The Pennsylvania State University

The Graduate School

Department of Statistics

ALMOST NONPARAMETRIC AND NONPARAMETRIC ESTIMATION IN MIXTURE MODELS

A Thesis in

Statistics

by

Isidro Roberto Cruz_Medina

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2001

We approve the thesis of Isidro Roberto Cruz_Medina.

Date of Signature

Thomas P. Hettmansperger Professor of Statistics Chair of Committee Thesis Advisor	
Calyampudi R. Rao Eberly Professor of Statistics	
Bruce G. Lindsay Distinguised Professor of Statistics	
Hoben Thomas Professor of Psychology	
James L. Rosenberger Professor of Statistics Head of the Department of Statistics	

Abstract

An almost nonparametric approach for the estimation of the mixing proportion in a mixture of two distributions, when we have a vector of observations on each subject, is to define a mixture of binomials. A mixture of binomials can be obtained if the vector of observations is mapped into a vector of zeroes and ones by selecting a cut point c. In this dissertation it is shown that the estimation of the cut point c, which minimizes the variance of the estimator of the mixing parameter, does not need to be very precise for some common distributions when the means of these distributions are more than two standard deviations apart. If more cut points are introduced a multinomial distribution is obtained and it is shown that the trinomial distribution is preferable to the binomial and the tetranomial is preferable to the trinomial distribution. In general, we prove that the multinomial distribution with $r + 1$ classes is preferable to the multinomial distribution with r classes. Nevertheless, it seems that if we introduce more than two cut points (a multinomial distribution with more than three regions) the gain in efficiency is not significant.

Nonparametric approaches are proposed for the estimation of the mixing parameter in a mixture of two continuous distributions with equal shapes and unimodal symmetric densities. In these approaches some cut points c_i are introduced in order to define a multinomial distribution, three cut points for a tetranomial distribution and five cut points for a sextinomial distribution. The assumed symmetry of the component distributions is exploited in order to obtain the probabilities for each class of the multinomial approach and five methods of estimation of the parameters of the multinomial mixture are studied. These methods basically measure the concordance among the observed frequencies and

the expected frequencies. We present Mathematica and S-plus program codes in order to obtain the estimates of the parameters in the multinomial mixture. A Monte Carlo study shows that for normal components, the estimators of the mixing proportion in the sextinomial approaches are comparable with the EM algorithm estimator if the means are 1.75 standard deviations apart, but the estimators of the sextinomial approaches have an efficiency of 50% with respect to the EM estimator when the distance between the means is 2.32 standard deviations. When the component distribution are not normal, the sextinomial approaches outperform the EM algorithm that assumes that the components are normal.

These tetranomial and sextinomial approaches can be easily adapted for use with training samples and three methods of sampling are considered. With training samples and normal components, the estimators from the sextinomial methods are comparable with the EM algorithm estimator. However, when component distributions are not normal, the sextinomial estimators outperform the EM algorithm estimator which assumes that the component distributions are normal.

Table of Contents

List of Tables

Acknowledgments

I wish to express my sincere gratitude to my thesis advisor, Professor Thomas Hettmansperger, for his teaching, guidance, encouragement and friendship. I want to thank my committee members, Professors C. R. Rao, Bruce Lindsay and Hoben Thomas for their time, teaching and valuable comments. I would also like to thank David Hunter for his help with my Splus codes.

I want to give thanks to my family: my wife, Rosa Maria, my eldest son Roberto, my daughter Paola and my youngest son Luis Gerardo for their support, patience and unconditional love. I would also like to thank my parents, my mother Socorro and my late father Ezequiel, and my uncles Victor and Juan and my late aunt Carmen for their encouragement throughout my childhood.

Finally, I want to thank the Instituto Tecnológico de Sonora, México, for their support and allowing me to take a leave of absence, to the Consejo Nacional de Ciencia y Tecnologia, México for having granted me a scholarship and to the Department of Statistics at Penn State for giving me this great opportunity.

Chapter 1

Background and Motivation

1.1 Introduction

The objective of many scientific investigations can be formulated as the following interesting problem:

• If we know that a population has K distinct subpopulations, how to estimate the proportions of these subpopulations and the parameters that define these subpopulations?

• How to identify the number of distinct subpopulations or groups? In genetics, for example, such groups may be determined by the distinct genotypes in the population and the number of genotypes may be of primary interest. In a fish population, the size distribution depends on the age in years of the fish, and the population from which samples are taken is a mixture of K age groups.

• How to specify an individual as a member of one group to which he can possibly belong?

• How to classify the groups themselves into some significant system based on the configuration of the most important characteristics?

When we consider the problem of assigning an individual to one of a finite number of groups to which he may belong on the basis of a set of random variables that we have observed we are in the realm of decision theory and in order to look for a satisfactory solution we need to know the following: The probability density functions for the set of measurements on each individual for each of the populations, the prior probabilities for

the populations and a loss function that specifies the loss in identifying an individual from one of the populations when it belongs to another population.

If a set of observations is modeled satisfactorily by a mixture model, the former requirements are fulfill and we can use this mixture model for classifying observations.

1.2 Mixture Models

Although mixtures are one of statistic's oldest structures, going back at least to Karl Pearson in 1894, until more recently they had not been studied very extensively, nor have they been applied often to practical settings. Clearly one main problem is computation; parameter estimates cannot in general be obtained in closed form from mixture structures. Conventional algorithms, such as the Newton-Raphson, have long been known to lead to difficulties; see Lindsay (p.65, [21]). The computational issue has largely been resolved, however, with the development of the EM algorithm by Dempster, Laird and Rubin [7]. See McLachlan and Krishnan [25] for a detailed account of the EM algorithm.

Mixture models arise in many areas of application where it is known or suspected that observations come from different populations, each of which has a different distribution. A common problem in practice is to estimate the proportions in which a specified number of populations occur, for example the proportions of some birds or mammals in some specific geographical region.

Titterington, Smith and Makov [34] describe in detail several areas of application of finite mixtures models and present an extensive summary of examples of applications.

We shall consider only finite mixtures in which all component distributions come from the same parametric class.

$$\mathcal{F} = \{\ F(x,\theta),\ \theta \in \Omega,\ x \in R^{\,n}\ \}$$

Let $\mathcal{F}$ be the class of m-dimensional distribution functions from which mixtures are to be formed. We identify the class of finite mixtures of $\mathcal{F}$ with the appropriate class of distributions functions, $\mathcal{G}$, defined by

$$\mathcal{G} = \left\{ G(x) :\ G(x) = \sum_{j=1}^{k} \pi_j\ F(x,\ \theta_j),\ \pi_j > 0,\ \sum_{j=1}^{k} \pi_j = 1,\ x \in R^{\,n} \right\}$$

Mixtures of distributions present at least three types of problems.

The first is the problem of identifiability; that is, given that a distribution function G is a probability mixture of distribution functions belonging to some family $\mathcal{F}$, is the mixture unique ?

In general, a necessary and sufficient condition for G to be identifiable is that $\mathcal{F}$ be a linearly independent set over the field of real numbers R, Yakowitz and Spragins [38].

For a mixture of binomials $b(y;\ m,\ p_j)$, this condition is equivalent to the condition that $m \geq 2K - 1$, where K is the number of components of the binomial mixture (Blischke [2]).

The second problem occurs when the number of components is unknown and we need to estimate it. McLachlan [23] mentions that an obvious way of assessing the number of components in a mixture model is to use the likelihood ratio test statistic. Unfortunately with mixture models, regularity conditions do not hold for this statistic to have the usual asymptotic distribution. Everitt [10] mentions that the usual asymptotic distribution

is attained only for sample sizes above fifty, and for data where the sample size is ten times the number of variables. For a mixture of multinomial models Lindsay [21] shows that the distribution of the LRT of 1 versus 2 components is asymptotically equivalent to the distribution of the squared length of the projection of standardized data onto the corresponding corrected score tangent cone.

Bozdogan [3], discusses the utilization of the Akaike's Information Criterion (AIC). Lindsay and Roeder [22] and Roeder [29], proposes graphical (residual and gradient plots) techniques for diagnostic purposes. Chen and Kalbfleisch [4] discuss a penalized minimum-distance estimate for estimating consistently the number of components in a finite mixture model, and Chen, Chen and Kalbfleisch [5] presents a modified likelihood ratio test where the ratio test statistic has the simple χ^2-type null limiting distribution.

The third problem is that of estimating the parameters of the individual distribution functions comprising the mixture and the mixing measure when the number of components is known. Karl Pearson [27] considered the problem of estimating the parameters of a mixture of two normal distributions with the method of moments. The estimation of the five parameters in this case depend on a suitably chosen root of a ninth degree equation constructed from the first five moments of the observed frequency distribution.

Tan [31] mentions that the method of moments was used for estimating the parameters of a mixture of two normal distributions until Rao [28] considered the same problem with the method of maximum likelihood. Considering equal variances there are only four parameters to be estimated. Rao estimated the four parameters with the method of computation known as the scoring system. The computational issue, as mentioned before, was partially solved with the development of the EM algorithm by Dempster, Laird and Rubin [7].

1.3 Some Applications of Mixture Models

In general it seems that mixture models represent a fertile approach for classifying individuals in many areas. Thomas and Lohaus [33], used binomial mixtures for classifying individuals in performance tests and for modeling judgments and strategies. Do and McLachlan [9] successfully solve the problem of determining the diet of the owls in some region of Malaysia estimating the proportion of each species of rat consumed by the owls. This problem was successfully solved with a mixture model assuming a multivariate normal distribution. The conclusion was that the rat diet of the owls is composed mainly for one species with the remainder of the diet consisting mostly of another species.

This example is similar to the one described by Titterington et al [34], in which the underlying categories of fish (age groups) were described by the length distributions of fish modeled by a mixture model of normal distributions.

Fowlkes [12] describes the use of mixture models in industry. Some lasers are employed in telephone communication system in which coherent laser light is used to transmit telephone conversations. It is necessary to predict how long all manufactured lasers should be life tested to assure that the shipped product contain no products with very short life. For this purpose an experiment was established in which 103 laser devices were operated in an elevated temperature environment until all had failed. Since most of the devices were extremely reliable, the experiment ran longer than one year before all the devices failed. Fortunately there was early recognition that the sample probably represented two distinct populations, one with a very short mean life (group with "infant mortality") and one with a much longer life (group of normal devices). Survival times were satisfactorily modeled by a mixture of lognormal distributions.

The main pitfall in applications of mixture models is that usually there is no justification for choosing the form of the distribution of the components of the mixture (the usual mixture of normal distributions is not appropriate in many cases). In the last example it is apparent also that many distribution functions can be used to model survival times.

There are many situations in which it would be desirable not to have to specify parametric distributions for the components. The default option in mixture decomposition, at least in the continuous case, is to assume a mixture of normal distributions. Hall [13] proposed a nonparametric approach to the estimation of the mixing proportions. His method, however, requires knowledge of the number of components and possession of training samples from each of the components.

Hettmansperger and Thomas [16] started with a vector of observations on each subject [the vector may have independent and identically distributed (iid) components or they may be exchangeable] in order to avoid the specification of component distributions. They transform the data to a binary sequence. This is achieved by replacing the data vector by a vector of 1s and 0s determined by whether the respective data vector components are below or above of a cut off point c, for $i = 1, \ldots, m$ and $j = 1, \ldots, n$. Let.

$$Y_{ij} = \begin{cases} 1 & \text{if} \quad X_{ij} \leq c \\ 0 & \text{otherwise.} \end{cases}$$

where $X_{1j}, \ldots, X_{mj}$ are iid for $j = 1, 2, \ldots, n$, then $Y_j = \sum_{i=1}^{m} Y_{ij}$ is simply the number of times the observations in the jth observation vector are less than or equal to c. In order to have identifiability in the binomial mixtures we must have $m \geq 2K - 1$ where k is the

number of components of the binomial mixture; see for example Blischke [2]. Then, for fixed c, $Y_1, \ldots, Y_n$ are iid and

$$h(y) = \lambda\, b\,(\, y; m, F_1(c)\,) + (1-\lambda)\, b(\, y; m, F_2(c)\,)$$

where $F_r\,(c) = \int_{-\infty}^{c} f_r(x)dx$, $r = 1, 2$ and $b(\,.\,; m, p)$ denotes the binomial mass function with parameters m and p. Hence, given a value of c, they use the mixed binomial likelihoods and the EM algorithm to estimate λ. In addition, using the Fisher information, they derive an estimate of the standard error of the estimate of λ for comparison with other approaches. Additionally they derived the asymptotic distribution of the estimator and found that it only weakly depends on the underlying mixture model. Given a value of c, we do not need to specify a model for f_1 and f_2 to estimate λ. Of course, the asymptotic variance of the estimate depends on the underlying distribution, but only through $F_1(c)$ and $F_2(c)$.

1.4 Thesis Outline

In Chapter 2, we will obtain the optimal cut points c for the normal distribution and for three common different distributions in order to minimize the asymptotic variance of the mixing parameter.

In Chapter 3, we will explore the inclusion of more cut points in order to define firstly a trinomial distribution, secondly a tetranomial and in general a multinomial approach in order to minimize the asymptotic variance of the mixing parameter. We will obtain expressions for the asymptotic variance of the mixing parameter for the binomial, trinomial and in general for the multinomial approach when $m \longrightarrow \infty$. We will compare

the asymptotic variance of the binomial approach with the asymptotic variance of the multinomial approach for finite number of components in the binomial and multinomial distributions (finite sample size) and for m, when $m \longrightarrow \infty$.

In Chapter 4, we will analyze the properties of a completely non parametric approach to estimate the mixing parameter in a mixture of two symmetric component distributions. We will compare this general approach with the parametric approach when the assumptions hold and when they are not true for at least four common distributions. We will also analyze the case in which training samples (samples where the component of origin of each observation is known with certainty) are available. In this case, two types of training samples are distinguished, depending whether or not the training sample contains information about the mixing proportion.

Finally, in Chapter 5, a summary will be presented and we will describe how the nonparametric approach proposed in Chapter 4 (which exploits the symmetry of the component distributions) can be generalized for a mixture of three symmetric component distributions.

Chapter 2

Binomial Approach

In this chapter, in order to estimate the mixing proportions in a finite mixture distribution, we consider an approach that does not make parametric assumptions about the component distributions. We require a vector of observations on each subject. This vector is mapped into a vector of zeros and ones and summed. The resulting distribution of sums under certain conditions will be modeled as a mixture of binomials.

2.1 Introduction

Let $p_1(x)$ and $p_2(x)$ be probability distributions with λ_1 and λ_2 positive real numbers that sum to one. Then $p(x) = \lambda_1\, p_1(x) + \lambda_2\, p_2\,(x)$ defines a two component mixture distribution with components $p_1(x)$ and $p_2(x)$ and mixture weights λ_1 and λ_2. Mixture distributions arise when observations come from different populations, each with a different distribution, and the population sampled is unobserved. Interesting examples include those in the social sciences in which important variables are latent and unobserved. As an example, consider the problem of trying to infer the solution strategies young children employ in solving a cognitive task when different solution strategies can be employed. A group of same age children can be considered a sample from a mixture distribution in which the components correspond to the various strategies; see Thomas and Horton [32]. Recently, there has been an expanding interest in mixtures and there is now a large literature.

Examples and theory may be found in many sources including Lindsay [21], McLachlan and Basford [24], Titterington, Smith, and Makov [34], and Everitt and Hand [11].

2.2 Nonparametric Estimation in Mixture Models

Tamura [30] makes an analogy between robust estimation and the solution of an insurance problem. A classical estimator that is optimal for some ideal model may be quite sensitive to small deviations from this model. In order to insure against "accidents" caused by these deviations, one has to pay for this security by sacrificing some efficiency at the model. Murray and Titterington [26] mention that parametric approaches run into difficulties when there are many observations from the mixture and the data are multivariate. Furthermore, the underlying data may clearly not come from parametric distributions amenable to analysis by the EM algorithm or by Bayesian methods. They consider kernel functions for the estimation of the mixing parameter. Hall [13] pointed out some significant drawbacks to the methods based on non-parametric density estimators and mentioned that the most obvious is the need to specify the "window size". Then he proposed estimators based on the empirical distribution function. Hall and Titterington [14] defined a whole class of estimators by "pseudo-convex" combinations of nonparametric estimators.

Hettmansperger and Thomas [16] proposed an almost nonparametric approach to the estimation of the mixing parameters. They make no assumptions about the parametric form of the component distributions but they require a vector of observations on each subject. The vector may have independent and identically distributed (iid) components or they may be exchangeable. In order to avoid the specification of component distributions, they transform the data to a binary sequence. This is achieved by replacing the data vector by a vector of 1s and 0s determined by whether the respective data vector components

are inside or outside of a specified region. Then each vector is replaced by a sum of 0s and 1s and the sums are modeled with mixtures of binomial distributions. The number and values of the component weights in the binomial mixture are the same as in the original model. In the next section we describe the model and restrict attention to vectors that have iid components. We also present the limiting distribution of the estimator of the mixing parameter in a two component mixture given by Hettmansperger and Thomas [16]. We then discuss how to determine the optimal cut point c.

2.3 A Model-Free Approach with the Binomial

Suppose $\mathbf{X}_1, ..., \mathbf{X}_n$ are independent and identically distributed (iid) m-variate random vectors with joint probability density function

$$p(\mathbf{x}) = \lambda\, p_1(\mathbf{x}) + (1-\lambda)\, p_2(\mathbf{x})\ , \quad 0 \le \lambda \le 1$$

We restrict discussion to the 2-component mixture model for the sake of simplicity; the results for the general case can be readily written down. The initial part of the discussion will be limited to the conditional iid case defined next.

Definition: Conditional Independence. When the multivariate component density $p_r(\mathbf{x})$ is equal to $\prod_{i=1}^{m} f_r(x_i)$ for $r = 1, 2$ we say that, given the component distribution, the observations are independent and identically distributed.

The proposal of Hettmansperger and Thomas [16] is to reduce the multivariate data to sums of binary responses. They introduce a cut point c and define for $i = 1, \ldots, m$ and $j = 1, \ldots, n$.

$$Y_{ij} = \begin{cases} 1 & \text{if} \quad X_{ij} \leq c \\ 0 & \text{otherwise.} \end{cases}$$

Then $Y_j = \sum_{i=1}^{m} Y_{ij}$ is simply the number of times the observations in the jth observation vector are less than or equal to c. In order to have identifiability in the binomial mixtures we must have $m \geq 2K - 1$ where K is the number of components of the binomial mixture; see for example Blischke [2]. They proved the next two theorems.

THEOREM 2.1. *Assume the mixture model with conditional iid structure and* $m \geq 3$. *Further, assume that* $F_1(c) \neq F_2(c)$. *Then, for fixed* c, $Y_1, ..., Y_n$ *are iid*

$$h(y) = \lambda\, b(y; m, F_1(c)) + (1 - \lambda)\, b(y; m, F_2(c)) \tag{2.1}$$

where $F_r(c) = \int_{-\infty}^{c} f_r(x)dx$, $r = 1, 2$ *and* $b(\,.\,; m, p)$ *denotes the binomial mass function with parameters* m *and* p.

The extension to more than two components is immediate. Note that we have the same λ in the binomial mixture as in the original mixture model. Hence, given a value of c, we can use the mixed binomial likelihoods and the EM algorithm to estimate λ; see McLachlan and Krishnan [25] for the EM algorithm. In addition, using the Fisher information, we can derive an estimate of the standard error of the estimate of λ for comparison with other approaches.

When c is specified, it is straight forward to compute the information from the binomial mixture. Further, the asymptotic distribution of $\widehat{\lambda}$, the maximum likelihood estimator, is normal. This is summarized in the following theorem.

THEOREM 2.2. *Assume the mixture model with conditional iid structure. For a given c suppose $F_1(c) \neq F_2(c)$ and define*

$$g(y, \lambda, c) = \frac{b(y; m, F_1(c)) - b(y; m, F_2(c))}{\lambda\, b(y; m, F_1(c)) + (1-\lambda)\, b(y; m, F_2(c))}$$

where $b(\,.\,; m, p)$ is the binomial mass function with parameters m and p. Let $0 < \lambda_0 < 1$ denote the true value of the parameter λ. Then $\widehat{\lambda}$ solves $\sum g(y_j, \lambda, c) = 0$ and, as $n \to \infty$,

$$\sqrt{n}(\widehat{\lambda} - \lambda_0) \xrightarrow{D} Z \sim N\left(0, \frac{1}{Eg^2(Y, \lambda_0, c)}\right) \tag{2.2}$$

where

$$Eg^2(Y, \lambda_0, c) = \sum_{i=0}^{m} \frac{[b(i; m, F_1(c)) - b(i; m, F_2(c))]^2}{\lambda_0\, b(i; m, F_1(c)) + (1-\lambda_0)\, b(i; m, F_2(c))}. \tag{2.3}$$

The asymptotic distribution of the estimator only weakly depends on the underlying mixture model. Given a value of c, we do not need to specify a model for f_1 and f_2 to estimate λ. Of course, the asymptotic variance of the estimate depends on the underlying distribution, but only through $F_1(c)$ and $F_2(c)$.

A generalization of theorems 2.1 and 2.2 for the trinomial approach are given in theorems 3.1 and 3.2 in Chapter 3.

We now turn to the problem of choosing c, the cut point that defines the binary variables.

2.4 Optimal Choice of c

Our optimal choice of c is determined by minimizing the asymptotic variance of $\sqrt{n}(\widehat{\lambda} - \lambda_0)$, in other words, for equations (2.2) and (2.3) the value of c which maximizes

$$Eg^2(Y,\lambda_0,c)=\sum_{i=0}^{m} g_i^2=\sum_{i=0}^{m}\frac{[b(i;m,F_1(c))-b(i;m,F_2(c))]^2}{\lambda_0\, b(i;m,F_1(c))+(1-\lambda_0)\, b(i;m,F_2(c))}.$$

If f_1 and f_2 are symmetric distributions with the same shape but different location parameters μ_1 and μ_2, for $\lambda_0 = 0.5$ the expectation $Eg^2(Y,\lambda_0,t)$ is symmetric with respect to the value $c=\frac{\mu_1+\mu_2}{2}$, because in this case $F_1(c-t)=1-F_2(c+t)$. This implies that $b\,(i;m,F_1(c)\ )=b\,(m-i;m,F_2(c)\ )$. The following Theorem shows the function $Eg^2(Y,1/2,c)$ has a local maximum at $c=\frac{\mu_1+\mu_2}{2}$.

THEOREM 2.3. *$Eg^2(Y,1/2,c)$ has a local maximum at $c=\frac{\mu_1+\mu_2}{2}$ when λ_0 =0.5 .*

Proof : First we will show that the derivative of the function $Eg^2(Y,1/2,c)$ at $c=\frac{\mu_1+\mu_2}{2}$ is zero.

Let $b\ (i;m,F_1(c)\ )=b_i\ (F_1(c)\ )$.

$$\frac{\partial}{\partial c}g_i^2=u_i-v_i$$

$$=\frac{[b_i(F_1(c))-b_i(F_2(c))]\{[b_i^{'}(F_1(c))]F_1^{'}(c)-[b_i^{'}(F_2(c))]F_2^{'}(c)\}}{b_i(F_1(c))+b_i(F_2(c))}$$

$$-\frac{[b_i(F_1(c))-b_i(F_2(c))]^2\{[b_i^{'}(F_1(c))]F_1^{'}(c)+[b_i^{'}(F_2(c))]F_2^{'}(c)\}}{2\,[b_i(F_1(c))+b_i(F_2(c))]^2}$$

For $c=\frac{\mu_1+\mu_2}{2}$, if $p=F_1(c)$, then $F_2(c)=1-p$, besides $F_1^{'}(c)=F_2^{'}(c)=f_1(c)$ therefore:

$$\sum_{i=0}^{m} u_i = \sum_{i=0}^{m} \frac{[b_i(p) - b_i(1-p)][b_i^{'}(p) - b_i^{'}(1-p)]f_1(c)}{b_i(p) + b_i(1-p)} = 0$$

Because $b_i^{'}(p) = i\binom{m}{i}p^{i-1}(1-p)^{m-i} - (n-i)\binom{m}{i}p^i(1-p)^{m-i-1}$, we have that $b_i^{'}(1-p) = -\, b_{m-i}^{'}(p)$. We have also:

$$\sum_{i=0}^{m} v_i = \sum_{i=0}^{m} \frac{[b_i(p) - b_i(1-p)]^2[b_i^{'}(p) - b_{m-i}^{'}(p)]f_1(c)}{[b_i(p) + b_i(1-p)]^2} = 0$$

Now we will show that the second derivative of $Eg^2(Y, 1/2, c)$ with respect to c is negative.

Observe that $F_1^{''}(c) = -\, F_2^{''}(c) = \ a$ at $c = \frac{\mu_1+\mu_2}{2}$.

$$\frac{\partial^2}{\partial c^2} g_i^2 = u_i + v_i - w_i$$

$$= \frac{4\,[f_1(c)]^2\ \{\ b_i(1-p)\ b_i^{'}(p) - b_i(p)\ b_i^{'}(1-p)\}}{[a\ b_i(p) + a b_i(1-p)\ ]^3}$$

$$+ \frac{a^2\ [b_i(p) - b_i(1-p)]\ [b_i^{''}(p) - b_i^{''}(1-p)]}{b_i(F_1(c)) + b_i(F_2(c))}$$

$$- \frac{a^2\ [b_i(p) - b_i(1-p)]^2\ [b_i^{''}(p) + b_i^{''}(1-p)]}{b_i(F_1(c)) + b_i(F_2(c))}$$

We have that $F_1^{''}(c) < 0$, therefore $\sum_{i=0}^{m} u_i < 0$, by the argument given before $b_i^{''}(p) = -\, b_i^{''}(1-p)$ this implies by symmetry that $\sum_{i=0}^{m} v_i = 0$ and $\sum_{i=0}^{m} w_i = 0$.

When f_1 and f_2 are symmetric distributions with the same shape and different location parameters but λ_0 is different from 0.5 the expectation $Eg^2(Y, \lambda_0, t)$ is no longer symmetric with respect to $c = \frac{\mu_1+\mu_2}{2}$ and this value does not minimize the asymptotic variance of $\sqrt{n}(\widehat{\lambda} - \lambda_0)$.

For f_1 and f_2 symmetric distributions with the same shape and different location parameter and λ_0 different from 0.5 we will have another type of symmetry. If $Eg^2(Y, \lambda_0, t)$ is minimized for $c = c_0$ by symmetry $Eg^2(Y, 1-\lambda_0, t)$ will be minimized for $c = \mu_1+\mu_2-c_0$ and for these values of c, $Eg^2(Y, \lambda_0, t) = Eg^2(Y, 1-\lambda_0, t)$ or in other words the asymptotic variances will have the same value.

This value of $c = \frac{\mu_1+\mu_2}{2}$ coincides with the value given by Hettmansperger and Thomas [16] determined by maximizing $|F_1(c) - F_2(c)|$. These authors proved also a theorem that shows how to construct an optimal region to maximize the separation between the binomial parameters.

When f_1 and f_2 are not symmetric densities this theorem does not give the optimal value of c.

In general it seems difficult to minimize analytically the asymptotic variance $A.V$ of $\sqrt{n}(\widehat{\lambda} - \lambda_0)$ in terms of c. This minimization can be done numerically and it seems that the $A.V$ will depend on $\delta = \mu_1 - \mu_2$ and on $\rho = \sigma_1 / \sigma_2$.

For example, if f_1 is a $N(0, 1)$ and f_2 is a $N(1, 1)$ and $m = 3$, the asymptotic variance can be expressed in terms of c and λ. Figure 2.1 shows that the asymptotic variance does not change very much for $c \in (0, 1)$ when $\lambda = .5$. This characteristic is very important because in a real situation we do not know the parameters of the distribution and therefore we can not obtain the optimal point c.

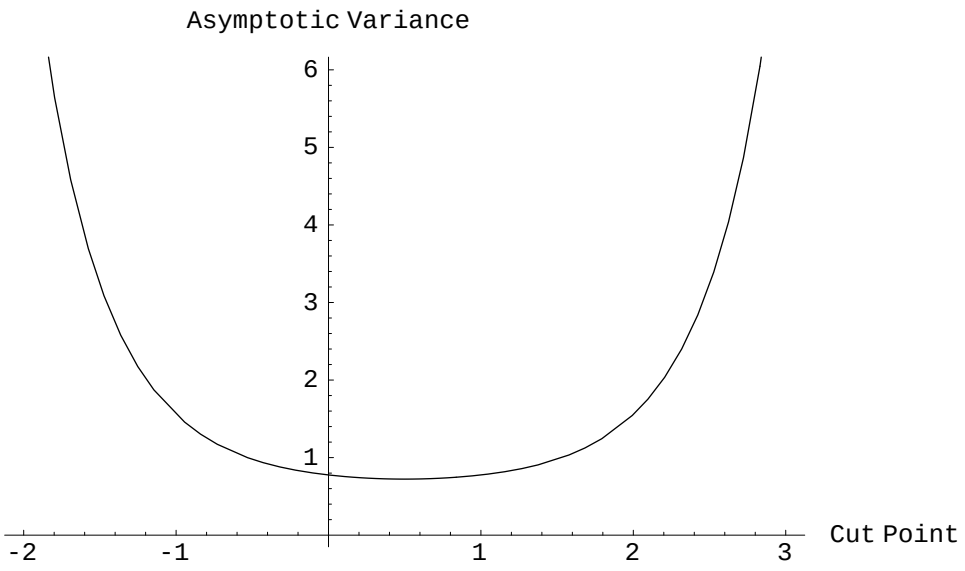

Fig. 2.1. $A.V$ of $\sqrt{n}(\widehat{\lambda} - \lambda_0)$, in terms of c when λ=0.5 and $\mu_1 - \mu_2 = \sigma$ for normal distributions

This characteristic of lack of sensitivity of the $A.V$ is even more apparent if f_1 is a $N(0, 1)$ and f_2 is a $N(2, 1)$ for $m = 3$. In this case the $A.V$ is also smaller than in the former case. Figure 2.2 shows this fact.

Figure 2.3 shows the asymptotic variance when f_1 is a Cauchy distribution (0, 1) and f_2 is a Cauchy distribution (2, 1) for λ =0.5, where the density function for a Cauchy distribution (θ, 1) with median θ and scale parameter equal to 1 is given by:

$$f(x \mid \theta) = \frac{1}{\pi \ [1 + (x - \theta)^2]}$$

Figure 2.4 shows the asymptotic variance for non-symmetric distributions; in this figure f_1 is an Exponential distribution, Gamma (1, 1) and f_2 is a Gamma (1, 3) for λ =0.5, where the density function for a Gamma distribution (1, θ) with scale parameter equal to θ is given by:

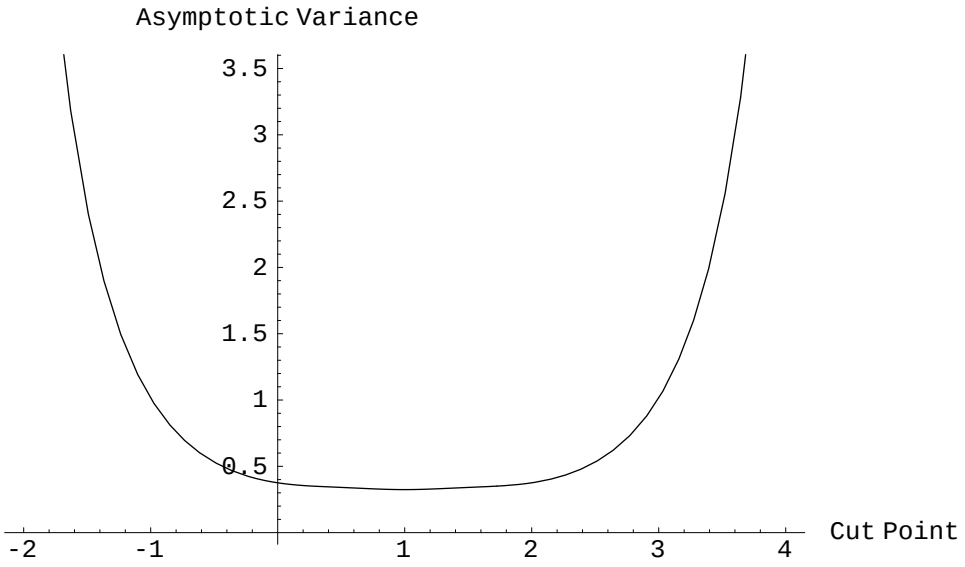

Fig. 2.2. *A.V* of $\sqrt{n}(\widehat{\lambda} - \lambda_0)$, in terms of c when λ=0.5 and $\mu_1 - \mu_2 = 2\sigma$ for normal distributions

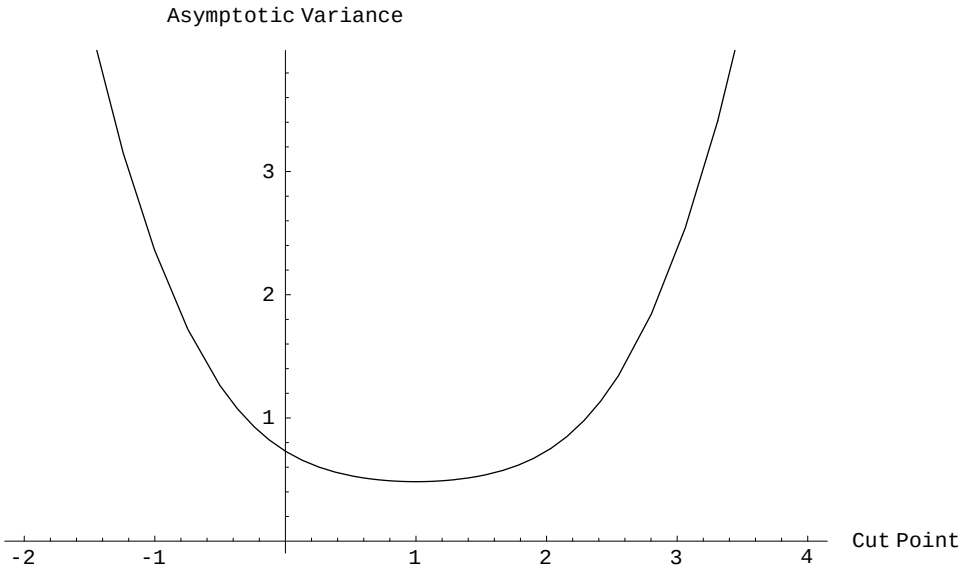

Fig. 2.3. *A.V* of $\sqrt{n}(\widehat{\lambda} - \lambda_0)$, in terms of c when λ=0.5 for Cauchy distributions

$$f(x \mid \theta) = \frac{1}{\theta}\, e^{-x/\theta}$$

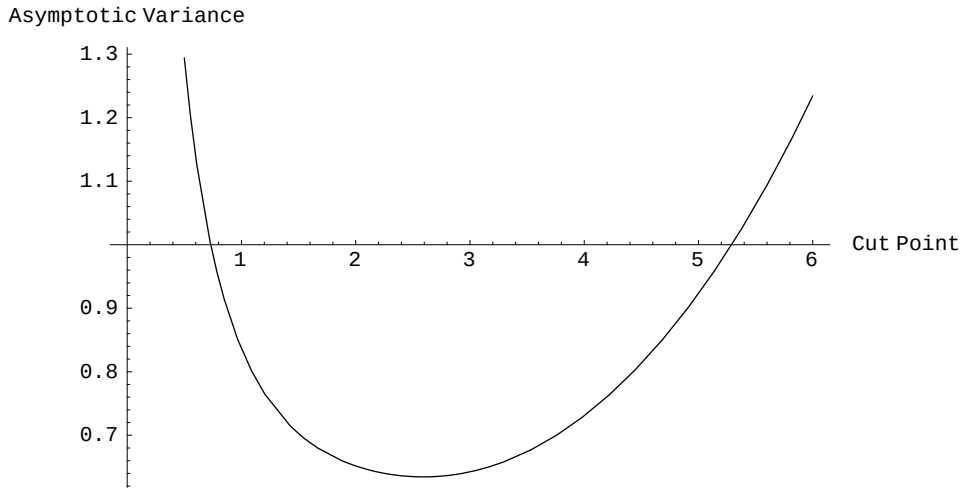

Fig. 2.4. *A.V* of $\sqrt{n}(\widehat{\lambda} - \lambda_0)$, in terms of c when λ=0.5 for Gamma distributions

These graphics show that the asymptotic variance of $\sqrt{n}(\widehat{\lambda}-\lambda_0)$ is not very sensitive to small changes of c for some common distributions.

