follow certain probability distribution, can be estimated by maximum likelihood method. We can obtain efficient estimators, against which to calibrate alternatives, in the situation where the μ's are independent realisation from a probability distribution with density function $p(\mu)$. Then the probability density of $Y_i \equiv (Y_{i1}, \cdots, Y_{im})$ is

$$p(Y_i) = \int \prod_{j=1}^{m} \frac{1}{\sigma_j}\phi\left(\frac{y_{ij}-\mu\beta_j}{\sigma_j}\right) p(\mu)d\mu, \tag{6.37}$$

where ϕ denotes the standard Gaussian probability density function. If p is known then we can, at least in principle, obtain efficient estimators by numerically maximising likelihood (6.37). Computations can be simplified if, in particular,

$$\mu \sim N(\zeta, \tau^2), \tag{6.38}$$

because then

$$Y_i \sim N_m(\zeta\beta, V), \tag{6.39}$$

and the likelihood for Y's is

$$\prod_{i=1}^{n}\left\{\frac{1}{|V|^{1/2}}\exp\left[-\frac{1}{2}(Y_i-\zeta\beta)^T V^{-1}(Y_i-\zeta\beta)\right]\right\}. \tag{6.40}$$

For simulation purpose we can estimate the β's and σ^2's, with ζ and τ^2 set to their true values, by maximising the likelihood function (LF) (6.40), to obtain efficient estimators, and then use (6.5) to estimate the μ's. Alternatively, we can simultaneously estimate ζ and τ^2 when maximising (6.40), which is equivalent to Factor Analysis with one factor. If assumption (6.38) is valid, either with or without $\zeta \equiv 0$, then one or both of these estimators should lead to efficient estimates of the μ's. But it remains to investigate how efficient the estimators are if assumption (6.38) is not valid.

A more flexible approach is to model the distribution of the μ's by a mixture of Gaussian distributions, *i.e.*,

$$\mu \sim N(\zeta_l, \tau_l^2), \qquad \text{with probability } \pi_l \qquad \text{for } i = 1, \cdots, L, \tag{6.41}$$

then the likelihood for the Y's is

$$\prod_{i=1}^{n} \left\{ \sum_{i=1}^{L} \frac{\pi_l}{|V_l|^{1/2}} \exp\left[-\frac{1}{2}(Y_i - \zeta_l\beta)^T V_l^{-1}(Y_i - \zeta_l\beta) \right] \right\}. \tag{6.42}$$

We use the likelihood (6.41) as a baseline for comparing efficiency of the estimators by alternative methods through simulation. We can estimate the β's and σ^2's by numerically maximising this likelihood, either assuming the ζ's, τ^2 and π's are known, or estimating them at the same time.

6.3.6 EM algorithm for estimating structural relationship

Aitkin and Rocci (2002) proposed an EM algorithm for maximum likelihood estimation of generalised linear structural models. For this method, the measurement error distribution can be of any specified form, although the implementation has been described using normal measurement error. The method does not necessarily require the distribution of the true-score (μ) of the variables with measurement error to be known.

For the description of the method for a simple structural regression, let y_{ij} and y_{ih} be the ith observations on the response and explanatory variables, and μ_i be the unobserved true-score corresponding to the observed y_{ih}. In addition to y_{ih}, observation z_i on an error-free covariate was assumed to be given to describe the estimation method. By allowing the true-score (μ) to depend on the error-free covariate z, the structural model was defined as

$$\begin{aligned} y_{ij}|y_{ih}, \mu_i, z_i &\sim N(\alpha + \beta\mu_i + \gamma_1 z_i, \sigma_j^2), \\ y_{ih}|\mu_i, z_i &\sim N(\mu_i, \sigma_h^2), \\ \mu_i|z_i &\sim N(\zeta + \gamma_2 z_i, \tau^2). \end{aligned} \tag{6.43}$$

For the subsequent analyses the model (6.43) was transformed so that the transformed true-score $\mu^* = \mu - \gamma_2 z$ has a homogeneous $N(\zeta, \tau^2)$ distribution. Now the original true-score can be expressed as $\mu = \mu^* + \gamma_2 z$. Defining $\gamma_1^* = \gamma_1 + \beta\gamma_2$, dropping the stars and suppressing notationally the conditioning on z, the model (6.43) can be expressed as

$$\begin{aligned} y_{ij}|y_{ih}, \mu_i &\sim N(\alpha + \beta\mu_i + \gamma_1 z_i, \sigma_j^2), \\ y_{ih}|\mu_i, &\sim N(\mu_i + \gamma_2 z_i, \sigma_h^2), \\ \mu_i &\sim N(\zeta, \tau^2). \end{aligned} \tag{6.44}$$

To construct the likelihood, Aitkin and Rocci (2002) treated the true-scores (μ_i) as missing data. The complete data log-likelihood function for model (6.44) was defined as

$$\begin{aligned} L = \sum_{i=1}^{n} \Bigg\{ & -\frac{1}{2}\log(2\pi) - \log\sigma_j - \frac{1}{2\sigma_j^2}(y_{ij} - \alpha - \beta\mu_i - \gamma_1 z_i)^2 \\ & -\frac{1}{2}\log(2\pi) - \log\sigma_h - \frac{1}{2\sigma_h^2}(y_{ih} - \mu_i - \gamma_2 z_i)^2 \\ & -\frac{1}{2}\log(2\pi) - \log\tau - \frac{1}{2\tau^2}(\mu_i - \zeta)^2 \Bigg\}. \end{aligned} \tag{6.45}$$

The authors proposed EM algorithm for estimating the parameters using the log-likelihood function (6.45) considering σ_h^2 as fixed. The missing data appear in the complete data log-likelihood as μ_i and μ_i^2 which are replaced in the E step by their conditional expectations. A modified mixture maximum likelihood approach of the above method was suggested for the full maximum likelihood estimation of generalised linear structural model with unknown σ_h^2 which does not require making any assumption about the distribution of μ_i other than that it is non-normal.

6.4 Application

We apply the methods to the subset of murine macrophage data plotted in Figure 6.1 which shows plot of all four scans of data $(y_{\cdot 1}, \cdots, y_{\cdot 4})$ against the data of the first scan $(y_{\cdot 1})$. The subset represents the linear range of the full data set described in Chapter 3 and is free from extreme outliers as seen in the full data. The way we select the subset has been described in Section 6.1. To implement the methods of Morton (1981) and Chan and Mak (1983) we use the IMSL routine `NEQNF`. The routine solves a system of nonlinear equations using a modified Powell hybrid algorithm and a finite-difference approximation to the Jacobian. The algorithm is a variation of Newton's method, which uses a finite-difference

approximation to the Jacobian and takes precautions to avoid large step sizes or increasing residuals. Further details of the algorithm can be found in More *et al.* (1980). For applying the method of second moments, Gaussian LF (6.40) and Gaussian mixture (6.42) methods, we use the simplex method of Nelder and Mead (1965) as the optimisation tool. The method is implemented using the IMSL routine `UMPOL`. Results of applying the methods of Section 6.3 are summarised in Table 6.1.

To obtain the estimates based on Gaussian mixture likelihood (6.42), the parameters of (6.41) were regarded as fixed at their maximum likelihood estimates with $L = 5$, which are tabulated in Table 6.2. These estimates give the best fit of the model (6.41) to the distribution of $\hat{\mu}$'s obtained from Morton's (1981) method. We have used 500 randomly chosen multiple starts from the plausible parameter space, and the estimates tabulated in Table 6.2 correspond to the solution with the highest value of the likelihood function. We see that only a small proportion ($\approx 3\%$) of the density belong to the fifth component distribution with large mean and standard deviation ($\zeta_5 = 3382, \tau_5 = 1634$). A histogram of the distribution of $\hat{\mu}$ along with the fitted model (6.41) to the distribution is shown in Figure 6.2. To have a clear idea about the fit at the right tail, we compare the estimated kernel density and the fitted mixture model by plotting them against log-scale of the horizontal axis (Figure 6.3). The plots suggest a good fit of the 5-component Gaussian mixture to the distribution of $\hat{\mu}$. A Q-Q type plot of the fit shown in Figure 6.2 is displayed in Figure 6.4. To obtain the approximate theoretical quantiles in the plot, we compute n_l quantiles from the l-th component distribution $N(\zeta_l, \tau_l^2)$ for $l = 1, \cdots, 5$, where n_l is the nearest integer to $n\pi_l$ satisfying the condition $\sum_{l=1}^{5} n_l = n$. The combined sets of quantiles are used as the theoretical quantiles of the corresponding Gaussian mixture distribution. Titterington *et al.* (1985, pp. 58–65) discussed similar approach to compute quantiles of mixture distributions. We see from Figure 6.4 that a few points ($\approx 0.3\%$) at the lower end (≤ 150) and a few more ($\approx 1.4\%$) at the upper end (≥ 3300) deviate from the mixture model. Apart from that the mixture model (6.41) fits the distribution of $\hat{\mu}$ reasonably well. A better impression of the fit can be seen from the log-scale version, omitting the points corresponding to 8 nonpositive theoretical quantiles, of the plot in Figure 6.5.

