1

PM-1 3 1/2"x4" PHOTOGRAPHIC MICROCOPY TARGET
NBS 1010a ANSI/ISO #2 EQUIVALENT

PRECISIONSM RESOLUTION TARGETS

National Library
of Canada

Acquisitions and
Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale
du Canada

Direction des acquisitions et
des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file *Votre référence*

Our file *Notre référence*

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

Canada

Estimation Methods for Finite Mixture Distributions

by

Stephen T. Sum, B.Sc.

**A thesis submitted to
the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of
Master of Science**

**School of Computer Science
Carleton University
Ottawa, Ontario**

11 January, 1993

© copyright

1993, Stephen T. Sum

National Library
of Canada

Acquisitions and
Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale
du Canada

Direction des acquisitions et
des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file Votre référence

Our file Notre référence

The author has granted an irrevocable non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of his/her thesis by any means and in any form or format, making this thesis available to interested persons.

L'auteur a accordé une licence irrévocable et non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de sa thèse de quelque manière et sous quelque forme que ce soit pour mettre des exemplaires de cette thèse à la disposition des personnes intéressées.

The author retains ownership of the copyright in his/her thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without his/her permission.

L'auteur conserve la propriété du droit d'auteur qui protège sa thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

ISBN 0-315-84168-0

Canada

The undersigned recommend to the Faculty of Graduate Studies
and Research acceptance of the thesis

**"Estimation Methods for
Finite Mixture Distributions"**

submitted by Stephen T. Sum, B.Sc.
in partial fulfillment of the requirements for
the degree of Master of Science

Thesis Supervisor

Chair, School of Computer Science

Carleton University

28 JAN 93

Date

Abstract

A finite mixture distribution consists of the superposition of a finite number of component probability densities. Such a mixture models a population composed of two or more subpopulations mixed in fixed proportions. Mixture models find utility in situations where there is a difficulty in observing directly the underlying components of the population of interest. This thesis studies general estimation methods used to estimate the parameters of the component distributions and their mixing proportions. The primary focus is on the method of moments and the method of maximum likelihood. Solutions for these methods are derived and discussed in detail for a large class of probability densities. An empirical study of the methods is performed through the use of simulated mixture distributions. Several other existing methods are also discussed for comparison.

Acknowledgements

I wish to extend my sincere thanks to Professor John Oommen for providing me with the opportunity to carry out this work. I thank him for his guidance and his seemingly boundless patience and understanding. His infectious enthusiasm and cheerful spirit were a constant source of motivation and encouragement, without which this thesis would not have come to fruition. I would also like to express my gratitude to Dr. Radhakrishna Valiveti for graciously taking the time to review the draft of the manuscript and to provide helpful suggestions. Finally, I am indebted to Marlene Wilson for her assistance in the final stages of submitting this thesis.

Table of Contents

	Page
Acceptance Sheet	ii
Abstract.....	iii
Acknowledgements	iv
Table of Contents.....	v
List of Tables	viii
List of Figures	x
List of Acronyms.....	xiii
 CHAPTER 1: INTRODUCTION	 1
1.1 Statistical Pattern Recognition.....	2
1.2 Mixture Models	3
1.3 Parameter Estimation.....	6
1.3.1 Estimation Methods.....	7
1.4 Organization of the Thesis.....	12
 CHAPTER 2: FUNDAMENTAL CONCEPTS	 15
2.1 Introduction	15
2.2 Mixture Distributions.....	15
2.2.1 Definition.....	15
2.2.2 Estimators and their Sampling Properties	17
2.2.3 Identifiability.....	18
2.3 Exponential Family of Distributions.....	21
2.3.1 Normal Distribution.....	22
2.3.2 Exponential Distribution.....	23
2.3.3 Gamma Distribution	23
2.3.4 Weibull Distribution	25
2.3.5 Poisson Distribution	26
2.3.6 Binomial Distribution.....	26
2.3.7 Geometric distribution.....	27

	Page
2.4 Simulation Technique	28
2.4.1 Generating Random Variables.....	28
2.4.2 Generating Mixtures of Random Variables	30
2.4.3 Selecting Population Parameter Values.....	31
2.4.4 Evaluating Parameter Estimators	33
2.5 Summary	35
 CHAPTER 3: ESTIMATION USING THE METHOD OF MOMENTS.....	 37
3.1 Introduction	37
3.2 Method Of Moments	38
3.3 Difficulties.....	39
3.4 Mixtures of Normal Distributions	40
3.4.1 Explicit Solution of Two-Component Normal Mixtures	41
3.4.2 Iterative Solution of Two-Component Normal Mixtures.....	48
3.4.3 Simulation Results	52
3.5 Mixtures of Other Continuous Distributions	58
3.5.1 Exponential Mixtures.....	62
3.5.2 Gamma Mixtures.....	68
3.5.3 Weibull Mixtures.....	73
3.6 Mixtures of Discrete Distributions.....	76
3.6.1 Poisson Mixtures.....	79
3.6.2 Binomial Mixtures	84
3.6.3 Geometric Mixtures.....	90
3.7 Summary	93
 CHAPTER 4: ESTIMATION USING MAXIMUM LIKELIHOOD.....	 95
4.1 Introduction	95
4.2 Maximum Likelihood	96
4.3 Iterative Procedures	101
4.3.1 The Method of Steepest Ascent.....	102
4.3.2 Newton's Method.....	104
4.3.3 The Method of Scoring	106
4.3.4 Selecting Initial Estimates	107
4.3.5 Convergence of Iterative Procedures	109

	Page
4.4 Mixture of Distributions from the Exponential Family.....	113
4.4.1 Normal Mixtures.....	115
4.4.2 Exponential Mixtures.....	119
4.4.3 Gamma Mixtures.....	122
4.4.4 Weibull Mixtures.....	125
4.4.5 Poisson Mixtures.....	127
4.4.6 Binomial Mixtures.....	130
4.4.7 Geometric Mixtures.....	133
4.5 Difficulties.....	136
4.6 Summary.....	139
 CHAPTER 5: ESTIMATION USING OTHER METHODS.....	 141
5.1 Introduction.....	141
5.2 Moment-Like Methods.....	142
5.2.1 Finite Intervals.....	143
5.2.2 Fractional Moments.....	149
5.3 Error Minimization Methods.....	151
5.4 Bayesian Method.....	154
5.5 Summary.....	156
 CHAPTER 6: SUMMARY AND CONCLUSIONS.....	 157
 REFERENCES.....	 164

List of Tables

	Page
Table 3.1 The component means of a mixture of two normal distributions for the three cases when Δ assumed the values 1.0, 1.7, and 3.0. The two standard deviations and the mixing proportion were $(\sigma_1, \sigma_2, \alpha) = (5.79, 3.03, 0.332)$	52
Table 3.2 The means, standard deviations, and mixing proportion of a two-component normal mixture for various separation measures.	56
Table 3.2 (a) Parameters for the case when $\Delta = 1.0$	55
Table 3.2 (b) Parameters for the case when $\Delta = 1.7$	55
Table 3.2 (c) Parameters for the case when $\Delta = 3.0$	56
Table 3.3 The scale parameters and mixing proportion of a two-component exponential mixture. The estimation was performed using the MM.	66
Table 3.4 The MM estimates and true values of the scale parameters and mixing proportion of a two-component gamma mixture with shape parameter $a=2$	71
Table 3.5 The scale parameters and mixing proportion of a two-component Weibull mixture with estimates produced using the MM. The shape parameter was $a=2$ for both densities.....	74
Table 3.6 The rate parameters and mixing proportion of a two-component Poisson mixture. The estimates were obtained using the MM.	83
Table 3.7 The probabilities and mixing proportion of a two-component binomial mixture with $m=10$. The estimation was carried out using the MM.	88
Table 3.8 The probabilities and mixing proportion of a two-component geometric mixture. The estimates result from using the MM.....	91
Table 4.1 The parameters of a mixture of two normal distributions using two different sets of initial values.	117
Table 4.1 (a) The initial values are the moment estimates of Table 3.2 (b).....	117
Table 4.1 (b) The initial values for ϕ_2 and σ_1 are the moment estimates of Table 3.2 (b). The initial values for ϕ_1 , σ_2 , and α are completely arbitrary.....	117

	Page
Table 4.2 The scale parameters and mixing proportion of a two-component exponential mixture. The estimates were obtained using ML.....	120
Table 4.3 The scale parameters and mixing proportion of a two-component gamma mixture including ML estimates. The shape parameter was $a = 2$	123
Table 4.4 The initial values, ML estimates, and true values of the scale parameters and mixing proportion of a two-component Weibull mixture with shape parameter $a = 2$	126
Table 4.5 The rate parameters and mixing proportion of a two-component Poisson mixture using ML estimation.....	129
Table 4.6 The probabilities and mixing proportion of a two-component binomial mixture including ML estimates. The number of independent trials was $m = 10$ for both densities	132
Table 4.7 The probabilities and mixing proportion of a two-component geometric mixture. The estimates were obtained using the ML method.	135
Table 5.1 Error functions typically used to evaluate the squared difference between CDFs and PDFs. Note that w is an arbitrary weight function.	153

List of Figures

	Page
Figure 1.1 Pattern classification model with feature vector $\mathbf{x} = [x_1, \dots, x_d]^T$ and c feature classes.....	2
Figure 1.2 Mixture model of the distribution of the length of fish for a given age class. The two components correspond to the male and female populations.....	5
Figure 2.1 Two normal distributions and their corresponding mixture in which $\alpha=0.5$	32
Figure 3.1 (a) The individual normal distributions and the corresponding mixture for the case when $\Delta = 1.0$ The parameter values of the distribution are given in Table 3.1.....	53
Figure 3.1 (b) The individual normal distributions and the corresponding mixture for the case when $\Delta = 1.7$. The parameter values of the distribution are given in Table 3.1.....	53
Figure 3.1 (c) The individual normal distributions and the corresponding mixture for the case when $\Delta = 3.0$ The parameter values of the distribution are given in Table 3.1.....	54
Figure 3.2 The average χ^2 and EMIM metrics as a function of sample size for the normal distribution of Table 3.2(b) ($\Delta = 1.7$).....	57
Figure 3.3 The population and estimated mixture distributions of two exponential PDFs with the parameters tabulated in Table 3.3.....	67
Figure 3.4 The average χ^2 and EMIM metrics as a function of sample size for the exponential distribution of Table 3.3.....	68
Figure 3.5 The population and estimated mixture distributions of two gamma PDFs with the parameters listed in Table 3.4.	71
Figure 3.6 The average χ^2 metric as a function of sample size for the gamma distribution of Table 3.4.	72
Figure 3.7 The population and estimated mixture distributions of two Weibull PDFs with the parameters shown in Table 3.5.....	75
Figure 3.8 The average χ^2 metric as a function of sample size for the Weibull distribution of Table 3.5.....	76

	Page
Figure 3.9 The population and estimated mixture distributions of two Poisson pmfs with the parameters tabulated in Table 3.6.....	83
Figure 3.10 The average EMIM metric as a function of sample size for the Poisson distribution of Table 3.6.....	84
Figure 3.11 The population and estimated mixture distributions of two binomial pmfs with the parameters listed in Table 3.7.....	89
Figure 3.12 The average EMIM metric as a function of sample size for the binomial distribution of Table 3.7.....	89
Figure 3.13 The population and estimated mixture distributions of two geometric pmfs with the parameters indicated in Table 3.8.....	92
Figure 3.14 The average EMIM metric as a function of sample size for the geometric distribution of Table 3.8.....	92
Figure 4.1 The population and estimated mixture distributions of two normal PDFs with the parameters tabulated in Table 4.1 (a).	118
Figure 4.2 The average χ^2 metric as a function of sample size for the normal distribution of Table 4.1(a).....	118
Figure 4.3 The population and estimated mixture distributions of two exponential PDFs with the parameters tabulated in Table 4.2.	121
Figure 4.4 The average χ^2 metric as a function of sample size for the exponential distribution of Table 4.2.	121
Figure 4.5 The population and estimated mixture distributions of two gamma PDFs with the parameters listed in Table 4.3.	124
Figure 4.6 The average χ^2 metric as a function of sample size for the gamma distribution of Table 4.3.....	124
Figure 4.7 The population and estimated mixture distributions of two Weibull PDFs with the parameters shown in Table 4.4.....	126
Figure 4.8 The average χ^2 metric as a function of sample size for the Weibull distribution of Table 4.4.....	127
Figure 4.9 The population and estimated mixture distributions of two Poisson pmfs with the parameters tabulated in Table 4.5.....	129
Figure 4.10 The average EMIM metric as a function of sample size for the Poisson distribution of Table 4.5.....	130

	Page
Figure 4.11 The population and estimated mixture distributions of two binomial pmfs with the parameters listed in Table 4.6.....	132
Figure 4.12 The average EMIM metric as a function of sample size for the binomial distribution of Table 4.6.....	133
Figure 4.13 The population and estimated mixture distributions of two geometric pmfs with the parameters indicated in Table 4.7.....	135
Figure 4.14 The average EMIM metric as a function of sample size for the geometric distribution of Table 4.7.....	136

List of Acronyms

CDF	Cumulative Distribution Function
EMIM	Expected Mutual Information Measure
IID	Independent and Identically Distributed
ML	Maximum Likelihood
MM	Method of Moments
PDF	Probability Density Function
PMF	Probability Mass Function

CHAPTER 1

INTRODUCTION

One particularly intriguing application of the digital computer is in the area of pattern recognition. A pattern is loosely defined as some structured arrangement of the salient characteristics of an object or phenomenon. Pattern recognition is concerned with the machine identification of such arrangements. When the number of classes is known *a priori*, the main thrust of pattern recognition is to categorize individual objects with similar properties into one of the set of classes. This function is aptly known as pattern classification [24].

The conceptual model of pattern classification comprises two principal components. The first consists of the feature selector which measures the pertinent characteristics of the objects in question. Such measurements are usually represented in the form of a d -dimensional feature vector $\mathbf{x} = [x_1, \dots, x_d]^T$. The second component constitutes the classifier which ultimately classifies the various objects on the basis of their respective feature vectors into their proper classes [31]. The classifier decides the class-membership of an object by using a discriminant function. The function is typically formulated through a preliminary training stage during which the classifier is presented with a sample of objects of specified classes. This learning phase is commonly referred to as supervised learning and will be elaborated upon in Section 1.3. Figure 1.1 depicts a schematic diagram of the pattern classification model.

Figure 1.1: Pattern classification model with feature vector $\underline{x} = [x_1, \dots, x_d]^T$ and c feature classes.

1.1 Statistical Pattern Recognition

The ideal pattern classifier would be endowed with the ability to categorize an object into one of the possible classes with unfailing certainty. Unfortunately, such a classifier is not feasible, and consequently, a decision as to which class an object belongs to must be made with a certain probability of error. Hence, pattern classification is often viewed as a field closely related to hypothesis testing and statistical decision theory. In doing so, one provides a mathematical foundation on which the pattern classification model can be built [20].

Consider the set of c classes $\{s_1, \dots, s_c\}$ whose elements partition the object space. Then, using the terminology of statistical pattern recognition, the probability of a given object belonging to the class s_i ($i = 1, \dots, c$) is denoted by $P(s_i)$. Further, each processed object is characterized by a feature vector $\underline{x} = [x_1, \dots, x_d]^T$. The measurements of the features of an object will vary for a given class. Thus, $\underline{x}$ constitutes a random variable whose distribution is dependent on the class membership of the object. Such a distribution is termed a class-conditional probability density and is denoted by $p(\underline{x} | s_i)$.

If we adopt the Bayesian school of thought and consider $P(s_i)$ to be an *a priori* probability, then the *a posteriori* probability of the random vector $\mathbf{x}$ belonging to the class s_i is given by Bayes Rule:

$$P(s_i | \mathbf{x}) = \frac{p(\mathbf{x} | s_i) P(s_i)}{p(\mathbf{x})}, \quad (1.1)$$

where

$$p(\mathbf{x}) = \sum_{i=1}^c p(\mathbf{x} | s_i) P(s_i). \quad (1.2)$$

Thus, given the *a priori* probability $P(s_i)$ and the class-conditional probability density function (PDF) $p(\mathbf{x} | s_i)$, we have the basis for classifying a given object whose features are represented by $\mathbf{x}$. That is, given $\mathbf{x}$, the object is classified into the class s_i for which $p(s_i | \mathbf{x})$ ($i=1, \dots, c$) is maximized [24].

Note, however, that the realization of the pattern classifier depends on knowing both the *a priori* probabilities $P(s_i)$ and the conditional densities $p(\mathbf{x} | s_i)$. In reality, these are rarely known, and must therefore be estimated in some manner. These estimates can be calculated on the basis of feature measurements of a sampling of the objects of interest. The methods utilized in arriving at these estimates are many and varied. Before broaching this topic, however, we will first consider the significance of the PDF, $p(\mathbf{x})$.

1.2 Mixture Models

Expression (1.2) essentially represents a weighted average of the class-conditional densities, $p(\mathbf{x} | s_i)$, with the weights as the *a priori* class probabilities, $P(s_i)$. Such a relation, termed a mixture distribution, describes a population which comprises two or more

subpopulations combined in a fixed proportion. Thus, the generation of the random vector, $\mathbf{x}$, can be viewed as a two-step process. First, the subpopulation s_i is selected according to its proportion of the mixed population, i.e., with probability $P(s_i)$. Second, the sample measurement, $\mathbf{x}$, is obtained according to the distribution, $p(\mathbf{x}|s_i)$, of the selected subpopulation, s_i . Mixture models are utilized in situations where the sample observations can only be drawn from the whole population and not separately from the individual component populations themselves. Such cases typically occur due to a difficulty in distinguishing the underlying subpopulations forming the mixed population [48]. These situations arise frequently in practice, and hence, mixture distributions serve as statistical models in numerous applications. In addition to pattern recognition, areas in which mixture models are commonly employed include biology, chemistry, geology, medicine, and actuarial science [71].

An example taken from the discipline of biology will serve to illustrate the type of situation in which mixture models arise [27]. Biologists are often faced with the task of inferring population parameters on the basis of samples collected during the course of field experiments. A particular characteristic of the species of interest is measured for each individual in the sample. An important consideration when taking these measurements is the age group or sex of the individual, since the distribution of many characteristics is dependent on these. However, the age or sex of a species is not always readily determined when measurements are performed in the natural environment. Hence, observations must be drawn from the mixed population, where the component populations are the different age groups or sexes of the species. Figure 1.2 illustrates a typical mixture distribution representing the distribution of the length of a species of fish for a given age class. The continuous mixture is composed of two normal distributions mixed in the proportions corresponding to the relative frequencies of males and females in that age class. Such a mixture model can be used to estimate not only the component means and variances of the

lengths of male and female fish, but also the proportions of each sex in the population. Various methods for arriving at parameter estimates will be discussed in the following section.

Figure 1.2: Mixture model of the distribution of the length of fish for a given age class. The two components correspond to the male and female populations.

An example of a discrete mixture model can be drawn from the field of actuarial science. Consider, for instance, the mortality rate of seniors over a certain age. The number of deaths per day may reasonably be expected to be greater during the winter than during the milder seasons. Thus, the distribution of the number of deaths may be modeled according to a mixture of two Poisson distributions, each distribution representing a different period of the year. Schilling [63] conducted a statistical study of the number of daily deaths of London women aged 80 years or older which conforms to this type of model. By using an estimation technique for mixture distributions, he was able to estimate the mean death rates and proportions of women in the individual seasonal categories.

1.3 Parameter Estimation

Ultimately, the goal in studying distributions of the foregoing nature is to decompose the mixture population into its component densities. That is, we wish to determine the class-conditional densities, $p(\mathbf{x}|s_i)$, which represent the distributions of the individual components, as well as the class probabilities, $P(s_i)$, which indicate the proportions in which the components are present in the mixture. Most commonly, the only information available to aid in achieving such a goal is a sample of measurements representative of the population in question. If one can assume a functional form for the component densities, then the problem becomes one of estimating the parameters of the individual distributions and their mixing weights on the basis of the given sample. The probability model defined by these parameter estimates becomes our best estimate of the "true" mixture population.

Before proceeding to delineate the different existing parameter estimation methods, an important notion regarding sample labeling should be noted. A sample observation, $\mathbf{x}$, is termed labeled if the component class to which it belongs to is known, and unlabeled otherwise. In the context of pattern recognition, for example, parameter estimation using these types of samples is referred to as supervised and unsupervised learning, respectively. As one might expect, supervised learning presents considerably less difficulty than unsupervised learning [24]. In the former case, knowledge of the distribution from which the sample observation is drawn allows us to use one of the straightforward estimation methods for individual densities. In fact, the problem can simply be regarded as one of estimating the parameters of c separate single distributions, where c is the number of components in the mixture. Estimation of the mixing proportions is just a matter of noting the relative number of occurrences for each class. In the second case, however, lack of knowledge of the subpopulation of origin requires us to consider all of the mixture

subpopulations together. Parameter estimation in such a case is not a straightforward extension of the methods for a single distribution. This thesis shall be concerned with the latter problem.

1.3.1 Estimation Methods

The parameter estimation of mixtures of distributions has been the subject of considerable interest for nearly a century. Karl Pearson [53], who reported a study of mixtures of two univariate normal densities in 1894, is generally credited with being the pioneer in this area. Since that time, numerous researchers from a variety of different fields have investigated mixture distributions from both theoretical and practical perspectives (see bibliographies of [60,71]). As a result, several different techniques have been developed in order to estimate the parameters of mixtures. Some of these techniques are of general applicability, whereas others are designed for specific cases. This specificity may be with respect to a particular form of the probability densities, or with respect to particular parameters of the mixture. A majority of the literature concerns mixtures of normal distributions, which is partly a consequence of the relative importance of this distribution, and partly due to the greater difficulty presented by such mixtures. Often constraints on specific parameters of a mixture are imposed in order to simplify the estimation problem. For example, in mixtures of normal distributions, one might assume that the means or variances are equal or known. Alternatively, one might be concerned solely with estimating the mixing proportions in a mixture of known component densities. In such particular cases, either specialized or general procedures could be employed.

In the first few decades of this century, a number of procedures based on graphical techniques were formulated. These techniques, which were generally confined to mixtures of normal distributions, owed their popularity largely to the relative ease with which

parameter estimates could be effected compared to other more computationally intensive methods. However, generally speaking, the graphical methods afford only a very rough estimate of the mixture parameters. Moreover, the fit of the resulting probability model may not be adequate unless there is only minimal overlap of the component densities [26]. Nonetheless, graphical techniques have demonstrated a utility in determining whether or not the given data show evidence of a mixture, and if so, the number of components present.

Several of the graphical methods involve normal probability plots. Harding [33] and Cassie [10] reported estimation techniques which utilize plots of sample versus standard normal quantiles. In such a scheme, single normal distributions yield a straight line, whereas mixtures produce a non-linear curve. Estimates of the mixture parameters are obtained on the basis of the points of inflection of the curve. The determination of these points is typically accomplished visually, and hence, the method involves a significant amount of subjectivity. Fowlkes [29] proposed a modified logistic distribution function which can be used to calculate the inflection point in the two-component case. However, the accuracy of the resulting parameter estimates remains questionable.

A somewhat more appealing graphical approach for grouped data was later developed by Bhattacharya [5]. In this case, parameter estimates are obtained from a plot of the logarithm of the ratio of adjacent histogram class frequencies. This technique produces a series of straight lines with negative slopes where each line is associated with a different component of the mixture. A certain amount of subjectivity is also implicated here in drawing the lines through scattered sample points.

A general analytic technique used for the parameter estimation of mixtures is known as the method of moments (MM). This method is applicable to any type of probability density. Moreover, it can be employed to estimate any of the parameters of the mixture,

including the mixing proportions and the individual component parameters. Estimates are obtained through the use of a system of equations constructed by equating sample moments to their expected values. The moment equations are normally selected in increasing order from the first moment to an order equivalent to the number of parameters to be estimated. Solving the system of estimating equations for the various parameters of the mixture yields the estimates of these parameters.

As discussed earlier, many investigators focused primarily on normal mixtures. These distributions are particularly difficult to resolve using the MM due to the algebraic complexity of the problem. Pearson's work on mixtures of two univariate normal densities, for example, was carried out using the moments approach. Even for such a relatively simple mixture, calculation of the parameter estimates requires the laborious reduction of the system of equations to a ninth-degree polynomial, and subsequently determining the negative root of this nonic equation. Consequently, a number of researchers strove to simplify the problem of estimation by making assumptions about the parameters.

Other continuous as well as discrete mixtures can also be decomposed using the MM. Parameter estimation of these distributions tends to be less complicated than for normal mixtures. In particular, many different mixtures of densities from the exponential family can be resolved using the same basic solution. Typically, the moment equations can be solved by the same generic approach, varying only a factor which depends on the particular form of the component densities. In all cases, the system of estimating equations for c -component mixtures is reduced to a c -degree polynomial whose roots correspond to the parameter estimates. Although less computationally tedious than for normal mixtures, estimation of the parameters of other distributions from the exponential family is still only treated for the two-component case in most of the published work. Studies of continuous and discrete mixtures by Rider [61,62], John [40,41], and Cohen [15], are typical of this.

Blischke [7] generalized the method to c components for binomial distributions, while Everitt and Hand [26] did so for Poisson mixtures. Titterington et al. [71] mentioned some results for several general continuous and discrete distributions.

An increased accessibility to digital computers complemented by an improvement in numerical methods incited researchers to consider more closely an alternate estimation method generally applicable to mixture distributions. First proposed by Fisher [28] in the early 1900s for single distributions, the method of maximum likelihood (ML) later gained favour over the MM amongst some investigators due the properties of the resulting estimators. ML estimates are defined as the choice of parameters which maximize the joint probability density function of the mixed sample. This function, termed the likelihood function, is maximized by differentiating with respect to each of the parameters and equating the result to zero. Solving these likelihood equations for the various parameters of the mixture generates expressions for the parameter estimates.

Initially, the method of ML in the context of mixtures was principally of theoretical interest due to the daunting computations associated with the method. The mixtures studied were simplified by placing constraints on some of the parameters. For example, Rao [59] estimated the parameters of a mixture of two univariate normal densities with common variances. In time, more complicated normal mixtures ensued, such as the general c -component univariate case first reported by Hasselblad [34]. Although normal mixtures appear to be of primary interest in the literature, mixtures of other distributions from the exponential family aside from the Gaussian are similarly decomposable using the method of ML.

The likelihood equations which lead to the parameter estimates generally constitute a complicated system of coupled equations which can not be solved analytically. Consequently, iterative procedures are utilized to numerically approximate their solutions.

Starting parameter estimates must be specified for these schemes, the choice of which may be critical to the final estimates obtained.

In addition to the MM and the method of ML, several other general methods have been used to estimate the parameters of mixture distributions, albeit less extensively. Kabir [42] reported an estimation technique for general mixtures of exponential densities. A sample statistic defined over a finite interval of the sample space is equated to its expected value to set up a system of estimating equations similar to that of the method of moments. The use of fractional moments instead of the regular integer moments was investigated by Tallis and Light [67]. They tested their method on mixtures of two exponential distributions.

Several methods estimate mixture parameters by minimizing a certain error function. This error function is typically defined as the sum of squares of differences between sample and theoretical functions of the mixture. These include the moment generating function reported by Quandt and Ramsay [58], and the characteristic function studied by Kumar [45]. The minimum χ^2 method [60] which uses the cumulative distribution function also falls in this group of methods.

Another method which is conceptually appealing in the context of pattern recognition is known as Bayesian estimation. The basis of this method is the *a posteriori* probability density, $p(\underline{\Theta} | X^n)$, where $X^n = \{x_1, \dots, x_n\}$ is the set of sample observations. As the sample size increases, $p(\underline{\Theta} | X^n)$ hopefully converges to a Dirac delta function centered at the "true" parameter value, $\underline{\Theta}$ [24]. Although useful for single distributions, this method is not practically feasible in the case of mixtures due to the prohibitive amount of computation involved when dealing with unlabeled samples.

1.4 Organization of the Thesis

Despite the considerable interest manifested in mixture models over the years, there exist surprisingly few sources of literature that address the problem of general mixture distributions in a comprehensive manner. In treatments of the method of moments, a significant number of papers have confined their discussion to a single type of mixture distribution. Moreover, many authors have focused exclusively on two-component mixtures. Typically, expressions have been derived specifically for mixtures of two densities of a particular functional form. There has been little reported on the generalization of the solution to mixtures of any number of component densities of an arbitrary form. Discussions on the use of the maximum likelihood method have been geared to more general mixtures, but have principally featured the Gaussian mixture distribution.

This thesis presents an in depth study of estimation techniques for finite mixture distributions. Given the number of component densities and their functional forms, as well as a sample of observations from the population of interest, the procedure for estimating the parameters of the component densities and their mixing proportions is shown. Solutions of the method of moments and the method of maximum likelihood are derived in detail for general c -component mixture densities from the exponential family. An empirical study using simulated mixtures of these densities is performed to evaluate the properties of the estimators produced by the two methods. Several other general estimation techniques are also described for comparison.

Chapter 2 explains some of the fundamental concepts associated with the parameter estimation of mixture distributions. Finite mixture distributions and the problem of resolving these mixtures into the individual component populations through parameter estimation are formally defined. The importance of the *unique* characterization of a mixture to the feasibility of the estimation process is emphasized. Several oft-encountered

probability distributions from the exponential family are presented. These distributions include the normal, exponential, gamma, Weibull, Poisson, binomial, and geometric densities. The simulation workbench utilized for the experimental component of the thesis is also described in this chapter. In particular, the manner in which random variables and the simulated mixtures are generated is given. Also, the χ^2 test and the Expected Mutual Information Measure (EMIM) utilized to evaluate the goodness of fit of the models to the simulated data are defined.