We will analyze now the variation of the asymptotic variance of $\sqrt{n}(\widehat{\lambda} - \lambda_0)$ for normal distributions. In Figure 2.5 we have the asymptotic variance of $\sqrt{n}(\widehat{\lambda} - \lambda_0)$ for different values of c and λ. In this figure due to the scale factor, it seems that the surface is flat in the interval $(0, 1)$, but in fact, the form of the function is a saddle. Figure 2.6 presents with more detail the variance in this interval. In this figure we can observe that

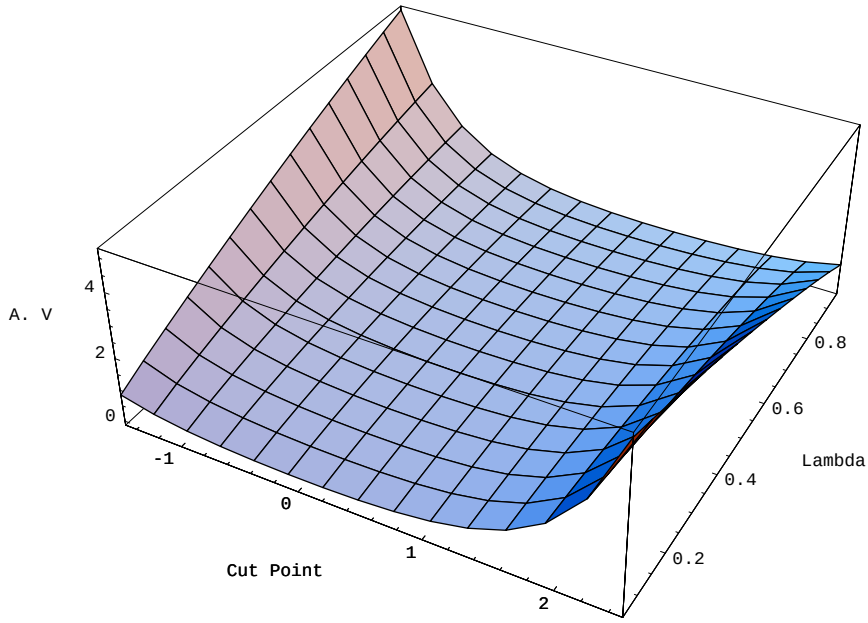

Fig. 2.5. $A.V$ of $\sqrt{n}(\widehat{\lambda} - \lambda_0)$, in terms of c and λ for normal distributions

the asymptotic variance of $\sqrt{n}(\widehat{\lambda} - \lambda_0)$ decreases when λ approaches the limits of the interval $(0, 1)$.

Figure 2.6 shows that the minimum asymptotic variance of $\sqrt{n}(\widehat{\lambda} - \lambda_0)$ when $\lambda_0 = 0.5$ is attained at $c = 0.5$ for all values of m in accordance with Theorem 2.3. We can observe also that the minimum asymptotic variance of $\sqrt{n}(\widehat{\lambda} - \lambda_0)$ is maximum when $\lambda_0 = 0.5$. Table 2.1 presents the optimal values of c and the asymptotic variance of $\sqrt{n}(\widehat{\lambda} - \lambda_0)$ attained for different values of λ and m for normal distributions.

In Table 2.1 we can observe some interesting features: first we note that the asymptotic variance of $\sqrt{n}(\widehat{\lambda} - \lambda_0)$ decreases with m for all values of λ_0. Secondly, the optimal cut point c is equal to 0.5 for λ_0=0.5 in accordance with theorem 2.3 at the beginning of this section. Finally, we can observe that the optimal cut point c is symmetric with respect to c and λ_0 in the sense that for λ_0=0.4 and λ_0=0.6 the optimal cut point is

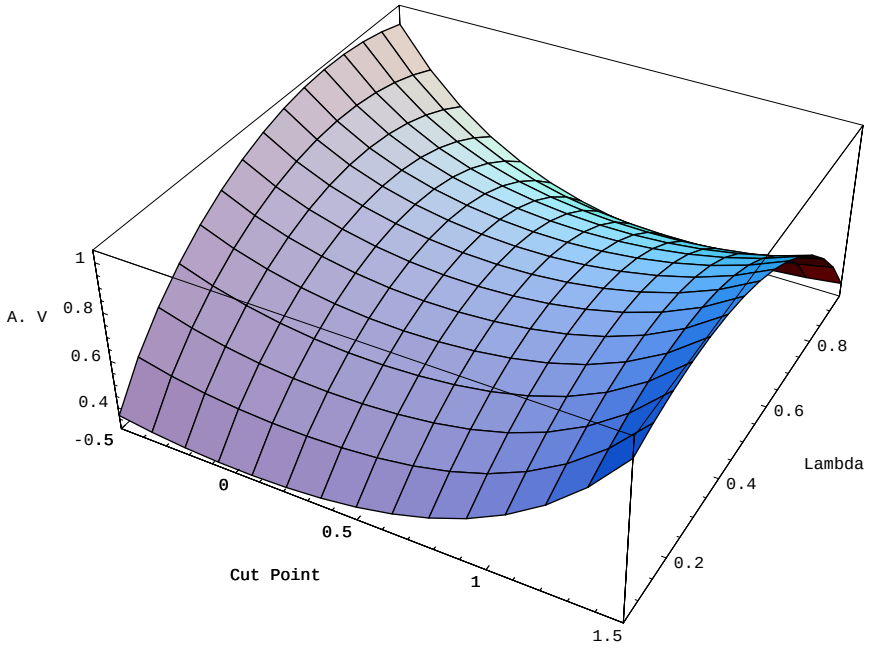

Fig. 2.6. $A.V$ of $\sqrt{n}(\widehat{\lambda} - \lambda_0)$, in terms of $c \, \epsilon \, (-0.5, 1.5)$ for normal distributions

Table 2.1. $A.V$ of $\sqrt{n}(\widehat{\lambda} - \lambda_0)$ for different values of λ_0 and m.

λ_0/m	3 $(c, A.V)$	4	6	10	15
0.1	0.1392, 0.4239	0.1950, 0.3209	0.2669, 0.2239	0.3409, 0.1528	0.3873, 0.1213
0.2	0.2638, 0.5632	0.3099, 0.4472	0.3550, 0.3329	0.4013, 0.2443	0.4272, 0.2029
0.3	0.3526, 0.6537	0.3854, 0.5301	0.4127, 0.4057	0.4406, 0.3069	0.4547, 0.2597
0.4	0.4287, 0.7056	0.4457, 0.5780	0.4586, 0.4481	0.4719, 0.3438	0.4781, 0.2934
0.5	0.5000, 0.7227	0.5000, 0.5937	0.5000, 0.4621	0.5000, 0.3560	0.5000, 0.3046
0.6	0.5712, 0.7056	0.5542, 0.5780	0.5413, 0.4481	0.5280, 0.3438	0.5218, 0.2934
0.7	0.6473, 0.6537	0.6145, 0.5301	0.5873, 0.4057	0.5593, 0.3069	0.5452, 0.2597
0.8	0.7361, 0.5632	0.6900, 0.4472	0.6449, 0.3329	0.5986, 0.2443	0.5727, 0.2029
0.9	0.8607, 0.4239	0.8049, 0.3209	0.7330, 0.2239	0.6590, 0.1528	0.6126, 0.1213

symmetric with respect to 0.5. In the next chapter we will investigate how the inclusion of more cut points affects the asymptotic variance of the estimator, and we will obtain the limit of this variance when $m \longrightarrow \infty$.

2.5 Summary

In this chapter, we introduced the model-free approach given by Hettmansperger and Thomas [16] for the estimation of the mixing parameter of a two component mixture distribution. We proposed an optimal choice for the cut point c, which defines the binomial parameters, as the value which minimizes the asymptotic variance of $\sqrt{n}(\widehat{\lambda}-\lambda_0)$. We found that for normal, Cauchy and gamma component distributions the estimation of c does not need to be precise because the asymptotic variance does not change significantly in a wide range of values. This characteristic is very important because in a real problem we do not know the parameters of the distribution and therefore we can not obtain the optimal point c. This characteristic of lack of sensitivity of the $A.V$ is even more apparent when the distance between the means of the two component distributions increases and, in this case, not only the $A.V$ decreases (as we expect) but also the length of the interval (in which the $A.V$ does not change significantly) increases.

The numerical results presented in Table 2.1 give some interesting facts: First we note that the asymptotic variance of $\sqrt{n}(\widehat{\lambda} - \lambda_0)$ increases when the distance between λ_0 and 1/2 increases. Secondly, the asymptotic variance of $\sqrt{n}(\widehat{\lambda} - \lambda_0)$ decreases with m for all values of λ_0. Finally, we can observe that the optimal cut point c is symmetric with respect to c and λ_0 in the sense that for λ_0=0.4 and λ_0=0.6 the optimal cut point is symmetric with respect to 0.5.

Chapter 3

Multinomial Approach

3.1 Introduction

As we mentioned before, we can include more cut points, provided there is enough data (large m). This leads to multinomial rather than binomial mixtures. In this chapter we will analyze the effect of including additional cut points on the estimation efficiency of λ.

3.2 Identifiability in multinomial mixtures

For the binomial distribution, Blischke [2] used the method of moments in order to determine the conditions for identifiability in a mixture of binomials, we will use the same method for multinomial mixtures.

Let $b(y; m, p)$ be the binomial mass function with parameters m and p. For this distribution, we have $m+1$ equations (probabilities): $b(i; m, p),\ i = 0, 1, 2, \ldots, m$ and m out of them are independent because $\sum_{i=0}^{m} b(i; m, p) = 1$. On the other hand in a mixture of K binomials we have $2K-1$ parameters. K parameters $p_j,\ j = 1, 2, \ldots, K$; and $K-1$ independent parameters λ_j because $\sum_{i=1}^{K} \lambda_j = 1$. Therefore the mixture of binomials will be identifiable if we have more independent equations $b(i; m, p)$ than parameters or equivalently if $m \geq 2K-1$.

For the trinomial distribution the following table permits us to see that there are $(m+1)(m+2)/2$ equations, with one restriction $\sum_{x=0}^{m} \sum_{y=0}^{x-y} t(x\,, y\,;\, m\,, p_1,\, p_2) = 1$.

Where $t\,(x, y;\ m,\ p_1,\ p_2)$ denotes the trinomial mass function with parameters m, p_1 and p_2.

Table 3.1. Identifiability for the trinomial distribution

x/y	0	1	2	...	$m-1$	m
0	$t(0,0;m)$	$t(0,1;m)$	$t(0,2;m)$	...	$t(0,m-1;m)$	$t(0,m;m)$
1	$t(1,0;m)$	$t(1,1;m)$	$t(1,2;m)$	...	$t(1,m-1;m)$	
2	$t(2,0;m)$	$t(2,1;m)$	$t(2,2;m)$			
...	...	...	...	...		
$m-2$	$t(m-2,0;m)$	$t(m-2,1;m)$	$t(m-2,2;m)$			
$m-1$	$t(m-1,0;m)$	$t(m-1,1;m)$				
m	$t(m,0;m)$					

In a mixture of K trinomials we have $3K-1$ parameters, $2K$ parameters from p_{1j} and $p_{2j},\ j=1,2,\ldots,K$; and $K-1$ parameters λ_j because $\sum_{i=1}^{K}\lambda_j = 1$. Therefore the mixture of trinomials will be identifiable if we have more equations $t(x\,,y\,;m\,,p)$ than parameters or equivalently if $(m+1)(m+2)/2\ \ -1\ \geq\ 3K-1$. In a mixture of two trinomials we need only $m \geq 2$. This method suggests that as the number of classes in the multinomial distribution increases we require fewer observations in the multinomial distribution. In general, we need $m \geq 2$ in order to estimate a mixture of two multinomials.

3.3 A Model-Free Approach with the trinomial

Our proposal now is to reduce the multivariate data to sums of trinomial responses. We introduce two cut points: c_1 and c_2 and define for $i=1,\ldots,m$ and $j=1,\ldots,n$.

$$Y_{1ij} = \begin{cases} 1 & \text{if} \quad X_{ij} \leq c_1 \\ 0 & \text{otherwise.} \end{cases}$$

$$Y_{2ij} = \begin{cases} 1 & \text{if} \quad c_1 < X_{ij} \leq c_2 \\ 0 & \text{otherwise.} \end{cases}$$

$$Y_{3ij} = \begin{cases} 1 & \text{if} \quad c_2 < X_{ij} \\ 0 & \text{otherwise.} \end{cases}$$

In this way $Y_{1j} = \sum_{i=1}^{m} Y_{1ij}$ is the number of times the observations in the jth observation vector are less than or equal to c_1, $Y_{2j} = \sum_{i=1}^{m} Y_{2ij}$ is the number of components in the jth observation vector that fall in the interval $(c_1 \ , \ c_2]$ and finally $Y_{3j} = \sum_{i=1}^{m} Y_{3ij} = \ m - Y_{1j} - Y_{2j}$ is the number of times the observations in the jth observation vector that are greater than c_2. As we saw in the former section, in order to have identifiability in the mixture of two trinomial distributions we must have $m \geq 2$.

THEOREM 3.1. *Assume the mixture model with conditional iid structure and* $m \geq 2$. *Further, assume that* $F_1(c_k) \neq F_2(c_k)$ *for* $k = 1, 2$. *Then, for fixed* $c_1 < c_2, Y_1, \ldots, Y_n$ *where* $Y_j^T = (Y_{1j}, Y_{2j}, Y_{3j})$ *are iid*

$$h(y) = \lambda \, t_1 \, (y_1, \, y_2 \, ; \, m \,) + (1 - \lambda) \, t_2 \, (y_1, \, y_2 \, ; \, m \,)$$

where $F_r\,(c_k) = \int_{-\infty}^{c_k} f_r(x)dx,\ r = 1, 2\ ;\ k\ =\ 1, 2$ *and*

$$t_1\,(y_1,\ y_2\ ;\ m\,) = t\,(y_1,\ y_2\ ;\ m,\ F_1(c_1),\ F_1(c_2) - F_1(c_1)\,)$$

$$t_2\,(y_1,\ y_2\ ;\ m\,) = t\,(y_1,\ y_2\ ;\ m, F_2(c_1),\ F_2(c_2) - F_2(c_1)\,)$$

$t_k\,(y_1,\ \ y_2\ ;\ \ m,\ \ p_{\,1k},\ \ p_{\,2k})$ *denotes the trinomial mass function with parameters* m, $p_{\,1k}$ *and* $p_{\,2k}$.

Proof: Define a random variable R such that $P(R = r) = \lambda$ for $r = 1$ and $P(R = r) = 1 - \lambda$ for $r = 2$ then:

$$p(x) = \sum_{r=1}^{2} \prod_{i=1}^{m} f_r(x_i) P(R = r)$$

By the definition of $Y_j^T\ =\ (Y_{1j}\ , Y_{2j},\ Y_{3j})$ given $R = r$, Y_j has a trinomial distribution with parameters $m,\ F_r(c_1),\ F_r(c_2)$. Then:

$$P(Y_j = y) = \sum_{r=1}^{2} P(Y_j = y \mid R = r)\ P(R = r)$$

and the theorem follows.

Observe that if c_1 and c_2 are given, we can use the mixed trinomial likelihood and the EM algorithm to estimate λ.

When c_1 and c_2 are specified, it is straight forward to compute the information from the trinomial mixture. Further, the asymptotic distribution of $\widehat{\lambda}$, the maximum likelihood estimator, is normal. This is summarized in the following theorem.

THEOREM 3.2. *Assume the mixture model with conditional iid structure. Given* c_1 *and* c_2 *suppose* $F_1(c_k) \neq F_2(c_k)$ *for* $k = 1, 2$; *and define*

$$g(y_1, y_2, \lambda, c_1, c_2) = \frac{t_1(y_1, y_2; m) - t_2(y_1, y_2; m)}{\lambda\, t_1(y_1, y_2; m) + (1-\lambda)\, t(y_1, y_2; m)}$$

let $0 < \lambda_0 < 1$ *denote the true value of the parameter* λ.

Then $\widehat{\lambda}$ *solves* $\sum_{j_1} \sum_{j_2} g(\, y_{1 j_1}\, ,\, y_{2\, j_2}\, ,\, \lambda,\, c_1\, ,\, c_2) = 0$ *and, as* $n \to \infty$,

$$\sqrt{n}(\widehat{\lambda} - \lambda_0) \xrightarrow{D} Z \sim N\left(0\, ,\, \frac{1}{E g_T^2(Y, \lambda_0, c_1, c_2)}\right)$$

where

$$E g_T^2(Y, \lambda_0, c_1, c_2) = \sum_{i=0}^{m} \sum_{k=0}^{m-i} \frac{[t_1(i, k; m) - t_2(i, k; m)]^2}{\lambda_0\, t_1(i, k; m) + (1-\lambda_0)\, t_2(i, k; m)}$$

Proof: Observe that $g(y_1, y_2, \lambda, c_1, c_2)$ is the estimating equation for the trinomial mixture given by the maximum likelihood method. The following expression denotes the partial derivative with respect to λ:

$$\frac{\delta}{\delta\lambda} g(y_1, y_2, \lambda, c_1, c_2) = g'_{\lambda}\, (y_1, y_2, \lambda, c_1, c_2) = -\, g^2(y_1, y_2, \lambda, c_1, c_2)$$

Expanding the estimating equation about λ_0 we get:

$$\sum_j \sum_k g(y_{1j}, y_{2k}, \lambda, c_1, c_2) = \sum_j \sum_k g(y_{1j}, y_{2k}, \lambda_0, c_1, c_2) + \\ + (\lambda - \lambda_0) \sum_j \sum_k g'_{\lambda}\, (y_{1j}, y_{2k}, \lambda, c_1, c_2) \;+\; o_p(1/\sqrt{n})$$

but:

$$\sum_j \sum_k g(y_{1j}, y_{2\,k}, \widehat{\lambda}, c_1, c_2) = 0$$

therefore:

$$\begin{aligned}\sqrt{n}(\widehat{\lambda} - \lambda_0) &= -\sqrt{n}\,\frac{\sum_j \sum_k g(y_{1j}, y_{2\,k}, \lambda_0, c_1, c_2)}{\sum_j \sum_k g'_\lambda\,(y_{1j}, y_{2k}, \lambda, c_1, c_2)} \quad + \quad o_p(1) \\ &= -\frac{n^{-1/2}}{n^{-1}}\frac{\sum_j \sum_k g(y_{1j}, y_{2\,k}, \lambda_0, c_1, c_2)}{\sum_j \sum_k g'_\lambda\,(y_{1j}, y_{2k}, \lambda, c_1, c_2)} + o_p(1)\end{aligned}$$

The result now follows from the Central Limit Theorem and Slutsky's Theorem.

The asymptotic distribution of the estimator only weakly depends on the underlying mixture model. Given values of c_1 and c_2, we do not need to specify a model for f_1 and f_2 to estimate λ. Of course, the asymptotic variance of the estimate depends on the underlying distribution, but only through $F_1(c_i)$ and $F_2(c_i)$.

With arguments similar to the ones given before, it is possible to prove the following general theorem. In this theorem, for $k = 1, 2$.

$$Mr_k(y_1, y_2, \ldots, y_r\,;\, m) = Mr_k(y_1, y_2, \ldots, y_r\,;\, m\,, p_{k1}, p_{k2}, \ldots, p_{kr})$$

is the multinomial mass function with $r + 1$ classes and parameters m, p_{k1}, $p_{k2}, \ldots,$ p_{kr}, such that $\sum_{j=1}^{r} p_{kj} < 1$.

THEOREM 3.3. *Assume the mixture model with conditional iid structure and* $m \geq 2$. *Given* $c_1 < c_2 < \ldots < c_r$ *suppose* $F_1(c_k) \neq F_2(c_k)$ *for* $k = 1, 2, \ldots, r$ *and define*

$$g(y_1, y_2, \ldots, y_r\ , \lambda, c_1, c_2, \ldots, c_r) =$$

$$= \quad \frac{Mr_1(y_1, y_2, \ldots, y_r\ ; m) - \ Mr_2(y_1, y_2, \ldots, y_r; m)}{\lambda\ Mr_1(y_1, y_2, \ldots, y_r\ ; m) + (1-\lambda)\ Mr_2(y_1, y_2, \ldots, y_r; m)}$$

let $0 < \lambda_0 < 1$ *denote the true value of the parameter* λ.

Then $\widehat{\lambda}$ *solves* $\sum_{j_1} \cdots \sum_{j_r}\ g(\ y_{1j_1}\ , y_{2j_2}\ , \ldots, y_{rj_r}\ \lambda,\ c_1,\ c_2, \ldots, c_r) = 0$ *and, as* $n \to \infty$,

$$\sqrt{n}(\widehat{\lambda} - \lambda_0)\ \xrightarrow{D}\ Z \sim N\left(0, \frac{1}{Eg^2_{M_r}(Y, \lambda_0, c)}\right)$$

where

$$Eg^2_{M_r}(Y, \lambda_0, c_1, c_2, \ldots c_r) =$$

$$= \sum_{i_1=0}^{m} \sum_{i_2=0}^{m-i_1} \cdots \sum_{i_r=0}^{m-i_{r-1}} \frac{[Mr_1(i_1, i_2, \ldots, i_r\ ; m\ ) - Mr_2(i_1, i_2, \ldots, i_r\ ; m\ )]^2}{\lambda_0\ Mr_1(i_1, i_2, \ldots, i_r;\ m\ ) + (1-\lambda_0) Mr_2(i_1, i_2, \ldots, i_r; m)}.$$

In this theorem for k = 1, 2.

$$Mr_k(y_1, y_2, \ldots, y_r\ ; m\ ) =$$

$$= \quad Mr_k[y_1, y_2, \ldots, y_r\ ; m\ , F_k(c_1), F_k(c_2) - F_k(c_1), \ldots, F_k(c_r) - F_k(c_{r-1})]$$

If we have have a binomial distribution with one cut point c_1 (the optimal c for the binomial), then we include an additional point $c_2 > c_1$ in order to define a trinomial distribution (a multinomial distribution with two independent random variables). The trinomial approach will be more efficient than the binomial if:

$$A.V_{Bin}\left[\sqrt{n}(\widehat{\lambda}-\lambda_0)\right] = \frac{1}{Eg_B^2(Y,\lambda_0,c_1)}$$

$$\geq A.V_{Tri}\left[\sqrt{n}(\widehat{\lambda}-\lambda_0)\right] = \frac{1}{Eg_T^2(Y,\lambda_0,c_1,c_2)} \tag{3.1}$$

3.4 Comparisons for small m

The trinomial approach will be more efficient than the binomial, as we mentioned before if Equation 3.1 holds or equivalently if:

$$Eg_B^2(Y,\lambda_0,c_1) = \sum_{i=0}^{m} \frac{\left[b\left(i\,;m\,,F_1(c_1)\right) - b\left(i\,;m\,,F_2(c_1)\right)\right]^2}{\lambda_0\, b(i\,;m\,,F_1(c_1)) + (1-\lambda_0)\, b(i\,;m\,,F_2(c_1))}$$

$$\leq Eg_T^2(Y,\lambda_0,c_1,c_2)$$

$$= \sum_{i=0}^{m}\sum_{k=0}^{m-i} \frac{\left[t_1\left(i,k\,;m\,\right) - t_2\left(i,k;m\right)\right]^2}{\lambda_0\; t_1(i,k\,;m\,) + (1-\lambda_0)\; t_2\left(i,k\,;m\,\right)} \tag{3.2}$$

We know that:

$$b(i\,;m\,,F_1(c_1)) = \sum_{k=0}^{m-i} t\ (i,k\,;m,F_1(c_1),F_1(c_2)-F_1(c_1)\ )$$

Then if we let:

$$A_k = t\ (i,k;m,F_1(c_1),F_1(c_2)-F_1(c_1)\ )$$

$$-\,t\ (i,k;m,F_2(c_1),F_2(c_2)-F_2(c_1)\ )$$

$$B_k = \lambda_0\ t\ (i,k;m,F_1(c_1),F_1(c_2)-F_1(c_1)\ )$$

$$+\,(1-\lambda_0)\ t\ (i,k;m,F_2(c_1),F_2(c_2)-F_2(c_1)\ )$$

Inequality 3.2 will be proved if we can show that:

$$\frac{(\sum_k A_{\,k})^2}{\sum_k B_{\,k}} \leq \sum_k \frac{A_k^2}{B_k}$$

But this inequality is just another expression of the Cauchy-Schwartz inequality.

$$\left(\sum_k X_k\ Y_k\right)^2 \leq \sum_k X_k^2 \sum_k Y_k^2$$

If we take (this identification is valid because $B_k \geq 0$.):

$$X_k = \frac{A_{\,k}}{\sqrt{B_k}},\ Y_k = \sqrt{B_k}$$

We have proved that for finite m, the trinomial approach is more efficient than the binomial approach.

In general we let:

$$M_r\,(y_1,\ldots,y_r\,;\,m\,) \;=\; Mr\,(y_1,\ldots,y_r\,;\,m,\,p_1\,,\ldots,p_r\,)$$

$$M_{r-1}\,(y_1,\ldots,y_{r-1}\,;\,m\,) \;=\; M_{r-1}\,(y_1,\ldots,y_{r-1}\,;\,m,\,,p_1\,,\ldots,p_{r-1}\,)$$

where $M_r\,(y_1,\ldots,y_r\,;\,m\,)$ and $M_{r-1}\,(y_1,\ldots,y_{r-1}\,;\,m\,)$ denotes the multinomial with r and $(r-1)$ independent variables mass functions respectively.

We know that:

$$M_{r-1}\,(y_1,\ldots,y_{r-1}\,;\,m\,) = \sum_{y_r=0}^{m-y_{r-1}} M_r\,(y_1,\ldots,y_r\,;\,m\,)$$

In other words this means that the marginal of an r-th multinomial distribution is an $(r-1)$ - th multinomial distribution. With the help of the Cauchy-Schwartz inequality we can prove the following general theorem.

THEOREM 3.4. *Assume the mixture model with conditional iid structure. Given* $c_1 < c_2 < \ldots < c_r$ *suppose* $F_1(c_k) \neq F_2(c_k)$ *for* $k = 1, 2, \ldots, r$. *Let* $0 < \lambda_0 < 1$ *denote the true value of the parameter* λ.

If $\widehat{\lambda}_r$ *and* $\widehat{\lambda}_{r-1}$ *represent the estimators of the parameter* λ *for the* r*-th and* $(r-1)$ *- th*

multinomial mass functions then:

$$A.V_{M_{r-1}}\left[\sqrt{n}(\widehat{\lambda}_{r-1}-\lambda_0)\right]=\frac{1}{Eg^2_{M_{r-1}}(Y,\lambda_0,c_1,c_2,\ldots,c_{r-1})}$$

$$\geq A.V_{M_r}\left[\sqrt{n}(\widehat{\lambda}_{r}-\lambda_0)\right]=\frac{1}{Eg^2_{M_{r}}(Y,\lambda_0,c_1,c_2,\ldots,c_{r})}$$

Theorem 3.4 proves that the trinomial approach gives more efficient estimators than the binomial approach when the cut point in the binomial is included as one of the two cut points in the trinomial distribution. One important question arises: How large is this improvement? For answering this question we define a function for the asymptotic variance in Mathematica, in terms of m and λ, when f_1 is a $N(0,1)$, f_2 is a $N(1,1)$, $\lambda=0.5$ and $m=4$. A trinomial distribution can be obtained including an additional cut point c_2 to the right of the optimal cut point $c_1=0.5$ given for the binomial distribution. The asymptotic variance, for the trinomial distribution described before, can be expressed in terms of c_2 and λ. Figure 3.1 shows that the asymptotic variance for $\sqrt{n}(\widehat{\lambda}-\lambda_0)$ decreases for $\lambda=0.5$, for all values of $c_2>c_1$. This figure also show that is possible to obtain the optimal cut point $c_2=1.4471$ for this trinomial distribution.

The asymptotic variance of $\sqrt{n}(\widehat{\lambda}-\lambda_0)$ associated with the optimal trinomial distribution for $\lambda=0.5$, $c_1=0.5$ and $c_2=1.4471$ is equal to 0.5311 which is smaller than 0.5937 the asymptotic variance of $\sqrt{n}(\widehat{\lambda}-\lambda_0)$ associated with the optimal binomial distribution for $\lambda=0.5$ and $c_1=0.5$ (Table 2.1).

It is interesting to mention that it is possible to obtain the same asymptotic variance of $\sqrt{n}(\widehat{\lambda}-\lambda_0)$ associated with this optimal trinomial distribution for $\lambda=0.5$, when the additional cut point c_2 is located to the left of $c_1=0.5$, in this case we get another optimal

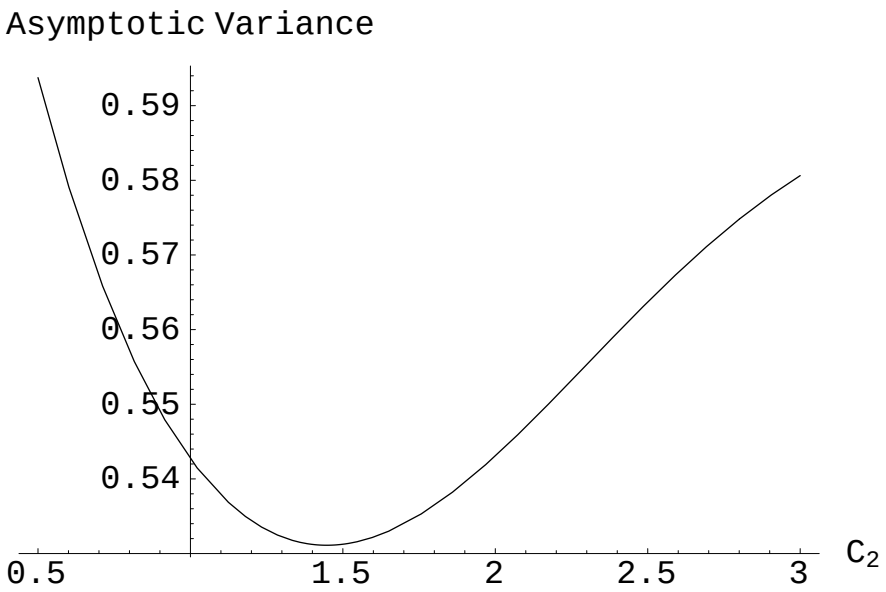

Fig. 3.1. *A.V* of $\sqrt{n}(\widehat{\lambda} - \lambda_0)$, for $\lambda = 0.5$ and $c_2 > c_1$ for normal distributions.

trinomial distribution with $c_2 = -0.4471$ and the asymptotic variance of $\sqrt{n}(\widehat{\lambda} - \lambda_0)$ equal to 0.5311. Note that this value of c_2 (to the left of c_1 for this optimal trinomial distribution is symmetric to the value of c_2 (to the right) with respect to $c_1 = 0.5$ for the first optimal trinomial distribution. Figure 3.2 shows this fact.

