2009

Mixture distributions with application to microarray data analysis

O'Neil Lynch
University of South Florida

Mixture Distributions with Application to Microarray Data Analysis

by

O'Neil Lynch

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Mathematics and Statistics
College of Arts and Sciences
University of South Florida

Co-Major Professor: Kandethody Ramachandran, Ph.D.
Co-Major Professor: Wonkuk Kim, Ph.D.
Chris Tsokos, Ph.D.
Tapas Das, Ph.D.

Date of Approval:
September 4, 2008

Keywords: Likelihood ratio test; Modified likelihood, Penalized likelihood;
Asymptotic chi-square distribution; Consistency

Dedication

To my wife Lonnette

Acknowledgements

It is a great pleasure for me to express my profound respect, gratitude, and admiration for my major supervisor Dr. K. Ramachandran and co-major supervisor Dr. W. Kim. They have given unstintingly of their time, advise, and encouragement ever since they both undertook the task to make me further understand the fields of microarray data analysis and mixture distribution.

I also would like to thank the other members of my committee namely Dr. C. Tsokos, and Dr. T. K. Das who all have given of their time and have always encouraged me to do my best. The staff at the Department of Mathematics and Statistics is deserving of my thanks for the help and advice they have given me over the years.

I owe thanks to my family and a large number of friends for the encouragement and complete understanding in my quest to excel in the field of Statistics.

Table of Contents

List of Figures

List of Tables

Mixture Distributions with Application to Microarray Data Analysis

O'Neil Lee Lynch

ABSTRACT

The main goal in analyzing microarray data is to determine the genes that are differentially expressed across two types of tissue samples or samples obtained under two experimental conditions. In this dissertation we proposed two methods to determine differentially expressed genes. For the penalized normal mixture model (PMMM) to determine genes that are differentially expressed, we penalized both the variance and the mixing proportion parameters simultaneously. The variance parameter was penalized so that the log-likelihood will be bounded, while the mixing proportion parameter was penalized so that its estimates are not on the boundary of its parametric space. The null distribution of the likelihood ratio test statistic (LRTS) was simulated so that we could perform a hypothesis test for the number of components of the penalized normal mixture model. In addition to simulating the null distribution of the LRTS for the penalized normal mixture model, we showed that the maximum likelihood estimates were asymptotically normal, which is a first step that is necessary to prove the asymptotic null distribution of the LRTS. This result is a significant contribution to field of normal mixture model.

The modified p-value approach for detecting differentially expressed genes was also discussed in this dissertation. The modified p-value approach was implemented so that a hypothesis test for the number of components can be conducted by using the modified likelihood ratio test. In the modified p-value approach we penalized the mixing proportion so that the estimates of the mixing proportion are not on the boundary of its

parametric space. The null distribution of the (LRTS) was simulated so that the number of components of the uniform beta mixture model can be determined. Finally, for both modified methods, the penalized normal mixture model and the modified p-value approach were applied to simulated and real data.

1 Introduction

In recent years microarray technology has made it possible to simultaneously analyze thousands of genes. Although an enormous volume of data is being produced by microarray technologies (Schena et al., 1995; DeRisi et al., 1997; Hughes et al., 2001; Lockhart et al., 1996), one remaining challenge is how to analyze and interpret the large amounts of data. A major challenge is to detect genes with differentially expressed profiles under two different experimental conditions, which may refer to samples drawn from two types of tissues, tumors or cell lines, or at two time points during important biological processes.

Many of the methods used for such analysis, including the method of identifying genes with fold changes are known to be unreliable because in such methods the statistical variability of the data is not properly addressed [8]. While various parametric methods and tests such as the two-sample t-test [12] and regression model have been applied for microarray data analysis, strong parametric assumptions made in these methods as well as strong dependency on large sample sets restrict the reliability of such techniques in microarray problems where only a small number of replications are available. The non parametric statistical methods, including the Empirical Bayes (EB) method [14], the significance analysis for microarray data (SAM [39]) and mixture model method (MMM) [27, 42, 25] have been applied to microarray data analysis. It is claimed and argued that the new extensions of the (MMM) are among the available methods producing biologically-meaningful results [27, 43].

In this dissertation we extended the mixture model method (MMM) by penalizing the mixing proportions and the component variances simultaneously. The mixing proportion was penalized so that a modified likelihood ratio test similar to that of Chen et al. (2001, 2004) for testing the number of components of the fitted normal mixture model can be implemented. The variance was penalized so that the log-likelihood is bounded resulting in the existence of the MLE's. In a similar fashion the p-value approach (Allison et al. (2002)) for the detection of differentially expressed genes of microarray data was also modified. For the p-value approach only the mixing proportion was modified so that the MLE of the mixing proportion was not on the boundary of its parametric space. This modification was done so that a modified likelihood ratio test similarly to what was done by Chen et al. may be implemented so that we may test the hypothesis for the number of components.

This dissertation is organized as follows. Chapter 2 describes in some detail the genetic background of DNA and two of the leading microarray experiments, cDNA and Oligonucleotide. In Chapter 2 we also discussed some of the statistical challenges we have in analyzing microarray data and gives a description of some of the methods used to analyze microarray data. The methods that were discussed are (1) Cluster analysis (2) T-test (3) Regression analysis (4) Significant analysis of microarray (SAM) (5) Mixture model method (MMM) and (6) A p-value approach for detecting differentially expressed microarray data.

In Chapter 3 we present the theory of finite mixture methods and discussed how the parameters can be estimated by (1) expectation maximization algorithm (EM) and (2) the robust parameter estimation - which is of interest if the data contains outliers. One of the challenges of finite mixture distributions is to determine the number of components therefore we discussed some techniques used to determine the number of components which are namely AIC, BIC, simulation and the modified likelihood ratio test. The box-cox transformation for distinquishing skewed distributions from commingled distributions was also presented in chapter 3.

The penalized modified approach will be discussed in chapter 4. The estimators of the parameters of the penalized normal mixture model when both the variance and mixing proportion were simultaneously penalized was illustrated. The evaluation of the estimators for the two penalty functions for the variance, the inverse gamma and inverse chi-square distributions were addressed. The asymptotic property namely asymptotic normality of the normal mixture model was also proved in Chapter 4. Chapter 5 discussed the applications of the penalized/modified approach of the normal mixture model to detecting differentially expressed genes and illustrated its applications to simulated and real data. The results of the penalized/modified normal mixture model approach were compared to that of SAM and was shown to out perform SAM.

Chapter 6 discussed the modified p-value approach for detecting differentially expressed genes. Similar to the work done in Chapter 6 we applied our method to simulated and real data. The motivation for modifying the p-value approach of Allison et al. was that the MLE of the mixing proportion was on the boundary point of its parametric space, therefore we applied the technique of Chen et al., that is, we applied a penalty function for the mixing proportion so that the MLE of the mixing proportion will not be on the boundary points of its parametric space. The conclusions of this study were summarized in Chapter 7.

2 Microarray Data and Some Statistical Analysis

2.1 DNA Microarray Experiments

2.1.1 Genetic Background

The double-stranded molecules deoxyribonucleic acid (DNA) (Watson and Crick, 1953) contains all the genetic information of living organisms. Each strand or helix of DNA is a chain of nucleotides that consists of a sugar, a phosphate and a nitrogenous base molecule. The information in DNA is stored as a code made up of four chemical bases: Adenine (A), Guanine (G), Cytosine (C) and Thymine (T). These four bases are responsible for the DNA molecule having four distinct types of nucleotides. The bases are coupled in the following manner: A with T and C with G, by a hydrogen bond which is called complementary base pairing. The nucleotides are arranged in two long strands that form a spiral called a double helix. The double helix structure of DNA is similar to a ladder, with the base pairs forming the ladder's rungs and the sugar and phosphate molecules forming the vertical sidepieces of the ladder.

In cells, genes consist of a long strand of DNA that contains a promoter, which controls the activity of a gene. Additionally, all living cells contain chromosomes, that are, large pieces of genes containing hundreds or thousands of genes, each of which specifies the composition and structure of a protein (or several related proteins). The workhorse molecules of the cell are protein polymers of amino acids which are responsible for cellular structure, producing energy and important biomolecules like DNA and proteins, and for reproducing the cell chromosomes. The cohort of chromosomes are almost the same in every cell in an organism, and contains the same repertoire of proteins. However, cells have remarkably distinct properties, such as the difference between human eye cells, hair cells, and liver cells; these distinctions are the result of differences in the abundance, distribution, and state of the cell proteins.

When a gene is active, the coding and non-coding sequence is copied in a process called transcription, producing messenger RNA (mRNA) which is a copy of the gene's information. The mRNA, a small and relatively unstable nucleic acid polymers, can then direct the synthesis of proteins through the genetic code. However, mRNAs can also be used directly, for example as part of the ribosome. The resulting molecules from the gene expression, mRNA or protein are known as gene products. There is therefore a logical connection between the state of a cell and the details of its protein and mRNA composition.

Whereas it remains difficult to measure the abundance of a cell's proteins, DNA microarray makes it possible to quickly and efficiently measure the relative representation of each mRNA species in the total cellular mRNA population, or in more familiar terms, to measure gene expression levels. There are several types of microarray systems including the cDNA microarray (Schena et al., 1995; DeRisi et al., 1997: Hughes et al., 2001) and oligonucleotide array (Lockhart et al., 1996).

2.1.2 cDNA Microarray Experiment

In this experiment, the cDNA sequence corresponding to a set of genes pertinent to the biological question under investigation are obtained and printed onto a glass slide or substrate using a robotic arrayer. Second, the sample RNA is isolated, a critical step in the experiment in order to ensure that a sufficient amount of each cDNA clone is printed on the array where each clone is amplified by a technique called polymerase chain reaction (PCR). In practice the printed amount of cDNA is not the same, therefore the cDNA on the array, which is a double-stranded probe, needs to be denatured and this is achieved by heating the array so that a target cDNA can bind to it.

In the third step the cDNA is synthesized, a procedure that also involves labeling the isolated mRNA from the biological samples. Usually in the most current cDNA microarray experiments, cDNAs from the experimental and reference samples are labeled with red-fluorescent dye, Cy5 and green-fluorescent dye, Cy3 respectively, mixed and hybridized on the slide. There are several different labeling methods including Primer

Tagging, Direct Incorporation Labeling and Amino-Modified Nucleotide. Nguyen et al. (2002), Wong et al. (2001) and Stears et al. (2000) discuss the advantages and disadvantages of these methods.

Fourth, the labeled probe cDNA is hybridized to target the cDNA on the microarray. That is, if a particular gene is expressed in the target cell, where the cDNAs corresponding to this gene are found in the target cDNA pool, these cDNAs will bind with the complementary cDNA probes printed on a specific spot on the microarray. Hybridization refers to the binding ability of two complementary DNA strands by the base-pairing rule thus reforming the DNA double helix.

Finally, the hybridization results are imaged and analyzed using a fluorescent microscope, the log(red/green) intensities of mRNA hybridization at each site is measured. The result is tens of thousands of gene expressions, typically ranging from -4 to 4, which is a measure of the expression level of each gene in the experimental sample relative to the reference sample. Positive values indicate higher expression in the target versus the reference, and vice versa for negative values.

2.1.3 Oligonucleotide Microarray Experiment

Another widely used microarray technology is high density oligonucleotide arrays known as Affymetrix (Lockhart et al., 1996). This method is based on the fact that each gene is represented by 14 to 20 features (Lipshutz et al., 1999). for example, Affymetrix array used 20 features. Each feature is a short sequence of nucleotides, an oligonucleotide, and it is a perfect match (PM) to a segment of the gene. Paired with the 20 PM oligonucleotides to the gene sequence are 20 other oligonucleotides having the same sequence corresponding to the 20 PMs except for a single mismatch (MM) at the central base of the nucleotide. When the gene is expressed in the cell sample, high intensity is expected for the PM feature and low intensity for the MM feature. Given the 20 PM and MM feature pairs for the gene, many methods have been proposed to quantify the expression level of the gene. For example, Affymetrix originally proposed the average difference $x = avg\{d_k = (PM_k - MM_k), k = 1, 2, \ldots, 20 = K\}$ to quantify expression level of a gene in a particular array. The average is usually based only on the differences, d_k, with

3 standard deviations from the mean of $d_{(2)}, \ldots, d_{(K-1)}$, where $d_{(k)}$ is the k^{th} smallest difference, but there are various other ways to filter the outliers, Efron et al. (2001) suggested $x = avg\{d_k = \log(PM_k) - c\log(MM_k), k = 1, 2, \ldots, 20\}$ for several different scale factors c. Naef et al. (2001) proposed to use only the PM features. In an attempt to obtain more sensitive measure of gene expression, Li and Wong (2001) proposed a model-based estimate of the expression level using the least square method. The method for sample labeling and image processing in the Affymetrix arrays are found in Lockart et al. (1996). Refer to "The Chipping Forecast" (Lander et al., 1999) for more details on cDNA microarrays and oligonucleotide chips.

2.2 Some Statistical Challenges With Analyzing Microarray Data

Microarray technologies allow scientists to monitor the mRNA transcript levels of thousands of genes in a single experiment. However, the tremendous amount of data that is obtained from microarray studies presents challenges for data analysis. One challenge in the development of statistical methods for microarray data analysis is that sample sizes under two different experimental conditions are typically small. We can depict this situation by defining the data as follows: for each gene i, $i = 1, 2, \ldots, N$, we have expression levels $(Y_{i1}, \ldots, Y_{im})$ from m microarrays under condition 1, and $(Y_{i,m+1}, \ldots, Y_{i,m+n})$ from n arrays under condition 2. Usually the total number of genes N is large (> 1000) whereas the number of replications, m and n are small (typically < 20).

Since statisticians are primarily interested in genes that are differentially expressed across two different experimental conditions, which may refer to samples drawn from two types of tissues, tumors or cell lines, or at two time points during important biological processes, we need to make an adjustment for the type I error rate when doing simultaneous hypothesis tests. This adjustment is done by means of the Bonferroni method, to deal with multiple comparisons. If we use α as the significance level then the test or gene specific significance level for a two sided test is therefore $\alpha^* = \alpha/2n$.

Investigators may need to have the answer for the following question "Is the difference in expression level for a particular gene statistically significant?" However, there are a number of equally important questions that need to be answered (Allison et al. (2001)):

1. Is there statistically significant evidence that any of the genes under study exhibit a difference in expression across the conditions?

2. What is the best estimate of the number of genes for which there is a true difference in gene expression?

3. What is the confidence interval around that particular estimate?

4. If we set some threshold for which we expect particular genes to be *interesting* and worthy of follow-up study, what proportion of those genes are likely to be genes for which there is a real difference in expression and what proportion are likely to be false leads?

5. What proportion of those genes not declared *"interesting"* are likely to be genes for which there is a real difference in expression (i.e., misses or false negatives)?

In analyzing microarray data the assumptions made are (1) For each gene, the measurements of gene expression have a finite population mean and variance; (2) For each gene under study, there is a measure of expression level available for each sample and this measure has sufficient reliability and validity (i.e. the measurements of the expression levels are a true reflection of the true state of nature); (3) The most important assumption that is made is that gene expression levels across the two groups are independent - which implies that we may able to evaluate the likelihood function which will become important later in this dissertation.

2.3 Methods of Analyzing Microarray Data

2.3.1 Cluster Analysis

One method used in the analysis of microarray data is Cluster analysis. Cluster analysis groups genes or samples into "clusters" based on similar expression profiles and provides clues to the function or regulation of genes or similarity of samples via shared cluster membership [34, 35, 18]. Several clustering methods have been usefully applied to analyzing genome-wide expression data and can be classified largely into three categories. The three-based approach uses distance measures between genes such as correlation coefficients to group genes into a hierarchical tree [15]. The second category clusters genes so that within-cluster variation is minimized and between-cluster variation is maximized [34, 35]. The third category group genes into blocks, in which the correlation is maximized and between which the correlation is minimized [3]. The power of cluster analysis in the analysis of microarray data lies in discovering gene transcripts or samples that show similar expression profiles. However, identification of ”like” groups is not necessarily the objective in a microarray study, because the interest is to discover genes that are differentially expressed between predefined sample groups, such as normal versus cancerous tissues.

Data

Let Y_{ik} be the expression level of gene i in array k $(i = 1, \ldots, N; k = 1, \ldots, m, m + 1, \ldots, m + n)$. Suppose that the first m and the last n arrays are both obtained under two different conditions, that is $Y_{i(1)} = (Y_{i1}, \ldots, Y_{im})$ and $Y_{i(2)} = (Y_{i,m+1}, \ldots, Y_{i,m+n})$. Since we are interested to determine which genes are differentially express between $Y_{i(1)}$ and $Y_{i(2)}$, we let

$$Y_{ik} = a_i + b_i x_k, \tag{2.3.1}$$

where

$$x_k = \begin{cases} 1 & \text{for } 1 \leq k \leq m \\ 0 & \text{for } m+1 \leq k \leq m+n. \end{cases}$$

Therefore the mean expression levels of gene i under the two conditions are $a_i + b_i$ and a_i respectively. Hence to determine the genes that are differentially expressed is equivalent to testing the hypothesis

$$\begin{aligned} H_0 &: b_i = 0, \text{ there is no gene with altered expression} \\ H_1 &: b_i \neq 0, \text{ otherwise} \end{aligned} \tag{2.3.2}$$

Using the data construction of equation (2.3.1) for Y_{ik} we will now present the t-test and regression models used in microarray data analysis.

2.3.2 The T-Test

There are several versions of the two-sample t-test, depending on whether the sample size (i.e m and n) is large and whether it is reasonable to assume that the gene expression levels have an equal variance under the two conditions. Both m and n are usually small, and there is evidence to support unequal variance (Thomas et. al. 2001), we will only discus the t-test with two independent small Normal samples with unequal variances.

Let the sample means and variances of Y_{ik} for gene i under the two conditions be

$$\bar{Y}_{i(1)} = \frac{\sum_{k=1}^{m} Y_{ik}}{m},\ \bar{Y}_{i(2)} = \frac{\sum_{k=m+1}^{m+n} Y_{ik}}{n} \tag{2.3.3}$$

and

$$\begin{aligned} s^2_{i(1)} &= \frac{\sum_{k=1}^{m}(Y_{ik} - \bar{Y}_{i(1)})^2}{m-1}, \\ s^2_{i(2)} &= \frac{\sum_{k=m+1}^{m+n}(Y_{ik} - \bar{Y}_{i(2)})^2}{n-1}. \end{aligned} \tag{2.3.4}$$

The t-statistic is

$$Z_i = \frac{\bar{Y}_{i(1)} - \bar{Y}_{i(2)}}{\sqrt{\frac{s^2_{i(1)}}{m} + \frac{s^2_{i(2)}}{n}}}, \tag{2.3.5}$$

Under the normality assumption for Y_{ik}, Z_i approximately has a t-distribution with degrees of freedom

$$d_j = \frac{\left(\frac{s^2_{i(1)}}{m} + \frac{s^2_{i(2)}}{n}\right)^2}{\frac{\left(\frac{s^2_{i(1)}}{m}\right)^2}{m-1} + \frac{\left(\frac{s^2_{i(2)}}{n}\right)^2}{n-1}}$$

This t-test was proposed by Welch (1947). Its method of calculating the degrees of freedom is similar to the idea of the Satterthwaite approximation.

2.3.3 Regression Model

The regression model estimates the values of (a_i, b_i) using the weighted least square method, and then estimates the variance of $\hat{b}_i$ using the robust or sandwich variance estimator.

$$Var(\hat{b}_i) = \Big(\frac{s^2_{i(1)}}{m}\Big)\Big(\frac{m-1}{m}\Big) + \Big(\frac{s^2_{i(2)}}{n}\Big)\Big(\frac{n-1}{n}\Big),$$

and the estimate of $\hat{b}_i = \bar{Y}_{i(1)} - \bar{Y}_{i(2)}$. Therefore the test statistics is

$$Z'_i = \frac{\hat{b}_i}{Var(\hat{b}_i)} = \frac{\bar{Y}_{i(1)} - \bar{Y}_{i(2)}}{\sqrt{\frac{s^2_{i(1)}}{m}\frac{m-1}{m} + \frac{s^2_{i(2)}}{n}\frac{n-1}{n}}}. \tag{2.3.6}$$

This test statistic compares well with that of the t-test. In the case of the t-test the test statistic is

$$Z_i = \frac{\bar{Y}_{i(1)} - \bar{Y}_{i(2)}}{\sqrt{\frac{s^2_{i(1)}}{m} + \frac{s^2_{i(2)}}{n}}}, \tag{2.3.7}$$

where $\bar{Y}_{i(1)}, \bar{Y}_{i(2)}, s^2_{i(1)}$ and $s^2_{i(2)}$ are defined as in (2.3.3) and (2.3.4). Note that the two tests are the same as $m, n \rightarrow \infty$, however in microarray data analysis both m, n are small, which makes the t-test better because of the unbiasedness of its variance estimator.

Note that the strong parametric assumptions that needs to be made to use both the t-test and the regression approach is often times violated for microarray data analysis. Therefore, the Significance Analysis of Microarrays (SAM) is an important method developed for microarray data analysis that seeks to over theses strong parametric assumptions.

2.3.4 Significance Analysis of Microarrays (SAM)

The significance analysis of microarrays (SAM) is one statistical technique for finding significant genes in a set of micoarray experiments. It was proposed by Tusher, Tibshirani and Chu [39]. This approach was based on analysis of random fluctuations in the data. However, even for a given level of expression, the fluctuations were gene specific. To account for gene-specific fluctuations, a statistic based on the ratio of change in gene expression to standard deviation in the data for that gene was defined. The ”relative difference” $d(i)$ in the gene expression is:

$$d(i) = \frac{\bar{Y}_{i(1)} - \bar{Y}_{i(2)}}{s(i) + s_0} \tag{2.3.8}$$

where $\bar{Y}_{i(1)}$ and $\bar{Y}_{i(2)}$ are defined as the average expression levels of the i^{th} gene from conditions 1 and 2, respectively. The "gene-specific scatter" $s(i)$ is the standard deviation of repeated expression measurements:

$$s(i) = \sqrt{\frac{1/m + 1/n}{m+n-2}\Big(\sum_{k=1}^{m}(Y_{ik} - \bar{Y}_{i(1)})^2 + \sum_{k=m+1}^{m+n}(Y_{ik} - \bar{Y}_{i(1)})^2\Big)} \qquad (2.3.9)$$

where m and n are the numbers of measurements in conditions 1 and 2 respectively.

In order to compare values of $d(i)$ across all genes, the distribution of $d(i)$ should be independent of the level of gene expression. At low expression levels, variance in $d(i)$ can be high because of small values of $s(i)$. To ensure that the variance of $d(i)$ is independent of the gene expression, a small positive constant s_0 (exchangeability factor) was added to the denominator of equation (2.3.8). The coefficient of variation of $d(i)$ was computed as a function of $s(i)$ in moving windows across the data. The value for s_0 was chosen to minimize the coefficient of variation.

To minimize the effects of potential confounders between the conditions, the data was analyzed by taking B sets of permutations. For each permutation b the statistic d_i^{*b} and the corresponding order statistics $d_{(1)}^{*b} \leq d_{(2)}^{*b} \ldots \leq d_{(N)}^{*b}$ was computed. The expected relative difference, $\bar{d}_i = \frac{\sum_b d_i^{*b}}{B}$, was defined as the average over the set of B permutations.

To identify potentially significant changes in expression levels, they used a scatter plot of the observed relative difference $d(i)$ versus the expected relative difference $\bar{d}_i$. For a fixed threshold Δ, starting at the origin, and moving up to the right find the first $i = i_1$ such that $d_i - \bar{d}_i > \Delta$. All genes pass i_1 are called "significant positive". Similarly, start at the origin, move down to the left and find the first $i = i_2$ such that $\bar{d}_i - d_i > \Delta$. All genes pass i_2 are called "significant negative". For each Δ the upper cutoff point $\text{cut}_{up}(\Delta)$ was defined as the smallest d_i among the significant positive genes, and similarly defining the lower cutoff point $\text{cut}_{low}(\Delta)$.

To determine the number of falsely significant genes generated by SAM, the total number of falsely significant genes corresponding to each permutation was computed by counting the number of genes that exceeded the cutoffs $\text{cut}_{up}(\Delta)$ and $\text{cut}_{low}(\Delta)$. The estimated number of falsely significant genes was the median (or 90^{th} percentile) of the number of genes called significant from the B sets of permutations. Such genes are called false positive (FP). This information will then be used to estimate the false Discovery Rate (FDR)

$$FDR = \pi_0 FP/TP \tag{2.3.10}$$

where π_0 is the true proportion of equivalent expressed (EE) genes in the data set and TP is the number of total (true) positives discovered from the test statistic, that is, TP is the total number of genes claimed to be differentially expressed (DE).

2.3.5 Mixture Model Method (MMM)

The mixture model method (MMM) was introduced to handle the problem when a small number of replications under two experimental conditions exist, which is exactly the case for the data in a microarray experiment. The main purpose of the MMM is to estimate the distribution of a t-type test statistic and its null statistic using finite normal mixture models, which results in the method being non-parametric. Additionally, the strong parametric assumption made when analyzing microarray when the traditional statistical test is applied is often violated, hence this make the MMM statistically safer because the assumption of normality is not made.

Consider the situation where, for each gene i, $i = 1, 2, \ldots, N$, we have expression levels $Y_{i(1)} = (Y_{i1}, \ldots, Y_{im})$ from m microarrays under condition 1, and $Y_{i(2)} = (Y_{i,m+1}, \ldots, Y_{i,m+n})$ from n arrays under condition 2. Here we need to assume that both m and n are even integers, this will become obvious later.

The goal is to identify genes such that $(Y_{i1}, \ldots, Y_{im})$ and $(Y_{i,m+1}, \ldots, Y_{i,m+n})$ have different means. This appears to be a two sample comparison however, in microarray data, that has small m and n with a large N, renders the traditional statistical tests such as the t-test or rank-based nonparametric tests, ineffective. One alternative is to draw statistical inference based on the distributions of quantities related to $(Y_{i1}, \ldots, Y_{im})$ or $(Y_{i,m+1}, \ldots, Y_{i,m+n})$, for $1 \leq i \leq N$, to take advantage of the large population size N.

The model assumes a nonparametric approach for gene expression data:

$$Y_{i(1)} = \mu_{i(1)} + \varepsilon_{i(1)} \qquad Y_{i(2)} = \mu_{i(2)} + e_{i(2)}$$

where $\mu_{i(1)}$ and $\mu_{i(2)}$ are the mean expression levels for gene i under the two conditions respectively, and $\varepsilon_{i(1)}$ and $e_{i(2)}$ are independent random errors with means and variances, such that

$$E(\varepsilon_{i(1)}) = E(e_{i(2)}) = 0, \qquad Var(\varepsilon_{i(1)}) = \sigma^2_{i(1)}, \qquad Var(e_{i(2)}) = \sigma^2_{i(2)},$$

for any $j = 1, \ldots, m, m+1, \ldots, m+n$ and $i = 1, \ldots, N$. Note, we do not assume equality of variance of the gene expression levels, because the variance $\sigma^2_{i(c)}$ of gene expression level depends on the mean expression $\mu_{i(c)}$. Also, we do not assume $\mu_{i(1)} = \mu_{i(2)}$.

The basis of the model is to compare two distributions of two similar statistics (after being suitably standardized) to infer whether some genes are differentially expressed. Let m and n be even such that p_i (q_i) is a column vector containing random permutation of $m/2$ 1's and $m/2$ -1's ($n/2$ 1's and $n/2$ -1's). Let $Y_{i(1)} = (Y_{i1}, \ldots, Y_{im})$ and $Y_{i(2)} = (Y_{i,m+1}, \ldots, Y_{i,m+n})$ then assume that

$$z_i = \frac{Y_{i(1)}p_i/m + Y_{i(2)}q_i/n}{\sqrt{s^2_{i(1)}/m + s^2_{i(2)}/n}} \sim f_0, \tag{2.3.11}$$

which does not depend on $\mu_{i(1)}$ and $\mu_{i(2)}$ since its mean is 0. Furthermore, suppose that

$$\begin{aligned} Z_i &= \frac{\sum_{k=1}^{m} Y_{ik}/m - \sum_{k=m+1}^{m+n} Y_{ik}/n}{\sqrt{s^2_{i(1)}/m + s^2_{i(2)}/n}} \\ &= \frac{\bar{Y}_{i(1)} - \bar{Y}_{i(2)}}{\sqrt{s^2_{i(1)}/m + s^2_{i(2)}/n}} \sim f_1. \end{aligned} \tag{2.3.12}$$

The hypothesis is of the form

$$\begin{aligned} H_0 &: f_0 = f_1, \text{ there is no gene with altered expression} \\ H_1 &: f_0 \neq f_1, \text{ otherwise} \end{aligned} \tag{2.3.13}$$

and is valid only if the random errors are independent and their distribution is symmetric about 0. Since $m, n > 1$ then we can estimate $s^2_{i(1)}$ and $s^2_{i(2)}$ using the sample variances $s^2_{i(1)} = \frac{\sum_{k=1}^{m}(Y_{ik} - \bar{Y}_{i(1)})^2}{m-1}$ and $s^2_{i(2)} = \frac{\sum_{k=m+1}^{m+n}(Y_{ik} - \bar{Y}_{i(2)})^2}{n-1}$ respectively. Note the data z_i's and

Z_i's are used to estimate f_0 and f_1 by normal mixture model respectively, which will be discussed in more details in chapter 3.