The results in Table 6.1 show that estimates obtained by the methods are very close to each other. Estimates of the regression coefficients are very similar for all methods. In terms of the scale estimates, the method of second moments produces slightly different estimates compared to the other methods.

Table 6.1: Estimated parameters of the model (6.1) applied to the data in Figure 6.1.

Estimation method	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\sigma}_1$	$\hat{\sigma}_2$	$\hat{\sigma}_3$	$\hat{\sigma}_4$
Method of 2nd moments	1.559	2.747	4.294	26.438	28.80	49.64	149.4
Morton (1981)	1.559	2.747	4.294	25.647	27.52	54.99	151.1
Chan and Mak (1983)	1.559	2.747	4.294	25.647	27.52	54.99	151.1
Gaussian LF (6.40)	1.559	2.747	4.294	25.588	27.56	55.08	151.0
Gaussian mixture (6.42)	1.559	2.747	4.294	25.642	27.68	54.84	150.8

Table 6.2: Maximum likelihood estimates of $\mu \sim \sum_{l=1}^{5} \pi_l N(\zeta_l, \tau_l^2)$. $\hat{\mu}$'s are obtained from Morton's (1981) method.

	Component (l)				
Parameter estimates	1	2	3	4	5
$\hat{\zeta}_l$	320.60	453.54	809.34	1531.55	3382.06
$\hat{\tau}_l$	47.66	90.06	241.18	536.93	1634.38
$\hat{\pi}_l$	0.25	0.27	0.28	0.17	0.03

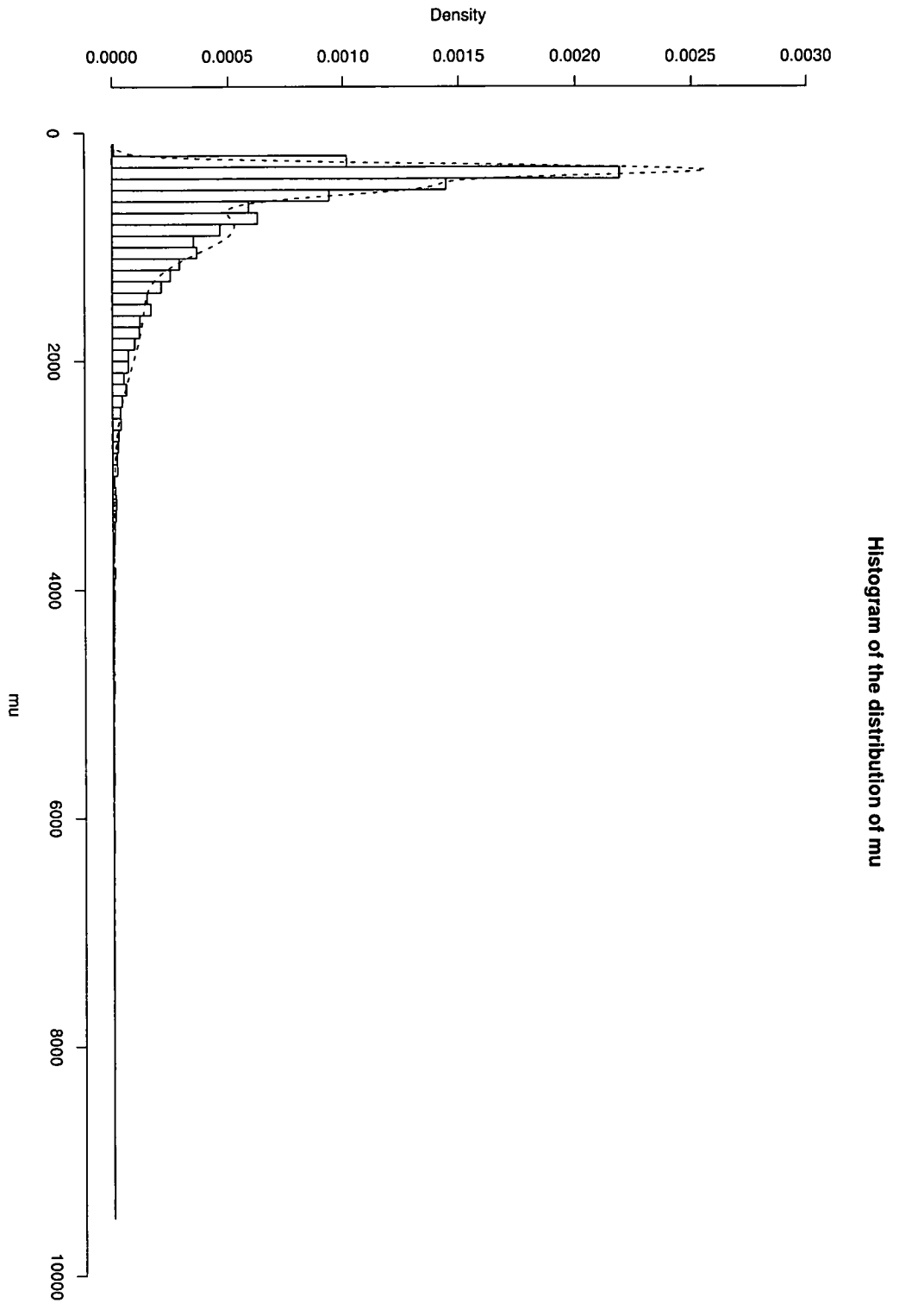


Figure 6.2: Histogram of the distribution of $\hat{\mu}$ obtained from Morton's (1981) method. The dashed line indicates the fitted mixture model (6.41) to the distribution.

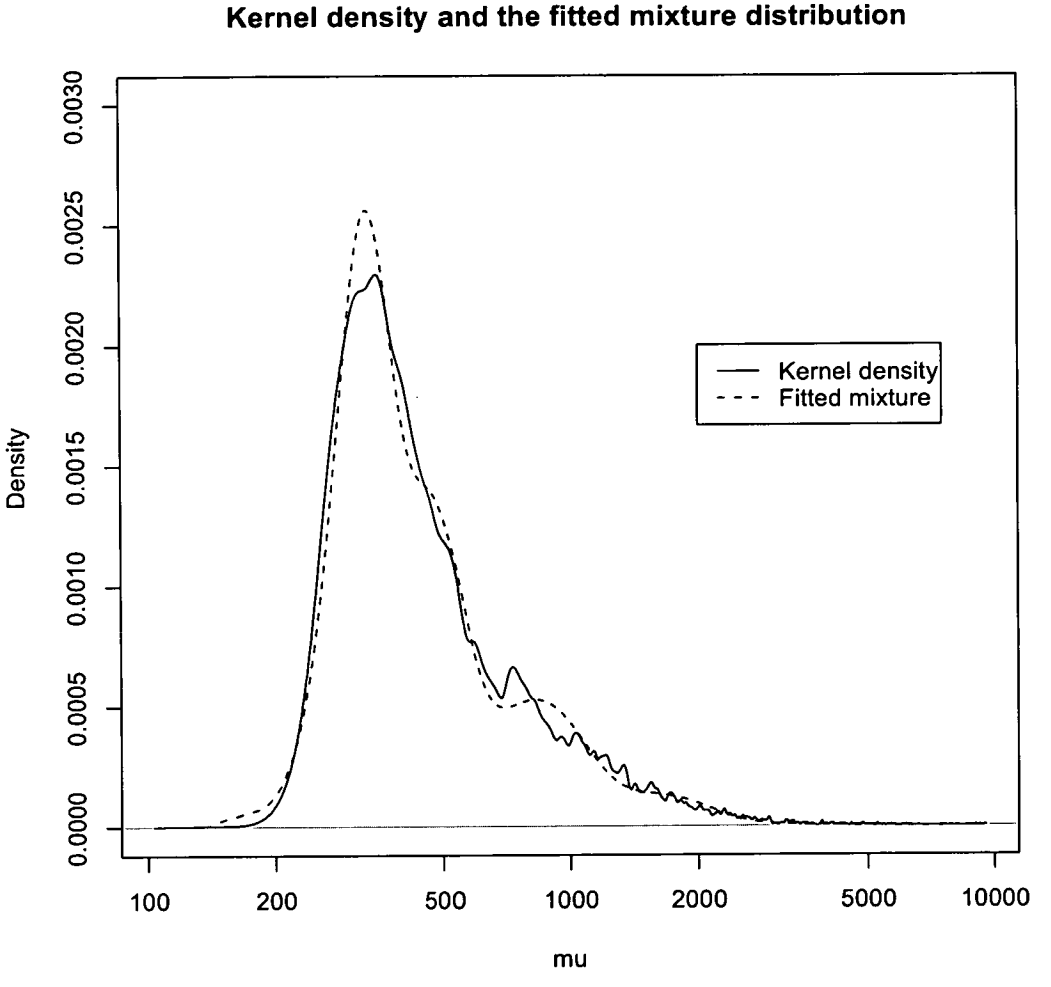


Figure 6.3: Kernel density estimates and the fitted Gaussian mixture plotted using log-scale of horizontal axis.