The trinomial distributions mentioned before were generated by adding one additional point to the optimal binomial distribution. Will it be possible to decrease this asymptotic variance with another cut point $c_1 \neq 0.5$?. This question is answered numerically also. Figure 3.3 shows a tridimensional plot of the asymptotic variance associated with the trinomial distribution in terms of c_1 and c_2 for $\lambda = 0.5$. In order to generate this graphic we define $c_2 = c_1 + t,\ t > 0$.

Figure 3.3 shows that the asymptotic variance is not very sensitive to changes in c_1 and c_2. This result is very informative and useful because in a real situation we need to estimate these cut points and this result tells us that this estimation does not need to

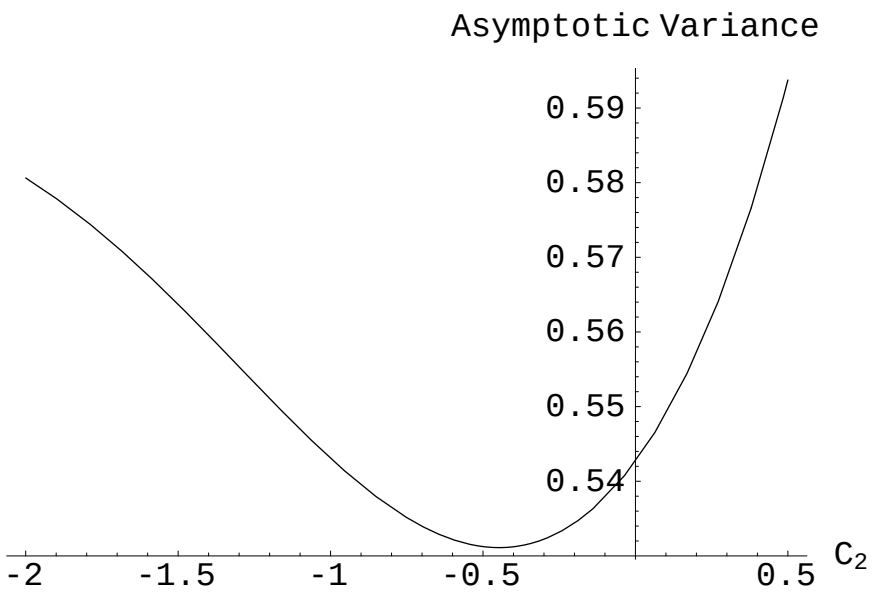

Fig. 3.2. *A.V* of $\sqrt{n}(\widehat{\lambda} - \lambda_0)$, for λ=0.5 and $c_2 < c_1$ for normal distributions.

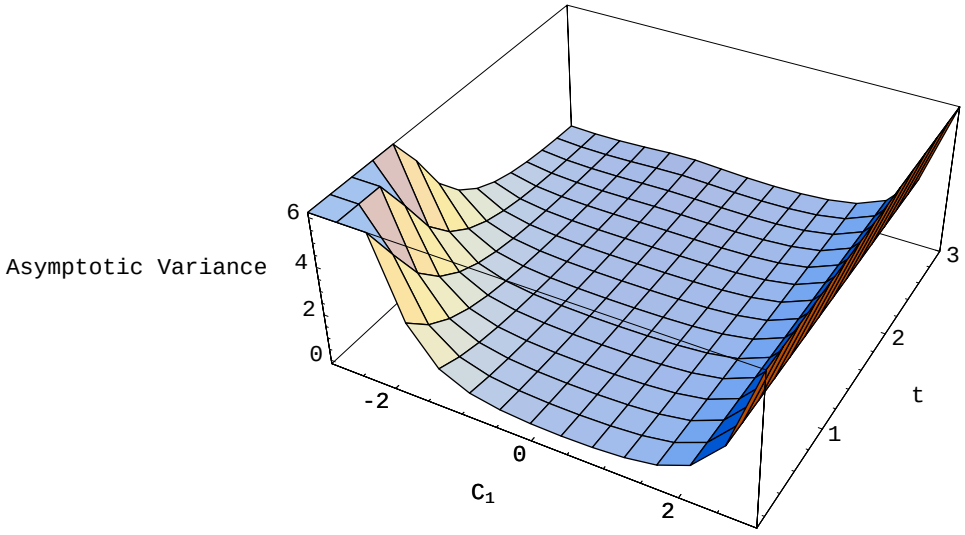

Fig. 3.3. *A.V* of $\sqrt{n}(\widehat{\lambda} - \lambda_0)$, for λ=0.5, c_1 and $c_2 = c_1 + t$ for normal distributions.

be very precise. Figure 3.4 shows more resolution in the vicinity of the optimum values of $c_1 = -0.0908$ and $c_2 = 1.0908$ ($t = 1.1817$) where the asymptotic variance is equal to 0.5106. Note that this value is not far from the asymptotic variance equal to 0.5311 attained with the inclusion of a second optional point along with the original optimal binomial point.

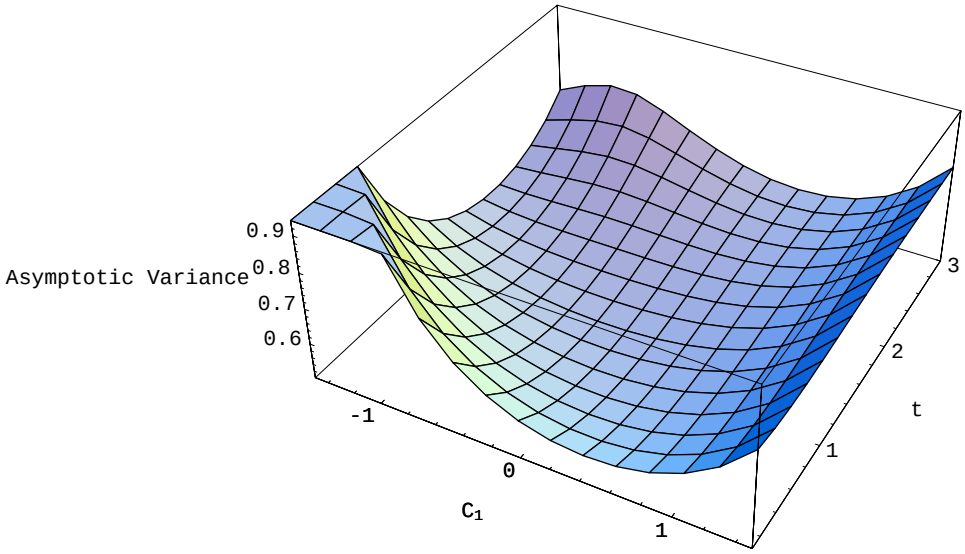

Fig. 3.4. *A.V* of $\sqrt{n}(\widehat{\lambda} - \lambda_0)$, for λ=0.5, in the vicinity of the optimal values of c_1 and $c_2 = c_1 + t$ for normal distributions.

3.5 Comparisons for large m

In this section the binomial approach will be compared with the trinomial and in general with the multinomial approaches when $m \to \infty$. First we will present the asymptotic information matrix for the binomial mixture.

3.5.1 Asymptotic Information Matrix for the Binomial Mixture

In this subsection with the argument that for large m, the supports of the the binomial components are disjoint, we will obtain the asymptotic distribution of the MLE estimators of a mixture of two binomial distributions when we are interested in the parameters: $\lambda,\ p_1$ and p_2. Blischke [2] found that this limiting distribution is also attained for the moment estimators of the parameters.

From Chapter 2, we know that the density function for the binomial mixture is:

$$b_\lambda(y) = \lambda b(y\,;\ m, p_1) + (1-\lambda) b(y\,;\ m, p_2)$$

If we define $b_1(y) = b(y\,; m,\ p_1)\ ,\ b_2(y) = b(y\,; m,\ p_2)$ and $\theta' = (\lambda,\ p_1,\ p_2)$. The loglikehood function is:

$$\log L(\lambda, \theta) = \sum_{i=1}^{n} \log[b_\lambda(y_i)]$$

Therefore:

$$\frac{\partial}{\partial \lambda} \log[\, b_\lambda(y)\,] = \frac{b_1(y) - b_2(y)}{b_\lambda(y)}$$

$$\left[\frac{\partial}{\partial \lambda} \log[\, b_\lambda(y)\,]\right]^2 = \frac{[b_1(y) - b_2(y)]^2}{b_\lambda^2(y)}$$

$$\frac{\partial^2}{\partial\lambda\,\partial p_1}\log[\,b_\lambda(y)\,] =$$

$$= \frac{\binom{m}{y}\left[y\,p_1^{y-1}(1-p_1)^{m-y} - (m-y)\,p_1^{y}(1-p_1)^{m-y-1}\right]}{b_\lambda(y)}$$

$$- \frac{\lambda\binom{m}{y}\left[y\,p_1^{y-1}(1-p_1)^{m-y} - (m-y)\,p_1^{y}(1-p_1)^{m-y-1}\right][b_1(y)-b_2(y)]}{b_\lambda^2(y)}$$

$$= \frac{\binom{m}{y}\,p_1^{y}(1-p_1)^{m-y}(y-m\,p_1)}{p_1(1-p_1)b_\lambda(y)} - \lambda\frac{\binom{m}{y}\,p_1^{y}(1-p_1)^{m-y}(y-m\,p_1)}{\theta_1(1-p_1)b_\lambda^2(y)}[b_1(y)-b_2(y)]$$

$$= \frac{b_1(y)\,(y-m\,p_1)}{p_1(1-p_1)b_\lambda(y)} - \lambda\frac{b_1(y)\,(y-m\,p_1)}{p_1(1-p_1)b_\lambda^2(y)}[b_1(y)-b_2(y)]$$

and

$$\frac{\partial}{\partial\,p_1}\log[\,b_\lambda(y)] \;= \frac{\lambda\,b_1(y)(y-m\,p_1)}{p_1(1-p_1)\,b_\lambda(y)}$$

$$\left[\frac{\partial}{\partial\,p_1}\log[\,b_\lambda(y)]\right]^2 = \left[\frac{\lambda\,b_1(y)(y-m\,p_1)}{p_1(1-p_1)\,b_\lambda(y)}\right]^2$$

$$\frac{\partial^2}{\partial\,p_1\partial\,p_2}\log[\,b_\lambda(y)] = \frac{\lambda\,b_1(y)(y-m\,p_1)}{p_1(1-p_1)}\,\frac{(1-\lambda)\,b_2(y)(y-m\,p_2)}{p_2(1-p_2)\,b_\lambda^2(y)}$$

For the well known properties of MLE estimators if $\widehat{\phi}' = (\,\widehat{\lambda}\,,\,\widehat{p}_1,\,\widehat{p}_2)$:

$$\sqrt{n}(\widehat{\phi}-\phi_0) \;\xrightarrow{D}\; Z\sim N\left(0\,,\,I^{\,-1}\right)$$

Where:

$$I = \begin{bmatrix} E\left(\left[\frac{\partial}{\partial\lambda}\log[\,b_\lambda(y)]\right]^2\right) & -E\left(\frac{\partial^2}{\partial\lambda\,\partial p_1}\log[b_\lambda(y)]\right) & -E\left(\frac{\partial^2}{\partial\lambda\,\partial p_2}\log[b_\lambda(y)]\right) \\ . & E\left(\left[\frac{\partial}{\partial p_1}\log[b_\lambda(y)]\right]^2\right) & -E\left(\frac{\partial^2}{\partial p_1\,\partial p_2}\log[b_\lambda(y)]\right) \\ . & . & E\left(\left[\frac{\partial}{\partial p_2}\log[b_\lambda(y)]\right]^2\right) \end{bmatrix}$$

is the Fisher information matrix, whose elements were obtained above. Now we will obtain the $\lim_{m\to\infty} I$.

$$E\left(\left[\frac{\partial}{\partial\lambda}\log(b_\lambda(y))\right]^2\right) = \sum_{i=0}^{m}\frac{[b_1(i)-b_2(i)]^2}{b_\lambda(i)} = \sum_{i=0}^{m}\frac{b_1^2(i)+b_2^2(i)-2b_1(i)\;b_2(i)}{b_\lambda(i)}$$

Observe that for m sufficiently large the difference of the means of the binomial distributions $b_1(y)$ and $b_2(y)$ is of order of magnitude m, but their standard deviations are of order $m^{1/2}$, then their probability masses are concentrated over different values of integers $i's$. Therefore for m sufficiently large:

$$\sum_{i=0}^{m}\frac{b_1^2(i)+b_2^2(i)-2b_1(i)\;b_2(i)}{b_\lambda(i)}$$
$$\approx \sum_{i=0}^{m}\left\{\frac{b_1^2(i)}{\lambda_0\;b_1(i)}\right\} + \sum_{i=0}^{m}\left\{\frac{b_2^2(i)}{(1-\lambda_0)\;b_2(i)}\right\} = \frac{1}{\lambda_0}+\frac{1}{1-\lambda_0}$$

Therefore:

$$\lim_{n\to\infty} E\left(\left[\frac{\partial}{\partial\lambda}\log(b_\lambda(y))\right]^2\right) = \frac{1}{\lambda_0(1-\lambda_0)}$$

In a similar way:

$$E\left(\frac{\partial^2}{\partial\,\lambda\partial\,p_1}\log[b_\lambda(y)]\right)=E\left(\frac{b_1(y)\;(y-m\,p_1)}{p_1(1-p_1)b_\lambda(y)}\right)-\lambda\,E\left(\frac{b_1(y)\;(y-m\,p_1)}{p_1(1-p_1)b_\lambda^2(y)}[b_1(y)-b_2(y)]\right)$$

but

$$E\left(\frac{b_1(y)\;(y-m\,p_1)}{p_1(1-p_1)b_\lambda(y)}\right)=\frac{1}{p_1(1-p_1)}\sum_{i=0}^{m}\;b_1(i)(i-m\,p_1)=0$$

and

$$\lim_{n\to\infty}\lambda E\left(\frac{b_1(y)\;(y-m\,p_1)}{p_1(1-p_1)b_\lambda^2(y)}[b_1(y)-b_2(y)]\right)=$$

$$=\lim_{n\to\infty}\lambda\sum_{i=0}^{m}\frac{b_1^2(i)\;(i-m\,p_1)-b_1(i)\;b_2(i)\;(i-m\,p_1)}{p_1(1-p_1)b_\lambda(i)}\approx\lambda\sum_{i=0}^{m}\left[\frac{b_1(i)\;(i-m\,p_1)}{p_1(1-p_1)\lambda_0}\right]=0$$

Therefore

$$\lim_{n\to\infty}E\left(\frac{\partial^2}{\partial\,\lambda\partial\,p_1}\log[b_\lambda(y)]\right)=0$$

Also

$$\lim_{n\to\infty}\left(E\left[\frac{\partial}{\partial\,p_1}\log(b_\lambda(y))\right]^2\right)=\lim_{n\to\infty}\sum_{i=0}^{m}\frac{\lambda^2\;b_1^2(i)(i-m\,p_1)^2}{p_1^2(1-p_1)^2\;b_\lambda(i)}$$

$$\approx\sum_{i=0}^{m}\frac{\lambda_0^2\;b_1(i)(i-m\,p_1)^2}{p_1^2(1-p_1)^2\;\lambda_0}=\frac{\lambda_0}{p_1(1-p_1)}$$

and finally

$$\lim_{n\to\infty}\left(E\frac{\partial^2}{\partial\, p_1\partial\, p_2}\log[b_\lambda(y)]\right)=$$

$$=\lim_{n\to\infty}\sum_{i=0}^{m}\frac{\lambda\, b_1(y)(y-m\,p_1)}{p_1(1-p_1)}\frac{(1-\lambda)\; b_2(y)(y-m\,p_2)}{p_2(1-p_2)\; b_\lambda^2(i)}\approx 0$$

We have shown that when $m\to\infty$:

$$\lim_{m\to\infty} I=\begin{bmatrix}\frac{1}{\lambda_0(1-\lambda_0)} & 0 & 0\\ 0 & \frac{\lambda_0}{p_1(1-p_1)} & 0\\ 0 & 0 & \frac{\lambda_0}{p_2(1-p_2)}\end{bmatrix}$$

Or equivalently:

$$\sqrt{n}(\widehat{\phi}-\phi_0)\ \xrightarrow{D}\ Z\sim N\left(0\ ,\ \begin{bmatrix}\lambda_0(1-\lambda_0) & 0 & 0\\ 0 & \frac{p_1(1-p_1)}{\lambda_0} & 0\\ 0 & 0 & \frac{p_2(1-p_2)}{\lambda_0}\end{bmatrix}\right)$$

We have shown that when $m\to\infty$ the estimator of the mixing parameter is independent of the parameters of the binomial components of the mixture distribution.

3.5.2 Asymptotic Variance of the Mixing Parameter Estimator

With the help of this result we can prove the following theorem.

THEOREM 3.5. *The asymptotic variance of* $\sqrt{n}\left(\widehat{\lambda}-\lambda_0\right)$ *, when* $m\to\infty$ *for the Binomial, Trinomial and, in general, for the Multinomial Approaches is equal to* $\lambda_0(1-\lambda_0)$.

Proof:

For c, the optimal cut point for the binomial distribution let $p_1 = F_1(c) \neq p_2 = F_2(c)$, then if we denote the asymptotic variance of $\sqrt{n}\left(\widehat{\lambda} - \lambda_0\right)$ by $AV\{\sqrt{n}\left(\widehat{\lambda} - \lambda_0\right)\}$, we have:

$$\begin{aligned} AV\{\sqrt{n}(\widehat{\lambda} - \lambda_0)\} &= \lim_{m\to\infty} \frac{1}{\sum_{i=0}^{m} \left\{ \frac{[\, b(i;\, m,\, p_1) - b(i;\, m,\, p_2)\,]^2}{\lambda_0\, b(i;\, m,\, p_1) + (1-\lambda_0)\, b(i;\, m,\, p_2)} \right\}} \\ &= \frac{1}{\lim_{m\to\infty} \sum_{i=0}^{m} \left\{ \frac{[\, b(i;\, m,\, p_1) - b(i;\, m,\, p_2)\,]^2}{\lambda_0\, b(i;\, m,\, p_1) + (1-\lambda_0)\, b(i;\, m,\, p_2)} \right\}} \\ &= \lambda_0(1 - \lambda_0) \end{aligned}$$

In the last equality we use the same argument as before. The proof for the trinomial and, in general, for the multinomial approach is similar, we only need to substitute the binomial density for the multinomial density.

We observed that the asymptotic variance of the multinomial estimator of the mixing parameter is equal to the minimum value of the Cramér-Rao minimum variance bound for an unbiased estimator given by Hill [17] and equal to

$$\frac{\lambda_0\,(1-\lambda_0)}{1 - I(\lambda)}.$$

In our case

$$I(\lambda) = \sum_{i=0}^{m} \frac{b_1(i)\, b_2(i)}{b_\lambda(i)}$$

and for m sufficiently large is equal to zero. Therefore the asymptotic variance of the multinomial estimator of the mixing parameter is equal to the minimum value of the Cramér-Rao minimum variance bound.

This result is very logical; it tells us that the maximum information that any of these approaches can give about the mixture distribution is the correct classification of the observation in each component. This is the number of observations of a binomial distribution with parameter λ_0, and that is why the limiting variance is equal to $\lambda_0\,(1-\ \lambda_0)/n$.

Bernardo [1] shows that this is also the lowest bound for the estimator of the mixing parameter in a Bayesian setting when we use the Jeffrey's prior for any component densities $f_1(x)$ and $f_2(x)$ with disjoint supports almost everywhere.

Table 3.2 shows the asymptotic variance of $\sqrt{n}(\widehat{\lambda}-\lambda_0)$ when $n\to\infty$ for some values of m for the binomial distribution when f_1 is a $N(0,1)$, and f_2 is a $N(1,1)$. In this table we can observe that the asymptotic variance given by Theorem 3.5 and equal to $\lambda_0(1-\lambda_0)$ is almost attained for m=40, we will see that the value m at which this limit is almost attained depends also on the separation of the medians of the component distributions. If the medians of the normal distributions are separated by two standard deviations, the asymptotic variance is almost attained at m=10.

We observed also that for the binomial approach, for small values of m, there is only one optimal value for the cut point c, but for large values of m, there exists an infinite number of values of c, for which the $AV\ \{\sqrt{n}(\widehat{\lambda}-\lambda_0)\}$ (the minimum Asymptotic Variance of $\sqrt{n}(\widehat{\lambda}-\lambda_0)$) is almost attained.

Table 3.3 shows the asymptotic variance of $\sqrt{n}(\widehat{\lambda}-\lambda_0)$ for the binomial, trinomial, tetranomial and pentanomial approaches for $\lambda=0.5$ and different values of m, when f_1

Table 3.2. Asymptotic Variance $(A.V)$ of $\sqrt{n}(\widehat{\lambda} - \lambda_0)$ for the Binomial Approach

λ_0/m	15	20	30	40	50	60	100	∞
.1	0.1213	0.1077	0.0964	0.0926	0.0910	0.0904	0.09001	0.09
.2	0.2029	0.1843	0.1689	0.1635	0.1615	0.1606	0.16002	0.16
.3	0.2597	0.2382	0.2203	0.2141	0.2117	0.2107	0.21002	0.21
.4	0.2934	0.2704	0.2511	0.2444	0.2418	0.2408	0.24002	0.24
.5	0.3046	0.2810	0.2614	0.2545	0.2519	0.2508	0.25002	0.25

is a $N(0,1)$, and f_2 is a $N(1,1)$. This table shows that the largest decrement in variance occurs when we change the binomial for the trinomial approach.

This result is very interesting because it tells us to use the trinomial approach instead of the binomial approach whenever possible.

Table 3.3. Asymptotic variance $(A.V)$ of $\sqrt{n}(\widehat{\lambda} - \lambda_0)$ for the Binomial, Trinomial, Tetranomial and Pentanomial Approaches

m	Binom	Trinom[1]	Trinom[2]	Tetranom	Pentanom
3	0.7227	0.6415	0.6144	0.5819	0.5675
4	0.5937	0.5311	0.5106	0.4857	0.4745
5	0.5149	0.4643	0.4478	0.4276	0.4185
6	0.4621	0.4195	0.4058	0.3889	0.3813
10	0.3560	0.3308	0.3228	0.3130	0.3087
15	0.3046	0.2891	0.2843	0.2786	0.2761
20	0.2810	0.2708	0.2677	0.2642	0.2626
30	0.2614	0.2566	0.2553	0.2539	0.2533
40	0.2545	0.2523	0.2517	0.2511	0.2509
50	0.2519	0.2508	0.2506	0.2503	0.2503
60	0.2508	0.2503	0.2502	0.2501	0.2501
100	0.2500	0.2500	0.2500	0.2500	0.2500

[1]Trinomial obtained by inserting an additional optimal point with the optimal binomial
[2]Optimal trinomial

3.6 Numerical Results

We have analyzed in some detail the case when f_1 is a $N(0,1)$ and f_2 is a $N(1,1)$. Now we will analyze the general case when f_1 is a $N(\mu_1, \sigma_1)$ and f_2 is a $N(\mu_2, \sigma_2)$. Observe that if we use the transformation $Z = (X - \mu_1) / \sigma_1$, the observations of f_1 will follow a distribution $N(0,1)$ and the observations of f_2 follow a distribution $N\ [\ (\mu_2 - \mu_1) / \sigma_1\ ,\ \sigma_2/\sigma_1\ ]$. Therefore if we know the parameters of these distributions it would be possible to obtain the optimal cut points and to construct tables similar to Table 2.1, for different values of λ_0, $(\mu_2 - \mu_1) / \sigma_1$ and different ratios σ_2 / σ_1. These tables, of course, are not very useful for selecting the optimal cut points because, in a real situation, we do not know the values of these parameters. On the other hand these tables can be very useful in order to analyze the behavior of the asymptotic variance for different values of λ_0, $\mu_2 - \mu_1$ in terms of σ_1, and different ratios σ_2 / σ_1.

When f_1 and f_2 have equal means (μ) but different variances, the optimal two cut points for the binomial distribution are symmetric with respect to the mean value μ. This result is in accordance with the theorem given by Hettmansperger and Thomas [16]. If we consider that $\mu = 0$ we obtain two optimal cut points $\pm c$ ($c_1 = -c,\ c_2 = c$), with two regions $R_1 = \{x : |x| < c\}$ and $R_2 = \{x : |x| \geq c\}$ (the complement of the region R_1). Figure 3.5 helps to illustrate this case.

Table 3.4 to Table 3.7 show the optimal cut points and the $A.V$ of the mixing parameter for the equal means case. Observe that the $A.V$ for the same values of m and λ_0 decreases when the ratio σ_2/σ_1 increases and the $A.V$ of the mixing parameter almost attains its limit value for m=10 when $\sigma_2/\sigma_1 \geq 3$. Considering a trinomial distribution in these cases did not decrease the asymptotic variance of the mixing estimator significantly

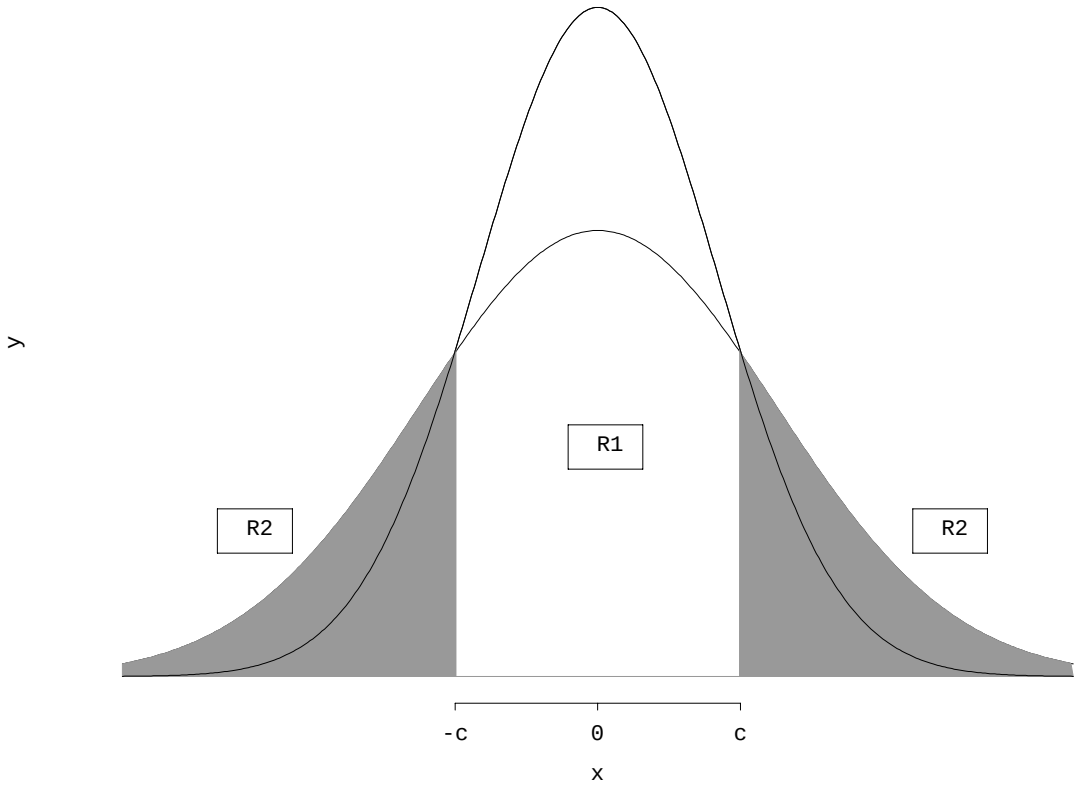

Fig. 3.5. Regions for the Binomial Distribution, Equal Means case

but, considering a binomial distribution with only one cut point, can increase indefinitely the variance of the estimator of the mixing parameter (if $\mu_1 = \mu_2 = 0$ and we choose the cut point $c = 0$, then $F_1(c) = F_2(c) = 0.5$ and $Eg^2(Y, \lambda_0, c) = 0$ in equation 2.3). See Figure 3.6.

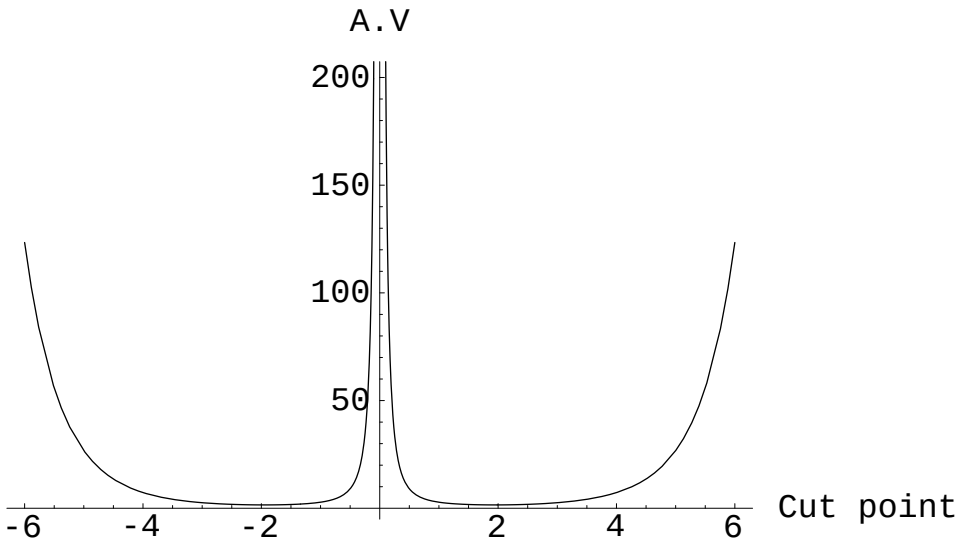

Fig. 3.6. *A.V* of $\sqrt{n}(\widehat{\lambda} - \lambda_0)$, in terms of c when $\mu_1 = \mu_2 = 0$ for normal distributions

Table 3.4. *A.V* of $\sqrt{n}(\widehat{\lambda} - \lambda_0)$ and optimal cut points for $\mu_2 = \mu_1$ and $\sigma_2/\sigma_1 = 1.5$ for different values of λ_0 and m.

λ_0/m	3 ($\pm c$, *A.V*)	4	6	8	10
.1	±1.454, 1.899	±1.466, 1.373	±1.494, 0.860	±1.523, 0.615	±1.550, 0.476
.3	±1.585, 1.936	±1.605, 1.470	±1.639, 1.010	±1.664, 0.785	±1.682, 0.652
.5	±1.722, 1.818	±1.737, 1.413	±1.756, 1.008	±1.765, 0.806	±1.772, 0.683
.7	±1.889, 1.529	±1.888, 1.195	±1.878, 0.858	±1.870, 0.687	±1.865, 0.583
.9	±2.175, 0.994	±2.130, 0.757	±2.075, 0.518	±2.041, 0.399	±2.011, 0.329

Table 3.5. $A.V$ of $\sqrt{n}(\widehat{\lambda} - \lambda_0)$ and optimal cut points for $\mu_2 = \mu_1$ and $\sigma_2/\sigma_1 = 2$ for different values of λ_0 and m.

λ_0/m	3 ($\pm c$, $A.V$)	4	6	8	10
.1	$\pm$1.510, 0.659	$\pm$1.560, 0.464	$\pm$1.667, 0.290	$\pm$1.757, 0.216	$\pm$1.797, 0.177
.3	$\pm$1.724, 0.775	$\pm$1.791, 0.604	$\pm$1.894, 0.443	$\pm$1.916, 0.368	$\pm$1.912, 0.323
.5	$\pm$1.919, 0.760	$\pm$1.980, 0.616	$\pm$2.041, 0.476	$\pm$1.985, 0.406	$\pm$2.004, 0.363
.7	$\pm$2.152, 0.617	$\pm$2.196, 0.508	$\pm$2.145, 0.398	$\pm$2.074, 0.339	$\pm$2.110, 0.304
.9	$\pm$2.592, 0.320	$\pm$2.604, 0.260	$\pm$2.244, 0.197	$\pm$2.258, 0.163	$\pm$2.293, 0.143

Table 3.6. $A.V$ of $\sqrt{n}(\widehat{\lambda} - \lambda_0)$ and optimal cut points for $\mu_2 = \mu_1$ and $\sigma_2/\sigma_1 = 3$ for different values of λ_0 and m.

λ_0/m	3 ($\pm c$, $A.V$)	4	6	8	10
.1	$\pm$1.688, 0.265	$\pm$1.833, 0.193	$\pm$2.108, 0.137	$\pm$2.334, 0.116	$\pm$2.074, 0.106
.3	$\pm$2.001, 0.395	$\pm$2.158, 0.328	$\pm$2.422, 0.270	$\pm$2.126, 0.246	$\pm$2.230, 0.231
.5	$\pm$2.238, 0.417	$\pm$2.387, 0.360	$\pm$2.633, 0.308	$\pm$2.212, 0.285	$\pm$2.336, 0.271
.7	$\pm$2.493, 0.338	$\pm$2.629, 0.296	$\pm$2.851, 0.257	$\pm$2.319, 0.238	$\pm$2.447, 0.227
.9	$\pm$2.913, 0.152	$\pm$3.021, 0.133	$\pm$2.383, 0.116	$\pm$2.503, 0.105	$\pm$2.626, 0.099

When the distributions have different means (we will consider the case $\mu_1 = 0 < \mu_2$, assuming that the transformation Z defined at the beginning of this section was applied) and different variances, the optimal cut points for the binomial distribution are no longer symmetric with respect to zero ($c_1 < 0 < c_2$). We found that for the same values of m, λ_0 and the ratio σ_2/σ_1, the $A.V$ decreases when the difference of $\mu_1 - \mu_2$ increases from 0 to σ_1. We also found that for the same values of m, λ_0 and equal differences in means $\mu_1 - \mu_2$ the $A.V$ decreases when the ratio σ_2/σ_1 increases. Table 3.8 to Table 3.11 show

Table 3.7. $A.V$ of $\sqrt{n}(\widehat{\lambda} - \lambda_0)$ and optimal cut points for $\mu_2 = \mu_1$ and $\sigma_2/\sigma_1 = 5$ for different values of λ_0 and m.