To test the null hypothesis H_0 that Z is from f_0 (which is equivalent to testing the hypothesis (2.3.13)), we can construct a likelihood ratio test (LRT) based on the following statistic:

$$LR(Z) = \frac{f_0(Z)}{f_1(Z)}. \tag{2.3.14}$$

A large value of $LR(Z)$ gives no evidence against H_0, whereas a too small value of $LR(Z)$ leads to rejecting H_0. With the normal mixture model, it is possible to numerically determine the rejection region. For any given false positive rate α, we can use the bisection method [29] to solve

$$\alpha = \int_{LR(z)<s} f_0(z)dz$$

to obtain the suitable cut off point s. Then the rejection region is $RR(\alpha) = \{Z : LR(Z) < s\}$. We call the method of using the LRT in MMM as MMM-LRT. Similar to SAM (Tusher et al. 2001), we can estimate the numbers of false positive (FP) and total (TP) directly. In MMM-LRT, for any given s, we have:

$$FP(s) = \frac{1}{B}\sum_{b=1}^{B} n(i : LR(z_i^{(b)}) < s), \quad TP(s) = n(i : LR(Z_i) < s)$$

where $n(i)$ represents the number of genes. In estimating FP, one can also use median, rather than mean, FP over the permuted data. Based on the estimated FP and TP, one can also calculate the false discovery rate as $FDR = FP/TP$ (Benjamini and Hochberg 1995; Storey 2001; Tusher et al. 2001).

2.3.6 A Mixture Model Approach Using P-Value

In is well known that the distribution of the p-values is uniformly distributed on the interval $[0, 1]$, regardless of the statistical test used and the sample size. Therefore if investigators uses a valid statistical test to produce p-values for testing the null hypothesis

H_0 there is no difference between the two experiments for the i^{th} gene, $i = 1, \ldots, N$, then the distribution of the p-value can be used to determine the genes that were differentially expressed. The assumption of independence of the gene expression levels across genes was made under the null hypothesis. Additionally, under the alternative hypothesis, the distribution of p-values will tend to cluster closer to zero than to one, as opposed to be uniformly distributed under the null hypothesis. Then, the question "Is there statistically significant evidence that any of the genes under study exhibit a difference in expression across the two experimental conditions?" can be answered by conducting a test to determine if the observed p-values are significantly different from the uniform distribution. This is done by mixture model approach [2].

The mixture model is a g-component of beta distributions $\beta(r_j, s_j)$ for $j = 1, \ldots, g$ with the parameters r_j and s_j, where the beta distribution is defined as follows

$$\beta(y|r, s) = \frac{\Gamma(r+s)y^{r-1}(1-y)^{s-1}}{\Gamma(r)\Gamma(s)}.$$

The reason for the choice of the beta distribution is because of its great flexibility in modeling any shaped distribution on the interval $[0, 1]$. Note that the uniform distribution is a special case of the beta distribution with $r = s = 1$. The likelihood for the collection of N, p-values from a model with g components is given as

$$L_g = \prod_{i=1}^{N} \left[p_1 \beta(y_i|1, 1) \prod_{j=2}^{g} p_j \beta(y_i|r_j, s_j) \right],$$

Therefore the log likelihood for the N p-values can be expressed as

$$l_g = \sum_{i=1}^{N} \ln \left[p_1 \beta(y_i|1, 1) + \sum_{j=2}^{g} p_j \beta(y_i|r_j, s_j) \right],$$

where y_i is the p-value for the i^{th} test, p_1 is the probability that a randomly chosen test from the collection of tests is for a gene where there is no population difference in gene expression (i.e., a test of a true null hypothesis), and p_j is the probability that a randomly chosen test from the collection of tests is for a gene where there is a population difference in gene expression (i.e., a test of a false null hypothesis). The above model now requires the calculation of the MLE of the parameters p_j, r_j and s_j through iterative procedure subject to the constraints $\sum_{j=1}^{g} p_j = 1$ and $0 \leq p_j \leq 1$ for all $j = 1, \ldots, g$.

The estimate of the number of genes for which there is a difference in gene expression is evaluated as $N(1 - \hat{p}_1)$, where $\hat{p}_1$ is the MLE of p_1. Let T be some threshold below which the results for particular genes are declared *"interesting"* and worthy of follow-up study, the proportion of those genes that are likely to be genes for which there is a real difference is

$$P(\bar{H}_{0,i}|y_i \leq T) = 1 - P(H_{0,i}|y_i \leq T) = 1 - \frac{P(H_{0,i}|y_i \leq T)}{P(y_i \leq T)},$$

where

$$P(y_i \leq T) = p_1 T + \sum_{j=2}^{g} p_j \int_0^T \frac{\Gamma(r_j + s_j) y^{r_j - 1}(1-y)^{s_j - 1}}{\Gamma(r_j)\Gamma(s_j)} dy$$

and $P(\bar{H}_{0,i} \cap y_i \leq T) = p_1 T$. The estimated proportion of genes declared interesting that are likely to be false leads is simply

$$P(H_{0,i}|y_i \leq T) = \frac{P(H_{0,i} \cap y_i \leq T)}{P(y_i \leq T)}.$$

Similarly the proportion of those genes not declared *"interesting"* that are likely to be genes for which there is a real difference is

$$P(\bar{H}_{0,i}|y_i \geq T) = 1 - P(H_{0,i}|y_i \geq T) = 1 - \frac{P(H_{0,i}|y_i \geq T)}{P(y_i \geq T)},$$

where

$$P(y_i \geq T) = p_1(1 - T) + \sum_{j=2}^{g} p_j \int_T^1 \frac{\Gamma(r_j + s_j) y^{r_j - 1}(1-y)^{s_j - 1}}{\Gamma(r_j)\Gamma(s_j)} dy$$

and $P(\bar{H}_{0,i} \cap y_i \geq T) = p_1(1 - T)$.

2.4 Conclusion

This chapter discussed a few of the methods used to analyze microarray data. An introduction to cluster analysis was presented, but, cluster analysis was not an effective method to determine differentially express genes. Hence the need to make use of the more classical statistical methods such as the t-test and regression analysis. However, with strong parametric assumptions that will be necessary for microarray analysis, these methods has some limitations. Microarray data are many times consist of a few replications for case and control groups, although the number of genes are usually greater than 1000. The assumption that the genes are independent is one assumption that is typical in the analysis of microarray data. Note that in chapters 3 and 5 the development of the modified approaches use the independence assumption, therefore we are prepared to deal with the consequences of assuming the genes are independent.

The Significance Analysis of Microarrays (SAM) and the Mixture Model Method (MMM) presented in this chapter uses a t-type statistics to determine the number of differentially expressed genes. However, the MMM has one advantage in that it is a non-parametric approach. The MMM determines the distributions under the null and alternative and then uses these distributions to determine the number of differentially expressed genes by means of a likelihood ratio test.

The p-value approach of Allision relies on parametric assumptions that are made to determine the p-values. If the p-values are not valid then its distributions under the null hypothesis may not be uniform on the interval [0,1]. In discussing the modified p-value approach presented in chapter 6, we are aware that the t-test used to determine the p-values must be valid for the modified p-value approach to be valid. However, for this dissertation we assume all the assumptions are satisfied with respect to the modified p-value approach.

In addition, to the method used to analyze microarray data, we presented the biological background that the reader needs so that he may fully understand the challenges statisticians have in the analysis of microarray data.

3 Finite Mixture Distribution

In this chapter we will give a brief background on mixture distributions. Mixture models are vital in statistical practice and research because many problems in statistics have mixture structures. Furthermore they are useful in describing complex population with observed or unobserved heterogeneity. Some examples are that human heights may be modeled as a two-component mixture, one component for men and one for women. Substructures in galaxy may be modeled as contaminations of big initial galaxy; the evidence of substructures is important in modern galaxy formation theory (Sun, Morrison, Harding and Woodroofe 2002). There are also applications in actuarial science, biological science, econometrics, medicine, agriculture, zoology, population studies and microarray data analysis.

K. Pearson (1894) was the first to study mixture of two normal distributions, where he modeled the mixing of different crab species. Mixture model has become popular because: (1) they provide a simple mechanism to incorporate extra variation and correlation in the model (2) they add model flexibility and (3) they are a natural approach for modeling data that arise in multiple stages or when populations are composed of sub populations. In addition the theory, applications, history and importance of mixture models have been discussed in journal articles, monographs and textbooks. Everitt and Hand (1981), Titterington, Smith and Makov (1985), Böhning (1999), and McLachlan and Peel (2000) provided models, statistical methods and references for finite mixtures problems.

3.1 Definition and Preliminary

Definition 3.1.1 *A stochastic variable* $\{Y_i : 1 \leq i \leq n\}$ *with density function* $f(y_i|\theta_j)$ *follows a finite mixture distribution if*

$$\begin{aligned} Y_i &\sim \pi_1 f_{i1}(y_i|\theta_1) + \pi_2 f_{i2}(y_i|\theta_2) + \ldots + \pi_g f_{ig}(y_i|\theta_g) \\ &= \sum_{j=1}^{g} \pi_j f_{ij}(y_i|\theta_j), \end{aligned} \tag{3.1.1}$$

where $f_{i1}(y_i|\theta_1), \ldots, f_{ig}(y_i|\theta_g)$ are g density functions and $\pi_1, \ldots, \pi_g$ are called mixing proportions, satisfying the following properties $0 \leq \pi_j \leq 1$ and $\sum_{j=1}^{g} \pi_j = 1$. The densities $f_{ij}(y)$ for $j = 1, \ldots, g$ may be continuous or discrete, or a combination of both.

From Definition 3.1.1 we observe that finite mixture distribution is the marginal distribution of a random variable which follows different distributions in different sub-populations of a general population. Therefore, if a population S is defined as

$$S = \{S_1, S_2, \ldots, S_g\}, \text{ such that } S_j \cap S_k = \emptyset, \; j \neq k.$$

Then the distribution in each sub-population is given to be

- In $S_1 : Y|S_1 \sim f_1(Y|\theta_1)$
- In $S_2 : Y|S_2 \sim f_2(Y|\theta_2)$
- $\ldots$
- In $S_g : Y|S_g \sim f_g(Y|\theta_g)$

Furthermore, let X represent the statistic in each sub-population i.e.,

$$\begin{cases} X = x_1, & \text{if in} S_1; \\ X = x_2, & \text{if in} S_2; \\ \ldots, & \ldots; \\ X = x_3, & \text{if in} S_3. \end{cases}$$

Then X follows a discrete distribution with support $\{x_1, x_2, \ldots, x_g\}$ and corresponding probabilities (weights) $\{\pi_1, \pi_2, \ldots, \pi_g\}$, that is $P(X = x_j) = \pi_j$ for $j = 1, \ldots, g$.

Therefore for the finite mixture

$$Y_i \sim \sum_{j=1}^{g} \pi_j f_{ij}(y_i|\theta_j),$$

we have

$$Y_i|(X = x_j) \sim f_{ij}(y_i|\theta_j),\ j = 1, 2, \ldots, g$$

where X is denoted as follow,

$$X \sim \begin{pmatrix} x_1 & x_2 & \ldots & x_g \\ \pi_1 & \pi_2 & \ldots & \pi_g \end{pmatrix}.$$

Note the random variable X is called **latent** because, in most applications, it is not observed. We now present some examples of finite mixture distributions.

3.1.1 Examples of Mixture Distributions

Example 3.1.2 *Normal with common variance, that is,*

$$Y \sim \sum_{j=1}^{g} \pi_j N(\mu_j, \sigma^2)$$

where the parameters for this mixture are $\theta_j = (\mu_j, \sigma^2)$ *and* π_j *for* $j = 1, \ldots, g$. *Note that*

$$Y|(X = \mu_j) \sim N(\mu_j, \sigma^2)$$

where

$$X \sim \begin{pmatrix} \mu_1 & \mu_2 & \ldots & \mu_g \\ \pi_1 & \pi_2 & \ldots & \pi_g \end{pmatrix}.$$

Example 3.1.3 *Normal with common mean, that is,*

$$Y \sim \sum_{j=1}^{g} \pi_j N(\mu, \sigma_j^2)$$

where the parameters for this mixture are $\theta_j = (\mu, \sigma_j^2)$ *and* π_j *for* $j = 1, \ldots, g$. *Note that*

$$Y|(X = \sigma_j^2) \sim N(\mu, \sigma_j^2)$$

where

$$X \sim \begin{pmatrix} \sigma_1^2 & \sigma_2^2 & \ldots & \sigma_g^2 \\ \pi_1 & \pi_2 & \ldots & \pi_g \end{pmatrix}.$$

Example 3.1.4 *Normal with general mean and variance, that is,*

$$Y \sim \sum_{j=1}^{g} \pi_j N(\mu_j, \sigma_j^2)$$

where the parameters for this mixture are $\theta_j = (\mu_j, \sigma_j^2)$ *and* π_j *for* $j = 1, \ldots, g$. *Note that*

$$Y|(X_1 = \mu_j, X_2 = \sigma_j^2) \sim N(\mu_j, \sigma_j^2)$$

where

$$X = (X_1, X_2) \sim \begin{pmatrix} (\mu, \sigma_1^2) & (\mu, \sigma_2^2) & \ldots & (\mu, \sigma_g^2) \\ \pi_1 & \pi_2 & \ldots & \pi_g \end{pmatrix}.$$

3.1.2 Mean and Variance of Mixtures

Let $Y \sim \sum_{j=1}^{g} \pi_{ij} f_{ij}(y_i|\theta_j)$ be a random variable that has a mixture distribution. Using the latent variable definition above, the mean and variance have the following known basic probability results for any random variables.

Proposition 3.1.5 $E(Y) = E(E(Y|X))$

Proposition 3.1.6 $Var(Y) = Var(E(Y|X)) + E(Var(Y|X))$

This implies that the mean and variance of Examples (3.1.2), (3.1.3) and (3.1.4) are given as: For Example (3.1.2) we have

$$E(Y) = E(E(Y|X)) = E(X) = \sum_{j=1}^{g} \pi_j \mu_j$$

and

$$\begin{aligned} Var(Y) &= Var(E(Y|X)) + E(Var(Y|X)) \\ &= Var(X) + E(Var(X)) \\ &= \sum_{j=1}^{g} \pi_j \mu_j^2 - \Big(\sum_{j=1}^{g} \pi_j \mu_j\Big)^2 + E(\sigma^2) \\ &= \sum_{j=1}^{g} \pi_j \mu_j^2 - \Big(\sum_{j=1}^{g} \pi_j \mu_j\Big)^2 + \sigma^2. \end{aligned}$$

Example (3.1.3) results in

$$E(Y) = E(E(Y|X)) = E(\mu) = \mu$$

and

$$\begin{aligned} Var(Y) &= Var(E(Y|X)) + E(Var(Y|X)) \\ &= Var(\mu) + E(\sigma_j^2) \\ &= \sum_{j=1}^{g} \pi_j \sigma_j^2. \end{aligned}$$

For Example (3.1.4) we have

$$E(Y) = E(E(Y|X)) = E(X_1) = \sum_{j=1}^{g} \pi_j \mu_j$$

and

$$\begin{aligned} Var(Y) &= Var(E(Y|X)) + E(Var(Y|X)) \\ &= Var(X_1) + E(X_2) \\ &= \sum_{j=1}^{g} \pi_j \mu_j^2 - \Big(\sum_{j=1}^{g} \pi_j \mu_j\Big)^2 + E(\sigma_j^2) \\ &= \sum_{j=1}^{g} \pi_j \mu_j^2 - \Big(\sum_{j=1}^{g} \pi_j \mu_j\Big)^2 + \sum_{j=1}^{g} \pi_j \sigma_j^2. \end{aligned}$$

3.1.3 Comparison of Two Groups: Iris Data

Here we will use data to illustrate the importance of mixture distribution. The iris data is found in the statistical software package R consisting of 100 sample points of two species of flowers, Versicolor and Virginica was used for this illustrative purpose. For each species the measurements of the sepal length of 50 flowers were reported. It is clear that we have a dataset that is composed of two different populations. Since mixture distribution is applicable in the case where the data has sub-populations, we use this example to illustrate the idea of fitting mixture distribution. Note that in dealing with real life problems one will not have any information as to whether the data is composed of different populations. The histograms for both samples are presented in Figure 3.1.

The summary statistics is given in Table 3.1. For this data we have no evidence that the data is not normally distributed, because the Kolmogorov-Smirnov tests for normality resulted in a p-value > 0.5 for both groups. The Q-Q plots are displayed in Figures 3.2 and 3.3. Additionally, the assumption of equal variance is satisfied because the p-value for the F-test is 0.148.

Table 3.1: Summary statistics of data.

Species	Sepal Means	Sepal Std. Dev.
Versicolor	5.94	0.516
Virginica	6.59	0.636

The known normal mixture distribution using the summary statistics displayed in Table 3.1 is

$$0.5N(5.94, 0.516^2) + 0.5N(6.59, 0.636^2)$$

and represented graphically in Figure 3.4. However, when a two-component mixture of normals with equal variance was fitted to the data, the following fitted distribution was obtained (Figure 3.5)

$$0.83N(6.08, 0.526^2) + 0.17N(7.13, 0.526^2)$$

Figure 3.6 shows the comparison of the fitted mixture model with equal variance and

Figure 3.1: Histogram of Sepal length of the two species of flowers

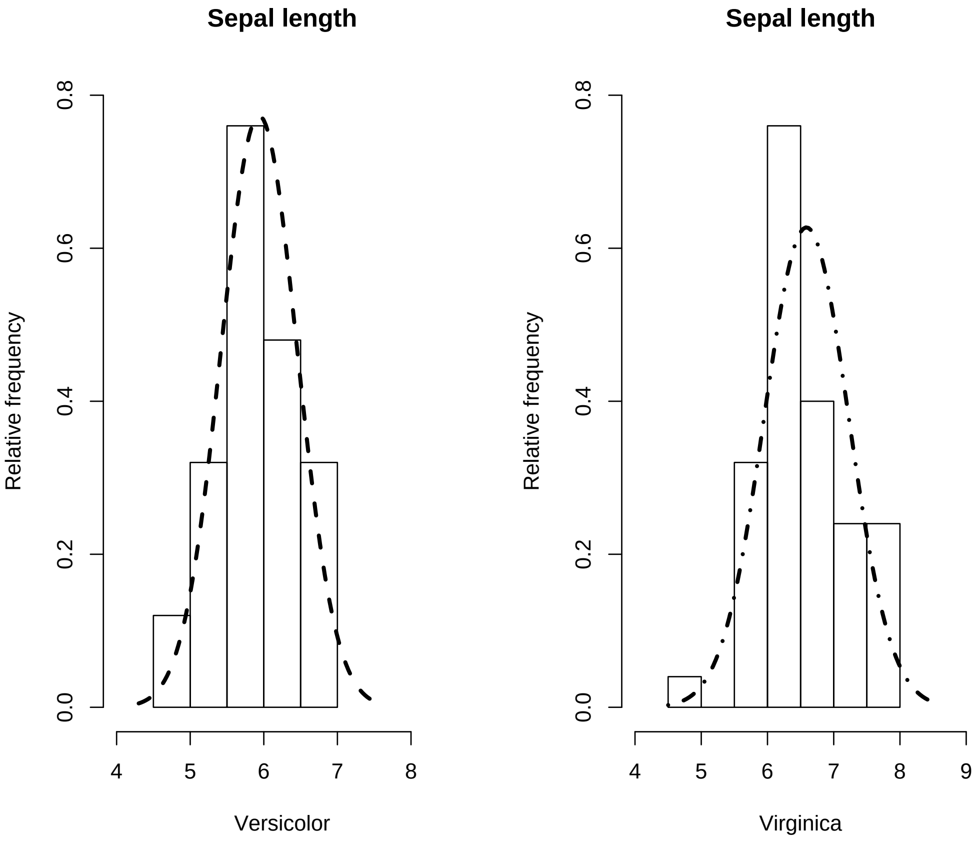


Figure 3.2: Q-Q plots of Sepal lengths for versicolor flowers

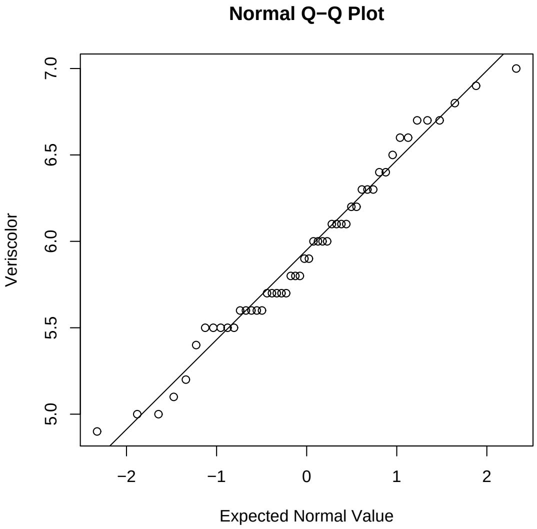


Figure 3.3: Q-Q plots of Sepal lengths for verginica flowers

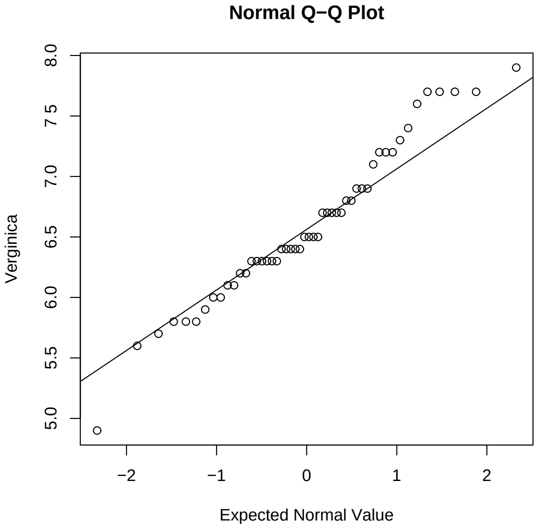
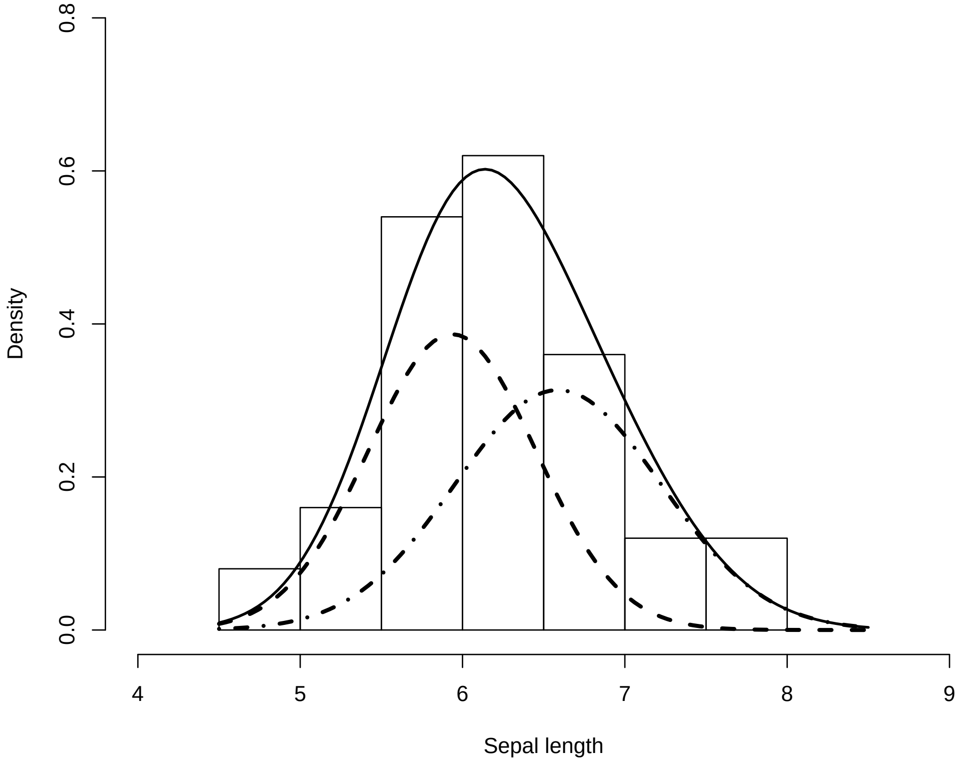


Figure 3.4: Histogram and known mixture distribution

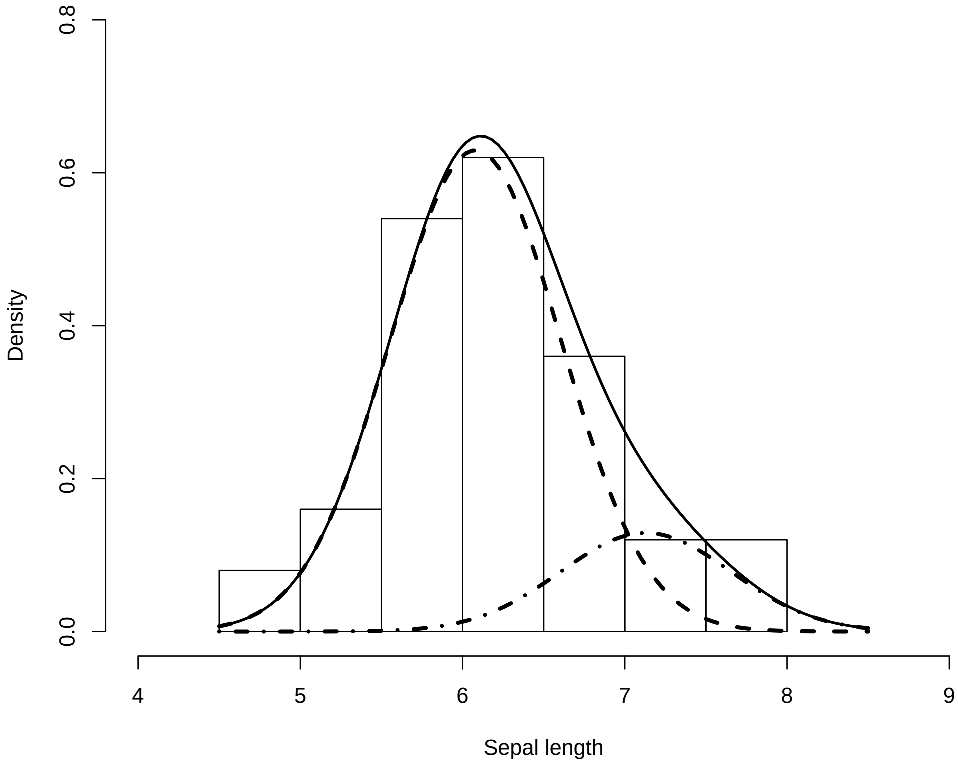


Dotted lines to the left and right represents the known distributions of versicolor and virginica respectively. The known mixture structure is $0.5N(5.94, 0.516^2) + 0.5N(6.59, 0.636^2)$.

the known mixture model. This example illustrates that the fitted mixture distribution does not necessarily reflect prior known group structures in the data.