In Chapter 3, the general estimation scheme known as the method of moments (MM) is discussed in the context of mixture distributions. The method is shown to be based on a system of estimating equations obtained by equating sample moments to their corresponding theoretical moments. The difficult problem of the parameter estimation of normal mixtures is first discussed. In particular, the derivation of the solution for Pearson's mixture of two univariate normal densities is given in detail. An iterative solution of this same mixture which reduces some of the computational burden is also given. Chapter 3 then proceeds to detail the solution of general mixtures of an arbitrary number of component distributions from the exponential family. It is shown that the parameters of many continuous and discrete distributions from this family can be estimated in a similar fashion. To our knowledge this generalization is novel. The results of simulations of two-component mixtures of all the distributions discussed in the chapter are also presented.

The second general estimation method named the method of maximum likelihood (ML) is the subject of Chapter 4. The premise of this method is shown to be that the best choice of the parameter estimate is the parameter which maximizes the probability density function of the sample. Expressions for the parameter estimates of general c -component mixtures of distributions from the exponential family are derived explicitly. It is shown that these expressions do not yield a closed form for the estimates, and hence, their solution

must be approximated numerically. Several iterative procedures for accomplishing this are given: the hill-climbing method, Newton's method, and the method of scoring. Issues associated with the numerical methods such as their starting values and convergence criteria are briefly discussed. The method of ML is investigated empirically by estimating the parameters of the same mixtures simulated in Chapter 3.

Chapter 5 describes several other general methods less commonly used in the parameter estimation of mixture distributions. Two of these methods use sample and theoretical moments in similar fashion to the MM, and hence, are grouped as moment-like methods. A number of estimation techniques can be collectively called error minimization methods. Parameter estimation using these methods is shown to be based on the minimization of the difference between sample and population functions. Lastly, Bayesian estimation is also discussed and the reasons why this method is not generally practicable for mixtures are documented.

CHAPTER 2

FUNDAMENTAL CONCEPTS

2.1 Introduction

In the present chapter some basic concepts associated with the parameter estimation of mixture distributions are introduced. These concepts form the framework within which the estimation problem is discussed at length in the ensuing chapters. Mixture distributions are defined and the general estimation problem is stated. An important consideration regarding the unique characterization of a given mixture is explained. Individual distributions from the exponential family which are commonly found in mixtures are also presented. Finally, the simulation technique for the empirical studies conducted in this thesis are described.

2.2 Mixture Distributions

2.2.1 Definition

Consider the random vector $\mathbf{X}$ whose probability distribution can be modeled according to the probability density function (PDF)

$$f(\mathbf{x}; \boldsymbol{\Theta}, \boldsymbol{\alpha}) = \sum_{i=1}^c \alpha_i f_i(\mathbf{x}; \boldsymbol{\theta}_i) \quad \mathbf{x} \in \mathbb{R}^d, \quad (2.1)$$

where the constraints on α_i ($i=1, \dots, c$), are $\alpha_i > 0$ and $\sum_{i=1}^c \alpha_i = 1$. The density function

(2.1) defines a finite mixture distribution consisting of a superposition of c component densities, $f_i(\mathbf{x}; \boldsymbol{\theta}_i)$.

In addition to its constituent functions, $f_i(\cdot)$, the mixture density is characterized by two vectors, namely, $\underline{\theta}$ and $\underline{\alpha}$. The former denotes the c -component parameter vector $\underline{\theta} = (\underline{\theta}_1, \dots, \underline{\theta}_c)$ whose elements represent the parameter vectors $\underline{\theta}_i = (\theta_{i1}, \dots, \theta_{ir})$ of the individual densities for $i = 1, \dots, c$. The vector, $\underline{\alpha} = (\alpha_1, \dots, \alpha_c)$, comprises what are alternatively known as the mixing weights, mixing parameters, or mixing proportions of the component densities forming the mixture.

Implicit in the use of the term PDF is the assumption that the component densities, $f_i(\cdot)$, are continuous functions. However, mixture distributions may also be composed of discrete component densities. In this case, we shall talk about the probability mass function (PMF) of the individual densities and denote these by $P_i(\underline{x}; \underline{\theta}_i)$.

Several notes should be made at this juncture concerning the specific class of distributions under study in this paper. First, the expression (2.1) for mixture densities actually represents a specific case of a more general definition of mixtures. Consider the parameter space Ω defining the possible values of the vectors $\underline{\theta}_i$ ($i=1, \dots, c$). Suppose we adopt the view that the mixing proportions α_i define a probability measure over Ω such that $\alpha_i = \Pr[\underline{\theta} = \underline{\theta}_i]$. Expressing this measure by the cumulative distribution $H(\underline{\theta})$, we then have an alternative expression for the mixture density given by

$$f(\underline{x}; \underline{\theta}, \underline{\alpha}) = \int_{\Omega} f(\underline{x}; \underline{\theta}) dH(\underline{\theta}). \quad (2.2)$$

If $H(\underline{\theta})$ is a continuous function and has positive probability for each of an infinite number of points $\underline{\theta}_i$, then equation (2.2) defines an infinite mixture distribution [71]. In this thesis, the discussion shall be confined to the finite case.

Second, although there is no such restriction inherent in the definitions of the mixture density (2.1) or (2.2), the component densities, $f_i(\cdot)$, are commonly assumed to

belong to the same parametric family of distributions. This is the case in most instances of practical interest, and subsequent discussion shall be predicated on this assumption.

Finally, the random variate, X , of the mixture distribution may either be a random variable or vector, depending on whether the component densities are univariate or multivariate, respectively. For the most part, we shall restrict ourselves to the univariate case, unless otherwise indicated.

2.2.2 Estimators and their Sampling Properties

In considering finite mixture densities of the form (2.1), the main objective is to estimate the relevant parameters. In doing so, two principal assumptions are made. Indeed, we shall assume that the functional form of the component densities, $f_i(\cdot)$, and their number, c , are known prior to estimation. Based on a sample $X^n = \{x_1, \dots, x_n\}$ drawn from the population of interest, the parameter vectors, θ_i , and the mixing weights, α_i , are to be estimated.

The process of inferring the population characteristics from those of the sample is known as statistical inference. The estimators resulting from this process must be evaluated in some way in order to determine how well they approximate the “true” parameters. This is accomplished by examining the sampling distributions of the estimators and the moments of these distributions. In this way, one can ascertain the long run properties of the estimators averaged over a large number of samples drawn from the population [70].

Several properties are generally accepted as being desirable for an estimator of a population parameter. These include unbiasedness, consistency, efficiency, and sufficiency. Unfortunately, estimators are rarely found to possess all of these properties.

Generally, finding a “best” estimator for any given situation, in the strict sense, is not feasible. Thus, the “best” estimator must be chosen in the context of what is most suitable for the particular problem in question. A detailed explanation of the significance of the foregoing properties can be found in most intermediate statistical texts [32,37].

A number of methods for constructing parameter estimators from the sample data have been devised. These standard techniques have been studied extensively and the properties of the estimators which they produce have been established. In this study, estimators constructed by the method of moments (MM) and maximum likelihood principle (ML) are investigated.

2.2.3 Identifiability

An aspect of particular importance when dealing with the parameter estimation of mixture distributions is the concept of identifiability. A given estimation method can not reasonably be expected to yield reliable estimators if the mixture in question is not identifiable. That is, there must exist a set of parameters, $(\underline{\Theta}, \underline{\alpha})$, which uniquely characterize the mixture distribution in order for the method to provide meaningful results.

A more formal definition of identifiability in the context of finite mixture distributions is as follows [71]: A class $\mathbf{F}$ of finite mixture densities is termed identifiable if and only if for any two $f(\underline{x}; \underline{\Theta}, \underline{\alpha})$, $f(\underline{x}; \underline{\Theta}^*, \underline{\alpha}^*) \in \mathbf{F}$, the equality

$$\sum_{i=1}^c \alpha_i f_i(\underline{x}; \underline{\theta}_i) = \sum_{j=1}^{c^*} \alpha_j^* f_j(\underline{x}; \underline{\theta}_j^*)$$

implies that $c = c^*$, and that the summations can be ordered such that $\alpha_i = \alpha_j^*$, and $\underline{\theta}_i = \underline{\theta}_j^*$ for $i, j = 1, \dots, c$.

For the most part, commonly used distributions do not pose a problem of non-identifiability. In particular, the difficulty with identifiability seldomly arises with continuous distributions except in certain simple or degenerate cases. On the other hand, discrete densities have a greater likelihood of forming non-identifiable mixtures [24].

As an illustration of a continuous case, consider the simple mixture of two uniform densities, where $U(a,b)$ denotes a uniform distribution in the interval (a,b) . From the given sample data, one might conjecture that the correct solution is given by

$$f(x;a,b,\alpha) = \frac{1}{3}U(-1,1) + \frac{2}{3}U(-2,2).$$

However, for the same data, the solution

$$f(x;a,b,\alpha) = \frac{1}{2}U(-2,1) + \frac{1}{2}U(-1,2),$$

is equally plausible. Hence, in this case the mixture can not be uniquely defined by the proper choice of parameters [26].

Generally, mixtures of discrete distributions are more troublesome. Consider a random variable X which can assume the values 0, 1, or 2, and whose distribution is given by

$$P(x;p_1,p_2,\alpha) = \alpha \binom{2}{x} p_1^x (1-p_1)^{2-x} + (1-\alpha) \binom{2}{x} p_2^x (1-p_2)^{2-x} \quad x = 0,1,2.$$

Thus, we can write the two independent equations

$$P[X=0] = \alpha (1-p_1)^2 + (1-\alpha) (1-p_2)^2,$$

$$P[X=1] = 2\alpha p_1(1-p_1) + 2(1-\alpha)p_2(1-p_2).$$

The third equation, $P[X=2]$, is of course given by $1 - P[X=0] - P[X=1]$. With only 2 independent equations, the 3 parameters (p_1, p_2, α) can not be uniquely determined given the values of $P[X=0]$, and $P[X=1]$.

The problem of non-identifiability of discrete mixtures, as exemplified above, arises because of an insufficient number of independent equations necessary for a unique solution. For m possible categories of a discrete density there are only $(m-1)$ independent equations. Hence, if there are more than $(m-1)$ unknown parameters, the lack of identifiability becomes a problem [71].

Several authors, including Yakowitz and Spragins [74], and Teicher [69], have extensively investigated the difficulty of identifiability and their work is well summarized in [71]. A common basis for determining the identifiability of a given mixture is a theorem provided by Yakowitz and Spragins [74]. This theorem is stated below:

Theorem 2.1 A necessary and sufficient condition that the class of all finite mixtures formed from the set

$$\{F(\underline{x}; \underline{\theta}), \underline{x} \in \mathbb{R}^d, \underline{\theta} \in \mathbb{R}^r\}$$

be identifiable is that the set be linearly independent over the set of real numbers, $\mathbb{R}$.

Proof: The proof of this theorem can be found in [74].

◆◆◆

In the remainder of this thesis, it shall be tacitly assumed that the mixture distributions under study are identifiable.

2.3 Exponential Family of Distributions

A large number of probability densities belong to a family of distributions known as the exponential family. These functions may be either continuous or discrete and form an important class of commonly used distributions both in theory and in practice. A detailed mathematical treatise on the exponential family is found in Barndorff-Nielsen [4]. A more informal discussion is given by Dobson [23].

Consider the random variable X whose PDF is a function of the parameter θ . Then a probability density $f(x;\theta)$ is said to belong to the exponential family if $f(x;\theta)$ has the form

$$f(x;\theta) = g(x) \exp[d(\theta)t(x) + c(d(\theta))] \quad (2.3)$$

where $g(x)$ is nonnegative and is independent of the parameter θ . In many instances it is more convenient to express a probability density from the exponential family in the above form where the reparametrization, $\theta \rightarrow d(\theta)$, is made. The parameter $d(\theta)$ is called the natural parameter of the density. This form of the PDF is equivalent to one known as the Koopman-Darmois or the Pittman-Koopman form [43].

Equation (2.3) is easily generalized to the r -parameter case where c and d are now functions of the parameter vector $\boldsymbol{\theta} = (\theta_1, \dots, \theta_r)$. Thus, $f(x;\boldsymbol{\theta})$ is given by

$$f(x; \boldsymbol{\theta}) = g(x) \exp \left[\sum_{h=1}^s d_h(\boldsymbol{\theta}) t_h(x) + c(d_1(\boldsymbol{\theta}), \dots, d_s(\boldsymbol{\theta})) \right], \quad 1 \leq r \leq s. \quad (2.4)$$

Further, if $X_1, \dots, X_n$ are independent and identically distributed (IID) random variables, each with parameter vector $\boldsymbol{\theta}$, then the joint PDF is expressed as [75]

$$f_n(x; \boldsymbol{\theta}) = \prod_{k=1}^n g(x_k) \cdot \exp \left[d_1 \sum_{k=1}^n t_1(x_k) + \dots + d_s \sum_{k=1}^n t_s(x_k) + n \cdot c(d_1(\boldsymbol{\theta}), \dots, d_s(\boldsymbol{\theta})) \right]. \quad (2.5)$$

Examples of probability functions which belong to the exponential family include the normal, exponential, gamma, Weibull, Poisson, binomial, and geometric distributions. With the exception of the binomial, mixtures of any of these densities are normally identifiable [71]. A necessary and sufficient condition for the identifiability of binomial mixtures is that $m \geq 2c-1$, where m is the number of independent trials (cf. §2.3.6) [7,69].

2.3.1 Normal Distribution

The ubiquity of the normal (or Gaussian) random variable attests to its importance in modeling random behavior in a miscellany of applications. This is particularly true due to the relevance of the normal distribution to the well-known central limit theorem. The PDF of a single univariate normal random variable is expressed as

$$f(x; \phi, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\phi)^2}{2\sigma^2}\right], \quad -\infty < x < \infty, \quad (2.6)$$

where $\phi = E[X]$ and $\sigma^2 = \text{VAR}[X]$. The Gaussian function is a bell-shaped curve centered symmetrically about its mean, ϕ .

Equation (2.6) can be written in the natural form as

$$f(x; \phi, \sigma^2) = \exp\left[\frac{\phi}{\sigma^2} x - \frac{x^2}{2\sigma^2} - \frac{\phi^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)\right] \quad (2.7)$$

where the natural parameters are given by

$$d_1(\theta) = \frac{\phi}{\sigma^2}, \quad d_2(\theta) = -\frac{1}{2\sigma^2}.$$

and the remaining parameters are

$$t_1(x) = x, \quad t_2(x) = x^2,$$

$$c(d_1, d_2) = -\frac{\phi^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2).$$

2.3.2 Exponential Distribution

The exponential random variable is typically utilized to model the time response of physical and biological systems. In particular, the exponential distribution can often be fitted to failure data [64]. If failure is the result of a random event whose probability of occurrence is proportional to the length of the time interval, X , with a constant rate of occurrence, b^{-1} , then the probability density of X is given by

$$f(x; b) = \frac{1}{b} \exp(-x/b), \quad x \geq 0, b > 0, \quad (2.8)$$

where the mean time to failure is given by $b = E[X]$.

In natural parametrization the function (2.8) takes the form

$$f(x; b) = \exp[-x/b - \ln(b)], \quad (2.9)$$

where, from (2.3), we see that $d(\theta) = -b^{-1}$, $t(x) = x$, and $c(d) = -\ln(b)$.

2.3.3 Gamma Distribution

The gamma random variable lends itself well to the modeling of numerous different distributions in various applications such as lifetime testing and queuing theory [64]. Consider, for example, a system whose failure results from the cumulative effect of n random occurrences of damage. Then the time to failure of the system is given by

$X = T_1 + \dots + T_n$ where T_i is the time interval between the i th and $(i-1)$ st occurrence of damage. If each T_i is exponentially distributed with mean b , then the total time to failure, X , has a gamma distribution expressed as

$$f(x; n, b) = \frac{b^n x^{n-1} \exp(-x/b)}{(n-1)!}, \quad 0 < x < \infty,$$

where $n > 0$ and $b > 0$ represent the shape and scale parameters, respectively. More generally, the shape parameter is a real-valued constant, $a > 0$, and so the distribution is

$$f(x; a, b) = \frac{b^a x^{a-1} \exp(-x/b)}{\Gamma(a)}, \quad 0 < x < \infty. \quad (2.10)$$

The special function, $\Gamma(z)$, is the well-known gamma function given by

$$\Gamma(z) = \int_0^{\infty} x^{z-1} \exp(-x) dx, \quad z > 0.$$

The gamma random variable can be made to adapt to many different applications simply by varying the shape and scale parameters, a and b . For $a > 1$ the gamma PDF is bell-shaped, while for $a \leq 1$ it is reverse J-shaped. Several common random variables are particular cases of the gamma random variable. For instance, setting $a=1$ yields the exponential distribution, whereas setting $a=k/2$ and $b=2$ leads to the chi-square function with k degrees of freedom [46].

The density of this member of the exponential family in the natural form (2.4) is given by

$$f(x; a, b) = \exp[-x/b + (a-1)\ln(x) - a\ln(b) - \ln(\Gamma(a))], \quad (2.11)$$

where the natural parameters are

$$d_1(\theta) = -b^{-1}, \quad d_2(\theta) = a-1.$$

Further, we have the parameters

$$t_1(x) = x, \quad t_2(x) = \ln(x),$$

$$c(d_1, d_2) = -a \ln(b) - \ln(\Gamma(a)).$$

2.3.4 Weibull Distribution

Like the exponential, and more generally the gamma random variable, the Weibull random variable is also utilized in failure models. Consider the transformation of the time random variable X such that $Z = X^a$ with $a > 0$. Further, suppose that failures occur randomly in the interval $0 < Z < \infty$. Then it can be shown [64] that X has the Weibull distribution

$$f(x; a, b) = ab^{-a} x^{a-1} \exp[-(x/b)^a], \quad x > 0, \quad (2.12)$$

where $a, b > 0$ are the usual shape and scale parameters, respectively.

The form of the Weibull PDF is similar to that of the gamma PDF, in that, for $a > 1$ the function is bell-shaped, and for $a \leq 1$ it is reverse J-shaped. In addition, letting $a=1$ in (2.12) leads to the familiar exponential function.

The probability density (2.12) can also be expressed in natural parametrization as

$$f(x; a, b) = \exp[-(x/b)^a + (a-1)\ln(x) + \ln(a) - a\ln(b)], \quad x > 0. \quad (2.13)$$

Comparing with the general form (2.4), we have

$$d_1(\theta) = -b^{-a}, \quad d_2(\theta) = (a-1),$$

$$\text{and} \quad t_1(x) = x^a, \quad t_2(x) = \ln(x),$$

$$c(d_1, d_2) = \ln(a) - a\ln(b).$$

2.3.5 Poisson Distribution

Numerous applications may be cited in which a count of the number of random occurrences of an event in a given time interval or region of space is desired. For example, one might be interested in the number of telephone calls arriving at a terminal during a certain period, or the number of bacterial colonies striking a slide [46]. Such instances can be modeled by the Poisson random variable whose PMF is defined by

$$P(x; \lambda) = \frac{\lambda^x \exp(-\lambda)}{x!}, \quad x = 0, 1, \dots \quad (2.14)$$

with the expected number of events occurring in a given time period or space specified by the parameter λ .

Expressing (2.14) using the natural parametrization, we have from (2.1),

$$P(x; \lambda) = \frac{1}{x!} \exp[x \ln(\lambda) - \lambda], \quad x = 0, 1, \dots \quad (2.15)$$

where $t(x) = x$, $c(d) = -\lambda$, and $d(\theta) = \ln(\lambda)$ is the natural parameter.

2.3.6 Binomial Distribution

The binomial random variable arises in situations which can be modeled as a set of n independent trials, each trial resulting in one of two possible outcomes. If the probability of occurrence of one outcome is p then the probability of occurrence of the other outcome is $(1-p)$. Suppose that the former outcome is termed a success and the latter a failure. Then for one arrangement, the probability of x successes and $(m-x)$ failures is given by $p^x (1-p)^{m-x}$. There are $\binom{m}{x}$ such possible arrangements, however, and so the probability of x successes in m trials is given by the binomial random variable as

$$P(x; m, p) = \binom{m}{x} p^x (1-p)^{m-x}, \quad x = 0, 1, \dots, m. \quad (2.16)$$

Determining the natural form of this member of the exponential family is a straightforward matter. According to equation (2.3) we obtain for the binomial distribution,

$$P(x; p, m) = \binom{m}{x} \exp[x \ln(p/(1-p)) + m \ln(1-p)] \quad x = 0, 1, \dots, m. \quad (2.17)$$

Here, $t(x) = x$, $c(d) = m \ln(1-p)$, and $d(\theta) = \ln(p/(1-p))$.

2.3.7 Geometric Distribution

The geometric random variable represents the discrete counterpart of the exponential random variable. It is found in applications in which one is concerned with the number of independent trials between the occurrences of a random event. Consider the probability p of occurrence of the event in question and the probability $(1-p)$ of it not occurring. Then the probability of performing x trials before the occurrence of the event is given by the PMF of the geometric random variable

$$P(x; p) = (1-p)^x p \quad x = 0, 1, \dots \quad (2.18)$$

where $0 \leq p \leq 1$.

Converting (2.18) to its natural form, we have for the geometric distribution

$$P(x; p) = \exp[x \ln(1-p) + \ln(p)], \quad x = 0, 1, \dots \quad (2.19)$$

where according to (2.3), we have $t(x) = x$, $c(d) = \ln(p)$, and $d(\theta) = \ln(1-p)$.

2.4 Simulation Technique

In the course of this thesis, an empirical study of the parameter estimation of mixture distributions was undertaken. This was achieved by computer simulation. Two different estimation techniques, namely, the method of moments and the maximum likelihood method were implemented. These methods were applied to a number of continuous and discrete distributions and their performance was evaluated.

Before embarking on a description of the specific methods, we shall first elaborate on the salient aspects of the computer simulation workbench.

2.4.1 Generating Random Variables

The computer generation of random variables can be broadly divided into two principal methods. The first, known as the transformation or inversion method, is based on the cumulative distribution function (CDF) of the random variable. The second, called the rejection method, is based on the PDF of the random variable [46].

(a) Transformation method

Suppose that U is a uniform random variable in the interval $(0,1)$. Further, suppose that $F_X(x)$ is the CDF of the random variable we wish to generate. Let

$$Z = F_X^{-1}(U)$$

be the random variable obtained by first generating U and then evaluating the inverse of the CDF $F_X(x)$ at U . The value of U will always fall in the domain of $F_X^{-1}(x)$ since, by

definition of a CDF, $0 \leq F_X(x) \leq 1$. It can be shown that Z is the desired random variable. The algorithm for the transformation method is given below:

Transformation Algorithm**Output:**

Random variable with CDF $F_X(x)$.

Method:

1. Generate $U(0,1)$
2. Determine $Z = F_X^{-1}(U)$.

End Transformation Algorithm

This method is extremely attractive in its simplicity. Unfortunately, it requires a closed expression for the inverse of the CDF, $F_X(x)$. This is not always possible, and hence, an alternate method must be found. The following technique is somewhat more complicated but will work for most commonly encountered distributions.

(b) Rejection method

Suppose that Z is the random variable of interest and that $f_X(x)$ is its PDF. Further, let W be another random variable with an easily generated PDF $f_W(x)$ such that the following inequality holds for all x :

$$kf_W(x) \geq f_X(x) \quad k > 1.$$

In other words, $kf_W(x)$ encompasses the PDF $f_X(x)$ over its entire range. Then it can be shown that the random variable we seek is obtained according to the following algorithm:

Rejection Algorithm**Output:**

Random variable with PDF $f_X(x)$.

Method:

1. Generate X_1 with PDF $f_W(x)$
2. Generate $U(0, kf_W(X_1))$
3. If $U \leq f_X(X_1)$ then

$$Z = X_1$$

Else

Reject X_1 and go to step 1.

End Rejection Algorithm

For this study, mixtures of seven different distributions from the exponential family were generated. These included four continuous density functions: normal, exponential, gamma, and Weibull; and three discrete density functions: Poisson, binomial, and geometric. The generation of these random variables was based on one of the foregoing methods. Specific algorithms for simulating distributions of each of the random variables are found in [21] and [57].

2.4.2 Generating Mixtures of Random Variables

Consider the mixture density

$$f(\mathbf{x}; \boldsymbol{\theta}, \boldsymbol{\alpha}) = \sum_{i=1}^c \alpha_i f_i(\mathbf{x}; \boldsymbol{\theta}_i),$$

where $f_i(\cdot)$ is the i th component density and α_i is the corresponding mixing proportion as defined by (2.1). A sample $X^n = \{x_1, \dots, x_n\}$ is to be drawn from the population mixture.

The generation of each mixture random variable, X_k ($k = 1, \dots, n$), consists of a two-step process. First, the component from which the sample value is to be drawn is selected according to a discrete PMF described by the probability vector $\underline{\alpha} = (\alpha_1, \dots, \alpha_c)$, where α_i is the i th mixing weight. Next, the random variable from the selected component density is generated. This process is repeated n times in order to generate the complete sample, X^n .

For the purpose of the empirical study, two-component mixtures ($c=2$) were generated in all cases.

2.4.3 Selecting Population Parameter Values

The success of a given method in estimating the population parameters with reasonable accuracy and precision is evidently influenced by the nature of the mixture. That is, one might intuitively expect that the extent of mixing will affect how well the estimators approximate the "true" parameters.

The degree of mixing can be quantified on the basis of two measures. The first is the mixing weight; that is, the proportion α_i of each component in the mixture. The second is the extent of separation of the components as measured by the generalized distance [18]

$$\Delta = \frac{|\phi_i - \phi_j|}{\max(\sigma_i, \sigma_j)} \quad i, j = 1, \dots, c,$$

where ϕ_i and σ_i are the mean and standard deviation of the i th component density, respectively. Figure 2.1 illustrates this concept of a separation distance with a plot of a mixture of two unit variance normal distributions of equal mixing weights.

Figure 2.1 Two normal distributions and their corresponding mixture in which $\alpha = 0.5$.

In our experiments, three benchmark definitions were used. In the first, the distributions were poorly separated, in the second they were moderately separated, and in the last mixture the individual distributions were well separated. The details of these distributions are found in the appropriate sections which report the experimental results.

2.4.4 Evaluating Parameter Estimators

As mentioned earlier, the parameter estimators were obtained by using two estimation methods, namely, the method of moments and the maximum likelihood method. The resulting estimators were evaluated by examining the mean and variance of their sampling distributions. These statistics were compared with the true parameters and mixing weights of the component densities.

The goodness of fit of the estimated distribution to the “true” population distribution was also evaluated by the use of two metrics, namely, the chi-square metric (χ^2) and the Expected Mutual Information Measure (EMIM).

(a) Chi-square test

The χ^2 test is a commonly used metric for determining the goodness of fit of a distribution to data. The test can be broken down into the following steps [46]:

1. Divide the sample space into K bins
2. Compute the probability, p_k , of an outcome falling in the k th bin. Then the expected number of outcomes falling in the k th bin in n trials is given by

$$m_k = np_k.$$

3. Calculate the χ^2 statistic as the weighted difference between the observed and expected number of outcomes in the k th bin:

$$D^2 = \sum_{k=1}^K \frac{(N_k - m_k)^2}{m_k},$$

where N_k is the k th observed frequency

and m_k is the k th expected frequency.

4. Assess whether or not the fit is good. A small value of D^2 indicates a good fit.

The relative size of D^2 is based on its comparison to a threshold, t_β . The particular value of the threshold is determined by the significance level, β , set by the investigator. For a large sample size, the random variable, D^2 , is approximately a chi-square random variable, X , with $K-1$ degrees of freedom. Hence, the threshold, t_β , is determined by the point at which

$$P[X \geq t_\beta] = \beta.$$

Thus, if $D^2 \geq t_\beta$, the fit is deemed unsatisfactory.

The χ^2 metric as defined above is typically utilized for comparing a “true” distribution to a given set of data. In the present study, the objective is to compare a known distribution, $f(x; \underline{\theta}, \underline{\alpha})$, to an estimated distribution, $f(x; \hat{\underline{\theta}}, \hat{\underline{\alpha}})$. The “true” PDF denotes the simulated population prescribed by the parameter vectors $\underline{\theta}$ and $\underline{\alpha}$. The estimated PDF represents the distribution characterized by the estimators $\hat{\underline{\theta}}$ and $\hat{\underline{\alpha}}$. These estimated parameter vectors are determined on the basis of sample data, $X^n = (x_1, \dots, x_n)$, drawn from the population. Hence, if the population space is subdivided into K bins, we can make use of a χ^2 -based metric defined as

$$I_\chi = \sum_{k=1}^K \frac{(p_k - \hat{p}_k)^2}{p_k},$$

where p_k is the “true” probability of an outcome falling in the k th bin

and $\hat{p}_k$ is the estimated probability of an outcome falling in the k th bin.