λ_0/m	3 ($\pm c$, $A.V$)	4	6	8	10
.1	$\pm$2.037, 0.142	$\pm$2.307, 0.116	$\pm$2.753, 0.099	$\pm$2.297, 0.094	$\pm$2.523, 0.092
.3	$\pm$2.406, 0.267	$\pm$2.659, 0.240	$\pm$3.059, 0.221	$\pm$2.463, 0.215	$\pm$2.685, 0.212
.5	$\pm$2.648, 0.300	$\pm$2.881, 0.277	$\pm$2.350, 0.264	$\pm$2.571, 0.255	$\pm$2.786, 0.252
.7	$\pm$2.889, 0.247	$\pm$3.101, 0.231	$\pm$2.449, 0.221	$\pm$2.680, 0.214	$\pm$2.888, 0.212
.9	$\pm$3.263, 0.107	$\pm$3.441, 0.099	$\pm$2.630, 0.095	$\pm$2.853, 0.092	$\pm$3.049, 0.091

these results. Observe that when $\mu_1 - \mu_2 = \sigma_1$, the $A.V$ of the mixing parameter almost attains its limit value for m=8 when $\sigma_2/\sigma_1 \geq 3$.

It is interesting to note that when $\mu_2 - \mu_1 = \sigma_1$ and $\sigma_2/\sigma_1 = 1.5$, the $A.V$ given by the binomial distribution with only one cut point is very similar with $A.V$ given by the binomial with two cut points (Figure 3.5 helps also to illustrate this case, the only difference is that here the means are different). For example when m=3, the optimal cut point for the binomial with one cut point is $c = 1.020$ and the $A.V$ is equal to 0.587. Compare this value with 0.576, the value given by the binomial with two cut points in Table 3.8. Observe also that one cut point $c_2 = 1.006$ of this binomial, is very close to the optimal cut point $c = 1.020$ of the binomial with one cut point. When $\mu_2 - \mu_1 = 2\,\sigma_1$ and $\sigma_2/\sigma_1 = 2$, the $A.V$ given by the binomial distribution with only one cut point differs from the $A.V$ given by the binomial distribution with two cut points in the third decimal place. Considering a trinomial distribution in these cases did not decrease the asymptotic variance of the mixing estimator significantly.

Table 3.12 to Table 3.15 show that when $\mu_2 - \mu_1 = 2\,\sigma_1$ and $1 \leq \sigma_2/\sigma_1 \leq 2$, the $A.V$ given by the binomial distribution with only one cut point is equal to the $A.V$ given by the binomial with two cut points. It is also possible to show that the $A.V$ for

Table 3.8. $A.V$ of $\sqrt{n}(\widehat{\lambda}-\lambda_0)$ and optimal cut points for $\mu_2-\mu_1=\sigma_1$ and $\sigma_2/\sigma_1=1.5$ for different values of λ_0 and m.

λ_0/m	3 (c_1/c_2, $A.V$)	4	6	8	10
.1	−2.606, 0.576	−2.672, 0.410	−2.794, 0.263	−2.855, 0.200	−2.891, 0.166
	+1.006	+1.072	+1.194	+1.255	+1.291
.3	−2.851, 0.718	−2.920, 0.567	−2.979, 0.423	−2.990, 0.353	−3.033, 0.313
	+1.251	+1.320	+1.379	+1.390	+1.433
.5	−3.052, 0.726	−3.094, 0.594	−3.070, 0.462	−3.092, 0.395	−3.116, 0.355
	+1.452	+1.498	+1.470	+1.492	+1.516
.7	−3.270, 0.605	−3.249, 0.495	−3.172, 0.389	−3.202, 0.333	−3.171, 0.300
	+1.670	+1.649	+1.572	+1.602	+1.571
.9	−3.599, 0.332	−3.399, 0.265	−3.374, 0.195	−3.353, 0.162	−3.271, 0.142
	+2.000	+1.799	+1.774	+1.754	+1.671

Table 3.9. $A.V$ of $\sqrt{n}(\widehat{\lambda}-\lambda_0)$ and optimal cut points for $\mu_2-\mu_1=\sigma_1$ and $\sigma_2/\sigma_1=2$ for different values of λ_0 and m.

λ_0/m	3 (c_1/c_2, $A.V$)	4	6	8	10
.1	−1.971, 0.434	−2.057, 0.307	−2.231, 0.200	−2.355, 0.157	−2.320, 0.134
	+1.304	+1.390	+1.564	+1.688	+1.654
.3	−2.237, 0.563	−2.341, 0.450	−2.502, 0.346	−2.434, 0.298	−2.450, 0.269
	+1.571	+1.674	+1.835	+1.767	+1.784
.5	−2.461, 0.571	−2.558, 0.476	−2.687, 0.383	−2.489, 0.337	−2.556, 0.309
	+1.794	+1.892	+2.020	+1.823	+1.889
.7	−2.716, 0.463	−2.801, 0.391	−2.885, 0.321	−2.591, 0.282	−2.670, 0.259
	+2.049	+2.135	+2.218	+1.925	+2.003
.9	−3.170, 0.225	−3.233, 0.189	−2.722, 0.151	−2.786, 0.129	−2.864, 0.117
	+2.504	+2.567	+2.055	+2.119	+2.198

Table 3.10. *A.V* of $\sqrt{n}(\widehat{\lambda}-\lambda_0)$ and optimal cut points for $\mu_2-\mu_1=\sigma_1$ and $\sigma_2/\sigma_1=3$ for different values of λ_0 and m.

λ_0/m	3 (c_1/c_2, *A.V*)	4	6	8	10
.1	−1.8620, 0.239	−2.024, 0.176	−2.333, 0.128	−2.574, 0.111	−2.264, 0.102
	+1.612	+1.774	+2.076	+2.324	+2.014
.3	−2.188, 0.370	−2.359, 0.310	−2.645, 0.259	−2.297, 0.239	−2.425, 0.226
	+1.938	+2.109	+2.395	+2.047	+2.175
.5	−2.429, 0.394	−2.591, 0.343	−2.856, 0.297	−2.392, 0.278	−2.532, 0.266
	+2.179	+2.341	+2.606	+2.142	+2.282
.7	−2.685, 0.319	−2.832, 0.282	−3.048, 0.248	−2.502, 0.233	−2.642, 0.223
	+2.435	+2.582	+2.834	+2.251	+2.392
.9	−3.101, 0.143	−3.219, 0.126	−3.414, 0.109	−2.687, 0.101	−2.820, 0.097
	+2.851	+2.969	+3.164	+2.437	+2.570

Table 3.11. *A.V* of $\sqrt{n}(\widehat{\lambda}-\lambda_0)$ and optimal cut points for $\mu_2-\mu_1=\sigma_1$ and $\sigma_2/\sigma_1=5$ for different values of λ_0 and m.

λ_0/m	3 (c_1/c_2, *A.V*)	4	6	8	10
.1	−2.096, 0.139	−2.372, 0.114	−2.823, 0.098	−2.355, 0.094	−2.586, 0.091
	+2.013	+2.288	+2.740	+2.272	+2.503
.3	−2.467, 0.264	−2.723, 0.238	−3.128, 0.220	−2.522, 0.215	−2.748, 0.212
	+2.384	+2.640	+3.045	+2.439	+2.665
.5	−2.625, 0.298	−2.945, 0.276	−2.403, 0.263	−2.630, 0.255	−2.850, 0.252
	+2.625	+2.861	+2.320	+2.547	+2.767
.7	−2.949, 0.245	−3.164, 0.229	−2.504, 0.220	−2.739, 0.213	−2.951, 0.211
	+2.866	+3.080	+2.422	+2.656	+2.868
.9	−3.322, 0.106	−3.502, 0.099	−2.685, 0.095	−2.913, 0.092	−3.111, 0.091
	+3.238	+3.419	+2.601	+2.830	+3.028

these binomial distributions are equal when $\mu_2 - \mu_1 = 4\,\sigma_1$ and $1 \leq \sigma_2/\sigma_1 \leq 3$. It seems that when the difference of means increases it is necessary that the ratio of variances also increases for the binomial with two cut points to be more efficient that the binomial with only one cut point.

Observe that the cut points for $\mu_2 - \mu_1 = \sigma_1$ and $\sigma_2/\sigma_1 = 5$ given in Table 3.11, are very similar to the cut points when $\mu_2 - \mu_1 = 2\,\sigma_1$ and $\sigma_2/\sigma_1 = 5$ given in Table 3.15. If we use the optimal cut points for $\mu_2 - \mu_1 = \sigma_1$ and $\sigma_2/\sigma_1 = 5$ for the case when in fact $\mu_2 - \mu_1 = 2\,\sigma_1$ the $A.V$ given by the binomial distribution in the second case differs from the $A.V$ given by the optimal cut points only in the third decimal place. This fact is very important because it shows that the estimation of the cut points does not need to be very precise.

It is also possible to show that when $\mu_2 - \mu_1 = 4\sigma_1$ and $\sigma_2/\sigma_1 \leq 3$, the $A.V$ given by the binomial distribution with only one cut point differs from the variance given by the binomial with two cut points in the third decimal place.

In general we observe that when the distance between the means μ_1 and μ_2 is large in comparison with σ_1 , the variance given by the binomial distribution with only one cut point is similar to the variance given by the binomial with two cut points. In this case also if we define a trinomial distribution the decrement in the $A.V$ is not significant.

3.7 Summary

In this chapter, following Hettmansperger and Thomas [16], we proposed the inclusion of more cut points, provided m is large, in order to define a multinomial approach for

Table 3.12. *A.V* of $\sqrt{n}(\widehat{\lambda}-\lambda_0)$ and optimal cut point for $\mu_2-\mu_1$=2 σ_1 and $\sigma_2/\sigma_1 = 1.5$ for different values of λ_0 and m.

λ_0/m	3 (c, *A.V*)	4	6	8	10
.1	1.179, 0.178	1.371, 0.145	1.290, 0.115	1.506, 0.103	1.409, 0.097
.3	1.517, 0.327	1.605, 0.288	1.481, 0.246	1.382, 0.230	1.540, 0.221
.5	1.738, 0.369	1.390, 0.330	1.606, 0.288	1.465, 0.270	1.624, 0.261
.7	1.967, 0.313	1.512, 0.277	1.733, 0.242	1.554, 0.227	1.708, 0.220
.9	1.656, 0.150	1.739, 0.127	1.596, 0.109	1.702, 0.100	1.842, 0.096

Table 3.13. *A.V* of $\sqrt{n}(\widehat{\lambda}-\lambda_0)$ and optimal cut point for $\mu_2-\mu_1$=2 σ_1 and $\sigma_2/\sigma_1 = 2$ for different values of λ_0 and m.

λ_0/m	3 (c, *A.V*)	4	6	8	10
.1	1.361, 0.209	1.555, 0.159	1.881, 0.121	1.653, 0.107	1.819, 0.099
.3	1.711, 0.344	1.899, 0.295	2.188, 0.253	1.821, 0.233	1.993, 0.223
.5	1.958, 0.376	2.130, 0.332	1.778, 0.294	1.938, 0.273	2.104, 0.263
.7	2.215, 0.309	2.367, 0.277	1.885, 0.245	2.058, 0.229	2.217, 0.221
.9	2.632, 0.141	2.752, 0.126	2.085, 0.108	2.253, 0.100	2.397, 0.096

Table 3.14. *A.V* of $\sqrt{n}(\widehat{\lambda}-\lambda_0)$ and optimal cut points for $\mu_2-\mu_1$=2 σ_1 and $\sigma_2/\sigma_1 = 3$ for different values of λ_0 and m.

λ_0/m	3 (c_1/c_2, *A.V*)	4	6	8	10
.1	−2.137, 0.186	−2.349, 0.143	−2.720, 0.111	−2.435, 0.102	−2.585, 0.096
	+1.637	+1.849	+2.220	+1.933	+2.085
.3	−2.493, 0.315	−2.701, 0.272	−3.040, 0.238	−2.575, 0.225	−2.753, 0.217
	+1.993	+2.201	+2.540	+2.075	+2.433
.5	−2.740, 0.345	−2.934, 0.308	−3.245, 0.277	−2.683, 0.265	−2.861, 0.257
	+2.240	+2.434	+2.745	+2.183	+2.361
.7	−2.994, 0.281	−3.169, 0.255	−2.628, 0.235	−2.796, 0.222	−2.969, 0.216
	+2.494	+2.669	+2.129	+2.296	+2.469
.9	−3.396, 0.124	−3.538, 0.112	−2.803, 0.102	−2.980, 0.096	−3.142, 0.093
	+2.896	+3.039	+2.303	+2.480	+2.642

Table 3.15. $A.V$ of $\sqrt{n}(\widehat{\lambda}-\lambda_0)$ and optimal cut points for $\mu_2-\mu_1$=2 σ_1 and $\sigma_2/\sigma_1=5$ for different values of λ_0 and m.

λ_0/m	3 (c_1/c_2, $A.V$)	4	6	8	10
.1	−2.191, 0.133	−2.480, 0.111	−2.949, 0.096	−2.449, 0.093	−2.693, 0.091
	+2.025	+2.314	+2.782	+2.282	+2.526
.3	−2.565, 0.257	−2.832, 0.234	−2.405, 0.222	−2.617, 0.215	−2.856, 0.211
	+2.398	+2.665	+2.239	+2.451	+2.686
.5	−2.806, 0.292	−3.052, 0.272	−2.480, 0.261	−2.726, 0.253	−2.954, 0.251
	+2.640	+2.885	+2.314	+2.559	+2.788
.7	−3.045, 0.241	−3.268, 0.226	−2.587, 0.218	−2.834, 0.213	−3.055, 0.211
	+2.879	+3.102	+2.419	+2.668	+2.889
.9	−3.414, 0.104	−3.603, 0.097	−2.767, 0.094	−3.006, 0.091	−3.214, 0.095
	+3.247	+3.435	+2.601	+2.840	+3.047

the estimation of the mixing parameter of a two component mixture distribution. We proposed an optimal choice for these cut points (which defines the multinomial parameters), as the values which minimize the asymptotic variance of $\sqrt{n}(\widehat{\lambda}-\lambda_0)$.

We compare the binomial versus the trinomial approach for small m and large n, and we found that the trinomial approach is more efficient than the binomial approach and, in general, the multinomial approach with r classes is more efficient than the multinomial approach with $r-1$ classes.

For large m and n we found that all multinomial approaches are equivalent and that the asymptotic variance of $\sqrt{n}(\widehat{\lambda}-\lambda_0)$ has a very simple expression equal to $\lambda_0(1-\lambda_0)$. As mentioned earlier, this result is very logical, since it tells us that the maximum information that any of these approaches can give about the mixture distribution is the correct classification of the number of observations in each component. This is the number of observations of a binomial distribution with parameter λ_0, and that is why the limiting variance is equal to $\lambda_0\,(1-\lambda_0)/n$.

The numerical results show that the largest decrement in variance occurs when we introduce an additional cut point c to the binomial and we define a trinomial distribution. This result is very interesting because it tells us to use the trinomial approach instead of the binomial approach whenever possible, that is whenever $m \geq 2$.

When the component distributions are normal, with equal variances and the means are one standard deviation apart, the limiting variance is almost attained for $m = 40$, but when the means are two standard deviations apart the limiting variance is almost attained for $m = 10$. This behavior is also observed when the component distributions are normal with different variances and different means, as the distance between the means increases the variance of the estimator decreases and the limiting variance is attained with a smaller m. When we have different variances and different means, a binomial distribution with two cut points (see Figure 3.5) is better than the binomial distribution with only one cut point. If the separation between the means increases the difference between the asymptotic variances of the estimators of the mixing parameter for these two binomial approaches decreases, in other words, if the distance between the means is large the binomial with one cut point and the binomial with two cut points produce equivalent estimators of the mixing parameter. If the distance between the means is large, there exist an interval in which, for any cut point in this interval, the estimator of the mixing parameter almost attains the limiting variance.

Chapter 4

Nonparametric Inference in Mixture Model

4.1 Introduction

In the last two chapters, our approach for estimating the mixing parameter λ was to reduce the multivariate data to binomial, trinomial or in general multinomial responses. For this purpose we introduced some cut points c_i . In this chapter we will assume that we have a mixture of two continuous distribution functions with equal shapes and symmetric unimodal densities $f_1(x)$ and $f_2(x)$ which belong to the same location family with medians M_1 and M_2, and that we have univariate responses. We will try to obtain estimates of the mixing parameter λ by means of a tetranomial or a sextinomial distribution defined by some cut points c_i'.

4.2 Tetranomial Approach

The definition of the tetranomial distribution can be better explained with the help of Figure 4.1. In this figure it is assumed that we have two symmetric distributions with equal shapes and densities $f_1(x)$ and $f_2(x)$ which belong to the same location family with unknown medians M_1 and M_2.

If the medians were known we could select the cut points: $c_1 = M_1$, $c_2 = M_c = (M_1 + M_2)/2$, and $c_3 = M_2$. These cut points define four regions R_k, $k = 1, 2, 3, 4$, where $R_1 = \{x \mid x \leq M_1\}$, $R_2 = \{x \mid M_1 < x \leq M_c\}$, $R_3 = \{x \mid M_c < x \leq M_2\}$ and $R_4 = \{x \mid M_2 < x \}$.

Let x_j for $j = 1, \ldots, n$, be the observations of the continuous random variable whose distribution $f(x)$ is a mixture of the two densities $f_1(x)$ and $f_2(x)$.

$$f(x) = \lambda f_1(x) + (1 - \lambda) f_2(x) \tag{4.1}$$

In order to define a tetranomial distribution let:

$$z_{jk} = \begin{cases} 1 & \text{if observation } x_j \text{ is in region k=1, 2, 3, 4} \\ 0 & \text{otherwise.} \end{cases}$$

Observe now that $z_{jk} = z_{jk1} + z_{jk2}$, where:

$$z_{jki} = \begin{cases} 1 & \text{if observation } x_j \text{ has density } f_i(x), \text{ i=1,2 and is in region k=1, 2, 3, 4.} \\ 0 & \text{otherwise.} \end{cases}$$

In this way (Figure 4.1):

$$\begin{aligned} P(z_{j11} = 1) &= P(z_{j42} = 1) = .5 \\ P(z_{j21} = 1) &= P(z_{j32} = 1) = p_1 \\ P(z_{j31} = 1) &= P(z_{j22} = 1) = p_2 \\ P(z_{j41} = 1) &= P(z_{j12} = 1) = .5 - p_1 - p_2 \end{aligned} \tag{4.2}$$

If we let $\mathbf{z}_{j1} = (z_{j11},\ z_{j21},\ z_{j31},\ z_{j41})$ be the tetranomial random variable which corresponds to the density function $f_1(x)$ and $\mathbf{z}_{j2} = (z_{j12},\ z_{j22},\ z_{j32},\ z_{j42})$ is the tetranomial random variable which corresponds to the density function $f_2(x)$, then $\mathbf{z}_j =$

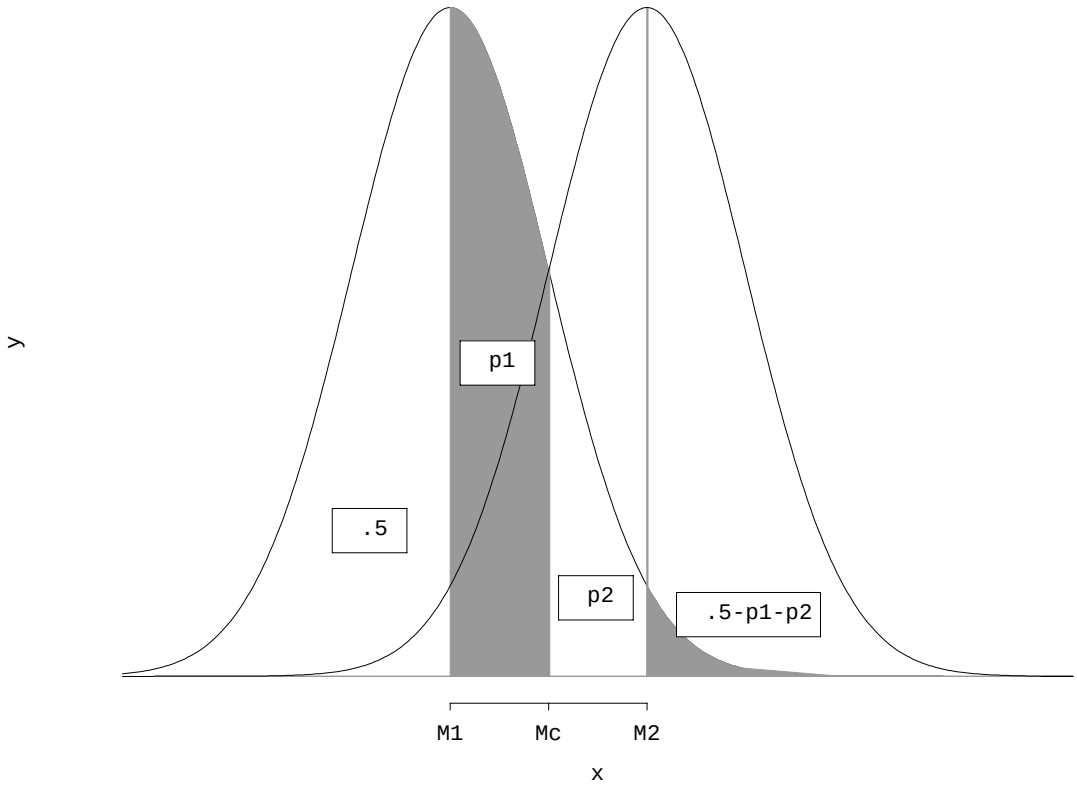

Fig. 4.1. Regions for the Tetranomial Distribution

$\mathbf{z}_{j1}+\mathbf{z}_{j2} = (z_{j1},\ z_{j2},\ z_{j3},\ z_{j4})$, is the tetranomial variable which corresponds to the continuous variable y_j, and has mixture of two tetranomial distributions $m_1(z)$ and $m_2(z)$:

$$m(\mathbf{z}) = \lambda m_1(\mathbf{z}) + (1-\lambda)m_2(\mathbf{z}),$$

where:

$$\begin{aligned} m_1(\mathbf{z}_j\ ; p_1,\, p_2) &= m_1(z_{j1},\, z_{j2},\, z_{j3};\, p_1,\, p_2) \\ &= (.5)^{\,z_{j1}}\ p_1^{z_{j2}}\ p_2^{z_{j3}}\ (.5-p_1-p_2)^{1-z_{j1}-z_{j2}-z_{j3}} \\ m_2(\mathbf{z}_j;\, p_1,\, p_2) &= m_2(z_{j1},\, z_{j2},\, z_{j3};\, p_1,\, p_2) \\ &= (.5-p_1-p_2)^{z_{j1}}\ p_2^{z_{j2}}\ p_1^{z_{j3}}\ (.5)^{\,1-z_{j1}-z_{j2}-z_{j3}} \end{aligned}$$

If we let $n_k = \Sigma_{j=1}^{n} z_{jk}$, then n_k is the number of observations in region R_k, and $\mathbf{n}' = (n_1,\ n_2\ , n_3,\ n_4)$ can be interpreted as a single observation from a tetranomial distribution with parameters $n = \Sigma_{k=1}^{4} n_k$, $P(R_1),\ P(R_2),\ P(R_3)$; where $P(R_k)$, defined below, is the probability that one observation falls in region $R_k,\ k = 1,\ 2,\ 3,\ 4$.

$$\begin{aligned} P(R_1) &= .5 - (1-\lambda)(p_1+p_2) \\ P(R_2) &= p_2 + \lambda(p_1-p_2) \\ P(R_3) &= p_1 + \lambda(p_2-p_1) \\ P(R_4) &= .5 - \lambda(p_1+p_2) \end{aligned}$$

We note again that the parameters λ, p_1 and p_2 are unknown. If they were known they determine uniquely the values of M_1 and M_2 the medians of the component distributions. If $F(x)$ is the distribution function of the mixture:

$$M_1 = F^{-1}[P(R_1)] = F^{-1}[.5 - (1-\lambda)(p_1 + p_2)]$$

$$M_2 = F^{-1}[P(R_1) + P(R_2) + P(R_3)] = F^{-1}[.5 + \lambda(p_1 + p_2)] \tag{4.3}$$

4.3 Sextinomial Approach

The sextinomial approach can be better explained with the help of Figure 4.2. In this figure it is assumed again that we have two distributions with equal shapes and symmetric unimodal densities $f_1(x)$ and $f_2(x)$ which belong to the same location family with unknown medians M_1 and M_2. If the medians were known we could, with $D = (M_2 - M_1)/2$, select the cut points: $c_1 = M_i = M_1 - D$, $c_2 = M_1$, $c_3 = M_c = (M_1 + M_2)/2$, $c_4 = M_2$ and $c_5 = M_s = M_2 + D$. These cut points define six regions R_k, $k = 1, 2, \ldots, 6$, where $R_1 = \{x \mid x \leq M_i\}$, $R_2 = \{x \mid M_i < x \leq M_1\}$, $R_3 = \{x \mid M_1 < x \leq M_c\}$, $R_4 = \{x \mid M_c < x \leq M_2\}$, $R_5 = \{x \mid M_2 < x \leq M_s\}$, and $R_6 = \{x \mid M_s < x\ \}$.

Let x_j for $j = 1, \ldots, n$, be the observations of the continuous random variable whose distribution $f(x)$ is the mixture of the two densities $f_1(x)$ and $f_2(x)$ given by Equation 4.1.

In order to define a sextinomial distribution let:

$$z_{jk} = \begin{cases} 1 & \text{if observation } y_j \text{ is in region k=1, 2, . . . ,6} \\ 0 & \text{otherwise.} \end{cases}$$

Observe now that $z_{jk} = z_{jk1} + z_{jk2}$, where:

$$z_{jki} = \begin{cases} 1 & \text{if observation } y_j \text{ has density } f_i(x), \text{ i=1,2 and is in region k=1, 2, . . . ,6} \\ 0 & \text{otherwise.} \end{cases}$$

Let $\mathbf{z}_{j1} = (z_{j11},\ z_{j21},\ z_{j31},\ z_{j41},\ z_{j51},\ z_{j61}\ )$ be the sextinomial random variable which corresponds to the density function $f_1(x)$ and $\mathbf{z}_{j2} = (z_{j12},\ z_{j22},\ z_{j32},\ z_{j42},\ z_{j52},\ z_{j62})$ be the sextinomial random which corresponds to the density function $f_2(x)$, then $\mathbf{z}_j = \mathbf{z}_{j1} + \mathbf{z}_{j2} = (z_{j1},\ z_{j2},\ z_{j3},\ z_{j4},\ z_{j5},\ z_{j6})$, is the sextinomial variable which corresponds to the continuous variable y_j, which distribution is a mixture of two sextinomial distributions $s_1(z)$ and $s_2(z)$:

$$s(\mathbf{z}) = \lambda s_1(\mathbf{z}) + (1-\lambda) s_2(\mathbf{z})$$

Where:

$$s_1(\mathbf{z}_j\ ; p_1, p_2, p_3) = s_1(z_{j1}, z_{j2}, z_{j3}, z_{j4}, z_{j5}; p_1, p_2, p_3)$$
$$= (.5 - p_1)^{\,z_{j1}}\ p_1^{z_{j2}+z_{j3}}\ p_2^{z_{j4}}\ p_3^{z_{j5}} (.5 - p_1 - p_2 - p_3)^{1-z_{j1}-z_{j2}-z_{j3}-z_{j4}-z_{j5}}$$

$$s_2(\mathbf{z}_j\ ; p_1, p_2, p_3) = s_2(z_{j1}, z_{j2}, z_{j3}, z_{j4}, z_{j5}; p_1, p_2, p_3)$$
$$= (.5 - p_1 - p_2 - p_3)^{z_{j1}}\ p_3^{z_{j2}}\ p_2^{z_{j3}}\ p_1^{z_{j4}+z_{j5}}\ (.5 - p_1)^{\,1-z_{j1}-z_{j2}-z_{j3}-z_{j4}-z_{j5}}$$

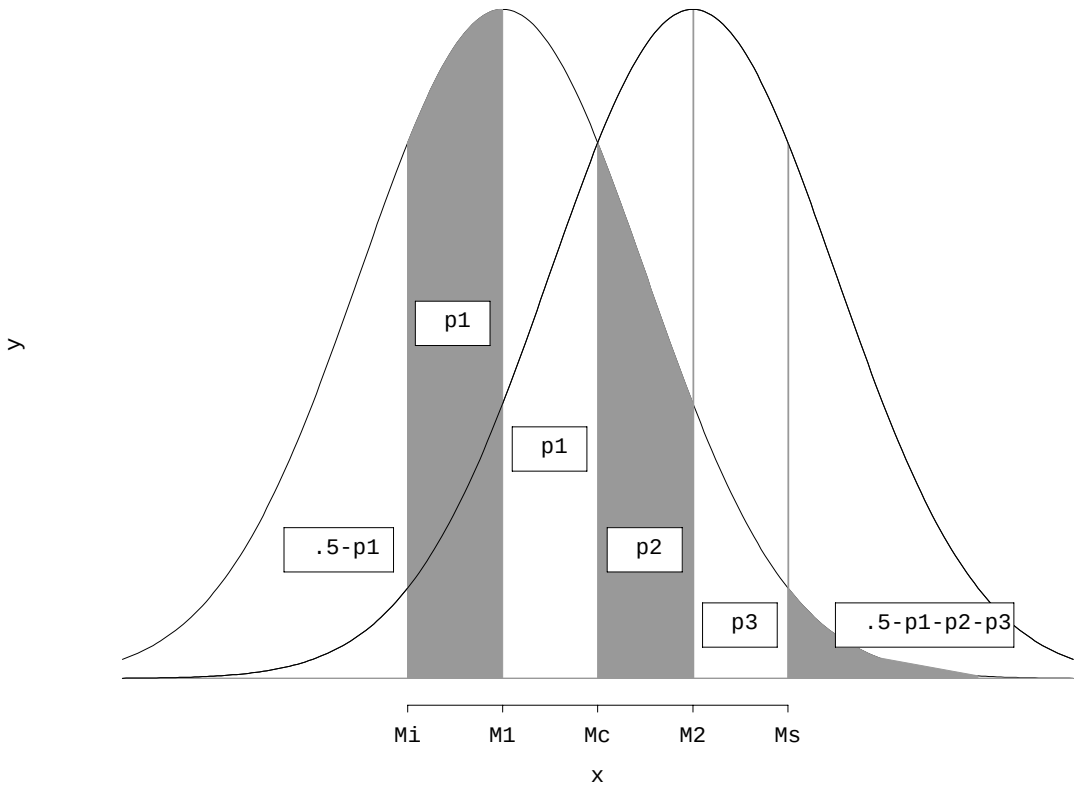

Fig. 4.2. Regions for the Sextinomial Distribution

and:

$$P(\ z_{j11}=1)=P(1-z_{j12}-z_{j22}-z_{j32}-z_{j42}-z_{j52}=1)=.5-p_1$$

$$P(\ z_{j12}=1)=P(1-z_{j11}-z_{j21}-z_{j31}-z_{j41}-z_{j51}=1)=.5-p_1-p_2-p_3$$

$$P(\ z_{j21}=1)=P(z_{j31}=1)=P(\ z_{j42}=1)=P(z_{j52}=1)=p_1$$

$$P(\ z_{j41}=1)=P(z_{j32}=1)=p_2$$

$$P(\ z_{j51}=1)=P(z_{j22}=1)=p_3 \tag{4.4}$$

If we let $n_k = \Sigma_{j=1}^{n} z_{jk}$, then n_k is the number of observations in region R_k, and $\mathbf{n}' = (n_1,\ n_2\ , n_3,\ n_4,\ n_5,\ n_6)$ can be interpreted as a single observation from a sextinomial distribution with parameters $n = \Sigma_{k=1}^{6} n_k$, and $P(R_1),\ P(R_2), \ldots, P(R_6)$, where $P(R_k)$, given below, is the probability that one observation falls in region R_k, $k = 1, 2, \ldots, 6$.

$$P(R_1)=.5-p_1-(1-\lambda)(p_2+p_3)$$

$$P(R_2)=\lambda p_1+(1-\lambda)p_3$$

$$P(R_3)=\lambda p_1+(1-\lambda)p_2$$

$$P(R_4)=\lambda p_2+(1-\lambda)p_1$$

$$P(R_5)=\lambda p_3+(1-\lambda)p_1$$

$$P(R_6)=.5-p_1-\lambda(p_2+p_3)$$

We note again that the parameters $\lambda,\ p_1,\ p_2$ and p_3 are unknown. If they were known, along with $F(x)$ the distribution function of the mixture, they determine uniquely the values of M_1 and M_2 the medians of the component distributions. Note that M_1 and M_2 are given again by Equation 4.3.

$$M_1 = F^{-1}[P(R_1) + P(R_2)] = F^{-1}[.5 - (1-\lambda)(p_1 + p_2)]$$

$$M_2 = F^{-1}[1 - P(R_5) - P(R_6)] = F^{-1}[.5 + \lambda(p_1 + p_2)]$$

4.4 Methods of Estimation

We will analyze six approaches in order to estimate the parameters in the tetranomial and sextinomial mixtures. Our first attempt was to obtain the maximum likelihood estimators of the parameters in these mixture models.