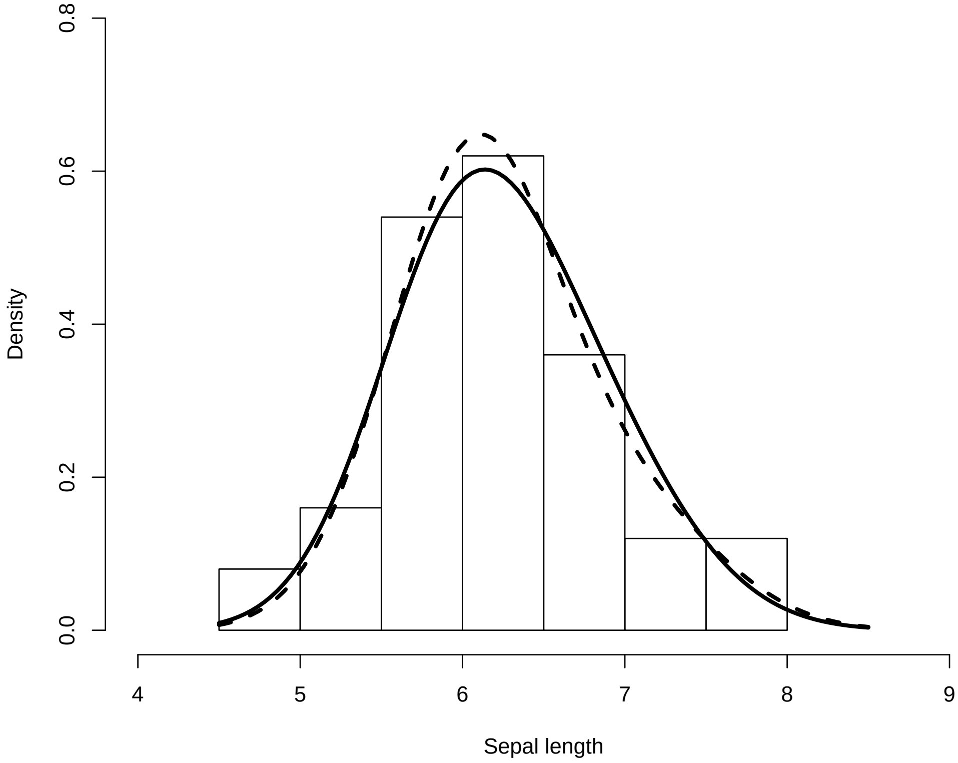
In reality the estimated mixture distribution obtained for the illustrative example may be symmetric. The distribution may be bimodal or multimodal in the case where we have more than two components.

Figure 3.5: Histogram and estimated mixture distribution



Dotted lines to the left and right represents the fitted distributions of versicolor and virginica respectively. The fitted model is given by $0.83N(6.08, 0.526^2) + 0.17N(7.13, 0.526^2)$.

Figure 3.6: Histogram with known and estimated mixture distribution



Dotted line represents the fitted mixture model while the bold line is the known mixture structure.

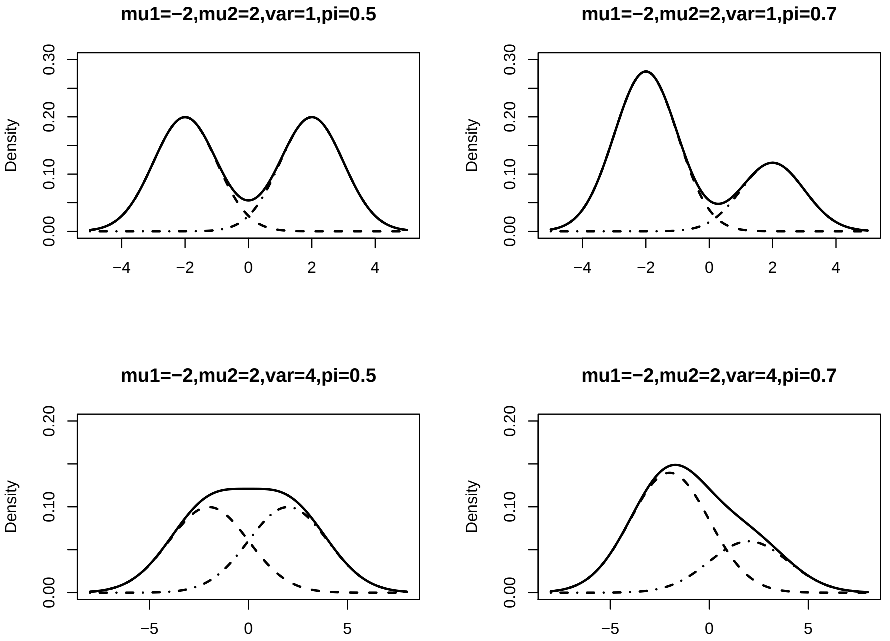
Figure 3.7 depicts that mixtures have very flexible class of models, that is:

1. They are symmetric as well as skewed

2. Unimodal as well as multimodal.

Figure 3.7: Graphical representations of two component normal with equal variance

$$\pi N(\mu_1, \sigma^2) + (1 - \pi)N(\mu_2, \sigma^2)$$



From Figure 3.7 we see that the following proposition below determines the modality of a 2-component mixture if the parameters are known, but in general we do not know μ_1, μ_2 and σ.

Proposition 3.1.7 *The modality of the 2-component mixture of normals with equal variance is determined as follows.*

$$If \quad \frac{|\mu_1 - \mu_2|}{\sigma} \quad \begin{cases} \leq 2 & \text{then the mixture is unimodal } \forall\ \pi \\ > 2 & \text{then the modality of the mixture depends on } \pi. \end{cases}$$

3.2 Parameter Estimation

3.2.1 Expectation Maximization Algorithm

This section describes how the parameters of a g-component finite mixture distribution can be estimated using maximum likelihood estimation (MLE) [10]. Let $\{Y_i\}_{1\leq i\leq n}$ be distributed as

$$\begin{aligned} Y_i &\sim \pi_1 f_{i1}(y_i|\theta_j) + \pi_2 f_{i2}(y_i|\theta_j) + \ldots + \pi_g f_{ig}(y_i|\theta_g) \\ &= \sum_{j=1}^{g} \pi_j f_{ij}(y_i|\theta_j), \end{aligned}$$

where $f_{ij}(y_i|\theta_j)$ are density functions of Y_i in a g-component mixture. The parameters of interest are that of the density functions $f_{ij}(y_i|\theta_j)$ which we denote as a vector θ and the proportion probability $\pi' = (\pi_1, \ldots, \pi_g)$. In short, the mixture distribution parameters can be denoted as a vector $\psi' = (\pi', \theta')$. Let $y' = (y_1, \ldots, y_g)$ be a vector of observed values, then the observed likelihood function is given to be:

$$L(\psi|y) = \prod_{i=1}^{n} \left\{ \sum_{j=1}^{g} \pi_j f_{ij}(y_i|\theta) \right\}, \tag{3.2.2}$$

additionally, the observed log-likelihood is given by:

$$l(\psi|y) = \sum_{i=1}^{n} \ln \left\{ \sum_{j=1}^{g} \pi_j f_{ij}(y_i|\theta) \right\}. \tag{3.2.3}$$

We now need to maximize the log-likelihood $l(\psi|y)$ with respect to ψ. This is done by using the Expectation-Maximization (EM) (Dempster et al., 1977) algorithm as an alternative to the Newton-Raphson which involves the calculation of first and second derivatives of $l(\psi|y)$. The EM algorithm was developed for missing observation, in our case we considered the component membership as missing. This can be seen if we define indicators Z_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, g$ such that

$$Z_{ij} = \begin{cases} 1 & \text{if observation } i \text{ belongs to component } j \\ 0 & \text{otherwise} \end{cases}$$

Therefore we have that $P(Z_{ij} = 1) = \pi_j$, and hence the joint density of Y_i and all Z_{ij} is given by

$$
\begin{aligned}
& f_i(y_i, Z_{i1} = z_{i1}, \ldots, Z_{ig} = z_{ig}) \\
&= f_i(y_i | Z_{i1} = z_{i1}, \ldots, Z_{ig} = z_{ig}) P(Z_{i1} = z_{i1}, \ldots, Z_{ig} = z_{ig}) \\
&= \left\{ \prod_{j=1}^{g} [f_{ij}(y_i|\theta)]^{z_{ij}} \right\} \left\{ \prod_{j=1}^{g} \pi_j^{z_{ij}} \right\} \\
&= \left\{ \prod_{j=1}^{g} [\pi_j f_{ij}(y_i|\theta)]^{z_{ij}} \right\}
\end{aligned}
$$

Therefore the likelihood of the complete data is

$$
L(\psi|y, z) = \prod_{i=1}^{n} \prod_{j=1}^{g} [\pi_j f_{ij}(y_i|\theta)]^{z_{ij}} \tag{3.2.4}
$$

and the log-likelihood of the complete data is

$$
l(\psi|y, z) = \sum_{i=1}^{n} \sum_{j=1}^{g} z_{ij} [\ln \pi_j + \ln f_{ij}(y_i|\theta)]. \tag{3.2.5}
$$

It is therefore obvious that maximizing $l(\psi|y, z)$ ("the complete log likelihood") is easier than maximizing $l(\psi|y)$ ("the observe log likelihood"). Note that (3.2.2) and (3.2.3) are referred to as the observe data likelihood and observe log-likelihood respectively, while (3.2.4) and (3.2.5) are referred to as the complete data likelihood and complete log-likelihood respectively. Instead of maximizing $l(\psi|y, z)$ we maximize $E(l(\psi|y, z)|y)$, which is interpreted intuitively as replacing the missing observations z_{ij} by their expected values.

The EM algorithm acts iteratively, in the sense that, starting from a "first guess estimate" (starting value) $\psi^{(0)}$ for ψ, a series of estimates $\psi^{(t)}$ is constructed, which converges to the MLE $\hat{\psi}$ of ψ

$$
\psi^{(0)} \to \psi^{(1)} \to \ldots \to \psi^{(t)} \to \psi^{(t+1)} \to \ldots \to \psi^{(\infty)} = \hat{\psi}
$$

Given $\psi^{(t)}$, the updated estimate $\psi^{(t+1)}$ is obtained through one E-step and one M-step.

Definition 3.2.1 *The E-step is the calculation of* $Q(\psi|\psi^{(t)}) = E(l(\psi|y,z)|y,\psi^{(t)})$.

Definition 3.2.2 *The M-step is defined as the maximization of* $Q(\psi|\psi^{(t)})$ *with respect to* ψ *to obtain the updated value* $\psi^{(t+1)}$.

The EM procedure keeps iterating between the E-step and the M-step until convergence is attained, that is, until

$$|l(\psi^{(t+1)}|y) - l(\psi^{(t)}|y)| < \varepsilon.$$

for some small, pre-specified, $\varepsilon > 0$.

We now present the calculation of the E-step, therefore from definition 3.2.1, we have

$$\begin{aligned} Q(\psi|\psi^{(t)}) &= E(l(\psi|y,Z)|y,\psi^{(t)}) \\ &= E\Big(\sum_{i=1}^{n}\sum_{j=1}^{g} Z_{ij}[\ln\pi_j + \ln f_{ij}(y_i|\theta)]\Big| y,\psi^{(t)}\Big) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{g} E[Z_{ij}|y,\psi^{(t)}][\ln\pi_j + \ln f_{ij}(y_i|\theta)] \end{aligned}$$

Note the E-step requires only the calculation of

$$\begin{aligned} E[Z_{ij}|y,\psi^{(t)}] &= P(Z_{ij}=1|y_i,\psi^{(t)}) \\ &= \frac{f_i(y_i|Z_{ij}=1)P(Z_{ij}=1)}{f_i(y_i|\theta)}\Big|_{\psi^{(t)}} \\ &= \frac{\pi_j f_{ij}(y_i|\theta)}{\sum_j \pi_j f_{ij}(y_i|\theta)}\Big|_{\psi^{(t)}} \\ &= \pi_{ij}(\psi^{(t)}). \end{aligned}$$

Therefore the E-step results in

$$\pi_{ij}(\psi^{(t)}) = \frac{\pi_j f_{ij}(y_i|\theta)}{\sum_{j=1}^{g} \pi_j f_{ij}(y_i|\theta)}\Bigg|_{\psi^{(t)}} \tag{3.2.6}$$

where $\pi_{ij}(\psi^{(t)})$ is called the posterior probabilities and π_j is called the prior probabilities. Note the E-step reduces to calculating all the posterior probabilities $\pi_{ij}(\psi^{(t)})$ for $i = 1,\ldots,n$, $j = 1,\ldots,g$.

The M-step maximizes $Q(\psi|\psi^{(t)})$ with respect to ψ to obtain the updated estimates

$\psi^{(t+1)}$. Since

$$Q(\psi|\psi^{(t)}) = \sum_{i=1}^{n}\sum_{j=1}^{g} \pi_{ij}(\psi^{(t)})[\ln \pi_j + \ln f_{ij}(y_i|\theta)]$$

we first maximize with respect to π_j. This requires maximization of

$$\sum_{i=1}^{n}\sum_{j=1}^{g} \pi_{ij}(\psi^{(t)}) \ln \pi_j = \sum_{i=1}^{n}\sum_{j=1}^{g-1} \pi_{ij}(\psi^{(t)}) \ln \pi_j + \sum_{i=1}^{n} \pi_{ig}(\psi^{(t)}) \ln \Big[1 - \sum_{j=1}^{g-1} \pi_j\Big]$$

with respect to $\pi_1, \ldots, \pi_{g-1}$. Setting

$$\frac{\partial}{\partial \pi_j}\Big\{ \sum_{i=1}^{n}\sum_{j=1}^{g-1} \pi_{ij}(\psi^{(t)}) \ln \pi_j + \sum_{i=1}^{n} \pi_{ig}(\psi^{(t)}) \ln \Big[1 - \sum_{j=1}^{g-1} \pi_j\Big]\Big\} = 0$$

we have that

$$\begin{aligned} \sum_{i=1}^{n} \frac{\pi_{ij}(\psi^{(t)})}{\pi_j^{(t+1)}} &= \sum_{i=1}^{n} \frac{\pi_{ig}(\psi^{(t)})}{\pi_g^{(t+1)}} \\ \Rightarrow \frac{\pi_j^{(t+1)}}{\pi_g^{(t+1)}} &= \frac{\sum_{i=1}^{n} \pi_{ij}(\psi^{(t)})}{\sum_{i=1}^{n} \pi_{ig}(\psi^{(t)})} \end{aligned}$$

Note that

$$\begin{aligned} 1 &= \sum_{j=1}^{g} \pi_j^{(t+1)} \\ &= \sum_{j=1}^{g} \frac{\pi_g^{(t+1)} \sum_{i=1}^{n} \pi_{ij}(\psi^{(t)})}{\sum_{i=1}^{n} \pi_{ig}(\psi^{(t)})} \\ &= \frac{\pi_g^{(t+1)} \sum_{i=1}^{n} \sum_{j=1}^{g} \pi_{ij}(\psi^{(t)})}{\sum_{i=1}^{n} \pi_{ig}(\psi^{(t)})} \end{aligned}$$

since $\sum_{j=1}^{g} \pi_{ij}(\psi^{(t)}) = 1$, therefore

$$1 = \frac{\pi_g^{(t+1)} n}{\sum_{i=1}^{n} \pi_{ig}(\psi^{(t)})}$$

hence $\pi_g^{(t+1)}$ is given by

$$\pi_g^{(t+1)} = \frac{\sum_{i=1}^{n} \pi_{ig}(\psi^{(t)})}{n}$$

It follows that all $\pi_j^{(t+1)}$ are given by

$$\pi_j^{(t+1)} = \frac{\sum_{i=1}^{n} \pi_{ij}(\psi^{(t)})}{n} \tag{3.2.7}$$

Note that the updated mixture component probabilities are the average posterior probabilities. The M-step also requires the maximization of

$$\sum_{i=1}^{n}\sum_{j=1}^{g} \pi_{ij}(\psi^{(t)}) \ln f_{ij}(y_i|\theta) \tag{3.2.8}$$

with respect to θ. This maximization step is often times non-trivial. In such cases, the EM algorithm is double iterative. Below are some examples when the M-step is trivial (c.f. [40]).

Example 3.2.3 *Poisson, let* $Y_i \sim \sum_{j=1}^{g} \pi_j Poisson(\lambda_j)$ *with* $\theta = (\lambda_1, \ldots, \lambda_g)$

From (3.2.8), and for simplicity we let $\pi_{ij}(\psi^{(t)}) = \pi_{ij}$, then we have

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{j=1}^{g} \pi_{ij} \ln f_{ij}(y_i|\theta) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{g} \pi_{ij} \ln \left(\frac{e^{-\lambda_j} \lambda_j^{y_i}}{y_i!} \right) \\
&\propto \sum_{i=1}^{n}\sum_{j=1}^{g} \pi_{ij} \left(-\lambda_j + y_i \ln \lambda_j \right)
\end{aligned}$$

therefore

$$\begin{aligned}
&\frac{\partial}{\partial \lambda_j} \Big\{ \sum_{i=1}^{n}\sum_{j=1}^{g} \pi_{ij} \left(-\lambda_j + y_i \ln \lambda_j \right) \Big\} = 0, \quad \forall\, j \\
\Leftrightarrow \qquad & \lambda_j = \frac{\sum_{i=1}^{n} \pi_{ij} y_i}{\sum_{i=1}^{n} \pi_{ij}}
\end{aligned}$$

Example 3.2.4 *Normals with common variance, let* $Y_i \sim \sum_{j=1}^{g} \pi_j N(\mu_j, \sigma^2)$ *with* $\theta = (\mu_1, \ldots, \mu_g, \sigma^2)$

Similar as in Example (3.2.3), we have that

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{j=1}^{g}\pi_{ij}\ln f_{ij}(y_i|\theta)\\
&=\sum_{i=1}^{n}\sum_{j=1}^{g}\pi_{ij}\ln\left[\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(y_i-\mu_j)^2\right\}\right]\\
&\propto\sum_{i=1}^{n}\sum_{j=1}^{g}\pi_{ij}\left[-\ln(\sigma^2)/2-(y_i-\mu_j)^2/(2\sigma^2)\right]
\end{aligned}$$

Therefore, we minimize

$$\sum_{i=1}^{n}\sum_{j=1}^{g}\pi_{ij}\left[\ln(\sigma^2)/2+(y_i-\mu_j)^2/(2\sigma^2)\right]$$

therefore

$$\frac{\partial}{\partial\mu_j}\Big\{\sum_{i=1}^{n}\sum_{j=1}^{g}\pi_{ij}\left[\ln(\sigma^2)/2+(y_i-\mu_j)^2/(2\sigma^2)\right]\Big\}=0,\quad\forall\, j$$

$$\Leftrightarrow\mu_j=\frac{\sum_{i=1}^{n}\pi_{ij}y_i}{\sum_{i=1}^{n}\pi_{ij}}. \tag{3.2.9}$$

Also

$$\frac{\partial}{\partial\sigma_j^2}\Big\{\sum_{i=1}^{n}\sum_{j=1}^{g}\pi_{ij}\left[\ln(\sigma^2)/2+(y_i-\mu_j)^2/(2\sigma^2)\right]\Big\}=0,\quad\forall\, j$$

$$\begin{aligned}
\Leftrightarrow\quad&\sum_{i=1}^{n}\sum_{j=1}^{g}\pi_{ij}\left[1/\sigma^2-(y_i-\mu_j)^2/\sigma^4\right]=0\\
\Leftrightarrow\quad&\sum_{i=1}^{n}\sum_{j=1}^{g}\pi_{ij}=\sum_{i=1}^{n}\sum_{j=1}^{g}\pi_{ij}(y_i-\mu_j)^2/\sigma^2\\
\Leftrightarrow\quad&\sigma^2=\frac{\sum_{i=1}^{n}\sum_{j=1}^{g}\pi_{ij}(y_i-\mu_j)^2}{\sum_{i=1}^{n}\sum_{j=1}^{g}\pi_{ij}}\\
\Leftrightarrow\quad&\sigma^2=\frac{\sum_{i=1}^{n}\sum_{j=1}^{g}\pi_{ij}(y_i-\mu_j)^2}{n}
\end{aligned}$$

$$\Leftrightarrow \qquad \sigma^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{g}\pi_{ij}\left(y_i - \frac{\sum_{i=1}^{n}\pi_{ij}y_i}{\sum_{i=1}^{n}\pi_{ij}}\right)^2}{n}. \tag{3.2.10}$$

Example 3.2.5 *Normals with general mean and variance, let* $Y_i \sim \sum_{j=1}^{g}\pi_j N(\mu_j, \sigma_j^2)$ *with* $\theta = (\mu_1, \ldots, \mu_g, \sigma_1^2, \ldots, \sigma_g^2)$

Similar to Example (3.2.4), we can show that the mean is given by

$$\mu_j = \frac{\sum_{i=1}^{n}\pi_{ij}y_i}{\sum_{i=1}^{n}\pi_{ij}}$$

Note that the variance estimator is only achieved if we assume that all the variances are equal, i.e $\sigma_j^2 = \sigma^2$. Since the log-likelihood of this model is

$$\begin{aligned}
l(\psi|y) &= \sum_{i=1}^{n}\ln\left\{\sum_{j=1}^{g}\pi_j\frac{1}{\sqrt{2\pi\sigma_j^2}}\exp\left[-\frac{1}{2\sigma_j^2}(y_i-\mu_j)^2\right]\right\} \\
&= \sum_{i=2}^{n}\ln\left\{\sum_{j=2}^{g}\pi_j\frac{1}{\sqrt{2\pi\sigma_j^2}}\exp\left[-\frac{1}{2\sigma_j^2}(y_i-\mu_j)^2\right]\right. \\
&\qquad + \left.\pi_1\frac{1}{\sqrt{2\pi_1\sigma_1^2}}\exp\left[-\frac{1}{2\sigma_1^2}(y_i-\mu_1)^2\right]\right\} \\
&\quad + \ln\left\{\sum_{j=2}^{g}\pi_j\frac{1}{\sqrt{2\pi\sigma_j^2}}\exp\left[-\frac{1}{2\sigma_j^2}(y_1-\mu_j)^2\right]\right. \\
&\qquad + \left.\pi_1\frac{1}{\sqrt{2\pi_1\sigma_1^2}}\exp\left[-\frac{1}{2\sigma_1^2}(y_1-\mu_1)^2\right]\right\}
\end{aligned}$$

Let μ_1 equal y_1, then we have

$$\begin{aligned}
l(\psi|y) &= \sum_{i=2}^{n}\ln\left\{\sum_{j=2}^{g}\pi_j\frac{1}{\sqrt{2\pi\sigma_j^2}}\exp\left[-\frac{1}{2\sigma_j^2}(y_i-\mu_j)^2\right]\right. \\
&\qquad + \left.\pi\frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left[-\frac{1}{2\sigma_1^2}(y_i-\mu_1)^2\right]\right\} \\
&\quad + \ln\left\{\sum_{j=2}^{g}\pi_j\frac{1}{\sqrt{2\pi\sigma_j^2}}\exp\left[-\frac{1}{2\sigma_j^2}(y_1-\mu_j)^2\right]\right.
\end{aligned}$$

$$+ \quad \pi_1 \frac{1}{\sqrt{2\pi\sigma_1^2}} \Bigg\}$$

It is straight forward to see that the $l(\psi|y)$ is unbounded if $\sigma_1^2 = 0$. This is the reason why it is vital that we have all the variances equal i.e. $\sigma_j^2 = \sigma^2$ see Example 3.2.4. We will show in the section 3.2.3 how we can apply mixture of normals with unequal variances by implementing a penalty term .

3.2.2 Robust Parameter Estimation

In the previous section the EM algorithm was presented to find the parameters of the mixture models. The parameters however are sensitive to the presence of statistical outliers [33]. In microarray data analysis we are not immune to statistical outliers, therefore the parameter estimation problem where the presence of outliers exist should be addressed. The solution to this problem is accomplished by the Robust parameter estimation for mixture model, which will be presented below.

There are several factors affecting the convergence of the EM algorithm to the maximum likelihood estimates. These factors are:

1. the initial estimates can affect the convergence greatly and
2. the presence of statistical outliers defined to be those observations that are substantially different from the distributions of the mixture components.

The EM algorithm assigns each observation to one of the components with the sample's posterior probability as its weight. Although an outlying sample is inconsistent with the distributions of all the defined components, it may still have a large posterior probability for one or more of the components. Therefore the iteration converges to erroneous solutions.

A common approach to eliminating the presence of outliers in the EM algorithm is to apply a chi-square threshold test. This test eliminates observations with distances greater than some threshold value. These observations are considered to be outliers and subsequently excluded from updating the parameter estimates. This chi-square threshold χ_α^2 for a given probability α is defined as the square distance between the sample $y \in \Re$

and the mean of the j^{th} component based on the chi-square test shown below:

$$P\left\{y \left| \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{\sigma^2} \leq \chi_\alpha^2 \right.\right\} = \alpha$$

The threshold approach can be regarded as performing a hard decision to eliminate outlying sample points before initiating the EM algorithm. Furthermore, a suitable threshold value is often difficult to select and is usually arbitrary. In view of this difficulty, an alternative would be to assign different weight to each data points and use all available data points for updating the estimates. This method may be regarded as applying a soft decision. The Robust Parameter Estimation For Mixture Model will be discussed next.

It should be noted that the EM algorithm first estimates the posterior probabilities of each sample belonging to each of the component distributions, and then computes the parameter estimates using these posterior probabilities as weights. With this method, each sample is assumed to come from all components. The robust estimation attempts to circumvent this problem by including the typicality of the sample with respect to the component densities in updating the estimates in the EM algorithm.

A measure of typicality is incorporated in the parameter estimation of the mixture density, if we assume that each component density $f_j(y_i|\mu_j, \sigma^2)$ is a member of the family of symmetric densities with mean μ_j and σ^2, i.e.

$$\left(2\pi\sigma^2\right)^{-1/2} f_s\{\delta_j(x|\mu_j, \sigma^2)\},$$

where $\delta_j^2 = \frac{(y-\mu_j)^2}{\sigma^2}$, and $f_s(\delta_j)$ is assumed to be the exponential of some symmetric function $\rho(\delta_j)$, i.e.

$$f_s(\delta_j) = \exp\{-\rho(\delta_j)\}.$$

$$\pi_j^{(t+1)} = \frac{\sum_{i=1}^{g} \pi_{ij}}{n},$$

$$\mu_j^{(t+1)} = \frac{\sum_{i=1}^{n} \pi_{ij} w_{ij} y_i}{\sum_{i=1}^{n} \pi_{ij}},$$

$$(\sigma^2)^{(t+1)} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{g}\pi_{ij}w_{ij}\left(y_i-\mu_j\right)^2}{n}.$$

where $w_{ij} = \psi(\delta_{ij})/\delta_{ij}$ is the weight function and $\psi(\delta_{ij}) = \rho'(\delta_{ij})$ is the first derivative of $\rho(\delta_{ij})$. To limit the influence of large atypical data points, the variance estimator is modified to be

$$(\sigma^2)^{(t+1)} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{g}\pi_{ij}w_{ij}^2\left(y_i-\mu_j\right)^2}{n}.$$

The weight function has been chosen to be $\psi(s)/s$ where $s = \delta_{ij}$. A popular choice of $\psi(s)$ is the Huber's ψ-function that is defined by $\psi(s) = -\psi(-s)$ where for $s > 0$

$$\psi(s) = \begin{cases} s & 0 \le s \le k \\ k & s > k \end{cases}$$

and k is called a tuning constant, and needs to be appropriately chosen. Furthermore we have

$$\rho(s) = \begin{cases} \frac{1}{2}s^2 & 0 \le s \le k \\ ks - \frac{1}{2}k^2 & s > k. \end{cases}$$

In the case of normal mixture distributions, the value of the tuning k is chosen to be 3 standard deviation from the mean as most data point should fall within this band and is given a unit weight. The outliers are then given weights which are inversely proportional to their distances from the class mean. Hence, the weights can be expressed as:

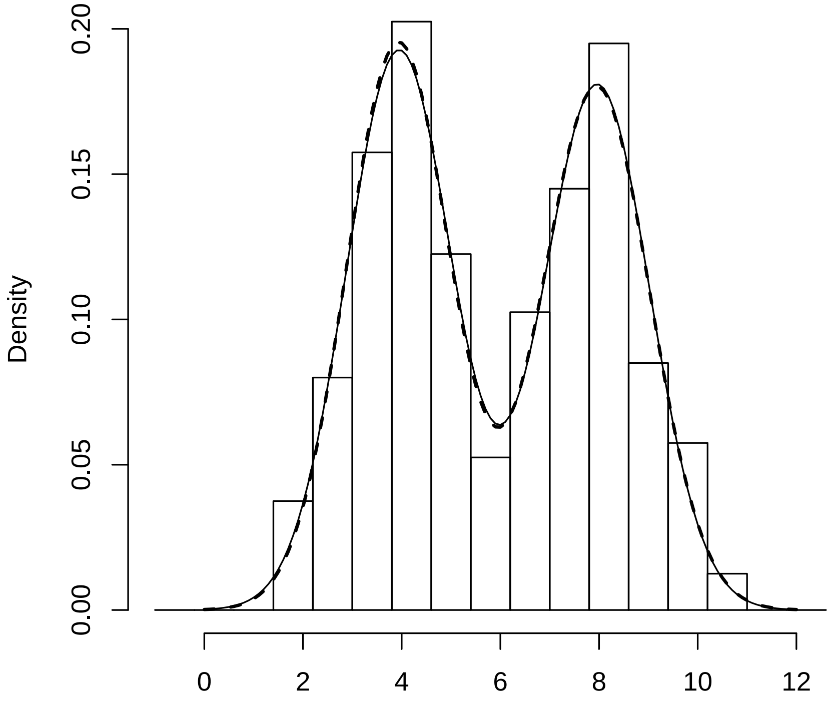
$$w_{ij} = \begin{cases} 1 & 0 \le d_{ij} \le 3 \\ 3/d_{ij} & 3 < d_{ij} < \infty \end{cases}$$

where $d_{ij} = \frac{(y_i-\mu_j)}{\sigma_j}$

3.2.3 Penalized Maximum Likelihood Estimator for Normal Mixture Models

We illustrated through Example 3.2.5 in section 3.2 that we can only fit mixture of normals with equal variance which was proved by Kiefer and Wolfowitz (1954). However, Ciuperca et al. (2003) overcame this difficulty by penalizing the variance, which allowed the likelihood function of the normal mixture model to be bounded, hence the existence of the MLE. If we fit a mixture model with equal variance if in fact the mixture heteroscedasticity we observe that homoscedastic model does not result in a good fit as compared to the heteroscedasticity fit.