(b) Expected Mutual Information Measure (EMIM)

Another way of expressing the goodness of fit of an estimated distribution to the “true” population distribution is through the EMIM metric [72]. This method measures an

effective distance between the estimated and "true" densities. For continuous distributions, the EMIM metric is defined as

$$I[f(x; \underline{\Theta}), f(x; \hat{\underline{\Theta}})] = \int_{\mathbf{x}} f(x; \underline{\Theta}) \cdot \log \frac{f(x; \underline{\Theta})}{f(x; \hat{\underline{\Theta}})} d\mathbf{x},$$

where $f(x; \underline{\Theta})$ and $f(x; \hat{\underline{\Theta}})$ are the "true" and estimated mixture densities, respectively. Similarly, for discrete densities, $P(x; \underline{\Theta})$ and $P(x; \hat{\underline{\Theta}})$,

$$I[P(x; \underline{\Theta}), P(x; \hat{\underline{\Theta}})] = \sum_{\mathbf{x}} P(x; \underline{\Theta}) \cdot \log \frac{P(x; \underline{\Theta})}{P(x; \hat{\underline{\Theta}})}.$$

It can be shown that, based on the properties of PDFs (or PMFs),

$$I[f(x; \underline{\Theta}), f(x; \hat{\underline{\Theta}})] \geq 0,$$

where equality holds only when $f(x; \underline{\Theta}) = f(x; \hat{\underline{\Theta}})$ for all $\mathbf{x}$.

2.5 Summary

The foregoing discussion dealt with some fundamental concepts involved in the parameter estimation of mixture distributions. A formal definition of the mixtures studied in this work was given. The general objective of obtaining "good" parameter estimators was briefly outlined. Estimators are judged on the basis of their sampling properties as measured by their sampling distributions and the moments of these distributions. The problem of identifiability is a significant consideration in the parameter estimation of mixture densities. Meaningful estimators can not be obtained by any method if the mixture is not uniquely characterized by some set of parameters, $\{\underline{\Theta}, \underline{\alpha}\}$.

The simulation technique adopted in this paper was described. Algorithms for generating random variables largely based on two methods, namely, the transformation and rejection methods were highlighted. In the forthcoming experiments, the parameter values of the simulated populations will be selected according to a distance measure separating the components. Also, in addition to determining the sampling means and variances of the estimators, the goodness of fit of the estimated distribution to the population distribution will be evaluated by using the χ^2 and EMIM metrics.

CHAPTER 3

ESTIMATION USING THE METHOD OF MOMENTS

3.1 Introduction

The method of moments (MM) constitutes the first general method applied to the parameter estimation of mixture distributions. It was introduced almost a century ago in 1894 by Karl Pearson [53] in order to resolve a mixture of two univariate normal densities. From its inception, the method of moments remained, by far, the most popular estimation method for decades during which normal mixtures were the principal focus of study [11,15]. In 1948, Rao's [59] application of maximum likelihood estimation to a two-component normal density with equal variances signaled the first departure from the moments method as the general technique of choice. However, due to the computationally intensive nature of the maximum likelihood equations, the method of moments continued to enjoy widespread use until the common availability of electronic computers. Today the method of moments retains its appeal in that it usually yields a tractable estimator and can provide an initial approximation for more efficient iterative estimation techniques.

In the following sections, parameter estimation by the MM is explored. A general description of the method and some of its shortcomings are first briefly introduced. The method as applied to several different compound distributions is subsequently described in greater detail. In addition to a solution based on the scheme derived by Pearson for a two-component normal mixture, an alternative solution which noticeably reduces the algebraic complexity is given. Mixtures of any arbitrary number of continuous distributions from the exponential family, including the exponential, gamma, and Weibull densities, are also resolved. Similarly, the parameter estimation of multi-component discrete mixtures of

Poisson, binomial, and geometric distributions is presented. Further, as part of an empirical study, the performance of the method on simulated mixtures of the foregoing probability densities is evaluated. As outlined in Section 2.4.4, the performance measures include the chi-square test and the Expected Mutual Information Measure (EMIM), as well as the sample means and variances of the estimators.

3.2 Method Of Moments

It can be shown that the theoretical moments of a distribution are reasonably approximated by their observational counterparts when the number of sample values is sufficiently large [66]. The method of moments utilizes this fact to estimate the parameters of a distribution. Theoretical moments, which are functions of the unknown parameters, are equated to their corresponding sample moments. Estimates of the parameters are obtained by setting up as many equations as there are parameters, and solving these equations in terms of the sample moments.

Suppose that there are r parameters $\underline{\theta} = (\theta_1, \theta_2, \dots, \theta_r)$ to be estimated. Then a system of r equations is created as follows:

$$\mu_k(\hat{\underline{\theta}}) = m_k(X^n) \quad k = 1, 2, \dots, r \quad (3.1)$$

where $\mu_k(\hat{\underline{\theta}})$ is the k th population moment

$m_k(X^n)$ is the k th sample moment

$\hat{\underline{\theta}} = (\theta_1, \dots, \theta_r)$ is the estimated parameter vector

and $X^n = \{x_1, \dots, x_n\}$ is the sample data.

Expressing the above system of equations as

$$\mu(\hat{\theta}) = m(X^n), \quad (3.2)$$

the estimated parameters are obtained by inversion

$$\hat{\theta} = \mu^{-1} [m(X^n)]. \quad (3.3)$$

Of course the moment equations may not always be invertible and, in such cases, one must resort to a different estimation method. Fortunately, for mixtures of most of the common parametric densities of theoretical and practical interest, this is not a concern. The moment estimates for several such mixture densities are derived further below.

3.3 Difficulties

There are several difficulties associated with parameter estimation of compound distributions by the method of moments. First, it should be noted that population moments of all orders do not necessarily exist for all distributions [44]. In such cases, nonconsecutive and/or higher moments than the first r may be required. The danger, however, is that estimates resulting from the use of higher moments may unfortunately be suspect, due to the inherent large sampling variance of these moments [44].

Another difficulty regarding the method of moments is that solutions given by equation (3.3) may be indeterminate. That is, the moment estimates may not necessarily fall in the permissible range of values for the parameters. In particular, imaginary estimates of real-valued parameters may result, and estimates of mixing proportions may not be in the interval $(0,1)$. No general theoretical conditions exist under which estimation by the method of moments can be assured to yield determinate solutions for the general mixture problem [27].

Finally, the moment estimates are sensitive to the sampling proportions of the various components of the mixture. Components present only in very small proportions, such as in the case of outliers, may lead to poor estimates. This is of particular concern in mixtures where higher order moments are required to obtain the estimates.

3.4 Mixtures of Normal Distributions

Consider a mixture of c univariate normal distributions with component means ϕ_i and variances σ_i^2 ($i=1, \dots, c$), respectively. The PDF of such a mixture is denoted by

$$f(x; \underline{\phi}, \underline{\sigma}^2, \underline{\alpha}) = \sum_{i=1}^c \alpha_i f_i(x; \phi_i, \sigma_i^2) \quad -\infty < x < \infty \quad (3.4)$$

where α_i is the proportion of component i satisfying the conditions

$$\sum_{i=1}^c \alpha_i = 1 \quad \text{and} \quad 0 < \alpha_i < 1,$$

and $f_i(\cdot)$ ($i=1, \dots, c$) is the i th density function given by (2.6). The unknown parameter vectors of interest are $\underline{\alpha} = (\alpha_1, \dots, \alpha_{c-1})$, $\underline{\phi} = (\phi_1, \dots, \phi_c)$, and $\underline{\sigma}^2 = (\sigma_1^2, \dots, \sigma_c^2)$.

As mentioned earlier, the bulk of the work involving mixture parameter estimation by the method of moments concerns mixtures of normal distributions. This is, in part, due to the wide-ranging applicability of the Gaussian random variable, but also due to the considerably more difficult problem normal mixtures present over other compound distributions. In a c -component mixture of Gaussian densities, where none of the parameters are known in advance, there are $3c-1$ parameters to be estimated. This

represents c more unknowns than in most other types of mixtures with $2c-1$ parameters. As a result, many investigators have sought to reduce the algebraic complexity of the problem by making simplifying assumptions about the parameters.

For instance, Charlier and Wicksell [11] investigated mixtures of two univariate normals in which either the two component means are known, or the two individual variances are equal. Cohen [15] studied the symmetric cases of (i) equal means, as well as (ii) equal variances and equal proportions in two-component normal mixtures. Day [18], Tan and Chang [68] reported work on mixtures of multivariate normals with equal covariance matrices. Cooper [16] also studied a number of special cases involving normal mixtures in the context of unsupervised adaptive detection.

Because of the suspect nature of the higher sample moments, as well as the computational difficulty associated with multi-component normal mixtures, use of the method of moments for this type is commonly restricted to two-component compound distributions. As an illustration of the method, explicit and iterative solutions for a mixture of two univariate normal densities are derived next.

3.4.1 Explicit Solution of Two-Component Normal Mixtures

In the case before us, where no assumptions are made, five parameters must be estimated. In this case, the first five sample moments are equated to their corresponding theoretical moments. The resulting system of equations, after a substantial amount of algebra, leads to a ninth degree polynomial whose solution yields the desired parameter estimates. The following mathematical treatment is that of Cohen [15]

Recall that the central population moment of order k is written as

$$\mu_k = \int_{-\infty}^{\infty} (x - \mu_1')^k f(x) dx \quad k = 0, \dots, 5,$$

while the central sample moment is given by

$$m_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k \quad k = 0, \dots, 5,$$

where μ_1' and $\bar{x}$ are the population and sample means, respectively. The population mean of the mixture, μ_1' , is defined as

$$\mu_1' = \alpha\phi_1 + (1 - \alpha)\phi_2. \quad (3.5)$$

Let

$$\delta_1 = \phi_1 - \mu_1' \quad \text{and} \quad \delta_2 = \phi_2 - \mu_1'. \quad (3.6)$$

Then the first five central moments of the distribution (3.4) are given by

$$\mu_1 = \alpha\delta_1 + (1 - \alpha)\delta_2 = 0 \quad (3.7a)$$

$$\mu_2 = \alpha(\sigma_1^2 + \delta_1^2) + (1 - \alpha)(\sigma_2^2 + \delta_2^2) \quad (3.7b)$$

$$\mu_3 = \alpha\delta_1(3\sigma_1^2 + \delta_1^2) + (1 - \alpha)\delta_2(3\sigma_2^2 + \delta_2^2) \quad (3.7c)$$

$$\mu_4 = \alpha[3\sigma_1^4 + 6\sigma_1^2\delta_1^2 + \delta_1^4] + (1 - \alpha)[3\sigma_2^4 + 6\sigma_2^2\delta_2^2 + \delta_2^4] \quad (3.7d)$$

$$\mu_5 = \alpha\delta_1[15\sigma_1^4 + 10\sigma_1^2\delta_1^2 + \delta_1^4] + (1 - \alpha)\delta_2[15\sigma_2^4 + 10\sigma_2^2\delta_2^2 + \delta_2^4]. \quad (3.7e)$$

Equating theoretical and empirical moments, and rearranging, equations (3.5) and (3.7) lead to

$$\alpha[\phi_1 - \bar{x}] + (1 - \alpha)[\phi_2 - \bar{x}] = 0, \quad (3.8)$$

and

$$\alpha\delta_1 + (1 - \alpha)\delta_2 = 0 \quad (3.9a)$$

$$\alpha[\sigma_1^2 + \delta_1^2 - m_2] + (1 - \alpha)[\sigma_2^2 + \delta_2^2 - m_2] = 0 \quad (3.9b)$$

$$\alpha[3\sigma_1^2\delta_1 + \delta_1^3 - m_3] + (1 - \alpha)[3\sigma_2^2\delta_2 + \delta_2^3 - m_3] = 0 \quad (3.9c)$$

$$\alpha[3\sigma_1^4 + 6\sigma_1^2\delta_1^2 + \delta_1^4 - m_4] + (1 - \alpha)[3\sigma_2^4 + 6\sigma_2^2\delta_2^2 + \delta_2^4 - m_4] = 0 \quad (3.9d)$$

$$\alpha[15\sigma_1^4\delta_1 + 10\sigma_1^2\delta_1^3 + \delta_1^5 - m_5] + (1 - \alpha)[15\sigma_2^4\delta_2 + 10\sigma_2^2\delta_2^3 + \delta_2^5 - m_5] = 0. \quad (3.9e)$$

Estimates of the mixed distribution parameters $(\phi_1, \phi_2, \sigma_1, \sigma_2, \alpha)$ are thus obtained by the simultaneous solution of the above system of five equations.

Using equation (3.9a) and solving for α , we substitute

$$\alpha = \frac{\delta_2}{\delta_2 - \delta_1}$$

into the remaining expressions in (3.9) to reduce the system to four equations:

$$\delta_1^{-1}[\sigma_1^2 + \delta_1^2 - m_2] = \delta_2^{-1}[\sigma_2^2 + \delta_2^2 - m_2], \quad (3.10a)$$

$$\delta_1^{-1}[3\sigma_1^2\delta_1 + \delta_1^3 - m_3] = \delta_2^{-1}[3\sigma_2^2\delta_2 + \delta_2^3 - m_3], \quad (3.10b)$$

$$\delta_1^{-1}[3\sigma_1^4 + 6\sigma_1^2\delta_1^2 + \delta_1^4 - m_4] = \delta_2^{-1}[3\sigma_2^4 + 6\sigma_2^2\delta_2^2 + \delta_2^4 - m_4], \quad (3.10c)$$

$$\delta_1^{-1}[15\sigma_1^4\delta_1 + 10\sigma_1^2\delta_1^3 + \delta_1^5 - m_5] = \delta_2^{-1}[15\sigma_2^4\delta_2 + 10\sigma_2^2\delta_2^3 + \delta_2^5 - m_5]. \quad (3.10d)$$

Note that the above derivation clearly holds only for $\delta_1 \neq 0$ and $\delta_2 \neq 0$. This implies that the symmetric case $\phi_1 = \phi_2$ is not applicable here. When $\phi_1 = \phi_2$, we see from (3.6) that $\delta_1 - \delta_2 = 0$. From equation (3.7a) then $\delta_2 = 0$ and hence $\delta_1 = 0$. Consequently, we assume that $\phi_1 < \phi_2$ which implies from (3.5) and (3.6) that $\phi_1 < \mu_1 < \phi_2$ and $\delta_1 < 0 < \delta_2$. The special case of $\phi_1 = \phi_2$ is a considerably simpler problem than the one considered here and is treated by Cohen [15].

Setting equation (3.10a) equal to β , expressions for σ_1^2 and σ_2^2 can be written as

$$\sigma_1^2 = \delta_1\beta + m_2 - \delta_1^2, \quad (3.11a)$$

$$\sigma_2^2 = \delta_2\beta + m_2 - \delta_2^2. \quad (3.11b)$$

Substituting these equations into (3.10b) - (3.10d), the system becomes

$$\delta_1\delta_2[3\beta - 2(\delta_1 + \delta_2)] = -m_3, \quad (3.12a)$$

$$\delta_1\delta_2[3\beta^2 - 2(\delta_1^2 + \delta_1\delta_2 + \delta_2^2)] = -k_4, \quad (3.12b)$$

$$\begin{aligned} \delta_1\delta_2[15(\delta_1 + \delta_2)\beta^2 - 20(\delta_1^2 + \delta_1\delta_2 + \delta_2^2)\beta \\ + 6(\delta_1 + \delta_2)(\delta_1^2 + \delta_2^2)] = -k_5. \end{aligned} \quad (3.12c)$$

The above fourth and fifth order sample cumulants for moments about the mean, denoted by k_4 and k_5 respectively, are given by [44]

$$k_4 = m_4 - 3m_2^2, \quad (3.13a)$$

$$k_5 = m_5 - 10m_2m_3. \quad (3.13b)$$

Introducing the substitutions

$$r = \delta_1 + \delta_2 \quad \text{and} \quad v = \delta_1\delta_2, \quad (3.14)$$

equations (3.12) transform to

$$3\beta v - 2rv = -m_3 \quad (3.15a)$$

$$3\beta^2 v - 2v(r^2 - v) = -k_4 \quad (3.15b)$$

$$15vr\beta^2 - 20v(r^2 - v)\beta + 6vr(r^2 - 2v) = -k_5. \quad (3.15c)$$

Further, letting

$$\omega = rv \quad (3.16)$$

the system of equations (3.15) becomes

$$3\beta v - 2\omega = -m_3 \quad (3.17a)$$

$$3\beta^2 v^2 - 2\omega^2 + 2v^3 = -k_4 v \quad (3.17b)$$

$$15v^2\omega\beta^2 - 20v\omega^2\beta + 20v^4\beta + 6\omega^3 - 12v^3\omega = -k_5 v^2. \quad (3.17c)$$

Eliminating β between (3.17a) and (3.17b), as well as between (3.17a) and (3.17c), the system is reduced to

$$2\omega^2 - 6v^3 - 3k_4v + 4\omega m_3 - v_3^2 = 0. \quad (3.18a)$$

$$5\omega(2\omega - m_3)^2 - 20\omega^2(2\omega - m_3) + 20v^3(2\omega - m_3) + 18\omega^3 - 36\omega v^3 = -3k_5v^2. \quad (3.18b)$$

Introducing yet another transformation

$$z = \omega + m_3, \quad (3.19)$$

(3.18) becomes

$$2z^2 = 6v^3 + 3k_4v + 3m_3^2, \quad (3.20a)$$

$$2z^2(z - 3m_3) + z(m_3^2 - 4v^3) + 3m_3^3 + 24m_3v^3 = 3k_5v^2. \quad (3.20b)$$

Substituting the equation (3.20a) for $2z^2$ into equation (3.20b) yields

$$z = \frac{-6v^3m_3 + 3k_5v^2 + 9m_3k_4v + 6m_3^3}{2v^3 + 3k_4v + 4m_3^2}. \quad (3.21)$$

Finally, replacing z in the first expression of (3.20) by the above expression leads to the desired nonic equation

$$\sum_{i=0}^9 a_i v^i = 0 \quad (3.22)$$

where the coefficients are given by

$$\begin{aligned}
a_9 &= 24 & a_4 &= 444k_4m_3^2 - 18k_5^2 \\
a_8 &= 0 & a_3 &= 288m_3^4 - 108m_3k_4k_5 + 27k_4^3 \\
a_7 &= 84k_4 & a_2 &= -(63m_3^2k_4^2 + 72m_3^3k_5) \\
a_6 &= 36m_3^2 & a_1 &= -96m_3^4k_4 \\
a_5 &= 90k_4^2 + 72k_5m_3 & a_0 &= -24m_3^6
\end{aligned} \tag{3.23}$$

Since $\delta_1 < 0 < \delta_2$, it follows that $v = \delta_1\delta_2 < 0$. Hence, the estimate $\hat{v}$ is a negative real root of the polynomial (3.22). Determining the root of this ninth degree polynomial to obtain $\hat{v}$ then leads to $\hat{\omega}$ using (3.19) and (3.21) as

$$\hat{\omega} = \frac{-8m_3\hat{v}^3 + 3k_5\hat{v}^2 + 6k_4m_3\hat{v} + 2m_3^3}{2\hat{v}^3 + 3k_4\hat{v} + 4m_3^2}. \tag{3.24}$$

From relation (3.16), $\hat{r} = \hat{\omega} / \hat{v}$, and so

$$\hat{r} = \frac{-8m_3\hat{v}^3 + 3k_5\hat{v}^2 + 6k_4m_3\hat{v} + 2m_3^3}{\hat{v}(2\hat{v}^3 + 3k_4\hat{v} + 4m_3^2)}. \tag{3.25}$$

Since $\hat{r} = \hat{\delta}_1 + \hat{\delta}_2$ and $\hat{v} = \hat{\delta}_1\hat{\delta}_2$, where $\hat{\delta}_1 < 0 < \hat{\delta}_2$, estimates of δ_1 and δ_2 are respectively the negative and positive roots of the quadratic equation

$$\hat{\delta}^2 - \hat{r}\hat{\delta} + \hat{v} = 0, \tag{3.26}$$

and hence,

$$\hat{\delta}_1 = \frac{1}{2}(\hat{r} - \sqrt{\hat{r}^2 - 4\hat{v}}) \tag{3.27a}$$

$$\hat{\delta}_2 = \frac{1}{2}(\hat{r} + \sqrt{\hat{r}^2 - 4\hat{v}}). \tag{3.27b}$$

From (3.15), $\hat{\beta}$ can be determined, and hence, estimates of the two component standard deviations are given by (3.11) as

$$\hat{\sigma}_1^2 = \hat{\delta}_1 \hat{\beta} + m_2 - \hat{\delta}_1^2 \quad (3.28a)$$

$$\hat{\sigma}_2^2 = \hat{\delta}_2 \hat{\beta} + m_2 - \hat{\delta}_2^2. \quad (3.28b)$$

Estimates of the two component means follow from (3.6)

$$\hat{\phi}_1 = \hat{\delta}_1 + \bar{x}, \quad (3.29a)$$

$$\hat{\phi}_2 = \hat{\delta}_2 + \bar{x}. \quad (3.29b)$$

Finally, the estimate of the component proportion is obtained from (3.9a) as

$$\hat{\alpha} = \frac{\hat{\delta}_2}{\hat{\delta}_2 - \hat{\delta}_1}. \quad (3.30)$$

It should be noted that multiple solutions to the system of equations (3.9) are possible due to the existence of two or more negative real roots of the nonic equation (3.22) with some samples. Determining the appropriate root may be achieved by selecting the one for which the sixth central sample moment most closely agrees with its theoretical counterpart. Alternatively, the correct root may be chosen on the basis of a chi-square test in which observed and expected frequencies of the classed data are compared [36].

3.4.2 Iterative Solution of Two-Component Normal Mixtures

Although less of an issue today in the age of the modern computer, many early authors of the subject of normal mixtures were concerned with the daunting task of

determining the negative root of a ninth degree polynomial. Consequently, considerable effort was devoted to deriving alternate solutions to the estimation problem so as to obviate the need for the nonic equation. One such solution presented here, essentially replaces the need to solve the nonic estimating equation for v by iteratively evaluating a cubic polynomial.

Let $r(i)$ denote the i th approximation to $\hat{r}$ and $\mu_5(i)$ the corresponding approximation to the fifth population moment. The objective is to find $r(i)$ and $r(i+1)$ such that $\mu_5(i) < m_5 < \mu_5(i+1)$. Then by linear interpolation an estimate, $\hat{r}$, can be determined to obtain the parameter estimates $(\hat{\phi}_1, \hat{\phi}_2, \hat{\sigma}_1, \hat{\sigma}_2, \hat{\alpha})$ of the mixture.

A first approximation to $\hat{r}$ can readily be obtained by initially assuming that the two standard deviations are equal. Let $\sigma_1 = \sigma_2$ in (3.11). Then,

$$\sigma^2 = \delta_1\beta + m_2 - \delta_1^2 = \delta_2\beta + m_2 - \delta_2^2 \quad (3.31)$$

Effecting the substitution

$$r = \delta_1 + \delta_2 \text{ and } v = \delta_1\delta_2$$

defined in (3.14), and rearranging yields the expressions

$$\beta = r, \quad (3.32a)$$

$$\sigma^2 = v + m_2. \quad (3.32b)$$

Thus, replacing β by r in the first two equations of (3.15), the latter reduces to

$$rv = -m_3 \quad (3.33a)$$

$$r^2v + 2v^2 = -k_4. \quad (3.33b)$$

Eliminating r in (3.33) leads to the cubic estimating equation

$$2v^3 + k_4v + m_3^2 = 0, \text{ with } m_3^2 \neq 0 \quad (3.34)$$

Finally, solving for the negative real root, $v(1)$, enables us to determine from (3.33a) a first approximation to $\hat{r}$ as

$$r(1) = \frac{-m_3}{v(1)}. \quad (3.35)$$

Using the foregoing values, $v(1)$ and $r(1)$, initial approximations of the mixture parameters can be determined. Values of $\delta_1(1)$ and $\delta_2(1)$ result from expressions (3.27), and the variance $\sigma^2(1)$ is found from (3.31). Substituting these results into the fifth moment equation of (3.7) yields the first approximation $\mu_5(1)$ of the population moment.

Subsequent approximations $r(i+1)$ are selected such that the corresponding moment approximation, $\mu_5(i+1)$, satisfies $\mu_5(i) < m_5 < \mu_5(i+1)$. By assuming that r is known in the moment equations derived in Section 3.4.1 for the general case, only the first two expressions of (3.15), namely,

$$3\beta v - 2rv = -m_3,$$

$$3\beta^2 v - 2v(r^2 - v) = -k_4,$$

are required to determine a solution. Eliminating β reduces the above system to another cubic estimating equation given by

$$6v^3 - 2r^2v^2 + (3k_4 - 4rm_3)v + m_3^2 = 0. \quad (3.36)$$

Again solving for the negative real root leads to the $(i+1)$ st approximation, $v(i+1)$.

Using the approximations $r(i+1)$ and $v(i+1)$ in expressions (3.15), (3.27), (3.28), and (3.30) yields the parameters $\delta_1(i+1)$, $\delta_2(i+1)$, $\sigma_1^2(i+1)$, $\sigma_2^2(i+1)$, and $\alpha(i+1)$. The moment approximation $\mu_5(i+1)$ results from substituting these values in equation (3.7e).

The foregoing procedure using (3.36) as the estimating equation for $\hat{v}$ can be repeated until the interval $(\mu_5(i), \mu_5(i+1))$ is suitably narrow. Linear interpolation then yields the estimator $\hat{r}$ at the sample moment m_5 . Solving for the negative root of (3.36) produces $\hat{v}$ which, in conjunction with $\hat{r}$, leads to the parameter estimates $\hat{\phi}_1$, $\hat{\phi}_2$, $\hat{\sigma}_1^2$, $\hat{\sigma}_2^2$, and $\hat{\alpha}$ from the equations (3.27) to (3.30).

3.4.3 Simulation Results

The estimation of the parameters of two-component normal mixtures was effected according to the simulation workbench described in Section 2.4. The process of obtaining samples from a mixture was explained in Section 2.4.2. The five parameters of these mixtures, $(\phi_1, \phi_2, \sigma_1, \sigma_2, \alpha)$, were estimated using the explicit solution of the method of moments.

Three different separation distances between the components were selected to investigate the influence of the extent of mixing on the estimation method (cf. §2.4.3). The distances $\Delta = 1.0$, $\Delta = 1.7$, and $\Delta = 3.0$ were chosen to create mixtures of poorly separated, moderately separated, and well separated components, respectively. In each case, the standard deviations and the mixing proportion remained fixed at values $(\sigma_1, \sigma_2, \alpha) = (5.79, 3.03, 0.332)$. Only the two means were varied to achieve the different separations between the constituent densities. These means are listed in Table 3.1. For illustration, the individual component densities are plotted along with their corresponding mixture for the three cases in Figures 3.1(a) - 3.1(c).

	$\Delta = 1.0$	$\Delta = 1.7$	$\Delta = 3.0$
ϕ_1	39.51	47.72	38.02
ϕ_2	45.30	57.61	55.40

Table 3.1: The component means of a mixture of two normal distributions for the three cases when Δ assumed the values 1.0, 1.7, and 3.0. The two standard deviations and the mixing proportion were $(\sigma_1, \sigma_2, \alpha) = (5.79, 3.03, 0.332)$.

Figure 3.1 (a): The individual normal distributions and the corresponding mixture for the case when $\Delta = 1.0$. The parameter values of the distribution are given in Table 3.1.

Figure 3.1 (b): The individual normal distributions and the corresponding mixture for the case when $\Delta = 1.7$. The parameter values of the distribution are given in Table 3.1.

Figure 3.1 (c): The individual normal distributions and the corresponding mixture for the case when $\Delta = 3.0$. The parameter values of the distribution are given in Table 3.1.

In each of the above cases, the sampling distributions of the moment estimators were determined from 10 samples of size $n=1000$. The resulting means and variances of the estimates are tabulated in Tables 3.2(a) - 3.2(c).

	Mean Parameter Estimates	Variance of Parameter Estimates	True Parameter Values
ϕ_1	34.68	2.3599	39.51
ϕ_2	44.92	0.07664	45.30
σ_1	3.79	0.4723	5.79
σ_2	3.37	0.05311	3.03
α	0.1578	2.023×10^{-3}	0.332

Table 3.2 (a): Parameters for the case when $\Delta = 1.0$

	Mean Parameter Estimates	Variance of Parameter Estimates	True Parameter Values
ϕ_1	46.18	0.9824	47.72
ϕ_2	57.37	0.02972	57.61
σ_1	4.91	0.3291	5.79
σ_2	3.20	9.259×10^{-3}	3.03
α	0.2649	1.485×10^{-3}	0.332

Table 3.2 (b): Parameters for the case when $\Delta = 1.7$

	Mean Parameter Estimates	Variance of Parameter Estimates	True Parameter Values
ϕ_1	37.85	0.2140	38.02
ϕ_2	55.37	0.05091	55.40
σ_1	5.66	0.1348	5.79
σ_2	3.04	0.01304	3.03
α	0.3250	1.816×10^{-4}	0.332

Table 3.2 (c): Parameters for the case when $\Delta = 3.0$

Table 3.2: The means, standard deviations, and mixing proportion of a two-component normal mixture for various separation measures.

Examination of the results reveals that the method of moments was reasonably successful in estimating the parameters of the mixtures. Further, in all three cases, the variances of the estimates of the first component were significantly greater than the corresponding variances of the estimates of the second component. This is reasonable since the mixing weight of $\alpha = 0.332$ leads to a sample consisting of a smaller proportion of observations from the first population on which to base the estimation.

In addition, the accuracy and precision of the estimates generally improved with increasing distance between the components. Estimation of the mixture characterized by

the largest separation ($\Delta = 3.0$) yielded estimates for which the relative differences ranged from 0.054% (for ϕ_2) to 2.30% (for σ_1).