4.4.1 Likelihood Function

4.4.1.1 Likelihood function for the Tetranomial distribution

We define the indicator variable t_j as:

$$t_j = \begin{cases} 1 & \text{if observation } y_j \text{ belongs to population 1} \\ 0 & \text{otherwise.} \end{cases}$$

The likelihood function in terms of individual observations is:

$$L(\theta, z_{jk}) = \Pi_{j=1}^{n} \left[m_1(z_{jk}; p_1, p_2)\right]^{t_j} \left[m_2(z_{jk}; p_1, p_2)\right]^{1-t_j} \lambda^{t_j} (i-\lambda)^{1-t_j}$$

therefore:

$$\begin{aligned}
\log[L(\theta, z_{jk})] &= l(\theta, z_{jk}) \\
&= \left[\Sigma_{j=1}^{n}\, z_{j1}\, t_j + \Sigma_{j=1}^{n}\, z_{j4}\,(1-t_j)\right] \log(.5) + \\
&+ \left[\Sigma_{j=1}^{n}\, z_{j2}\, t_j + \Sigma_{j=1}^{n}\, z_{j3}\,(1-t_j)\right] \log(p_1) + \\
&+ \left[\Sigma_{j=1}^{n}\, z_{j3}\, t_j + \Sigma_{j=1}^{n}\, z_{j2}\,(1-t_j)\right] \log(p_2) + \\
&+ \left[\Sigma_{j=1}^{n}\, z_{j4}\, t_j + \Sigma_{j=1}^{n}\, z_{j1}\,(1-t_j)\right] \log(.5 - p_1 - p_2) + \\
&+ \Sigma_{j=1}^{n}\, t_j \log(\lambda) + \Sigma_{j=1}^{n}\,(1-t_j) \log(1-\lambda)
\end{aligned}$$

where $\theta' = (\lambda, p_1, p_2)$. Observe that $z_{j1k} = z_{jk}\, t_j$, and $\Sigma_{j=1}^{n}\, z_{jk}\, t_j = \Sigma_{j=1}^{n}\, z_{j1k} = n_{1k}$, is the number of observations of population 1 which belongs to region k.

4.4.1.2 Likelihood function for the Sextinomial distribution

With the same definition of the indicator variable t_j given before, the likelihood and log-likelihood functions for the sextinomial distribution in terms of individual observations are:

$$L(\theta, z_{jk}) = \Pi_{j=1}^{n} \left[s_1(z_{jk}; p_1, p_2, p_3)\right]^{t_j} \left[s_2(z_{jk}; p_1, p_2, p_3)\right]^{1-t_j} \lambda^{t_j} (i-\lambda)^{1-t_j}.$$

$$\log\left[L(\theta, z_{jk})\right] = l(\theta, z_{jk})$$

$$= \left[\Sigma_{j=1}^{n}\, z_{j1}\, t_j + \Sigma_{j=1}^{n}\, z_{j6}\,(1-t_j)\right]\log(.5-p_1) +$$

$$+ \left[\Sigma_{j=1}^{n}\, z_{j2}\, t_j + \Sigma_{j=1}^{n}\, z_{j3}\, t_j + \Sigma_{j=1}^{n}\, z_{j4}\,(1-t_j) + \Sigma_{j=1}^{n}\, z_{j5}\,(1-t_j)\right]\log(p_1) +$$

$$+ \left[\Sigma_{j=1}^{n}\, z_{j4}\, t_j + \Sigma_{j=1}^{n}\, z_{j3}\,(1-t_j)\right]\log(p_2) +$$

$$+ \left[\Sigma_{j=1}^{n}\, z_{j5}\, t_j + \Sigma_{j=1}^{n}\, z_{j2}\,(1-t_j)\right]\log(p_3) +$$

$$+ \left[\Sigma_{j=1}^{n}\, z_{j6}\, t_j + \Sigma_{j=1}^{n}\, z_{j1}\,(1-t_j)\right]\log(.5-p_1-p_2-p_3) +$$

$$+ \Sigma_{j=1}^{n}\, t_j\, \log(\lambda) + \Sigma_{j=1}^{n}\,(1-t_j)\log(1-\lambda)$$

where $\theta^{'} = (\lambda,\, p_1,\, p_2,\, p_3)$. Observe that $z_{jk1} = z_{jk}\, t_j$, and $\Sigma_{j=1}^{n}\, z_{jk}\, t_j = \Sigma_{j=1}^{n}\, z_{jk1}$ is the number of observations of population 1 which belongs to region k.

It is important to note that we do not know the indicator variable t_j, and we do not know the population medians M_1 and M_2, therefore z_{jki} are conditional observations given M_1 and M_2. We need a prior distribution on λ, M_1 and M_2 (or λ, p_1 and p_2) in order to obtain the likelihood function for z_{jki} (there is already literature about Bayesian methods in mixture models that can be used for this case, for example Bernardo [1] or Diebolt [8]). We will not assume a prior distribution on the medians so we will not have MLE of the parameters in our mixture models. In order to obtain estimates of the parameters $\lambda,\ p_1,\ p_2,\ M_1$ and M_2 we will use the following methods:

4.4.2 Least Squares Estimates.

In this method we will minimize $\Sigma_{k=1}^{4}[n_k - E(n_k)]^2$ for the tetranomial distribution or $\Sigma_{k=1}^{6}[n_k - E(n_k)]^2$ for the sextinomial distribution. Where n_k as mentioned before, is

the observed number of observations in Region k, and $E(n_k)$ is the expected value of the number of observations in that region:

$$E(n_k) = nP(R_k)$$

$P(R_k)$ was given before for the tetranomial and the sextinomial approaches.

4.4.3 Chi-Square Method

In this method we will minimize: $\Sigma_{k=1}^{4}[n_k - E(n_k)]^2/E(n_k)$ for the tetranomial distribution or $\Sigma_{k=1}^{6}[n_k - E(n_k)]^2/E(n_k)$ for the sextinomial distribution.

4.4.4 Modified Chi-Square Method

Here we will minimize $\Sigma_{k=1}^{4}[n_k - E(n_k)]^2/[n_k + E(n_k)]$ for the tetranomial distribution or $\Sigma_{k=1}^{6}[n_k - E(n_k)]^2/[n_k + E(n_k)]$ for the sextinomial distribution.

4.4.5 Hellinger 1

In this approach we will try to minimize $\Sigma_{k=1}^{r}\left[\sqrt[2]{n_k/n} - \sqrt[2]{E(n_k)/n}\right]^2$, $r = 4$ for the tetranomial distribution and $r = 6$ for the sextinomial approach. Note that:

$$\Sigma_{k=1}^{r}\left[\sqrt[2]{n_k/n} - \sqrt[2]{E(n_k)/n}\right]^2 = 2 - 2\,\Sigma_{k=1}^{r}\sqrt[2]{n_k E(n_k)/n^2}$$

Therefore the minimization of $\Sigma_{k=1}^{r}\left[\sqrt[2]{n_k/n} - \sqrt[2]{E(n_k)/n}\right]^2$ is equivalent to the maximization of $\Sigma_{k=1}^{r}\sqrt[2]{n_k E(n_k)/n^2}$ or to the minimization of any decreasing function of this last expression.

4.4.6 Hellinger 2

Here, we will minimize $\cos^{-1}(\Sigma_{k=1}^{r} \sqrt[2]{n_k E(n_k)/n^2})$, note that the minimum value of this expression is *zero* when $n_k = E(n_k)$ for $k = 1, 2, \ldots, r$.

4.5 Estimators

As we have seen in Sections 4.2 and 4.3, the parameters λ, p_1 and p_2 determine uniquely the medians M_1 and M_2 of the component distributions if the distribution function $F(x)$ of the mixture is known. In our case this distribution is unknown but if we assumed that λ, p_1 and p_2 are known we could estimate M_1 and M_2 with the help of the empirical *cdf* :

$$\begin{aligned} \widehat{M_1} &= m_1 = y_{(q_1)} \\ \widehat{M_2} &= m_2 = y_{(q_2)} \end{aligned} \tag{4.5}$$

where $q_1 = n[.5 - (1 - \lambda)(p_1 + p_2)$, and $q_2 = n[.5 + \lambda(p_1 + p_2)]$. See equation 4.3.

The estimates of the parameters in the mixture model will be obtained with a grid search described below.

The estimators of the parameters M_1, M_2 and the mixing parameter λ permit us to obtain an estimator of the common variance σ_0^2 (if this variance exists), of the component populations. Assuming the mixture model given by Equation 4.1 and considering that X_i has distribution $f_i(x)$ and $E(X_i)$ exists for $i = 1, 2$, we have:

$$E(X) = \mu = \lambda\mu_1 + (1 - \lambda)\mu_2$$

$$E(X^2) = \lambda E({X_1}^2) + (1-\lambda)E({X_2}^2)$$

with this results it can be shown that the variance of the mixture is:

$$\sigma^2 = \sigma_0^2 + \mu_1^2\lambda(1-\lambda) + \mu_2^2(1-\lambda)[1-(1-\lambda)] - 2\lambda(1-\lambda)(\mu_1-\mu_2)^2$$
$$\sigma^2 = \sigma_0^2 + \lambda(1-\lambda)(\mu_1-\mu_2)^2$$

therefore an estimator of the common variance σ_0^2, of the component populations is:

$$\hat{\sigma}_0^2 = \hat{\sigma}^2 - \hat{\lambda}(1-\hat{\lambda})(m_1-m_2)^2 \tag{4.6}$$

where m_1 and m_2 are defined in Equation 4.5 and $\hat{\sigma}^2$ is the variance of the sample that comes from the mixture model.

Observe that we are using m_i as an estimator of μ_i for $i = 1, 2$, this can be done if we assumed that the μ_i's exist for each component population, because we have assumed that distribution of these populations are symmetric.

4.5.1 Tetranomial Approach

Our numerical method of estimation for the tetranomial distribution will assume values for λ, p_1 and p_2; these values will give the estimates m_1 and m_2 (Equation 4.5) and $m_c = \widehat{M_c} = (m_1 + m_2)/2$, which in turn will be used to determine the four regions for the tetranomial distribution (Figure 4.1). The estimates of the parameters λ, M_1 and M_2 will be the values $\widehat{\lambda}$, m_1 and m_2 that optimize the criteria listed in subsections 4.4.2 to 4.4.6.

For the tetranomial distribution in the grid search the parameter estimates will take the values:

$\lambda = 0.05, 0.06, \ldots, 0.95.$

$p_1 = 0.20, 0.21, \ldots, 0.50.$

$p_2 = 0.01, 0.02, \ldots, \min(p_1, .5 - p_1)$

With the restrictions $p_1 > p_2 > .5 - p_1 - p_2$.

It is interesting to observe that for the tetranomial distribution (given the definition of m_1 and m_2 as percentiles of the mixed distribution), only regions R_2 and R_3 contribute to the sum of squares for the methods described above. In Appendix B we present the codes in Mathematica and in Splus for the minimization of the criteria for the tetranomial and the sextinomial approaches.

4.5.2 Sextinomial Approach

In this case, as in the tetranomial distribution, we will assume values for λ, p_1, p_2 and p_3; these values will give the estimates m_1, m_2 (defined before), $m_i = \widehat{M_i} = m_1 - d$, $m_c = \widehat{M_c} = m_1 + d$, and $m_s = \widehat{M_s} = m_2 + d$ (where $d = (m_1 + m_2)/2$). These values will be used to determine the six regions for the sextinomial distribution (Figure 4.2). The estimates of the parameters λ, M_1 and M_2 will be the values $\widehat{\lambda}$, m_1 and m_2 that optimize the criteria listed in subsections 4.4.2 to 4.4.6. For this approach in the grid search the parameter estimates will take the values:

$\lambda = 0.05, 0.06, \ldots, 0.95.$

$p_1 = 0.20, 0.21, \ldots, 0.50.$

$p_2 = 0.01, 0.02, \ldots, \min(p_1, .5 - p_1)$

$p_3 = 0.01,\ 0.02, \ldots, \min(p_1,\ p_2,\ .5 - p_1 - p_2)$

4.5.3 Lower Bound for the Variance of the Multinomial Estimators

It seems difficult to obtain exact expressions for the variances of the estimators of the mixing parameter given by the methods mentioned above. Considering the original mixture model 4.1, we can obtain a lower bound for the variances of the estimators of the mixing parameter with the results given by Hill [17], Hill found that the lower bound for the asymptotic variance of $\sqrt{n}(\widehat{\lambda} - \lambda_0)$ is equal to:

$$L_b = I(\lambda)^{-1} = \lambda\,(1-\lambda)\left[1 - \int_{\infty}^{\infty} \frac{f_1(x)\,f_2(x)}{f(x)}dx\right]^{-1}$$

where $I(\lambda)$ is the Fisher information for λ. We can approximate the integral in this expression by

$$\sum_{k=1}^{r} \frac{p_{k1}\,p_{k2}}{\lambda\,p_{k1} + \lambda\,p_{k2}}$$

where $p_{ki} = P(z_{jki} = 1)$ for $i = 1, 2$ are given in Equation 4.2 for the tetranomial distribution ($r = 4$) and in Equation 4.4 for the sextinomial distribution ($r = 6$).

If we denote $\left[1 - \int_{\infty}^{\infty} \frac{f_1(x)\,f_2(x)}{f(x)}dx\right]^{-1}$ as the inflation factor for the variance, Table 4.1 shows that the approximation to the integral by the summation is good for normal components when $M_1 - M_2 \geq 1.5$. In this table $\lambda = 0.25$, the first component is $N(0,\,1)$ and the second component is on the first row. The sum $\sum_{k=1}^{r} \frac{p_{k1}\,p_{k2}}{\lambda\,p_{k1}+\lambda\,p_{k2}}$ can be estimated plugging in the estimates of the $p's$.

Table 4.1. Inflation factors for the variance for normal components when $\lambda = 0.25$

Component	$N(1, 1)$	$N(1.75, 1)$	$N(2.32, 1)$	$N(3.11, 1)$
Normal	5.8343	2.3414	1.6351	1.2464
Tetranomial	6.8444	2.4938	1.7428	1.3314
Sextinomial	6.2026	2.4501	1.7410	1.3314

4.6 Monte Carlo Study

4.6.1 Normal Component Distributions

In order to compare the performance of the multinomial estimates with the MLE estimates (obtained with the EM algorithm) when the components of the mixture are normal distributions we performed 500 simulations considering "overlaps" equal to: 0.03, 0.05, 0.10 and 0.15. Woodward et al. [36] defined overlap (ov) as the probability of misclassification using this rule: Classify an observation x as being from population 1 if $x < x_c$ and from population 2 if $x \geq x_c$, where x_c is the unique point between μ_1 and μ_2 such that $\lambda f_1(x_c) = (1-\lambda) f_2(x_c)$. For example if the distributions have the same variance and $f_1(x)$ is the standard normal distribution, for an overlap equal to 0.10 ($ov = 0.10$) when $\lambda = 0.5$, necessarily $\mu_2 = 2.56$ and $x_c = 1.28$, because in this case $f_1(x_c) = f_2(x_c)$ and $\lambda P_1(X \geq x_c = 1.28) = 0.05$ (where P_1 is associated with $f_1(x)$). Figure 4.3 shows this fact.

4.6.1.1 Simulations for Normal Component Distributions

The results of the simulation study for a sample size equal to 100, mixing parameter $\lambda = 0.25$, and overlap = 0.15 ($M_1 = 0$ and $M_2 = 1.75$) are presented in Table 4.2. For

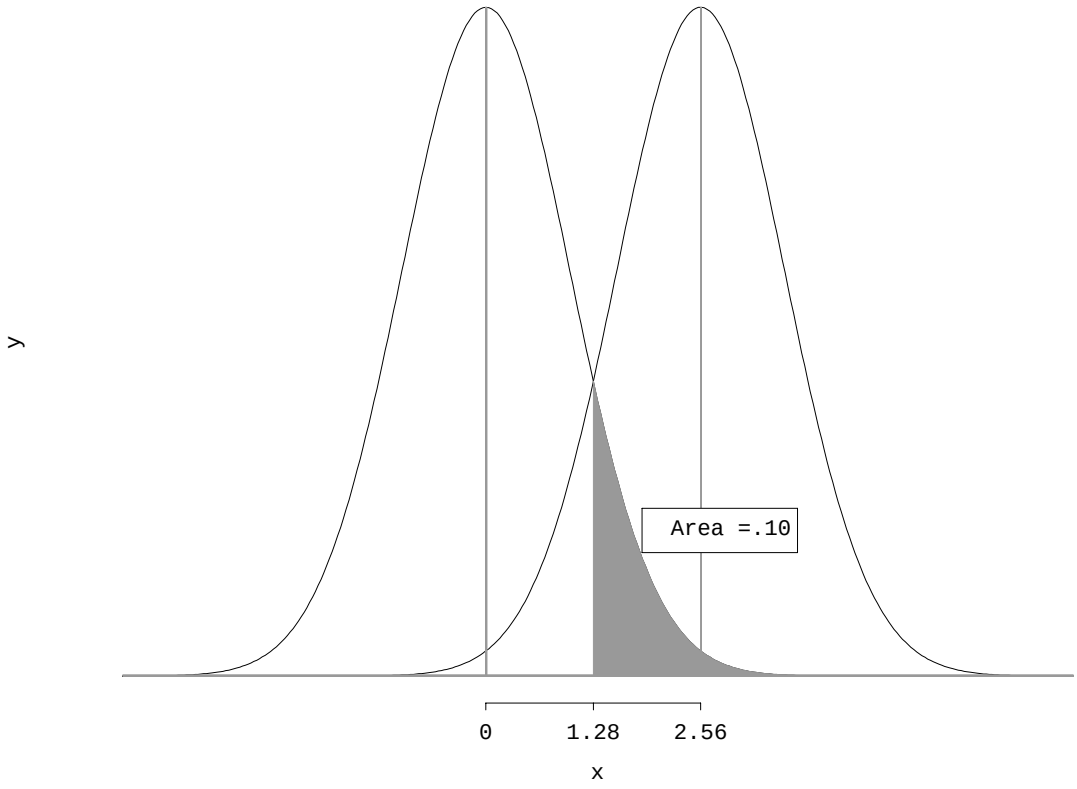

Fig. 4.3. Overlap equal to 0.10 for Normal Distributions when $\lambda = 0.5$. In this figure the total probability of misclassification is equal to 0.20, but if we consider that $\lambda = 0.50$, the probability of misclassification in the mixture distribution is equal to $\lambda \times 0.20 = 0.10$.

this case, the Hellinger 2 method for the tetranomial distribution seems to be better, its variance is 30% of the variance of the MLE estimator. Observe that in 51% of the cases the estimate coincides with the true value of the parameter. The performance of the sextinomial criteria are similar to the performance of the MLE estimator.

Table 4.2. Properties of $\widehat{\lambda}$ for normal components, n=100, $\lambda = 0.25$, and ov[1]=0.15

Method	LQ[2]	Median	UQ[3]	Mean	P[4]	Var $(\widehat{\lambda})$[5]
EM	.1909	.2942	.4074	.3187		0.03485
			Tetranomial			
LS	.25	.25	.50	.3607	0.522	0.03332
χ^2	.25	.25	.50	.3668	0.514	0.03516
Mod. χ^2	.25	.25	.50	.3668	0.514	0.03516
H1	.25	.25	.25	.3068	0.616	0.02219
H2	.20	.25	.25	.2458	0.510	0.01247
			Sextinomial			
LS	.17	.31	.4725	.3398	0.022	0.04180
χ^2	.18	.32	.4800	.3497	0.024	0.04109
Mod. χ^2	.18	.32	.4800	.3496	0.024	0.04113
H1	.18	.32	.4800	.3496	0.024	0.04113
H2	.18	.32	.4800	.3496	0.024	0.04113

Results of the simulation study for a sample size equal to 100, mixing parameter $\lambda = 0.25$, and overlap = 0.10 ($M_1 = 0$ and $M_2 = 2.32$) are presented in Table 4.3. In this case the MLE estimator seems to be the best and the Hellinger 2 method for the

[1]Overlap defined in Subsection 4.6.1
[2]Lower Quartile
[3]Upper Quartile
[4]Proportion of times the estimate coincides with the parameter
[5]Variance of $\widehat{\lambda}$ based on 500 samples of size $n = 100$

tetranomial the second best. Observe that the sextinomial approach works well for all the criteria.

Table 4.3. Properties of $\widehat{\lambda}$ for normal components, n=100, $\lambda = 0.25$, and ov=0.10

Method	LQ	Median	UQ	Mean	P	Var $(\widehat{\lambda})$
EM	.2138	.2651	.3328	.2746		0.01014
Tetranomial						
LS	.25	.25	.50	.3972	0.462	0.03983
χ^2	.25	.25	.50	.3982	0.458	0.03982
Mod. χ^2	.25	.25	.50	.3982	0.458	0.03982
H1	.25	.25	.25	.3013	0.550	0.02223
H2	.25	.25	.25	.2870	0.582	0.01761
Sextinomial						
LS	.15	.24	.3325	.2612	0.036	0.02371
χ^2	.16	.25	.3425	.2732	0.038	0.02488
Mod. χ^2	.1675	.25	.3500	.2732	0.038	0.02480
H1	.1675	.25	.3500	.2732	0.038	0.02480
H2	.1675	.25	.3500	.2732	0.038	0.02480

The results for 500 simulations for a sample size equal to 50 when $\lambda = 0.25$, overlap = 0.10, $M_1 = 0$ and $M_2 = 1.75$, are presented in Table 4.4. The results for this sample size are similar to the results obtained for a sample size equal to 100 observations.

Results for a sample size equal to 50, $\lambda = 0.25$, and overlap = 0.10 ($M_1 = 0$ and $M_2 = 2.32$) are presented in table 4.5

Results of the simulation study for a sample size equal to 100, mixing parameter $\lambda = 0.25$, and overlap = 0.05 ($M_1 = 0$ and $M_2 = 3.11$) are presented in Table 4.6.

Results of the simulation study for a sample size equal to 100, mixing parameter $\lambda = 0.25$, and overlap = 0.03 ($M_1 = 0$ and $M_2 = 3.6$). are presented in Table 4.7.

Table 4.4. Properties of $\widehat{\lambda}$ for normal components, n=50, $\lambda = 0.25$, and ov=0.15

Method	LQ	Median	UQ	Mean	P	Var $(\widehat{\lambda})$
EM	.1909	.2942	.4074	.3187		0.03485
			Tetranomial			
LS	.3	.40	.50	.4145	0.128	0.03421
χ^2	.3	.40	.55	.4212	0.122	0.03435
Mod. χ^2	.3	.40	.55	.4212	0.122	0.03435
H1	.2	.25	.25	.2508	0.562	0.01407
H2	.2	.25	.25	.2189	0.480	0.01048
			Sextinomial			
LS	.15	.31	.5425	.3571	0.010	0.05434
χ^2	.18	.34	.58	.3762	0.016	0.05192
Mod. χ^2	.18	.34	.58	.3785	0.016	0.05241
H1	.18	.345	.58	.3788	0.016	0.05241
H2	.18	.345	.58	.3788	0.016	0.05241

Table 4.5. Properties of $\widehat{\lambda}$ for normal components, n=50, $\lambda = 0.25$, and ov=0.10

Method	LQ	Median	UQ	Mean	P	Var $(\widehat{\lambda})$
EM	.1877	.2692	.3573	.2890		0.02233
			Tetranomial			
LS	.25	.40	.60	.4212	0.188	0.03673
χ^2	.25	.40	.60	.4237	0.176	0.03687
Mod. χ^2	.25	.40	.60	.4237	0.176	0.03687
H1	.25	.25	.35	.3264	0.478	0.02812
H2	.25	.25	.30	.3056	0.464	0.02475
			Sextinomial			
LS	.13	.23	.3725	.2824	0.008	0.04007
χ^2	.1575	.27	.41	.3067	0.012	0.04277
Mod. χ^2	.16	.27	.42	.3079	0.012	0.04272
H1	.16	.27	.42	.3082	0.012	0.04262
H2	.16	.27	.42	.3082	0.012	0.04262

Table 4.6. Properties of $\widehat{\lambda}$ for normal components, n=100, $\lambda = 0.25$, and ov=0.05

Method	LQ	Median	UQ	Mean	P	Var $(\widehat{\lambda})$
EM	.2229	.2575	.2946	.2602		0.00283
			Tetranomial			
LS	.25	.25	.75	.4467	0.388	0.05708
χ^2	.25	.30	.75	.4514	0.378	0.05820
Mod. χ^2	.25	.30	.75	.4514	0.378	0.05820
H1	.25	.25	.60	.3869	0.434	0.05023
H2	.20	.25	.50	.3437	0.422	0.04738
			Sextinomial			
LS	.19	.24	.29	.2420	0.088	0.00743
χ^2	.19	.25	.30	.2499	0.054	0.01079
Mod. χ^2	.19	.25	.30	.2546	0.066	0.01281
H1	.1975	.25	.30	.2563	0.068	0.01335
H2	.1975	.25	.30	.2563	0.068	0.01335

Table 4.7. Properties of $\widehat{\lambda}$ for normal components, n=100, $\lambda = 0.25$, and ov=0.03

Method	LQ	Median	UQ	Mean	P	Var $(\widehat{\lambda})$
EM	.2252	.2547	.2882	.2569		0.00212
			Tetranomial			
LS	.25	.50	.75	.4739	0.296	0.06676
χ^2	.25	.50	.75	.4754	0.286	0.06700
Mod. χ^2	.25	.50	.75	.4754	0.286	0.06700
H1	.25	.25	.75	.4318	0.338	0.06343
H2	.20	.25	.70	.4163	0.292	0.06763
			Sextinomial			
LS	.21	.25	.29	.2538	0.098	0.00565
χ^2	.22	.25	.30	.2757	0.078	0.01343
Mod. χ^2	.22	.26	.30	.2860	0.080	0.01734
H1	.23	.27	.31	.3005	0.070	0.02221
H2	.23	.27	.31	.3005	0.070	0.02221

It can be seen that the Hellinger estimator in the tetranomial method is a good alternative to the Maximum Likelihood estimator when the variables in the mixture model are normal. Tables 4.6 and 4.7 are in accordance with Everitt [10] and Hosmer [18] that mention that inferences with the MLE method may be not good for sample sizes ≤ 300 and $|M_1 - M_2| \leq 3\sigma$. Observe that in Tables 4.6 and 4.7 the MLE method shows its superiority. In the next section we will compare these estimators when the variables are not normal and we estimate the parameters in the mixture model with the Maximum Likelihood Method assuming normality.

4.6.1.2 Example for Normal Component Distributions

In this subsection we will analyze Professor Hoben Thomas's data, these data consist of 126 observations of his students' heights in inches presented in Table 4.8.

Table 4.8. Height data.

Females heights $\overline{x}_1 = 65.93$, $s_1^2 = 9.4198$									
55.00	60.00	60.25	61.00	61.75	62.25	62.25	62.63	62.75	63.00
63.25	63.25	63.25	63.25	63.38	64.00	64.25	64.25	64.50	64.75
64.75	65.00	65.00	65.13	65.13	65.17	65.25	65.25	65.25	65.25
65.25	65.50	65.75	66.00	66.00	66.25	66.25	66.25	66.50	66.75
66.75	66.75	66.75	67.00	67.13	67.25	67.38	67.50	67.50	67.75
67.75	67.75	68.13	68.75	69.00	69.00	69.25	69.50	69.88	70.00
70.00	70.25	70.38	71.00	71.00	71.25	71.75			
Males heights $\overline{x}_2 = 71.52$, $s_2^2 = 9.6098$									
65.50	65.75	66.00	66.25	66.75	67.00	67.13	67.25	68.00	68.13
68.75	68.75	69.13	69.25	69.25	69.38	69.75	70.00	70.13	70.50
70.75	70.75	71.00	71.00	71.50	71.50	71.63	71.75	71.75	72.00
72.00	72.00	72.00	72.00	72.00	72.25	72.25	72.25	72.38	72.50
72.50	72.88	73.00	73.00	73.00	73.25	73.25	73.63	74.00	74.63
75.00	75.00	75.05	76.00	77.00	77.13	77.50	77.50	78.50	

We will estimate the proportion of female students in this data set in which, of course, we will assume that we only know the students heights and that we do not know the sex associated with the height; the estimates are presented in Table 4.9.

Table 4.9. Estimates for Height data.

Method	$\widehat{\lambda}$	m_1	m_2	$\widehat{\sigma}_0^2$
EM	0.551	66.48	71.08	11.9222
Tetranomial				
LS	0.54	66.02	71.53	9.7456
χ^2	0.54	66.02	71.53	9.7456
Mod. χ^2	0.54	66.02	71.53	9.7456
H1	0.54	66.02	71.53	9.7456
H2	0.54	66.02	71.53	9.7456
Sextinomial				
LS	0.05	61.26	68.77	14.6080
χ^2	0.50	65.82	71.43	9.4190
Mod. χ^2	0.50	65.82	71.43	9.4190
H1	0.50	65.82	71.43	9.4190
H2	0.50	65.82	71.43	9.4190

The pooled variance s_0^2 is equal to 9.5087 and the sample means are separated $1.83\, s_0$. Observed that all tetranomial methods and all sextinomial methods with the exception of the least squares method give good estimates. The tetranomial methods give a closer estimate for $\widehat{\lambda}$, in this example the proportion of females is equal to 0.5317. The EM algorithm gives a good estimate of the mixing parameter but overestimate the common variance.

In order to construct the histogram of the medians and the covariance matrix for the estimates of the mixing parameter and the medians, we bootstrap the data. Results

for 100 bootstraps samples for female and male heights are presented in the Table 4.10. Figure 4.4 presents the histogram for the estimates of the male height medians with Hellinger 2 method.

Table 4.10. Properties of the medians for Height data

Method	Females		Males	
	Mean	Variance	Mean	Variance
EM	64.10	16.1159	71.23	2.8476
Tetranomial				
LS	64.82	2.5510	70.58	1.5364
χ^2	64.91	2.0104	70.63	1.4393
Mod. χ^2	64.91	2.0104	70.63	1.4393
H1	64.91	2.0104	70.63	1.4393
H2	64.91	2.0104	70.63	1.4393
Sextinomial				
LS	65.60	2.2675	71.06	1.0101
χ^2	65.80	1.2215	71.15	0.7419
Mod. χ^2	65.80	1.2214	71.15	0.7419
H1	65.80	1.2214	71.15	0.7419
H2	65.80	1.2214	71.15	0.7419

Bootstrapping the data permit us also to obtain an estimate of the covariance (correlation) matrix for the estimators of the mixing parameter and the medians of the populations. We found that the estimates of the mixing parameter and the estimates of the medians are very highly correlated for all methods. The covariance (correlation) matrix for the estimators of the mixing parameter and the medians of the heights of female and male students for the χ^2, Mod. χ^2 and Hellinger methods, denoted COV is:

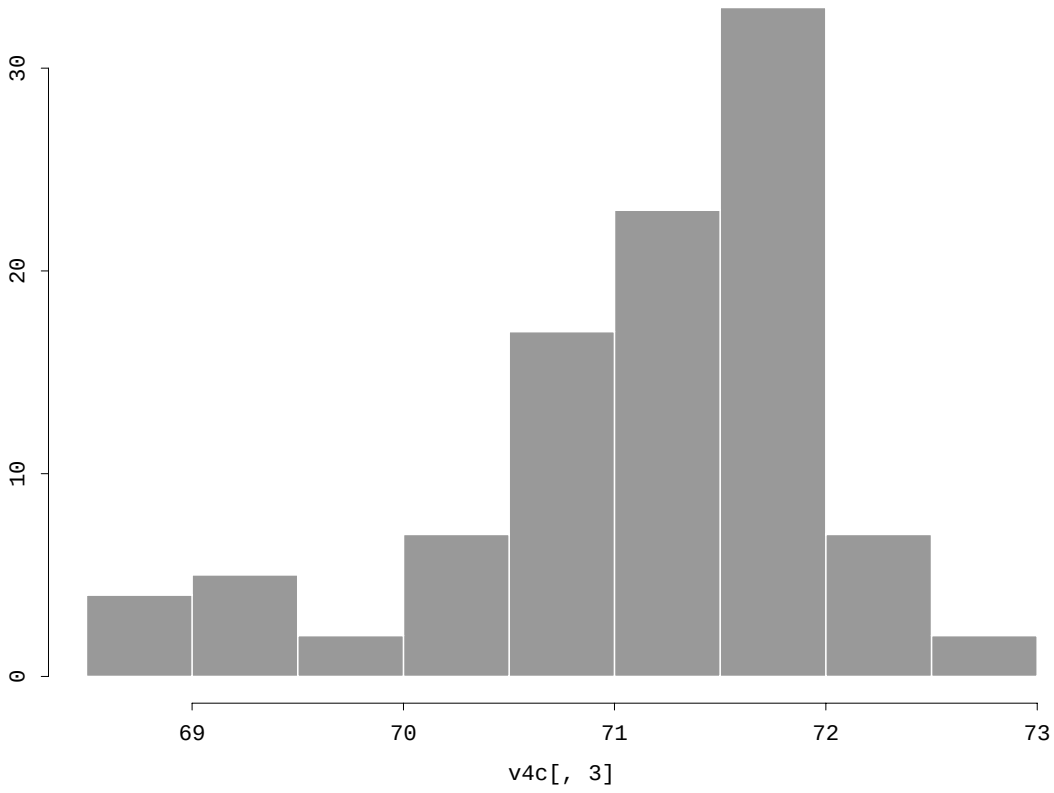

Fig. 4.4. Bootstrap histogram for the estimates of male height medians for Hellinger 2 method

$$COV = \begin{pmatrix} 0.02551\ (1) & 0.1403\ (.8011) & 0.1032\ (.7565) \\ . & 1.2214\ (1) & 0.6259\ (.6574) \\ . & . & 0.7419\ (1) \end{pmatrix}$$

This estimated covariance matrix for the estimators of the mixing parameter and the medians could be used for a rough test of the hypothesis: $H_0 : M_1 = M_2$ which can be interpreted as a test for only one component in the mixture model. For the Hellinger 2 method we have:

$$\begin{aligned} m_2 - m_1 &= 71.43 - 65.82 = 5.61 \\ \widehat{Var}(m_1 - m_2) &= \widehat{Var}(m_1) + \widehat{Var}(m_2) - 2\,\widehat{Cov}(m_1, m_2) \\ &= 1.2214 + 0.7419 - 2 \times 0.6259 = 0.7115 \end{aligned}$$

The medians are separated more than six standard deviations, therefore, we have strong support for the hypothesis that the population is a mixture of two densities. The same conclusion can be reach with the Boxplot of the estimates of the medians presented in Figure 4.5.