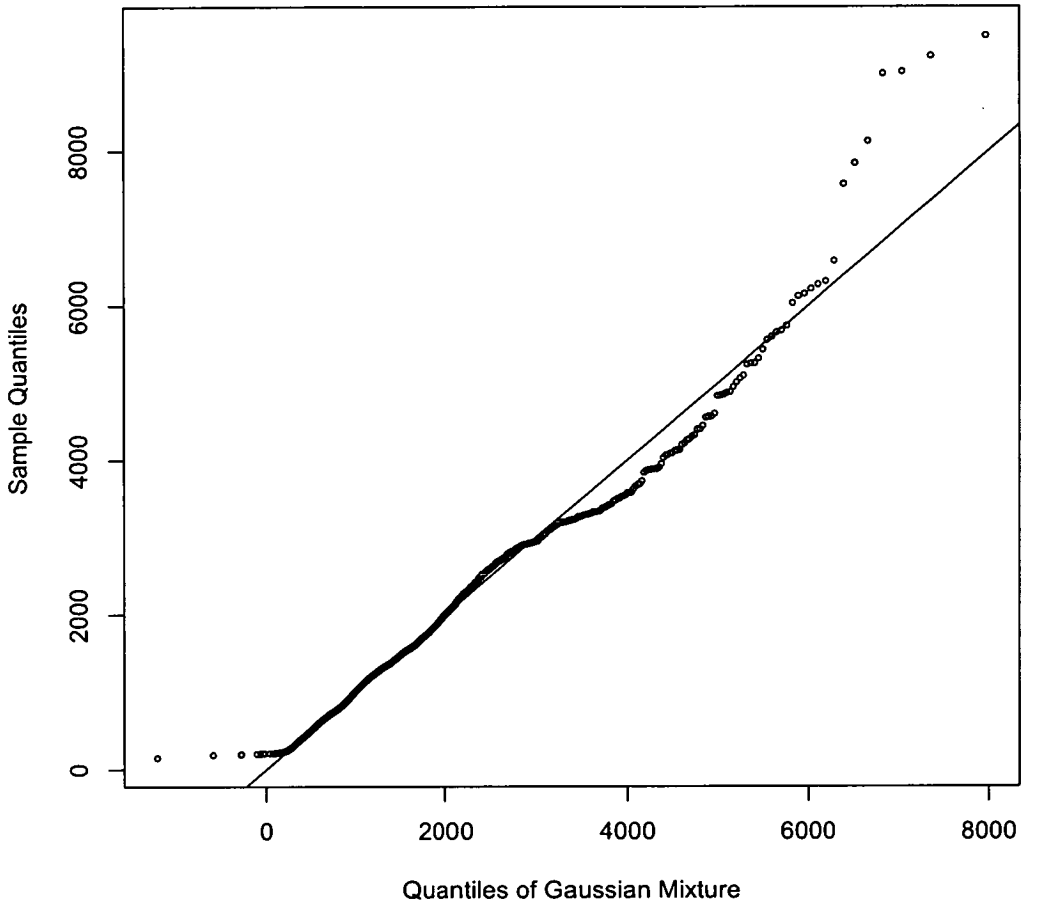


Figure 6.4: Q-Q type plot of the fitted Gaussian mixture (6.41) to the distribution of $\hat{\mu}$ with $L = 5$. Theoretical quantiles in the plot are obtained by combining quantiles proportionately from each component of the mixture distribution.

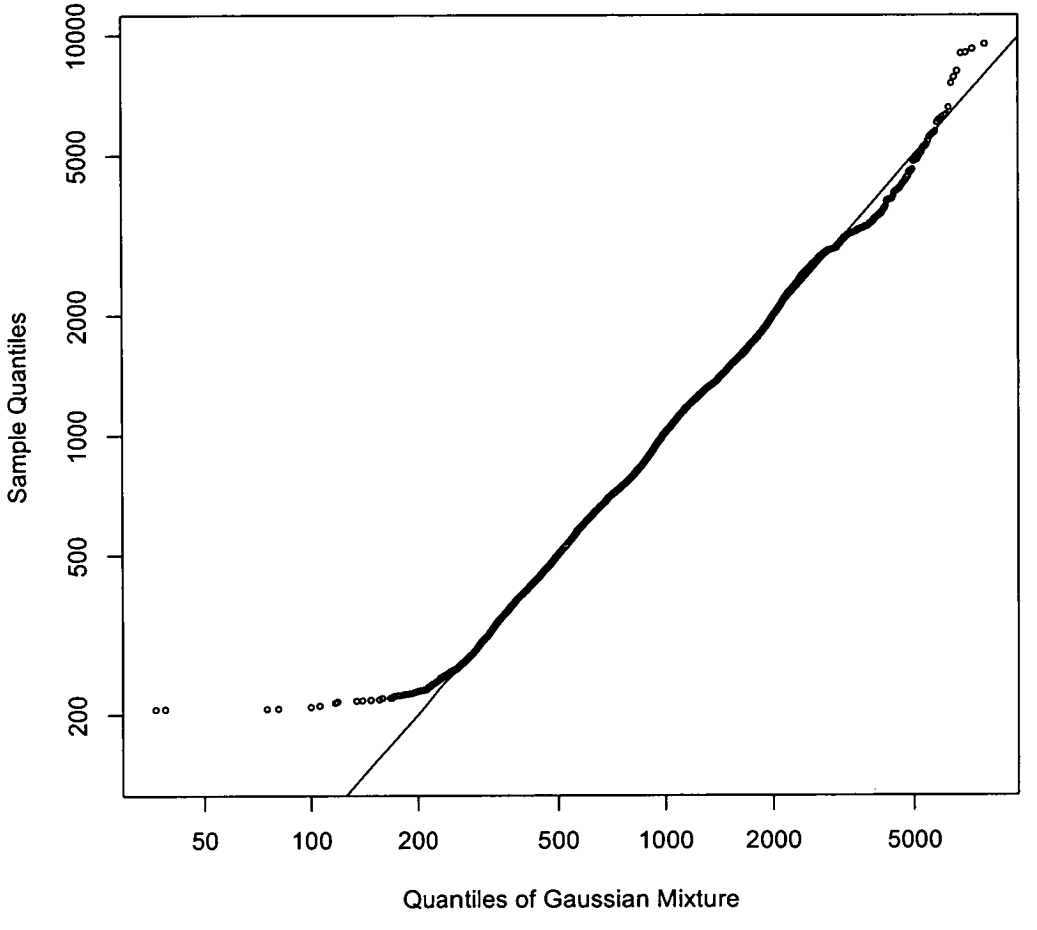


Figure 6.5: Log-scale version of Figure 6.4

The methods of Morton (1981) and Chan and Mak (1983) lead to the same estimates for all parameters. From this numerical finding and the theoretical equivalence of the methods for $m = 2$ shown in Section 6.3, it can be expected that the two methods lead to identical estimating equations for the special case model (6.2).

6.5 Simulation study

We conducted a simulation study to compare estimators from the methods described in Section 6.3 in the context of the special case model (6.2). We used estimates of β's and σ^2's from Morton's (1981) method (Table 6.1), and the fitted parameters of the 5-component Gaussian mixture to the $\hat{\mu}$'s (Table 6.2) as the true values, and then generated 1000 data sets according to the Gaussian mixture model (6.42). We then estimated the parameters by each of the methods of Section 6.3 and compared the estimators in terms of bias and efficiency. We computed % root-mean-squares to compare the efficieny of the individual parameter estimates, and also a generalised variance type summary measure $|\Delta|$, which is the determinant of the $(2m-1) \times (2m-1)$ approximate covariance matrix Δ, given by

$$\Delta = \frac{1}{N} \sum_{r=1}^{N} (\hat{\lambda}_r - \lambda)(\hat{\lambda}_r - \lambda)^T, \tag{6.46}$$

where N is the number of simulated data sets and

$$\lambda = (\beta_2, \cdots, \beta_m, \sigma_1^2, \cdots, \sigma_m^2)^T$$

is the $(2m-1)$-vector of parameters.

Table 6.3 summarises the estimated % bias, % root-mean-squares, % standard errors (ESE) and the summary efficiency measure $|\Delta|$ of the parameter estimates according to each of the methods based on 1000 simulated data sets. The final column gives the relative efficiency (RE) as % of the efficiency $(1/|\Delta|)$ of the Gaussian mixture method. The generalised variance measures $|\Delta|$'s are estimated to be very small. We present them in Table 6.3 after multiplying by the factor 10^5 for ease of comparison.