The estimators were further evaluated by determining the goodness of fit of the estimated distributions by the χ^2 and EMIM tests. Plots of these metrics as a function of sample size from $n = 100$ to $n = 1000$ are illustrated in Figure 3.2 for mixtures of moderately separated components ($\Delta = 1.7$). Each value constitutes an average of results from ten samples.

Figure 3.2: The average χ^2 and EMIM metrics as a function of sample size for the normal distribution of Table 3.2(b) ($\Delta = 1.7$).

Clearly, the fit improves considerably with increasing number of samples. From Figure 3.2 we see that the average χ^2 metric decreased from a value of 0.070 for $n = 100$ to 0.0072 for $n = 1000$. This represents a drop of approximately 90%. The EMIM metric manifested similar behavior by also falling by roughly 90% in the same interval.

3.5 Mixtures of Other Continuous Distributions

Although normal mixtures have generated the most interest, other heterogeneous continuous distributions have also been the subject of study. For example, several members of the exponential scale parameter family, namely, the exponential, gamma, and Weibull densities have been investigated in the context of the method of moments.

The general form of the single distribution in natural parametrization is given by

$$f(x; a, b) = g(x) \exp\left[\sum_{h=1}^s d_h(a, b) t_h(x) + c(d_1(a, b), \dots, d_s(a, b))\right], \quad 1 \leq s \leq 2, \quad (3.37)$$

where $a > 0$ and $b > 0$ are the called the shape and scale parameters respectively.

A general mixture of c distributions with common shape parameter, a , and component scale parameter, b_i ($i = 1, \dots, c$), has the PDF

$$f(x; a, \underline{b}, \underline{\alpha}) = \sum_{i=1}^c \alpha_i f_i(x; a, b_i) \quad x > 0 \quad (3.38)$$

where α_i ,

$$0 < \alpha_i < 1 \quad \text{and} \quad \sum_{i=1}^c \alpha_i = 1,$$

is the i th component mixing proportion, and, $f_i(\cdot)$ is the i th component density function shown in (3.37). For such a mixture there are $2c - 1$ parameters to be estimated, specifically, $\underline{\alpha} = (\alpha_1, \dots, \alpha_{c-1})$ and $\underline{b} = (b_1, \dots, b_c)$.

The parameters of compound distributions of this type are estimated using non-central moments. The k th population moment about the origin is defined as

$$\mu'_k = \int_{-\infty}^{\infty} x^k f(x) dx, \quad k = 0, 1, 2, \dots$$

The corresponding sample moment about the origin is given by

$$m'_k = \frac{1}{n} \sum_{j=1}^n x_j^k \quad k = 0, 1, 2, \dots$$

where $x_1, x_2, \dots, x_n$ are the n sample values drawn from the population of interest.

Following the method of moments for parameter estimation, the first $2c-1$ sample moments are equated to their theoretical counterparts to yield the system of equations

$$m'_k = \sum_{i=1}^c \alpha_i \int_0^{\infty} x^k f_i(x) dx \quad k = 0, 1, 2, \dots, 2c-1. \quad (3.39)$$

Evaluating the right-hand side of (3.39) leads to a system of equations linear in the mixing proportions, α_i , of the form

$$\phi_k = \sum_{i=1}^c \alpha_i b_i^k \quad k = 0, 1, \dots, 2c-1 \quad (3.40)$$

where ϕ_k is a function of the k th order sample moment, m'_k , and of the shape parameter, a .

The c scales, b_i ($i = 1, \dots, c$) are taken to be distinct.

After some algebraic manipulations, the system in (3.40) can be reduced to a single c -degree polynomial of the form

$$b^c + \beta_{c-1} b^{c-1} + \dots + \beta_1 b + \beta_0 = 0 \quad (3.41)$$

where β_i ($i=1,\dots,c$) are the constants such that the component scale parameters $b_1,\dots,b_c$ are the ordered roots of (3.41).

Now taking the sum of the product of β_k and the k th equation in (3.40) for $k=0,1,\dots,c-1$, and then adding the c th equation results in the expression

$$\sum_{k=0}^{c-1} \beta_k \phi_k + \phi_c = \sum_{k=0}^{c-1} \beta_k \left(\sum_{i=1}^c \alpha_i b_i^k \right) + \sum_{i=1}^c \alpha_i b_i^c.$$

Rearranging, we have

$$\sum_{k=0}^{c-1} \beta_k \phi_k + \phi_c = \sum_{i=1}^c \alpha_i \left(\sum_{k=0}^{c-1} \beta_k b_i^k + b_i^c \right)$$

Applying (3.41), this reduces to

$$-\phi_c = \sum_{k=0}^{c-1} \beta_k \phi_k.$$

In similar fashion, taking the sum of the product of β_k and the $(k+1)$ st equation in (3.40) for $k=0,1,\dots,c-1$, and then adding the $(c+1)$ st equation leads to

$$-\phi_{c+1} = \sum_{k=0}^{c-1} \beta_k \phi_{k+1}.$$

Proceeding in this manner, a system of c linear equations is created such that

$$-\phi_{c+h} = \sum_{k=0}^{c-1} \beta_k \phi_{k+h} \quad h = 0, 1, \dots, c-1,$$

$$\text{or} \quad \tilde{\Phi} \beta = \Phi \quad (3.42)$$

where

$$\tilde{\Phi} = \begin{bmatrix} \phi_0 & \phi_1 & \dots & \phi_{c-1} \\ \phi_1 & \phi_2 & \dots & \phi_c \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{c-1} & \phi_c & \dots & \phi_{2c-2} \end{bmatrix}, \quad \underline{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{c-1} \end{bmatrix}, \quad \underline{\phi} = \begin{bmatrix} \phi_c \\ \phi_{c+1} \\ \vdots \\ \phi_{2c-1} \end{bmatrix}.$$

Under proper conditions, it can be shown that $\tilde{\Phi}$ is non-singular [6], and hence, the coefficients β_k ($k=0, \dots, c-1$) may be determined by inverting (3.42):

$$\underline{\beta} = \tilde{\Phi}^{-1} \underline{\phi}. \quad (3.43)$$

Subsequently, the c coefficients β_k can be substituted into equation (3.41) and the estimates of the parameters b_i ($i=1, \dots, c$) determined as the roots of the polynomial. A number of root-finding routines such as Laguerre's method [55] can be used to accomplish this.

Once the estimates $\hat{b}_1, \dots, \hat{b}_c$ have been determined, the mixing proportions are obtained by solving the first c equations in (3.40), i.e.,

$$\phi_k = \sum_{i=1}^c \alpha_i b_i^k \quad k=0, 1, \dots, c-1.$$

This system may be expressed as

$$\tilde{B} \underline{\alpha} = \underline{\phi} \quad (3.44)$$

where

$$\tilde{B} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ b_1 & b_2 & \dots & b_c \\ b_1^2 & b_2^2 & \dots & b_c^2 \\ \vdots & \vdots & \ddots & \vdots \\ b_1^{c-1} & b_2^{c-1} & \dots & b_c^{c-1} \end{bmatrix}, \quad \underline{\alpha} = \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_{c-1} \end{bmatrix}, \quad \underline{\phi} = \begin{bmatrix} \phi_0 \\ \phi_1 \\ \vdots \\ \phi_{c-1} \end{bmatrix}.$$

The matrix $\tilde{B}$ is known as the Vandermonde matrix and the system can be solved by Rybicki's algorithm [55].

To recapitulate, the first $2c-1$ sample moments about the origin are first calculated. Then, using the calculated moments in (3.43) the constants β_k ($k=0, \dots, c-1$) are determined. Next, the constants are substituted into the c -degree polynomial and the c ordered roots corresponding to the parameters $0 < b_1 < b_2 < \dots < b_c$ are found. Finally, the estimated mixing proportions $\hat{\alpha}_i$ ($i=1, \dots, c$) are determined using the scale parameter estimates $\hat{b}_i$ ($i=1, \dots, c$) in (3.44). A precautionary note should be reiterated, however, concerning the fact that estimators obtained by the method of moments are not guaranteed to yield values from within the boundaries of the parameter space of the mixed distribution.

Solutions of the exponential, gamma, and Weibull mixtures will be examined in further detail in the following sections.

3.5.1 Exponential Mixtures

As described in Chapter 2, the exponential random variable commonly finds utility in failure modeling. Intuitively, one would expect mixtures of exponential distributions to arise frequently in lifetime testing since system failures are often attributable to a combination of causes. The general compound exponential PDF, $f(x; \underline{b}, \underline{\alpha})$, for such a distribution is denoted by (3.38) where the i th frequency function is given by (2.8). Using the mixed frequency function, we can proceed to estimate the $2c-1$ parameters by the method of moments.

Equating the first $2c-1$ observational and theoretical moments yields

$$m_1' = 1 \cdot \sum_{i=1}^c \alpha_i b_i,$$

$$m_2' = 2 \cdot 1 \cdot \sum_{i=1}^c \alpha_i b_i^2,$$

$$m_3' = 3 \cdot 2 \cdot 1 \cdot \sum_{i=1}^c \alpha_i b_i^3, \dots$$

or for $k = 0, 1, 2, \dots, 2c-1$,

$$m_k' = k! \cdot \sum_{i=1}^c \alpha_i b_i^k.$$

Letting $\phi_k = (m_k' / k!)$ we have the system of (3.40), i.e.,

$$\phi_k = \sum_{i=1}^c \alpha_i b_i^k \quad k=0, 1, 2, \dots, 2c-1.$$

Hence, the parameter vectors $\underline{b}$ and $\underline{\alpha}$ can be estimated by the process described in Section 3.5.

As an illustration of the estimation technique, consider a two-component exponential mixture:

$$f(x; b_1, b_2, \underline{\alpha}) = \alpha b_1^{-1} \exp(-x/b_1) + (1-\alpha) b_2^{-1} \exp(-x/b_2)$$

where $x \geq 0$ and $0 < b_1 < b_2$. The objective is to estimate the three parameters b_1 , b_2 , and α . Hence, the first three moments are required:

$$m_1' = \alpha b_1 + (1-\alpha) b_2,$$

$$\frac{m_2'}{2} = \alpha b_1^2 + (1-\alpha) b_2^2,$$

$$\frac{m_3'}{6} = \alpha b_1^3 + (1-\alpha) b_2^3.$$

Letting $\phi_k = (m_k' / k!)$ for $k=1,2,3$ and rearranging we have

$$\alpha(b_1 - b_2) = \phi_1 - b_2, \quad (3.45a)$$

$$\alpha(b_1^2 - b_2^2) = \phi_2 - b_2^2, \quad (3.45b)$$

$$\alpha(b_1^3 - b_2^3) = \phi_3 - b_2^3. \quad (3.45c)$$

We now seek to reduce the above system to a quadratic equation of the form of (3.41). Replacing $\alpha(b_1 - b_2)$ by $(\phi_1 - b_2)$ in the last two equations of (3.45) yields

$$\phi_1(b_1 + b_2) - b_1 b_2 = \phi_2,$$

$$\phi_1[(b_1 + b_2)^2 - b_1 b_2] - b_1 b_2(b_1 + b_2) = \phi_3.$$

Then substituting $b_1 b_2 = \phi_1(b_1 + b_2) - \phi_2$ from the first equation into the second, and solving for $(b_1 + b_2)$ we obtain

$$(b_1 + b_2) = \frac{\phi_3 - \phi_1 \phi_2}{\phi_2 - \phi_1^2}.$$

Hence,

$$b_1 b_2 = \frac{\phi_1 \phi_3 - \phi_2^2}{\phi_2 - \phi_1}.$$

Letting $\beta_0 = b_1 b_2$ and $\beta_1 = -(b_1 + b_2)$ we then see that

$$b^2 + \beta_1 b + \beta_0 = 0.$$

which is the quadratic of interest.

Thus, the two means are given by [†]

$$\hat{b}_{1,2} = \frac{-\beta_1 \pm \sqrt{\beta_1^2 - 4\beta_0}}{2} \quad (3.46)$$

In the two-component case the mixing proportion is most readily obtained using equation (3.45a)

$$\hat{\alpha} = \frac{\phi_1 - \hat{b}_2}{b_1 - \hat{b}_2}. \quad (3.47)$$

Thus, the three mixture parameters b_1 , b_2 , and α can be estimated by using the estimators (3.46) and (3.47).

As discussed previously, the estimators obtained by the method of moments may be indeterminate. That is, any of the above estimates may be negative or even complex. In

[†] We assume here that $b_1 < b_2$. The only constraint is that b_1 and b_2 be distinct.

the two-component case, Everitt and Hand [26] show that the conditions for $\hat{b}_1$ and $\hat{b}_2$ to be real and positive are:

(i) $\beta_1^2 > 4\beta_0$, and,

(ii) the sequence $1/\phi_1, \phi_1/\phi_2, \phi_2/\phi_3$ must be monotonically increasing or decreasing.

3.5.1.1 Simulation Results

The parameter estimation of two-component exponential mixtures was also investigated using the method of moments. The three parameters, (b_1, b_2, α) , were all assumed to be unknown. Although experiments were carried out using various separation measures, only the results for the case when $\Delta = 0.9$ are shown here, in the interest of brevity.

Table 3.3 lists the sampling distribution of the estimates of the three parameters of the mixture. Ten samples of 1000 observations were used.

	Mean Parameter Estimates	Variance of Parameter Estimates	True Parameter Values
b_1	0.4393	2.903×10^{-3}	0.45
b_2	3.3316	1.2378	3.29
α	0.8804	2.658×10^{-3}	0.90

Table 3.3: The scale parameters and mixing proportion of a two-component exponential mixture. The estimation was performed using the MM.

Figure 3.3: The population and estimated mixture distributions of two exponential PDFs with the parameters tabulated in Table 3.3.

From the table, we see that the mean estimates of the exponential mixture parameters were relatively close to the true parameter values. The small sampling proportion of the second population, however, resulted in a fairly large variance in the estimate of the second scale parameter. Figure 3.3 compares graphically the true and estimated mixture distributions.

The χ^2 and EMIM metrics, as averaged for 10 samples of 10 different sizes ranging from 100 to 1000 observations, are illustrated in Figure 3.4. From this figure, the increased accuracy of the moment estimates based on larger samples is evident. The χ^2 value decreased by about 85%, while the EMIM value fell by roughly 92% in increasing the sample size from $n = 100$ to $n = 1000$.

Figure 3.4: The average χ^2 and EMIM metrics as a function of sample size for the exponential distribution of Table 3.3.

3.5.2 Gamma Mixtures

The general mixture density of c component gamma distributions with scale parameter $b_i > 0$ ($i=1, \dots, c$) and common shape parameter $a > 0$ is defined by (3.38) with $f_i(\cdot)$ denoted by (2.10):

$$f(x; a, b) = \frac{x^{a-1} \exp(-x/b)}{b^a \Gamma(a)}, \quad 0 < x < \infty.$$

The resulting compound distribution is utilized in determining the moment estimates of the $2c-1$ parameters.

Setting the first few sample moments equal to their expected values we have

$$m_1 = a \sum_{i=1}^c \alpha_i b_i,$$

$$m_2' = a(a+1) \sum_{i=1}^c \alpha_i b_i^2,$$

$$m_3' = a(a+1)(a+2) \sum_{i=1}^c \alpha_i b_i^3, \dots$$

The emerging pattern for the k th moment equation is thus

$$m_k' = a(a+1)\cdots(a+k-1) \cdot \sum_{i=1}^c \alpha_i b_i^k, \quad k = 0, 1, 2, \dots, 2c-1.$$

Letting $\phi_k = m_k' \{a(a+1)\cdots(a+k-1)\}^{-1}$ we are led to the familiar system of equations of (3.40),

$$\phi_k = \sum_{i=1}^c \alpha_i b_i^k, \quad k=0, 1, 2, \dots, 2c-1.$$

Thus, the mixing proportions $\underline{\alpha} = (\alpha_1, \dots, \alpha_{c-1})$ and the scales $\underline{b} = (b_1, \dots, b_c)$ can be estimated as shown in Section 3.5. As mentioned earlier, the exponential random variable is simply the gamma random variable with shape parameter $a=1$. Consequently, using the above expression for ϕ_k with $a=1$ yields exactly the same results as those of the preceding section.

To obtain explicit expressions for the two-component case, consider the PDF of a mixture of two gamma distributions with known shape parameter, a ,

$$f(x; a, b_1, b_2, \underline{\alpha}) = \alpha \frac{x^{a-1} \exp(-x/b_1)}{b_1^a \Gamma(a)} + (1-\alpha) \frac{x^{a-1} \exp(-x/b_2)}{b_2^a \Gamma(a)},$$

with $x > 0$ and $0 < b_1 < b_2$.

Three parameters b_1 , b_2 , and α are to be estimated and so again the first three non-central moments are needed:

$$\phi_1 = \alpha b_1 + (1-\alpha) b_2,$$

$$\phi_2 = \alpha b_1^2 + (1-\alpha) b_2^2,$$

$$\phi_3 = \alpha b_1^3 + (1-\alpha) b_2^3,$$

where we have defined $\phi_k = m_k' \{a(a+1)\dots(a+k-1)\}^{-1}$ for $k=1, 2, 3$. These three equations can easily be rearranged into exactly the same form as the system of (3.45). Hence, proceeding as for the exponential distribution, estimates of the mixture parameters b_1 , b_2 and α are found using equations (3.46) and (3.47).

3.5.2.1 Simulation Results

In using the method of moments to estimate the parameters of two-component gamma mixtures, it was assumed that the shape parameter, a , was known and common to both components. Thus, the two scale parameters b_1 and b_2 , and the mixing weight, α , were to be estimated.

The results given in Table 3.4 are for component densities with a shape parameter of $a=2$ and a distance of separation of $\Delta = 1.0$.

It should be noted that although the final estimates were reasonably close to the true parameters, the method of moments was somewhat unreliable in resolving gamma mixtures. On several occasions, indeterminate estimates were produced by the method. That is, the component density parameter estimates, $\hat{b}_1$, $\hat{b}_2$, were found to be negative.

Consequently, a considerable number of simulated samples were rejected and the estimates derived from them were not included in the sampling distribution.

	Mean Parameter Estimates	Variance of Parameter Estimates	True Parameter Values
b_1	0.1937	8.445×10^{-5}	0.20
b_2	0.7120	1.281×10^{-2}	0.80
α	0.8571	2.658×10^{-3}	0.90

Table 3.4: The MM estimates and true values of the scale parameters and mixing proportion of a two-component gamma mixture with shape parameter $\alpha = 2$.

Figure 3.5: The population and estimated mixture distributions of two gamma PDFs with the parameters listed in Table 3.4.

As for the simulation of exponential mixtures, the estimate of the second scale parameter was noticeably poorer than that for the first component due to the smaller number of observations drawn from the second population. This larger deviation from the true parameter value is not evident in the plots of the theoretical and estimated mixture densities in Figure 3.5

The average χ^2 metric was calculated for each of 10 different sample sizes of $n = 100$ to $n = 1000$ observations. As shown in Figure 3.6, the goodness of fit improved with increasing sample size. The approximately 75% drop in χ^2 value was noticeably smaller and less rapid than the 85% decrease evinced by the exponential mixture with a comparable degree of mixing.

Figure 3.6: The average χ^2 metric as a function of sample size for the gamma distribution of Table 3.4.

3.5.3 Weibull Mixtures

Mixtures of c Weibull distributions in the proportions α_i ($i=1, \dots, c$) are defined by the density function of (3.38) where the i th component function, $f_i(\cdot)$, is denoted by

$$f(x; a, b) = ab^{-a} x^{a-1} \exp[-(x/b)^a] \quad x > 0.$$

As usual, the $2c-1$ parameters of the compound distribution can be estimated by the method of moments.

Setting up the $2c-1$ moment equations as before we have

$$m'_1 = \Gamma(1 + \frac{1}{a}) \sum_{i=1}^c \alpha_i b_i,$$

$$m'_2 = \Gamma(1 + \frac{2}{a}) \sum_{i=1}^c \alpha_i b_i^2,$$

$$m'_3 = \Gamma(1 + \frac{3}{a}) \sum_{i=1}^c \alpha_i b_i^3, \dots$$

$$\text{i.e.,} \quad m'_k = \Gamma(1 + \frac{k}{a}) \sum_{i=1}^c \alpha_i b_i^k, \quad k = 0, 1, 2, \dots, 2c-1.$$

Defining $\phi_k = m'_k \left\{ \Gamma(1 + \frac{k}{a}) \right\}^{-1}$ we obtain

$$\phi_k = \sum_{i=1}^c \alpha_i b_i^k, \quad k=0, 1, 2, \dots, 2c-1,$$

which is again the system of equations of (3.40). Hence, employing the same process detailed earlier, estimates $\hat{\alpha} = (\hat{\alpha}_1, \dots, \hat{\alpha}_{c-1})$ and $\hat{b} = (\hat{b}_1, \dots, \hat{b}_c)$ can be determined. In

particular, estimates for the two-component case are obtained by use of the expressions (3.46) and (3.47). Note that with $a=1$ the gamma function in the moment equations becomes a function of a positive integer, yielding $\Gamma(1+k) = k!$. As a result, ϕ_k reduces to the same expression as that obtained for the exponential mixture in Section 3.5.1.

3.5.3.1 Simulation Results

The estimation of the parameters of Weibull mixtures by the method of moments was carried out in similar fashion to that of the gamma mixtures. The shape parameter was the same for both Weibull components and was known prior to estimation. Ten different samples of 1000 values were generated for the experiments.

	Mean Parameter Estimates	Variance of Parameter Estimates	True Parameter Values
b_1	0.6863	5.368×10^{-5}	0.69
b_2	1.6950	9.581×10^{-3}	1.69
α	0.8979	1.782×10^{-4}	0.90

Table 3.5: The scale parameters and mixing proportion of a two-component Weibull mixture with estimates produced using the MM. The shape parameter was $a=2$ for both densities.

The investigation of a mixture with shape parameter $a = 2$, scale parameters $(b_1, b_2) = (0.69, 1.69)$, and mixing weight $\alpha = 0.90$, yielded the estimates shown in Table 3.5. The components were moderately separated by a distance measure of $\Delta = 1.7$. The resulting estimates were quite good for all three parameters. The largest deviation was for the estimate of the second scale parameter which had a mean value of 1.6950 and a variance of only 9.581×10^{-3} .

Figure 3.7 illustrates the close match between the theoretical and estimated mixture densities.

Figure 3.7: The population and estimated mixture distributions of two Weibull PDFs with the parameters shown in Table 3.5.

The determination of the χ^2 values further confirmed the good fit of the estimated distribution to the true population (Fig.3.8). In particular, for a sample size of $n = 1000$, the χ^2 metric was merely 2.10×10^{-3} .

Figure 3.8: The average χ^2 metric as a function of sample size for the Weibull distribution of Table 3.5.

3.6 Mixtures of Discrete Distributions

A number of investigators have also studied the parameter estimation of discrete distributions by the method of moments. In particular, binomial [6,7,62] and Poisson [26,62] mixtures have been of interest.

The general PMF of a mixture of c discrete distributions with unknown component parameters, $0 \leq p_i \leq 1$, and mixing proportions $0 < \alpha_i < 1$, $\sum_{i=1}^c \alpha_i = 1$, is given by

$$P(x; \underline{p}, \underline{\alpha}) = \text{Prob}[X_j = x] = \sum_{i=1}^c \alpha_i P_i(x; p_i) \quad x = 0, 1, \dots \quad (3.48)$$

where $X_1, \dots, X_n$ are n independent, identically distributed (IID) random variables each with PMF $P_i(\cdot)$. Such a mixture is characterized by a total of $2c-1$ parameters, that is, $\underline{\alpha} = (\alpha_1, \dots, \alpha_{c-1})$ and $\underline{p} = (p_1, \dots, p_c)$.

The use of factorial rather than regular moments for parameter estimation of discrete mixtures greatly simplifies the moment equations. The k th order theoretical factorial moment about the origin is defined as [44]

$$\mu'_{[k]} = \sum_{j=1}^{\infty} x_j(x_j-1)\dots(x_j-k+1) f(x_j) \quad k=1,2,\dots$$

where $\mu'_{[0]}=1$, while its observational analog is given by

$$m'_{[k]} = \frac{1}{n} \sum_{j=1}^n x_j(x_j-1)\dots(x_j-k+1) \quad k=1,2,\dots; \quad m'_{[0]}=1.$$

For mixtures of discrete distributions of the form of (3.48), the population factorial moment can be expressed as a linear equation in the mixing proportions. Thus, parameter estimation is effected by substituting the first $2c-1$ sample moments for their theoretical counterparts to create a system of equations analogous to that of (3.40):

$$\phi_k = \sum_{i=1}^c \alpha_i p_i^k, \quad k=0,1,2,\dots,2c-1, \quad (3.50)$$

where ϕ_k is a function of the sample factorial moment of order k and of some known distribution parameter, ρ , common to all components. The c parameters $p_i > 0$ ($i=1,\dots,c$) are assumed to be distinct.

Following the same process as described earlier for continuous distributions (cf. §3.5), the estimators $\hat{p}_1, \dots, \hat{p}_c$ can be obtained as the roots of a polynomial of the form

$$p^c + \beta_{c-1}p^{c-1} + \dots + \beta_1p + \beta_0 = 0 \quad (3.51)$$

The constants β_i ($i=0,\dots,c-1$) are determined from an equation similar to (3.42),

$$\underline{\beta} = \tilde{\Phi}^{-1} \underline{\phi}, \quad (3.52)$$

$$\text{where} \quad \tilde{\Phi} = \begin{bmatrix} \phi_0 & \phi_1 & \dots & \phi_{c-1} \\ \phi_1 & \phi_2 & \dots & \phi_c \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{c-1} & \phi_c & \dots & \phi_{2c-2} \end{bmatrix}, \quad \underline{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{c-1} \end{bmatrix}, \quad \underline{\phi} = \begin{bmatrix} \phi_c \\ \phi_{c+1} \\ \vdots \\ \phi_{2c-1} \end{bmatrix}.$$

using the factorial moment functions, ϕ_k , defined above for $k=c, \dots, 2c-1$. The estimated mixing proportions, $\hat{\alpha}_1, \dots, \hat{\alpha}_c$, result from solving an equation analogous to (3.44) with the estimators $\hat{p}_i$ ($i=1, \dots, c$) and factorial sample moments, ϕ_k , for $k=0, \dots, c-1$:

$$\bar{P} \underline{\alpha} = \underline{\phi}, \quad (3.53)$$

$$\text{where} \quad \bar{P} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ p_1 & p_2 & \dots & p_c \\ p_1^2 & p_2^2 & \dots & p_c^2 \\ \vdots & \vdots & \ddots & \vdots \\ p_1^{c-1} & p_2^{c-1} & \dots & p_c^{c-1} \end{bmatrix}, \quad \underline{\alpha} = \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_{c-1} \end{bmatrix}, \quad \underline{\phi} = \begin{bmatrix} \phi_0 \\ \phi_1 \\ \vdots \\ \phi_{c-1} \end{bmatrix}.$$

Note once again that the previous caveats concerning the validity of the estimates resulting from moment estimation are equally applicable to discrete mixtures (cf. §3.3). That is, the estimated mixing proportion, $\hat{\alpha}_i$, may not fall in the interval (0,1) and the $\hat{p}_i$ may be negative, imaginary, or otherwise outside their defined parameter space.

The following sections illustrate the technique for mixtures of Poisson, binomial and geometric distributions.

3.6.1 Poisson Mixtures

The PMF of a c -component mixture of Poisson distributions is given by equation (3.48) where $P_i(\cdot)$ is the i th component density (2.14). The parameter vectors, $\underline{\alpha} = (\alpha_1, \dots, \alpha_{c-1})$ and $\underline{\lambda} = (\lambda_1, \dots, \lambda_c)$, are estimated by use of the factorial moments.

Recalling the k th sample factorial moment

$$m'_{[k]} = \frac{1}{n} \sum_{j=1}^n x_j(x_j-1)\dots(x_j-k+1), \quad k = 1, 2, \dots$$

the expected value is given by

$$\begin{aligned} \mu'_{[k]} &= E[m'_{[k]}] = \sum_{x=0}^{\infty} x(x-1)\dots(x-k+1) \sum_{i=1}^c \alpha_i \frac{\lambda_i^x \exp(-\lambda_i)}{x!} \\ &= \sum_{i=1}^c \alpha_i \exp(-\lambda_i) \lambda_i^k \sum_{x=0}^{\infty} \frac{1}{x!} \frac{d^k \lambda_i^x}{d\lambda_i^k} \\ &= \sum_{i=1}^c \alpha_i \exp(-\lambda_i) \lambda_i^k \frac{d^k \exp(\lambda_i)}{d\lambda_i^k} \\ &= \sum_{i=1}^c \alpha_i \exp(-\lambda_i) \lambda_i^k \exp(\lambda_i) \\ &= \sum_{i=1}^c \alpha_i \lambda_i^k \quad k = 1, 2, \dots \end{aligned}$$

Equating the first $2c-1$ observed and theoretical moments we obtain

$$m'_{[k]} = \sum_{i=1}^c \alpha_i \lambda_i^k \quad k=0,1,2,\dots,2c-1.$$

Setting $\phi_k = m'_{[k]}$, the resulting system of equations is of the form of (3.50).

Consequently, the parameter estimates given by $\hat{\alpha}=(\hat{\alpha}_1,\dots,\hat{\alpha}_{c-1})$ and $\hat{\lambda}=(\hat{\lambda}_1,\dots,\hat{\lambda}_c)$ follow as previously with the appropriate factorial moments and estimators for this particular mixture.