Figure 3.8: Histogram, heteroscedastic and homoscedastic fit for simulated data from the mixture $0.5\phi(y|4,1) + 0.5\phi(y|8,1)$

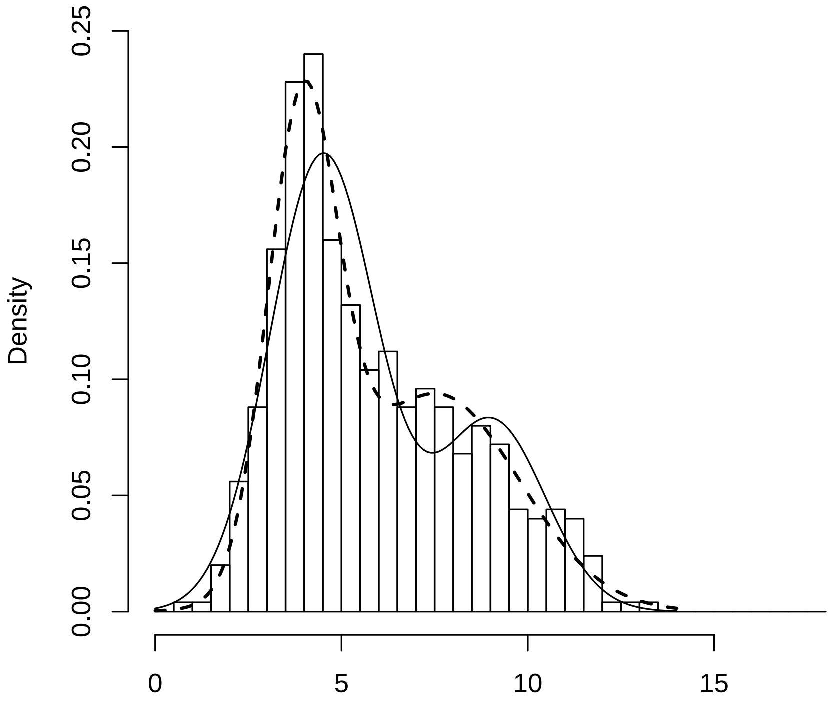


The dotted and bold lines represent the heteroscedastic and homoscedastic models respectively.

We simulated the following mixture distributions from a sample of size $n = 500$ from

$$Y \sim 0.5\phi(y|4,1) + 0.5\phi(y|8,1)$$

Figure 3.9: Histogram, heteroscedastic and homoscedastic fit for simulated data from the mixture $0.5\phi(y|4,1) + 0.5\phi(y|8,2)$



The dotted and bold lines represent the heteroscedastic and homoscedastic models respectively.

and

$$Y \sim 0.5\phi(y|4,1) + 0.5\phi(y|8,2)$$

and then fit the simulated data with equal and unequal variances. The model with unequal variances seems to be a better fit in the case where the simulated data with unequal variance was fitted with unequal variance as oppose to when fitted using equal variance Figure 3.8. However the results for the data that was simulated using equal variance see Figure 3.9

This example shows that we attain better fit to our data if the data is heteroscedastic, hence fitting heteroscedastic mixture model is vital. Ciuperca et al. considered mixture densities of g univariate normal densities, with g known, defined as in (3.1.1), i.e.,

$$f_1(Y|\psi) = \sum_{j=1}^{g} \pi_j f_{ij}(y_i|\theta_j) \tag{3.2.11}$$

where

$$f_{ij}(y_i|\theta_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left\{-\frac{1}{2\sigma_j^2}(y_i - \mu_j)^2\right\} \quad j = 1, \ldots, g$$

are normal densities with mean μ_j and standard deviation σ_j. The parameter set of the mixture is

$$\Psi = (\pi_1, \ldots, \pi_j, \mu_1, \ldots, \mu_j, \sigma_1, \ldots, \sigma_j) \tag{3.2.12}$$

such that $0 \leq \pi_j \leq 1, \sum_{j=1}^{g} \pi_j = 1, -\infty < \mu_j < \infty$, $\sigma_j > 0$ and the true parameters defined as $\psi_0 \in \Psi$.

In their analysis the maximum likelihood (ML) framework was used to estimate the parameters of the mixture, with likelihood function given by

$$\tilde{L}(\psi|y) = f_n(Y_1, \ldots, Y_n|\psi) = \prod_{i=1}^{n} f_1(Y|\psi). \tag{3.2.13}$$

Since the likelihood function (3.2.13) is unbounded on Ψ because if one of the variance parameter in the denominator of (3.2.13) approaches 0 as μ_j approaches y_i (c.f. Example 3.2.5) then the likelihood is unbounded.

They circumvented this problem by considering a penalized likelihood function defined as

$$\check{L}_n(\psi|y) = f_n(Y_1, \ldots, Y_n|\psi) \prod_{j=1}^{g} h(\sigma_j) \tag{3.2.14}$$

where the penalized function h was chosen so that $\check{L}_n$ is bounded over the parameter space Ψ. The penalized function was assumed to have satisfied the following conditions:

(1) $\lim_{\sigma_j \to 0} \frac{1}{\sigma_j^n} h(\sigma_j) = 0$, for all n, which ensures that for any fixed n,

the maximum argument of $\check{L}_n$, that is the penalized MLE

$$\arg\max_{\psi\in\Psi} \check{L}_n \text{ exists.}$$

The consistency of the estimator was also a concern. In order to prove the consistency they required that h also satisfied the following conditions:

(2) $h(\sigma)$ is many-to-one from $(0, \infty)$ onto $(0, G], G > 0$,

(3) h is strictly increasing in an open interval $(0, \delta)$ of the origin which has a non-null measure,

(4) h is continuously differentiable on $(0, \infty)$.

3.3 Estimating the Number of Components g

One interesting but difficult problem is to determine the number of components g. This can be accomplished through using various model selection criteria, of which the most well known are the Akaike Information Criterion (AIC) (Akaike 1973) and the Bayesian Information Criterion (BIC) (Schwartz 1978)

$$AIC = -2\log L(\Psi_g) + 2\nu_g,$$

$$BIC = -2\log L(\Psi_g) + \nu_g \log(n),$$

where ν_g is the number of independent parameters in Ψ_g. In using the AIC or BIC, one first fits a series of models with various values of g, then picks up the g corresponding to the first local minimum of AIC or BIC (Fraley and Raftery 1998). Some other criteria have been studied but it does not appear that there exists a clear winner (Biernacki and Govaert 1999). Some empirical studies seem to favor the use of BIC (Fraley and Raftery 1998). With this in mind the AIC and BIC may not agree with each other in some cases, therefore it often means that several models are reasonable and that no one can dominate the others. Therefore we seek other methods which are more reliable in the selection of g, the number of components. A different approach to selecting g is through hypothesis testing. This could be done through the use of the likelihood ratio test (LRT) to test for the null hypothesis $H_0 : g = k$ against $H_1 : g = k + 1$ for any given positive integer k. The LRT statistic is $2(\ln L(\Psi_{k+1}) - \ln L(\Psi_k))$, which, however, does not have the usual asymptotic chi-squared distribution because of the loss of identifiability of the null distribution and also that the null hypothesis lies on the boundary of the parameter space ($\pi = 0$). Without loss of generality, let us assume that a random sample $Y_1, \ldots, Y_n$ is from the mixture

$$(1-\pi)f_{i1}(y_i|\theta_1) + \pi f_{i2}(y_i|\theta_2) \tag{3.3.15}$$

where $\theta_1 \leq \theta_2$ and $0 \leq \pi \leq 1$. The hypothesis we wish to test is

$$H_0 : \theta_1 = \theta_2,$$

therefore we see that the two statements $\pi = 0$ and $\theta_1 = \theta_2$ are equivalent hence the parameters π, θ_1 and θ_2 are not identifiable under the null model. In the next few sections we shall discuss how we may achieve the asymptotic null distribution of the log likelihood ratio statistic through the use of: (1) simulation and (2) the modified likelihood ratio test.

3.3.1 Simulation Approach

Here we shall describe how to simulate the degrees of freedom of the null distribution of the likelihood ratio test (LRT)

$$2(\ln L(\Psi_{k+1}) - \ln L(\Psi_k))$$

from a univariate normal mixture distribution for the hypothesis H_0 versus H_1, see Everitt et al. (1981). Without loss of generality we assume distribution under null hypothesis H_0 is normally distributed that is the number of component $g = k = 1$ and the distribution under the alternate is a two component mixture of normal distribution, that is, $g = k = 2$. Note that the distribution of

$$\ln L(\Psi_{k+1}) - \ln(\Psi_k)$$

and

$$2(\ln L(\Psi_{k+1}) - \ln L(\Psi_k)),$$

clearly depends on n. McLachlan et al. (1987) simulated the homoscedastic case, that is, mixture of normal with equal variances for each component, using 500 replicates for samples of sizes $n = 25, 50$ and 100, under H_0. The mean(variance) of the simulated null distribution of $2(\ln L(\Psi_{k+1}) - \ln L(\Psi_k))$ was found to be equal to $2.47(5.66), 2.36(5.06)$, and 2.16(4.30) for $n = 25, 50$, and 100 respectively. The empirical distribution function of $2(\ln L(\Psi_{k+1}) - \ln L(\Psi_k))$, generated from the 500 replicated simulated values of the test statistic for $n = 100$, was shown to be similar in distribution of the χ^2_2 distribution function. McLachlan et al. (1987) explain that the choice of the χ^2_2 distribution corresponds to the approximation of Wolfe (1971), where the degrees of freedom of the chi-squared distribution is taken to be twice the difference in the number of parameters under H_0 and H_1, excluding the mixing proportions.

McLachlan further stated that the Wolfe's approximation to the null distribution of $2(\ln L(\Psi_{k+1}) - \ln L(\Psi_k))$ was not applicable in the heteroscedastic case (i.e where the component variances were unequal). McLachlan evaluated the empirical distribution function of $2(\ln L(\Psi_{k+1}) - \ln L(\Psi_k))$ by constructing 500 replicates with a sample size of $n = 100$ generated under H_0 using the normal component densities having unequal variances under H_1. When Wolfe's approximation was applied, the resulting chi-squared distribution was χ^2_4 however, the χ^2_6 distribution function was found to provide a much better fit. Furthermore, the simulated null distribution of $2(\ln L(\Psi_{k+1}) - \ln L(\Psi_k))$ had mean and variance equal to 5.96 and 13.86 respectively which further solidified that the χ^2_6 distribution function characterizes the empirical null distribution. Wolf's approximation was not applicable in the case where heteroscedastic was considered.

In the case of heteroscedasticity the regression approach of Thode et al. (1988) is more appropriate to be used to remedy the aforemention situation of unequal variances. The approach is to fit a regression model as a function of the sample size n, using different sample sizes which results in the regressed degrees of freedom to be

$$f = \beta_0 + \beta_1 \frac{1}{\sqrt{n}}. \tag{3.3.16}$$

From equation (3.3.16) we observe that the asymptotic degrees of freedom is β_0.

The regression technique of Thode et al. (1988) was presented to determine the degrees of freedom of the asymptotic distribution of the likelihood ratio test. Thode et al. found the empirical null distribution of the likelihood ratio test for the sample sizes 15, 20, 25, 40, 50, 75, 80, 100, 150, 250, 500 and 1,000. However, their approach did not account for skewness which was addressed by MacLean et al. (1976). Furthermore, for each sample size, percentile points and moments were evaluated using 2,500 normal samples. Thode et al. also used an iterative procedure to determine the maximum likelihood estimates of the normal mixture distribution. They also applied the random starting point method of Thode, Finch and Mendell (1987) by using five random starting points so that the global maximum is achieved, instead of the local maximum of the MLE of the parameters in the normal mixture model.

Thode et al. mentioned that since the regularity conditions do not hold in the case of mixture of normal distribution, therefore the asymptotic distribution is not chi-squared with degrees of freedom 2. Therefore they found the means and variances for the sample sizes 15, 20, 25, 40, 50, 75, 80, 100, 150, 250, 500 and 1,000. Note that the mean is equal to the number of degrees of freedom for the chi-squared random variable, and the variance is twice the degrees of freedom. They also estimated the asymptotic distribution of the likelihood ratio test by regressing the mean, variance and simulated percentiles of the LRT against various functions of the sample size n. Thode et al. further divided the 2,500 samples generated for each of the sample into 5 subsamples of size 500 each, and applied the goodness-of-fit test described in Draper and Smith (1981) and considered a regression model as a function of $(1/n)^t$ for $t = 0.125, 0.25, 0.50, 1, 2$ and 3. The regression model is

$$E(Y_{PNs}) = a_{P,t} + b_{P,t}(1/n)^t, \tag{3.3.17}$$

where Y_{PNs} is the P^{th} percentile of the s^{th} subsample of size n. From model (3.3.17) they fitted regression model for $t = 0.125, 0.25, 0.50, 1, 2$ and 3 and found that the intercepts estimated for various powers of t were essentially the same therefore indicating the convergence of the asymptotic distribution. However, Thode et al. concluded that the regression model of the mean on $(1/n)^{0.5}$ suggested a very good goodness-of-fit statistics and a value of R^2 around 0.6. Therefore in this dissertation we will regress the means on $(1/n)^{0.5}$ and use $a_{\bar{x},0.5}$ as our asymptotic degrees of freedom.

In the next section we will describe the approach of Chen et al. that was used to determine the exact distribution of the null distribution of the likelihood ratio statistic in the case were there was equal variance in each component of mixture of normal distributions. This approach for our purposes was modified so that we accounted for differences in the variances for each component. It should be stated that the method of Chen can not be applied directly to the problem of heteroscedasticity, that is, in the case where the variances are different in each component which is the case used in this dissertation. Therefore, the asymptotic distribution of the penalized modified likelihood method used in this dissertation, will be estimated using the regression model of Thode et al. (1988). The theoretical distribution of the penalized modified likelihood ratio statistic in the case of unequal variances for each component is an open problem which I hope to solve in the near future. The next section describes the method of Chen which is the method in this dissertation we modified to account for heteroscedasticity (unequal component variances).

3.3.2 Modified Likelihood Ratio Test for Homogeneity in Finite Mixture Models

Finite mixture models are often used to study data from a population that is suspected to be composed of a number of homogeneous sub-populations. For example, when a disease has a simple genetic cause, the population may be divided into two or three homogeneous groups. In the initial stage of these investigations, it is important to have a sensitive test for the number g of sub-populations included in the data. The construction of such a test, however, is often more challenging than might be expected.

Chen and Kalbfleisch (1996), Chen (1998) and Chen et al. (2001, 2002) suggest a modification of the likelihood by incorporating a penalty term that forces certain estimates away from the boundary of the parameter space. The likelihood ratio statistic based on the modified estimators is shown, in many instances, to yield relatively simpler limiting distributions and hence simpler tests.

We consider a finite mixture distribution with probability density function as defined in (3.1.1), i.e.,

$$f(Y|\psi) = \sum_{j=1}^{g} \pi_j f(y|\theta_j)$$

where $f(y|\theta_j)$, is a probability density function with parameter $\theta_j \in \Theta$. Let $\theta_1, \ldots, \theta_g \in \Theta$ be the support points of $f(y|\theta_j)$ and let $\pi_1, \ldots, \pi_g$ be the corresponding weights with $\pi_j \geq 0$ and $\sum \pi_j = 1$. If we consider $g = 2$ then we have $\pi f(y|\theta_1) + (1-\pi) f(y|\theta_2)$ where $\pi \in [0,1]$ and $\theta_1 \leq \theta_2$. We wish to test the hypothesis

$$H_0 : \theta_1 = \theta_2 \quad \text{versus} \quad H_0 : \theta_1 \neq \theta_2$$

however the parameters under the null is not identifiable. Therefore, Chen penalized the log-likelihood, hence the modified likelihood approach is given by

$$l_n^*(\psi|y) = \tilde{l}_n(\psi|y) + C \ln 4\pi(1-\pi). \tag{3.3.18}$$

where C is a positive constant and

$$\tilde{l}_n(\psi|y) = \sum_{i=1}^{n} \ln \left\{ \sum_{j=1}^{2} \pi_j f(y|\theta_j) \right\}. \tag{3.3.19}$$

is the ordinary log likelihood. The purpose of the "penalty term", $C \ln 4\pi(1-\pi)$ in (3.3.18) is to restore regularity to the problem by avoiding estimates of π on or near the boundary. The modified likelihood ratio statistic is thus

$$R_n^* = 2\{l_n^*(\hat{\pi}, \hat{\theta}_1, \hat{\theta}_2) - l_n^*(1/2, \hat{\theta}, \hat{\theta})\}. \tag{3.3.20}$$

and the null distribution is given by

$$\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2.$$

The finite mixture distribution (3.1.1) can also be written as

$$f(y|G) = \int f(y|\theta) dG(\theta), \tag{3.3.21}$$

where $G(\theta)$ is a discrete cumulative distribution function (called the mixing distribution) with a finite number of support points. The class of all finite mixing distributions with g support points is

$$\mathsf{M}_g = \Big\{ G(\theta) = \sum_{j=1}^{g} \pi_j I(\theta_j \leq \theta) : \theta_1 \leq, \ldots, \leq \theta_g, \sum_{j=1}^{g} \pi_j = 1, \pi_j \geq 0 \Big\} \tag{3.3.22}$$

where $I(\cdot)$ is an indicator function and $g = 1, 2, \ldots$. The class of all finite mixing distributions is $\mathsf{M} = \bigcup_{g \geq 1} \mathsf{M}_g$.

We consider the test with null hypothesis $g = 1$ versus the alternative $g \geq 2$; or more precisely we consider a test of the hypothesis $G \in \mathsf{M}_1$ versus $G \in \mathsf{M}_{g \geq 2}$. Furthermore, let $\hat{G}_0$ and $\hat{G}_1$ denote the estimates under the null and alternate hypothesis respectively, hence the modified likelihood ratio statistic for testing $G \in \mathsf{M}_1$ against $G \in \mathsf{M}_{g \geq 2}$ is given by

$$R_n^* = 2\{l_n^*(\hat{G}_1) - l_n^*(\hat{G}_0)\}$$

where

$$l_n^*(\psi|y) = \sum_{i=1}^{n} \ln \left\{ \sum_{j=1}^{g} \pi_j f(y|\theta_j) \right\} + C \sum_{j=1}^{g} \ln(g\pi_j). \tag{3.3.23}$$

The Theorems below summarize the above arguments.

Theorem 3.3.1 *If the regularity conditions hold (c.f. Chen et al. 2001), the asymptotic null distribution of the modified LRT statistics*

$$R_n^* = 2\{l_n^*(\hat{G}_1) - l_n^*(\hat{G}_0)\}$$

for testing $G \in \mathsf{M}_1$ against $G \in \mathsf{M}_{g \geq 2}$, is the mixture of χ_1^2 and χ_0^2 with equal weights, i.e.

$$\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$$

where χ_0^2 is a degenerate distribution with all its mass at 0.

Constructing a test of the hypothesis $G \in \mathsf{M}_2$ or $(g = 2)$ is similar in principle to $g = 1$ but perhaps because of its mathematical complexity has a less extensive literature. Some approaches can be found in the diagnostic method of Roeder (1994) and Lindsay and Roeder (1997), and model selection approach (Chen and Kalbfleisch, 1996: Henna, 1985).

Theorem 3.3.2 *(Chen et al. 2004) If the regularity conditions hold, and the true distribution is a 2-component model. Then the asymptotic null distribution of the modified LRT statistics*

$$R_n^* = 2\{l_n^*(\hat{G}_1) - l_n^*(\hat{G}_0)\}$$

for testing $G \in \mathsf{M}_2$ against $G \in \mathsf{M}_{k\geq 2}$, is the mixture of

$$\Big(\frac{1}{2} - \frac{\alpha}{2\pi}\Big)\chi_0^2 + \frac{1}{2}\chi_1^2 + \frac{\alpha}{2\pi}\chi_2^2,$$

where $\alpha = cos^{-1}(\rho)$, ρ is the correlation coefficient between the two components of the null hypothesis and χ_0^2 is a degenerate distribution with all its mass at 0.

One of the important issues of this dissertation is to obtain the asymptotic null distribution of the likelihood ratio tests for the penalized modified mixture model and the modified p-value approach. Note that in the case of the penalized modified mixture model both the mixing proportion and the variance parameters are simultaneous penalized, therefore changing the assumptions of Theorems 3.3.1 and 3.3.2. Since the assumptions of Theorems 3.3.1 and 3.3.2 were not satisfied we determined the asymptotic null distribution of the likelihood ratio test by simulation.

For the modified p-value approach the assumption that the mixing distribution is from the exponential family has been violated since the beta distribution is not of the exponential family. To this end, the asymptotic null distribution of the likelihood ratio statistic will be determine by simulation.

Note that the asymptotic null distribution is absolutely necessary so that we can carry out a hypothesis test to determine the number of components of the mixture model.

3.3.3 Regularity Conditions

Suppose that $Y_1, \ldots, Y_n$ is an independent and identically distributed sample from (3.3.21), and suppose that (3.1.1) is identifiable in the sense that $f(y|G_1) = f(y|G_2)$, for all y, implies $G_1 = G_2$. We consider the hypothesis $H_0 : G \in \mathsf{M}_g$, $(g = 1 \text{ or } 2)$. We assume throughout that the true mixing distribution is

$$G_0 = \sum_{j=1}^{g} \pi_j I(\theta_{0j} \leq \theta), \quad (g = 1 \text{ or } 2), \tag{3.3.24}$$

where $\theta_{0j}, (j = 1, 2)$ are distinct interior points of Θ and $0 < \pi_0 < 1$. All expectation and probabilities are with respect to this null distribution. We also assume that the distance between two mixing distributions G and Q is measured by the supremum distance, i.e.,

$$|G - Q| = \sup_{\theta} |G(\theta) - Q(\theta)|.$$

Condition 1 *Wald's integrability conditions.*
The function $f(y|\theta)$ satisfies Wald's integrability conditions for consistency of the maximum likelihood estimation, i.e. for each $\theta \in \Theta$, (i) $E|\log f(y|G_0)| < \infty$, and (ii) there exists $\rho > 0$ such that $E[\log f(y|G, \rho)] < \infty$, where

$$f(x, |G, \rho) = 1 + \sup_{|G-Q| \leq \rho} \{f(y|Q)\}.$$

Condition 2 *Smoothness.*
The function $f(y|\theta)$ has common support and is three times continuously differentiable with respect to θ. The first three derivatives are denoted by $f'(y|\theta)$ $f''(y|\theta)$ and $f'''(y|\theta)$, respectively.

Condition 3 *Strong identifiability.*
For any $\theta_1 \neq \theta_2 \in \Theta$,

$$\sum_{j=1}^{2} \{a_j f(y|\theta_j) + b_j f'(y|\theta_j) + c_j f''(y|\theta_j)\} = 0, \quad \forall x,$$

implies that $a_j = b_j = c_j = 0, j = 1, 2$.

Condition 4 *Uniform strong law condition of large numbers.*
There exists integrable g with some $\delta > 0$ such that $|X_i(\theta)|^{4+\delta} \leq g(Y_i)$, $|X_i'(\theta)|^3 \leq g(Y_i)$, $|X_i''(\theta)|^3 \leq g(Y_i)$ and $|X_i'''(\theta)|^3 \leq g(Y_i)$ $\forall \theta \in \Theta$, where for $i = 1, \ldots, n$ and $j = 1, 2$ we define

$$\begin{aligned} X_{ij}(\theta) &= \frac{f(Y_i|\theta) - f(Y_i|\theta_{0j})}{f(Y_i|G_0)}, \qquad X_i'(\theta) = \frac{f'(Y_i|\theta)}{f(Y_i|G_0)} \\ X_i''(\theta_0) &= \frac{f''(Y_i|\theta_0)}{f(X_i|G_0)}, \qquad X_i'''(\theta) = \frac{f'''(Y_i|\theta)}{f(Y_i|G_0)}. \end{aligned} \tag{3.3.25}$$

Condition 5 *Tightness.*
For $j = 1, 2$ the processes

$$\frac{\sum X_{ij}(\theta)}{n^{1/2}}, \quad \frac{\sum X_i'(\theta)}{n^{1/2}}, \quad \frac{\sum X_i''(\theta)}{n^{1/2}} \quad \textit{and} \quad \frac{\sum Y_i'''(\theta)}{n^{1/2}}$$

are tight

The tightness condition ensures the weak converges of the process.

In the next section we will describe the Box-Cox transformation that is used to distinguish skewed from commingled distribution in mixture models. Note in this dissertation we did not account for skewness as was the case of the regression method of Thode et al. (1988). However, it is important to the reader to be aware that in mixture distribution we can normalize mixture of any distributions, that is to mixture of normal distributions if that need arises.

3.4 Box-Cox transformation

One challenge in applying mixture models is the difficulty of distinguishing commingled distributions from distribution that are skewed. MacLean et al. (1976) proposed a likelihood ratio test to distinguish skewness from commingled distributions, using the Box-Cox transformation (Box and Cox (1964)) to eliminate skewness for each of the hypothesis to be tested. The hypothesis to be test is that the transformed data is from one normal or a mixture of normal homoscedastic distributions. The Box-Cox transformation will now be presented.

Let $Y_1, \ldots, Y_n$ be a random sample which has been standardized to mean 0 and variance 1, the Box-Cox type transformation is then applied with the power parameter λ, where

$$z = g(y) = \begin{cases} \frac{r}{\lambda}\Big[\Big(\frac{y}{r}+1\Big)^{\lambda}-1\Big], & \text{when } \lambda \neq 0 \\ r\ln\Big(\frac{y}{r}+1\Big), & \text{when } \lambda = 0 \end{cases} \tag{3.4.26}$$

The scale parameter r is necessary only to ensure that every $\frac{y}{r}+1$ is positive in the sample, however it slightly affect the distribution of Y. MacLean et al. (1976) suggested, using a fixed value of r because, while simultaneous estimation of r and λ might improve the approximation to normality, it might exacerbate convergence problems.

In the case of a 2-component normal mixture model given by

$$f(y) = \pi N(\mu_1, \sigma) + (1-\pi)N(\mu_2, \sigma) \tag{3.4.27}$$

The MLE's of the parameters $\pi, \mu_1, \mu_2, \sigma$, and λ are estimated iteratively by maximizing the log likelihood function

$$\begin{aligned} l(y) = &\sum_{i=1}^{n} \ln\Big(\frac{y}{r}+1\Big)^{\lambda-1} + n\ln\sigma \\ &\sum_{i=1}^{n} \ln\Big[\pi\exp\Big\{-\frac{(z_i-\mu_1)^2}{2\sigma^2}\Big\} + (1-\pi)\exp\Big\{-\frac{(z_i-\mu_2)^2}{2\sigma^2}\Big\}\Big] \end{aligned} \tag{3.4.28}$$

where $z = \frac{r}{\lambda}\Big[\Big(\frac{y}{r}+1\Big)^{\lambda}-1\Big]$.

Note that after the Box-Cox transformation has been applied the data is now either normally distributed or is a mixture of normal distributions see MacLean et al. (1976) for detail.

3.5 Conclusion

The fundamental theory of mixture models was discussed in chapter 3. We illustrated how to determine the parameters of the mixture model by: (1) the expectation maximization (EM) algorithm and (2) the robust parameter estimation approaches. Furthermore the parameters of the normal mixture model with unequal variances was discussed. The method of Ciuperca et al. the penalized likelihood for normal mixture was perused.

One of the many challenges for researchers in the field in the field of mixture distribution is to determine the number of components. The model selection criteria BIC and AIC were discussed, however, for mixture distribution there has not been any theoretical justification for their use. Therefore simulation and the modified likelihood ratio test are methods that had no such theoretical drawback. All three approaches were discussed in this chapter, with the modified likelihood ratio test used in this dissertation to determine the number of components for the mixture models. Note the asymptotic null distribution for the modified likelihood test is done by means of simulation.