We see by comparing the simulation results for the individual regression parameters β's that all the methods are almost unbiased and equally efficient in estimating the regression coefficients. Performance of the methods mainly varies in terms of the variance estimators. However, except for the method of second moments, all the methods also perform similarly in terms of bias and efficiency of the variance estimators.

Table 6.3: Estimated % bias, % root-mean-squares, % standard errors (ESE) and the summary efficiency measure $|\Delta|$ of the parameter estimates according to each of the methods. The final column gives the relative efficiency (RE) as % of the efficiency of the Gaussian mixture method. Results are based on 1000 simulated data sets.

Estimation method	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\sigma}_3^2$	$\hat{\sigma}_4^2$	$\lvert\Delta\rvert \times 10^5$	RE(%)
Method of 2nd moments–Section 6.3.1, equation (6.4)									
% root-MSE	0.0343	0.0355	0.0482	2.7063	3.7576	3.1440	1.8715	12.9018	29.3
% ESE	0.0343	0.0355	0.0482	2.7060	3.7554	3.1427	1.8713		
% Bias	0.0007	0.0014	0.0014	−0.0414	−0.1292	0.0926	0.0215		
Morton (1981)–Section 6.3.2, equations (6.14)–(6.15)									
% root-MSE	0.0343	0.0354	0.0482	2.0758	2.8616	2.4430	1.8489	3.8472	98.2
% ESE	0.0343	0.0354	0.0482	2.0739	2.8616	2.4405	1.8488		
% Bias	0.0007	0.0014	0.0014	0.0890	−0.0030	0.1102	0.0164		
Chan and Mak (1983)–Section 6.3.3, equations (6.23)–(6.24)									
% root-MSE	0.0343	0.0354	0.0482	2.0746	2.8559	2.4409	1.8493	3.8251	98.8
% ESE	0.0343	0.0354	0.0482	2.0726	2.8559	2.4384	1.8492		
% Bias	0.0007	0.0014	0.0014	0.0909	0.0011	0.1097	0.0163		
Gaussian LF–Section 6.3.5, equation (6.40)									
% root-MSE	0.0343	0.0354	0.0482	2.0794	2.8660	2.4449	1.8497	3.8819	97.3
% ESE	0.0343	0.0354	0.0482	2.0775	2.8660	2.4424	1.8496		
% Bias	0.0007	0.0014	0.0014	0.0888	−0.0054	0.1124	0.0163		
Gaussian mixture–Section 6.3.5, equation (6.42)									
% root-MSE	0.0343	0.0354	0.0482	2.0763	2.8513	2.4314	1.8470	3.7784	100.0
% ESE	0.0343	0.0354	0.0482	2.0763	2.8513	2.4314	1.8470		
% Bias	0.0007	0.0014	0.0014	0.0894	0.0048	0.1071	0.0127		

The Gaussian mixture method (6.42) is efficient and can be used as a baseline for comparing the performance of the methods. Simulation shows that the method achieves the highest efficiency, *i.e.*, the smallest generalised variance (3.7784×10^{-5}) among the alternatives compared. In terms of the overall performance, method of second moments is the least efficient (RE= 29.3%). The method is also more biased than the other methods with respect to the estimators of variance parameters. Methods of Morton (1981) and Chan and Mak (1983), that are expected to be equivalent for the functional model (6.1), show similar performance in terms of overall efficiency with estimated generalised variances 3.8472×10^{-5} and 3.8251×10^{-5}, and relative efficiencies 98.2% and 98.8% respectively. Furthermore, the methods are almost as efficient as the Gaussian mixture method. The method of single component Gaussian LF (6.40) is also highly efficient ($|\Delta| = 3.8819 \times 10^{-5}$, and RE= 97.3%), but is slightly outperformed by the methods of Morton (1981) and Chan and Mak (1983).

6.6 Discussion

Measurement error models, in particular functional relationship models, are a generalisation of standard regression models. Such models can be applied more appropriately than the standard regression technique in many data analysis problems, but the analysts commonly prefer to use the more familiar and easy to use standard regression models. These models have both advantages and disadvantages compared to standard regression models. One major limitation is that the conventional maximum likelihood estimation leads to some anomalies. For example, parameters are often not estimable by direct application of maximum likelihood estimation, and sometimes lead to inconsistent estimators when they are estimable. An extensive literature is available suggesting alternative approaches mainly focusing on the problem of estimability and consistency, which result from the presence of incidental parameters whose numbers increases with the sample size. The alternative estimators suggested in the literature have not been compared with respect to other important properties of estimators, such as unbiasedness and efficiency.

Although application of the methods to real data produces similar estimates, our simulation study shows that the methods vary in their performance with respect to efficiency. It is known that, when μ's can be assumed to be distributed as Gaussian, the parameters of model (6.1) can be estimated efficiently by maximum likelihood method. In our study, we have used efficiency of such a method as a baseline to compare the performance of the alternative methods. Because the

distribution μ is positively skewed and a Gaussian assumption may not be valid in most applications, we consider a Gaussian mixture to approximate the probability distribution of μ in model (6.1). We see from the simulation results that the performance of the methods studied mainly vary with respect to the efficiency of the estimators of variance parameters. Estimated % bias and % root-mean-squared errors of the estimators of the regression parameters have been found to be similar for all methods. Comparing the estimated generalised variances of the alternative estimators it can be concluded that the methods of Morton (1981) and Chan and Mak (1983) give the most efficient estimators of the functional regression model (6.1). Because the methods also produce consistent estimators, they are likely to be the best options for estimating functional models. We have shown that the two methods lead to the same estimating equations for the model in bivariate case. They also gave the same numerical results when applied to the data plotted in Figure 6.1. From these findings, together with the observation that they produce very similar estimate of generalised variance, we can expect that the methods are equivalent for the functional model (6.1). Therefore, the methods can be considered equally good for the analysis of functional relationships. However, in microarray applications main interest is in the estimation of the gene expression parameters (μ), and in contrast to the general applications of functional models, β and $\theta = (\sigma_1^2, \cdots, \sigma_m^2)$ are really nuisance parameters in our case. It therefore remains to investigate if the methods provide consistent and efficient estimators of the μ's, and if not, how the estimating equations for μ can be modified for consistency and efficiency.

It is however surprising to note that, despite having elegant properties, the methods have not found much application in the literature. Application of standard regressions, when observations of both response and explanatory variables are subject to measurement errors, may be misleading. There is a general preference of using standard regression approach because of its familiarity and ease of application even when functional regression would be more appropriate. One such example is Dudley *et al.* (2002) who used standard linear regression approach to calibrate multiple scans of microarray data. Because all scans of microarray data are subject to measurement errors, the functional model (6.1), with estimation method of Morton (1981) or Chan and Mak (1983) would be a more appropriate choice for doing such analysis.

It may however be mentioned that this investigation is limited to a special case, unreplicated functional models without intercepts. Methods of Morton (1981) and Chan and Mak (1983) were proposed in a more general context, and for models with intercepts. It would be worth investigating if the omission of intercept

term has any effect on the consistency property of the estimating equations. We modified the data, *i.e.*, used uncorrected first and second moments in the estimating equations, for exclusion of the intercept, but further investigation is needed to see if deletion of intercept has any impact on the consistency property of the estimating equations.

Chapter 7

Discussion and future work

The main aim of the research reported in this thesis was to develop statistical methods to provide better alternatives for resolving some of the existing problems in the data preprocessing step of microarray analysis. This chapter provides an overall discussion of the work presented in this thesis. Indications for possible further research on the problems studied are also given at the end of this chapter.

7.1 Review

Two of the major objectives of this thesis were to

- develop statistical model for combining multiple scan data to correct for "signal saturation" and "signal deterioration" effects in the gene expression measurement, and
- suggest unified approach for nonparametric location and scale normalisation of microarray data.

We have also investigated the efficiency of different estimation methods for functional regression models through simulation studies. Discussions are organised in the following three subsections corresponding to these objectives.

7.1.1 Combining multiple scan data

The motivation behind combining multiple scan data was to deal with the systematic bias in the gene expression measurements, induced due to the limitation of acquisition device (scanner), during the scanning process of the hybridised microarrays. The two conflicting problems at this stage affecting highly expressed and weakly expressed genes respectively are termed as 'signal saturation' and 'signal deterioration'. Signal saturation occurs when the pixel intensity exceeds $2^{16} - 1$, the threshold for a 16-bit (double precision) image converting software,

and the recorded pixel intensity is censored at this threshold value. As a result, estimators of gene expression are biased, with the amount of bias increasing as a function of the proportion of pixels saturated. Signal deterioration refers to the problem at the other extreme where the noise dominates the very weak signal intensity. There is a trade-off between the problems and are related to the photomultiplier tube (PMT) voltage settings of the scanner. Because low-intensity spots are better measured at high PMT settings and highly expressed genes are better measured at low settings, no single setting can be optimal for both ends. We therefore consider combining multiple scan data obtained at a series of increasing PMT settings to get improved gene expression estimates.