Another approach for constructing the estimators is derived by Rider [62] for a mixture of two Poisson densities. In this solution, ordinary non-central moments defined in Section 3.5 are utilized instead of factorial moments. The resulting coefficients of the estimating polynomial are given by slightly more complicated expressions than those obtained with factorial moments.

The mixed distribution with means $0 < \lambda_1 < \lambda_2$ is expressed as

$$P(x; \lambda_1, \lambda_2, \alpha) = \alpha \frac{\lambda_1^x \exp(-\lambda_1)}{x!} + (1 - \alpha) \frac{\lambda_2^x \exp(-\lambda_2)}{x!}$$

The first three moment equations are thus obtained as

$$m'_1 = \alpha \lambda_1 + (1 - \alpha) \lambda_2,$$

$$m'_2 = \alpha(\lambda_1 + \lambda_1^2) + (1 - \alpha)(\lambda_2 + \lambda_2^2),$$

$$m'_3 = \alpha(\lambda_1 + 3\lambda_1^2 + \lambda_1^3) + (1 - \alpha)(\lambda_2 + 3\lambda_2^2 + \lambda_2^3).$$

Upon rearranging we have

$$\alpha(\lambda_1 - \lambda_2) = m'_1 - \lambda_2, \quad (3.54a)$$

$$\alpha(\lambda_1 - \lambda_2)(\lambda_1 + \lambda_2 + 1) = m_2' - \lambda_2 - \lambda_2^2, \quad (3.54b)$$

$$\alpha(\lambda_1 - \lambda_2)[1 + 3(\lambda_1 + \lambda_2) + \lambda_1^2 + \lambda_1\lambda_2 + \lambda_2^2] = m_3' - \lambda_2 - 3\lambda_2^2 - \lambda_2^3 \quad (3.54c)$$

Replacing $\alpha(\lambda_1 - \lambda_2)$ in the last two equations above by (3.54a) yields

$$m_1'[1 + (\lambda_1 + \lambda_2)] - \lambda_1\lambda_2 = m_2' \quad (3.55a)$$

$$m_1'[1 + 3(\lambda_1 + \lambda_2) + (\lambda_1 + \lambda_2)^2 - \lambda_1\lambda_2] - \lambda_1\lambda_2[3 + (\lambda_1 + \lambda_2)] = m_3' \quad (3.55b)$$

Substituting $\lambda_1\lambda_2 = m_1'[1 + (\lambda_1 + \lambda_2)] - m_2'$ from equation (3.55a) into (3.55b), and

solving for $(\lambda_1 + \lambda_2)$ leads to

$$(\lambda_1 + \lambda_2) = \frac{m_1'^2 + 2m_1' - m_1'm_2' - 3m_2' + m_3'}{m_2' - m_1'^2 - m_1'}.$$

Hence,

$$\lambda_1\lambda_2 = \frac{m_1'^2 - m_1'm_2' + m_1'm_3' - m_2'^2}{m_2' - m_1'^2 - m_1'}.$$

If we let $\beta_0 = \lambda_1\lambda_2$ and $\beta_1 = -(\lambda_1 + \lambda_2)$ then the means λ_1, λ_2 are given by the roots of the quadratic $\lambda_2 + \beta_1\lambda + \beta_0 = 0$. That is,

$$\hat{\lambda}_{1,2} = \frac{-\beta_1 \pm \sqrt{\beta_1^2 - 4\beta_0}}{2}. \quad (3.56)$$

The mixing proportion estimate is found using (3.54a)

$$\hat{\alpha} = \frac{m_1' - \hat{\lambda}_2}{\hat{\lambda}_1 - \hat{\lambda}_2} \quad (3.57)$$

The estimates of Poisson distributions suffer from the same lack of certainty regarding their validity as do the estimates of the previously discussed continuous distributions.

3.6.1.1 Simulation Results

Mixtures of two Poisson distributions were simulated using the workbench detailed in Section 2.4. The three parameter estimators, $(\hat{\lambda}_1, \hat{\lambda}_2, \hat{\alpha})$, were constructed using factorial moments calculated from a sample of 1000 observations.

In the following table, the results for a mixture of two components separated by a distance of $\Delta = 1.3$ are presented. The sampling distribution was derived from ten different samples. Similar difficulties with indeterminate estimates as those encountered with gamma mixtures arose with Poisson distributions. A number of samples yielded negative values for component density estimates. These estimates were excluded from the results, which were otherwise reasonably good. A histogram representing the true and estimated Poisson distributions is shown in Figure 3.9 for comparison.

2 of/de 2

PM-1 3½"x4" PHOTOGRAPHIC MICROCOPY TARGET
NBS 1010a ANSI/ISO #2 EQUIVALENT

PRECISIONSM RESOLUTION TARGETS

	Mean Parameter Estimates	Variance of Parameter Estimates	True Parameter Values
λ_1	0.4583	6.315×10^{-3}	0.50
λ_2	2.4138	0.1050	2.50
α	0.7776	3.825×10^{-3}	0.80

Table 3.6: The rate parameters and mixing proportion of a two-component Poisson mixture. The estimates were obtained using the MM.

Figure 3.9: The population and estimated mixture distributions of two Poisson PMFs with the parameters tabulated in Table 3.6.

The average EMIM metric was calculated for a range of sample sizes to evaluate the goodness of fit of the estimated PMF (Fig. 3.10). The fit improved measurably in

increasing the sample size from $n = 100$ to $n = 300$. The EMIM value was observed to drop by about 62% in this short interval. In contrast, the EMIM metric remained more or less constant for sample sizes $n = 700$ to $n = 1000$.

Figure 3.10: The average EMIM metric as a function of sample size for the Poisson distribution of Table 3.6.

3.6.2 Binomial Mixtures

General mixtures of c densities, $P_i(\cdot)$ defined by (2.16), are expressed by the compound PMF (3.48). The number of trials, m , is assumed to be known and common to all component distributions. Hence, binomial mixtures involve the estimation of $2c-1$ parameters, namely, the individual probabilities of success, $p_1, \dots, p_c$ and the mixing proportions $\alpha_1, \dots, \alpha_{c-1}$.

In the case of the binomial mixture, the theoretical factorial moment is given by

$$\mu'_{[k]} = m(m-1)\cdots(m-k+1) \cdot \sum_{i=1}^c \alpha_i p_i^k, \quad k = 1, 2, \dots$$

Applying the moments method for the first $2c-1$ moments we obtain

$$\phi_k = \sum_{i=1}^c \alpha_i p_i^k, \quad k = 0, 1, \dots, 2c-1$$

where $\phi_k = m_{[k]} \cdot \{m(m-1)\cdots(m-k+1)\}^{-1}$. This, of course, leads to the familiar system of equations (3.50) needed to estimate the mixing proportions, $(\alpha_1, \dots, \alpha_{c-1})$, and the probabilities, $(p_1, \dots, p_c)$.

Rider [62] also describes a solution for the two-component binomial mixture given by

$$P(x; p_1, p_2, \alpha) = \binom{m}{x} [\alpha p_1^x (1-p_1)^{m-x} + (1-\alpha) p_2^x (1-p_2)^{m-x}] \quad x=0, 1, \dots, m$$

where $0 \leq p_1 < p_2$. The author's use of ordinary rather than factorial moments leads to a somewhat more involved construction of the three parameter estimators, p_1 , p_2 , and α .

Setting up the first three moment equations we have

$$m_1' = \alpha m p_1 + (1-\alpha) m p_2,$$

$$m_2' = \alpha [m p_1 + m(m-1) p_1^2] + (1-\alpha) [m p_2 + m(m-1) p_2^2],$$

$$m_3' = \alpha [m p_1 + 3m(m-1) p_1^2 + m(m-1)(m-2) p_1^3] + (1-\alpha) [m p_2 + 3m(m-1) p_2^2 + m(m-1)(m-2) p_2^3].$$

Rearranging the above system we obtain

$$\alpha m (p_1 - p_2) = m_1' - m p_2, \quad (3.58a)$$

$$\alpha m (p_1 - p_2) [(m-1)(p_1 + p_2) + 1] = m_2' - m p_2 - m(m-1) p_2^2, \quad (3.58b)$$

$$\alpha m (p_1 - p_2) [1 + 3(m-1)(p_1 + p_2) + (m-1)(m-2)(p_1^2 + p_1 p_2 + p_2^2)] =$$

$$m_3' - mp_2 - 3m(m-1)p_2^2 - m(m-1)(m-2)p_2^3. \quad (3.58c)$$

Substituting equation (3.58a) into the last two reduces the system to

$$m_1'[1 + (m-1)(p_1 + p_2)] = m_2' + m(m-1)p_1p_2$$

$$m_1'[1 + 3(m-1)(p_1 + p_2) + (m-1)(m-2)[(p_1 + p_2)^2 - p_1p_2]] = m_3' + p_1p_2[3m(m-1) + m(m-1)(m-2)(p_1 + p_2)].$$

Using

$$p_1p_2 = \frac{m_1'[1 + (m-1)(p_1 + p_2)] - m_2'}{m(m-1)}$$

in the second equation above and solving for $(p_1 + p_2)$, we have, after considerable algebra:

$$(p_1 + p_2) = \frac{m(m_3' - 3m_2' + 2m_1') - (m-2)m_1'(m_2' - m_1')}{(m-2)[m(m_2' - m_1') - (m-1)m_1'^2]}.$$

It follows then that

$$p_1p_2 = \frac{m_1'(m-1)(m_3' - 3m_2' + 2m_1') - (m-2)(m_2' - m_1')^2}{(m-1)(m-2)[m(m_2' - m_1') - (m-1)m_1'^2]}.$$

Let $\beta_0 = p_1p_2$ and $\beta_1 = -(p_1 + p_2)$. Then $p^2 + \beta_1p + \beta_0 = 0$, and hence, the component probability estimates are obtained from

$$\hat{p}_{1,2} = \frac{-\beta_1 \pm \sqrt{\beta_1^2 - 4\beta_0}}{2}. \quad (3.59)$$

Finally, the mixing proportion estimate is given by (3.58a) as

$$\hat{\alpha} = \frac{m_1 - mp_2}{m(p_1 - p_2)}. \quad (3.60)$$

The method as explained previously is not infallible. Both the estimators, $\hat{\alpha}_i$ and $\hat{p}_i$ ($i=1,\dots,c$), may assume values outside of (0,1) and $\hat{p}_i$ may be complex. In the two-component case, the conditions

$$(i) \beta_1^2 - 4\beta_0 > 0 \text{ and,}$$

$$(ii) \sqrt{\beta_1^2 - 4\beta_0} \leq \min(-\beta_1, 2+\beta_1)$$

must be satisfied in order for the estimates of the parameters to be valid [6].

3.6.2.1 Simulation Results

The method of moments was also applied experimentally to mixtures of binomial densities. Estimation of the component density parameter, p_i , and the mixing weight, α_i , were based on factorial moments of 1000 observations.

A number of two-component mixtures with various separation distances were utilized to investigate the method of moments. Table 3.7 lists the estimates derived for a mixture characterized by a distance between components of $\Delta = 2.1$. The number of Bernoulli trials per given binomial experiment was set at $m = 10$ for both components.

The results of Table 3.7 reveal a close agreement between the moment estimates and the prescribed parameters. The estimate of the mixing proportion differed by just over 2% from the population value, while the component probability estimates differed by less than 1% from their theoretical counterparts. The variances of the estimators were in the order of 10^{-3} or less. Figure 3.11 illustrates the population and estimated mixtures of the two binomial functions.

	Mean Parameter Estimates	Variance of Parameter Estimates	True Parameter Values
p_1	0.3970	4.575×10^{-4}	0.40
p_2	0.6957	1.992×10^{-4}	0.70
α	0.3910	2.819×10^{-3}	0.40

Table 3.7: The probabilities and mixing proportion of a two-component binomial mixture with $m = 10$. The estimation was carried out using the MM.

The determination of the EMIM metric for the binomial mixture served to quantify the goodness of fit of the estimated distribution to the true density. The plot of this metric as a function of sample size shown in Figure 3.12 reveals a fairly rapid decrease of 70% in the average EMIM value in the interval $n = 100$ to $n = 400$. Samples of 500 or more observations afforded only modest improvements in fit over smaller samples.

Figure 3.11: The population and estimated mixture distributions of two binomial PMFs with the parameters listed in Table 3.7.

Figure 3.12: The average EMIM metric as a function of sample size for the binomial distribution of Table 3.7.

3.6.3 Geometric Mixtures

Mixture distributions of c geometric densities are defined by (3.48) with $P_i(\cdot)$ given by

$$P(x;p) = (1-p)^x p \quad x = 0, 1, \dots$$

Compound probability functions of this type involve the estimation of the $2c-1$ parameters $(p_1, \dots, p_c)$, and $(\alpha_1, \dots, \alpha_{c-1})$.

By a similar derivation to that for Poisson mixtures, factorial moments of geometric mixtures can be shown to have the form

$$\mu'_{[k]} = k! \cdot \sum_{i=1}^c \alpha_i \left(\frac{q_i}{p_i} \right)^k \quad k = 1, 2, \dots$$

where $q_i = 1 - p_i$.

Constructing the $2c-1$ moment equations in the usual manner leads to the system

$$\phi_k = \sum_{i=1}^c \alpha_i \left(\frac{q_i}{p_i} \right)^k \quad k = 0, 1, 2, \dots, 2c-1$$

with $\phi_k = m'_{[k]} / k!$. Progressing as described in Section 3.6, this system ultimately yields the desired parameter estimates of the component probabilities, p_i , and the mixing proportions, α_i ($i=1, \dots, c$).

3.6.3.1 Simulation Results

Simulations of geometric distributions were also performed and their parameters were subsequently estimated using the method of moments. For two-component mixtures, the parameters of interest were the Bernoulli probabilities, p_1 and p_2 , and the mixing weight, α .

The results tabulated in Table 3.8 are for mixtures with a component separation of $\Delta = 0.8$. From the table, we see that the estimates for the geometric densities were noticeably less accurate than for the other discrete, as well as continuous, distributions presented earlier. The method of moments manifested a particular difficulty in resolving geometric mixtures. Estimates of both the probability of occurrence of an event, p_i , and the mixing proportion, α_i , were frequently outside the range (0,1) for sample sizes less than 500 observations. Consequently, larger samples were drawn from the population for geometric mixtures than for other types of distributions. Specifically, the results given in Table 3.8 are for samples of $n = 1400$. Figure 3.13 compares the true and estimated mixture distributions.

	Mean Parameter Estimates	Variance of Parameter Estimates	True Parameter Values
p_1	0.7873	1.598×10^{-3}	0.8
p_2	0.1824	1.900×10^{-3}	0.2
α	0.9079	8.035×10^{-4}	0.9

Table 3.8: The probabilities and mixing proportion of a two-component geometric mixture. The estimates result from using the MM.

Figure 3.13: The population and estimated mixture distributions of two geometric PMFs with the parameters indicated in Table 3.8.

Figure 3.14: The average EMIM metric as a function of sample size for the geometric distribution of Table 3.8.

The goodness of fit of the estimated distribution was measured by the EMIM metric. Figure 3.14 shows the change in the metric with sample sizes ranging from $n = 500$ to $n = 1400$. The greatest improvement in fit occurred in increasing the number of

observations from 500 to 900 where the EMIM value decreased by about 50%. A comparatively smaller change was measured for larger sample sizes.

3.7 Summary

The focus of this chapter was the parameter estimation of mixture distributions by the method of moments. This general estimation method was derived in detail and its application to a number of continuous and discrete distributions from the exponential family was shown. In particular, estimators for normal, exponential, gamma, Weibull, Poisson, binomial, and geometric mixtures were constructed. It was noted that the principal difficulty with estimation by the method of moments is the possibility of producing indeterminate estimates [26].

A number of experiments were performed on simulated mixtures to investigate the estimation of parameters using the method of moments. Seven different types of densities from the exponential family were utilized to form two-component mixtures.

A generalized distance measure between components was previously defined to characterize the extent of mixing in a mixture. Several different distances were tested for various distributions. In particular, results provided for the normal distribution showed that estimates generally improved with increasing separation of the component densities.

Overall, estimates obtained by the method of moments agreed quite well with the true values. Certain types of mixtures were more problematic than others in that the method was prone to yielding undefined estimates. In particular, a number of gamma, Poisson, and geometric samples led to estimates which were outside their parameter space.

The goodness of fit of the estimated mixtures to the true distributions was quantified through the use of the χ^2 and EMIM metrics. The calculation of these quantities demonstrated that the parameter estimates improved significantly with increasing sample size.

CHAPTER 4

ESTIMATION USING MAXIMUM LIKELIHOOD

4.1 Introduction

The method of maximum likelihood (ML) estimation constitutes another general technique applicable to the problem of parameter estimation. Although the method is generally attributed to the work of Fisher [28] dating back to the early part of this century, ML estimation in mixture density problems did not come to the fore until the last couple of decades. The solution of the nonlinear system of equations required for ML estimates presented a computational task of such formidable proportions that its use was not generally practicable. As a result, ML estimation was initially restricted to simple mixtures in which assumptions served to reduce the magnitude of the computations. An example of such an endeavor is the study by Rao [59] of mixtures of two univariate normal densities with equal variances.

As improved numerical methods emerged and electronic computers became more widely available, the method of ML was subsequently applied to increasingly general and varied mixtures. The late 1960's and 1970's witnessed a substantial number of investigators studying the mixture density estimation problem by ML. For example, Hasselblad [34] first reported on the estimation of c -component mixtures of univariate normals, and then Duda and Hart [24], Wolfe [73], and Peters and Walker [56] followed with investigations of general multivariate normal mixtures. Today, due in large part to the desirable sampling properties of its estimates, the method of ML is commonly considered the method of choice for the parameter estimation of mixture distributions.

The following sections describe ML estimation for mixtures of probability densities. The derivation of the expressions for the ML estimates is presented in the

general case. Several iterative procedures typically used in obtaining approximations to the parameter estimates are subsequently described. Further, their different approaches and convergence properties are briefly discussed. Following this, the estimating equations for the class of distributions from the exponential family are developed specifically. In particular, ML estimates for general c -component univariate mixtures of common continuous and discrete distributions are given. The results of simulation studies involving two-component mixtures of such distributions are presented. In closing, certain practical difficulties associated with the method of ML are explained.

4.2 Maximum Likelihood

Consider the n independent observations $x_1, \dots, x_n$ drawn from the PDF $f(x; \theta)$ where θ is the fixed unknown parameter vector of interest. Then the joint probability of the sample $X^n = \{x_1, \dots, x_n\}$ is given by the function

$$f_n(X^n; \theta) = \prod_{k=1}^n f(x_k; \theta). \quad (4.1)$$

Taking the perspective that (4.1) is a function of θ , the principle of estimation by ML is to seek an estimator $\hat{\theta}$ such that f_n is maximized. That is, we select an estimator $\hat{\theta}$ such that for any value of θ in its parameter space,

$$f_n(X^n; \hat{\theta}) \geq f_n(X^n; \theta).$$

Thus we say that there is a greater likelihood that the observed sample X^n originated from the density $f_n(X^n; \hat{\theta})$ than from $f_n(X^n; \theta)$ for any allowable θ . In this context f_n is called the likelihood function.

The estimator, $\hat{\theta}$, giving rise to the maximum value of the likelihood function, f_n , can be determined in the following manner. Assuming that the first and second partial derivatives of f_n with respect to θ exist, then the stationary points of the likelihood function are given by the roots of

$$\nabla_{\theta} f_n(X^n; \theta) = 0,$$

where ∇_{θ} denotes the gradient of $\partial/\partial\theta_i$. The roots, $\tilde{\theta}$, obtained above yield local maxima provided that

$$\nabla_{\theta} \nabla_{\theta}^T f_n(X^n; \tilde{\theta}) < 0,$$

where $\nabla_{\theta} \nabla_{\theta}^T$ indicates the grad of $\partial^2/\partial^2\theta_i$. The ML estimator, $\hat{\theta}$, is selected as the largest of the local maxima.

Due to the exponential nature of many PDFs, one typically works with the logarithm of the likelihood function in order to simplify the algebra. Using (4.1), the log-likelihood function is defined as

$$L(X^n; \theta) = \log f_n(X^n; \theta) = \sum_{k=1}^n \log f(x_k; \theta). \quad (4.2)$$

This is a reasonable transformation since the maxima of the functions f_n and L are seen to coincide

$$\nabla_{\theta} L = \frac{\nabla_{\theta} f_n}{f_n}, \quad f_n > 0.$$

Hence solutions of

$$\nabla_{\theta} L = 0, \quad (4.3)$$

known as the likelihood equations, for which

$$\nabla_{\theta} \nabla_{\theta}^T L < 0$$

yield the ML estimator, $\hat{\theta}$ [44].

The foregoing definition of ML is applicable to the problem of parameter estimation for mixtures of distributions. Consider now n independent samples $x_1, \dots, x_n$ taken from a mixed population of c component densities $f_i(x; \theta_i)$ ($i=1, \dots, c$). The compound PDF is given by†

$$f(x; \Theta) = \sum_{i=1}^c \alpha_i f_i(x; \theta_i)$$

where $\Theta = (\theta_1, \dots, \theta_c, \alpha_1, \dots, \alpha_c)$ denotes the vector of component parameters, $\theta_i = (\theta_{i1}, \dots, \theta_{ir})$, and of mixing proportions, α_i , satisfying

$$\alpha_i > 0, \quad \sum_{i=1}^c \alpha_i = 1.$$

The log-likelihood function of a sample $X^n = \{x_1, \dots, x_n\}$ is thus defined as

$$L(X^n; \Theta) = \sum_{k=1}^n \log f(x_k; \Theta) = \sum_{k=1}^n \log \left\{ \sum_{i=1}^c \alpha_i f_i(x_k; \theta_i) \right\}. \quad (4.4)$$

As explained above, the ML estimate of a parameter vector, Θ , for a given distribution is the choice of the parameter vector which maximizes the log-likelihood function, $L(X^n; \Theta)$. The ML estimator, $\hat{\Theta}$ is determined by first setting up the system of likelihood equations satisfied by Θ and then solving these likelihood equations for the ML estimator. The likelihood equations are obtained by equating to zero the first partial

† Note that Θ is defined slightly differently here than in Section 2.2.1 for notational convenience.

derivatives of the log-likelihood function with respect to the components of Θ . Thus, if $\hat{\Theta} = (\hat{\theta}_1, \dots, \hat{\theta}_c, \hat{\alpha}_1, \dots, \hat{\alpha}_c)$ is a ML estimator of Θ , we have the likelihood equations

$$\nabla_{\Theta} L(\hat{\Theta}) = 0. \quad (4.5)$$

Using (4.4), the first partial derivative of the log-likelihood function with respect to the mixing proportion α_i is given by

$$\frac{\partial}{\partial \alpha_i} L(\Theta) = \sum_{k=1}^n \frac{1}{f(x_k; \Theta)} \frac{\partial}{\partial \alpha_i} \left[\sum_{i=1}^c \alpha_i f_i(x_k; \theta_i) \right] \quad i=1, \dots, c.$$

Recalling the constraints $\alpha_i > 0$ and $\sum_{i=1}^c \alpha_i = 1$, we have

$$\begin{aligned} \frac{\partial}{\partial \alpha_i} L(\Theta) &= \sum_{k=1}^n \frac{1}{f(x_k; \Theta)} \frac{\partial}{\partial \alpha_i} \left[\sum_{i=1}^{c-1} \alpha_i f_i(x_k; \theta_i) + (1 - \sum_{i=1}^{c-1} \alpha_i) f_c(x_k; \theta_c) \right], \\ \frac{\partial}{\partial \alpha_i} L(\Theta) &= \sum_{k=1}^n \frac{f_i(x_k; \theta_i) - f_c(x_k; \theta_c)}{f(x_k; \Theta)} = 0. \end{aligned} \quad (4.6)$$

Multiplying both sides of (4.6) by α_i and summing over i we obtain

$$\sum_{i=1}^c \alpha_i \sum_{k=1}^n \frac{f_i(x_k; \theta_i)}{f(x_k; \Theta)} = \sum_{i=1}^c \alpha_i \sum_{k=1}^n \frac{f_c(x_k; \theta_c)}{f(x_k; \Theta)}.$$

Simplifying then leads to

$$n = \sum_{k=1}^n \frac{f_c(x_k; \theta_c)}{f(x_k; \Theta)}. \quad (4.7)$$

If $\hat{\alpha}_i$ denotes the ML of the mixing proportion, α_i , then it follows from (4.6) and (4.7) that

$$\hat{\alpha}_i = \frac{1}{n} \sum_{k=1}^n \hat{\alpha}_i \frac{f_i(x_k; \hat{\theta}_i)}{f(x_k; \hat{\Theta})} \quad i=1, \dots, c. \quad (4.8)$$

Let the component parameters θ_i ($i=1, \dots, c$) be mutually independent vectors. Then taking the gradient of the likelihood function with respect to θ_i we have

$$\begin{aligned} \nabla_{\theta_i} L(\Theta) &= \sum_{k=1}^n \frac{1}{f(x_k; \Theta)} \nabla_{\theta_i} \left[\sum_{i=1}^c \alpha_i f_i(x_k; \theta_i) \right] \\ \nabla_{\theta_i} L(\Theta) &= \sum_{k=1}^n \alpha_i \frac{f_i(x_k; \theta_i)}{f(x_k; \Theta)} \nabla_{\theta_i} \log f_i(x_k; \theta_i) = 0. \end{aligned}$$

Thus the ML estimators, $\hat{\theta}_i$, of the component parameters are given by

$$\sum_{k=1}^n \hat{\alpha}_i \frac{f_i(x_k; \hat{\theta}_i)}{f(x_k; \hat{\Theta})} \nabla_{\theta_i} \log f_i(x_k; \hat{\theta}_i) = 0. \quad (4.9)$$

By Bayes' rule the posterior probability that the sample value x_k was drawn from the i th class is expressed as

$$P(i | x_k) = \alpha_i \frac{f_i(x_k; \theta_i)}{f(x_k; \Theta)}. \quad (4.10)$$

Hence, equations (4.8) and (4.9) can be written in the intuitive forms

$$\hat{\alpha}_i = \frac{1}{n} \sum_{k=1}^n \hat{P}(i | x_k) \quad (4.11)$$

and

$$\sum_{k=1}^n \hat{P}(i | x_k) \nabla_{\theta_i} \log f_i(x_k; \hat{\theta}_i) = 0, \quad (4.12)$$

respectively. That is, the ML estimate of the mixing proportion, α_i , is simply the sample mean of the posterior probabilities, $\hat{P}(i | x_k)$. Similarly, ML estimates of the component parameters, $\hat{\theta}_i$, can be viewed as the average of the likelihood equations of the individual components of the mixture with weights $\hat{P}(i | x_k)$ [24].

In most cases, unique expressions for the $\hat{\theta}_i$ are easily obtained from (4.12). In particular, when the individual component densities, $f_i(\cdot)$, belong to the exponential family of distributions, the ML estimates are readily determined. Mixtures of such distributions are treated in more detail in Section 4.4.

4.3 Iterative Procedures

The expressions for the ML estimates given in the foregoing section cannot normally be solved analytically. The nonlinear likelihood equations for mixture distributions comprise a complex web of coupled simultaneous equations for which an explicit solution cannot be determined. However, provided with reasonable initial values for the estimates (i.e. in the neighborhood of the "true" parameter values), the likelihood equations can be solved numerically by a number of different iterative schemes. General procedures include the method of steepest ascent, Newton's method, and the method of scoring. A description of each of these algorithms follows.

4.3.1 The Method of Steepest Ascent

This hill climbing method has been studied by several different investigators from two distinct perspectives. The first view arises intuitively from the form of the likelihood equations and was conceived in the context of the mixture density estimation problem. The procedure is also typically referred to by various other names including steepest descent, gradient ascent, and deflected gradient [35,56]. The second view is a more general one in that the problem is approached as one of incomplete data where the samples of a mixture are treated as missing the labels that would otherwise indicate their population of origin. From this point of view the procedure is termed the EM algorithm where the acronym, EM, comes from E for expectation and M for minimization. The theoretical formulation for this algorithm can be found in Dempster, Laird, and Rubin [19]. For the present purpose, the first view of this iterative scheme is adopted.

Let $\underline{\Theta}^{(v)} = (\underline{\theta}_1^{(v)}, \dots, \underline{\theta}_c^{(v)}, \alpha_1^{(v)}, \dots, \alpha_c^{(v)})$ denote the parameter vector approximation at the v th iteration stage. Then from (4.8) and (4.9) the next approximation $\underline{\Theta}^{(v+1)} = (\underline{\theta}_1^{(v+1)}, \dots, \underline{\theta}_c^{(v+1)}, \alpha_1^{(v+1)}, \dots, \alpha_c^{(v+1)})$ is given by

$$\alpha_i^{(v+1)} = \frac{1}{n} \sum_{k=1}^n \alpha_i^{(v)} \frac{f_i(x_k; \underline{\theta}_i^{(v)})}{f(x_k; \underline{\Theta}^{(v)})} \quad i=1, \dots, c. \quad (4.13)$$

and

$$\sum_{k=1}^n \alpha_i^{(v)} \frac{f_i(x_k; \underline{\theta}_i^{(v)})}{f(x_k; \underline{\Theta}^{(v)})} \nabla_{\theta_i} \log f_i(x_k; \underline{\theta}_i^{(v)}) = \underline{0}. \quad (4.14)$$

Thus initial values $\alpha_i^{(0)}$ and $\theta_i^{(0)}$ are first inserted into the right-hand side of (4.13) and (4.14) to obtain first approximations of the ML estimates. These approximations are then substituted into the same equations to yield second approximations and the process is repeated until convergence to the estimates $\hat{\alpha}_i$ and $\hat{\theta}_i$. Techniques for obtaining starting values that are reasonably close to the "true" parameter values for the iterative procedure are discussed in Section 4.3.4.