In some cases in mixture distribution researchers may not be able to distinguish commingled distributions from distribution that are skewed. In this situation, the likelihood ratio test to distinguish skewness from commingled distributions, using the Box-Cox transformation to eliminate skewness for each of the hypothesis to be tested is one available method. Throughout this dissertation we assume that mixture distribution is distinguishable from skewed distribution, therefore we need not apply the Box-Cox transformation.

4 The Penalized Modified Likelihood for Normal Mixture Model

In chapter 3 we introduce both the penalized likelihood approach and modified likelihood approach. The main reason for the penalization of the variance as discussed in chapter 3 was that the log likelihood will be bounded guaranteeing the existence of the MLE where normal mixture models with unequal variances needed to be implemented. The modification for the mixing proportion was done so that the estimates will not be on the boundary point of its parameter space and more importantly the resulting modified likelihood ratio test statistic will enjoy the simple χ^2-type null limiting distribution.

In this chapter one of our major contribution is the building of a model with both the above mentioned capabilities, that is, we penalize both the mixing proportion and the variance parameters simultaneously. Therefore, we are able to fit normal mixture models with unequal variances and be able to conduct a test of hypothesis for the number of components that characterizes the model.

Firstly, estimators for the parameters of the penalized modified likelihood approach will be illustrated. These estimators are necessary so that we can implement the expectation maximization algorithm when simulating the null distribution for the likelihood ratio statistic (LRTS) for the penalized normal mixture model. Another major contribution in this dissertation is that we proved asymptotic normality of the MLE's (estimators) for the penalized normal mixture model. Asymptotic normality of the MLE's (estimators) is a major contribution of this dissertation and is a first step to determine the asymptotic null distribution of the likelihood ratio test statistic (which is an open problem).

4.1 Penalized Modified Likelihood

Let $Y_1, \ldots, Y_n$ be a random sample of size n from the mixture model

$$f_1(Y|\psi) = \sum_{j=1}^{g} \pi_j f_{ij}(y_i) \tag{4.1.1}$$

where

$$f_{ij}(y_i) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left\{ -\frac{1}{2\sigma_j^2}(y_i - \mu_j)^2 \right\} \quad j = 1, \ldots, g$$

are normal densities with mean μ_j and standard deviation σ_j. The parameter set of the mixture is given as

$$\Psi = (\pi_1, \ldots, \pi_j, \mu_1, \ldots, \mu_j, \sigma_1, \ldots, \sigma_j) \tag{4.1.2}$$

such that $0 \leq \pi_j \leq 1, \sum_{j=1}^{g} \pi_j = 1, -\infty < \mu_j < \infty$, $\sigma_j > 0$ and the true parameters defined as $\psi_0 \in \Psi$. The penalized modified likelihood for a g-component normal mixture model is given by

$$\mathcal{L}_n(\psi|y) = \prod_{i=1}^{n} \sum_{j=1}^{g} \pi_i f_{ij}(y_i|\theta) \prod_{j=1}^{g} h(\sigma_j) \prod_{j=1}^{g} (g\pi_j)^C \tag{4.1.3}$$

for the observed data, where C is a positive constant that control the level of modification of the mixing proportion π_j (the last term of equation (4.1.3)). The function h as mentioned in the previous chapter, was chosen so that $\mathcal{L}_n$ is bounded over the parameter space Ψ. The penalized function h was assumed to have satisfied the following conditions:

(1) $\lim_{\sigma \to 0} \frac{1}{\sigma^n} h(\sigma) = 0$, for all n, which ensures that for any fixed n, the maximum argument of $\mathcal{L}_n$, that is the penalized MLE

$$\arg\max_{\psi \in \Psi} \mathcal{L}_n \text{ exists.}$$

The consistency of the estimator was also a concern. In order to prove the consistency it was required that h also satisfied the following conditions:

(2) $h(\sigma)$ is many-to-one from $(0, \infty)$ onto $(0, G]$, $G > 0$,

(3) h is strictly increasing in an open interval $(0, \delta)$ of the origin which has a

non-null measure,

(4) h is continuously differentiable on $(0, \infty)$.

In this dissertation we consider two distributions that satisfy the aforementioned conditions on the penalized function $h(\sigma)$ for the variance. These distributions are (1) the inverse gamma and (2) the inverse chi-square distributions.

In the next section we will evaluate the estimators for the penalized modified likelihood for normal mixture models. These estimates are vital because in chapter 5 we used these estimators in the expectation maximization algorithm to evaluate the log likelihood which is then used in the simulation of the asymptotic null distribution of the modified likelihood ratio test, see section 5.2 of chapter 5.

4.2 Parameter Estimation of Penalized Modified Likelihood

The penalized modified likelihood for a g-component normal mixture model is given by (4.1.3) for the observed data. Furthermore, the likelihood for the complete data is given by (c.f. section 3.2 chapter 3)

$$\mathcal{L}_n(\psi|y, Z) = \prod_{i=1}^{n}\prod_{j=1}^{g}\left[\pi_i f_{ij}(y_i|\theta)\right]^{z_{ij}}\prod_{j=1}^{g}h(\sigma_j)\prod_{j=1}^{g}(g\pi_j)^C,$$

and the complete log-likelihood is

$$l_n(\psi|y, Z) = \sum_{i=1}^{n}\sum_{j=1}^{g}z_{ij}\left[\ln \pi_j + \ln f_{ij}(y_i|\theta)\right] + \sum_{j=1}^{g}\ln h(\sigma_j) + C\sum_{j=1}^{g}\ln(g\pi_j). \qquad (4.2.4)$$

Similar to the approach in section 3.2, we need only to maximize the expectation of the log-likelihood

$$Q(\psi|\psi^{(t)}) = E\left[l_n(\psi|y, Z)\Big|y, \psi^{(t)}\right].$$

Note that the E-step resulted in

$$\pi_{ij}^{(t)} = \frac{\pi_j f_{ij}(y_i|\theta)}{\sum_{j=1}^{g}\pi_j f_{ij}(y_i|\theta)} \qquad (4.2.5)$$

hence, the log likelihood is

$$\sum_{i=1}^{n}\sum_{j=1}^{g}\left[\pi_{ij}^{(t)}\ln\pi_j+\pi_{ij}^{(t)}\ln f_{ij}(y_i|\theta)\right]+\sum_{j=1}^{g}\ln h(\sigma_j)+C\sum_{j=1}^{g}\ln(g\pi_j)$$

$$=\sum_{j=1}^{g}\left[\sum_{i=1}^{n}\pi_{ij}^{(t)}\ln f_{ij}(y_i|\theta)+\ln h(\sigma_j)\right]+\sum_{j=1}^{g}\left[\sum_{i=1}^{n}\pi_{ij}^{(t)}\ln\pi_j+C\ln(g\pi_j)\right] \qquad (4.2.6)$$

Now we maximize with respect to π_j, therefore we consider the last term of equation (4.2.6), since

$$\sum_{j=1}^{g}\left[\sum_{i=1}^{n}\pi_{ij}^{(t)}\ln\pi_j+C\ln(g\pi_j)\right]\propto\sum_{j=1}^{g}\left[\sum_{i=1}^{n}\pi_{ij}^{(t)}\ln\pi_j+C\ln(\pi_j)\right]$$

therefore we have

$$\begin{aligned}
&\sum_{j=1}^{g}\left[\sum_{i=1}^{n}\pi_{ij}^{(t)}\ln(\pi_j)+C\ln(\pi_j)\right]\\
=\quad&\sum_{j=1}^{g}\left[\left(\sum_{i=1}^{n}\pi_{ij}^{(t)}+C\right)\ln(\pi_j)\right]\\
=\quad&\sum_{j=1}^{g-1}\left[\left(\sum_{i=1}^{n}\pi_{ij}^{(t)}+C\right)\ln(\pi_j)\right]+\left(\sum_{i=1}^{n}\pi_{ig}^{(t)}+C\right)\ln\left(1-\sum_{j=1}^{g-1}\pi_j\right) \qquad (4.2.7)
\end{aligned}$$

then taking the derivative w.r.t π_j of equation (4.2.7) and then equating to 0 we get

$$\begin{aligned}
&\frac{\sum_{i=1}^{n}\pi_{ij}^{(t)}+C}{\pi_j}=\frac{\sum_{i=1}^{n}\pi_{ig}^{(t)}+C}{\pi_g}\\
\Rightarrow\quad&\frac{\pi_j}{\pi_g}=\frac{\sum_{i=1}^{n}\pi_{ij}^{(t)}+C}{\sum_{i=1}^{n}\pi_{ig}^{(t)}+C}\\
\Rightarrow\quad&\frac{\sum_{j=1}^{g}\pi_j}{\pi_g}=\frac{\sum_{j=1}^{g}\left(\sum_{i=1}^{n}\pi_{ij}^{(t)}+C\right)}{\sum_{i=1}^{n}\pi_{ig}^{(t)}+C}\\
\Rightarrow\quad&\frac{1}{\pi_g}=\frac{n+gC}{\sum_{i=1}^{n}\pi_{ig}^{(t)}+C}\\
\Rightarrow\quad&\pi_g=\frac{\sum_{i=1}^{n}\pi_{ig}^{(t)}+C}{n+gC}.
\end{aligned}$$

It follows that all the $\pi_j^{(t+1)}$ are given by

$$\pi_j^{(t+1)} = \frac{\sum_{i=1}^n \pi_{ig}^{(t)} + C}{n + gC} \tag{4.2.8}$$

For normal mixture i.e f_{ij} is normally distributed we have that

$$\ln f_{ij}(x|\mu_j, \sigma_j^2) \propto -\frac{\ln(\sigma_j^2)}{2} - \frac{(y_i - \mu_j)^2}{2\sigma_j^2}.$$

Furthermore maximizing equation (4.2.7) w.r.t. μ results in

$$\sum_{j=1}^g \sum_{i=1}^n \pi_{ij}^{(t)} \left[-\frac{\ln(\sigma_j^2)}{2} - \frac{(y_i - \mu_j)^2}{2\sigma_j^2} \right]$$

to be

$$\sum_{i=1}^n \pi_{ij}^{(t)} (y_i - \mu_j) = 0$$

therefore the estimate for μ_j is given by

$$\mu_j^{(t+1)} = \frac{\sum_{i=1}^n \pi_{ij}^{(t)} y_i}{\sum_{i=1}^n \pi_{ij}^{(t)}}. \tag{4.2.9}$$

We now turn our attention to maximizing the variance of the normal mixture model when the inverse gamma function is used as the penalty function.

4.2.1 Inverse Gamma Penalty Function for σ

In this section we will find the estimate for the variance when the inverse gamma function is used as the penalty function. The inverse gamma distribution for the penalty term for the variance σ_j^2 is

$$h(\sigma_j) = \frac{\alpha^\beta}{\Gamma(\beta)\sigma_j^{2(\beta+1)}} \exp(-\frac{\alpha}{\sigma_j^2}),$$

therefore

$$\ln h(\sigma_j) \propto -(\beta + 1)\ln(\sigma_j^2) - \frac{\alpha}{\sigma_j^2}.$$

We therefore find the derivative w.r.t. σ_j^2 of equation (4.2.10) then equating to zero,

$$\sum_{j=1}^{g}\left\{\sum_{i=1}^{n}\pi_{ij}^{(t)}\left[-\frac{\ln(\sigma_j^2)}{2}-\frac{(y_i-\mu_j)^2}{2\sigma_j^2}\right]-(\beta+1)\ln(\sigma_j^2)-\frac{\alpha}{\sigma_j^2}\right\}. \tag{4.2.10}$$

Thus we have

$$\sum_{i=1}^{n}\pi_{ij}^{(t)}\left[-\frac{1}{\sigma_j^2}+\frac{(y_i-\mu_j)^2}{\sigma_j^4}\right]-\frac{2(\beta+1)}{\sigma_j^2}+\frac{2\alpha}{\sigma_j^4}=0$$

$$\Rightarrow\frac{\sum_{i=1}^{n}\pi_{ij}^{(t)}(y_i-\mu_j)^2}{\sigma_j^2}+\frac{2\alpha}{\sigma_j^2}=\sum_{i=1}^{n}\pi_{ij}^{(t)}+2(\beta+1)$$

$$\Rightarrow(\sigma_j^2)^{(t+1)}=\frac{\sum_{i=1}^{n}\pi_{ij}^{(t)}(y_i-\mu_j)^2+2\alpha}{\sum_{i=1}^{n}\pi_{ij}^{(t)}+2(\beta+1)}. \tag{4.2.11}$$

To estimate the null parameters we maximize the log likelihood under the null which is given by

$$\sum_{i=1}^{n}\ln f_{ij}(y_i|\mu,\sigma^2)+\sum_{j=1}^{g}\ln h(\sigma).$$

For μ, we have that

$$\frac{\partial}{\partial\mu}\left\{\sum_{i=1}^{n}\left[-\frac{\ln(\sigma^2)}{2}-\frac{(y_i-\mu)^2}{2\sigma^2}\right]\right\}=0$$

to be

$$\sum_{i=1}^{n}(y_i-\mu)=0$$

therefore

$$\hat{\mu}=\frac{\sum_{i=1}^{n}y_i}{n}. \tag{4.2.12}$$

Under the null hypothesis the estimate of σ^2 is evaluated as follow. Using the inverse gamma penalty term, that is $h(\sigma)=\frac{\alpha^\beta}{\Gamma(\beta)\sigma^{2(\beta+1)}}\exp(-\frac{\alpha}{\sigma^2})$, therefore
$\ln h(\sigma)\propto-(\beta+1)\ln(\sigma^2)-\frac{\alpha}{\sigma^2}$. We want the derivative w.r.t. σ^2 of the following

$$\sum_{i=1}^{n}\left[-\frac{\ln(\sigma^2)}{2}-\frac{(y_i-\mu)^2}{2\sigma^2}\right]-(\beta+1)\ln(\sigma^2)-\frac{\alpha}{\sigma^2}$$

then equating the derivative to zero, resulted in

$$\sum_{i=1}^{n}\left[-\frac{1}{\sigma^2}+\frac{(y_i-\mu)^2}{\sigma^4}\right]-\frac{2(\beta+1)}{\sigma^2}+\frac{2\alpha}{\sigma^4}=0$$

$$\Rightarrow \frac{\sum_{i=1}^{n}(y_i-\mu)^2}{\sigma^2}+\frac{2\alpha}{\sigma^2}=n+2(\beta+1)$$

$$\Rightarrow \hat{\sigma}^2=\frac{\sum_{i=1}^{n}(y_i-\mu)^2+2\alpha}{n+2(\beta+1)} \tag{4.2.13}$$

The other penalty term of interest is the inverse chi-square distribution, which in the next section has be used in the evaluation of the MLE for variance parameter of the normal mixture model.

4.2.2 Inverse Chi-Square Penalty Function for σ

The inverse chi-square distribution for the penalty term for σ_j^2 is

$$h(\sigma_j)=\frac{2^{\nu/2}}{\Gamma(\nu/2)\sigma_j^{2(\nu/2+1)}}\exp\Big(-\frac{1}{2\sigma_j^2}\Big),$$

therefore

$$\ln h(\sigma_j)\propto -(\nu/2+1)\ln(\sigma_j^2)-\frac{1}{2\sigma_j^2}$$

We want the derivative w.r.t. σ_j^2 of the following

$$\sum_{j=1}^{g}\left\{\sum_{i=1}^{n}\pi_{ij}^{(t)}\left[-\frac{\ln(\sigma_j^2)}{2}-\frac{(y_i-\mu_j)^2}{2\sigma_j^2}\right]-(\nu/2+1)\ln(\sigma_j^2)-\frac{1}{2\sigma_j^2}\right\}.$$

After taking the derivative and equating to zero, we have that

$$\sum_{i=1}^{n}\pi_{ij}^{(t)}\left[-\frac{1}{\sigma_j^2}+\frac{(y_i-\mu_j)^2}{\sigma_j^4}\right]-\frac{(\nu/2+1)}{\sigma_j^2}+\frac{1}{2\sigma_j^4}=0$$

$$\Rightarrow \frac{\sum_{i=1}^{n}\pi_{ij}^{(t)}(y_i-\mu_j)^2}{\sigma_j^2}+\frac{1}{2\sigma_j^2}=\sum_{i=1}^{n}\pi_{ij}^{(t)}+(\nu/2+1)$$

$$\Rightarrow (\sigma_j^2)^{(t+1)} = \frac{\sum_{i=1}^n \pi_{ij}^{(t)}(y_i - \mu_j)^2 + 1/2}{\sum_{i=1}^n \pi_{ij}^{(t)} + (\nu/2+1)} \tag{4.2.14}$$

To estimate the null parameters we maximize the log likelihood under the null, given by

$$\sum_{i=1}^n \ln f_{ij}(y_i|\mu, \sigma^2) + \sum_{j=1}^g \ln h(\sigma).$$

Maximizing

$$\sum_{i=1}^n \left[-\frac{\ln(\sigma^2)}{2} - \frac{(y_i-\mu)^2}{2\sigma^2} \right]$$

w.r.t. μ, we have that

$$\sum_{i=1}^n (y_i - \mu) = 0.$$

The result for μ is similar to that of equation (4.2.12), i.e.,

$$\hat{\mu} = \frac{\sum_{i=1}^n y_i}{n}.$$

In the case of σ^2 using the inverse chi-square penalty term i.e., $h(\sigma) = \frac{2^{\nu/2}}{\Gamma(\nu/2)\sigma^{2(\nu/2+1)}} \exp(-\frac{1}{2\sigma^2})$, therefore $\ln h(\sigma) \propto -(\nu/2+1)\ln(\sigma^2) - \frac{1}{2\sigma^2}$. Therefore taking the derivative w.r.t. σ^2 of equation (4.2.15) and equation to zero,

$$\sum_{i=1}^n \left[-\frac{\ln(\sigma^2)}{2} - \frac{(y_i-\mu)^2}{2\sigma^2} \right] - (\nu/2+1)\ln(\sigma^2) - \frac{1}{2\sigma^2} \tag{4.2.15}$$

we have

$$\sum_{i=1}^n \left[-\frac{1}{\sigma^2} + \frac{(y_i-\mu)^2}{\sigma^4} \right] - \frac{(\nu/2+1)}{\sigma^2} + \frac{1}{2\sigma^4} = 0$$

$$\Rightarrow \frac{\sum_{i=1}^n (y_i-\mu)^2}{\sigma^2} + \frac{1}{2\sigma^2} = n + (\nu/2+1)$$

$$\Rightarrow \hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i-\mu)^2 + 1/2}{n + (\nu/2+1)} \tag{4.2.16}$$

4.3 Consistency and Asymptotic Normality

Let $Y_1, \ldots, Y_n$ be a random sample of size n from the mixture model with density given by (4.1.1), where the parameters $\psi \in \Psi$ defined in (4.1.2) and $\bar{\Psi}$ denote the closure of set Ψ. From chapter 3 Example 9 we illustrated that the likelihood function is unbounded on Ψ. This was circumvented by adding a penalty term for the variance parameter with the properties mentioned in section 1 of this chapter. From Redner (1980), we know that if a likelihood function has a strongly consistent maximizer over a compact set, then penalizing it with a penalty term that is continuously differentiable and that has a bounded logarithm, does not alter its asymptotic property. G. Ciuperca et al. (2003) stated that Redner's results can be applied on every compact set that excludes a neighbourhood of $\sigma = 0$. However, this resulted in considering the problem in a neighbourhood of the origin of the parameters σ_j, where the MLE does not exist and, therefore, Redner's property does not apply.

Consequently G. Ciuperca et al. (2003) focused their study of the asymptotic properties in a neighbourhood of the origin of the parameters σ_j. In this section we applied their idea to prove that there exists a constant $\eta > 0$, not dependent on n, so that the probability that the penalized modified likelihood $\mathcal{L}_n$ is maximized by a $\sigma_j \in [0, \eta)$ is zero. Similar to the approach in G. Ciuperca et al. (2003), from (4.1.3) we consider $\mathcal{L}_n$ and extended its definition to $\bar{\Psi}$, i.e,

$$\mathcal{L}_n = \begin{cases} 0 & \text{if } \sigma_k = 0, \infty \text{ or } \mu_k = \pm\infty \\ f_n(Y_1, \ldots, Y_n|\psi) \prod_{j=1}^{g} h(\sigma_j) \prod_{j=1}^{g} (g\pi_j)^C & \text{otherwise,} \end{cases}$$

where $f_1(Y|\psi)$ is a mixture of normal distributions (definition 4.1.1), $f_n(Y_1, \ldots, Y_n|\psi) = \prod_{i=1}^{n} f_1(Y|\psi)$ the ordinary likelihood and $1 \leq k \leq g$. Let

$$\psi_0 = (\pi_{01}, \ldots, \pi_{0j}, \mu_{01}, \ldots, \mu_{0j}, \sigma_{01}, \ldots, \sigma_{0j}) \in \Psi$$

be the true value of the parameter and let us define the Banach space

$$H = \mathbf{L}^1(f_1(y, \psi_0))$$

where $\mathbf{L}^1$ is a linear space such that for a function $f_1 \in \mathbf{L}^1$ we define

$$||f_1(y, \psi)|| = \int |f_1|.$$

Note that the operator E_H denotes the expectation in the space H. The reason for introducing a Banach space will be clear from the definition below.

Definition 4.3.1 *A normed linear space is called* **complete** *if every Cauchy sequence in the space converges, that is, if for each Cauchy sequence $\{a_n\}$ in the space there is an element a in the space such that $a_n \to a$. A complete normal linear space is called a* **Banach space**.

Therefore from definition 4.3.1 we have that the expectation E_H will be finite.

4.3.1 Preliminary Results

In this section we will present preliminary results along with their proofs, that will be useful to prove asymptotic normality of the modified penalized method. The results presented in this section are similar to work presented in G. Ciuperca (2003) which have been slightly modified for our approach. First we consider a random variable Y with density $f_1(y|\psi_0)$, then the following Lemmas hold:

Lemma 4.3.2 *(c.f. [9]) If $\{\psi_m\} \in \bar{\Psi}$ and $\psi^* \in \bar{\Psi}$ is such that $\lim_{m\to\infty} \psi_m = \psi^*$, then*

$$\mathcal{L}_1(y|\psi_m) \to \mathcal{L}_1(y|\psi^*), \quad \textit{as } m \to \infty$$

Lemma (4.3.3) is similar to that stated in [9] with exception that we penalized both the mixing proportion and the variance parameters. We present our proof which accounts for the addition mixing proportion which is the major difference to that prosented in [9].

Lemma 4.3.3 *(c.f. [9]) There exists $\eta > 0$ with the property*

$$\eta < \sigma_{0j} \quad \forall j = 1, \ldots, g \tag{4.3.17}$$

such that

$$E_H[\ln \mathcal{L}_1(Y|\psi)] < E_H[\ln \mathcal{L}_1(Y|\psi_0)], \tag{4.3.18}$$
$$\forall\, \psi \in \bar{\Psi}|\min_{j=1,\ldots,g} \sigma_j \in [0,\eta)$$

Proof. Let $\nu = \ln \mathcal{L}_1(Y|\psi) - \ln \mathcal{L}_1(Y|\psi_0)$, where $\psi \in \bar{\Psi}$. We therefore need to prove that $E_H[\nu] < 0$. Given $\psi \in \Psi$, we have that

$$E_H[e^\nu] = E_H\Big[\frac{\mathcal{L}_1(Y|\psi)}{\mathcal{L}_1(Y|\psi_0)}\Big]$$
$$= \int_{\Re} f_1(y,\psi)\prod_{j=1}^{g}\frac{h(\sigma_j)}{h(\sigma_{0j})}\Big(\frac{\pi_j}{\pi_{0j}}\Big)^C dy = \prod_{j=1}^{g}\frac{h(\sigma_j)}{h(\sigma_{0j})}\Big(\frac{\pi_j}{\pi_{0j}}\Big)^C$$
$$= \prod_{j=1}^{g}\Big(\frac{\pi_j}{\pi_{0j}}\Big)^C\prod_{j=1}^{g}\frac{h(\sigma_j)}{h(\sigma_{0j})} = \kappa\prod_{j=1}^{g}\frac{h(\sigma_j)}{h(\sigma_{0j})}$$

where $\kappa = \prod_{j=1}^{g}\left(\frac{\pi_j}{\pi_{0j}}\right)^C > 0$ is a constant, and we defined a function $w : (0,+\infty) \to (0,\frac{1}{2}]$ to be

$$w(\sigma) = \frac{h(\sigma)}{2G}.$$

Since κ is positive we therefore have that

$$E_H[e^\nu] = \prod_{j=1}^{g}\frac{w(\sigma_j)}{w(\sigma_{0j})},$$

noting that ν is taken such that $w(\nu) = \prod_{j=1}^{g} w(\sigma_{0j})$. Because of the many-to-one character of the function w (see assumption 2 of the penalized function h) then the existence of $\nu \in (0,+\infty)$ is guaranteed. For us to define η and to prove inequality (4.3.3), we considered two cases

1. $\nu \leq \delta$. Then, we set $\eta = \nu$;

2. $\nu > \delta$. Then, if $w(\nu) \leq w(\delta)$, from the on-to-one character of the function w over $(0,\delta)$ (see assumption 3 of the penalty function h) there exists $\eta \in (0,\delta]$ such that $w(\eta) = w(\nu)$. Else, if $w(\nu) > w(\delta)$ we take $\eta = \delta$.

In both cases

$$w(\eta) < w(\sigma_0) \quad \forall j = 1, ..., g \tag{4.3.19}$$

For $\sigma_{0k} > \delta$ and $k \in \{1, \dots, g\}$, we can see that $\eta \leq \sigma_{0k}$. On the hand, i.e. when $\sigma_{0k} \leq \delta$, $k \in \{1, \dots, g\}$, from (4.3.19) we have $\eta < \sigma_{0k}$. Hence it follows that, inequality (4.3.3) holds.

If $\min_{j=1,\dots,g} \sigma_j \in (0, \eta)$, then by taking the definition of w and the assumption (3) on h into account, we have

$$E_H[e^\nu] < \max\Big(1, \frac{w(\min_{j=1,\dots,g} \sigma_j)}{w(\eta)}\Big) = 1$$

where $\psi \in \Psi |\min_{j=1,\dots,g} \sigma_j \in (0, \eta)$. If we now consider the definition by prolongation of Ψ (for $\sigma_j = 0$, $\nu = -\infty$), we get

$$E_H[e^\nu] < 1 \quad \forall\, \psi \in \bar{\Psi} | \min_{j=1,\dots,g} \sigma_j \in (0, \eta)$$

From Lemma (4.3.2) and by noting that $y < e^y$ $\forall y \in \Re$ implies

$$E_H[y] \leq E_H[e^y] \quad \forall\, y \in \Re,$$

we obtain

$$E_H[\nu] \leq E_H[e^\nu] < 1 \quad \forall\, \psi \in \bar{\Psi} | \min_{j=1,\dots,g} \sigma_j \in (0, \eta).$$

Observe that $E_H[\nu] \leq E_H[\ln e^v] < 1$, and since the function $f(y) = \ln y$ is concave, then applying Jensen's inequality we get

$$E_H[\nu] \leq \ln E_H[e^v] < 0.$$

Therefore

$$E_H[\nu] < 0 \quad \forall\, \psi \in \bar{\Psi} | \min_{j=1,\dots,g} \sigma_j \in (0, \eta)$$

therefore the Lemma proved.

Lemma 4.3.3 is important for the proof of consistency of the estimator $\hat{\psi}$ and illustrates that the true state of nature ψ_0 is indeed the global maximum.

For $\psi \in \Psi$ let us define the following functions

$$\begin{cases} w_1(y,\psi,\rho) = \sup_{\psi',|\psi'-\psi|<\rho} \mathcal{L}_1(y,\psi'), \quad \rho > 0 \\ w_n(y_1,\ldots,y_n;\psi,\rho) = \sup_{\psi',|\psi'-\psi|<\rho} \mathcal{L}_n(y_1,\ldots,y_n|\psi') \end{cases}$$

We have the following Lemma

Lemma 4.3.4 *(c.f. [9]) For all $\psi \in \bar{\Psi}$ we have*

$$\lim_{\rho\to 0^+} E_H[\ln w_1(Y|\psi,\rho)] = E_H[\ln \mathcal{L}_1(Y|\psi)] \tag{4.3.20}$$

Let us introduce two results which will be useful to characterize the speed of convergence of the penalized estimator. First, note that since $\pi_g = \sum_{j=1}^{g-1} \pi_j$, the vector ψ contains $3g-1$ parameters

$$\psi = (\pi_1,\ldots,\pi_{g-1},\mu_1,\ldots,\mu_g,\sigma_1,\ldots,\sigma_g)^T$$

These $3g-1$ elements is denoted with $\psi_l, l = 1,\ldots,3g-1$.