We have observed that when the medians of the populations are separated less than two standard deviations the tetranomial methods give better estimates than the sextinomial methods and when the separation between the medians is greater than two standard deviations the sextinomial methods are better. In order to show this statement we will modify this data set increasing the males' heights in order to have a separation

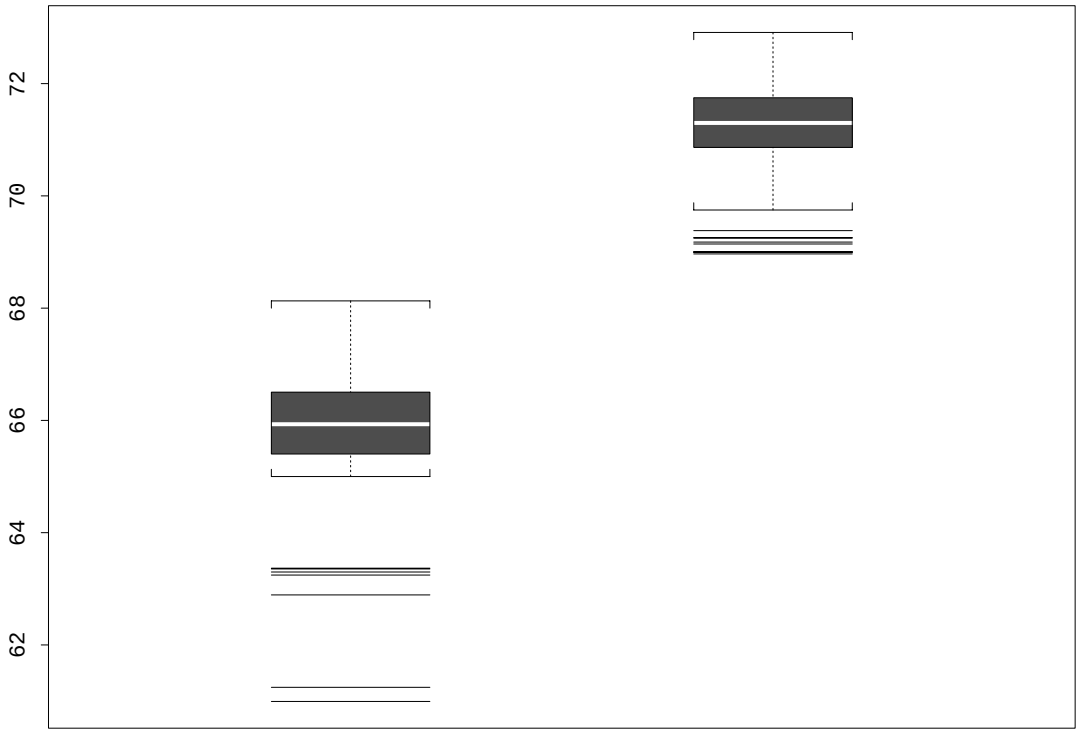

Fig. 4.5. Boxplot for the estimates of female and male height medians for Hellinger 2 method

of 2.32 standard deviations ($ov = 0.10$), in this way the females heights are the same but for the males heights $\overline{x}_2 = 73.08$, with the same variance $s_2^2 = 9.6098$. Table 4.11 shows that in this case the sextinomial methods are comparable with the EM algorithm but are better than the tetranomial methods.

Table 4.11. Estimates for the modified height data.

Method	$\widehat{\lambda}$	m_1	m_2	$\widehat{\sigma}_0^2$
EM	0.548	66.13	73.10	10.0698
Tetranomial				
LS	0.57	66.27	73.32	10.0969
χ^2	0.57	66.27	73.32	10.0969
Mod. χ^2	0.57	66.27	73.32	10.0969
H1	0.57	66.27	73.32	10.0969
H2	0.57	66.27	73.32	10.0969
Sextinomial				
LS	0.54	66.05	73.31	9.1670
χ^2	0.54	66.05	73.31	9.1670
Mod. χ^2	0.54	66.05	73.31	9.1670
H1	0.54	66.05	73.31	9.1670
H2	0.54	66.05	73.31	9.1670

4.6.2 Symmetric Component Distributions

In order to investigate the performance of our estimators when the component populations of the mixture are not normal we performed simulations with a sample size $n = 100$ for five distributions: The Cauchy distribution, t-Student distribution with 2, 4 and 10 d.f., and the Laplace distribution.

The results of the simulation study when we have a Cauchy distribution with an overlap of 0.10 ($M_1 = 0$ and $M_2 = 5.7$) in Table 4.12 shows that the least squares estimator is much better that the MLE estimator that assumes that the component populations are normal.

Table 4.12. Properties of $\widehat{\lambda}$ for Cauchy components, $\lambda = 0.25$, and ov=0.10

Method	LQ	Median	UQ	Mean	P	Var $(\widehat{\lambda})$
EM	.487	.498	.500	.4269		0.04432
			Tetranomial			
LS	.25	.60	.75	.4747	0.224	0.06997
χ^2	.25	.60	.75	.4819	0.208	0.06759
Mod. χ^2	.25	.60	.75	.4819	0.208	0.06759
H1	.20	.25	.625	.3542	0.322	0.05995
H2	.20	.25	.55	.3212	0.280	0.05541
			Sextinomial			
LS	.1675	.22	.27	.2106	0.048	0.00707
χ^2	.19	.23	.28	.2269	0.046	0.00664
Mod. χ^2	.19	.23	.28	.2278	0.050	0.00644
H1	.19	.23	.28	.2282	0.050	0.00641
H2	.19	.23	.28	.2282	0.050	0.00641

The results of the simulation study when we have a $t-$student distribution with 2 d.f. and an overlap of 0.10 ($M_1 = 0$ and $M_2 = 3.48$) are presented in Table 4.13

We can observe that in this case also the Hellinger 2 Method is much better than the EM algorithm that assumes that the populations are normal. We expect also that as the degrees of freedom in the $t-$Student distribution increases the maximum likelihood method will improve. The results of the simulation study when we have a $t-$Student distribution with 4 d.f. and an overlap of 0.10 ($M_1 = 0$ and $M_2 = 2.82$) are in Table 4.14

Table 4.13. Properties of $\widehat{\lambda}$ for t_2-student components, $\lambda = 0.25$, and ov=0.10

Method	LQ	Median	UQ	Mean	P	Var $(\widehat{\lambda})$
EM	.0264	.2992	.5043	.3646		0.11331
			Tetranomial			
LS	.25	.50	.75	.4636	0.254	0.06293
χ^2	.25	.50	.75	.468	0.234	0.06192
Mod. χ^2	.25	.50	.75	.468	0.234	0.06192
H1	.20	.25	.25	.2966	0.410	0.03915
H2	.20	.25	.25	.2605	0.380	0.03195
			Sextinomial			
LS	.18	.235	.29	.2362	0.058	0.00751
χ^2	.19	.24	.29	.2412	0.064	0.00839
Mod. χ^2	.19	.24	.29	.2410	0.060	0.00844
H1	.19	.24	.29	.2417	0.060	0.00839
H2	.19	.24	.29	.2417	0.060	0.00839

Table 4.14. Properties of $\widehat{\lambda}$ for t_4-student components, $\lambda = 0.25$, and ov=0.10

Method	LQ	Median	UQ	Mean	P	Var $(\widehat{\lambda})$
EM	.1653	.2535	.3667	.3145		0.06908
			Tetranomial			
LS	.25	.25	.62	.4023	0.406	0.04838
χ^2	.25	.30	.64	.4113	0.386	0.04878
Mod. χ^2	.25	.30	.64	.4113	0.386	0.04878
H1	.20	.25	.25	.2692	0.438	0.02845
H2	.20	.25	.25	.2350	0.378	0.02114
			Sextinomial			
LS	.18	.24	.30	.2465	0.034	0.01694
χ^2	.18	.24	.31	.2508	0.038	0.01195
Mod. χ^2	.18	.24	.31	.2508	0.036	0.01194
H1	.18	.24	.31	.2507	0.036	0.01194
H2	.18	.24	.31	.2507	0.036	0.01194

Table 4.15 presents the results of the simulation study when we have a $t-$student distribution with 10 d.f. and an overlap of 0.10 ($M_1 = 0$ and $M_2 = 2.5$). Observe now that the EM algorithm has improved and gives similar results as the sextinomial methods.

Table 4.15. Properties of $\widehat{\lambda}$ for t_{10}-student components, $\lambda = 0.25$, and ov=0.10

Method	LQ	Median	UQ	Mean	P	Var $(\widehat{\lambda})$
EM	.1964	.2587	.3233	.2681		0.01791
			Tetranomial			
LS	.25	.35	.60	.4084	0.394	0.04492
χ^2	.25	.35	.60	.4101	0.388	0.04439
Mod. χ^2	.25	.35	.60	.4101	0.388	0.04439
H1	.20	.25	.25	.2579	0.480	0.02106
H2	.20	.25	.25	.2307	0.406	0.01782
			Sextinomial			
LS	.15	.23	.31	.2407	0.036	0.01592
χ^2	.16	.24	.31	.2483	0.044	0.01721
Mod. χ^2	.16	.24	.31	.2490	0.044	0.01729
H1	.16	.24	.31	.2497	0.044	0.01746
H2	.16	.24	.31	.2497	0.044	0.01746

Table 4.16 presents the results of the simulation study when we have a double exponential distribution with an overlap of 0.10 ($M_1 = 0$ and $M_2 = 3.22$). Observed that the sextinomial methods outperform the EM algorithm.

For symmetric components we can conclude that the sextinomial methods outperform the EM algorithm unless the component distributions are close to the normal distribution (when the component distributions follow a t_{10}-student distribution the sextinomial methods and the EM method are comparable).

Table 4.16. Double exponential components, $\lambda = 0.25$, and ov=0.10

Method	LQ	Median	UQ	Mean	P	Var $(\widehat{\lambda})$
EM	.1986	.2538	3087	.2743		0.02938
			Tetranomial			
LS	.25	.40	.75	.4581	0.228	0.06486
χ^2	.25	.50	.75	.4651	0.264	0.06335
Mod. χ^2	.25	.50	.75	.4651	0.266	0.06335
H1	.20	.25	.30	.3205	0.414	0.04554
H2	.20	.25	.25	.2854	0.362	0.04184
			Sextinomial			
LS	.2075	.25	.30	.2455	0.080	0.00511
χ^2	.20	.25	.30	.2498	0.068	0.00649
Mod. χ^2	.21	.25	.30	.2501	0.070	0.00654
H1	.21	.25	.30	.2503	0.070	0.00658
H2	.21	.25	.30	.2503	0.070	0.00658

4.6.3 Asymmetric Component Distributions

In order to study the robustness of our estimators when the component distributions are not symmetric we performed simulations when the component distributions are similar in shape with the χ^2 distribution, but in such a way that the smallest median is equal to zero and the largest is equal to 2 or 2.5 standard deviations to the right. We performed this Monte Carlo study considering 10, 20, 30 and 40 degrees of freedom.

Table 4.17 presents the results when $f_1(x + 9.3418)$ and $f_2(x + .3975)$ are χ^2_{10} distributions (χ^2 distributions with 10 *df*) and the mixing parameter $\lambda = 0.25$. The medians of these distributions are 2 standard deviations apart.

Table 4.18 presents the results when $f_1(x + 9.3418)$ and $f_2(x - 1.8384)$ are χ^2_{10} distributions (χ^2 distributions with 10 *df*) and the mixing parameter $\lambda = 0.25$. The medians of these distributions are 2.5 standard deviations apart. We observe that for a

Table 4.17. Properties of $\widehat{\lambda}$ for χ^2_{10} components, $\lambda = 0.25$, and $\delta = 2\,\sigma$

Method	LQ	Median	UQ	Mean	P	Var $(\widehat{\lambda})$
EM	.2340	.4027	.9022	.5158		0.10302
			Tetranomial			
LS	.25	.25	.50	.3762	0.494	0.02657
χ^2	.25	.25	.50	.3776	0.490	0.02663
Mod. χ^2	.25	.35	.50	.3776	0.490	0.02663
H1	.20	.25	.25	.2527	0.458	0.01788
H2	.15	.25	.25	.2183	0.384	0.01465
			Sextinomial			
LS	.11	.23	.60	.3376	0.006	0.07034
χ^2	.14	.275	.64	.3639	0.008	0.06923
Mod. χ^2	.14	.28	.64	.3659	0.010	0.06876
H1	.14	.28	.64	.3659	0.010	0.06876
H2	.14	.28	.64	.3659	0.010	0.06876

χ^2_{10} Hellinger 1 and Hellinger 2 tetranomial methods give good estimates of the mixing parameter, When $\delta = 2\,\sigma$ the sextinomial methods are better than the EM algorithm that assumes that the component distributions are normal.

Table 4.19 presents the results when $f_1(x + 19.3374)$ and $f_2(x + 6.6883)$ are χ^2_{20} distributions and the mixing parameter $\lambda = 0.25$. The medians of these distributions are 2 standard deviations apart.

Table 4.20 presents the results when $f_1(x + 19.3374)$ and $f_2(x + 3.526)$ are χ^2_{20} distributions and the mixing parameter $\lambda = 0.25$. The medians of these distributions are 2.5 standard deviations apart. When the component distributions are χ^2_{20}, we observe a similar pattern as in the case mentioned before in the sense that when the medians are close, Hellinger 1 and Hellinger 2 tetranomial methods give good estimates of the mixing

Table 4.18. Properties of $\widehat{\lambda}$ for χ^2_{10} components, $\lambda = 0.25$, and $\delta = 2.5\,\sigma$

Method	LQ	Median	UQ	Mean	P	Var $(\widehat{\lambda})$
EM	.1885	.2378	.3110	.3130		0.05011
			Tetranomial			
LS	.25	.46	.60	.4500	0.328	0.03419
χ^2	.25	.46	.60	.4511	0.326	0.03419
Mod. χ^2	.25	.46	.60	.4511	0.326	0.03419
H1	.20	.25	.40	.3064	0.398	0.03385
H2	.1875	.25	.30	.2634	0.362	0.02801
			Sextinomial			
LS	.09	.19	.30	.2659	0.016	0.05518
χ^2	.11	.21	.485	.3103	0.014	0.06399
Mod. χ^2	.11	.22	.53	.3159	0.018	0.06512
H1	.12	.22	.55	.3225	0.016	0.06657
H2	.12	.22	.55	.3225	0.016	0.06657

Table 4.19. Properties of $\widehat{\lambda}$ for χ^2_{20} components, $\lambda = 0.25$, and $\delta = 2\,\sigma$

Method	LQ	Median	UQ	Mean	P	Var $(\widehat{\lambda})$
EM	.2114	.3166	.5229	.4241		0.08011
			Tetranomial			
LS	.25	.35	.50	.3880	0.460	0.02979
χ^2	.25	.35	.50	.3881	0.458	0.02977
Mod. χ^2	.25	.35	.50	.3881	0.458	0.02977
H1	.20	.25	.25	.2489	0.504	0.01498
H2	.20	.25	.25	.2184	0.446	0.01217
			Sextinomial			
LS	.10	.225	.50	.3102	0.016	0.06020
χ^2	.13	.24	.535	.3294	0.018	0.06078
Mod. χ^2	.13	.24	.55	.3308	0.018	0.06093
H1	.13	.24	.55	.3306	0.018	0.06088
H2	.13	.24	.55	.3306	0.018	0.06088

parameter. When $\delta = 2.5\,\sigma$ the EM algorithm that assumes normality and the sextinomial methods give good estimates of the mixing parameter.

Table 4.20. Properties of $\widehat{\lambda}$ for χ^2_{20} components, $\lambda = 0.25$, and $\delta = 2.5\,\sigma$

Method	LQ	Median	UQ	Mean	P	Var $(\widehat{\lambda})$
EM	.1863	.2272	.2850	.2615		0.02190
			Tetranomial			
LS	.25	.48	.60	.4599	0.270	0.03642
χ^2	.25	.50	.605	.4651	0.264	0.03598
Mod. χ^2	.25	.50	.605	.4648	0.266	0.03607
H1	.20	.25	.40	.3036	0.364	0.03476
H2	.15	.25	.2625	.2573	0.348	0.02713
			Sextinomial			
LS	.11	.18	.29	.2438	0.012	0.03943
χ^2	.12	.20	.3125	.2662	0.022	0.04445
Mod. χ^2	.12	.20	.32	.2669	0.026	0.04466
H1	.12	.20	.32	.2685	0.026	0.04511
H2	.12	.20	.32	.2685	0.026	0.04511

Table 4.21 presents the results when $f_1(x + 29.3360)$ and $f_2(x + 13.8441)$ are χ^2_{30} distributions and the mixing parameter $\lambda = 0.25$. The medians of these distributions are 2 standard deviations apart. When the $d.f = 30$, the χ^2 distribution is close to the normal distribution, but if $\delta = 2\,\sigma$ the EM algorithm and the sextinomial methods give unreliable estimates of the mixing parameter (Observe that again Hellinger 1 and Hellinger 2 tetranomial methods give good estimates of the mixing parameter).

Table 4.22 presents the results when $f_1(x + 29.3360)$ and $f_2(x + 6.0981)$ are χ^2_{30} distributions and the mixing parameter $\lambda = 0.25$. In this case, when $\delta = 2.5\,\sigma$ (the medians of these distributions are 2.5 standard deviations apart), the EM algorithm and

Table 4.21. Properties of $\widehat{\lambda}$ for χ^2_{30} components, $\lambda = 0.25$, and $\delta = 2\,\sigma$

Method	LQ	Median	UQ	Mean	P	Var $(\widehat{\lambda})$
EM	.2045	.3128	.4974	.3948		0.06695
			Tetranomial			
LS	.25	.25	.46	.3514	0.512	0.02760
χ^2	.25	.25	.48	.3547	0.506	0.02814
Mod. χ^2	.25	.35	.48	.3547	0.506	0.02814
H1	.20	.25	.25	.2329	0.488	0.01102
H2	.15	.20	.25	.2061	0.400	0.00976
			Sextinomial			
LS	.14	.25	.5000	.3328	0.018	0.05866
χ^2	.16	.27	.5625	.3469	0.022	0.05734
Mod. χ^2	.16	.27	.5625	.3469	0.022	0.05739
H1	.16	.27	.5625	.3469	0.022	0.05739
H2	.16	.27	.5625	.3467	0.022	0.05735

the sextinomial methods give similar estimates of the mixing parameter but the variance of the sextinomial methods is twice the variance of the EM estimate.

For χ2-type asymmetric components, we can conclude that the sextinomial methods are slightly better than the EM algorithm when the medians are two or 2.5 standard deviations apart. This was not the case for χ^2_{30} component distributions and medians 2.5 standard deviations apart; however, in this case the component distributions are close to the normal distribution.

4.7 Training Samples

In this section we will consider the estimation of the mixing parameter when training samples are available. We will distinguish between the three types of sampling considered by Hosmer [18]. The first case referred as model $M0$ is when there are only data

Table 4.22. Properties of $\widehat{\lambda}$ for χ^2_{30} components, $\lambda = 0.25$, and $\delta = 2.5\,\sigma$

Method	LQ	Median	UQ	Mean	P	Var $(\widehat{\lambda})$
EM	.1881	.2304	.2904	.2599		0.01939
			Tetranomial			
LS	.25	.40	.60	.4348	0.360	0.03537
χ^2	.25	.40	.60	.4372	0.352	0.03518
Mod. χ^2	.25	.40	.60	.4372	0.352	0.03518
H1	.20	.25	.25	.2671	0.463	0.02427
H2	.10	.25	.25	.2258	0.386	0.01899
			Sextinomial			
LS	.12	.1950	.2925	.2512	0.024	0.03773
χ^2	.14	.2200	.3125	.2790	0.026	0.04327
Mod. χ^2	.14	.2200	.3200	.2794	0.026	0.04327
H1	.14	.2200	.3200	.2794	0.026	0.04335
H2	14	.2200	.3125	.2790	0.028	0.04324

from the mixed distribution and no training sample is available we have considered this case already.

A sample where the component of origin of each observation is known with certainty will be called known data or training sample. Two types of training samples are possible according to whether or not the known data contains information about the mixing proportion. A sample which contains both mixed and known data but where the known data contains no information about the mixing proportion will be called model $M1$. An example of this model is when we arbitrarily choose a fixed number or elements from each component population. Finally model $M2$ will refer to the case when the sample contains both mixed and known data, and information about the mixing proportion is contained in the relative number of observations from the two component populations in the known data. An example of $M2$ is given by Hosmer [18], in which the purpose of the study

is to estimate the proportion of male fish using the variable fish length for classification purposes. Suppose that a random sample of 100 fish is selected and, from this sample, we select at random a subset of 20 fish, that are classified into males and females. The classified subset of 20 fish constitutes the training sample.

Our grid search program can be easily modified in order to use the information given by training samples. For the $M1$ method, as a first step, for each combination of λ and p_i's we obtain the cut points and the regions of the multinomial approach. In the second step, we count the number of observations for the unknown sample and for the known sample for each component, next we obtain the expected number of observations per region and finally we evaluate the criterion for selecting the estimators (observe that in this description, we are not using the possible information about the mixing proportion contained in the relative number of observations from the two component populations in the known data).

For the $M2$ method in the first phase, we bootstrap the sample and we apply the steps given for the $M1$ method in order to estimate the variance of the mixing parameter. In the second phase, we obtain a weighted estimate of the mixing parameter with the $M1$ estimate and the estimate given by the relative number of observations from the two component populations.

We perform 500 simulations with the Tetranomial and Sextinomial distribution when the total sample size is equal to 100, the mixing parameter $\lambda = 0.25$, $M_1 = 0$, $M_2 = 2.32$, and the percentage of known data is 20% (the training sample size is 20).

For the $M1$ method we assumed that 10 observation from each population were known. Table 4.23 presents the results. The variance of the EM method is about 50% of

the variance of the multinomial approaches. The variance of the mixing parameter for the EM algorithm decreased 25% under this sampling method.

Table 4.23. Method M1 training sample size equal to 20, $\lambda = 0.25$, and ov=0.10

Method	LQ	Median	UQ	Mean	Var $(\widehat{\lambda})$
EM	.1973	.2499	.3028	.2525	0.00764
			Tetranomial		
LS	.15	.23	.3125	.2406	0.01689
χ^2	.15	.23	.32	.2422	0.01715
Mod. χ^2	.15	.23	.32	.2422	0.01777
H1	.15	.23	.32	.2422	0.01777
H2	.15	.23	.32	.2422	0.01777
			Sextinomial		
LS	.14	.225	.30	.2325	0.01390
χ^2	.15	.23	.30	.2348	0.01484
Mod. χ^2	.15	.23	.30	.2359	0.01491
H1	.15	.23	.30	.2359	0.01491
H2	.15	.23	.30	.2359	0.01491

For the $M2$ method we assumed that a subsample of 20 observations was classified into the two component populations, therefore an unbiased estimator of the mixing parameter is available from this subsample. Note that we have two estimators, one of them is this unbiased estimator and the second one is the $M1$ estimator given by the 80 unclassified observations and the known sample without using the information that this known sample has about the mixing parameter. The final $M2$ estimator will be a weighted average of the above mentioned estimators. As the weights that minimized the variance of the mixing parameter are the variances of these individuals estimators we obtain two tables. Table 4.24 presents the estimator of the mixing parameter with the optimal weight, the

variance of the $M1$ estimator was obtained by an additional simulation with 100 repetitions and the variance of the known sample with the well known formula for the binomial distribution assuming that we know the value λ.

Table 4.24. Method M2 training sample equal to 20 with known weights, $\lambda = 0.25$, and ov=0.10

Method	LQ	Median	UQ	Mean	Var $(\widehat{\lambda})$
EM	.2099	.2498	.2955	.2535	0.00422
			Tetranomial		
LS	.2108	.2606	.3144	.2627	0.00641
χ^2	.2141	.2608	.3177	.2645	0.00629
Mod. χ^2	.2141	.2608	.3173	.2643	0.00628
H1	.2141	.2608	.3173	.2643	0.00628
H2	.2141	.2608	.3173	.2643	0.00628
			Sextinomial		
LS	.2079	.2566	.3013	.2587	0.00592
χ^2	.2092	.2579	.3039	.2619	0.00584
Mod. χ^2	.2092	.2592	.3032	.2624	0.00577
H1	.2092	.2592	.3032	.2623	0.00579
H2	.2092	.2592	.3032	.2623	0.00579

We note that optimal weights are not possible to compute in practice because we do not know the mixing parameter λ or the variances of the two estimators. Table 4.25 presents the results when we estimate the optimal weight of the binomial for each simulation.

For these cases the multinomial methods are similar and the variance of the estimator of the mixing parameter is about twice the variance of the estimator given by the EM method. We observe also that the minimum variance that the estimator of the mixing parameter can attain is 0.00187 and the EM method has an efficiency of 42%.

Table 4.25. Method M2 training sample equal to 20 with unknown weights, $\lambda = 0.25$, and ov=0.10

Method	LQ	Median	UQ	Mean	Var $(\widehat{\lambda})$
EM	.2009	.2446	.2874	.2453	0.00438
			Tetranomial		
LS	.1999	.2590	.3159	.2590	0.00702
χ^2	.2016	.2642	.3142	.2605	0.00674
Mod. χ^2	.2016	.2642	.3137	.2606	0.00671
H1	.2016	.2642	.3137	.2606	0.00671
H2	.2016	.2642	.3137	.2606	0.00671
			Sextinomial		
LS	.1964	.2484	.2955	.2517	0.00659
χ^2	.1994	.2500	.2996	.2550	0.00662
Mod. χ^2	.1997	.2519	.2996	.2558	0.00656
H1	.1997	.2519	.2996	.2557	0.00658
H2	.1997	.2519	.2996	.2557	0.00658

As we mentioned before in order to estimate the weights for the $M2$ method we perform 100 simulations from the mixture, but in a real life problem this is not possible because we have one data set only, nevertheless we can bootstrap.

In order to investigate how accurate is the estimate of the variance of the estimator of the mixing parameter when we bootstrap, we performed 100 simulations. In order that the particular sample from which we bootstrap will not affect the results, we use a "perfect sample" given by the following formula which gives the normal scores accurate to four decimal places; see Hettmansperger (p. 145, [15]).

$$Z_i = 4.91\,[\ p^{0.14} - (1-p)^{0.14}\ ]$$

where

$$p = \frac{i - 3/8}{n + 1/4}$$

The variances for the estimator of the mixing parameter for the $M0$ method when the sample size n is equal to 100 are presented in Table 4.26 (For the random sample the variance is equal to 0.01014, see table 4.2).

Table 4.26. Properties of $\widehat{\lambda}$ for normal components; $M0$ method with $\lambda = 0.25$, and ov=0.10

Method	Random Sample		Perfect Sample	
	Mean	Variance	Mean	Variance
EM	.2746	0.01014	.2627	0.00807
		Tetranomial		
LS	.3972	0.03983	.4785	0.04704
χ^2	.3982	0.03982	.4799	0.04652
Mod. χ^2	.3982	0.03982	.4799	0.04652
H1	.3013	0.02223	.4078	0.04723
H2	.2870	0.01761	.3307	0.03986
		Sextinomial		
LS	.2612	0.02371	.2621	0.02944
χ^2	.2732	0.02488	.2712	0.02989
Mod. χ^2	.2732	0.02480	.2679	0.02957
H1	.2732	0.02480	.2668	0.02862
H2	.2732	0.02480	.2668	0.02720

In this table we can observe that for an overlap of 0.10, the variance obtained by bootstrapping the perfect sample is 75% of the variance obtained with random samples (Table 4.2). For the multinomial approaches we get similar results. This fact is very

important because it tells us that we can obtain a precise estimator of the variance of the mixing parameter bootstrapping the data set.

The variances for the estimator of the mixing parameter for the $M1$ method are presented in Table 4.27. It is interesting to note that in this table we present the results when we sample (sample size equal to 100) from a perfect population of 100 observations (25 observations are from population 1 and 75 observations from population 2). If we sample (sample size equal to 100) from a perfect population of 130 observations (30 observations from population 1 and 90 observations from population 2) the variance of the mixing estimator is closer to the variance of the mixing estimator obtained from the random sample scheme.

Table 4.27. Properties of $\widehat{\lambda}$ for normal components; $M1$ method with $\lambda = 0.25$, and ov=0.10

Method	Random Sample		Perfect Sample	
	Mean	Variance	Mean	Variance
EM	.2525	0.00764	.2514	0.00601
		Tetranomial		
LS	.2406	0.01689	.2198	0.01189
χ^2	.2422	0.01715	.2159	0.01186
Mod. χ^2	.2422	0.01777	.2159	0.01194
H1	.2422	0.01777	.2159	0.01194
H2	.2422	0.01777	.2159	0.01194
		Sextinomial		
LS	.2325	0.01390	.2081	0.01059
χ^2	.2348	0.01484	.2167	0.01067
Mod. χ^2	.2359	0.01491	.2145	0.00957
H1	.2359	0.01491	.2145	0.00957
H2	.2359	0.01491	.2145	0.00957

The variances for the estimator of the mixing parameter for the $M2$ method before the weighted average (before we get the final estimator) are presented in Table 4.28. In this table we can observe that a precise estimator of the variance of the mixing parameter can be obtained by bootstrapping the sample.

Table 4.28. Properties of $\widehat{\lambda}$ for normal components before weights; $M2$ method with $\lambda = 0.25$, and ov=0.10

Method	Random Sample		Perfect Sample	
	Mean	Variance	Mean	Variance
EM	.2519	0.00701	.2532	0.00705
		Tetranomial		
LS	.3126	0.02800	.2669	0.02731
χ^2	.3191	0.02629	.2696	0.02555
Mod. χ^2	.3196	0.02608	.2696	0.02555
H1	.3196	0.02608	.2696	0.02555
H2	.3170	0.02604	.2696	0.02555
		Sextinomial		
LS	.2720	0.02462	.2624	0.02059
χ^2	.2951	0.02538	.2590	0.02084
Mod. χ^2	.2951	0.02538	.2590	0.02084
H1	.2951	0.02538	.2590	0.02084
H2	.2951	0.02538	.2590	0.02084

Table 4.29 presents the estimates of the mixing parameter and the variance of the estimator of the mixing parameter for the final estimator (the weighted estimator) for random samples (Table 4.25) and for the "perfect sample", we note that the results are similar.

Table 4.29. Properties of the final estimator $\widehat{\lambda}$ for normal components; $M2$ method with $\lambda = 0.25$, and ov=0.10

Method	Random Sample		Perfect Sample	
	Mean	Variance	Mean	Variance
EM	.2453	0.00438	.2465	0.00403
		Tetranomial		
LS	.2590	0.00702	.2576	0.00563
χ^2	.2605	0.00674	.2593	0.00557
Mod. χ^2	.2606	0.00671	.2593	0.00557
H1	.2606	0.00671	.2593	0.00557
H2	.2606	0.00671	.2593	0.00557
		Sextinomial		
LS	.2517	0.00659	.2527	0.00512
χ^2	.2550	0.00662	.2544	0.00493
Mod. χ^2	.2558	0.00656	.2545	0.00497
H1	.2557	0.00658	.2545	0.00497
H2	.2557	0.00658	.2545	0.00497

4.7.1 Example for Normal Components

We will analyze the Height Data. In Table 4.30 we have the results for the M0 method for sample sizes equal to 100 and 100 bootstraps. Observe that the sextinomial method overperformed both the tetranomial approaches and the EM algorithm (whose results are not precise because the means of the component populations are less than two standard deviations apart).

Table 4.31 present the results for the M1 method, observe the improvement of the estimator for the EM algorithm when we have a training samples. This is in accordance with Hosmer [18], who mentions that it is possible to obtain a precise estimator of the mixing parameter when the means of the component distributions are close, even with a small sample size if training samples from both component populations are available.

Table 4.30. Method M0 for height data

Method	LQ	Median	UQ	Mean	Variance
EM	.0123	.5221	.5617	.3817	0.06828
		Tetranomial			
LS	.1400	.4850	.5200	.3775	0.04222
χ^2	.2400	.5000	.5200	.4057	0.03613
Mod. χ^2	.2400	.5000	.5200	.4057	0.03613
H1	.2300	.4950	.5200	.3987	0.03575
H2	.1550	.4700	.5200	.3682	0.03829
		Sextinomial			
LS	.4800	.5200	.5525	.5092	0.01169
χ^2	.4900	.5200	.5600	.5248	0.00560
Mod. χ^2	.4900	.5200	.5600	.5248	0.00560
H1	.4900	.5200	.5600	.5248	0.00560
H2	.4900	.5200	.5600	.5248	0.00560

Table 4.31. Method M1 for height data

Method	LQ	Median	UQ	Mean	Variance
EM	.4789	.5378	.6031	.5411	0.01118
		Tetranomial			
LS	.4200	.5150	.6200	.5044	0.02609
χ^2	.4200	.5050	.6125	.5000	0.02562
Mod. χ^2	.4200	.5050	.6125	.4999	0.02563
H1	.4200	.5050	.6125	.4999	0.02563
H2	.4200	.5050	.6125	.4999	0.02563
		Sextinomial			
LS	.4800	.5300	.5900	.5318	0.01139
χ^2	.4775	.5250	.5800	.5279	0.01002
Mod. χ^2	.4775	.5250	.5800	.5279	0.01007
H1	.4775	.5250	.5800	.5279	0.01007
H2	.4775	.5250	.5800	.5279	0.01007

Table 4.32 shows the results for the M2 method. In this method, we bootstrap the combined sample in order to obtain an estimator of the variance of the estimator of the mixing parameter. This variance was used in order to obtain a weighted estimator of the mixing parameter, whose components are the estimators given by the training sample and the nonclassified sample. For this case, the tetranomial estimators are comparable with the EM algorithm estimator that assumes normal components, and both the EM algorithm estimator and the tetranomial estimators are outperformed by the sextinomial estimators.