As an illustration, consider the iterative expressions for mixtures of normal densities whose ML estimates will be derived in Section 4.4.1. For a given v th approximation, $\theta_i^{(v)} = (\phi_i^{(v)}, \sigma_i^{2(v)})$, the next approximation, $\theta_i^{(v+1)} = (\phi_i^{(v+1)}, \sigma_i^{2(v+1)})$, of the component parameter estimates is given by

$$\phi_i^{(v+1)} = \frac{\sum_{k=1}^n x_k \frac{f_i(x_k, \phi_i^{(v)}, \sigma_i^{2(v)})}{f(x_k; \phi^{(v)}, \sigma^2(v))}}{\sum_{k=1}^n \frac{f_i(x_k, \phi_i^{(v)}, \sigma_i^{2(v)})}{f(x_k; \phi^{(v)}, \sigma^2(v))}} \quad (4.15)$$

and

$$\sigma_i^{2(v+1)} = \frac{\sum_{k=1}^n (x_k - \phi_i^{(v+1)})^2 \frac{f_i(x_k, \phi_i^{(v)}, \sigma_i^{2(v)})}{f(x_k; \phi^{(v)}, \sigma^2(v))}}{\sum_{k=1}^n \frac{f_i(x_k, \phi_i^{(v)}, \sigma_i^{2(v)})}{f(x_k; \phi^{(v)}, \sigma^2(v))}}. \quad (4.16)$$

Peters and Walker [56] report on the theoretical reliability of a generalized steepest ascent procedure for approximating the ML estimates of a mixture of multivariate normal

densities. Their procedure is analogous to the foregoing hill climbing method with the exception of a variable iteration step size.

The method of steepest ascent possesses several desirable properties which often make it the preferred scheme for ML estimation. From a practical standpoint, the algorithm is notably the simplest to implement and requires minimal storage space. From the theoretical formulation it can be shown that the log-likelihood function $L(\Theta)$ increases monotonically with successive iterations of the algorithm. In addition, provided that initial parameters are chosen within their admissible range, the ensuing estimation values will remain in that range throughout the sequence of iterations. This is a distinct advantage over the method of moments which does not guarantee the validity of its estimates (cf. § 3.3).

4.3.2 Newton's Method

The Newton or Newton-Raphson iterative procedure for approximating the ML estimates of mixture distributions is based on the standard Newton scheme [54]. Consider the equation $g(x)=0$ whose roots, x , we wish to determine. Then let

$$x = x_0 + \epsilon$$

where x_0 is the approximate value of the root and ϵ is some small correction. Hence, we have for $g(x)=0$

$$g(x_0 + \epsilon) = 0.$$

By a Taylor's series expansion of $g(x_0 + \epsilon)$ we obtain for sufficiently small ϵ

$$g(x_0) + \epsilon \cdot g'(x_0) \approx 0$$

or

$$\varepsilon \approx -\frac{g(x_0)}{g'(x_0)}.$$

Then as a first approximation, $x^{(1)}$, to the root, x , we have

$$x^{(1)} = x_0 - \frac{g(x_0)}{g'(x_0)}.$$

Hence, by reiterating this process we can approximate the root by

$$x^{(v+1)} = x^{(v)} - \frac{g(x^{(v)})}{g'(x^{(v)})}.$$

Using this scheme to approximate the roots of the likelihood equations $\nabla_{\Theta} L(\Theta) = 0$ given by (4.5), we have

$$\Theta^{(v+1)} = \Theta^{(v)} - H(\Theta^{(v)})^{-1} \cdot \nabla_{\Theta} L(\Theta^{(v)}) \quad (4.17)$$

where, if m is the number of parameters to be estimated, $H(\Theta^{(v)})$ is an $(m \times m)$ Hessian matrix of second partial derivatives of $L(\Theta)$ evaluated at the v th approximation to the parameter values, Θ . That is

$$H(\Theta^{(v)}) = \sum_{k=1}^n \nabla_{\Theta} \nabla_{\Theta}^T \log f(x_k; \Theta^{(v)}). \quad (4.18)$$

For example, the second partial derivatives which form the elements of the Hessian matrix for a mixture of normal densities are shown in Hasselblad [34]. From the likelihood equations (4.8) and (4.9) he derives the six expressions required to obtain the ML estimates.

A significant difficulty with Newton's method in some cases is the necessity of choosing starting values for the estimates which are very close to the "true" values in order to achieve convergence (techniques for selecting initial estimates are discussed in § 4.3.4). Hasselblad [34] investigated mixtures of univariate normal densities and found empirically

that Newton's method did not always converge to a solution of the likelihood equations (4.5), whereas the method of steepest ascent always succeeded. Further, he reported that the former tended to be more effective with larger samples of data.

4.3.3 The Method of Scoring

The method of scoring belongs to a class of Newton-like procedures devised specifically for ML estimation [60]. The system of equations leading to ML estimates is similar to that of (4.17) for Newton's method with the exception that the Hessian matrix, $H(\Theta)$, is approximated by a simpler matrix, $B(\Theta)$:

$$\Theta^{(v+1)} = \Theta^{(v)} - B(\Theta^{(v)})^{-1} \cdot \nabla_{\Theta} L(\Theta^{(v)}). \quad (4.19)$$

The objective is to specify $B(\Theta)$ such that the computational complexity is reduced compared to that of the Hessian matrix.

The Hessian matrix of (4.18) can be written as

$$H(\Theta) = - \sum_{k=1}^n [\nabla_{\Theta} \log f(x_k; \Theta)] \cdot [\nabla_{\Theta} \log f(x_k; \Theta)]^T + \sum_{k=1}^n \frac{1}{f(x_k; \Theta)} \nabla_{\Theta} \nabla_{\Theta}^T f(x_k; \Theta) \quad (4.20)$$

where $f(x_k; \Theta) = \sum_{i=1}^n \alpha_i f_i(x_k; \Theta_i)$. However, the second term of (4.20) has an expected value

of zero when evaluated at the "true" parameters. Hence, in the vicinity of local maxima of the likelihood function, $H(\Theta)$ is dominated by the first term of (4.20) and so $B(\Theta)$ can be chosen as

$$B(\Theta) = - \sum_{k=1}^n [\nabla_{\Theta} \log f(x_k; \Theta)] \cdot [\nabla_{\Theta} \log f(x_k; \Theta)]^T. \quad (4.21)$$

The matrix $B(\Theta)$ is less computationally demanding than the Hessian matrix since only first partial derivatives of the log-likelihood function are needed, and these are already found in $\nabla_{\Theta} L(\Theta)$.

4.3.4 Selecting Initial Estimates

As for any iterative procedure, the foregoing schemes require initial parameter estimates. The choice of starting values may or may not be critical to the accuracy of the final estimates depending on the type and properties of the mixture. Generally speaking, mixtures of univariate or well separated components are indifferent to the choice of initial estimates, while multivariate or poorly separated densities are markedly affected by it.

Everitt and Hand [26] provide two examples that illustrate the relative importance of initial values in different cases. In the first example, data from Duda and Hart [24] for a two-component univariate normal mixture are utilized. Using the hill climbing procedure it was found that three widely differing sets of initial values resulted in exactly the same set of final estimates.

A second example consists of a mixture of three four-dimensional normal densities representing the distributions of four measurements of three different plant species. For a certain choice of initial values, the ML estimates were seen to converge by the method of steepest ascent to a limit approaching the "true" parameter values. However, for another set of starting estimates the procedure yielded ML estimates corresponding to a singular solution of the likelihood equations. Hence, in this multivariate case the selection of initial values was crucial to the outcome of the estimation even with loosely coupled component densities. Singularities of the likelihood function are discussed further in Section 4.5.

Several approaches to selecting reasonable initial parameter estimates exist. The

most obvious one is a selection based on prior knowledge of the problem at hand. This knowledge may be derived from established theory concerning the population or from previous empirical observations.

A second technique alluded to in the preceding chapter is the use of estimates generated by the method of moments. This method can furnish good starting values in many cases. However, its use is mainly confined to univariate distributions due to the poor sampling properties of the higher moments required for more complex mixtures.

A third approach has its basis in an old estimation technique conceived before the computer age in an effort to avoid the computational complexity of more accurate methods. These techniques are collectively known as graphical methods and are mainly of historical interest today. However, for certain mixtures they may provide reasonable starting values for the iterative schemes of ML. Of note is the method of Bhattacharya [5] for grouped data. A distinct advantage of these estimation methods is that no prior knowledge of the number of component densities in the mixture is required. In fact this approach may be useful in determining the number of subpopulations prior to the estimation of the other parameters by the MM or ML.

Finally, a rather subjective technique based on some statistic determined from the sample can be used to generate initial values for the iterative procedures. For example, the mode of the observations has been used successfully for discrete distributions by Hasselblad [35]. If, for instance, in a two-component Poisson mixture the mode is found to be 2 then the initial parameter estimates could be $\hat{\lambda}_1=1$, $\hat{\lambda}_2=3$, and $\hat{\alpha}=0.5$.

An approach based on function mode could prove useful for mixtures of other types of distributions as well. Several investigators have studied the influence of component parameters on the modality of the mixed density [2,25,26,71]. For example, for mixtures

of two univariate normals it can be shown [25] that a sufficient condition for the existence of values of the mixing proportion, α , such that the distribution is bimodal is that

$$(\phi_2 - \phi_1)^2 > \frac{8 \cdot \sigma_1^2 \sigma_2^2}{(\sigma_1^2 + \sigma_2^2)}. \quad (4.22a)$$

However, for any such two-component mixture there are parameter values for which the density is unimodal as well. Behboodian [2] shows that for a two-component normal mixture, a sufficient condition for the distribution to be unimodal is given by

$$|\phi_2 - \phi_1| \leq 2 \min(\sigma_1, \sigma_2). \quad (4.22b)$$

In fact, in practice, single density functions approximated by sample data may reveal more than one mode due to spurious or insufficient observations. Hence, the modality of a distribution is not usually a reliable indicator of the presence of a mixture. Nevertheless, this approach can be tried in the absence of other alternatives for obtaining initial values.

4.3.5 Convergence of Iterative Procedures

In the use of iterative procedures, several aspects of convergence are the subject of practical concern. One such aspect regards establishing a convergence criterion to determine when the algorithm has reached the limiting values of the estimates. A number of different criteria are typically utilized for this purpose [66].

For instance, changes in the value of the log-likelihood function between successive iteration steps can be monitored and when this change falls below some prescribed tolerance, ϵ , the algorithm terminates. Using the log-likelihood function defined in (4.4) we have as a stopping condition

$$0 \leq L(\underline{\Theta})^{(v+1)} - L(\underline{\Theta})^{(v)} < \epsilon, \quad \epsilon > 0.$$

Note, however, that if the value of the function changes only slowly in the neighborhood of its maximum, this condition may cause the algorithm to terminate prematurely.

A second convergence criterion is based on the likelihood equations defined in (4.5). Since at a local maximum the first partial derivatives of the likelihood function with respect to the component parameters must vanish, we can terminate the iteration when

$$\|\nabla_{\Theta} L(\underline{\Theta})^{(v+1)}\| < \epsilon, \quad \epsilon > 0,$$

is satisfied. A potential problem with this condition is that if the likelihood surface is relatively flat in the vicinity of its optimum, the choice of appropriate tolerance, ϵ , becomes critical.

Finally, the approximate ML estimates themselves can be examined to determine convergence. If the change between successive values of the estimates of θ_i for all ($i=1, \dots, c$) is less than a given tolerance then the procedure is terminated. That is, the stopping condition is given by

$$\|\underline{\Theta}^{(v+1)} - \underline{\Theta}^{(v)}\| < \epsilon, \quad \epsilon > 0, \quad i = 1, \dots, c.$$

In practice, a combination of the above convergence criteria is often used along with the safeguard of specifying a maximum allowable number of iterations.

Another important matter of concern with regards to iterative procedures is their convergence rate. The speed of convergence of a given algorithm is affected by the time complexity per iteration and the overall number of iterations required for the estimates to approach their limit. In the following, the cost per iteration is first evaluated and then the

number of iterations required for convergence in general is examined to compare the expected speed of convergence of each algorithm.

The method of steepest ascent exhibits the lowest cost per iteration of the three procedures described. Consider the expressions for the ML estimates for general mixture densities given in (4.13) and (4.14). For a c -component mixture of univariate distributions the method requires $O(c)$ time to compute the summands. These computations must be performed n times for a sample of n observations $\{x_1, \dots, x_n\}$. Hence, the total cost per iteration is $O(nc)$.

Newton's method manifests a considerably higher cost per iteration than that of the method of steepest ascent. Equation (4.17) indicates that at each iteration three principal sets of calculations must be performed. The first involves evaluation of the Hessian matrix $H(\underline{\Theta})$ at $\underline{\Theta}^{(v)}$, while the second includes the computation of the vector of first partial derivatives of the likelihood function, $L(\underline{\Theta}^{(v)})$. Each require $O(nc^2)$ time. The last set is the solution of a linear matrix equation which in general can be achieved in $O(c^3)$ time. Thus, the overall cost per iteration of Newton's method for univariate mixtures is $O(nc^2) + O(c^3)$.

The cost per iteration of the method of scoring can be determined from equations (4.19) and (4.21). The matrix $B(\underline{\Theta}^{(v)})$ requires $O(nc^2)$ operations and the linear system of equations is solved in $O(c^3)$ time. Thus, the asymptotic behavior of this method is nominally the same as that of Newton's method. However, the terms of $B(\underline{\Theta}^{(v)})$ are contained in the vector $\nabla_{\underline{\Theta}} L(\underline{\Theta}^{(v)})$, and so the evaluation of the former is likely to be less expensive than that of the Hessian matrix. Consequently, in practice, the cost per iteration of the method of scoring is somewhat less than that of Newton's method.

Although the method of steepest ascent possesses the lowest cost per iteration, it normally exhibits a slower speed of convergence than the other two methods due to the

greater number of iterations required for convergence. It can be shown [60] that given an initial parameter approximation, $\underline{\theta}^{(0)}$, sufficiently near the ML estimate, $\hat{\underline{\theta}}$ the following inequality holds for the method of steepest ascent:

$$\|\underline{\theta}^{(v+1)} - \hat{\underline{\theta}}\| \leq \lambda \|\underline{\theta}^{(v)} - \hat{\underline{\theta}}\|, \quad v = 0, 1, 2, \dots \quad (4.23)$$

for some constant $0 \leq \lambda < 1$. That is, the method of steepest ascent converges linearly.

The method of scoring also manifests linear convergence as defined in (4.23), but is often faster than the method of steepest ascent [60].

On the other hand, under the same conditions for $\underline{\theta}^{(0)}$, Newton's method is characterized by quadratic convergence:

$$\|\underline{\theta}^{(v+1)} - \hat{\underline{\theta}}\| \leq \beta \|\underline{\theta}^{(v)} - \hat{\underline{\theta}}\|^2, \quad v = 0, 1, 2, \dots \quad (4.24)$$

for some constant β [60]. Hence, Newton's method is by far the most rapid of the three procedures.

Nevertheless, in light of the practical problems with Newton's method regarding the possibility of nonconvergence and the considerably greater complexity of implementation, the method of steepest ascent is typically the preferred technique. A number of investigators have empirically studied the properties of this algorithm. Hasselblad [35] and Redner and Walker [60] reported that the speed of convergence of the procedure increases with greater separation between the component subpopulations of the mixture and with larger sample size, n .

4.4 Mixture of Distributions from the Exponential Family

In Section 2.3 the general family of exponential probability densities was briefly discussed. Recall that in natural parametrization these distributions have the form

$$f(x; \boldsymbol{\theta}) = g(x) \exp\left[\sum_{h=1}^s d_h(\boldsymbol{\theta}) t_h(x) + c(d_1(\boldsymbol{\theta}), \dots, d_s(\boldsymbol{\theta}))\right], \quad 1 \leq r \leq s,$$

where $d(\boldsymbol{\theta})$ is the natural parameter and $\boldsymbol{\theta} = (\theta_1, \dots, \theta_r)$. Thus, for a c -component mixture the i th component PDF is written as

$$f_i(x; \boldsymbol{\theta}_i) = g(x) \exp\left[\sum_{j=1}^s d_{ij}(\boldsymbol{\theta}_i) t_{ij}(x) + c(d_{i1}(\boldsymbol{\theta}_i), \dots, d_{is}(\boldsymbol{\theta}_i))\right]$$

for $i = 1, \dots, c$. Including the mixing proportions, α_i , there are then $r \cdot c + c - 1$ unknown parameters to be estimated, namely, $\theta_{11}, \dots, \theta_{1r}, \theta_{21}, \dots, \theta_{2r}, \dots, \theta_{c1}, \dots, \theta_{cr}, \alpha_1, \dots, \alpha_{c-1}$.

The ML estimates of the parameters are obtained by maximizing the log-likelihood function (4.4)

$$L(X^n; \boldsymbol{\Theta}) = \sum_{k=1}^n \log \left\{ \sum_{i=1}^c \alpha_i f_i(x_k; \boldsymbol{\theta}_i) \right\}.$$

Assuming that $d(\boldsymbol{\theta})$ and $c(d(\boldsymbol{\theta}))$ are differentiable, the first partial derivatives of L with respect to θ_{ij} can be determined:

$$\frac{\partial L}{\partial \theta_{ij}} = \sum_{k=1}^n \alpha_i \frac{f_i(x_k; \boldsymbol{\theta}_i)}{f(x_k; \boldsymbol{\Theta})} \left[\frac{\partial c_i}{\partial \theta_{ij}} + \sum_{h=1}^s \frac{\partial d_{ih}}{\partial \theta_{ij}} t_{ih}(x_k) \right] \quad (4.25)$$

for $i = 1, \dots, c$ and $j = 1, \dots, r$. Setting them to zero in the usual manner we have

$$-\frac{\partial c_i}{\partial \theta_{ij}} = \frac{\sum_{k=1}^n \frac{f_i(x_k; \theta_i)}{f(x_k; \Theta)} \sum_{h=1}^s \frac{\partial d_{ih}}{\partial \theta_{ij}} t_{ih}(x_k)}{\sum_{k=1}^n \frac{f_i(x_k; \theta_i)}{f(x_k; \Theta)}}. \quad (4.26)$$

From (4.26) the ML estimates, $\hat{\theta}_{ij}$, of the distribution parameters are calculated based on the $d(\theta)$ and $c(d(\theta))$ for the particular component densities in question. The ML estimates of the mixing proportions, $\hat{\alpha}_i$, follow from equation (4.8).

In the following sections, expressions for the ML estimates for mixtures of distributions from the exponential family are presented. Seven different densities are treated, namely, the normal, exponential, gamma, Weibull, Poisson, binomial, and geometric distributions. For each type of mixture, simulation studies based on the workbench of Section 2.4 were performed to investigate the ML principle. The data utilized for this investigation were the same as those employed for the method of moments. All tabulated results are based on ten samples of 1000 observations. The only exception is the geometric mixture for which 1400 observations were drawn from the population[†]. The iterative procedure selected to determine the ML estimates was the method of steepest ascent. Initial values for the procedure consisted of the parameter estimates arrived at by the method of moments. The convergence was defined by the change in successive ML estimates for all parameters being less than 0.001 up to a maximum of twenty iterations. That is, the convergence criterion was taken to be

$$\|\underline{\theta}^{(v+1)} - \underline{\theta}^{(v)}\| < 0.001, \quad i = 1, \dots, c.$$

[†] In applying the MM to geometric mixtures, it was found that larger samples were necessary to minimize the occurrence of indeterminate estimates. The same data were used here for comparison.

4.4.1 Normal Mixtures

The i th component univariate normal PDF in standard parametrization is expressed as (cf. § 2.3.1)

$$f_i(x; \theta_i) = \frac{1}{\sqrt{2\pi} \sigma_i} \exp \left[-\frac{(x - \phi_i)^2}{2\sigma_i^2} \right], \quad -\infty < x < \infty$$

where $\theta_i = (\phi_i, \sigma_i)$. Hence, $r = 2$ in this case and so there are $3c-1$ parameters to be estimated for the c -component mixture.

In natural parameter notation we recall from (2.7) that for the normal density,

$$t_1(x) = x, \quad t_2(x) = x^2,$$

$$d_1(\theta) = \frac{\phi}{\sigma^2}, \quad d_2(\theta) = -\frac{1}{2\sigma^2},$$

$$\text{and} \quad c(d_1, d_2) = -\frac{\phi^2}{2\sigma^2} - \frac{1}{2} \ln(2\pi\sigma^2).$$

Thus, using (4.26) leads to the ML estimates $\hat{\phi}_i$ and $\hat{\sigma}_i^2$:

$$\hat{\phi}_i = \frac{\sum_{k=1}^n x_k \frac{f_i(x_k; \hat{\theta}_i)}{f(x_k; \hat{\theta})}}{\sum_{k=1}^n \frac{f_i(x_k; \hat{\theta}_i)}{f(x_k; \hat{\theta})}}, \quad (4.27)$$

$$\text{and } \hat{\sigma}_i^2 = \frac{\sum_{k=1}^n (x_k - \hat{\phi}_i)^2 \frac{f_i(x_k; \hat{\theta}_i)}{f(x_k; \hat{\theta})}}{\sum_{k=1}^n \frac{f_i(x_k; \hat{\theta}_i)}{f(x_k; \hat{\theta})}}, \quad (4.28)$$

for $i = 1, \dots, c$.

4.4.1.1 Simulation Results

Expressions (4.27) and (4.28) for the component density parameters, as well as a expression (4.8) for the mixing weight, were utilized to estimate the parameters of mixtures of two normal distributions. In all, there were five parameter estimates to be determined: $\hat{\phi}_1$, $\hat{\phi}_2$, $\hat{\sigma}_1$, $\hat{\sigma}_2$, and $\hat{\alpha}$.

Different sets of initial values for the parameters were employed in order to investigate the impact of the choice of these values on the final outcome. Table 4.1 lists two sets of initial values and their corresponding mean parameter estimates. The first set consists of the estimates obtained from the method of moments. The second one constitutes a completely arbitrary choice in which $\hat{\phi}_1^{(0)}$, $\hat{\sigma}_2^{(0)}$, and $\hat{\alpha}^{(0)}$ were changed to differ significantly from their counterparts in the first set. The results of Table 4.1 (a) show that the final parameter estimates were reasonably close to the true values. The largest error occurred in the estimate of the mixing weight, which differed by about 15% from the population parameter. Examining Table 4.1 (b), we see that although the initial estimates were considerably different from those of Table 4.1 (a), the final estimates obtained were comparable. It should be noted, however, that the estimates of Table 4.1 (a)

resulted from an average of twenty iterations of the method of steepest ascent, whereas those of Table 4.1 (b) required forty iterations.

	Initial Values	Mean Parameter Estimates	Variance of Parameter Estimates	True Parameter Values
ϕ_1	46.18	46.76	0.3667	47.72
ϕ_2	57.37	57.45	0.0222	57.61
σ_1	4.91	5.23	0.1758	5.79
σ_2	3.20	3.19	5.63×10^{-3}	3.03
α	0.2649	0.2826	9.298×10^{-4}	0.332

Table 4.1 (a): The initial values are the moment estimates of Table 3.2 (b).

	Initial Values	Mean Parameter Estimates	Variance of Parameter Estimates	True Parameter Values
ϕ_1	23.15	46.57	0.4435	47.72
ϕ_2	57.37	57.42	0.0301	57.61
σ_1	4.91	5.13	0.1848	5.79
σ_2	2.11	3.21	5.87×10^{-3}	3.03
α	0.4621	0.2752	9.799×10^{-4}	0.332

Table 4.1 (b): The initial values for ϕ_2 and σ_1 are the moment estimates of Table 3.2 (b). The initial values for ϕ_1 , σ_2 , and α are completely arbitrary.

Table 4.1: The parameters of a mixture of two normal distributions using two different sets of initial values.

The estimated and true PDFs of the two-component normal mixture are illustrated in Figure 4.1. The estimated curve represents the final estimates resulting from the first set of initial values. The curve representing the second set of values (not shown) is virtually identical.

Figure 4.1: The population and estimated mixture distributions of two normal PDFs with the parameters tabulated in Table 4.1 (a).

Figure 4.2: The average χ^2 metric as a function of sample size for the normal distribution of Table 4.1(a)

The χ^2 metric was calculated for various sample sizes ranging from $n = 100$ to $n = 1000$. An average value based on ten samples for each sample size is shown in Figure 4.2. It was found that the χ^2 value decreased monotonically with increasing number of observations. The average value dropped from 0.044 for $n = 100$ to 0.0054 for $n = 1000$. This represented a decrease of roughly 88%.

4.4.2 Exponential Mixtures

For the exponential random variable the probability function (2.8) consists of a single scale parameter, b , and so we have $\theta_i = b_i$. Consequently, including the mixing proportions α_i , there are a total of $2c-1$ parameters to be estimated.

Consider the natural parametric form of the i th component exponential density defined by (2.9). We have the parameters $t_i(x) = x$, $d_i(\theta) = -b_i^{-1}$, and $c(d_i) = -\ln(b_i)$.

Hence, differentiating these parameters according to equation (4.26) results in an expression for the ML estimates of the b_i ($i = 1, \dots, c$):

$$\hat{b}_i = \frac{\sum_{k=1}^n x_k \frac{f_i(x_k; \hat{\theta}_i)}{f(x_k; \hat{\theta})}}{\sum_{k=1}^n \frac{f_i(x_k; \hat{\theta}_i)}{f(x_k; \hat{\theta})}}. \quad (4.29)$$

4.4.2.1 Simulation Results

Two-component mixtures of exponential distributions were resolved using ML estimation. In this case, there were two component parameters, b_1 and b_2 , and a mixing weight, α , to be estimated.

	Initial Values	Mean Parameter Estimates	Variance of Parameter Estimates	True Parameter Values
b_1	0.4393	0.4572	1.969×10^{-4}	0.45
b_2	3.3316	3.3287	0.4846	3.29
α	0.8804	0.8985	1.944×10^{-4}	0.90

Table 4.2: The scale parameters and mixing proportion of a two-component exponential mixture. The estimates were obtained using ML.

The results for a sampling distribution comprising ten samples of 1000 observations are listed in Table 4.2. The mean parameter estimates of the exponential mixture were quite close to the true values, the relative differences being less than 2%. The variance of the second component parameter estimate was several orders of magnitude larger than the variance of the other two parameters. This may be attributed to the comparatively small sampling proportion from the second population. On average, nineteen iterations of the estimating procedure were necessary to achieve convergence.

The population and estimated mixture densities were plotted for comparison. Figure 4.3 reveals that the two are almost indistinguishable when viewed in this manner.

Figure 4.3: The population and estimated mixture distributions of two exponential PDFs with the parameters tabulated in Table 4.2.

Figure 4.4: The average χ^2 metric as a function of sample size for the exponential distribution of Table 4.2.

The goodness of fit of the estimated distribution to the true distribution was evaluated by the χ^2 test. From Figure 4.4, we see that the χ^2 value dropped significantly from 0.17 for $n = 100$ to 0.0037 for $n = 200$, and then remained approximately the same for larger sample sizes.

4.4.3 Gamma Mixtures

The gamma random variable whose PDF is given by (2.10) has the shape and scale parameters a and b , respectively. Thus, $\theta_i = (a_i, b_i)$ for the i th component function. If we assume that the shape parameter is known (but not necessarily common to all components) then we again have a total of $2c-1$ parameters to estimate.

Based on the natural parameters of the gamma PDF defined in Section 2.3.3 we can apply equation (4.26) as before to yield the ML estimates for $i = 1, \dots, c$:

$$\hat{b}_i = \frac{\sum_{k=1}^n x_k \frac{f_i(x_k; \hat{\theta}_i)}{f(x_k; \hat{\theta})}}{a_i \sum_{k=1}^n \frac{f_i(x_k; \hat{\theta}_i)}{f(x_k; \hat{\theta})}}. \quad (4.30)$$

4.4.3.1 Simulation Results

Mixtures of two gamma distributions were also studied under ML estimation. The shape parameter, a , was taken to be the same for both components. Hence, the three parameters of interest were b_1 , b_2 , and α .

	Initial Values	Mean Parameter Estimates	Variance of Parameter Estimates	True Parameter Values
b_1	0.1937	0.2000	1.855×10^{-5}	0.20
b_2	0.7120	0.7671	5.987×10^{-3}	0.80
α	0.8571	0.8864	5.559×10^{-4}	0.90

Table 4.3: The scale parameters and mixing proportion of a two-component gamma mixture including ML estimates. The shape parameter was $a = 2$.

The results presented here are for bell-shaped gamma densities with shape parameter $a = 2$. Table 4.3 lists the mean and variance of the parameter estimates obtained for the mixture. The estimate of b_1 coincided with the population value up to four decimal places. The estimate of b_2 , however, was considerably less accurate. The small variances of the estimates were in the order of 10^{-3} or less. Approximately sixteen iterations were sufficient for convergence of the parameter estimates.

As shown in Figure 4.5, the true and estimated distributions were in close agreement.