Let us define

$$u(Y|\psi) = f_1(Y|\psi)\prod_{j=1}^{g} h(\sigma_j)^{1/n}\prod_{j=1}^{g}(g\pi_j)^{C/n}$$

and let us denote by $h^{(s)}$ the s-order derivative of the penalizing function h. In the following, $\partial/\partial\psi$ will denote the vector of partial derivatives $\partial/\partial\psi_l$, $l = 1,\ldots,3g-1$, with respect to the elements ψ_l, $l = 1,\ldots,3g-1$ of ψ. Therefore, by simple computations, we have the following two Lemmas which are very similar to that presented in [9] with the exception that we have penalized both the mixing proportion and the variance parameters. The proofs are stated accounting for the addition of the penalty term for the mixing proportion.

Lemma 4.3.5 *(c.f. [9]) The means, the variances and the covariances of* $(\partial \ln u(Y|\psi_0)/\partial\psi)$ *are*

$$E_H\Big[\frac{\partial \ln u(Y|\psi_0)}{\partial\psi_l}\Big] = \begin{cases} 0 & if \quad l = 1,\dots,2g-1 \\ \frac{h^{(1)}(\sigma_{0j})}{nh(\sigma_{0j})}, j = 3g-l & if \quad l = 2g,\dots,3g-1 \end{cases}$$

$$var_H\Big[\frac{\partial \ln u(Y|\psi_0)}{\partial\psi_l}\Big] = var_H\Big[\frac{\partial \ln f_1(Y|\psi_0)}{\partial\psi_l}\Big] = E_H\Big[\frac{\partial \ln f_1(Y|\psi_0)}{\partial\psi_l}\Big]^2$$

for all $l = 1,\dots,3g-1$.

$$\begin{aligned} & cov_H\Big[\frac{\partial \ln u(Y|\psi_0)}{\partial\psi_l}, \frac{\partial \ln u(Y|\psi_0)}{\partial\psi_m}\Big] \\ & = E_H\Big[\frac{\partial \ln f_1(Y|\psi_0)}{\partial\psi_l}\frac{\partial \ln f_1(Y|\psi_0)}{\partial\psi_m}\Big] \end{aligned} \qquad (4.3.21)$$

for all $l, m \in \{1,\dots,3g-1\}$, $l \neq m$.

Proof. Since

$$u(Y|\psi) = f_1(Y|\psi)\prod_{j=1}^{g} h(\sigma_j)^{1/n}\prod_{j=1}^{g}(g\pi_j)^{C/n}$$

therefore

$$\begin{aligned} \ln u(Y|\psi) &= \ln f_1(Y|\psi) + 1/n\sum_{j=1}^{g} h(\sigma_j) + C/n\sum_{j=1}^{g}(g\pi_j) \\ &= \ln f_1(Y|\psi) + 1/n\sum_{j=1}^{g} h(\sigma_j) + gC/n \end{aligned} \qquad (4.3.22)$$

and

$$\frac{\partial \ln u(Y|\psi)}{\partial\psi_l} = \begin{cases} \frac{f_1'(Y|\psi)}{f_1(Y|\psi)} & \text{if} \quad l = 1,\dots,2g-1 \\ \frac{f_1'(Y|\psi)}{f_1(Y|\psi)} + \frac{h^{(1)}(\sigma_j)}{nh(\sigma_j)}, j = 3g-l & \text{if} \quad l = 2g,\dots,3g-1. \end{cases}$$

thus we have that

$$E_H\Big[\frac{\partial \ln u(Y|\psi_0)}{\partial\psi_l}\Big] = \begin{cases} 0 & \text{if} \quad l = 1,\dots,2g-1 \\ \frac{h^{(1)}(\sigma_{0j})}{nh(\sigma_{0j})}, j = 3g-l & \text{if} \quad l = 2g,\dots,3g-1. \end{cases}$$

Additionally we have that

$$\frac{\partial \ln f_1(Y|\psi)}{\partial \psi_l} = \frac{f_1'(Y|\psi)}{f_1(Y|\psi)} \quad \text{for} \quad l = 1, \ldots, 3g-1.$$

therefore we have the result that

$$var_H\Big[\frac{\partial \ln u(Y|\psi_0)}{\partial \psi_l}\Big] = var_H\Big[\frac{\partial \ln f_1(Y|\psi_0)}{\partial \psi_l}\Big] = E_H\Big[\frac{\partial \ln f_1(Y|\psi_0)}{\partial \psi_l}\Big]^2$$

for all $l = 1, \ldots, 3g-1$.

From the definition of the covariance the result

$$\begin{aligned} & cov_H\Big[\frac{\partial \ln u(Y|\psi_0)}{\partial \psi_l}, \frac{\partial \ln u(Y|\psi_0)}{\partial \psi_m}\Big] \\ & = E_H\Big[\frac{\partial \ln f_1(Y|\psi_0)}{\partial \psi_l}\frac{\partial \ln f_1(Y|\psi_0)}{\partial \psi_m}\Big] \end{aligned}$$

for all $l, m \in \{1, \ldots, 3g-1\}$, $l \neq m$, immediately follows.

Lemma 4.3.6 *(c.f. [9]) Let* $A = \{(l,l)|l \in \{2g, \ldots, 3g-1\}\}$ *be and index set. Then,* $\forall\, l, m \in \{1, \ldots, 3g-1\}$ *and* $j = 3g - l$ *we have*

$$\begin{aligned} & E_H\Big[-\frac{1}{u^2(Y|\psi_0)}\frac{\partial \ln u(Y|\psi_0)}{\partial \psi_l}\frac{\partial \ln u(Y|\psi_0)}{\partial \psi_m} + \frac{\partial^2 \ln u(Y|\psi_0)}{\partial \psi_l \partial \psi_m}\Big] \\ & = -E_H\Big[\frac{\partial \ln f_1(Y|\psi_0)}{\partial \psi_l}\frac{\partial \ln f_1(Y|\psi_0)}{\partial \psi_m}\Big] \\ & \quad + \frac{1}{n}\Big[\frac{h^{(2)}(\sigma_{0j})}{h(\sigma_{0j})} + \Big(\frac{h^{(1)}(\sigma_{0j})}{h(\sigma_{0j})}\Big)^2\Big]\mathbb{I}_{(l,m)} \in A \end{aligned}$$

Proof. We have that

$$\frac{\partial^2 \ln u(Y|\psi)}{\partial \psi_l \partial \psi_m} = \frac{1}{u(Y|\psi)}\frac{\partial^2 u(Y|\psi)}{\partial \psi_l \partial \psi_m} - \frac{1}{u^2(Y|\psi)}\frac{\partial u(Y|\psi)}{\partial \psi_l}\frac{\partial u(Y|\psi)}{\partial \psi_m}$$

and

$$\frac{\partial^2 \ln u(Y|\psi)}{\partial\psi_l\partial\psi_m} = \begin{cases} \frac{f_1''(Y|\psi)}{f_1(Y|\psi)} - \left(\frac{f_1'(Y|\psi)}{f_1(Y|\psi)}\right)^2 & \text{if} \quad l=1,\dots,2g-1 \\ \frac{f_1''(Y|\psi)}{f_1(Y|\psi)} - \left(\frac{f_1'(Y|\psi)}{f_1(Y|\psi)}\right)^2 & + \ \frac{h^{(2)}(\sigma_j)}{nh(\sigma_j)} - \frac{1}{n}\left(\frac{h^{(1)}(\sigma_j)}{h(\sigma_j)}\right)^2, j=3g-l \\ & \text{if} \quad l=2g,\dots,3g-1. \end{cases}$$

Also we know that p.d.f integrates to 1,

$$\int f_1(Y|\psi) = 1, \tag{4.3.23}$$

and if derivatives of equation (4.3.23) is taken with respect to ψ (and interchange derivative and integral, which can usually be done) we have,

$$\int \frac{\partial}{\partial\psi} f_1(Y|\psi)dY = \int f_1'(Y|\psi)dY = 0$$

and

$$\int \frac{\partial^2}{\partial\psi^2} f_1(Y|\psi)dY = \int f_1''(Y|\psi)dY = 0.$$

Here we show that

$$\int \left[\frac{f_1''(Y|\psi_0)}{f_1(Y|\psi_0)} - \left(\frac{f_1'(Y|\psi_0)}{f_1(Y|\psi_0)}\right)^2\right] f_1(Y|\psi_0)dY = \int f_1''(Y|\psi_0)dY - E_H\left[\frac{f_1'(Y|\psi_0)}{f_1(Y|\psi_0)}\right]^2$$

$$= 0 - E_H\left[\frac{\partial \ln f_1(Y|\psi_0)}{\partial\psi}\right]^2 = -E_H\left[\frac{\partial \ln f_1(Y|\psi_0)}{\partial\psi_l}\frac{\partial \ln f_1(Y|\psi_0)}{\partial\psi_m}\right],$$

and

$$E_H\left[\frac{\partial^2 \ln u(Y|\psi_0)}{\partial\psi_l\partial\psi_m}\right] = \begin{cases} E_H\left[\frac{\partial f_1(Y|\psi_0)}{\partial\psi_l}\frac{\partial f_1(Y|\psi_0)}{\partial\psi_m}\right] & \text{if } l=1,\dots,2g-1 \\ E_H\left[\frac{\partial f_1(Y|\psi_0)}{\partial\psi_l}\frac{\partial f_1(Y|\psi_0)}{\partial\psi_m}\right] & + \ \frac{h^{(2)}(\sigma_{0j})}{nh(\sigma_{0j})} - \frac{1}{n}\left(\frac{h^{(1)}(\sigma_{0j})}{h(\sigma_{0j})}\right)^2, j=3g-l \\ & \text{if } l=2g,\dots,3g-1. \end{cases}$$

therefore we proved the Lemma.

Strong consistency of the penalized MLE is stated by means of the following two Theorems. They follow the structure of the Theorems proved by Wald (1949) for the classical MLE over a compact set.

Theorem 4.3.7 *(c.f. [9]) Let S be a closed subset of $\bar{\Psi}$ such that*

$$S = \{\psi \in \bar{\Psi} \mid \exists\{1,\dots,g\} \textit{ so that } \sigma_j \in [0,\eta)\}$$

and such that $\psi_0 \notin S$. Then

$$P\Big(\lim_{n\to\infty} \sup_{\psi\in S} \frac{\mathcal{L}_n(Y_1,\dots,Y_n|\psi)}{\mathcal{L}_n(Y_1,\dots,Y_n|\psi_0)} = 0\Big) = 1$$

Theorem 4.3.8 *(c.f. [9]) Let $\bar{\psi}_n = \bar{\psi}(Y_1,\dots,Y_n) \in \bar{\Psi}$ be a function of $Y_1,\dots,Y_n$ such that*

$$\frac{\mathcal{L}_n(Y_1,\dots,Y_n|\bar{\psi}_n)}{\mathcal{L}_n(Y_1,\dots,Y_n|\psi_0)} \geq \rho > 0, \quad \forall\, Y_1,\dots,Y_n, \quad \forall n$$

Then

$$P\Big(\lim_{n\to\infty} \bar{\psi}_n = \psi_0\Big) = 1$$

Corollary 4.3.9 *(c.f. [9]) The penalized maximum likelihood estimator is strongly consistent, i.e. the point $\bar{\psi}_n$ which maximizes $\mathcal{L}_n$ is such that $\psi_n \to \psi_0$ a.s.*

G. Ciuperca et al. (2003) considered the speed of convergence of the penalized estimator, in this section we will do the same. In their work, it was assumed that

$$\pi_k \neq 0 \quad \text{and} \quad (\mu_k, \sigma_k) \neq (\mu_m, \sigma_m) \quad \text{for } k \neq m, \forall k = 1,\dots,g \tag{4.3.24}$$

in order to have a non-singular information matrix

$$I(\psi_0) = E_H\Big[\Big(\frac{\partial \ln f_1(\psi_0)}{\partial \psi}\Big)\Big(\frac{\partial \ln f_1(\psi_0)}{\partial \psi}\Big)^T\Big]$$

Note, since we penalized the mixing proportion π_j by the addition of the penalty term $\prod_{j=1}^{g}(g\pi_j)^C$, we ensured that the estimates of π_j is not on the boundary points of its parametric space, i.e. π_j can never by equal to zero which make the assumption of $\pi_k \neq 0$ for $1 \leq k \leq g$ unnecessary. Therefore we only need to assume that

$$(\mu_k, \sigma_k) \neq (\mu_m, \sigma_m) \quad \text{for } k \neq m, \forall\, k = 1,\dots,g \tag{4.3.25}$$

4.3.2 Main results

Theorem 4.3.10 *If the parameters satisfy the condition (4.3.25) and the penalizing function is such that*

$$\frac{h^{(s)}(\sigma)}{h(\sigma)} \textit{ is bounded for } s = 1, 2, 3 \textit{ and } \forall\, \sigma \in \{\sigma_{01}, \ldots, \sigma_{0n}\}$$

then $\sqrt{n}(\bar{\psi}_n - \psi_0)$ is asymptotically normal distributed with mean zero and covariance matrix $I(\psi_0)^{-1}$.

Proof. Since $\bar{\psi}_n$ is consistent, we write Taylor's expansion of $\partial \ln \mathcal{L}_n(\bar{\psi})/\partial\psi$, in a neighbourhood of ψ_0, up to the second order. Hence, we obtain the vector equation

$$\begin{aligned} 0 &= \frac{\partial \ln \mathcal{L}_n(\bar{\psi}_n)}{\partial \psi} \\ &= \frac{\partial \mathcal{L}_n(\psi_0)}{\partial \psi} + (\bar{\psi}_n - \psi_0)^T \frac{\partial^2 \ln \mathcal{L}_n(\psi_0)}{\partial \psi^2} + \frac{1}{2} R_n(\psi_n^+) \end{aligned} \tag{4.3.26}$$

The vector $R_n(\psi^+)$ has the components

$$R_n(\psi_n^+)_k = (\psi_n^+ - \psi_0)^T B_k (\psi_n^+ - \psi_0), \quad k = 1, \ldots, 3g-1$$

where B_k is a square matrix with elements

$$B_{k(i,j)} = \Big(\frac{\partial^3 \ln \mathcal{L}_n(\psi_n^+)}{\partial\psi_i \partial\psi_j \partial\psi_k}\Big), \quad i, j \in \{1, \ldots, 3g-1\},$$

and ψ_n^+ is an intermediate point between $\bar{\psi}_n$ and ψ_0. Let us define the vector $T_k = B_k(\psi_n^+ - \psi_0)$ and the matrix $T_n(\psi_n^+) = (T_1, T_2, \ldots, T_{3g-1})$. By multiplying (4.3.2) by $1/n$, and by considering that the penalized log-likelihood function can be written as

$$\ln \mathcal{L}_n(\psi) = \sum_{i=1}^{n} \ln \Big[f_1(Y|\psi) \prod_{j=1}^{g} h(\sigma_j)^{1/n} \prod_{j=1}^{g} (g\pi_j)^{C/n} \Big] = \sum_{i=1}^{n} \ln \Big[u(Y_i|\psi) \Big],$$

we obtain

$$\sqrt{n}(\bar{\psi}_n - \psi_0)^T$$

$$= \Big[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \ln u(Y_i|\psi_0)}{\partial \psi}\Big]\Big[-\frac{1}{n}\frac{\partial^2 \ln \mathcal{L}_n(\psi_0)}{\partial \psi^2} - \frac{1}{2n}T_n(\psi_n^+)\Big]^{-1}. \tag{4.3.27}$$

Let us now focus on the first term in the bracket of (4.3.27). By means of Lemma 4.3.5, application of the central limit Theorem on the set of random variables

$$\Big(\partial \ln u(Y_i|\psi_0)/\partial \psi_l\Big)_{1\leq i\leq n},$$

$l = 1, \dots, 3g-1$ leads to

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \ln u(Y_i|\psi_0)}{\partial \psi_l} - \frac{1}{n}\frac{h^{(1)}(\sigma_0)}{h(\sigma_0)}\mathbb{I}_{l\geq 2g}$$

$$\to \mathbf{n}\Bigg(0, E_H\Big[\frac{\partial \ln f_1(Y|\psi_0)}{\partial \psi_l}\Big]^2\Bigg), \text{ as } n \to \infty$$

for $l = 1, \dots, 3g-1$, with $j = 3g - l$. Since $h^{(1)}(\sigma_0)/h(\sigma_0)$ is bounded, from (4.3.23) of Lemma (4.3.5) we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \ln u(Y_i|\psi_0)}{\partial \psi_l} \to$$
$$\mathbf{n}\Bigg(0, E_H\Bigg[\Big(\frac{\partial \ln f_1(Y|\psi_0)}{\partial \psi_l}\Big)\Big(\frac{\partial \ln f_1(Y|\psi_0)}{\partial \psi_l}\Big)^T\Bigg]\Bigg), \text{as } n \to \infty \tag{4.3.28}$$

Concerning the terms in the second factor of (4.3.27), $\partial^2 \ln \mathcal{L}_n(\psi_0)/\partial\psi^2$ is equal to

$$\sum_{i=1}^{n}\Bigg[-\frac{1}{u^2(Y_i|\psi_0)}\Big(\frac{\partial u(Y_i|\psi_0)}{\partial \psi}\Big)\Big(\frac{\partial u(Y_i|\psi_0)}{\partial \psi}\Big)^T + \frac{1}{u(Y_i|\psi_0)}\Big(\frac{\partial^2 u(Y_i|\psi_0)}{\partial \psi^2}\Big)\Bigg].$$

Then, from Lemma (4.3.6) and the strong law of large numbers, we obtain

$$\frac{1}{n}\frac{\partial^2 \ln \mathcal{L}_n(\psi_0)}{\partial \psi^2} \to -E_H\Bigg[\Big(\frac{\partial \ln f_1(Y|\psi_0)}{\partial \psi_l}\Big)\Big(\frac{\partial \ln f_1(Y|\psi_0)}{\partial \psi_l}\Big)^T\Bigg],$$
$$\text{as } n \to \infty \tag{4.3.29}$$

For the second of the two, since $h^{(3)}(\sigma_{0l})/h(\sigma_{0l})$ is bounded, we have

$$\frac{1}{n}T_n(\psi_n^+) = o(1). \tag{4.3.30}$$

By taking relations (4.3.28), (4.3.29) and (4.3.30) into account, the asymptotic variance of $\sqrt{n}(\bar{\psi}_n - \psi_0)^T$ is $I(\psi_0)^{-1}$.

4.4 Conclusion

We discussed how to estimate the parameters of the penalized modified likelihood for normal mixture model (with unequal variance) in this chapter. The expectation maximization (EM) algorithm was used for the parameter estimation. Note that the variance parameter was penalized by the addition of the penalty functions; the inverse gamma and the inverse chi-square distributions.

The main result of this chapter was the proof presented for asymptotic normality (see section 4.3). This proof is a vital first step needed to prove the asymptotic null distribution for the likelihood ratio test use to determine the number of components for a normal mixture model with unequal variance parameter. However, the proof of the asymptotic null distribution for the likelihood ratio test is left for future work. Since the asymptotic null distribution for the likelihood ratio test is an open problem, we therefore used simulation in chapter 5 to determine the asymptotic null distribution of the likelihood ratio test.

5 Penalized Modified Likelihood Approach to Microarray Data Analysis

In chapter 3 we introduced the mixture model method and chapter 2 explained how mixture models may be applied to microarray data to determine differentially expressed genes. Wei Pan et al. used normal mixture models, a nonparametric method, to detect differentially expressed genes [26, 27, 43]. Their approach implemented a normal mixture with unequal variances for each component. However, Keifer and Wolfowitz [21] showed that when applying mixture of normals with unequal variances for each component the likelihood approaches infinity as one of the variances approaches 0. The issue of fitting normal mixture with unequal variance was addressed by Hathaway [19], Ciuperca, Ridolfi and Idier [9]. In this chapter the penalized modified likelihood approach will be presented. This model, unlike that of Wei Pan et al. circumvents the possibility of the mixing proportion being on the boundary of the parametric space (that is, $\pi_i = 0$) and addressed the issue of normal mixture unequal variances by applying the technique of Ciuperca, Ridolfi and Idier [9].

Wei pan et al. (2002, 2003) used BIC as a criterion for model selection, to determine the number of components for the normal mixture model. However, there are no theoretical justification for the use of either the BIC or AIC model selection criteria for mixture models. Therefore we used the modified likelihood ratio test proposed by Chen et al. [6, 7] to test the hypotheses: a 1-component (null hypothesis) versus at least a 2-component model (alternative hypothesis) and a 2-component (null hypothesis) versus at least a 3-component model (alternative hypothesis). However, the modified likelihood ratio test of Chen et al. [6, 7] is not applicable in the heteroscedastic sense (that is, mixture of normal with unequal variances), hence we simulate the null distribution of the penalized modified likelihood ratio test (c.f chapter 3 page 45, where the simulation approach of Thode et al. (1988) was presented).

5.1 Penalized Modified Mixture Model (PMMM)

The null distribution of the penalized modified normal mixture model will be approximated by simulation. The observed likelihood was defined in chapter 4 as

$$\mathcal{L}_n(\psi|y) = \prod_{i=1}^{n}\sum_{j=1}^{g}\pi_i f_{ij}(y_i|\theta)\prod_{j=1}^{g}h(\sigma_j)\prod_{j=1}^{g}(g\pi_j)^C \tag{5.1.1}$$

where

$$f_{ij}(y_i) = \frac{1}{\sqrt{2\pi\sigma_j^2}}\exp\left\{-\frac{1}{2\sigma_j^2}(y_i-\mu_j)^2\right\} \quad j = 1,\ldots,g$$

is a normal density with mean μ_j and standard deviation σ_j. The parameter set of the mixture is given as

$$\Psi = (\pi_1,\ldots,\pi_j,\mu_1,\ldots,\mu_j,\sigma_1,\ldots,\sigma_j) \tag{5.1.2}$$

satisfying that $0 \leq \pi_j \leq 1, \sum_{j=1}^{g}\pi_j = 1, -\infty < \mu_j < \infty$, $\sigma_j > 0$ and the true parameters defined as $\psi_0 \in \Psi$. Furthermore, as in chapter 3 we let

$$\mathsf{M}_g = \Big\{G(\theta) = \sum_{j=1}^{g}\pi_j f_j(x_i|\theta_j) : \theta_1 \leq,\ldots,\leq \theta_g, \sum_{j=1}^{g}\pi_j = 1, \pi_j \geq 0\Big\}. \tag{5.1.3}$$

denote the class of all mixture probability density functions of which components are less than or equal to g.

5.2 PMMM Simulated Null Distribution

In this section we simulate the null distribution for the hypotheses

$$H_0 : G(\theta) \in \mathsf{M}_1 \text{ against } H_1 : G(\theta) \in \mathsf{M}_2. \tag{5.2.4}$$

and

$$H_0 : G(\theta) \in \mathsf{M}_2 \text{ against } H_1 : G(\theta) \in \mathsf{M}_3. \tag{5.2.5}$$

The simulation of the null distribution is done as follows. In the case of hypothesis (5.2.4) we simulated 1000 replicates of the standard normals $N(0, 1)$ of sample sizes $100, 250, 500, 750$ and 1000. Then we fitted 2-components normal mixture with unequal variances for each of the sample sizes and calculated the penalized modified log likelihood ratio test (PMLRT) define as

$$R_n = 2\{\ln \mathcal{L}_n(\hat{\pi}, \hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2) - \ln \mathcal{L}_n(1/2, \hat{\mu}, \hat{\mu}, \hat{\sigma}, \hat{\sigma})\}. \tag{5.2.6}$$

where $\mathcal{L}_n$ is defined in (5.1.1). A linear regression equation was fitted using the 5 values of the PMLRT to determine the degrees of freedom as a function of the sample size n (see section 3.3 of chapter 3). The degrees of freedom of the simulated chi-squared null distribution as a function of n for hypothesis (5.2.4) are given by

$$f = 3.1 + 10.2n^{-0.5}. \tag{5.2.7}$$

Table 5.1 shows the mean, variance and percentiles of the PMLRT for the sample sizes $100, 250, 500, 750$ and 1000 for hypothesis 5.2.4. The percentiles in brackets are that of the chi-squared distribution with degrees of freedom given by equation (5.2.7), while those percentiles not in brackets are the ordered simulated percentiles of PMLRT. We can see from Table 5.1 that the percentiles for the ordered simulated values compares well with that of the chi-squared distribution. The values for the 50^{th}, 75^{th}, 90^{th} and 95^{th} percentiles for sample sizes 100, 250, 500, 750 and 1000 are relatively close, suggesting that we have a good agreement between our simulated and theoretical distributions.

Note that the degrees of freedom of a chi-square distribution are integers, therefore a gamma distribution with mean $1.55 + 5.10n^{-1/2}$ and second parameter 0.5 was used. This was done because the chi-squared distribution, χ^2_f with degrees of freedom f, is a special case of the gamma distribution $G(f/2, 1/2)$ with parameters $f/2$ and $1/2$.

In the case of the hypothesis (5.2.5) we simulated 1000 replicates from the normal

Table 5.1: Mean, variance and percentiles for the penalized modified likelihood, based on 1000 replicates for each sample for testing the hypothesis a 1-component against 2-components.

Sample size	100	250	500	750	1000
Mean	4.00	3.90	3.71	3.33	3.29
Variance	8.05	8.04	7.71	7.03	7.02
			Percentiles		
50%	3.20(3.45)	3.22(3.08)	3.05(2.90)	2.70(2.81)	2.66(2.76)
75%	5.59(5.51)	5.33(5.04)	5.13(4.80)	4.28(4.69)	4.40(4.63)
90%	8.01(7.93)	7.90(7.37)	7.23(7.08)	6.82(6.95)	6.87(6.87)
95%	9.74(9.65)	9.40(9.04)	9.25(8.72)	8.79(8.56)	8.58(8.50)

The percentiles of $\chi^2_f = G(f/2, 1/2)$, $f = 3.1 + 10.2n^{-0.5}$ are displayed in brackets

mixture models

$$0.5\phi(y|0,1) + 0.5\phi(y|2,1) \quad \text{and} \quad 0.2\phi(y|0,1) + 0.8\phi(y|2,1) \tag{5.2.8}$$

of sample sizes $100, 250, 500, 750$ and 1000. We fitted 2-components and 3-components normal mixture distributions of data simulated from the normal mixture models of (5.2.8) and evaluate the PMLRT. The PMLRT for hypothesis (5.2.5), is given by

$$R_n = 2\{\ln \mathcal{L}_n(G_3(\hat{\theta})) - \ln \mathcal{L}_n(G_2(\hat{\theta}))\} \tag{5.2.9}$$

where $G_2(\hat{\theta})$ and $G_3(\hat{\theta})$ are the estimates under the null and alternate hypothesis of (5.2.5) respectively. The linear regression equation for the degrees of freedom as a function of n was determined to be

$$f = 4.89 + 11.84n^{-1/2} + 0.09I, \tag{5.2.10}$$

where

$$I = \begin{cases} 1 & \text{if means are from} \quad 0.5\phi(y|0,1) + 0.5\phi(y|2,1) \\ 0 & \text{if means are from} \quad 0.2\phi(y|0,1) + 0.8\phi(y|2,1). \end{cases}$$

Table 5.2 depicts the mean, variance and percentiles of the PMLRT for the sample sizes $100, 250, 500, 750$ and 1000 for hypothesis (5.2.5). The percentiles in brackets are that of the chi-squared distribution with degrees of freedom given by (5.2.10), while those percentile not in brackets are the ordered simulated percentiles of PMLRT.

Table 5.2 illustrates that the percentiles for the ordered simulated values compares well with that of the chi-squared distribution. The values for the 50^{th}, 75^{th}, 90^{th} and 95^{th} percentiles for sample sizes 100, 250, 500, 750 and 1000 are relatively close, suggesting that we have a good agreement between our simulated and theoretical distributions.

Note that a gamma distribution with mean $2.44 + 5.92n^{-1/2} + 0.045I$ and second parameter 0.5 is equivalent to χ^2_f, where $f = 4.89 + 11.84n^{-1/2} + 0.09I$.

Table 5.2: Mean, variance and percentiles for the penalized modified likelihood, based on 1000 replicates for each sample for testing the hypothesis 2-components against 3-components.