The exploratory analysis presented in Chapter 3 provided useful information regarding the basic patterns of multiple scan data and helped us decide the refined model, presented in Chapter 4, for combining such data. The major patterns we discovered through exploratory analysis are that spot intensity measurements, particularly corresponding to higher PMT settings, are not linearly related to the gene expression levels across the entire range of the data and the variability of the data depends quadratically on the expression levels. Our initial attempt of modelling the nonlinearity with a hyperbolic function was not fully successful as the function is not flexible enough to comply with the varying nature of the nonlinearity found in different applications. One limitation of the hyperbolic function is the lack of flexibility to model the case of extreme censoring when the nonlinearity resemble almost a piecewise linear relation as found in iron-deficiency data (Section 4.4.2). The censored mean function, which is the location function of our refined model in Chapter 4, has been found to overcome these limitations of the hyperbolic function. The function has a natural analogy with the data generation mechanism of the scanner and has been found to provide satisfactory fit to the data sets considered in this study.

Among the several robust options we have experimented with in this study, use of maximum likelihood method based on a heavy tailed Cauchy model and a censored Cauchy model seems to be the most reasonable choice. Although application of the Cauchy and censored Cauchy models to the data sets used in this thesis produces very similar results, the censored Cauchy model is a more realistic choice because it takes account of the fact that spot averages cannot exceed the censoring threshold T, and in case of moderate or heavy censoring, censored Cauchy model can be expected to give better results than the Cauchy model. M-estimation using a Gaussian likelihood type objective function, downweighting the observations outside 3 standard deviations, with the censored mean functional model did not fit the data well. We have however found that the amount

of noise and outlying observations in multiple scan data vary in different applications. For example, in our applications, murine macrophage data are noisier and have more outliers than the iron-deficiency data. It seems that Gaussian distributed error with the censored mean functional model would also be adequate for the iron-deficiency data. Although the Cauchy distribution might be found to have heavier tail than the data require in some applications, providing some extra robustness in the estimation procedure does not do any harm. We have noted some downward bias in the maximum likelihood estimates of the additive and multiplicative scales σ_1 and σ_2 of the refined model for combining multiple scan data based on Cauchy and censored Cauchy distributions. This bias seems to arise due to combining a large number n of small samples of size m for estimating common scale parameters. Pattern of bias in this case is different from that of maximum likelihood estimate of Cauchy scale based on single sample and there seems to be no standard way of correction for this bias. This bias however does not affect the estimation of the gene expression parameters which is the main focus of the model.

Noting that, in terms of tail behaviour, t-distribution lies between Cauchy and Gaussian, and the tail weight of a t-distribution depends on the degrees of freedom of the distribution, we investigated the possibility of using a t-distribution as an alternative to the Cauchy model. This introduces an extra parameter, the degrees of freedom, which ideally need to be estimated from the data. However, in addition to the bias in the scale estimates, maximum likelihood has been found to substantially underestimate the degrees of freedom.

The murine macrophage data and iron-deficiency data described in Chapter 4 show considerable dissimilarities in terms of the patterns of multiple scan data. The censored mean functional model based on Cauchy and censored Cauchy distributions have been found to be flexible enough to provide satisfactory fit to both data sets. Comparison of between replicate variations computed from the raw data and from the estimated gene expression suggests that combing multiple scans according to the proposed model can reduce the signal-to-noise ratio in the intensity measurements.

The major strengths of the proposed model over the few existing methods of its kind in the literature are that, it is

- simple, robust and objective,
- based on full information of multiple scan data, and
- defined realistically to be consistent with the natural behaviour of the data.

The method is based on the simple and familiar censored Cauchy model, which provides a basis for robust and objective method of estimation taking account of the fact that spot averages, like the indididual pixel values, cannot exceed the censoring threshold T. The method is objective in the sense that, unlike other robust methods, it does not require choosing any weight function or any tuning constant to control the amount of robustness needed in a particular application.

The proposed method utilises the complete information from multiple scan data, which is not the case for most of the existing methods for handling the problem. Dudley *et al.* (2002), for example, loses information by discarding observations in the nonlinear range. Bell's (2003) algorithm also considers data within a certain range that are not likely to be affected by signal censoring and deterioration. In addition to losing information, these methods involve certain degree of subjectivity in choosing the 'linear range' or the portion of data not affected by censoring or deterioration.

The model is realistically built to represent the behaviour of the data. The nonlinear relationship, the censored mean function, has a natural analogy with the data generation mechanism of the scanner and represents the trend of the data nicely. This we think is more realistic than the linear and gamma curve representations of García de la Nava *et al.* (2004) for modelling the relationship of multiple scan data.

Some weaknesses of the method are also worth mentioning. The model would have been more realistic if we could consider separate variance terms for individual scan of data. This is however not easy in the framework of functional model because of the identifiability problem of the parameters. Another drawback is that, in terms of tail behaviour, Cauchy or censored Cauchy distribution may not always be the appropriate match of the data. Choice of Cauchy model has further disadvantage of having a multimodal likelihood which poses complexity in the estimation procedure and reduce computational efficiency. Although the main interest is in the gene expression parameters, adjustment of the bias in the scale estimates is also desirable.

Nevertheless, the proposed method is a considerable improvement over the existing methods and provides a reliable and elegant way of combining multiple scan data to get improved gene expression estimates.

7.1.2 Nonparametric location and scale normalisation

Normalisation is a much discussed issue in the microarray literature. Generally location normalisation and scale normalisation (variance stabilisation) are treated separately. For example, loess smoothing is routinely used for normalising the lo-

cation of log-ratio data to remove intensity dependent and spatial effects. On the other hand, parametric models, *e.g.*, additive plus multiplicative model, are widely used for variance stabilisation. Noting the limitations of parametric models to be fully adequate for modelling the complex nature of the variability of microarray data, we studied the use of nonparametric methods for both location and scale normalisation under a common framework. The proposed model using GAMLSS (Chapter 5) applied to lymphoma and iron-deficiency data suggests that it can model the various trends in both location and scale of the data and therefore may be a suitable normalisation method in microarray applications. One advantage of the method is that, GAMLSS normalised data are comparable across arrays, although the method is applied individually to each array for in-slide normalisation. Comparison of the method with the parametric AVST method applied to the data sets considered in this study suggests that GAMLSS identifies relatively more genes as differentially expressed than the AVST method for certain arrays investigated. While comparing the power of GAMLSS normalisation with that of AVST method in inferring differential expression using simulated data, it has been found that GAMLSS is as good as the AVST method when data are generated considering the parametric model as true. Whereas, GAMLSS method has been found to be considerably more powerful than the parametric AVST method when GAMLSS location and scale models fitted to real data are considered as true. The simulations in Section 5.6 demonstrate that GAMLSS normalisation improves the inference on differential expression even when applied to the AVST normalised data.

As with loess normalisation, the method is however based on the assumption that, either the proportion of differentially expressed genes are small, or there is symmetry in the expression values between up-regulated and down-regulated genes. If at least one of these two assumptions is not satisfied, the method may incorrectly normalise the differential expression. Dependence on these assumptions may be alleviated by applying the method to a set of invariant genes that are likely to be constantly expressed.

7.1.3 Efficiency of functional regression estimators

We experienced some interesting problems while investigating the use of functional regression models, in Chapters 3 and 4, for combining multiple laser scans through maximum likelihood estimation. One of the problems was concerned with the estimability of the parameters. Parameters are not estimable by direct application of maximum likelihood estimation without prior information about the variance parameters. This problem has been discussed extensively in the lit-

erature, and we have also noticed in Section 3.4.2 that the likelihood function becomes infinite as any of the variance parameters approaches zero because of the rounding errors in the data (Copas, 1972). Although estimability can be restored by imposing certain restrictions on the parameter space, Sprent (1976) showed that the maximum likelihood may lead to inconsistent estimators in such cases. Furthermore, these restrictions, *e.g.*, the assumption that the variance ratios or a subset of the variance parameters are known, are not often feasible in most applications. Alternative estimation methods that are available in the literature, such as, Morton (1981) and Chan and Mak (1983), mainly address the problems of identifiability and consistency. It is therefore reasonable to investigate how good these estimators are in terms of efficiency. We limited our investigation to the no intercept linear functional model, and used the efficiency of the maximum likelihood estimators in the corresponding structural model, assuming a Gaussian mixture distribution for μ, as a baseline for calibrating the efficiency of the alternatives. These estimators are theoretically known to be efficient, and also achieved the highest efficiency among the alternatives compared in our investigation. Simulation shows that the methods of Morton (1981) and Chan and Mak (1983), based on the likelihood of pivots rather than the likelihood of the observations themselves and the likelihood equations modified for consistency respectively, are almost as efficient as the estimators of Gaussian mixture structural model. Another interesting result we found is that the two methods lead to identical estimating equations for our no intercept linear functional model in the bivariate case. Although we could not prove the equivalence in general because the problem seems analytically intractable in higher dimensions, we have seen numerically that the methods give identical results in four dimensions when applied to our example data set. It may therefore be the case that the two methods are actually equivalent for the no intercept linear functional models. Further investigation is however needed to establish this point.