The χ^2 metric was calculated for samples ranging in size from $n = 100$ to $n = 1000$. It was found that the goodness of fit improved steadily as the sample size increased. Figure 4.6 shows a 90% drop in the χ^2 value in increasing the number of observations from 100 to 1000.

Figure 4.5: The population and estimated mixture distributions of two gamma PDFs with the parameters listed in Table 4.3.

Figure 4.6: The average χ^2 metric as a function of sample size for the gamma distribution of Table 4.3.

4.4.4 Weibull Mixtures

The Weibull random variable has a similar PDF to that of the gamma random variable as shown by equation (2.12). The former also has shape and scale parameters a and b , and so $\theta_i = (a_i, b_i)$.

Again assuming that the a_i are known, we use the parameters of the distribution in its natural form (2.13) and carry out the differentiation (4.26) to obtain

$$\hat{b}_i^{a_i} = \frac{\sum_{k=1}^n x_k^{a_i} \frac{f_i(x_k; \hat{\theta}_i)}{f(x_k; \hat{\theta})}}{\sum_{k=1}^n \frac{f_i(x_k; \hat{\theta}_i)}{f(x_k; \hat{\theta})}} \quad i=1, \dots, c. \quad (4.31)$$

4.4.4.1 Simulation Results

Simulation studies included two-component mixtures of Weibull distributions. As for the gamma densities, the shape parameter was common to both components.

Using equations (4.8) and (4.31), ML estimates were obtained for the gamma mixture. Table 4.4 reveals excellent estimates of all three parameters. The largest relative difference between the estimated and the true parameters was only 0.1% for the second component. Furthermore, an average of only eight iterations were required to determine these estimates. Figure 4.7 illustrates the estimated and population density functions.

	Initial Values	Mean Parameter Estimates	Variance of Parameter Estimates	True Parameter Values
b_1	0.6863	0.6894	7.850×10^{-5}	0.69
b_2	1.6950	1.6917	5.824×10^{-3}	1.69
α	0.8979	0.8999	1.080×10^{-4}	0.90

Table 4.4: The initial values, ML estimates, and true values of the scale parameters and mixing proportion of a two-component Weibull mixture with shape parameter $a = 2$.

Figure 4.7: The population and estimated mixture distributions of two Weibull PDFs with the parameters shown in Table 4.4.

The χ^2 metric for the Weibull mixture manifested a similar behavior to that of the gamma mixture when plotted as a function of sample size. Figure 4.8 shows the value of

χ^2 decreasing from 0.025 for $n = 100$ to 0.0020 for $n = 1000$. This represents a roughly 92% decrease.

Figure 4.8: The average χ^2 metric as a function of sample size for the Weibull distribution of Table 4.4.

4.4.5 Poisson Mixtures

Discrete mixtures are treated in similar fashion to the foregoing continuous distributions. For example, the Poisson random variable with natural parametric form (2.15) possesses the parameters $t(x) = x$, $c(d) = -\lambda$, and $d(\theta) = \ln(\lambda)$. Consequently, from the first partial derivatives with respect to the λ_i , equation (4.26) leads to the ML estimates for $i = 1, \dots, c$:

$$\hat{\lambda}_i = \frac{\sum_{k=1}^n x_k \frac{P_i(x_k; \hat{\theta}_i)}{P(x_k; \hat{\Theta})}}{\sum_{k=1}^n \frac{P_i(x_k; \hat{\theta}_i)}{P(x_k; \hat{\Theta})}}. \quad (4.32)$$

4.4.5.1 Simulation Results

Two-component Poisson distributions were simulated and their parameters estimated using the ML principle. The three parameters of interest were the two component density parameters λ_1 and λ_2 , as well as the mixing proportion, α .

The results of Table 4.5 show reasonably good parameter estimates for a sampling of ten samples of size $n = 1000$. The largest discrepancy occurred in estimating the parameter λ_1 of the first component. A relative difference of 4.3% between the estimated and true value was calculated for this parameter. The larger variance, however, was evinced by the second parameter estimate which had a deviation of one order of magnitude greater than the first parameter. The average number of procedure iterations needed to reach convergence was just below the maximum of twenty.

A histogram illustrating the pmfs of both the estimated and true mixtures is given in Figure 4.9. The estimated probability is slightly overstated for $X = 0, 1$, and understated for larger values of X .

	Initial Values	Mean Parameter Estimates	Variance of Parameter Estimates	True Parameter Values
λ_1	0.4583	0.4787	3.143×10^{-3}	0.50
λ_2	2.4138	2.4747	0.04691	2.50
α	0.7776	0.7968	1.704×10^{-3}	0.80

Table 4.5: The rate parameters and mixing proportion of a two-component Poisson mixture using ML estimation..

Figure 4.9: The population and estimated mixture distributions of two Poisson pmfs with the parameters tabulated in Table 4.5.

The influence of sample size on the goodness of fit was determined by calculating the EMIM values for sample sizes of 100 to 1000 observations. A rapid decrease of

roughly 76% was observed in increasing the sample size from $n = 100$ to $n = 300$. The EMIM value remained at about 0.0034 for larger samples and then decreased further to 0.0020 for $n = 600$. A slight upward drift in the EMIM metric occurred when the sample size was increased further. The behavior of the EMIM metric with varying sample size is illustrated in Figure 4.10.

Figure 4.10: The average EMIM metric as a function of sample size for the Poisson distribution of Table 4.5.

4.4.6 Binomial Mixtures

The binomial random variable is characterized by the number of trials, m , and the probability, p , of success where $0 \leq p \leq 1$. Hence, for the i th component we have $\theta_i = (p_i, m_i)$. If the number of independent trials, m_i , is known then we have $2c-1$ parameters to estimate in a c -component mixture of binomial densities.

From Section 2.3.6 we obtain the parameters of the distribution when the component functions are written in their natural form. Proceeding with the differentiation (4.26) we have for the c components,

$$\hat{p}_i = \frac{\sum_{k=1}^n x_k \frac{P_i(x_k; \hat{\theta}_i)}{P(x_k; \hat{\theta})}}{m_i \sum_{k=1}^n \frac{P_i(x_k; \hat{\theta}_i)}{P(x_k; \hat{\theta})}}. \quad (4.33)$$

4.4.6.1 Simulation Results

Mixtures of two binomial distributions were also generated for the study of ML estimation. The number of independent trials, m , of the binomial random variable was fixed at the same value for both components. Hence, with $c = 2$, there were three estimates to be determined, namely, p_1 , p_2 , and α .

The results tabulated in Table 4.6 and plotted in Figure 4.11 are for a mixture of binomial densities with $m = 10$. The mean parameter estimates for this mixture are seen to be very close. Variances of the estimates were on the order of 10^{-4} or less. Convergence to these estimates was achieved in roughly sixteen iterations of the method of steepest ascent.

	Initial Values	Mean Parameter Estimates	Variance of Parameter Estimates	True Parameter Values
p_1	0.3970	0.3987	2.191×10^{-4}	0.40
p_2	0.6957	0.6963	5.917×10^{-5}	0.70
α	0.3910	0.3939	9.358×10^{-4}	0.40

Table 4.6: The probabilities and mixing proportion of a two-component binomial mixture including ML estimates. The number of independent trials was $m = 10$ for both densities.

Figure 4.11: The population and estimated mixture distributions of two binomial pmfs with the parameters listed in Table 4.6.

The goodness of fit of the estimated binomial mixture density was measured using the EMIM metric. The value of the metric was calculated for samples ranging in size from

$n = 100$ to $n = 1000$. As shown in Figure 4.12, the fit improved noticeably with increasing sample size. The EMIM value dropped by about 90% from 9.7×10^{-3} for $n = 100$ to 9.7×10^{-4} for $n = 1000$.

Figure 4.12: The average EMIM metric as a function of sample size for the binomial distribution of Table 4.6.

4.4.7 Geometric Mixtures

The geometric probability density defined by (2.18) is a function of the probability, p , $0 \leq p \leq 1$. Based on the natural parameters $t(x) = x$, $c(d) = \ln(p)$, and $d(\theta) = \ln(1-p)$ we can differentiate according to expression (4.26) to obtain for $i = 1, \dots, c$,

$$\hat{\frac{q_i}{p_i}} = \frac{\sum_{k=1}^n x_k \frac{P_i(x_k; \hat{\theta}_i)}{P(x_k; \hat{\Theta})}}{\sum_{k=1}^n \frac{P_i(x_k; \hat{\theta}_i)}{P(x_k; \hat{\Theta})}}, \quad (4.34)$$

where $q_i = 1 - p_i$.

4.4.7.1 Simulation Results

The ML principle was finally tested on mixtures of two geometric distributions. The three parameter estimates obtained were $\hat{p}_1$, $\hat{p}_2$, and $\hat{\alpha}$.

The mean and variance of a sampling distribution of parameter estimates based on ten samples of 1400 observations are shown in Table 4.7. The estimation results for a mixture of geometric pmfs were quite good. The mean estimate of the second component parameter was the least accurate with a relative error of just under 3%. The variance of the parameter estimates was on the order of 10^{-4} . Figure 4.13 compares graphically the estimated and true distributions.

The change in the average EMIM metric with sample size is shown in Figure 4.14. The metric was observed to decrease by approximately 70% from $n = 500$ to $n = 1400$.

	Initial Values	Mean Parameter Estimates	Variance of Parameter Estimates	True Parameter Values
p_1	0.7873	0.7989	2.243×10^{-4}	0.80
p_2	0.1824	0.1941	6.319×10^{-4}	0.20
α	0.9079	0.8981	1.476×10^{-4}	0.90

Table 4.7: The probabilities and mixing proportion of a two-component geometric mixture. The estimates were obtained using the ML method.

Figure 4.13: The population and estimated mixture distributions of two geometric pmfs with the parameters indicated in Table 4.7.

Figure 4.14: The average EMIM metric as a function of sample size for the geometric distribution of Table 4.7.

4.5 Difficulties

There are a number of difficulties of which one must be aware when applying ML estimation to mixture distributions. First, ML equations are not generally characterized by a unique solution. That is, the log-likelihood function, $L(\Theta)$, may possess several local maxima. Solution of the likelihood equations does not necessarily assure that the ML estimate associated with the global maximum of $L(\Theta)$ will be obtained. In fact estimates may correspond to any one of the several local optima of the likelihood surface. In such a case, the only recourse is to determine as many local maxima as possible and to select the largest of the set for the ML estimates [60]. Unfortunately, there appear to be no reports of any concerted effort directed towards establishing an efficient procedure for locating all the local maxima of the likelihood function.

Duda and Hart [24] provide an illustration of a mixture of two univariate normals for which two local maxima of the likelihood function are obtained. The PDF of the mixture is given by

$$f(x; \phi_1, \phi_2) = \frac{1}{3\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x - \phi_1)^2\right] + \frac{2}{3\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x - \phi_2)^2\right].$$

With $(\phi_1, \phi_2) = (-2, 2)$, 25 observations of the distribution were taken and used to evaluate the log-likelihood function at several different values of (ϕ_1, ϕ_2) . It was found that maxima occurred at both $(-2.130, 1.668)$ and $(2.085, -1.257)$ with the former peak being slightly larger. Convergence to either one of these solutions depended on the choice of starting values for the iterative procedure.

A second difficulty associated with ML estimation of mixtures is that a given maximum of $L(\underline{\Theta})$ may be satisfied by more than one choice of $\hat{\underline{\Theta}}$ when the component densities are of the same functional form. This is because interchanging the component parameters $(\alpha_i, \underline{\theta}_i)$ and $(\alpha_j, \underline{\theta}_j), i \neq j$, yields the same mixture density, $f(x|\underline{\Theta})$. That is, if the mixture consists of c component densities then there are $c!$ label permutations in $\underline{\Theta}$ yielding the same value for $f(x|\underline{\Theta})$. In practice this interchangeability is an issue only if one is seeking to identify particular component parameters with a specific subpopulation of the mixture rather than simply resolving the mixture into its constituent parts [24].

A third problem which may arise in mixture density estimation by ML is the existence of singular solutions to the likelihood equations. Most notably, such solutions occur in mixtures of normal densities where the log-likelihood function can increase without bound. Consider a two-component normal compound density from which a sample $X^n = \{x_1, \dots, x_n\}$ is drawn. Then the log-likelihood is given by

$$L(\phi_1, \phi_2, \sigma_1, \sigma_2, \alpha) = \sum_{k=1}^n \log f(x_k; \phi_1, \phi_2, \sigma_1, \sigma_2)$$

where

$$f(x_k; \phi_1, \phi_2, \sigma_1, \sigma_2) = \frac{\alpha}{\sqrt{2\pi} \sigma_1} \exp\left[-\frac{(x_k - \phi_1)^2}{2\sigma_1^2}\right] + \frac{(1-\alpha)}{\sqrt{2\pi} \sigma_2} \exp\left[-\frac{(x_k - \phi_2)^2}{2\sigma_2^2}\right].$$

Now suppose that a sample observation x_k ($k=1, \dots, n$) coincides with one of the component means ϕ_i ($i=1, 2$). We see that if the corresponding variance σ_i^2 is allowed to

approach zero then the log-likelihood function tends to infinity:

$$\lim_{\sigma_1 \rightarrow 0} L(x_k, \phi_2, \sigma_1, \sigma_2, \alpha) = \lim_{\sigma_2 \rightarrow 0} L(\phi_1, x_k, \sigma_1, \sigma_2, \alpha) \rightarrow \infty$$

for $k=1, \dots, n$. Hence, under these conditions the method of ML does not yield a viable solution.

Several authors have studied this difficulty with singular solutions of normal mixtures and have concluded that in practice ML estimation will lead to reliable estimates provided that proper care is taken. Duda and Hart [24] showed that if one focuses on finite maxima, accurate ML estimates can be obtained. Hosmer [38,39] reported that singularities can best avoided by drawing large samples from loosely coupled distributions and ensuring that reasonable initial estimates are used. Other investigators such as Day [18] suggested restricting the use of ML estimation of normal mixtures to those instances where the component variances (or covariance matrices in a multivariate mixture) are equal, thereby eliminating the problem altogether.

4.6 Summary

The estimation of mixture distribution parameters by ML was discussed at length in the present chapter. The general expressions for the ML estimates were derived. It was shown that the ML estimate of a parameter vector is obtained by choosing the vector which maximizes the log of the joint probability density function.

Three different iterative procedures for approximating the solutions to the ML expressions were described. These included the method of steepest ascent, Newton's method, and the method of scoring. Although the last two are normally quicker methods, in practice the method of steepest ascent is typically the procedure of choice due to its ease of implementation and its monotonic convergence to a local maximum. Generally speaking, the choice of initial estimates for the iterative schemes is more crucial for multivariate distributions than for univariate ones.

Several difficulties with estimating mixture parameters using the ML method were explained. In particular, the possibility of several local maxima of the likelihood surface, and the interchangeability of parameter estimates were of concern. In addition, singular solutions to the ML equations were shown to occur in theory for mixtures of normal densities.

ML estimators for mixtures of distributions from the exponential family were derived explicitly. Specifically, these included estimates for mixtures of c normal, exponential, gamma, Weibull, Poisson, binomial, and geometric densities.

Simulation studies were carried out on these mixtures for the two-component case to investigate estimation using the ML principle. Generally, estimates of the parameters showed reasonably good agreement with the population values. Tests performed with

normal mixtures demonstrated that the method of steepest ascent was not sensitive to the choice of initial values in the univariate case.

The χ^2 and EMIM metrics were utilized to assess the goodness of fit of the estimated distributions to the true densities. It was shown that the parameter estimates improved measurably when the number of observations drawn from the population was increased from 100 to 1000 (or 500 to 1400 for the geometric mixture).

CHAPTER 5

ESTIMATION USING OTHER METHODS

5.1 Introduction

The past few chapters have demonstrated the course of research that has been undertaken in estimating the parameters of mixture distributions. Over the years, as mentioned earlier, a number of different methods have been devised to solve the problem. Some of these methods are of a specialized nature in that they have been developed either for specific distribution forms or for particular parameters of the mixture. Also, many investigators focused primarily on normal mixtures due to the practical importance of this distribution [11,15,22,34]. Others concentrated on estimating only the mixing proportions of mixtures for which the component densities were known [12,52,55]. Still others were concerned with only obtaining a rough estimate of the parameters using simple approaches. An example of such an approach is the class of graphical methods mentioned briefly earlier. These methods were predominant before the introduction of the digital computer. They too were principally directed towards normal mixtures and required a fairly high degree of subjectivity [5,10,29,33].

Of the general estimation methods, two of them have been discussed in detail in the two preceding chapters. These are the method of moments and the method of maximum likelihood. The former is the older method and was widely used until the late 1960's. Since then, the latter method is normally preferred due to the more favorable properties of its estimators. Several other general techniques exist, however, and these will be discussed in the present chapter.

A number of methods have been constructed which have their basis in the method of moments. Two of these are presented in the next section. The first constrains the

sample space to a finite interval divided into $2c$ equal subintervals, and the second makes use of fractional moments. Another group of methods, categorized as error minimization techniques, is subsequently described. Lastly, Bayesian estimation, a method which is of greater theoretical than practical significance for mixture distributions, is explained.

5.2 Moment-Like Methods

A number of investigators have sought to improve the asymptotic properties of the moment estimators by developing techniques based on modifications of the method of moments. The manner in which the various approaches deviate from the method of moments varies appreciably from one technique to another.

One technique imposes a constraint in which all sample observations are assumed to fall within a finite interval. A statistic based on this interval rather than on the moments is utilized to estimate the parameters of the component densities. The estimates of the mixing proportions, however, are derived from the first c moments. The system of estimating equations for the mixture parameters resembles quite closely that of the method of moments.

Another technique equates sample and population moments in the same manner as the regular moments method in order to derive the estimators. However, fractional moments, that is moments of fractional orders, are employed in this case. The estimates of the component density parameters and the mixing weights are approximated using an iterative procedure.

In the next two subsections we shall discuss the salient features of these latter techniques.

5.2.1 Finite Intervals

Kabir [42] reported a method for the estimation of the parameters of c -component mixtures of distributions from a subset of the exponential family. In Section 2.3 it was shown that this family of distributions in the natural form is denoted in general by

$$f(x) = g(x) \exp[d(\theta)t(x) + c(d(\theta))], \quad x \geq 0.$$

Consider a mixture of c densities from the exponential family when $t(x) = x$. Such a mixture is expressed as

$$f(x; \Theta) = \sum_{i=1}^c \alpha_i g(x) \exp[d_i(\theta)x + c(d_i(\theta))], \quad (5.1)$$

where α_i represents the usual mixing weight

$$\sum_{i=1}^c \alpha_i = 1, \quad \alpha_i > 0,$$

and θ_i denotes the i th component parameter with the constraint $\theta_1 < \dots < \theta_c$. Further, assume that the sample observations $\{x_1, \dots, x_n\}$ fall in a finite interval $[a, b]$ such that the population is given by equation (5.1), or zero otherwise.

The mixture distribution of (5.1) can be written as

$$f(x; \Theta) \cdot g(x)^{-1} = \sum_{i=1}^c \alpha_i \exp[c(d_i(\theta))] \cdot \exp[d_i(\theta)x].$$

Let

$$p(x) = f(x; \Theta) g(x)^{-1},$$

$$\eta_i = \alpha_i \exp[c(d_i(\theta))],$$

and

$$B_i = \exp[d_i(\theta)]. \quad (5.2)$$

Then, using the above substitutions, expression (5.1) becomes

$$p(x) = \sum_{i=1}^c \gamma_i B_i^x, \quad x \in [a, b]. \quad (5.3)$$

We must now consider separately the two cases when the mixture distribution, $f(x; \Theta)$, is continuous or discrete.

(a) Continuous Mixture

In this case, the finite interval $[a, b]$ is divided into $2c$ equal subintervals, I_k ($k = 0, \dots, 2c-1$) by $a = a_0, a_1, \dots, a_{2c} = b$. Thus, we have

$$I_k = (a_k, a_{k+1}), \quad k = 0, 1, \dots, 2c-1$$

where $a_{k+1} - a_k = h$.

Take T_k to be a population statistic defined by

$$T_k = \int_{a_k}^{a_{k+1}} p(x) d(x), \quad k = 0, \dots, 2c-1 \quad (5.4)$$

Then, from (5.3) we have

$$T_k = \sum_{i=1}^c \gamma_i \int_{a_k}^{a_{k+1}} B_i^x d(x).$$

Evaluating the integral leads to

$$T_k = \sum_{i=1}^c \frac{\gamma_i}{\ln(B_i)} [B_i^{a_0+h} - B_i^{a_0}] B_i^{kh}.$$

If we then make the substitutions

$$Q_i = \frac{\gamma_i}{\ln(B_i)} [B_i^{a_0+h} - B_i^{a_0}],$$

and

$$\xi_i = B_i^h, \quad (5.5)$$

we obtain

$$T_k = \sum_{i=1}^c Q_i \xi_i^k, \quad k = 0, \dots, 2c-1. \quad (5.6)$$

(b) Discrete Mixture

In the discrete case, the interval $[a, b]$ is taken to comprise j integers grouped into $2c$ equal subintervals each containing $j - 2c + 1$ elements. Thus, the k th interval, I_k , encompasses the integers

$$(a + k, a + k + 1, \dots, a + j - 2c + k) \quad k = 0, \dots, 2c-1.$$

Again, using the population statistic T_k , we have for the discrete case

$$T_k = \sum_{x \in I_k} p(x) \quad k = 0, \dots, 2c-1. \quad (5.7)$$

Using (5.3) yields

$$T_k = \sum_{i=1}^c \gamma_i \sum_{x \in I_k} B_i^x$$

or

$$T_k = \sum_{i=1}^c \gamma_i (B_i^a + B_i^{a+1} + \dots + B_i^{a+j-2c}) B_i^k.$$

Let

$$Q_i = \gamma_i (B_i^a + B_i^{a+1} + \dots + B_i^{a+j-2c}),$$

and

$$\xi_i = B_i. \quad (5.8)$$

Then we have an equation similar to (5.6) for discrete mixtures:

$$T_k = \sum_{i=1}^c Q_i \xi_i^k, \quad k = 0, \dots, 2c-1. \quad (5.9)$$

Kabir [42] demonstrated that a consistent sample estimator, τ_k , of the continuous or discrete theoretical statistic, T_k , is given by

$$\tau_k = \frac{1}{n} \sum_{x \in I_k} [g(x)^{-1}], \quad (5.10)$$

where $g(x)$ is defined in (5.1). Thus, substituting τ_k for T_k in expression (5.6) or (5.9), a system of $2c$ estimating equations can be constructed.

Note that the systems of equations (5.6) and (5.9) are of the same form as those of (3.40) and (3.50) derived for the method of moments. Hence, drawing the analogy with the method of moments, (5.6) and (5.9) can be reduced to a single c -degree polynomial

$$\xi^c + \beta_{c-1}\xi^{c-1} + \dots + \beta_1\xi + \beta_0 = 0, \quad (5.11)$$

where β_i ($i = 1, \dots, c$) are constants of the polynomial with ordered roots $\xi_1, \dots, \xi_c$.

Using similar arguments to those in Section 3.5, it can be shown that the constants of the polynomial are given by the matrix equation

$$\mathbf{T} \mathbf{\beta} = \mathbf{L} \quad (5.12)$$

$$\text{where } \mathbf{T} = \begin{bmatrix} T_0 & T_1 & \dots & T_{c-1} \\ T_1 & T_2 & \dots & T_c \\ \vdots & \vdots & \ddots & \vdots \\ T_{c-1} & T_c & \dots & T_{2c-2} \end{bmatrix}, \quad \mathbf{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{c-1} \end{bmatrix}, \quad \mathbf{L} = \begin{bmatrix} T_c \\ T_{c+1} \\ \vdots \\ T_{2c-1} \end{bmatrix}.$$

Once the coefficients, β_k ($k = 0, \dots, c-1$), have been determined, the c roots, ξ_k ($k = 1, \dots, c$), can be found from equation (5.11). Finally, using (5.5) for continuous mixtures, or (5.8) for discrete mixtures, the estimates of the component density parameters are given by (5.2).

The mixing weight estimates, $\hat{\alpha}_i$ ($i = 1, \dots, c$), are determined by using the first c moment equations for the method of moments. Hence, using an expression similar to (3.44) or (3.53), the estimated mixing proportions are found from

$$\mathbf{\Theta} \mathbf{\alpha} = \mathbf{1} \quad (5.13)$$

where

$$\tilde{\Theta} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ \theta_1 & \theta_2 & \dots & \theta_c \\ \theta_1^2 & \theta_2^2 & \dots & \theta_c^2 \\ \vdots & \vdots & \ddots & \vdots \\ \theta_1^{c-1} & \theta_2^{c-1} & \dots & \theta_c^{c-1} \end{bmatrix}, \quad \alpha = \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_{c-1} \end{bmatrix}, \quad \mathbf{I} = \begin{bmatrix} T_0 \\ T_1 \\ \vdots \\ T_{c-1} \end{bmatrix}.$$

Both the estimators $\hat{\theta}_i$ and $\hat{\alpha}_i$ are consistent and asymptotically normally distributed.

As an example, consider a two-component mixture of exponential distributions given here in the natural form of equation (5.1):

$$f(x; b_1, b_2, \alpha) = \sum_{i=1}^2 \alpha_i \exp[-x/b_i - \ln(b_i)].$$

Thus, $g(x) = 1$ and we have from (5.10)

$$\tau_k = \frac{1}{n} \sum_{x \in I_k} = \frac{n_k}{n},$$

where n_k is the number of observations in the k th interval. Also, we have

$$B_i = \exp\left[-\frac{1}{b_i}\right].$$

From (5.12), we find the coefficients, β_0 and β_1 , of the quadratic

$$\xi^2 + \beta_1 \xi + \beta_0 = 0,$$

whose roots, ξ_1, ξ_2 , are readily found.

Using (5.5) and (5.2) then, we find that

$$\xi_i = B_i^h = \exp\left[-\frac{h}{b_i}\right]$$

so that the i th component parameter estimate is given by

$$\hat{b}_i = -\frac{h}{\ln(\xi_i)} \quad i = 1, 2.$$

The mixing weight estimate is found using the first moment equation, that is

$$m_1 = \hat{\alpha} \hat{b}_1 + (1 - \hat{\alpha}) \hat{b}_2,$$

or

$$\hat{\alpha} = \frac{m_1 - \hat{b}_2}{\hat{b}_1 - \hat{b}_2}.$$

5.2.2 Fractional Moments

Tallis and Light [67] devised an estimation technique based on the method of moments whereby the efficiency of the estimators is improved relative to the regular moment estimators by the optimal selection of fractional moment orders.

Consider the mixture density $f(x; \Theta)$ with parameters $\Theta = (\theta_1, \dots, \theta_c, \alpha_1, \dots, \alpha_c)$ where θ_i and α_i are the i th component density parameter and mixing proportion, respectively. Then the theoretical moment of order a_k is given by

$$\mu_{a_k}(\Theta) = \int_{-\infty}^{\infty} |x|^{a_k} f(x; \Theta) dx, \quad k = 1, \dots, 2c, \quad (5.14)$$

and the observed counterpart is determined by

$$m_{a_k} = \frac{1}{n} \sum_{j=1}^n |x_j|^{a_k}, \quad k = 1, \dots, 2c. \quad (5.15)$$

In both (5.14) and (5.15), a_k is a positive, real number.

A system of $2c$ estimating equations is constructed by equating the above population and sample moments. Thus, we have the vector equation

$$\mu(\Theta) = m. \quad (5.16)$$

Expressing (5.16) as

$$\Delta(\Theta) = m - \mu(\Theta),$$

an iterative scheme based on Newton's method can be derived by expanding $\Delta(\Theta)$ in a Taylor series and neglecting higher order term (cf. §4.3.2). As a result, estimates of the mixture parameters can be approximated by

$$\Theta^{(v+1)} = \Theta^{(v)} + H(\Theta)^{-1} [m - \mu(\Theta^{(v)})], \quad |H(\Theta)| \neq 0. \quad (5.17)$$

The k,j term of the matrix $H(\Theta)$ is given by

$$h_{k,j}(\Theta) = \frac{\partial \mu_{a_k}}{\partial \zeta_j}, \quad k,j = 1, \dots, 2c,$$

where for convenience, we have adopted the notation, $\Theta = (\zeta_1, \dots, \zeta_{2c})$.

Tallis and Light [67] determined empirically the optimal choice of moment orders as a function of the parameter values for mixtures of two exponential densities. In this case, the mixed PDF is given by

$$f(x; b_1, b_2, \alpha) = \alpha b_1^{-1} \exp[-x/b_1] + (1 - \alpha) b_2^{-1} \exp[-x/b_2].$$

They found that significant improvements in the efficiency of the moment estimators could be achieved by selecting fractional moment orders. For example, for $b_1/b_2 = 4$ and $\alpha = 0.3$, the usual moments with orders $a_1 = 1$, $a_2 = 2$, $a_3 = 3$ lead to an efficiency of only 0.138. However, choosing $a_1 = 1$, $a_2 = 0.75$, $a_3 = 1.25$ results in a marked increase in efficiency to 0.671.

5.3 Error Minimization Methods

There exists a large number of estimation methods which can be collectively called error minimization methods [26,71]. In general, the approach taken by these methods is to minimize the sum of differences between a "true" function and one estimated from the sample data.