Sample size	100	250	500	750	1000
Simulated results for $0.5\phi(y\|0,1) + 0.5\phi(y\|2,1)$					
Mean	6.11	5.84	5.60	5.33	5.26
Variance	11.92	11.72	11.04	10.69	10.53
Percentiles					
50%	5.42(5.50)	5.03(5.07)	4.84(4.85)	4.71(4.75)	4.55(4.69)
75%	8.46(8.03)	7.65(7.50)	7.34(7.23)	7.02(7.12)	7.07(7.05)
90%	10.60(10.86)	11.10(10.25)	9.95(9.94)	9.74(9.81)	9.56(9.72)
95%	11.78(12.82)	12.19(12.17)	12.47(11.84)	11.82(11.69)	11.29(11.60)
Simulated results for $0.2\phi(y\|0,1) + 0.8\phi(y\|2,1)$					
Mean	6.00	5.75	5.47	5.25	5.24
Variance	12.03	12.13	11.85	10.84	10.63
Percentiles					
50%	5.67(5.41)	5.09(4.98)	4.90(4.76)	4.67(4.66)	4.59(4.60)
75%	7.67(7.92)	7.31(7.39)	7.15(7.12)	7.13(7.01)	7.00(6.94)
90%	9.83(10.73)	10.60(10.13)	10.09(9.82)	9.45(9.68)	9.65(9.60)
95%	12.44(12.69)	12.03(12.03)	11.90(11.70)	11.10(11.55)	11.20(11.46)

The percentiles of $\chi^2_f = G(f/2, 1/2)$, $f = 4.89 + 11.84n^{-1/2} + 0.09I$ are displayed in brackets

In the next section we applied the asymptotic null distributions of the likelihood ratio test for hypotheses (5.2.4) and (5.2.5) to determine the number of components of normal mixture models with unequal variances. This approach is our contribution to the theory of mixture models instead of using the model selection criterion BIC. The model selection criterion BIC has been used to determine the number of components for normal mixture model with unequal variances, but to date there has not been any theoretical justification for the use of the BIC as a model selection criterion.

5.3 Simulating Microarray Data

To mimic the real gene data, we generated data for $N = 1176$ genes under the following setup. We used $m = 2, n = 6$ and simulated 200 differentially expressed (DE) genes. The choices for N, m and n were made to parallel data from a study, that applied radioactively labeled DNA microarrays (Friemert et al. 1998) to the mRNA analysis of $N = 1176$ genes in middle ear mucosa of rats with and without subacute pneumococcal middle ear infection. The data consists of eight experiments: two ($m = 2$) DNA microarray were run with controls while six ($n = 6$) were run with pneumococcal middle ear infection.

The data for the equally expressed (EE) genes are simulated from $N(\mu_{i1}, \sigma_{i1}^2)$ for $k = 1, \ldots, m$ and $N(\mu_{i2}, \sigma_{i2}^2)$ for $k = m+1, \ldots, m+n$, where $\mu_{i1} = \mu_{i2} \sim N(0, 2)$ and σ_{i1} and σ_{i2} are generated from $Gamma(2, 4)$, respectively. Note that such generated σ_{i1} and σ_{i2} take different values for each gene and are also different between genes. The data for DE genes were generated similarly. However, in this case, μ_{i1} and μ_{i2} were generated from $N(0, 2)$ separately. The variances μ_{i1} and μ_{i2} are generated the same way as in the EE gene case.

5.4 Application of PMMM to Simulated Microarray Data

The method that will be used to analyze the simulated microarray data is that of Wei Pan at al. introduced in section 2.3 of chapter 2. The method involved first calculating the test statistics (5.4.11) and its null distribution (5.4.12)

$$\begin{aligned} Z_i &= \frac{\sum_{k=1}^{m} Y_{ik}/m - \sum_{k=m+1}^{m+n} Y_{ik}/n}{\sqrt{s_{i(1)}^2/m + s_{i(2)}^2/n}} \\ &= \frac{\bar{Y}_{i(1)} - \bar{Y}_{i(2)}}{\sqrt{s_{i(1)}^2/m + s_{i(2)}^2/n}} \sim f_1, \qquad (5.4.11) \end{aligned}$$

$$z_i = \frac{Y_{i(1)} p_i/m + Y_{i(2)} q_i/n}{\sqrt{s^2_{i(1)}/m + s^2_{i(2)}/n}} \sim f_0, \tag{5.4.12}$$

where $Y_{i(1)} = (Y_{i1}, \ldots, Y_{im})$ are gene expression from m microarrays under condition 1, and $Y_{i(2)} = (Y_{i,m+1}, \ldots, Y_{i,m+n})$ are from n arrays under condition 2 of a microarray experiment. Note that m and n are assumed to be even and p_i (q_i) is a column vector containing random permutation of $m/2$, 1's and $m/2$, -1's ($n/2$, 1's and $n/2$, -1's).

The hypothesis to be tested is

$$\begin{aligned} H_0 &: f_0 = f_1, \text{ there is no gene with altered expression} \\ H_1 &: f_0 \neq f_1, \text{ otherwise} \end{aligned} \tag{5.4.13}$$

We therefore fitted $1, 2$ and 3-components normal mixture model and calculated R_n defined in 5.2.6 and 5.2.9 respectively to determine the distributions of f_0 and f. Table 5.3 displays the results of the hypothesis test for the number of components for f_0 and f. Hence, the choice for both f_0 and f are the 2-components normal mixture model which are stated below:

$$\begin{aligned} f_0(z) &= 0.01\phi(z|-0.287, 0.05673^2) + 0.99\phi(z|-0.004558, 0.40812^2), \\ f(z) &= 0.20\phi(z|-2.442, 0.43703^2) + 0.80\phi(z|0.0062961, 0.48583^2). \end{aligned}$$

Table 5.3: Hypothesis test for the number of components for the fitted normal mixture models of z and Z, for the simulated microarray data.

	1 vs. 2 component	2 vs. 3 component
f_0	4.56 ($P < 0.01$)	1.22 ($P > 0.05$)
f	6.78 ($P < 0.01$)	1.06 ($P > 0.05$)

Figure 5.1a shows the histograms of z with the fitted normal mixture models, which shows strong agreement. Similar observation for Z is shown in 5.1b) with the dotted line being that of the fitted mixture model of f_0. Figure 5.2 illustrates the likelihood ratio statistic as a function of the Z values.

Our main interest for applying the PMMM approach is to determine which genes are differentially expressed, therefore the median number of false positive (FP) were calculated from the null scores of $B = 29$ permutations of a data set simulated from the fitted null distribution f_0. Additionally, we compared the results of PMMM to that of SAM by using the t-test with 500 permutation. Results for SAM were obtained by using the R-package sam3.0. For the purpose of comparison, the cut-off points s (see section 2.3) of the PMMM approach are specifically chosen to match the number of true positive (TP) produced by sam3.0. It can be seen from Tables 5.4 and 5.5 that our method out perform SAM. Figure 5.3 displayed a graphical comparison of the numerical results presented in Tables 5.4 and 5.5.

Figure 5.1: Histograms of z, Z and fitted models for the simulated data

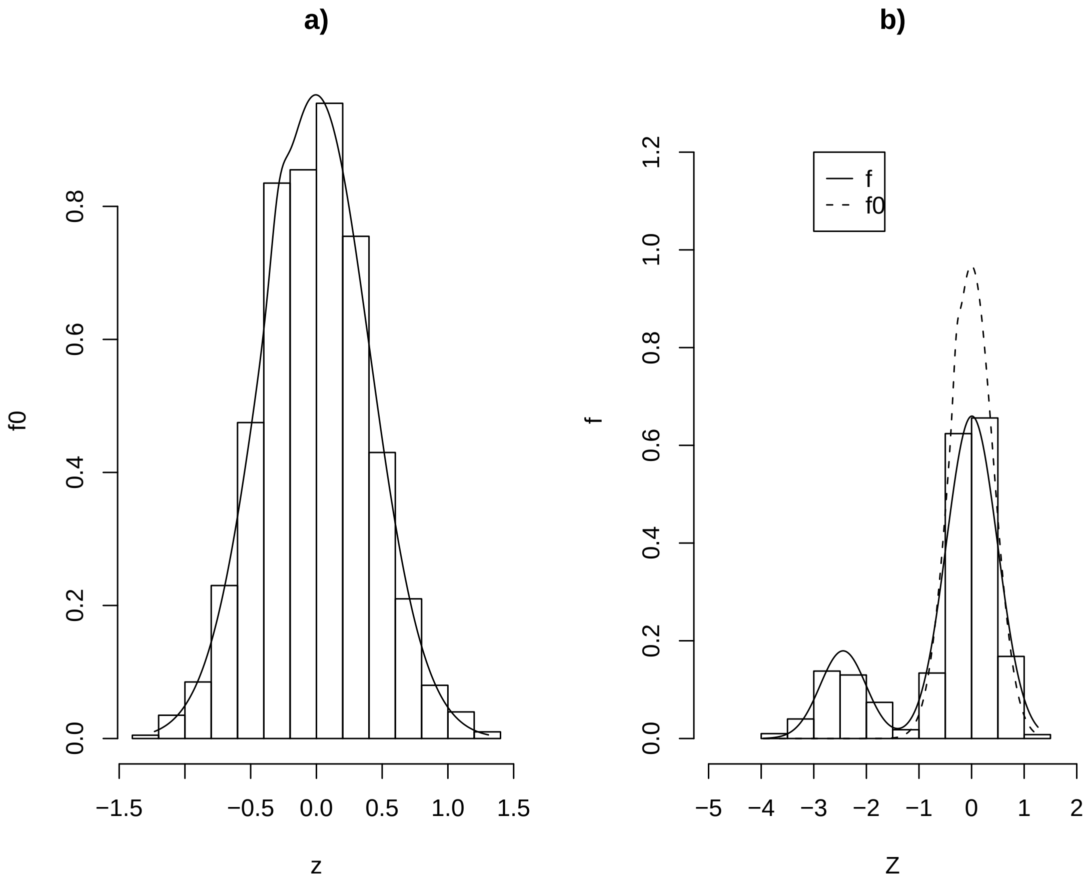


Table 5.4: Values of TP, FP and FDR from PMMM for the simulated data

s	$MedianFP$	$MeanFP$	TP	$FDR\%$
0.07	0	0.069	196	0.00
0.10	0	0.138	196	0.00
0.15	0	0.310	196	0.00
0.30	2	1.655	201	1.00
0.35	3	3.828	203	1.48
0.40	5	5.621	205	2.44
0.45	14	13.931	210	6.67
0.60	26	25.724	221	11.76
0.70	43	43.207	231	18.61
0.90	68	66.966	248	27.42
1.00	104	103.517	270	38.52

Table 5.5: Values of TP, FP and FDR from SAM for the simulated data

Δ	$Median\ FP$	$Mean\ FP$	TP	$FDR\%$
0.49	3.71	6.496	195	1.90
0.47	4.64	6.496	197	2.36
0.45	6.50	8.352	198	3.28
0.43	7.42	10.208	200	3.71
0.42	8.35	12.064	203	4.11
0.37	11.14	14.848	206	5.41
0.32	23.20	27.840	211	11.00
0.28	33.41	40.832	221	15.12
0.25	45.47	55.680	230	19.77
0.20	69.60	82.592	246	28.29
0.16	107.65	118.042	268	40.17

Figure 5.2: The likelihood ratio statistic as a function of Z value for the simulated data

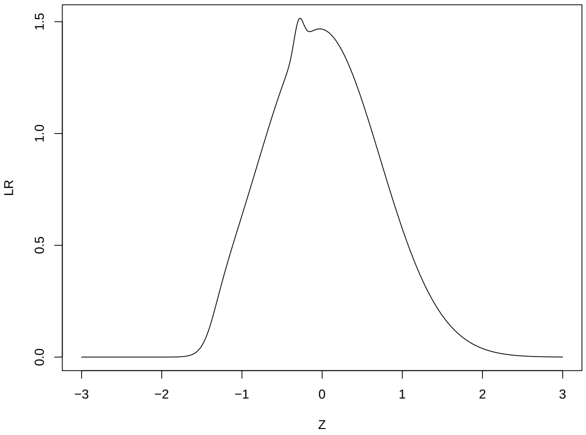
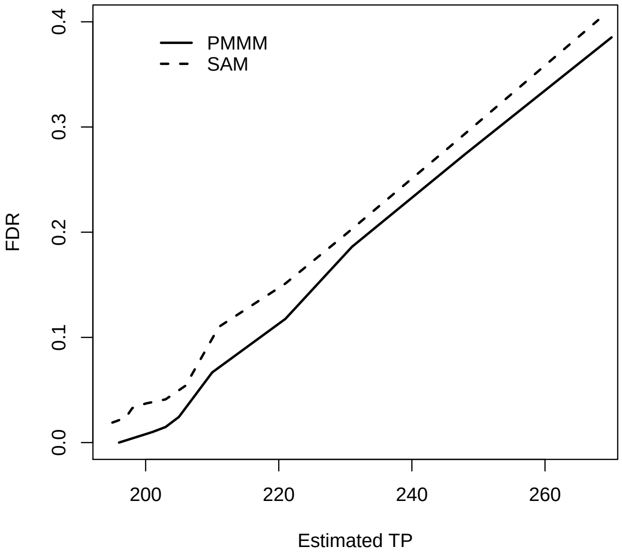


Figure 5.3: The values of FDR from our method and SAM for the simulated data



5.5 Application of PMMM to the Rat data

In this section, we apply the penalized modified likelihood method to the rat data of [26]. The data is from a study, that applied radioactively labeled DNA microarrays (Friemert et al. 1998) to the mRNA analysis of 1,176 genes in middle ear mucosa of rats with and without subacute pneumococcal middle ear infection. The data consists of eight experiments: two DNA microarray were run with controls while six were run with pneumococcal middle ear infection. The data was processed by first taking a natural logarithm transformation for all the observed gene expression levels so that the resulting data is less skewed. Then, for each microarray, we standardize the transformed gene expression levels by subtracting their mean and dividing by their standard deviation.

Table 5.6 presents the results of the test of hypothesis to determine the number of components of the normal mixture models for f_0 and f. We choose the 2-components normal mixture model for both f_0 and f which are stated below:

Table 5.6: Hypothesis test for the number of components for the fitted normal mixture models of z and Z, for the rat data.

	1 vs. 2 component	2 vs. 3 component
f_0	3.19 ($P < 0.01$)	0.92 ($P > 0.05$)
f	3.54 ($P < 0.01$)	1.16 ($P > 0.05$)

$$\begin{aligned} f_0(z) &= 0.983\phi(z|0.011, 0.430^2) + 0.017\phi(z|0.297, 0.263^2), \\ f(z) &= 0.958\phi(z|-0.032, 0.734^2) + 0.042\phi(z|0.246, 0.063^2). \end{aligned}$$

Figure 5.4a presented the histogram of z and the fitted f_0, which do not indicate strong discrepancy. The histogram of Z and the fitted mixture model are shown in Figure 5.4b with f_0 shown in the dotted line. The chi-square goodness of fit test was done resulting in p-values of 0.352 and 0.298 for f_0 and f respectively, supporting the claim that the fitted mixture models are f_0 and f for the null and alternative density functions respectively. The constructed LR statistics are plotted in Figure 5.5. It is not surprising to see as Z moves away from 0, $LR(Z)$ decreases.

Tables 5.7 and 5.8 report the results from our method and SAM. Figure 5.6 displays the FDR with respect to different values of TP. For $TP \leq 300$, the advantage of our method over SAM is obvious. For $TP > 300$, the FDR value of our method is higher than that of SAM. It is noteworthy that for this data set the number of genes that one wants to detect should be no greater than 300, hence the PMMM approach provides statistical significant results compared to that of SAM.

Figure 5.4: Histograms of z, Z and fitted models for the rat data

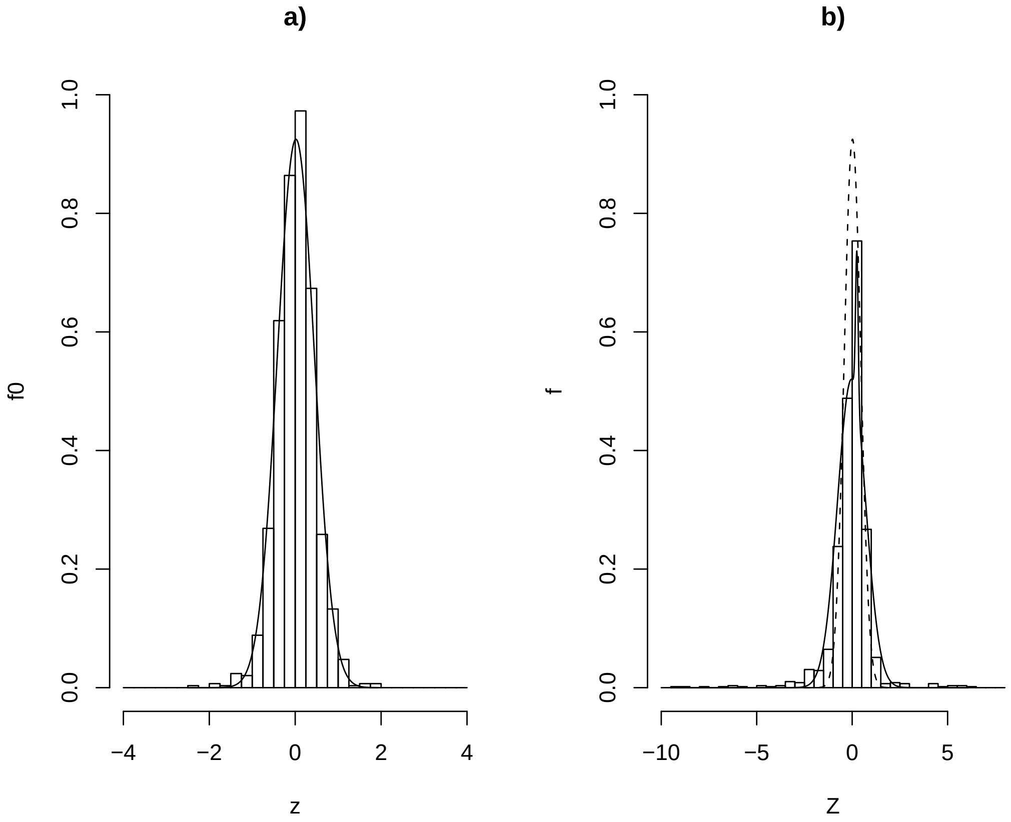


Table 5.7: Values of TP, FP and FDR from PMMM for the rat data

s	$MedianFP$	$MeanFP$	TP	$FDR\%$
0.07	0	0.03	94	0.00
0.10	0	0.07	103	0.00
0.15	0	0.28	113	0.00
0.30	3	3.17	144	2.08
0.35	8	8.75	168	4.76
0.40	12	12.44	178	6.74
0.45	29	27.86	215	13.49
0.60	44	46.17	248	17.74
0.70	65	65.96	288	22.57
0.90	95	95.59	323	29.41
1.00	134	133.83	368	36.41

Table 5.8: Values of TP, FP and FDR from SAM for the rat data

Δ	$Median\ FP$	$Mean\ FP$	TP	$FDR\%$
0.94	4.2	16.73	80	5.23
0.88	9.1	23.71	101	8.97
0.78	11.2	29.98	149	9.96
0.68	19.5	45.32	149	13.10
0.63	25.1	62.76	168	14.94
0.58	34.2	76.70	198	17.26
0.54	49.5	97.62	238	20.80
0.50	57.2	109.47	259	22.08
0.46	75.7	135.27	301	25.13
0.42	93.1	167.35	336	27.70
0.38	132.5	221.04	420	31.54

Figure 5.5: The likelihood ratio curve for the rat data

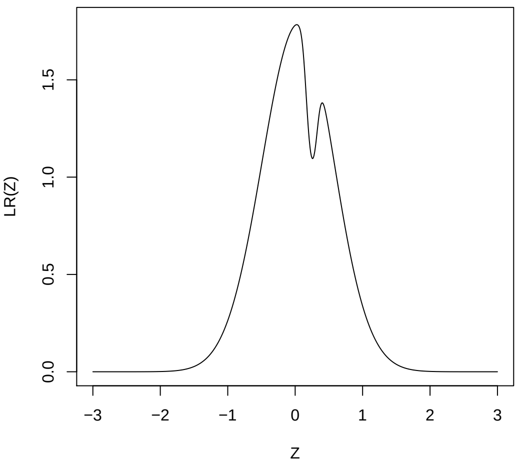
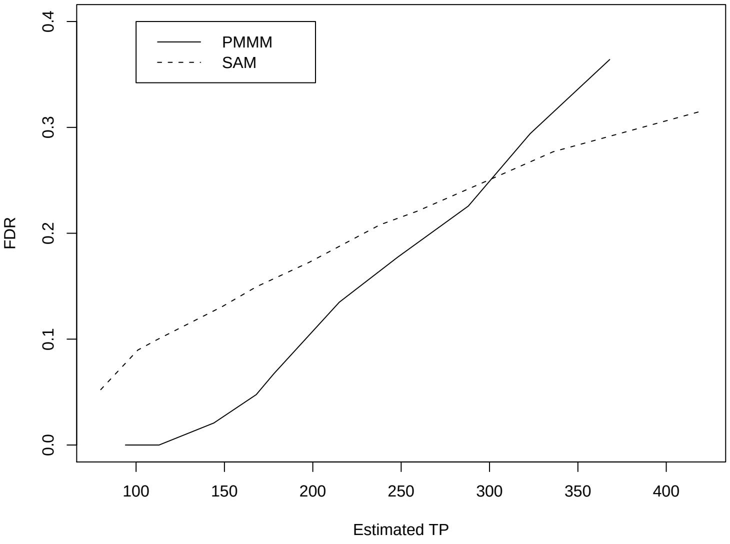


Figure 5.6: The values of FDR from our method and SAM for the rat data



5.6 Conclusion

In this chapter we presented the penalized modified likelihood approach. The advantage of this approach is that we can implement normal mixture models with unequal variances. Wei Pan et al. used normal mixture models with unequal variance parameters without any model justifications. They also used the BIC model selection criterion to determine the number of components of a normal mixture model. However, there are no theoretical justifications for the use of the model selection criterion, BIC for mixture models. For the penalized modified likelihood approach the likelihood ratio test can be applied to determine the number of components, because the mixing proportion have been penalized.

In this dissertation we have not determine the theoretical null distribution of the likelihood ratio test for the hypotheses: A one component normal mixture (H_0) against two components normal mixture (H_a). Two components normal mixture model (H_0) against three components normal mixture (H_a). Hence we simulated the null distribution of the likelihood ratio test. In chapter 3 section 3.3, the simulation of the null distribution was explain and the degrees of freedom for the chi-square statistic was determine by the regression approach of Thode et al. Since the degrees of freedom for the χ^2_f distribution with degrees of freedom f is equivalent to a gamma distribution with parameters $f/2$ and 0.5, the gamma distribution was used to determine the P-value of the hypotheses stated above.

The results of the penalized modified likelihood approach were compared to that of SAM. For simulated data the penalized modified likelihood approach outperformed that of SAM by comparing the false discovery rates (FDR) (see tables 5.4 and 5.5). The false discovery rates for the penalized modified likelihood approach were less than that of SAM. In the case of real data the penalized modified likelihood approached outperformed that of SAM for true positive (TP) less than or equal to 300. With $TP \leq 300$, the false discovery rates of the penalized modified likelihood were less than that of SAM (see tables 5.7 and 5.8).

6 Modified P-Value Approach to Microarray Data Analysis

In section 2.3 of chapter 2 we presented the p-value approach of Allison et al. (2002), used to determine differentially expressed genes in microarray data analysis. Our major contribution in this chapter is that we modified the p-value approach of Allison et al. by penalizing the mixing proportion. Note that, Theorems 3.3.1 and 3.3.2 of Chen et al. are not applicable, that is, the asymptotic null distribution is not

$$\frac{1}{2}\chi_1^2 + \frac{1}{2}\chi_0^2,$$

for test the hypothesis of 1-component against 2-component, and

$$\Big(\frac{1}{2} - \frac{\alpha}{2\pi}\Big)\chi_0^2 + \frac{1}{2}\chi_1^2 + \frac{\alpha}{2\pi}\chi_2^2,$$

in the case we test the hypothesis of 2-component against 3-component. Therefore, the null distribution of the modified p-value approach will be determined by simulation using the regression approach of Thode et al. (1988). The modified p-value approach was applied to both simulated and real micoarray data to determine the number of mixing components of a uniform-beta mixture model by means of likelihood ratio test.

6.1 The modified P-Value Approach

It is known that the distribution of p-values is uniform under the null hypothesis, therefore there exist a one component model, implying that the distribution characterizing the p-values is indeed uniformly distributed. Then we can safely say that the genes in the study are not differentially expressed. The hypothesis can be express as (c.f. Allison et al. (2002)):

$$\begin{aligned} H_0 &: \text{ uniformly distributed,} \\ H_1 &: \text{ mixture of uniform and beta distributions.} \end{aligned} \tag{6.1.1}$$

To test hypothesis (6.1.1), Allison et al. (2002) used bootstrapping to determine the number of components in the mixture model of uniform and beta distributions. The asymptotic null distribution of the likelihood ratio test for the p-value approach of Allison et al. can be determined by simulation if we penalize the mixing proportion. The penalization of the mixing proportion is important because hypothesis (6.1.1) can be easily misinterpreted as being unform if the estimates of the mixing component lie on the boundary of its parametric space, that is $p_j = 0$. Therefore we modified the p-value approach of Allison et al. (2002) by the addition of a penalty term for the mixing proportions as was done in Chen et al. The addition of this penalty term results in the parameter estimate of the mixing proportion p_j not being on the boundary of the parametric space (i.e. $p_j = 0$), hence circumventing the non-identifiability of the parametric space.

Let

$$\beta(y|r,s) = \frac{\Gamma(r+s)y^{r-1}(1-y)^{s-1}}{\Gamma(r)\Gamma(s)},$$

denote the beta distribution with parameters r and s, for $r = s = 1$ we have the special case of the beta which is uniform $U[0,1]$. The modified likelihood function can be expressed as

$$L_g = \prod_{i=1}^{n} \left[p_1\beta(y_i|1,1) \prod_{j=2}^{g-1} p_j\beta(y_i|r_j,s_j) \right] \prod_{j=1}^{g} (gp_j)^C, \tag{6.1.2}$$

hence the resulting modified log likelihood function is

$$l_g = \sum_{i=1}^{n} \ln \left[p_1\beta(y_i|1,1) + \sum_{j=2}^{g-1} p_j\beta(y_i|r_j,s_j) \right] + C\sum_{j=1}^{g} \ln(gp_j), \tag{6.1.3}$$

where $\sum_{j=1}^{g} p_j = 1$, $p_j \geq 0$ and y represent the p-value from a valid statistical test. From

(6.1.3) and let $g = 2$ for simplicity, we are interested in testing the hypothesis

$$\begin{aligned} &H_0 : r_2 = s_2 = 1 \\ &H_1 : r_2 \neq 1 \text{ or } s_2 \neq 1. \end{aligned} \tag{6.1.4}$$

Note that in the approach of Allison et al. (2002) the parameters r_2, s_2 and p_2 are not identifiable under the null as was mentioned early and the null hypothesis lies on the boundary of the parametric space ($p_2 = 0$).

With the addition of the penalty term $C \sum_{j=1}^{g} \ln(gp_j)$ we may be able to apply Theorem 3.3.1 of chapter 3, therefore the resulting null distribution is

$$\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2.$$

However, it should be noted that we are estimating the parameters p_2, r_2 and s_2 hence the Theorem 3.3.1 is not applicable. One way to address the problem is to fix r_2 preferable equal to 1.

With this done we will now need to estimate the parameters p_2 and s_2, noting that the parameter s_2 characterizes the behaviour of the p-values close to zero.