7.2 Future work

There is scope for further research on the problem for combining multiple scan data, on the nonparametric method of location and scale normalisation and on the efficiency investigation of the functional regression estimators. This is described in the following three sections.

7.2.1 Combining multiple scan data

This is again divided in to two heads, one describing possible further analysis of the proposed censored mean functional model and the other indicating the alternative models or statistical methodology that may be worth investigating for studying the problem.

Further analysis of censored mean model

One possible way of further investigating the utility of the censored mean functional model for combining multiple scans would be to apply the model in some actual scientific analysis. For example, results of formal analysis on some data set could be compared with and without applying the model of multiple scan to see how the method improves detection of differentially expressed genes, or the inference on the scientific question of interest for that particular analysis. It would also be interesting to investigate the performance of normalisation and variance stabilisation methods on the estimated gene expressions obtained from the multiple scan model.

This thesis illustrates the method using iron-deficiency and murine macrophage data consisting observations from 3 and 4 scans respectively. It would be useful to investigate the effect of number of scans considered on the performance of the method. This could be done by applying the method to other data sets with different multiplicity m of scan and through simulation study with a broader range of m values. This might eventually give guidelines about the ideal number of scans to be considered in such studies. Simulation studies with different m values would also be useful for a more detail study of the bias in the scale estimates of the censored mean model, which has been investigated for varying n but fixed m $(= 4)$. Since both n and m could vary in different applications, it would also be useful to see if m has any effect on this bias.

Alternative methods

Alternative measurement error (ME) models, *e.g.*, *ultrastructural* relationship (Dolby, 1976) models could be investigated as alternative to the functional relationship model. Although the functional and ultrastructural models are similar in appearance, they vary with respect to the underlying model assumptions and properties of the estimated model parameters. Functional models treat μ's as n unknown constants, whereas ultrastructural models consider μ's as independent random variables with different means and common variance. A comparison of these models in combining multiple scan data may therefore be interesting.

Other estimation methods, as alternative to maximum likelihood methods, could be investigated. We have seen in Chapter 6 that there are alternative methods, *e.g.*, Morton (1981) and Chan and Mak (1983) based on likelihood of certain functions of observations rather than the observations themselves and likelihood equations modified for consistency respectively, that provide consistent and highly efficient estimators for the functional regression models with Gaussian distributed errors. It would be interesting to apply these methods to our problem after modifying them for Cauchy distributed errors.

Ordinary least squares method does not work for ME models. Generalised least-squares has been shown in the literature (Sprent, 1966) to work with functional models. It seems reasonable to investigate such estimation method, because least squares gives such simple and elegant results for ordinary regression.

Another direction of extensive research on the problem would be to consider a Bayesian approach. One criticism often cited in the literature of ME models is the inconsistency of the maximum likelihood estimation for the functional and ultra-structural relationships. This is due to the incidental parameters whose number increases with the sample size. Bayesian approach to this problem, introduced by Lindley and El-Sayyad (1968), is suggested as an alternative to handle this problem. Although the Bayesian theory of estimation of functional relationship is well established (Lindley and El-Sayyad, 1968; Zellner, 1971; Florens *et al.*, 1974; Reilly and Patino-Leal, 1981), the way the method treats the incidental parameters, μ_i in our case, is not practical for the current problem. In the language of Neyman and Scott (1951) the parameters $\beta, \sigma_1, \sigma_2$ and ν of the censored mean functional model are called "structural" because they occur in the joint distribution of every observation. Whereas μ's are "incidental" parameters as μ_i is incidental to the ith observation alone. The Bayesian treatment of incidental parameters is to integrate them out from the likelihood with respect to a prior distribution conditioned on all remaining known or unknown parameters. Main focus of Bayesian studies of functional relationship therefore concerns the estimation of the common structural parameters only. This is the main thrust of frequentist approaches as well where the incidental parameters are mostly treated as nuisance. This however is not the case with our censored mean functional model. Main interest here is to estimate the gene expression parameters (μ_i). So the existing Bayesian approaches of handling functional relationships need to be modified for studying the censored mean functional model for combining multiple scan data.

7.2.2 Nonparametric location and scale normalisation

The GAMLSS normalisation method reported in this thesis can also be extended in many ways. P-splines was used for fitting the model in all the applications and simulations presented considering its desirable properties and computational advantage over the other smoothing methods. It would be interesting to see how the other smoothing methods, *e.g.*, *loess* and smoothing splines, compare with p-splines in fitting the model. Simulation studies could also be extended to compare the performance of the smoothing methods in inferring differential expression.

Although the GAMLSS method presented in this thesis models only the intensity-based trends in the location and scale of the log-ratio data, the method is flexible enough to accommodate the spatial and print-tip effects in both location and scale models. It therefore could be used to correct for spatial and print-tip bias as well by incorporating the row, column and print-tip factors.

In its presented form, GAMLSS method can only be applied for within-slide normalisation. Although the within-slide GAMLSS normalised data are comparable across slides, it would be worth investigating if the model could be generalised to multiple-slide normalisation, and if the multiple-slide normalisation has any advantage over the within-slide normalisation. One possible way of doing this would be to consider multivariate GAMLSS. This will however require extending the existing univariate GAMLSS theory to the multivariate case, which may not be very straight forward.

Another, possibly more realistic, extension of GAMLSS normalisation would be to consider the scale model as a combination of the smooth component $\sigma(x)$ and a parametric, fixed or random, gene specific component. This kind of formulation, termed semiparametric GAMLSS, is allowed within the GAMLSS framework but would require within-slide replication.

7.2.3 Efficiency of functional regression estimators

Our investigation on the efficiency of different estimators of functional regression was limited to only a special case, no intercept linear functional models for unreplicated data with Gaussian distributed errors. The problem could be further investigated in several directions. For example, the methods of Morton (1981) and Chan and Mak (1983) were proposed in a more general context, and for models with intercepts. One obvious question that needs further investigation is whether the methods require any modification for the omission of the intercept term for the estimators to be still consistent.

It is also important to compare the performance of the methods for more

general models, *e.g.*, linear models with intercept term and nonlinear functional models, and models with replicated observations.

Another, more extensive, direction of research would be to extend and compare efficiency of these estimation methods for non-Gaussian, such as Cauchy or t-distributed, errors.

As was discussed in Section 6.6, in microarray applications, the structural parameters $\theta = (\sigma_1^2, \cdots, \sigma_m^2)$ are really nuisance parameters, and main interest is in the estimation of the gene expression parameters. Consistency and efficiency of the μ's are more important than those of β and θ. An interesting problem of future research would therefore be to investigate if the estimation methods discussed in this thesis give consistent and efficient estimates of μ's. If they are found to be not consistent, it would be reasonable to investigate if the modifications as in (6.23) and (6.24) can be applied to the estimating equations of μ's to estimate them consistently.

References

Adam, R. and Bischof, L. (1994). Seeded region growing. *IEEE transactions on pattern analysis and machine intelligence*, **16**, 641–647.

Aitkin, M. and Rocci, R. (2002). A general maximum likelihood analysis of measurement error in generalised linear model. *Statistics and Computing*, **12**, 163–174.

Alizadeh, A. A. *et al.* (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.

Amaratunga, D. and Cabrera, J. (2004). *Exploration and Analysis of DNA Microarray and Protein Array Data.* New Jersey: Wiley.

Anderson, T. W. and Rubin, H. (1956). Statistical inference in factor analysis. In *Proceedings of the Thired Berkeley Symposium on Mathematical Statistics and Probability* (ed. J. Neyman), **5**, 111–150.

Axon Instruments Inc. (1999). *GenePix 400A User's Guide.*

Bai, Z. D. and Fu, J. C. (1987). On the maximum-likelihood estimator for the location parameter of a Cauchy distribution. *Canadian Journal of Statistics*, **15**, 137–146.

Barnett, V. D. (1966). Evaluation of the maximum-likelihood estimator where the likelihood equation has multiple roots. *Biometrika*, **53**, 151–165.

Bell, R. (2003). Help for the MWG MAVI PRO 2.6.0, MWG BIOTECH AG, Anzinger Strasse 7, 85560 Ebersberg, Germany.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, **57**, 289–300.

Beucher, S. and Meyer, F. (1993). The morphological approach to segmentation: the watershed transformation. In *Mathematical Morphology in Image Processing, Volume 34 of Optical Engineering.* New York: Marcel Dekker.