Table 4.32. Method M2 for height data

Method	LQ	Median	UQ	Mean	Variance
EM	.4792	.5314	.5807	.5281	0.00581
		Tetranomial			
LS	.4886	.5309	.5748	.5270	0.00589
χ^2	.4886	.5309	.5748	.5270	0.00589
Mod. χ^2	.4886	.5309	.5748	.5270	0.00589
H1	.4716	.5281	.5748	.5263	0.00593
H2	.4786	.5309	.5748	.5270	0.00589
		Sextinomial			
LS	.4900	.5301	.5786	.5342	0.00350
χ^2	.4900	.5274	.5785	.5334	0.00349
Mod. χ^2	.4900	.5274	.5807	.5336	0.00351
H1	.4900	.5274	.5807	.5336	0.00351
H2	.4900	.5274	.5807	.5336	0.00351

4.7.2 Example with Non-symmetric Components

We will analyze the Fish data given by Hosmer [18], in which a set of 209 observations is taken as the population (in this set, the true proportion of males is 35.40%). The purpose of the study is to estimate the true proportion of male fish with samples of size equal to 100. Results for the M0 method are presented in Table 4.33. Observe that the EM algorithm, and the tetranomial and sextinomial methods underestimate the true proportion of males. The reason for this fact lies in that the component populations not only do not seem to be normal distributions but seem to be asymmetric distributions. Observe also that the estimates from the sextinomial methods seem to be closer to the true proportion.

Table 4.33. Method M0 for Fish data

Method	LQ	Median	UQ	Mean	Variance
EM	.1884	.2082	.2489	.2136	0.00149
Tetranomial					
LS	.1500	.2050	.2700	.2309	0.01946
χ^2	.1500	.2000	.2700	.2305	0.01948
Mod. χ^2	.1500	.2000	.2700	.2305	0.01948
H1	.1500	.2000	.2700	.2305	0.01948
H2	.1500	.2000	.2700	.2305	0.01948
Sextinomial					
LS	.1600	.2200	.2725	.2604	0.03019
χ^2	.1600	.2350	.3000	.2871	0.04232
Mod. χ^2	.1600	.2350	.3000	.2871	0.04232
H1	.1600	.2350	.3000	.2871	0.04232
H2	.1600	.2350	.3000	.2871	0.04232

The M1 methods gives the results presented in Table 4.34. Observe that the EM algorithm estimator has improved and outperformed the tetranomial methods. The variance of the EM estimator is almost 30% of the variance of the sextinomial. The EM estimator and the sextinomial estimators underestimate the true proportion of male fish.

Table 4.34. Method M1 for Fish data

Method	LQ	Median	UQ	Mean	Variance
EM	.2246	.2636	.3003	.2703	0.00878
			Tetranomial		
LS	.1100	.1850	.2600	.2045	0.01851
χ^2	.1100	.1700	.2500	.2021	0.01810
Mod. χ^2	.1100	.1700	.2500	.2021	0.01810
H1	.1100	.1700	.2500	.2016	0.01810
H2	.1100	.1700	.2500	.2016	0.01810
			Sextinomial		
LS	.1000	.1800	.2600	.2197	0.02928
χ^2	.1500	.2000	.2625	.2420	0.02773
Mod. χ^2	.1500	.2000	.2775	.2468	0.02781
H1	.1575	.2000	.3200	.2540	0.02755
H2	.1575	.2000	.3200	.2540	0.02755

For the M2 method whose results are presented in Table 4.35, the EM algorithm (which assumes that the component populations are normal with equal variances) still underestimates the true proportion of male fish, and it is comparable with the tetranomial estimators. For the sextinomial methods the estimator has improved very much but still their variances are three times the variance of the EM estimator. Nevertheless if we base our comparisons on the mean quadratic error the sextinomial estimators are

preferable. Observe also that the sextinomial estimators are comparable with the EM algorithm estimator that assumes that the component populations have different variances.

Table 4.35. Method M2 for Fish data

Method	LQ	Median	UQ	Mean	Variance
EM	.2522	.2831	.3087	.2843	0.00249
EM[1]	.2931	.3467	.3887	.3472	0.00544
		Tetranomial			
LS	.2530	.2904	.3322	.3023	0.00532
χ^2	.2538	.2904	.3359	.3062	0.00604
Mod. χ^2	.2538	.2904	.3359	.3062	0.00604
H1	.2538	.2904	.3359	.3062	0.00604
H2	.2538	.2904	.3359	.3062	0.00604
		Sextinomial			
LS	.2828	.3329	.3779	.3347	0.00696
χ^2	.2847	.3362	.3912	.3401	0.00722
Mod. χ^2	.2847	.3371	.3967	.3417	0.00726
H1	.2847	.3379	.4081	.3430	0.00743
H2	.2847	.3379	.4081	.3430	0.00743

4.8 Comparison of Approaches

4.8.1 Tetranomial versus Sextinomial approach

We have mentioned before that the tetranomial approaches seem to give better estimates when the separation of the population medians is less than two standard deviations. In order to give a final recommendation we obtained additional simulations from normal component populations for a sample size equal to 100, mixing parameter $\lambda = 0.75$, and

[1]EM estimator that assumes that the component populations have different variances

overlap=0.10. The results in Table 4.36 show that the tetranomial methods fail in this case. A possible explanation is that the tetranomial approach does not check the symmetry of the distributions as the sextinomial approach does. Similar results were obtained for an overlap equal to 0.15. For this reason we recommend the sextinomial approach over the tetranomial approach. It is interesting to note that when training samples are available we do not have this problem and for sampling method $M2$ the results of the tetranomial and the sextinomial approaches are comparable. Table 4.37 presents the results when the training sample is equal to 20 and $\lambda = 0.75$.

Table 4.36. Properties of $\widehat{\lambda}$ for normal components, n=100, $\lambda = 0.75$, and $ov = 0.10$

Method	LQ	Median	UQ	Mean	Var($\widehat{\lambda}$)
EM	.6803	.7418	.8053	.7389	0.00931
Tetranomial					
LS	.25	.25	.60	.4115	0.04533
χ^2	.25	.25	.60	.4150	0.04644
Mod. χ^2	.25	.25	.60	.4140	0.04626
H1	.25	.25	.50	.3822	0.04276
H2	.20	.25	.4125	.3215	0.03615
Sextinomial					
LS	.66	.75	.8325	.7348	0.01981
χ^2	.66	.75	.8225	.7278	0.02116
Mod. χ^2	.66	.75	.8225	.7278	0.02116
H1	.66	.75	.8225	.7274	0.02116
H2	.66	.75	.8225	.7274	0.02116

Table 4.37. Properties of the final estimator $\widehat{\lambda}$ for normal components; $M2$ method, $\lambda = 0.75$, and ov=0.10

Method	LQ	Median	UQ	Mean	Var($\widehat{\lambda}$)
EM	.7036	.7504	.7994	.7498	0.00510
Tetranomial					
LS	.6897	.7375	.7864	.7351	0.00519
χ^2	.6893	.7363	.7864	.7341	0.00548
Mod. χ^2	.6897	.7375	.7864	.7345	0.00547
H1	.6897	.7375	.7864	.7345	0.00547
H2	.6897	.7375	.7864	.7345	0.00547
Sextinomial					
LS	.6804	.7415	.7908	.7316	0.00656
χ^2	.6823	.7415	.7936	.7333	0.00641
Mod. χ^2	.6860	.7415	.7941	.7333	0.00642
H1	.6860	.7415	.7941	.7333	0.00642
H2	.6860	.7415	.7941	.7333	0.00642

4.8.2 Hellinger 1 vs Hellinger 2

Subsection 4.4.5 shows that Hellinger 1 and Hellinger 2 methods should give the same results because the minimization of

$$\Sigma_{k=1}^{c} \left[\sqrt[2]{n_k/n} - \sqrt[2]{E(n_k)/n} \right]^2$$

is equivalent to the maximization of

$$\Sigma_{k=1}^{c} \sqrt[2]{n_k E(n_k)/n^2}$$

or to the minimization of

$$cos^{-1} \left[\Sigma_{k=1}^{c} \sqrt[2]{n_k E(n_k)/n^2} \right]$$

because cos^{-1} is a decreasing function. Nevertheless we observe that these two methods give different results in many cases (Tables 4.1, 4.2, 4.3 etc.). We found that these discrepancies are due to the approximation of the cos^{-1} function in Splus. We noted that for example cos^{-1}[1-(1e-10)]=4.470348e-08 but cos^{-1}[1-(1e-20)]=0. This means that with Hellinger 1 method the quantities 0 and $1e-25$ are different, but with Hellinger 2 method, the quantities $cos^{-1}(1)$ and cos^{-1}[1-(1e-25)] are equal. Therefore with the Hellinger 2 method we can have more zeroes of the objective function, as the minimum in our grid search is chosen from small to large values of λ. Method 2 can select a smaller value of λ than Method 1. This explains why the mean value of the estimates of the mixing parameter for Method 1 in our Monte Carlo studies is less than or equal to the mean value of the estimates of the mixing parameter for Method 2. As the results of Hellinger 2 method are different from Hellinger 1 method by rounding in Splus we recommend the Hellinger 2 method for estimating the mixing parameter in a mixture of two symmetric component distributions.

4.9 Summary

In this chapter we propose a nonparametric approach for the estimation of the mixing parameter when we have univariate responses. The component distributions are assumed to be continuous with equal shapes and unimodal symmetric densities which belong to the same location family with medians M_1 and M_2. We analyzed two multinomial approaches inserting cut points in the continuous univariate responses. In this way we define a tetranomial and a sextinomial distribution. The estimation procedures compare the distance between the observed and the expected number of observations per class (defined by the cut points mentioned before) of the multinomial distribution. We analyzed five

procedures: Least Squares, Chi Square, Modified Chi Square and two Hellinger distances. We present, in Appendix A, the program codes in Mathematica and Splus, for a grid search for the estimators in the mixture model.

A Monte Carlo study showed that for normal components, the estimators of the mixing proportion in the multinomial approaches are comparable with the EM algorithm estimator if the medians are 1.75 standard deviations apart. The efficiency of the EM algorithm increases when the separation of the medians increases. If the medians are 2.32 standard deviations apart the variance of the sextinomial estimators are two times the variance of the EM estimator but if the separation is three standard deviations the variance of the sextinomial estimators are four times the variance of the EM estimator.

We analyzed the data provided by professor Hoben Thomas in which the goal is to estimate the proportion of female and male students with the information given by their heights. In this example the Chi Square, Modified Chi Square and two Hellinger distances in the the sextinomial approach gave a very accurate estimation of the mixing parameter.

If the component populations are symmetric but not normal, for example the component populations follow Double Exponential, Cauchy , or t-student distributions with 2, 4 or 10 degrees of freedom, the sextinomial estimators outperformed the EM estimator that assumes normality. If the component populations are not symmetric (for example a χ^2 type distribution), the sextinomial approaches are slightly better than the EM algorithm when the medians are two or 2.5 standard deviations apart. This was not the case for χ^2_{30} component distributions and medians 2.5 standard deviations apart; however, in this case the component distributions are close to the normal distribution.

With a few modifications of the program codes, we can use the information of training samples. We analyze with some detail three sampling methods. With training

samples, the Monte Carlo study shows that when the component populations are normal the multinomial approaches are comparable with the EM method and in the case study presented by Hosmer [18], where the component populations seem to be nonnormal with different variances, the sextinomial methods are comparable with the EM algorithm that assumes unequal variances and outperformed the EM algorithm that assumes equal variances.

Chapter 5

Conclusions and possible extensions of the method

5.1 Conclusions

The main objective of this thesis was to develop almost nonparametric and nonparametric methods of estimation for the mixing parameter in a mixture model. In the first case a generalization of the method proposed by Hettmansperger and Thomas [16] is given. For the second case a nonparametric approach is proposed for estimating the mixing parameter in a mixture of two continuous and symmetric distributions.

In Chapter 2 is shown that the estimation of the optimal cut point c (the point that minimizes the variance of the estimator of the mixing parameter), which defines the mixtures of binomials in the Hettmansperger and Thomas approach, does not need to be very precise for some common distributions when the separation between the means of these distributions is more than two standard deviations.

In Chapter 3, more cut points are introduced and a multinomial approach is obtained. It is shown that in general, the multinomial distribution with $r + 1$ classes is preferable over the multinomial distribution with r classes. Nevertheless, it seems that if we introduce more than two cut points (a multinomial distribution with more than three regions) the gain in efficiency is not significant.

In Chapter 4, for a mixture of two continuous and symmetric distributions two non parametric approaches are proposed, in these approaches some cut points are introduced in order to define a tetranomial distribution and a sextinomial distribution. The assumed

symmetry of the component distributions is used in order to obtain the probabilities for each class of the multinomial approach and five methods of estimation of the parameters of the multinomial mixture are studied. Two program codes in Mathematica and S-plus are presented in order to obtain the estimates of the parameters in the multinomial mixture. These tetranomial and sextinomial approaches can be easily adapted in order to use the information of training samples and three methods of sampling are considered.

5.2 Future Research

In the last chapter, we proposed two approaches for estimating the mixing parameter λ in a mixture of two continuous and symmetric distribution functions with equal shapes and densities $f_1(x)$ and $f_2(x)$ which belong to the same location family with medians M_1 and M_2. In the following subsections we will try to describe how to estimate the mixing parameters when we have univariate responses from a mixture of three continuous distribution functions with equal shapes and symmetric unimodal densities $f_1(x)$, $f_2(x)$ and $f_3(x)$ which belong to the same location family with medians M_1, M_2 and M_3. We will describe how we can obtain estimates of the mixing parameters λ_i , $i = 1, 2, 3.$, by means of multinomial distribution defined by cut points c_i'.

5.2.1 Sextinomial Approach for three Populations

The sextinomial approach for three populations, can be better explained with the help of Figure 5.1. In this figure it is assumed that we have three distributions with equal shapes and symmetric unimodal densities $f_1(x)$, the distribution to the left; $f_2(x)$, the distribution in the middle, and $f_3(x)$ the distribution to the right, which belong to the same location family with unknown medians M_1, M_2 and M_3. If the medians were known

we could select the cut points: $c_1 = M_1,\ c_2 = M_c = (M_1 + M_2)/2,\ c_3 = M_2,\ c_4 = M_d = (M_2 + M_3)/2$ and $c_5 = M_3$. These cut points define six regions $R_k,\ k = 1, 2, \ldots, 6$, where $R_1 = \{y \mid y \leq M_1\},\ R_2 = \{y \mid M_1 < y \leq M_c\},\ R_3 = \{y \mid M_c < y \leq M_2\},\ R_4 = \{y \mid M_2 < y \leq M_d\},\ R_5 = \{y \mid M_d < y \leq M_3\}$, and $R_6 = \{y \mid M_3 < y\ \}$.

Let y_j for $j = 1, \ldots, n$, be the observations of the continuous random variable whose distribution $f(x)$ is the mixture of the three densities $f_1(x)$, $f_2(x)$ and $f_3(x)$.

$$f(x) = \lambda_1 f_1(x) + \lambda_2 f_2(x) + (1 - \lambda_1 - \lambda_2) f_3(x) \tag{5.1}$$

In order to define a sextinomial distribution let:

$$z_{jk} = \begin{cases} 1 & \text{if observation } y_j \text{ is in region k=1, 2, ... ,6} \\ 0 & \text{otherwise.} \end{cases}$$

Observe now that $z_{jk} = z_{jk1} + z_{jk2} + z_{jk3}$, where:

$$z_{jki} = \begin{cases} 1 & \text{if observation } y_j \text{ has density } f_i(x), \text{ i=1,2,3 and is in region k=1, 2, ... ,6} \\ 0 & \text{otherwise.} \end{cases}$$

Let $\mathbf{z}_{j1} = (z_{j11},\ z_{j21},\ z_{j31},\ z_{j41},\ z_{j51},\ z_{j61}\ )$ be the sextinomial random variable which corresponds to the density function $f_1(x)$, $\mathbf{z}_{j2} = (z_{j12},\ z_{j22},\ z_{j32},\ z_{j42},\ z_{j52},\ z_{j62})$ the sextinomial random variable which corresponds to the density function $f_2(x)$, and $\mathbf{z}_{j3} = (z_{j13},\ z_{j23},\ z_{j33},\ z_{j43},\ z_{j53},\ z_{j63})$ be the sextinomial random variable which corresponds to the density function $f_3(x)$, therefore:

$\mathbf{z}_j = \mathbf{z}_{j1} + \mathbf{z}_{j2} + \mathbf{z}_{j3} = (z_{j1},\ z_{j2},\ z_{j3},\ z_{j4},\ z_{j5},\ z_{j6})$, is the sextinomial variable which

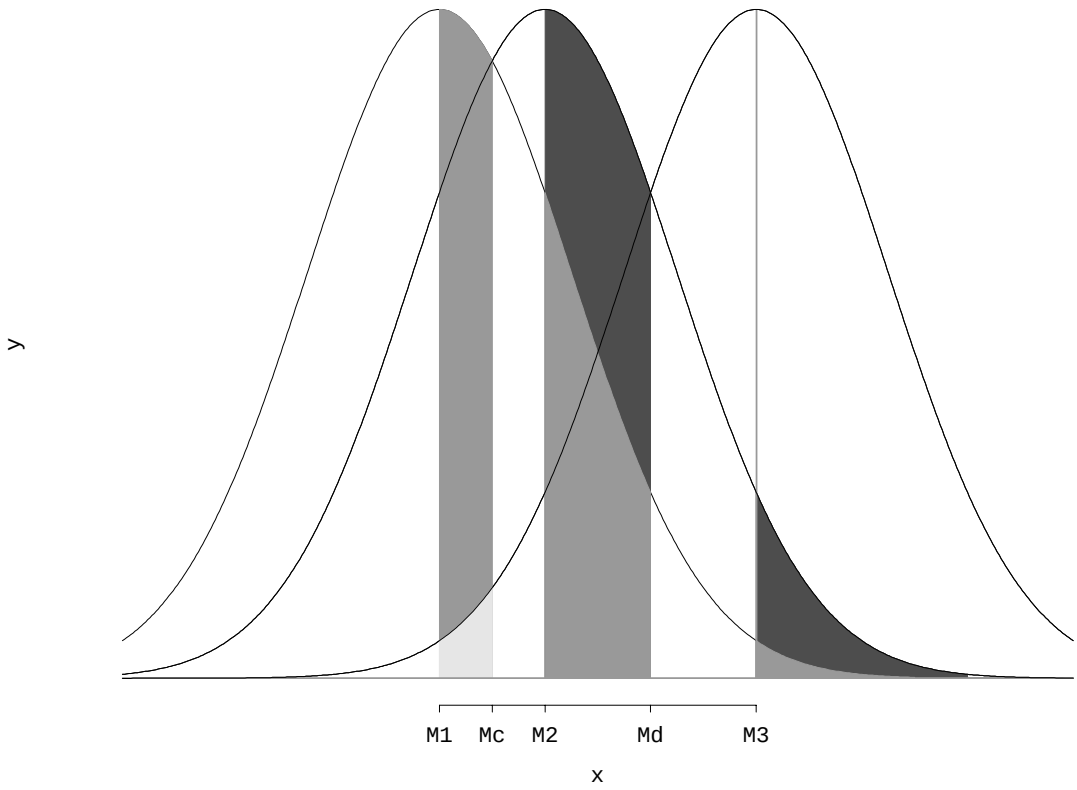

Fig. 5.1. Regions for the Sextinomial Distribution

corresponds to the continuous random variable y_j, whose distribution is a mixture of three sextinomial distributions $s_1(z)$, $s_2(z)$ and $s_3(z)$:

$$s(\mathbf{z}) = \lambda_1 s_1(\mathbf{z}) + \lambda_2 s_2(\mathbf{z}) + (1 - \lambda_1 - \lambda_2) s_3(\mathbf{z})$$

where:

$$s_i\,(\mathbf{z}_j\,; p_{ki}) = s_i(z_{j1},\, z_{j2},\, z_{j3},\, z_{j4},\, z_{j5};\, p_{1i},\, p_{2i},\, p_{3i},\, p_{4i},\, p_{5i})$$
$$= p_{1i}^{z_{j1}}\; p_{2i}^{z_{j2}}\; p_{3i}^{z_{j3}}\; p_{4i}^{z_{j4}}\; p_{5i}^{z_{j5}}\; (1 - p_{1i} - p_{2i} - p_{3i} - p_{4i} - p_{5i})^{1 - z_{j1} - z_{j2} - z_{j3} - z_{j4} - z_{j5}}$$

and:

$p_{ki} = P(z_{jki} = 1)$ for i=1, 2, 3., and k=1, 2, . . . ,6. Observe that:

$$P(\, z_{j11} = 1) = P(z_{j63} = 1) = .50$$

$$P(\, z_{j21} = 1) = P(z_{j32} = 1) = p_{21}$$

$$P(\, z_{j31} = 1) = P(z_{j22} = 1) = p_{31}$$

$$P(\, z_{j61} = 1) = P(z_{j13} = 1) = .5 - p_{21} - p_{31} - p_{41} - p_{51}$$

$$P(\, z_{j12} = 1) = .50 - p_{21} - p_{31}$$

$$P(\, z_{j42} = 1) = P(z_{j53} = 1) = p_{42}$$

$$P(\, z_{j52} = 1) = P(z_{j43} = 1) = p_{52}$$

$$P(\, z_{j62} = 1) = .5 - p_{42} - p_{52}$$

$$P(\, z_{j23} = 1) = P(z_{j62} = 1) - P(\, z_{j61} = 1) - p_{33}$$

If we let $n_k = \Sigma_{j=1}^{n} z_{jk}$, then n_k is the number of observations in region R_k, and $\mathbf{n}' = (n_1,\ n_2\ , n_3,\ n_4,\ n_5,\ n_6)$ can be interpreted as a single observation from a sextinomial distribution with parameters $n = \Sigma_{k=1}^{6} n_k$, and $P(R_1),\ P(R_2), \ldots, P(R_6)$, where $P(R_k)$, defined below, is the probability that one observation falls in Region $k = 1, 2, \ldots, 6$.

$$P(R_1) = .5\lambda_1 + (.5 - p_{21} - p_{31})\lambda_2 + (1 - \lambda_1 - \lambda_2)(.5 - p_{21} - p_{31} - p_{41} - p_{51})$$

$$P(R_2) = p_{21}\lambda_1 + p_{31}\lambda_2 + p_{23}(1 - \lambda_1 - \lambda_2)$$

$$P(R_3) = p_{31}\lambda_1 + p_{21}\lambda_2 + (.5 - p_{42} - p_{52} - p_{23}(1 - \lambda_1 - \lambda_2)$$

$$P(R_4) = p_{41}\lambda_1 + p_{42}\lambda_2 + p_{52}(1 - \lambda_1 - \lambda_2)$$

$$P(R_5) = p_{51}\lambda_1 + p_{52}\lambda_2 + p_{42}(1 - \lambda_1 - \lambda_2)$$

$$P(R_6) = (.5 - p_{21} - p_{31} - p_{41} - p_{51})\lambda_1 + (.5 - p_{42} - p_{52})(1 - \lambda_1 - \lambda_2)$$

We note that the parameters $\lambda_i,\ p_{ki}$ are unknown. If they were known they determine uniquely the values of M_1, M_2 and M_3 the medians of the component distributions. If $F(x)$ is the distribution function of the mixture:

$$M_1 = F^{-1}[P(R_1)] = F^{-1}(q_1)$$

$$M_2 = F^{-1}[P(R_1) + P(R_2) + P(R_3)] = F^{-1}(q_2)$$

$$M_3 = F^{-1}[P(R_1) + P(R_2) + P(R_3) + P(R_4) + P(R_5)] = F^{-1}(q_3)$$

Where:

$$q_1 = .5 - (1 - \lambda_1)(p_{21} + p_{31} + p_{41} + p_{51}) + \lambda_2(p_{41} + p_{51})$$

$$q_2 = .5 + \lambda_1(p_{21} + p_{31}) - (1 - \lambda_1 - \lambda_2)(p_{42} + p_{52})$$

$$q_3 = .5 - \lambda_1(p_{21} + p_{31} + p_{41} + p_{51}) + \lambda_2(.5 - p_{42} - p_{52})$$

We cannot obtain the M_i because $F(x)$ the distribution function of the mixture is unknown but, if we assumed that $\lambda_i,\ p_{ki}$ are known, we could estimate (as in Chapter 4) M_1, M_2 and M_3 with the help of the empirical *cdf* :

$$\widehat{M_1} = m_1 = y_{(q_1)}$$

$$\widehat{M_2} = m_2 = y_{(q_2)}$$

$$\widehat{M_3} = m_3 = y_{(q_3)}$$

Observe that if we are interested in estimating the mixing parameters when the separation among the medians of the distributions are more than two standard deviations apart, some of the probabilities p_{ki} given above can be set to zero.

5.2.2 Decinomial Approach for three Populations

The decinomial approach for three populations, can be better explained with the help of Figure 5.2. In this figure it is assumed that we have three distributions with equal shapes and symmetric unimodal densities $f_1(x)$, the distribution to the left; $f_2(x)$, the distribution in the middle, and $f_3(x)$ the distribution to the right, which belong to the same location family with unknown medians M_1, M_2 and M_3. If the medians were known we could select the cut points: $c_1 = M_a = M_1 - (M_2 - M_1),\ c_2 = M_b = M_1 - (M_2 - M_1)/2$,

$c_3 = M_1$, $c_4 = M_c = (M_1 + M_2)/2$, $c_5 = M_2$, $c_6 = M_d = (M_2 + M_3)/2$, $c_7 = M_3$, $c_8 = M_e = M_3 + (M_3 - M_2)/2$ and $c_9 = M_f = M_3 + (M_3 - M_2)$.

These cut points define ten regions R_k, $k = 1, 2, \ldots, 10$, where $R_1 = \{y \mid y \leq M_a\}$, $R_2 = \{y \mid M_a < y \leq M_b\}$, $R_3 = \{y \mid M_b < y \leq M_1\}$, $R_4 = \{y \mid M_1 < y \leq M_c\}$, $R_5 = \{y \mid M_c < y \leq M_2\}$, $R_6 = \{y \mid M_2 < y \leq M_d\}$, $R_7 = \{y \mid M_d < y \leq M_3\}$, $R_8 = \{y \mid M_3 < y \leq M_e\}$, $R_9 = \{y \mid M_e < y \leq M_f\}$, and $R_{10} = \{y \mid M_f < y\}$.

The estimates of M_i, assuming that λ_i, p_{ki} are known, are:

$\widehat{M_1} = m_1 = y_{(q_1)}$

$\widehat{M_2} = m_2 = y_{(q_2)}$

$\widehat{M_3} = m_3 = y_{(q_3)}$

Where:

$q_1 = .5 - (1 - \lambda_1)(p_{41} + p_{51} + p_{61} + p_{71}) + \lambda_2(p_{61} + p_{71})$

$q_2 = .5 + \lambda_1(p_{41} + p_{51}) - (1 - \lambda_1 - \lambda_2)(p_{62} + p_{72})$

$q_3 = .5 - \lambda_1(p_{41} + p_{51} + p_{61} + p_{71}) + \lambda_2(.5 - p_{62} - p_{72})$

5.2.3 Dodecanomial Approach for three Populations

We can propose also the Dodecanomial approach for three populations, this approach can be better explained with the help of Figure 5.3.In this figure it is assumed again that we have three distributions with equal shapes and symmetric unimodal densities $f_1(x)$, the distribution to the left; $f_2(x)$, the distribution in the middle, and $f_3(x)$ the distribution to the right, which belong to the same location family with unknown medians M_1, M_2 and M_3. If the medians were known we could select the cut points: $c_1 = M_i =$

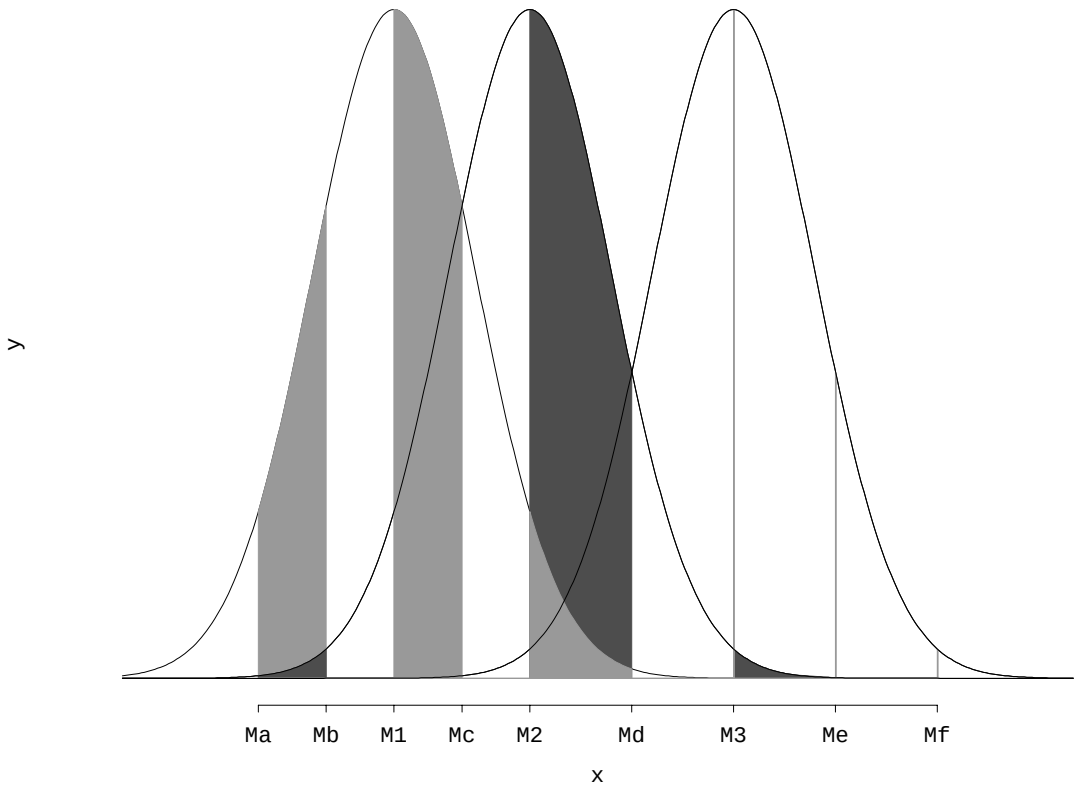

Fig. 5.2. Regions for the Decinomial Distribution

$M_1 - (3/2)(M_2 - M_1)$, $c_2 = M_a = M_1 - (M_2 - M_1)$, $c_3 = M_b = M_1 - (M_2 - M_1)/2$, $c_4 = M_1$, $c_5 = M_c = (M_1 + M_2)/2$, $c_6 = M_2$, $c_7 = M_d = (M_2 + M_3)/2$, $c_8 = M_3$, $c_9 = M_e = M_3 + (M_3 - M_2)/2$, $c_{10} = M_f = M_3 + (M_3 - M_2)$ and $c_{11} = M_f = M_3 + (3/2)(M_3 - M_2)$.

These cut points define twelve regions R_k, $k = 1, 2, \ldots, 12$, where $R_1 = \{y \mid y \leq M_i\}$, $R_2 = \{y \mid M_i < y \leq M_a\}$, $R_3 = \{y \mid M_a < y \leq M_b\}$, $R_4 = \{y \mid M_b < y \leq M_1\}$, $R_5 = \{y \mid M_1 < y \leq M_c\}$, $R_6 = \{y \mid M_c < y \leq M_2\}$, $R_7 = \{y \mid M_2 < y \leq M_d\}$, $R_8 = \{y \mid M_d < y \leq M_3\}$, $R_9 = \{y \mid M_3 < y \leq M_e\}$, $R_{10} = \{y \mid M_e < y \leq M_f\}$, $R_{11} = \{y \mid M_f < y \leq M_g\}$, and $R_{12} = \{y \mid M_g < y\ \}$.