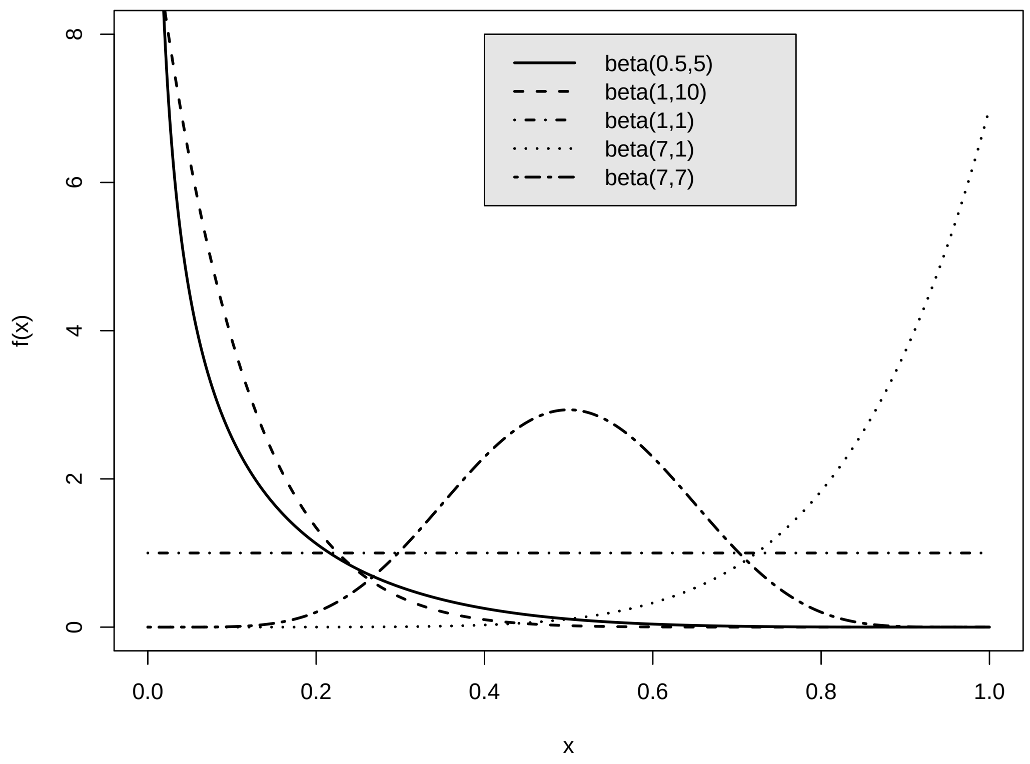
Examining Figure 6.1 we observe that the beta distributions that aptly describe the behaviour under the alternative hypothesis, that is, where the distribution of p-values tend to cluster closer to zero are given by $\beta(y|1, 10)$ and $\beta(y|0.5, 5)$. For this reason if we wish to apply the mixture of uniform-beta distributions for p-value approach we can implement a uniform-beta mixture of the form

$$p_1\beta(y_i|1, 1) + \sum_{j=2}^{g-1} p_j\beta(y_i|1, s_j),$$

with modified log likelihood function

$$l_g = \sum_{i=1}^{n} \ln \left[p_1\beta(y_i|1, 1) + \sum_{j=2}^{g-1} p_j\beta(y_i|1, s_j) \right] + C \sum_{j=1}^{g} \ln(gp_j). \tag{6.1.5}$$

Figure 6.1: Various Beta Distributions



Therefore the hypothesis to be tested, assuming $g = 2$, is given by

$$\begin{aligned} &H_0 : s_2 = 1 \\ &H_1 : s_2 \neq 1. \end{aligned} \tag{6.1.6}$$

However, the asymptotic null distribution is not

$$\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$$

as was shown in Chen et al. in the case of normal mixture models with equality of variance. We use simulation to determine the null distribution of hypothesis (6.1.6) by applying the regression method of Thode et al. (1988). Simulation was used because the theoretical asymptotic null distribution is an open problem as is the case for the penalized modified normal mixture model (see chapter 5).

6.2 Simulated Null Distribution of the Modified P-Value Approach

The simulation of the null distribution is done as follows, we generated 1000 replications of a uniform distribution on the interval $[0, 1]$, for each of the 5 sample sizes: 100, 250, 500, 750, 1000. We then fitted a two component uniform-beta model and evaluated the modified likelihood ratio statistic. The modified likelihood ratio statistic for the hypothesis test of a uniform versus a uniform and a beta is define to be

$$2(l_2 - l_1) \tag{6.2.7}$$

where l_g is defined in equation (6.1.5). Table 6.1 displays the results for the mean, variance and percentiles for the sample sizes stated above. From the results stated in Table 6.1 we see that the simulated results for the asymptotic null distribution is a χ^2 distribution, because the variance is twice the mean. Additionally, the simulated percentiles are approximately that of the χ^2_f distribution in brackets, where f is the regressed degrees of freedom which are stated below. The regression equation for the degrees of freedom as a function of the sample size n was evaluated using the means for each sample size. The regression equation was found to be

$$f = 1.32 + 4.01n^{-1/2} \tag{6.2.8}$$

Table 6.1: Mean, variance and percentiles for the likelihood ratio test for the modified p-value, based on 1000 replicates for each sample for testing the hypothesis a uniform against a uniform and one beta distribution.

Sample size	100	250	500	750	1000
Mean	1.69	1.63	1.54	1.49	1.37
Variance	3.52	3.29	3.17	2.84	2.80
			Percentiles		
50%	0.97(1.12)	0.96(0.98)	0.89(0.91)	0.94(0.88)	0.78(0.86)
75%	2.57(2.38)	2.31(2.17)	2.19(2.07)	2.11(2.02)	1.96(1.99)
90%	4.21(4.11)	4.22(3.84)	3.85(3.70)	3.66(3.64)	3.32(3.60)
95%	5.45(5.44)	5.32(5.13)	5.00(4.98)	5.01(4.91)	4.58(4.87)

The percentiles of $\chi^2_f = G(f/2, 0.5)$, $f = 1.32 + 4.01n^{-1/2}$ are in brackets

In a similar fashion if we wanted to test the hypothesis a uniform with a beta distribution versus a uniform with 2 beta distributions the modified likelihood ratio statistic are defined as

$$2(l_3 - l_2). \tag{6.2.9}$$

Table 6.2 depicts similar results as shown in Table 6.1. However the degrees of freedom for the hypothesis 2-component versus 3-component for the modified p-value approach is

$$f = 3.69 + 7.27n^{-1/2} \tag{6.2.10}$$

Table 6.2: Mean, variance and percentiles for the likelihood ratio test for the modified p-value, based on 1000 replicates for each sample for testing the hypothesis a uniform against a uniform and two beta distributions.

Sample size	100	250	500	750	1000
Mean	4.42	4.14	3.99	3.96	3.93
Variance	8.47	8.16	8.13	7.84	7.73
			Percentiles		
50%	3.94(3.77)	3.43(3.50)	3.36(3.37)	3.49(3.31)	3.17(3.28)
75%	5.96(5.91)	5.24(5.57)	5.37(5.40)	5.19(5.33)	5.46(5.28)
90%	8.01(8.39)	8.67(8.00)	7.96(7.80)	7.70(7.71)	7.83(7.66)
95%	9.51(10.16)	10.46(9.73)	9.12(9.51)	9.48(9.41)	9.37(9.36)

The percentiles of $\chi^2_f = G(f/2, 0.5)$, $f = 3.69 + 7.27n^{-1/2}$ are in brackets

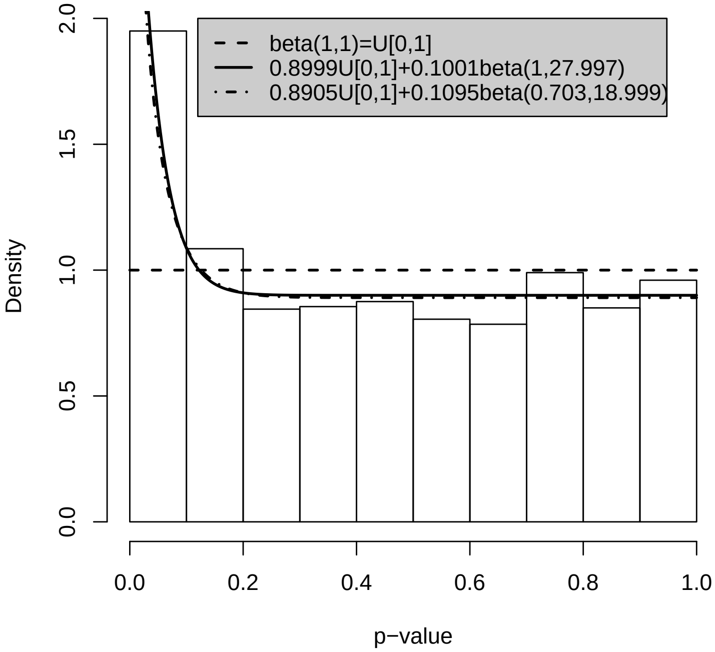
6.3 Application of Modified P-Value to Simulated Microarray Data

To illustrate that fixing $r = 1$ performs well we simulated data of sample size $n = 2,000$ such that 10% of the genes are differentially expressed (DE). In the control and case group we have a sample size of 10 each, and executed a t-test for each gene, then evaluated the distribution of p-values by fitting a mixture of uniform and beta distributions seen in Figure 6.2. The fitted mixture model is given by

$$0.89999416\beta(y|1,1) + 0.10000584\beta(y|1,25.997) \tag{6.3.11}$$

implying that the number of differentially expressed genes are 200 which compares well with the number of simulated differentially expressed genes which are 200. If $r \neq 1$,

Figure 6.2: Histogram of p-value and beta mixture distributions for 2000 simulated genes



the resulting fitted uniform-beta model was

$$0.89054436\beta(y|1,1) + 0.10945564\beta(y|0.703, 18.999) \tag{6.3.12}$$

yielding 219 differentially expressed genes which was out performed by the model that fixed $r = 1$. The graphs for both models are shown in Figure 6.2 where model (6.3.11) is the solid line and model (6.3.12) is the dotted line.

Under the model where $r = 1$, suppose we believed that the p-value from the distribution of p-values that are less than 0.10 are interesting and worthy of follow-up. The estimated proportion of these genes that are likely to be false leads is (see section 2.3 page 18 for details about formula)

$$\frac{0.89999416 \times 0.10}{0.89054436 \times 0.10 + 0.10000584 I_{0.10}(1, 25.997)} = 49\%$$

where $I_a(r, s)$ is the cumulative beta distribution with parameters r and s, evaluated at a. This proportion is 0.490, implying that there exist a 49% chance that any randomly selected genes with an ordinary p-value less than 0.10 will be a gene for which there is no real difference. The proportion not declared interesting that are likely to be genes for which there is a true significant difference in expression is

$$\frac{0.89999416 \times (1-0.10)}{0.89054436 \times (1-0.10) + 0.10000584[1 - I_{0.10}(1, 25.997)]} = 0.008.,$$

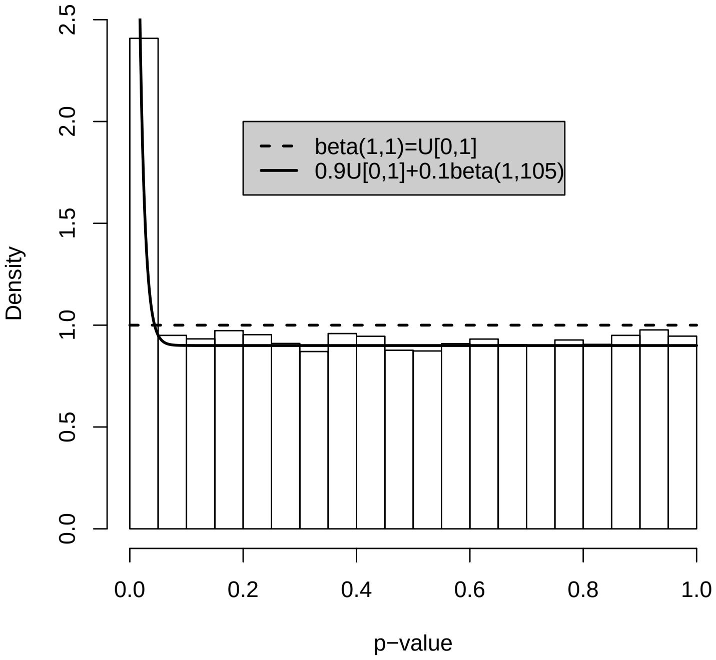
6.4 Application of Modified P-Value to simulated Prostate Data

To mimic the real prostate data we simulated 22,215 genes with the control group having 6 replicates and the case 5 replicates. The simulation of the data was done exactly as in the section entitled "Simulating microarray data". However, for this analysis we simulated 2,221 DE genes. The fitted uniform-beta model was

$$0.9\beta(y|1,1) + 0.1\beta(y|1,105)$$

and as can be seen in Figure 6.3, this model showed no discrepancy with the data.

Figure 6.3: Histogram of p-value and beta mixture distributions for simulated prostate data



Furthermore, the number of genes detected by this model is 2,221 which is exactly equal to the number we simulated.

With the MLE for p and s we have that the estimated proportion of genes that are likely to be false leads if we assume a threshold value of 0.10 is

$$\frac{0.9 \times 0.10}{0.9 \times 0.10 + 0.1 I_{0.10}(1, 105)} = 47\%,$$

and the proportion not declared interesting that are likely to be genes for which there is a true significant difference in expression is

$$\frac{0.9 \times (1 - 0.10)}{0.9 \times (1 - 0.10) + 0.1[1 - I_{0.10}(1, 105)]} = 0.000002.$$

6.5 Application of Modified P-Value to the Prostate Data

In this section we analyzed the prostate data set consisting 22,215 genes. The data has a sample size of 6 and 5 for the control and case group respectively, which is exactly the same as the simulated data. A t-test was done, generating p-values, which were then fitted by the unform-beta model to characterizing the distribution of the p-values. The distribution of the p-values is shown in Figure 6.4. From Figure 6.4 it can be seen that a uniform with one beta does not describe the distribution of the p-values as well as a uniform with two beta distributions model. To justify the choice of the 3-component model (a uniform with two beta model) a test of hypothesis was done, resulting in the rejection of the 2-component model (uniform with one beta), Table 6.3 illustrates the result. We found 6,753 differentially expressed genes for this prostate data.

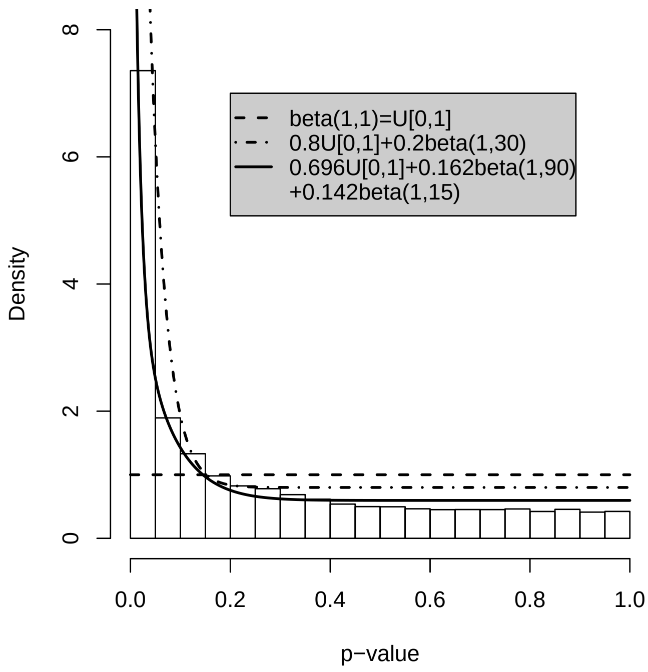
Table 6.3: Hypothesis test for the number of components for the fitted uniform-beta mixture models for the prostate data.

1 vs. 2 component	2 vs. 3 component
8.54 ($P < 0.01$)	1.67 ($P > 0.05$)

The uniform-beta model is therefore

$$0.696\beta(y|1, 1) + 0.162\beta(y|1, 90) + 0.142\beta(y|1, 15).$$

Figure 6.4: Histogram of p-value and beta mixture distributions for the prostate data



As was done before if we assumed a threshold value of 0.10 for which particular genes are declared "interesting" and worthy of follow-up study, the estimated proportion of genes declared interesting that are likely to be false leads is

$$\frac{0.696 \times 0.10}{0.696 \times 0.10 + 0.162 I_{0.10}(1, 90) + 0.142 I_{0.10}(1, 15)} = 20.2\%,$$

and the proportion not declared interesting that are likely to be genes for which there is a true significant difference in expression is

$$\frac{0.696 \times (1 - 0.10)}{0.696 \times (1 - 0.10) + 0.162[1 - I_{0.10}(1, 90)] + 0.142[1 - I_{0.10}(1, 15)]} = 0.045.$$

6.6 Conclusion

The Chapter examined the method of Allison et al. (2002) and applied the method of Chen et al. by adding a penalty term for the mixing proportion we used simulation to determine the degrees of freedom of the asymptotic null distributions for testing 1-component against 2-component (2-component against 3-component). By implementing a test of hypothesis we are statically certain of the distribution characterizing the behaviour of the *p*-values. One important observation that needs to be stated is that by modifying the mixing proportion the MLE of the mixing proportion cannot be on the boundary point of the parametric space. We carried out the same calculation as was done by Allison et al. that are (1) estimating the proportion of genes that are declared interesting that are likely to be false leads and (2) estimating the proportion of genes not declared interesting that are likely to be genes for which there is a real difference. We calculated these estimates when we applied our method to simulated data and the prostate data and we observed meaningful results.

7 Summary and Concluding Remarks

In this dissertation we modified the non-parametric normal mixture method of Wei Pan et al. for detecting differentially expressed genes in microarray data. In applying our modified non-parametric method, the penalized modified likelihood approach, we simulated the asymptotic null distribution of the likelihood ratio test for heteroscedastic normal mixture models where the mixing proportion and the variances were simultaneously penalized. Note that Wei Pan et al. used the model selection criterion BIC to determine the number of components in their normal mixture model. However, the BIC has no theoretical justification for mixture models.

The penalization techniques used in this dissertation was introduced by Chen et al. and they penalized the mixing proportion so that the asymptotic null distribution can be determined theoretically. In the non-parametric approach of Wei Pan they used heteroscedastic normal mixture models without addressing the unboundedness of the likelihood. Ciuperca et al. addressed the unboundedness of the MLE of the variance parameters with the addition of a penalty function for the variances. We combined both techniques so that we addressed the issues of non identifiability of the parameters under the null hypothesis and the unboundedness of the log likelihood simultaneously. Therefore our approach, the penalized modified likelihood approach is an important contribution to area of mixture models.

The proof that the penalized modified likelihood ratio statistic is asymptotically normal was presented in this dissertation. Asymptotical normality is an important property needed to prove the asymptotic null distribution of the penalized modified likelihood ratio statistic which was not proven in this dissertation. Since we did not prove the asymptotic null distribution of the penalized modified likelihood ratio test, we simulated the asymptotic null distribution and used the regression method of Thode et al. to determine its degrees of freedom.

The penalized modified likelihood approach for mixture of normal distribution with unequal variance was then used to determine the number of components for the null and alternative distributions for simulated and real world data. The results of the penalized modified likelihood approach were then compared to that of SAM. The results for the penalized modified method was found to out perform that of SAM.

In addition to the modified likelihood approach, we studied the p-value approach for detecting differentially express genes in microarray data introduced by Allison et al. We modified the p-value approach of Allison et al. by penalizing the mixing proportion. Similar argument as that presented above for the penalization of the mixing distribution in the case of normal mixture model applies. However, we made one simple modification by fixing the parameter, $r = 1$ of the beta distribution $\beta(r, s)$, because we observed that the distribution $\beta(1, s)$ describes the behaviour of the alternative hypothesis, where the distribution of p-values tends to be closer to zero. The challenge of proving the asymptotic distribution for the modified likelihood ratio statistic was not done in this dissertation. Therefore, the alternate approach of simulating asymptotic null distribution was done. The regression method of Thode at al. was used to determine the degrees of freedom of the asymptotic null distribution of the modified likelihood ratio statistics.

Allison et al. used the bootstrap approach to determine the number of components of a mixture of beta distribution. However, we simulated the asymptotic null distribution. Furthermore, Allison et al. did not state the empirical null distribution for the likelihood ratio test used to determine the number of components of the mixture of uniform and beta distributions. However, by using simulation we determined that the asymptotic null distribution has a χ^2 distribution, where the degrees of freedom was determine from the regression approach of Thode et al.

In the future I hope to prove the theoretical asymptotic null distribution of the penalized modified normal mixture model. Although the proof for the asymptotic null distribution of mixture of beta distributions will be more challenging, it is worth my focused attention. Furthermore, an interesting problem that needs serious consideration is that of using mixture of t-distributions instead of mixture of normals with unequal variance to describe the distributions of the null and alternative hypotheses used in the non-parametric approach of Wei Pan et al. Additionally, we would need to determine the power of the modified likelihood ratio test used throughout this dissertation.

References

[1] Akaike, H. (1973), *Information theory and an extension of maiximum likelihood principle,* 2^{nd} International Symposium on Infornmation Theory (eds. B. N. Petrov and F. Csaki), pp. 267-281, AKademiai Kiado, Budapest.

[2] Allison, D. B., Gadbury, G. L., Heo, M., (2002), *A mixture model approach for the analysis of microarray gene expression data.* Comput. Statis. & Data Analysis, Vol. 39, pp. 1-20

[3] Ben-Dor,A., Shaamir,R., and Yakhini,Z. (1999), *Clustering gene expression patterns.* J. Comput. Biol., Vol. 6, 281-297.

[4] Benjamini, Y., and Hochberg, Y., (1995), *Controlling the false discovery rate: a proctical and powerful approach to multiple testing,* J. R. Statist. Soc., Vol. 57, pp. 289-300.

[5] Böhning, D. (1999), *Computer Assisted Analysis of Mixture,* New York: Marcel Dekker.

[6] Chen, H., Chen, J., and Kalbfleischd, J. D., (2001), *A Modified Likelihood ratio Test for Homogeneity in Finite Mixture Models,* J. R. Statist. Soc., Vol. 63, Part 1, pp. 19-29.

[7] Chen, H., Chen, J., and Kalbfleischd, J. D., (2004), *Testing for a finite mixture model with two components,* J. R. Statist. Soc., Vol. 66, Part 1, pp. 95-115.

[8] Chen, Y., Doughterty, E., and Bitter, M., (1997), *Ratio-based decisions and the quantitative analysis of cDNA microarray images,* J. Biomedical Optics, Vol. 2, pp. 364-367.

[9] Ciuperca, G., Ridolfi, A. and Idier, J., (2003), *Penalized Maximum Likelihood Estimator for Normal Mixtures,* Scandinavian Journal of Statistics, Vol. 30, pp. 45-59.

[10] Dempster, A. P., Laird, N. M. and Rubin, D. B., (1977), *Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion)*, J. R. Statist. Soc. B, Vol. 39, pp. 1-38.

[11] DeRisi, J., Iyer, V. and Brown, P. O., (1997), *Exploring the metabolic and genetic control of gene expression on a genomic scale*, Science, Vol. 278, pp. 680-685.

[12] Devore J., and Peck, R., (1977), *Statistics: Exploration and Analysis of Data*, 3^{rd} edition. Pacific Grove, CA: Duxbury Press.

[13] Efron, B., Tibshirani, R. J., (1993), *An introduction to bootstrap*, London: Chapman and Hall.

[14] Efron, B., Tibshirani, R. J., Tusher, V., (2001), *Empirical Bayes Analysis of a Microarray Experiment*, J. of the Amer. Stat. Ass., Vol. 96, pp. 1151-1160.

[15] Eisen,M.B., Spellman,P.T., Brown,P.O., and Botstein,D. (1998), *Cluster analysis and display of genome-wide expression patterns.* Proc. Natl. Acad. Sci., Vol. 95, PP. 14863-14868.

[16] Everitt, B. S., and Hand, D. J. (1981), *Finite Mixture Distributions,* New York: Chapman and Hall.

[17] Fraley, C. and Raftery, A. E., (1998), *How many cluters? Which clustering methods? - Answer via model-based cluster analysis.*, The Computer Journal, Vol. 41, pp. 578-588.

[18] Gaasterland, T., and Bekiranov, S. (2000), *Making the most of mircoarray data.* Nat. Genet., Vol 24, pp. 204-206.

[19] Hathaway, R. J., (1985), *A contrained EM algorithm for univariate normal mixtures,* J. Stat. Comp. Simulation, Vol. 23, pp. 795-800.

[20] Hughes, T. R., Mao, M., Jones, A. R., Burchard, J. Marton, M.J., Shonnon, K. W., Lefkowitz, S. M., Ziman, M., Schelter, J. M., Meyer, M. R., Kobayashi, S., Davis,

C., Dai, H., He, Y. D., Stephaniants, S. B., Cavet, G., Walker, W. L., West, A., Coffey, E., Shoemaker, D. D., Stoughton, R., Blanchard, A. P., Friend, S. H. and Linsley, P.S., (2001), *Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer,* Nature Biotechnology, Vol. 19, pp. 342-347.

[21] Kiefer, J., Wolfowitz, J., (1956), *Consistency of the maximum-likelihood estimator in the presence of infinitely many incidental parameters,* Ann. Math. Stat., Vol. 27, pp. 888-906.

[22] 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 of monitoring by hybridization to high-density oligonucleotide arrays,* Nature Biotechnology, Vol. 14, pp. 1675-1996.

[23] MacLean, C. J., Morton, N. E., Elston, R. C., and Yee, S. (1976), *Skewness in commingled distributions* Biometrics, Vol. 51, No. 3, pp. 1461-1468.

[24] McLachlan, G. J., and Peel, D. (2000), *Finite Mixture Models,* New York: Wiley.

[25] Najarian, K., Zaheri, M., Rad, A. A., Najarian, S. and Dargahi, J., (2007), *A novel Mixture Model Methiod for identification of differentially expressed genes from DNA microarray data,* Bioinformatics, Vol. 5, pp. 201-211.

[26] Pan,W. (2002), *A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments.* Bioinformatics Vol. 18, pp. 546-554.

[27] Pan, W., Lin, J., Le, C., (2003), *A Mixture Model Approach to Detecting Differentially Expressed Genes with Microarray Data,* Functional & Integrative Genomics, Vol. 3, pp. 117-124.

[28] Pan,W., Lin,J. and Le,C., (2002), How many replication of Arrays are required to detect Gene Expression change in Microarray Experiment? A Mixture Model Approach, Division of Biostatistics, University of Minnesota. Available at http://www.biostat.umn.edu/cgi-bin/rrs?.

[29] Press, W. H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B. P., (1992), *Numerical Recipes in C, The Art of Scientific Computing,* 2^{nd} ed. New York: Cambridge University Press.

[30] Schena, M., Shalon, D., Davis, R. W. and Brown, P. O., (1995), *Quantitative monitoring of gene expression patterns with a complementary DNA microarray*, Science, Vol. 270, pp. 467-470.

[31] Schwartz, G., (1978), *Estimating the dimensions of a model*, Annals of Statistics, Vol. 6, pp. 461-464.

[32] Storey, J.D., (2003), *The positive false discovery rate; a Bayesian interpretation and q-value.*, The Annals of Statistics, Vol. 31, No. 6, pp. 2013-2035.

[33] Tadjudi, S., and Landgrebe, D. A., (2000), *Robust Parameter Estimation For Mixture Model*, IEEE Transactions on Geoscience and Remote Sensing, Vol. 38, No. 1, pp. 439-445.

[34] Tamyo,P., Slonim,D., Mesirov,J., Zhu,Q., Kitareewan,S., Dmitrovsky,E., Lander,E.S., and Golub, T.R. (1999), *Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation.* Proc. Natl. Acad. Sci., Vol. 96, pp. 2907-2912.

[35] Tavazoie,S., Hughes,J.D., Campbell,M.J., Cho,R.J., and Church,G.M. (1999), *Sytematic determination of genetic network architecture.* Proc. Natl. Acad. Sci., Vol. 22, pp. 281-285.

[36] Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985), *Statistical Analysis of Finite Mixture Distributions,* New York: Wiley.

[37] Thode, H. C., Finch S. J. and Mendell N. R. (1988), *Simulated Percentage Points for the Null Distribution of the Likelihood Ratio Test for a Mixture of Two Normals.* Biometrics, Vol. 44, No. 4, pp. 1195-1201.

[38] Thomas,J.G., Olson,J.M., Tapscott,S.J. and Zhao,L.P. (2001), *An Efficient and Robust Statistical Modeling Approach to Discover Differentially Expressed Genes Using Genomic Expression Profiles.* Genome Research, Vol. 11, pp. 1227-1236.

[39] Tusher, V., Tibshirani, R. J., Chu, G., (2001), *Significant Analysis of Microarrays Applied to the Ionizing Radiation Response.*, Proc. Nat. Acad. Sci., Vol. 98, pp. 5116-5121.

[40] Verbeke, G., (2000), *Inference for mixed populations.*, Biostatical Center, Catholic University of Leuven.

[41] Wald, A., (1949), *Note on the consistency of the maximum likelihood estimate.*, Ann. Math. Statist., Vol. 20, pp. 595-601.

[42] Zhang, S., Jiao, S., (2007), *The t-mixture model approach for detecting differentially expressed genes in microarrays,* Functional & Integrative Genomics.

[43] Zhao. Y, Pan, W., (2003), *Modified nonparametric approaches to detecting differentially expressed genes in replicated microarray experiments,* Bioinformatics, Vol. 19, pp. 1046-1054.

About the Author

O'Neil was born in the beautiful island paradise of Jamaica where he completed his B.Sc. (1998) and M.Phil (2003) at the University of the West Indies (UWI), Mona Campus. He lectured part-time in the Departments of Economics and Mathematics (UWI) for the period 2000-2004.

His graduate studies at the University of South Florida made him realize that teaching and researching was his true calling. O'Neil's teaching skills were further enhanced while serving as a Teaching Assistant in the Department of Mathematics and Statistics (USF). He was also a Research Assistant in the Department of Epidemiology and Biostatistics (USF) and volunteered at the Moffitt cancer center for 2 years.

O'Neil is a reggae fanatic and his favorite dance hall reggae artist is Buju Banton. He looks forward to the Olympic Games to see how well the Jamaican sprinters will perform. He also enjoys playing/watching soccer and cricket.