Box, G. E. P. and Cox, D. R. (1964). An anlysis of transformations. *Journal of the Royal Statistical Society, Series B*, **26**, 211–252.

Brooks, S. P. and Morgan, B. J. T. (1995). Optimization using simulated annealing. *The Statistician*, **44**, 241–257.

Buckley, M. J. (2000). *The Spot User's Guide.* CSIRO Mathematical and Information Sciences. Available at `http://www.cmis.csiro.au/IAP/Spot/spotmanual.htm`.

Chambers, J., Angulo, A., Amaratunga, D., Guo, H., Jiang, Y., Wan, J. S., Bittner, A., Frueh, K., Jackson, M. R., Peterson, P. A., Erlander, M. G. and Ghazal, P. (1999). DNA Microarrays of the complex human cytomegalovirus genome: profiling kinetic class with drug sensitivity of viral gene expression. *Journal of Virology*, **73**, 5757–5766.

Chan, N. N. and Mak, T. K. (1983). Estimation of multivariate linear functional relationships. *Biometrika*, **70**, 263–267.

Cheng, C.-L. and Van Ness, J. W. (1999). *Statistical Regression with Measurement Error*, London: Arnold.

Clarke, B. R. (1983). Uniqueness and Fréchet differentiability of functional solutions to maximum likelihood type equations. *The Annals of Statistics*, **11**, 1196–1205.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, **74**, 829–836.

Cleveland, W. S. and Devlin, S. J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, **83**, 596–610.

Cleveland, W. S. Grosse, E. and Shyu, M. (1993). Local regression models. In *Statistical Modelling in S* (eds I. Chambers and T. Hastie), pp. 309–376. New York: Chapman and Hall.

Copas, J. B. (1972). The likelihood surface in the linear functional relationship problem. *Journal of the Royal Statistical Society, Series B*, **34**, 274–278.

Copas, J. B. (1975). On the unimodality of the likelihood for the Cauchy distribution. *Biometrika*, **62**, 701–704.

Crick, F. H. C. (1958). On Protein Synthesis. *Symposia of the Society for Experimental Biology*, **12**, 138–163.

Crick, F. H. C. (1970). Central Dogma of Molecular Biology. *Nature*, **227**, 561–563.

Cui, X., Qiu, J., Blades, N. J. and Churchill, G. A. (2005). Improved statistical tests for differentail gene expression by shrinking variance components estimates. *Biostatistics*, **6**, 59–75.

Delmar, P., Robin, S., Roux, D. T-L. and Daudin, J. J. (2005). Mixture model on the variance for the differential analysis of gene expression data. *Applied Statistics*, **54**, 31–50.

DeRisi, J. L., Iyer, V. R. and Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.

DeRisi, J. L., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y., Su, Y. A. and Trent, J. M. (1996). Use of cDNA microarray to analyse

gene expression patterns in human cancer. *Nature Genetics*, **14**, 457–460.

Dolby, G. R. (1976). The ultrastructural relation: a synthesis of the functional and structural relations. *Biometrika*, **63**, 39–50.

Dudley, A. M., Aach, J., Steffen, M. A. and Church, G. M. (2002). Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 7554–7559.

Dudoit, S., Fridlynd, J. and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, **2002**, 77–87.

Dudoit, S., Yang, Y. H., Callow, M. J. and Speed, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, **12**, 111–139.

Durbin, B., Hardin, J., Hawkins, D. M., and Rocke, D. M. (2002). A variance stabilizing transformation for gene-expression microarray data. *Bioinformatics*, **18**, 105S–110S.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties (with discussion). *Statistical Science*, **11**, 89–121.

Eisen, M. B. (1999). *ScanAlyse*. Available at `http://rana.Standford.EDU/software/`.

Ekstrøm, C. T., Bak, S., Kristensen, C. and Rudemo, M. (2004). Spot shape modelling and data transformations for microarrays. *Bioinformatics*, **20**, 2270–2278.

Fan, J., Peng, H. and Huang, T. (2005). Semilinear high-dimensional model for normalisation of microarray data: a theoretical analysis and partial consistency (with discussion). *Journal of the American Statistical Association*, **100**, 781–813.

Ferguson, T. S. (1978). Maximum likelihood estimates of the parameters of the Cauchy distribution for samples of size 3 amd 4. *Journal of the American Statistical Association*, **73**, 211–213.

Florens, J. P., Mouchart, M. and Richard, J. F. (1974). Bayesian inference in error-in-variables models. *Journal of Multivariate Analysis*, **4**, 419–452.

Gabrielsen, G. (1982). On the unimodality of the likelihood for the Cauchy distribution: some comments. *Biometrika*, **69**, 677–678.

Garcia, J. N. and Wolkenhauer, O. (2001). Dynamic modelling of microarray time course data. Available at `http://www.umist.ac.uk/csc/people/wolkenhauer.htm`.

García de la Nava, J., van Hijum, S. A. F. T. and Trelles, O. (2004). Saturation and quantization reduction in microarray experiments using two scans at different sensitivities. *Statistical Applications in Genetics and Molecular Biology*, **3**, Article 11.

Gentleman, R., Garey, V. J., Huber, W., Irizarry, R. A. and Dudoit, S. (2005).

Bioinformatics and Computational Biology Solutions Using R and Bioconductor. New York: Springer.

Glasbey, C. A., Forster, T. and Ghazal, P. (2006). Estimation of expression levels in spotted microarrays with saturated pixels. *Submitted.*

Glasbey, C. A. and Ghazal, P. (2003). Combinatorial image analysis of DNA microarray features. *Bioinformatics*, **19**, 194–203.

Glonek, G. F. V. and Solomon, P. J. (2004). Factorial and time course designs for cDNA microarray experiments. *Biostatistics*, **5**, 89–112.

Gottardo, R., Raftery, A. E., Yeung, K. Y. and Bumgarner R. E. (2006). Bayesian robust inference for differential gene expression in microarrays with multiple samples. *Biometrics*, **62**, 10–18.

GSI Luminomics (1999). *QuantArray Analysis Software, Operator's Manual.*

Haas, G., Bain, L. and Antle, C. (1970). Inferences for the Cauchy distribution based on maximum likelihood estimators. *Biometrika*, **57**, p. 403.

Hardiman, G. (2002). Microarray technologies–an overview. *Pharmacogenomics*, **3**, 293–297.

Holder, D., Raubertas, R. F., Pikounis, V. B., Sventik, V. and Soper, K. (2001). Statistical analysis of high density oligonucleotide arrays: a SAFER approach. *GeneLogic Workshop on Low Level Analysis of Affymetrix GeneChip Data.*

Hong, F. and Li, H. (2006). Functional hierarchical models for identifying genes with different time-course expression profiles. *Biometrics*, **62**, 534–544.

Huang, J., Wang, D. and Zhang, C.-H. (2005). A two-way semilinear model for normalization and analysis of cDNA microarray data. *Journal of the American Statistical Association*, **100**, 814–829.

Huber, P. J. (1981). *Robust Statistics.* New York: Wiley.

Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A. and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18**, S96–S104.

Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A. and Vingron, M. (2003). Parameter estimation for the calibration and variance stabilization of microarray data. *Statistical Applications in Genetics and Molecular Biology*, Volume 2, No. 1, Article 3.

Ideker, T., Thorsson, V., Siegel, A. F. and Hood, L. E. (2000). Testing for differentially expressed genes by maximum-likelihood analysis of microarray data. *Journal of Computational Biology*, **7**, 805–818.

Ionides, E. L. (2005). Maximum smoothed likelihood estimation. *Statistica Sinica*, **15**, 1003–1014.

Jin, W., Riley, R. M., Wolfinger, R. D., White, K. P., Passador-Gurgel, G., Gibson,

G. (2001). The contributions of sex, genotype and age to transcriptional variance in drosophila melanogaster. *Nature Genetics*, **29**, 389–395.

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1994). *Continuous Univariate Distributions*, Volume 1. New York: Wiley.

Kendall, M. G. and Stuart, A. (1961). *The Advanced Theory of Statistics*, Volume 2. London: Griffin.

Kerr, M. K., Leiter, P. and Churchill, G. A. (2001). Analysis of a designed microarray experiments. In *Proceedings of the IEEE-Eurasip Nonlinear Signal and Image Processing Workshop*, June 3–6.

Kerr, M. K. and Churchill, G. A. (2001). Experimental design for gene expression microarrays. *Biostatistics*, **2**, 183–201.

Khondoker, M. R., Glasbey, C. A. and Worton, B. J. (2006a). Statistical estimation of gene expression using multiple laser scans of microarrays. *Bioinformatics*, **22**, 215–219.