The estimates of M_i for the dodecanomial approach, assuming that λ_i, p_{ki} are known, are:

$\widehat{M_1} = m_1 = y_{(q_1)}$

$\widehat{M_2} = m_2 = y_{(q_2)}$

$\widehat{M_3} = m_3 = y_{(q_3)}$

Where:

$q_1 = .5 - (1 - \lambda_1)(p_{51} + p_{61} + p_{71} + p_{81}+) + \lambda_2(p_{71} + p_{81})$

$q_2 = .5 + \lambda_1(p_{51} + p_{61}) - (1 - \lambda_1 - \lambda_2)(p_{72} + p_{82})$

$q_3 = .5 - \lambda_1(p_{51} + p_{61} + p_{71} + p_{81}) + \lambda_2(.5 - p_{72} - p_{82})$

5.2.4 Methods of Estimation

In these cases, as in the last chapter, it is also possible to analyze four approaches in order to estimate the parameters for the sextinomial, decinomial and dodecanomial

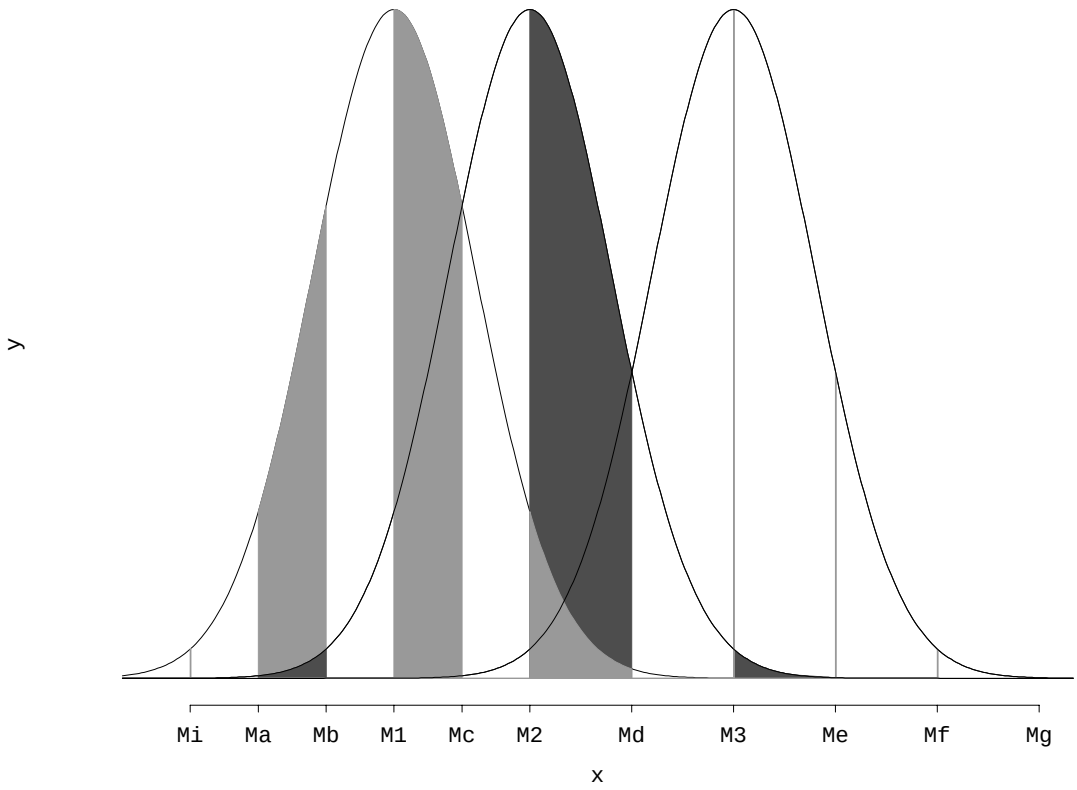

Fig. 5.3. Regions for the Dodecanomial Distribution

distributions for three populations, these approaches are: The Least Squares, Chi-Square, Modified Chi-Square and Hellinger methods. We proposed again, in liu of a better method, a grid search for the estimates of the parameters and a Monte Carlo study in order to compare them with the corresponding EM algorithm estimator. We consider that in this case, due to the large number of unknown parameters many computational problems have to be overcome in order to obtain the estimates of the mixing parameters.

A future interesting research topic is also the development of computational programs (no grid search) for the estimates of the mixing parameters.

Appendix A

Programs code for the multinomial approaches

A.1 Tetranomial Approach

A.1.1 Splus code for the Tetranomial Approach

```
i_1
#Vector v0 to v4 will have the results
#for the different criteria
v0_matrix(0,i,7)
v1_v0;  v2_v0;  v3_v0;  v4_v0

pvalues_(5:95)/100

p1values_(17:50)/100

p2values_(1:25)/100

plen_length(pvalues)

p1len_length(p1values)

p2len_length(p2values)

p_rep(pvalues,rep(p1len*p2len,plen))

p1_rep(rep(p1values,rep(p2len,p1len)),plen)

p2_rep(p2values,p1len*plen)

epsilon<-1/1000000

temp_(p2 < p1-epsilon) & (p2 < .5-p1+epsilon) &

(.5-p1-p2 <= p2 - epsilon)

p_p[temp]

p1_p1[temp]

p2_p2[temp]

#t is the number of observations in our data set
t_100

#Function which takes a scalar k and a vector vec and returns
#the number of entries of vec less than or equal to k

lessequal_function(k,vec) sum(vec<=k)

#da has the data set
```

```
da_c(75,78,...,90)

s2_var(da)

#Calculation of m0

q1_t*(.5-(1-p)*(p1+p2))
q2_floor(q1)
m0_da[q2]+(q1-q2)*(da[q2+1]-da[q2])

#Calculation of m1

q3_t*(.5+p*(p1+p2))
q4_floor(q3)
m1_da[q4]+(q3-q4)*(da[q4+1]-da[q4])
mc_(m0+m1)/2

# Note:  q1, q2, q3, and q4 will be reused later.
# They are very large vectors.

# Number of observations per region
n22_apply(matrix(mc),1,lessequal,da)-q1
n33_apply(matrix(m1),1,lessequal,da)-n22-q1

q2_t*(p2+p*(p1-p2))
q3_t*(p1+p*(p2-p1))
q4_t*(.5-p*(p1+p2))

# LS criterion function
z_(q2-n22)^2 + (q3-n33)^2
      mi_(1:length(z))[z==min(z)][1]
      v0[i,]_c(z[mi],p[mi],p1[mi],p2[mi],m0[mi],m1[mi],s2)

#Chi-Square criterion

z_(q2-n22)^2/q2 + (q3-n33)^2/q3
      mi_(1:length(z))[z==min(z)][1]
      v1[i,]_c(z[mi],p[mi],p1[mi],p2[mi],m0[mi],m1[mi],s2)

#Modified Chi-Square criterion

z_(q1-n11)^2/(q1+n11) + (q2-n22)^2/(q2+n22) +
      (q3-n33)^2/(q3+n33) + (q4-n44)^2/(q4+n44)
      mi_(1:length(z))[z==min(z)][1]
      v2[i,]_c(z[mi],p[mi],p1[mi],p2[mi],m0[mi],m1[mi],s2)

#Hellinger 1

z_(sqrt(q2/t)-sqrt(n22/t))^2 + (sqrt(q3/t)-sqrt(n33/t))^2
```

```
      mi_(1:length(z))[z==min(z)][1]
      v3[i,]_c(z[mi],p[mi],p1[mi],p2[mi],m0[mi],m1[mi],s2)
#Hellinger 2 distance
z_acos(q1/t + sqrt(n22*q2)/t + sqrt(n33*q3)/t + q4/t)
      mi_(1:length(z))[z==min(z)][1]
      v4[i,]_c(z[mi],p[mi],p1[mi],p2[mi],m0[mi],m1[mi],s2)
#It is necessary to remove the large matrices defined above
rm(p,p1,p2,m0,m1,mc,q1,q2,q3,q4,n11,n22,n33,n44)
```

A.1.2 Mathematica code for the tetranomial Approach

```
da={Data set}
n1=da;  n2=da;  n3=da;  n4=da;
t=Number of observations;
m=0;n=0;l=0;r=0;p=.05;p0=.98;p1=.20;p10=.50;p2=.01;d=.01;
w0=10000;w01=10000;w02=10000;w03=10000;w04=10000;
Clear[w];y=Array[w,6];
Clear[w1];y1=Array[w1,6];
Clear[w2];y2=Array[w2,6];
Clear[w3];y3=Array[w3,6];
Clear[w4];y4=Array[w4,6];
      While[p<p0,
n=n+1;
      While[p1<p10,
l=l+1;
p2=Max[N[Ceiling[100(.5-p1-.01)/2]/100],.01];
      While[p2<Min[p1,.5-p1],
r=r+1;
nm0=t*(.5-(1-p)*(p1+p2));
s=Floor[nm0];
m0=da[[s]]+ (nm0-s)*(da[[s+1]]-da[[s]]);
nm1=t*(.5+p*(p1+p2));
s=Floor[nm1];
m1=da[[s]]+ (nm1-s)*(da[[s+1]]-da[[s]]);
mc=(m0+m1)/2;
Do[If[da[[i]] ≤ mc,n2[[i]]=1,n2[[i]]=0],i,t];
Do[If[da[[i]] ≤ m1,n3[[i]]=1,n3[[i]]=0],i,t];
n22=Abs[Apply[Plus,n2]-nm0];
```

```
n33=Abs[Apply[Plus,n3]-n22-nm0];

z=((t*(p2+p*(p1-p2))-n22)^2)+((t*(p1+p*(p2-p1))-n33)^2);
      If[z<w0,w[1]=z,w[1]=w[1] ];
      If[z<w0,w[2]=p,w[2]=w[2] ];
      If[z<w0,w[3]=p1,w[3]=w[3] ];
      If[z<w0,w[4]=p2,w[4]=w[4] ];
      If[z<w0,w[5]=m0,w[5]=w[5] ];
      If[z<w0,w[6]=m1,w[6]=w[6] ];
      w0=Min[w0,z];

z1=((t*(p2+p*(p1-p2))-n22)^2)/(t*(p2+p*(p1-p2)))+
      ((t*(p1+p*(p2-p1))-n33)^2)/( t*(p1+p*(p2-p1)));
      If[z1<w01,w1[1]=z1,w1[1]=w1[1] ];
      If[z1<w01,w1[2]=p,w1[2]=w1[2] ];
      If[z1<w01,w1[3]=p1,w1[3]=w1[3] ];
      If[z1<w01,w1[4]=p2,w1[4]=w1[4] ];
      If[z1<w01,w1[5]=m0,w1[5]=w1[5] ];
      If[z1<w01,w1[6]=m1,w1[6]=w1[6] ];
      w01=Min[w01,z1];

z2=((t*(p2+p*(p1-p2))-n22)^2)/( t*(p2+p*(p1-p2))+n22)
      +((t*(p1+p*(p2-p1))-n33)^2)/( t*(p1+p*(p2-p1))+n33);
      If[z2<w02,w2[1]=z2,w2[1]=w2[1] ];
      If[z2<w02,w2[2]=p,w2[2]=w2[2] ];
      If[z2<w02,w2[3]=p1,w2[3]=w2[3] ];
      If[z2<w02,w2[4]=p2,w2[4]=w2[4] ];
      If[z2<w02,w2[5]=m0,w2[5]=w2[5] ];
      If[z2<w02,w2[6]=m1,w2[6]=w2[6] ];
      w02=Min[w02,z2];

z3=(Sqrt[n22/t]-Sqrt[p2+p*(p1-p2)])^2+(Sqrt[n33/t]-Sqrt[p1+p*(p2-p1)])^2;
      If[z3<w03,w3[1]=z3,w3[1]=w3[1] ];
      If[z3<w03,w3[2]=p,w3[2]=w3[2] ];
      If[z3<w03,w3[3]=p1,w3[3]=w3[3] ];
      If[z3<w03,w3[4]=p2,w3[4]=w3[4] ];
      If[z3<w03,w3[5]=m0,w3[5]=w3[5] ];
      If[z3<w03,w3[6]=m1,w3[6]=w3[6] ];
      w03=Min[w03,z3];
```

```
p2=p2+d;
];
p1=p1+d;
p2=.01;
r=0;
];
p=p+d;
p1=.10;
p2=.01;
r=0;
l=0;
];
```

A.2 Sextinomial Approach

A.2.1 Splus code for the Sextinomial Approach

```
i_1
#Vector v0 to v4 will have the results
#for the different criteria
v0_matrix(0,i,8)
v1_v0;  v2_v0;  v3_v0;  v4_v0;

pvalues_(5:80)/100
p1values_(20:50)/100
p2values_(1:15)/100
p3values_(1:15)/100
plen_length(pvalues)
p1len_length(p1values)
p2len_length(p2values)
p3len_length(p3values)

p_rep(pvalues,rep(p1len*p2len*p3len,plen))
p1_rep(rep(p1values,rep(p2len*p3len,p1len)),plen)
p2_rep(rep(p2values,rep(p3len,p2len)),p1len*plen)
p3_rep(p3values,p1len*p2len*plen)

epsilon<-1/1000000

temp <- (p2 < p1-epsilon) & (p2 <= .5-p1 + epsilon) &
( p2 >= (.5-p1)/3-epsilon) & ( p3 < p2 -epsilon) &
(p3<= .5-p1-p2+epsilon) & (p3 >=(.5-p1-p2)/2-epsilon)
```

```
p_p[temp]
p1_p1[temp]
p2_p2[temp]
p3_p3[temp]
#t is the number of observations in our data set
t_100
# Function which takes a scalar k and a vector vec and
# returns the number of entries of vec less than or equal to k
lessequal_function(k,vec) sum(vec<=k)
#da has the data set
da_c(75,78,...,90)
s2_var(da)
#Calculation of m0
q1_t*(.5-(1-p)*(p1+p2))
q2_floor(q1)
m0_da[q2]+(q1-q2)*(da[q2+1]-da[q2])
#Calculation of m1
q3_t*(.5+p*(p1+p2))
q4_floor(q3)
m1_da[q4]+(q3-q4)*(da[q4+1]-da[q4])
md_(m1-m0)/2
mb_m0-md
mc_(m0+m1)/2
ms_m1+md
# Note:  q1, q2, q3, and q4 will be reused later.
# They are very large vectors.
# Number of observations per region
n11_apply(matrix(mb),1,lessequal,da)
n22_apply(matrix(m0),1,lessequal,da)-n11
n33_apply(matrix(mc),1,lessequal,da)-n22-n11
n44_apply(matrix(m1),1,lessequal,da)-n33-n22-n11
n55_apply(matrix(ms),1,lessequal,da)-n44-n33-n22-n11
n66_t-n55-n44-n33-n22-n11
q1_t*(.5-p1-(1-p)*(p2+p3))
q2_t*(p*p1+(1-p)*p3)
q3_t*(p*p1+(1-p)*p2)
q4_t*(p*p2+(1-p)*p1)
```

```
q5_t*(p*p3+(1-p)*p1)
q6_t*(.5-p1-p*(p2+p3))

#LS criterion function
z_(q1-n11)^2 + (q2-n22)^2 + (q3-n33)^2 +
      (q4-n44)^2 + (q5-n55)^2 + (q6-n66)^2
      mi_(1:length(z))[z==min(z)][1]
      v0[i,]_c(z[mi],p[mi],p1[mi],p2[mi],p3[mi],m0[mi],m1[mi],s2)

#Chi-Square criterion
z_(q1-n11)^2/q1 + (q2-n22)^2/q2 + (q3-n33)^2/q3 +
      (q4-n44)^2/q4 +(q5-n55)^2/q5 +(q6-n66)^2/q6
      mi_(1:length(z))[z==min(z)][1]
      v1[i,]_c(z[mi],p[mi],p1[mi],p2[mi],p3[mi],m0[mi],m1[mi],s2)

# Modified Chi-Square criterion
z_(q1-n11)^2/(q1+n11) + (q2-n22)^2/(q2+n22) +
      (q3-n33)^2/(q3+n33) + (q4-n44)^2/(q4+n44) +
      (q5-n55)^2/(q5+n55) + (q6-n66)^2/(q6+n66)
      mi_(1:length(z))[z==min(z)][1]
      v2[i,]_c(z[mi],p[mi],p1[mi],p2[mi],p3[mi],m0[mi],m1[mi],s2)

#Hellinger 1 criterion
z_(sqrt(q1/t)-sqrt(n11/t))^2 + (sqrt(q2/t)-sqrt(n22/t))^2 +
      (sqrt(q3/t)-sqrt(n33/t))^2 + (sqrt(q4/t)-sqrt(n44/t))^2 +
      (sqrt(q5/t)-sqrt(n55/t))^2 + (sqrt(q6/t)-sqrt(n66/t))^2
      mi_(1:length(z))[z==min(z)][1]
      v3[i,]_c(z[mi],p[mi],p1[mi],p2[mi],p3[mi],m0[mi],m1[mi],s2)

# Hellinger 2 criterion
z_acos( (sqrt(n11*q1) + sqrt(n22*q2) + sqrt(n33*q3)
      + sqrt(n44*q4) + sqrt(n55*q5) + sqrt(n66*q6) )/t )
      mi_(1:length(z))[z==min(z)][1]
      v4[i,]_c(z[mi],p[mi],p1[mi],p2[mi],p3[mi],m0[mi],m1[mi],s2)

#It is necessary to remove the large matrices defined above
rm(p,p1,p2,p3,m0,m1,mb,mc,md,ms,q1,q2,q3,q4,q5,q6,
n11,n22,n33,n44,n55,n66)
```

A.2.2 Mathematica code for the Sextinomial Approach

```
da={Data set}
n1=da;  n2=da;  n3=da;
```

```
n4=da;  n5=da;  n6=da;
t=125;
m=0; n=0;l=0;r=0;r1=1;p=.05;p0=.98;p1=.20;p10=.50;p2=.01;
p3=.01;d=.01;
w0=10000;w01=10000;w02=10000;w03=10000;w04=10000;
Clear[w];y=Array[w,7];
Clear[w1];y1=Array[w1,7];
Clear[w2];y2=Array[w2,7];
Clear[w3];y3=Array[w3,7];
Clear[w4];y4=Array[w4,7];
      While[p<p0,
n=n+1;
      While[p1<p10,
l=l+1;
p2=Max[N[Ceiling[100(.5-p1-.01)/3]/100],.01];
      While[p2<Min[p1,.5-p1],
r=r+1;
p3=Max[N[Ceiling[100(.5-p1-p2-.01)/2]/100],.01];
      While[p3<Min[p2,.5-p1-p2],
r1=r1+1;

nm0=t*(.5*p+(1-p)*(.5-p1-p2));
s=0;
While[s<nm0,
s=s+1;
m0=da[[s-1]]+ (nm0-s+1)*(da[[s]]-da[[s-1]]);] ;


nm1=t*((.5+p1+p2)*p+.5*(1-p));
s=0;
While[s<nm1,
s=s+1;
m1=da[[s-1]]+ (nm1-s+1)*(da[[s]]-da[[s-1]]);];

dc=(m1-m0)/2;
mc=(m0+m1)/2;
mi=m0-dc;
md=m1+dc;
Do[If[da[[i]]≤ mi,n1[[i]]=1,n1[[i]]=0],i,t];
```

```
Do[If[mi<da[[i]]≤ m0,n2[[i]]=1,n2[[i]]=0],i,t];
Do[If[m0<da[[i]]≤ mc,n3[[i]]=1,n3[[i]]=0],i,t];
Do[If[mc<da[[i]]≤ m1,n4[[i]]=1,n4[[i]]=0],i,t];
Do[If[m1<da[[i]]≤ md,n5[[i]]=1,n5[[i]]=0],i,t];
Do[If[md<da[[i]],n6[[i]]=1,n6[[i]]=0],i,t];

n11=Apply[Plus,n1];
n22=Apply[Plus,n2];
n33=Apply[Plus,n3];
n44=Apply[Plus,n4];
n55=Apply[Plus,n5];
n66=Apply[Plus,n6];

z=((t*(.5-p1-(1-p)*(p2+p3))-n11)^2 )+ ((t*(p*p1+(1-p)*p3)-n22)^2)+
      (( t*(p*p1+(1-p)*p2)-n33)^2)+((t*(p*p2+(1-p)*p1)-n44)^2)+
      (( t*(p*p3+(1-p)*p1)-n55)^2)+(t*(.5-p1-p*(p2+p3))-n66)^2;
      If[z<w0,w[1]=z,w[1]=w[1] ];
      If[z<w0,w[2]=p,w[2]=w[2] ];
      If[z<w0,w[3]=p1,w[3]=w[3] ];
      If[z<w0,w[4]=p2,w[4]=w[4] ];
      If[z<w0,w[5]=p3,w[5]=w[5] ];
      If[z<w0,w[6]=m0,w[6]=w[6] ];
      If[z<w0,w[7]=m1,w[7]=w[7] ];
      w0=Min[w0,z];

z1=((t*(.5-p1-(1-p)*(p2+p3))-n11)^2 )/( t*(.5-p1-(1-p)*(p2+p3))) +
      ((t*(p*p1+(1-p)*p3)-n22)^2)/( t*(p*p1+(1-p)*p3)) +
      ((t*(p*p1+(1-p)*p2)-n33)^2)/( t*(p*p1+(1-p)*p2)) +
      ((t*(p*p2+(1-p)*p1)-n44)^2)/( t*(p*p2+(1-p)*p1)) +
      ((t*(p*p3+(1-p)*p1)-n55)^2)/( t*(p*p3+(1-p)*p1)) +
      ((t*(.5-p1-p*(p2+p3))-n66)^2)/( t*(.5-p1-p*(p2+p3)));
      If[z1<w01,w1[1]=z1,w1[1]=w1[1] ];
      If[z1<w01,w1[2]=p,w1[2]=w1[2] ];
      If[z1<w01,w1[3]=p1,w1[3]=w1[3] ];
      If[z1<w01,w1[4]=p2,w1[4]=w1[4] ];
      If[z1<w01,w1[5]=p3,w1[5]=w1[5] ];
      If[z1<w01,w1[6]=m0,w1[6]=w1[6] ];
      If[z1<w01,w1[7]=m1,w1[7]=w1[7] ];
```

```
      w01=Min[w01,z1];

z2=((t*(.5-p1-(1-p)*(p2+p3))-n11)^2 )/( t*(.5-p1-(1-p)*(p2+p3))+n11) +
      ((t*(p*p1+(1-p)*p3)-n22)^2)/( t*(p*p1+(1-p)*p3)+n22) +
      ((t*(p*p1+(1-p)*p2)-n33)^2)/( t*(p*p1+(1-p)*p2)+n33) +
      ((t*(p*p2+(1-p)*p1)-n44)^2)/( t*(p*p2+(1-p)*p1)+n44) +
      ((t*(p*p3+(1-p)*p1)-n55)^2)/( t*(p*p3+(1-p)*p1)+n55) +
      ((t*(.5-p1-p*(p2+p3))-n66)^2)/( t*(.5-p1-p*(p2+p3))+n66);
      If[z2<w02,w2[1]=z2,w2[1]=w2[1] ];
      If[z2<w02,w2[2]=p,w2[2]=w2[2] ];
      If[z2<w02,w2[3]=p1,w2[3]=w2[3] ];
      If[z2<w02,w2[4]=p2,w2[4]=w2[4] ];
      If[z2<w02,w2[5]=p3,w2[5]=w2[5] ];
      If[z2<w02,w2[6]=m0,w2[6]=w2[6] ];
      If[z2<w02,w2[7]=m1,w2[7]=w2[7] ];
      w02=Min[w02,z2];

z3=(Sqrt[ n11/t]-Sqrt[.5-p1-(1-p)*(p2+p3)])^2 +
      ( Sqrt[n22/t]-Sqrt[p*p1+(1-p)*p3])^2 +
      ( Sqrt[n33/t]-Sqrt[p*p1+(1-p)*p2])^2 +
      ( Sqrt[n44/t]-Sqrt[p*p2+(1-p)*p1])^2 +
      ( Sqrt[n55/t]-Sqrt[p*p3+(1-p)*p1])^2 +
      ( Sqrt[n66/t]-Sqrt[.5-p1-p*(p2+p3)])^2;
      If[z3<w03,w3[1]=z3,w3[1]=w3[1] ];
      If[z3<w03,w3[2]=p,w3[2]=w3[2] ];
      If[z3<w03,w3[3]=p1,w3[3]=w3[3] ];
      If[z3<w03,w3[4]=p2,w3[4]=w3[4] ];
      If[z3<w03,w3[5]=p3,w3[5]=w3[5] ];
      If[z3<w03,w3[6]=m0,w3[6]=w3[6] ];
      If[z3<w03,w3[7]=m1,w3[7]=w3[7] ];
      w03=Min[w03,z3];
```

```
z4=ArcCos[Sqrt[ (n11/t)*(.5-p1-(1-p)*(p2+p3)) ] +
      Sqrt[(n22/t)*(p*p1+(1-p)*p3)] + Sqrt[(n33/t)*(p*p1+(1-p)*p2)]+
      Sqrt[(n44/t)*(p*p2+(1-p)*p1)] + Sqrt[(n55/t)*(p*p3+(1-p)*p1)]+
      Sqrt[(n66/t)*(.5-p1-p*(p2+p3))]  ];
      If[z4<w04,w4[1]=z3,w4[1]=w4[1] ];
      If[z4<w04,w4[2]=p,w4[2]=w4[2] ];
      If[z4<w04,w4[3]=p1,w4[3]=w4[3] ];
      If[z4<w04,w4[4]=p2,w4[4]=w4[4] ];
      If[z4<w04,w4[5]=p3,w4[5]=w4[5] ];
      If[z4<w04,w4[6]=m0,w4[6]=w4[6] ];
      If[z4<w04,w4[7]=m1,w4[7]=w4[7] ];
      w04=Min[w04,z4];

p3=p3+d;
];
p2=p2+d;
p3=.01;
r1=0;
];
p1=p1+d;
p2=.01;
p3=.01;
r1=0;
r=0;
];
p=p+.01;
p1=.20;
p2=.01;
p3=.01;
r1=0;
r=0;
l=0;
];
```

Appendix B

Index of Selected Notation

AV, asymptotic variance of $\sqrt{n}(\widehat{\lambda} - \lambda_0)$ for the binomial distribution unless another multinomial distribution is stated.

$A.V_{Bin}\,[\sqrt{n}(\widehat{\lambda} - \lambda_0)]$, asymptotic variance of $\sqrt{n}(\widehat{\lambda} - \lambda_0)$ for the binomial approach.

$A.V_{Tri}\,[\sqrt{n}(\widehat{\lambda} - \lambda_0)]$, asymptotic variance of $\sqrt{n}(\widehat{\lambda} - \lambda_0)$ for the trinomial approach.

$A.V_{M_r}\,[\sqrt{n}(\widehat{\lambda}_r - \lambda_0)]$, asymptotic variance of $\sqrt{n}(\widehat{\lambda} - \lambda_0)$ for the multinomial approach with r classes.

$b_i(p) = b(i; m, p)$, binomial density function with parameters m and p.

c_i , cut point i for the multinomial distribution.

Methods of estimation for the mixing parameter:

χ^2, Chi-square method, section 4.4.3

EM, EM algorithm assuming that the component populations are normal

H1, Hellinger I method, section 4.4.5

H2, Hellinger II method, section 4.4.6

Mod. χ^2, modified Chi-square method, section 4.4.2

LS, least squares method, section 4.4.2

M_i , median of the $i-th$ component population.

m_i , estimator of the median of the $i-th$ population.

$M_r(y_1, y_2, \ldots, y_r\,;\, m\,, p_1, p_2, \ldots, p_r)$, multinomial mass function with $r+1$ classes and parameters $m, p_1, p_2, \ldots, p_r$, such that $\sum_{j=1}^{r} p_j < 1$.

ov, overlap defined in subsection 4.6.1.

$\rho = \sigma_2/\sigma_1$, ratio of standard deviations, it is assumed that $\sigma_2 > \sigma_1$.

$\mathrm{Var}\,(\widehat{\lambda})$, variance of $\widehat{\lambda}$ based on 500 samples of size $n = 100$ for the different approaches.

References

[1] Bernardo J. M. and Girón F. G. A Bayesian analysis of simple mixtures problems. *Bayesian Statistics*, 3, 67-78, 1988.

[2] Blischke, W. R. Estimating the parameters of mixtures of binomial. *J. Amer. Statist. Assoc.*, 59, 510-528, 1964.

[3] Bozdogan, H. Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika* , 52, 345-370, 1987

[4] Chen, J. and Kalbfleisch, J. D. Penalized minimum-distance estimates in finite mixture models. *Can. J. Statist.*, 24, 167-175, 1996.

[5] Chen H., Chen, J. and Kalbfleisch, J. D. A modified likelihood ratio test for homogeneity in the finite mixture models. *Working Paper University of Waterloo*, 2000.

[6] Cutler, A. and Cordero-Brañ̃a, O. I. Minimum Hellinger distance estimation for finite mixture models. *J. Amer. Statist. Assoc.*, 91, 1716-1723, 1996.

[7] Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B*, 39, 1-38. 1977.

[8] Diebolt J. and Christian P. R. Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society B*, 56, 363-375, 1994.

[9] Do K., and McLachlan, G. J. Estimation of mixing proportions: A case study. *Applied Statistics*, 33, 134-140, 1984.

[10] Everitt, B. S. A Monte Carlo investigation of the likelihood ratio test for the number of components in a mixture of normal distributions. *Multivariate Behavioral Research*, 16, 171-180, 1981.

[11] Everitt, B. S. and Hand, D. J. *Finite Mixture Distributions.* New York: John Wiley. 1981.

[12] Fowlkes, E. B. Some methods for studying the mixture of two normal (lognormal) distributions. *J. Am. Statist. Assoc.*, 74, 561-575, 1979.

[13] Hall, P. On the non-parametric estimation of mixing proportions. *Journal of the Royal Statistical Society B*, 43, 147-156, 1981.

[14] Hall, P. and Titterington D. M. Efficient nonparametric estimation of mixing proportions. *Journal of the Royal Statistical Society B*, 46, 465-473, 1984.

[15] Hettmansperger, T. P. *Statistical Inference based on Ranks.* John Wiley & Sons. 1984.

[16] Hettmansperger, T. P., and Thomas H. Almost nonparametric inference for repeated measures in mixture models. *Journal of the Royal Statistical Society B*, 62, 811-825, 2000.

[17] Hill, B. M. Information for estimating proportion in the mixtures of exponential and normal distributions. *J. Amer. Statist. Assoc.*, 58, 918-932, 1963.

[18] Hosmer, D. W. On MLE of the parameters of a mixture of two normal distributions when the sample size is small. *Communications in Statistics*, 1, 217-227, 1973

[19] Leroux, B. G. Consistent estimation of a mixing distribution. *Ann. Statist.*, 20, 1350-1360, 1992.

[20] Lindsay, B. G. Efficiency versus robustness: The case for minimum Hellinger distance and related methods. *Ann. Statist.*, 22, 1081-1114, 1994.

[21] Lindsay, B. G. *Mixture Models: Theory, Geometry, and Applications*. NSF-CBMS Regional Conference Series in Probability and Statistics v 5. Hayward, CA: Institute of Mathematical Statistics. 1995.

[22] Lindsay, B. G., and Roeder K. Residual diagnostics in the mixture model. *J. Am. Statist. Assoc.*, 87, 785-795, 1992.

[23] McLachlan, G. J. On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics*, 36, 318-324, 1987.

[24] McLachlan, G. J. and Basford, K. E. *Mixture Models: Inference and applications to clustering.* New York: Dekker. 1988.

[25] McLachlan, G. J. and Krishnan, T. *The EM Algorithm and Extensions.* New York: John Wiley. 1997.

[26] Murray, G. D. and Titterington D. M. Estimation problems with data from a mixture. *Applied Statistics*, 27, 325-334, 1978.

[27] Pearson, K. Contributions to the mathematical theory of evolution. *Phil. Trans. A.*, 185, 71-110, 1894.

[28] Rao, C. R. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society B*, 10, 159-203, 1948.

[29] Roeder, K. A graphical technique for determining the number of components in a mixture of normals. *J. Amer. Statist. Assoc.*, 89, 487-495, 1994.

[30] Tamura N. R. and Boos D. D. Minimum Hellinger distance estimation for multivariate location and covariance. *J. Amer. Statist. Assoc.*, 81, 223-229, 1984.

[31] Tan W. Y. and Chang W. C. Some comparisons of the method of moments and the method of maximum likelihood in estimating parameters of a mixture of two normal densities. *J. Amer. Statist. Assoc.*, 67, 702-708, 1972.

[32] Thomas, H. and Horton, J. J. Competency criteria and the class inclusion task: Modeling judgments and justifications. *Developmental Psychology*, 33, 1060-1073, 1997.

[33] Thomas, H. and Lohaus, A. *Modeling Growth and Individual Differences in Spatial Tasks.* Chicago: University of Chicago Press. Monographs of the Society for Research in Child Development, Serial No. 237. 1993.

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

[35] Windham, M. P. and Cutler, A. Information ratios for validating mixture analyses. *J. Amer. Statist. Assoc.,* 87, 1188-1192, 1992.

[36] Woodward, W. A., Parr W. C., Schucany W. R. and Lindsay H. A Comparison of minimum distance and maximum likelihood estimation of a mixture proportion. *J. Amer. Statist. Assoc.*, 79,590-598, 1984.

[37] Woodward, W. A., Whitney, P. and Eslinger, P. W. Minimum Hellinger distance estimation of mixture proportions. *J. Statist. Plan. Inf.*, 48, 303-319, 1995.

[38] Yakowitz, S. J., and Sprangins, J. D. On the identifiability of finite mixtures. *Ann. Math. Statist.*, 39, 209-214, 1968.

Vita

Isidro Roberto Cruz_Medina

- **Education**

  - Ph. D., Statistics, The Pennsylvania State University, May 2001 (anticipated)
  - M. S., Statistics, Colegio de Postgraduados, Chapingo. México, September 1978.
  - B. S., Agronomy Engineer, Universidad Autónoma Chapingo. México, January 1975.

- **Professional Experience**

  - 8/99 - Present: *Instructor of Statistics*, The Pennsylvania State University.
  - 8/97 - 7/99: *Teaching Assistant*, The Pennsylvania State University.
  - 1984 - 1997: *Lecturer of Statistics*, Instituto Tecnólogico de Sonora, México.
  - 1978 - 1984: *Statistical Consultant*, Centro Nacional de Investigaciones Agricolas del Noroeste INIA. Cd. Obregón. Sonora, México.

- **Publications**

  - Cruz I. R. Wheat yield response models to nitrogen and phosphorus fertilizer for rotation experiments in the Northwest of México. *Cereal Res. Communications*, Hungary 24:2, 239-245, 1996.
  - Cruz I. R. Generalización de modelos para el análisis de la interacción genotipo-ambiente. *Rev. Fitotec. Mex.*, 15, 149-158, 1992.
  - Cruz I. R. Some exact conditional tests for the multiplicative model to explain genotype-environment interaction. *Heredity*, 69, 128-132, 1992.
  - Cruz I. R. More about the multiplicative model for the analysis of genotype-environment interaction. *Heredity*, 68, 135-140, 1992.
  - Cruz I. R. et al. Un Estimador de Regresión Múltiple. *Agrociencia*, 33, 173-192, 1978.