Khondoker, M. R., Glasbey, C. A. and Worton, B. J. (2006b). A comparison of parametric and nonparametric methods for normalising cDNA microarray data. *Submitted.*

Kim, S. Y., Imoto S. and Miyano, S. (2003). Inferring gene networks from time series microarray data using dynamic Bayesian networks. *Briefings in Bioinformatics*, **4**, 228–235.

Koutrouvelis, I. A. (1980). Regression-type estimation of the parameters of stable laws. *Journal of the American Statistical Association*, **75**, 918–928.

Koutrouvelis, I. A. (1982). Estimation of location and scale in Cauchy distributions using the empirical characteristic function. *Biometrika*, **69**, 205–213.

Lewin, A., Richardson, S., Marshall, C., Glazier, A. and Aitman, T. (2006). Bayesian modeling of differential gene expression. *Biometrics*, **62**, 1–9.

Lindley, D. V. and El-Sayyad, G. M. (1968). The Bayesian estimation of a linear functional relationship. *Journal of the Royal Statistical Society, Series B*, **30**, 190–202.

Lipshutz, R. J., Fodor, S. P. A., Gingeras, T. R. and Lockhart, D. J. (1999). High-density synthetic oligonucleotide arrays. *Nature Genetics*, supplement **21**, 20–24.

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

Ma, S., Kosorok, M. R., Huang, J., Xie, H., Manzella, L. and Soares, M. B. (2006). Robust semiparametric microarray normalization and significance analysis. *Biometrics*, **62**, 555–561.

Madansky, A. (1959). The fitting of straight lines when both variables are subject to error. *Journal of the American Statistical Association*, **14**, 1675–1680.

Mäkeläinen, T., Schmidt, C. and Styan, G. P. H. (1981). On the existence and uniqueness of the maximum likelihood estimate of a vector-valued parameter in fixed size samples. *The Annals of Statistics*, **9**, 758–767.

Mallick, B. K., Ghosh, D. and Gosh, M. (2005). Bayesian classification of tumours by using gene expression data. *Journal of the Royal Statistical Society, Series B*, **67**, 219–234.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979). *Multivariate Analysis.* London: Academic Press.

Mardia, K. V., Southworth, H. R. and Taylor, C. C. (1999). On bias in maximum likelihood estimators. *Journal of Statistical Planning and Inference*, **76**, 31–39.

More, J. J., Garbow, B. S. and Hillstrom, K. E. (1980). *User guide for MINPACK-1*, Argonne National Labs Report ANL-80-74, Argonne, Illinois.

Morton, R. (1981). Estimating equations for an ultrastructural relationship. *Biometrika*, **68**, 735–737.

Müller, P., Parmigiani, G., Robert, C. and Rousseau, J. (2004). Optimal sample size for multiple testing: the case of gene expression microarrays. *Journal of the American Statistical Association*, **99**, 990–1001.

Munson, P. (2001). A "consistency" test for determining the significance of gene-expression changes on replicate samples and two convenient variance stabilizing transformations. *GeneLogic Workshop on Low Level Analysis of Affymetrix GeneChip Data.*

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *Computer Journal*, **7**, 308–313.

Neyman, J. and Scott, E. L. (1951). On certain methods of estimating linear structural relation. *Annals of Mathematical Statistics*, **22**, 352–361.

Parmigiani, G., Garrett, E. S., Anbazhagan, R. and Gabrielson, E. (2002). A statistical framework for expression-based molecular classification in cancer. *Journal of the Royal Statistical Society, Series B*, **64**, 717–736.

Parmigiani, G., Garrett, E. S., Irizarry, R. A. and Zeger, S. L. (2003). *The Analysis of Gene Expression Data: Methods and Software.* New York: Springer-Verlag.

Perou, C. M., Jeffrey, S. S., van de Rijn, M., Rees, C. A., Eisen, M. B., Ross, D. T., Pergamenschikov, A., Williams, C. F., Zhu, S. X., Lee, J. C. F., Lashkari, D., Shalon, D., Brown, P. O. and Botstein, D. (1999). Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proceedings of the National Academy of Sciences*, **96**, 9212–9217.

Qin, L-X. and Self, S. G. (2006). The clustering of regression models method with

applications in gene expression data. *Biometrics*, **62**, 526–533.

Reeds, J. A. (1985). Asymptotic number of roots of Cauchy location likelihood equation. *The Annals of Statistics*, **13**, 775–784.

Reilly, C., Wang, C. and Rutherford, M. (2003). A method for normalising microarrays using genes that are not differentially expressed. *Journal of the American Statistical Association*, **98**, 868–878.

Reilly, P. M. and Patino-Leal, H. (1981). A Bayesian study of the error-in-variables model. *Technometrics*, **23**, 221–231.

Rigby, R. A. and Stasinopoulos, D. M. (1996). Semiparametric additive model for variance heterogeneity. *Statistics and Computing*, **6**, 57–67.

Rigby, R. A. and Stasinopoulos, D. M. (2004). Smooth centile curves for skew and kurtotic data modelling using the Box-Cox power exponential distribution. *Statistics in Medicine*, **23**, 3053–3076.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape (with discussion). *Applied Statistics*, **54**, 507–554.

Rocke, D. M. and Durbin, B. (2001). A model for measurement error for gene expression arrays. *Journal of Computational Biology*, **8**, 557–569.

Rocke, D. M. and Durbin, B. (2003). Approximate variance-stabilizing transformations for gene expression microarray data. *Bioinformatics*, **19**, 966–972.

Romualdi, C., Trevisan, S., Celegato, B., Costa, G. and Lanfranchi, G. (2003). Improved detection of differentially expressed genes in microarray experiments through multiple scanning and image integration. *Nuclic Acids Research*, **31**, e149.

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection.* New York: Wiley.

Schena, M. (2000). *Microarray Biochip Technology.* Westborough, MA: BioTechniques Press.

Schena, M., Shalon, D., Davis, R. and Brown, P. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461–464.

Speed, T. P. (2003). *Statistical Analysis of Gene Expression Microarray Data.* Florida: Chapman and Hall/CRC.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, **9**, 3273–3297.

Southern, E. M. (2001). DNA microarrays: history and overview. *Methods in Molecular Biology*, **170**, 1–15.

Sprent, P. (1966). A generalized least-squares approach to linear functional relationships. *Journal of the Royal Statistical Society, Series B*, **28**, 278–297.

Sprent, P. (1976). Modified likelihood equation of a linear relationship. In *Studies in Probability and Statistics*, ed. E. J. Williams, pp. 109–120. Amsterdam: North-Holland.

Stasinopoulos, D. M., Rigby, R. A. and Akantziliotou, C. (2004). Instructions on how to use the GAMLSS package in R. *Technical Report 02/04*. STORM Research Centre, London Metropolitan University, London.

Stasinopoulos, D. M., Rigby, R. A. and Fahrmeir, L. (2000). Modelling rental data using mean and dispersion additive models. *Statistician*, **49**, 479–493.

Stigler, S. M. (1974). Studies in the history of probability and statistics. XXXIII. Cauchy and the witch of Agnesi: An historical note on the Cauchy distribution. *Biometrika*, **61**, 375–379.

Storey, J. D. and Tibshirani, R. (2001). Estimating false discovery rates under depencence, with applications to DNA microarrays. *Technical Report of the Stanford University, Department of Statistics.*

Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions.* Chichester: Wiley.

Tusher, V. G., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 5116–5121.

Wei, P. (2006). Incorporating gene functional annotations in detecting differential gene expression. *Applied Statistics*, **55**, 301–316.

Wierling, C. K., Steinfath, M., Elge, T., Schulze-Kremer, S., Aanstad, P., Clark, M., Lehrach, H. and Herwig, R. (2002). Simulation of DNA array hybridization experiments and evaluation of critical parameters during subsequent image and data analysis. *BMC Bioinformatics*, **3**, 29.

Wit, E. (2004). Using multiple scans to improve gene expression estimates. *From personal communication with the author.*

Wit, E. and McClure, J. (2003). Statistical adjustment of signal censoring in gene expression experiments. *Bioinformatics*, **19**, 1055–1060.

Wit, E. and McClure, J. (2004). *Statistics for Microarrays: Design, Analysis and Inference*, Chichester: Wiley.

Wit, E., Nobile, A. and Khanin, R. (2004). Simulated annealing near-optimal dual-channel microarray designs. *Technical Report 04–8.* Department of Statistics, University of Glasgow, UK.

Yang, Y., Buckley, M., Dudoit, S. and Speed, T. P. (2002). Comparison of methods for image analysis on cDNA microarray data. *Journal of Computational and Graphical Statistics*, **11**, 108–136.

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

Yang, Y. H. and Speed, T. P. (2002). Design issues for cDNA microarray experiments. *Nature Genetics Reviews*, **3**, 579–588.

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

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics.* New York: Wiley.

Zhou, C. and Wakefield, J. (2006). A Bayesian mixture model for partitioning gene expression data. *Biometrics*, **62**, 515–525.