Consider two functions, $F(\underline{\Theta})$ and $G(X^n)$, where the former is a population function dependent on the parameter vector, $\underline{\Theta}$, and the latter is the corresponding observed function based on the sample $X^n = \{x_1, \dots, x_n\}$. Then we can define an error function

$$\Delta[F(\underline{\Theta}), G(X^n)]. \quad (5.18)$$

A minimum error estimator $\hat{\underline{\Theta}}$ is the choice of $\underline{\Theta}$ for which (5.18) is minimized.

Minimization of the error function is achieved in the usual manner by differentiating $\Delta(F,G)$ with respect to the parameters $\underline{\Theta}$ and equating this to zero. Thus, we have

$$\nabla_{\underline{\Theta}} \Delta(\hat{\underline{\Theta}}) = 0 \quad (5.19)$$

Generally, mixture distributions involve nonlinear functions of the parameter vector $\underline{\Theta}$, and hence, explicit expressions for the estimator $\hat{\underline{\Theta}}$ are not possible. As a result, a numerical procedure is required to yield the parameter estimates. In a fashion analogous to

the Newton procedure derivation for ML (cf. §4.3.2), we can carry out a Taylor expansion of (5.19) about the parameter $\underline{\Theta}$. Neglecting higher order terms, we obtain a correction, ϵ , to the v th approximation to $\underline{\Theta}$ as:

$$\epsilon \approx \frac{-\nabla_{\underline{\Theta}} \Delta(\underline{\Theta}^{(v)})}{\nabla_{\underline{\Theta}} \nabla_{\underline{\Theta}}^T \Delta(\underline{\Theta}^{(v)})}.$$

Thus, the parameter $\underline{\Theta}$ can be estimated by iterating

$$\underline{\Theta}^{(v+1)} = \underline{\Theta}^{(v)} - [\nabla_{\underline{\Theta}} \nabla_{\underline{\Theta}}^T \Delta(\underline{\Theta}^{(v)})]^{-1} \nabla_{\underline{\Theta}} \Delta(\underline{\Theta}^{(v)}), \quad v = 0, 1, \dots \quad (5.20)$$

The functions F and G used in the error function can be either probability density functions (PDF), cumulative distribution functions (CDF), or transformations of the CDFs. In the first case, we denote the PDF of the mixture population in the usual way by

$$f(x; \underline{\Theta}) = \sum_{i=1}^c \alpha_i f_i(x; \theta_i).$$

The CDF of the mixture then is expressed as

$$F(x; \underline{\Theta}) = \int_{-\infty}^{\infty} f(x; \underline{\Theta}) dx = \sum_{i=1}^c \alpha_i \int_{-\infty}^{\infty} f_i(x; \theta_i) dx.$$

In the case of transforms of the distributions, consider some variable $s \in \mathbb{R}$ and a function $T(\cdot)$ such that

$$T(s | \underline{\Theta}) = E[t(s, X)] = \int_{-\infty}^{\infty} t(s, x) f(x; \underline{\Theta}) dx.$$

Further, assume that $t(s, X)$ has a finite variance for all $s \in \mathbb{R}$, and that a function

$$\hat{t}_n(s) = \frac{1}{n} \sum_{j=1}^n t(s, x_j),$$

exists. Then for large sample sizes, n , $\hat{i}_n(s)$ approaches $T(s | \Theta)$. Thus, parameter estimation by error minimization using transforms is accomplished by minimizing

$$\Delta[\hat{i}_n(s), T(s | \Theta)].$$

Note that in all three instances either continuous or discrete mixture distributions are applicable, where in the latter case a summation replaces the integration.

Many different variations of the error function are reported, but for the most part, the error is evaluated using the sums of squares. Table 5.1 lists several of the more common error functions for both continuous and discrete CDFs and PDFs [71]. Transforms of the distributions typically include the following: the characteristic function, the moment generating function, and the probability generating function (for discrete densities). Any one of these can also be used as arguments to the error functions tabulated in Table 5.1.

Continuous Distributions	Discrete Distributions
$\int [F(x) - G(x)]^2 dx$	$\sum_{i=1}^I \left(\sum_{j=1}^i p_j - \sum_{j=1}^i q_j \right)^2$
$\int [F(x) - G(x)]^2 w(x) dx$	$\sum_{i=1}^I \left(\sum_{j=1}^i (p_j - q_j) \right)^2 w_i$
$\int [F(x) - G(x)]^2 dF(x)$	$\sum_{i=1}^I \left(\sum_{j=1}^i (p_j - q_j) \right)^2 p_i$
$\int \frac{[f(x) - g(x)]^2}{f(x)} dx$	$\sum_{i=1}^I \frac{(p_i - q_i)^2}{p_i}$

Table 5.1: Error functions typically used to evaluate the squared difference between CDFs or PDFs. Note that w is an arbitrary weight function.

5.4 Bayesian Method

In all of the estimation methods considered so far, the parameter vector Θ is regarded as unknown and fixed. Further, any information on the possible values of Θ available prior to undertaking the estimation is not considered. The sole exception is the method of ML which utilizes *a priori* information to obtain initial parameter estimates for its iterative solution [24].

In contrast, Bayesian estimation methods adopt the view that Θ is an unknown random vector. Prior knowledge about the vector is expressed in the known PDF, $f(\Theta)$. Using the sample data $X^n = \{x_1, \dots, x_n\}$, the aim of Bayesian estimation is to transform the *a priori* density, $f(\Theta)$, into the *a posteriori* density, $f(\Theta | X^n)$:

$$f(\Theta | X^n) = \frac{f(X^n | \Theta) f(\Theta)}{\int f(X^n | \Theta) f(\Theta) d\Theta} \quad (5.21)$$

Equation (5.21) represents Bayes rule, where the integration in the denominator is performed over the entire parameter space. If the n observations $\{x_1, \dots, x_n\}$ are drawn independently from the mixture population, $f(x | \Theta)$, then $f(X^n | \Theta)$ denotes the joint PDF

$$f(X^n | \Theta) = \prod_{k=1}^n f(x_k | \Theta). \quad (5.22)$$

The *a posteriori* PDF of (5.21) can be expressed equivalently as a recursion. From equation (5.22) we have the identity

$$f(X^n | \Theta) = f(x_n | \Theta) f(X^{n-1} | \Theta).$$

Replacing the joint PDF in (5.21) by the above expression and reapplying Bayes rule yields

$$f(\underline{\Theta} | X^n) = \frac{f(x_n | \underline{\Theta}) f(\underline{\Theta} | X^{n-1})}{\int f(x_n | \underline{\Theta}) f(\underline{\Theta} | X^{n-1}) d\underline{\Theta}}. \quad (5.23)$$

Thus, using (5.23) one can generate a sequence $f(\underline{\Theta})$, $f(\underline{\Theta} | x_1)$, $f(\underline{\Theta} | x_1, x_2)$, etc. with each additional sample observation. In general, provided the mixture distribution $f(x | \underline{\Theta})$ is identifiable (cf. §2.2.3), the *a posteriori* density $f(\underline{\Theta} | X^n)$ can be made progressively sharper with increased sampling until it converges to a Dirac delta function at the “true” parameter vector $\underline{\Theta}$.

Although the foregoing scheme appears practicable, the estimation of mixture parameters using the Bayesian method presents a major difficulty. This difficulty arises in the computational complexity of the *a posteriori* PDF (5.21). As shown in (5.22), the function $f(X^n | \underline{\Theta})$ denotes the joint PDF of the n mixture densities. Thus, expanding (5.22), we have

$$f(X^n | \underline{\Theta}) = \prod_{k=1}^n \left[\sum_{i=1}^c \alpha_i f_i(x_k | \underline{\theta}_i) \right]. \quad (5.24)$$

In other words, for a c -component mixture of distributions from which n independent sample observations are drawn, the joint PDF constitutes a sum of c^n products of component densities and mixing proportions. Clearly, the *a posteriori* density $f(\underline{\Theta} | X^n)$ quickly becomes increasingly complex as the number of sample values, n , increases.

As a result of the computational difficulty with the Bayesian estimation method, it is not normally used in practice to resolve mixture distributions. Some investigators, however, have demonstrated some limited success using the scheme in simplified cases. For example, Smith and Makov [65] have obtained an approximate solution to the Bayesian estimation problem when only the mixing weight is of interest.

5.5 Summary

In this chapter, a number of generally applicable estimation methods for mixtures of distributions were presented. In one method, a generalized statistic defined over a finite interval of the sample space was employed. Equating the theoretical and observational statistics, a system of equations was created, which, by inversion, led to component density parameter estimates. The estimates of the mixing proportions were obtained by inversion of a system of moment equations.

A second method which utilized population and sample moments to create a system of estimating equations in similar fashion to the method of moments was also discussed. In this method, however, fractional orders of moments were used to significantly increase the efficiency of the estimators compared to those obtained from integer moments.

Several estimation methods which consisted of variations on a technique known as error minimization were also described. The objective of the methods was to estimate the parameters by minimizing a sum of squares of differences between PDFs, CDFs or transforms of the mixture distribution.

Finally, Bayesian estimation in the context of mixtures of distributions was explained. Generally speaking, the calculation of the *a posteriori* probability density, $f(\Theta|X^n)$, is observed to be computationally prohibitive.

CHAPTER 6

SUMMARY AND CONCLUSIONS

The subject of study of this thesis has been the parameter estimation of finite mixture distributions. Such distributions are denoted by

$$f(x; \underline{\theta}, \underline{\alpha}) = \sum_{i=1}^c \alpha_i f_i(x; \underline{\theta}_i). \quad (6.1)$$

The PDF (6.1) defines a composite population of two or more subpopulations. The mixing proportions, α_i , of the component densities, $f_i(\cdot)$, are themselves probabilities, and therefore, are positive and sum to one.

Of central interest in dealing with mixture distributions, is the decomposition of the mixture into its component densities. The decomposition is effected by estimating the vector of component parameters, $\underline{\theta} = (\underline{\theta}_1, \dots, \underline{\theta}_c)$, and the vector of mixing weights, $\underline{\alpha} = (\alpha_1, \dots, \alpha_c)$. The estimation of these quantities is based on the knowledge of the functional forms of the component densities, $f_i(\cdot)$, and on a sample of observations, $X^n = \{x_1, \dots, x_n\}$, drawn from the mixed population of interest.

Mixtures of distributions of both theoretical and practical interest are typically composed of densities from the exponential family. These densities, characterized by an r -parameter vector, $\underline{\theta} = (\theta_1, \dots, \theta_r)$, are expressed in the natural form

$$f(x; \underline{\theta}) = g(x) \exp \left[\sum_{h=1}^s d_h(\underline{\theta}) t_h(x) + c(d_1(\underline{\theta}), \dots, d_s(\underline{\theta})) \right], \quad 1 \leq r \leq s,$$

where $g(x) > 0$ is independent of the natural parameter $d(\underline{\theta})$. Common instances of continuous densities from the exponential family include the normal, exponential, gamma,

and Weibull distributions. Discrete densities frequently encountered include Poisson, binomial, and geometric distributions.

Prior to undertaking the estimation of parameters of a mixed distribution, one must consider the identifiability of the mixture density. That is, one must ensure the existence of parameter vectors, $\underline{\theta}$ and $\underline{\alpha}$, which uniquely characterize the mixture. Thus, if two or more distinct sets of parameters can give rise to the same mixture, the estimation is no longer a well defined procedure. It was discussed that, in general, mixtures of distributions from the exponential family presented no difficulty in this regard. In particular, the continuous densities considered here are all identifiable. Further, the discrete mixtures of Poisson and geometric densities were also identifiable. The only word of caution concerned mixtures of binomial densities, $\text{Bin}(m, p_i)$, which are identifiable provided that the condition $m \geq 2c-1$ is met.

Several generally applicable parameter estimation methods for mixture distributions exist. The two most widely accepted approaches to parameter estimation were derived and discussed in detail in this work. These are the method of moments (MM) and the method of maximum likelihood (ML). Although both of these methods can readily be applied to the estimation of parameters of single distributions, the extension of these methods to the mixture case is not straightforward.

The basis of parameter estimation by the MM is primarily one of providing a system of estimating equations constructed by equating theoretical moments to their sample counterparts. The number of moment equations in the system is equivalent to the number of parameters to be estimated. Since the population moments constitute functions of the parameters of interest, solving the system of equations in terms of the parameters yields the desired estimates.

It was shown that estimating the parameters of a normal mixture presents a particularly challenging task from a computational perspective. For a c -component mixture, there are $3c-1$ parameters to be estimated, and hence, $3c-1$ moment equations are to be manipulated. Following a considerable amount of algebra, this system of equations can be reduced to a single polynomial. Even for a simple two-component univariate normal mixture, the solution of a nonic equation is required to obtain the five desired parameter estimates. An iterative technique due to Cohen [15] for the two-component case which only requires the solution of a cubic estimating equation was subsequently described.

The parameter estimation of other continuous mixtures was also discussed. In particular, expressions for mixtures of exponential, gamma, and Weibull densities were derived. In these cases, the shape parameter, a , is considered known and common to all components. Hence, in these cases, $2c-1$ parameters are to be estimated, which are indeed, the component density parameters, b_i , and the mixing proportions, α_i , ($i=1, \dots, c$). Thus, a system of $2c-1$ moment equations is set up by equating non-central population and observational moments. The resulting system is of the form

$$\phi_k = \sum_{i=1}^c \alpha_i b_i^k, \quad k = 0, 1, \dots, 2c-1 \quad (6.2)$$

where $\phi_k = m_k \cdot [h(k, a)]^{-1}$ denotes the product of the k th non-central sample moment and the inverse of a function of a and k which depends on the particular form of the component densities. This system can be reduced to a single c -degree polynomial whose c ordered roots are the estimates of the component parameters of interest. Based on the system (6.2), it was shown that the coefficients of the polynomial are given by a matrix equation whose terms are functions of the non-central moments. Once the coefficients are determined, the roots of the polynomial can be calculated to yield the estimates, $\hat{b}_i$, of the component

density parameters. Lastly, the mixing proportions are found from the first $c-1$ moment equations (6.2) using the sample moments and the estimated scale parameters, $\hat{b}_i$.

The estimation of the parameters of discrete distributions was found to be remarkably similar. Attention was focused on Poisson, binomial, and geometric distributions. Again, a system of $2c-1$ moment equations is required to effect the estimation. In the discrete case, however, factorial rather than non-central moments afford the least cumbersome solution. Equating theoretical and sample moments in the usual manner, a system of equations of the same form as (6.2) is derived. Thus, the same solutions as for the continuous distributions can be applied to the discrete mixtures, taking care to use the appropriate function, $h(k,p)$, particular to the functional form of the given component densities.

It was found that a significant concern in utilizing the MM for mixture distributions is the possibility of producing indeterminate estimates. In such cases, the estimates fall outside of their parameter spaces. For instance, estimates, $\hat{\alpha}_i$, of the mixing weights or $\hat{p}_i$, of the Bernoulli probabilities may not be in the interval $(0,1)$. Also, estimates of the scale parameters may be negative. Moreover, the MM may yield imaginary values for any of the estimates.

The principle of ML essentially states that the best choice of parameter estimate, $\hat{\Theta}$, is the parameter vector which maximizes the log of the joint PDF of the mixture. The log of the joint PDF is commonly known as the log likelihood function, $L(\Theta)$. Thus, maximization of the log likelihood function is achieved by the differentiation of $L(\Theta)$ with respect to Θ and setting this equal to zero to define the likelihood equations. Rearranging the resulting equations, one obtains the expressions for the ML estimates of the component density parameters, $\hat{\theta}_i$, and of the mixing proportions, α_i .

Typically, the expressions for the ML estimates of mixture distributions can not be obtained in a closed form. Hence, iterative procedures are necessary to approximate the solution. The typical iterative schemes used in this capacity include the method of steepest ascent (EM algorithm), Newton's method, and the method of scoring. These were described and their relative convergence properties discussed. It was shown that the comparative ease with which the method of steepest ascent is implemented renders it a common procedural choice. Moreover, the method is known to converge monotonically to local maxima of the likelihood surface.

ML estimators for mixtures of distributions from the exponential family were derived using the natural forms of the component densities. The selected densities included the same continuous and discrete PDFs studied under the MM estimation. The construction of the estimators was shown to be quite straightforward using this form, unlike the comparatively more laborious derivation of the MM solution. However, from a real-time computational perspective, the MM was at least an order of magnitude faster than the ML method. This was a consequence of the analytic nature of the former technique which is inherently more rapid than the latter's numerical scheme.

Several difficulties associated with the ML method were noted. It was explained that the log-likelihood function $L(\Theta)$ may manifest more than one local maximum. Solutions to the likelihood equations do not necessarily yield the parameter estimates corresponding to the global maximum of the likelihood surface. Hence, it may be necessary to determine several sets of estimates, each associated with a different local maximum, and selecting the set corresponding to the largest maximum.

Different local maxima may be attained by changing the initial parameter estimates, the choice of which presents another difficulty. Several alternatives exist for making this choice, including *a priori* knowledge of the problem at hand and the use of approximated

MM estimates. Generally, it was found that the choice of initial estimates was not critical for univariate mixtures. Others have demonstrated that multivariate mixtures may be considerably more affected by this choice [26].

Another difficulty arises with mixtures of normal distributions. It was shown that, in theory, ML estimates of these densities may lead to singular solutions since the log-likelihood function is unbounded. In practice, however, this did not appear to pose a problem.

Extensive simulations of mixture distributions were performed in order to investigate the properties of the MM and ML estimation. Random variables of seven different distributions from the exponential family were simulated for use in mixtures. The two principal methods on which the generation of these random variables was based were the transformation and the rejection algorithms. Mixtures of distributions were effected by first selecting a component i with a probability specified by the mixing proportion, α_i . A random variable, X , was then chosen according the probability law given by the component density, $f_i(x; \theta_i)$.

Samples of 100 to 1000 observations were generally drawn from the mixed population of interest in order to determine estimates of the parameters. Sampling means and variances based on 10 experiments of 1000 observations each were reported for both estimation methods. It was shown that estimates of the parameters were generally good for both techniques. Overall, however, ML estimates demonstrated better precision.

The goodness of fit of the estimated distributions to the true distributions were also evaluated using the χ^2 and EMIM metrics. These metrics were averaged over 10 experiments for 10 different sizes ranging from 100 to 1000 observations. It was shown for both estimation methods that the fit of the estimated distribution improved significantly with increasing sample size. In general, this improvement occurred at approximately the

same rate for both techniques. ML estimates, however, tended to yield estimated distributions with a slightly better fit than those of the MM over the entire range of sample sizes. For example, the gamma mixture estimated using the method of moments had χ^2 values ranging from 0.023 to 0.0052 for $n=100$ to $n=1000$. In the same sample size range, the χ^2 values varied from 0.015 to 0.0025 when the ML method was used for this same mixture. These changes in the average χ^2 metric represented approximately 78% and 83% drops, for each method respectively (cf. Figs 3.6 and 4.6).

Several other general estimation methods were also briefly discussed. These included two moment-like methods, one of which utilized a statistic based on finite sample intervals, and another which employed fractional moments. In both cases, theoretical and sample statistics were equated to generate a system of estimating equations. The resulting estimators were consistent and asymptotically normal.

A number of estimation techniques were collectively called error minimization methods. The basic approach in these methods is to minimize a sum of differences or square of differences between a "true" function and the corresponding sample function. Functions used for such a purpose include PDFs, CDFs, and transforms of the PDFs.

Lastly, Bayesian estimation in the context of mixture distributions was discussed. It was shown that computation of the *a posteriori* probability $p(\Theta | X^n)$, was prohibitive due to the complexity added to the calculations by the joint mixture density, $p(X^n | \Theta)$.

REFERENCES

- [1] J. Behboodian, On a Mixture of Normal Distributions, *Biometrika*, 57:215-217 (1970).
- [2] J. Behboodian, On the Modes of a Mixture of Two Normal Distributions, *Technometrics*, 12:131-139 (1970).
- [3] J. Behboodian, Information Matrix for a Mixture of Two Normal Distributions, *J. Statist. Comp. Simul.*, 1:295-314 (1972).
- [4] O. Barndorff-Nielsen, Identifiability of Mixtures of Exponential Families, *J. Math. Anal. Appl.*, 12:115-121 (1965).
- [5] C.T. Bhattacharya, A Simple Method of Resolution of a Distribution into Gaussian Components, *Biometrics*, 23:115-137 (1967).
- [6] W.R. Blischke, Moment Estimators for the Parameters of a Mixture of Two Binomial Distributions, *Ann. Math. Statist.*, 33:444-454 (1962).
- [7] W.R. Blischke, Estimating the Parameters of Mixtures of Binomial Distributions, *J. Amer. Statist. Assoc.*, 59:510-528 (1964).
- [8] W.R. Blischke, Mixtures of Discrete Distributions in *Classical and Contagious Discrete Distributions* (Ed. G.P. Patil), Pergamon Press, New York, 1965, 351-372.
- [9] W.R. Blischke, Mixtures of Distributions in *International Encyclopedia of Statistics*, vol. 1, (Eds. W.H. Kruskal and J.M. Tanur), The Free Press, New York, 1978, 174-180.

-
- [10] R.M. Cassie, Some Uses of Probability Paper in the Analysis of Size of Frequency Distributions, *Austral. J. Mar. Fish. Freshw. Res.*, 5:513-522 (1954).
- [11] C.V.L. Charlier, S.D. Wicksell, On the Dissection of Frequency Functions, *Arkiv f. Matematik Astron. och Fysik.*, Bd. 18, No. 6 (1924).
- [12] K. Choi, W.G. Bulgren, An Estimation Procedure for Mixtures of Distributions, *J. Royal Statist. Soc. B*, 30:444-460 (1968).
- [13] A.C. Cohen, Estimation in Mixtures of Discrete Distributions in *Classical and Contagious Discrete Distributions* (Ed. G.P. Patil), Pergamon Press, New York, 373-378.
- [14] A.C. Cohen, Discussion of a Paper by V. Hasselblad, *Technometrics*, 8:445-446 (1966).
- [15] A.C. Cohen, Estimation in Mixtures of Two Normal Distributions, *Technometrics*, 9:15-28 (1967).
- [16] D.B. Cooper, P.W. Cooper, Nonsupervised Adaptive Signal Detection and Pattern Recognition, *Inform. Control*, 7:416-444 (1964).
- [17] G.R. Dattatreya, L.N. Kanal, Estimation of Mixing Probabilities in Multiclass Finite Mixtures, *IEEE Trans. Syst., Man, Cybern.*, 20:149-158 (1990).
- [18] N.E. Day, Estimating the Components of a Mixture of Normal Distributions, *Biometrika*, 56:463-474 (1969).
- [19] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum-Likelihood from Incomplete Data Via the EM Algorithm, *J. Royal Statist. Soc. Ser. B (methodological)*, 39:1-38 (1977).

-
- [20] P.A. Devijver, J. Kitter, *Pattern Recognition: A Statistical Approach*, Prentice-Hall International, Englewood Cliffs, 1982.
- [21] L. Devroye, *Non-Uniform Random Variate Generation*, Springer-Verlag, New York, 1986.
- [22] N.P. Wick, D.C. Bowden, Maximum Likelihood Estimation for Mixtures of Two Normal Distributions, *Biometrics*, 29:781-790 (1973).
- [23] A.J. Dobson, *An Introduction to Generalized Linear Models*, Chapman and Hall, London, 1990.
- [24] R.O. Duda, P.E. Hart, *Pattern Classification and Scene Analysis*, John Wiley and Sons, New York, 1973.
- [25] I. Eisenberger, Genesis of Bimodal Distributions, *Technometrics*, 6:357-363 (1964).
- [26] B.S. Everitt, D.J. Hand, *Finite Mixture Distributions*, Chapman and Hall, London, 1981.
- [27] B.S. Everitt, Mixture Distributions in *Encyclopedia of Statistical Sciences*, vol. 5, (Eds. S. Kotz and N.L. Johnson), New York, Wiley, 1985, 559-569.
- [28] R.A. Fisher, *Messenger of Mathematics*, 41:155-160 (1912).
- [29] E.B. Fowlkes, Some Methods for Studying the Mixture of Two Normal (Lognormal) Distributions, *J. Amer. Statist. Assoc.*, 74:561-575 (1979).
- [30] S.G. From, K.M. Lal Sexena, Estimating Parameters From Mixed Samples Using Sample Fractional Moments, *J. Statist. Plan. and Infer.*, 21:231-244 (1989).

-
- [31] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, New York, 1972.
 - [32] W. Gilchrist, *Statistical Modeling*, John Wiley and Sons, Chichester, 1984.
 - [33] J.P. Harding, The Use of Probability Paper for the Graphical Analysis of Polymodal Frequency Distributions, *J. Marine Biol. Assoc.*, 28:141-153 (1949).
 - [34] V. Hasselblad, Estimation of Parameters for a Mixture of Normal Distributions, *Technometrics*, 8:431-444 (1966).
 - [35] V. Hasselblad, Estimation of Finite Mixtures of Distributions from the Exponential Family, *J. Amer. Statist. Assoc.*, 64:1459-1471 (1969).
 - [36] R.H. Hawkins, A Note on Multiple Solutions to the Mixed Distribution Problem, *Technometrics*, 14:973-976 (1972).
 - [37] W.L. Hays, R.L. Winkler, *Statistics: Probability, Inference, and Decision*, vol.1, Holt, Rinehart, and Winston, New York, 1970.
 - [38] D.W. Hosmer, A Comparison of Iterative Maximum Likelihood Estimates of the Parameters of a Mixture of Two Normal Distributions Under Three Different Types of Sample, *Biometrics*, 29:761-770 (1973).
 - [39] D.W. Hosmer, Maximum Likelihood Estimates of the Parameters of a Mixture of Two Regression Lines, *Communications in Statistics*, 3:995-1006 (1974).
 - [40] S. John, On Identifying the Population of Origin of Each Observation in a Mixture of Observations from Two Normal Populations, *Technometrics*, 12:553-563 (1970)

-
- [41] S. John, On Identifying the Population of Origin of Each Observation in a Mixture of Observations from Two Gamma Populations, *Technometrics*, 12:565-568 (1970).
- [42] A.B.M.L. Kabir, Estimation of Parameters of a Finite Mixture of Distributions, *J.R. Statist. Soc. B*, 30:472-482 (1968).
- [43] O. Kempthorne, L. Folks, *Probability, Statistics, and Data Analysis*, Iowa State University Pres, Ames, 1971.
- [44] M.G. Kendall, A. Stuart, *The Advanced Theory of Statistics*, Vols.1 and 2, Charles Griffin and Company Ltd, London, 1967.
- [45] K.D. Kumar, E.H. Nicklin, A.S. Paulson, Comment on a Paper by Quandt and Ramsey, *J. Amer. Statist. Assoc.*, 74:52-55 (1979).
- [46] A. Leon-Garcia, *Probability and Random Processes*, Addison-Wesley, Reading, MA, 1989.
- [47] S. Lipschutz, *Theory and Problems of Linear Algebra*, McGraw-HillBook Company, New York, 1968.
- [48] P.D.M. Macdonald, Estimation of Finite Distribution Mixtures in *Applied Statistics*, (Ed. R.P. Gupta), North-Holland, Amsterdam, 1975, 231-245.
- [49] G.J. McLachlan, K.E. Basford, *Mixture Models: Inference and Applications to Clustering*, Marcel Dekker, New York, 1988.
- [50] W. Mendenhall, R.J. Beaver, *Introduction to Probability and Statistics*, PWS-KENT Publishing Company, Boston, 1991.

-
- [51] H. Muench, Discrete Frequency Distributions Arising from Mixtures of Several Single Probability Values, *J. Amer. Statist. Assoc.*, 33:390-398 (1938).
 - [52] P.L. Odell, J.P. Basu, Concerning Several Methods for Estimating Crop Acreages Using Remote Sensing Data, *Comm. Statist. Theor. Meth.*, A5:1091-1114 (1976).
 - [53] K. Pearson, Contribution to the Mathematical Theory of Evolution, *Phil. Trans. Roy. Soc. A*, 185:71-110 (1894).
 - [54] R.P. Pennington, *Introductory Computer Methods and Numerical Analysis*, The Macmillan Company, New York, 1968.
 - [55] B.C. Peters, Jr., W.A. Coberly, The Numerical Evaluation of the Maximum-Likelihood Estimate of Mixture Proportions, *Comm. Statist. Theor. Meth.*, A5:1127-1135 (1976).
 - [56] B.C. Peters, Jr., H.F. Walker, An Iterative Procedure for Obtaining Maximum Likelihood Estimates of the Parameters for a Mixture of Normal Distributions, *SIAM J. Appl. Math.*, 35:362-378 (1978).
 - [57] W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling, *Numerical Recipes: The Art of Scientific Computing*, New York, 1986.
 - [58] R.E. Quandt, J.B. Ramsey, Estimating Mixtures of Normal Distributions and Switching Regressions, *J. Amer. Statist. Assoc.*, 73:730-738 (1978).
 - [59] C.R. Rao, The Utilization of Multiple Measurements in Problems of Biological Classification, *J.R. Statist. Soc. B*, 10:159-203 (1948).
 - [60] R.A. Redner, H.F. Walker, Mixture Densities, Maximum Likelihood, and the EM Algorithm, *SIAM Rev.*, 26:195-239 (1984).

-
- [61] P.R. Rider, The Method of Moments Applied to a Mixture of Two Exponential Distributions, *Ann. Math. Statist.*, 32:143-47 (1961).
- [62] P.R. Rider, Estimating the Parameters of Mixed Poisson, Binomial, and Weibull Distributions by the Method of Moments, *Bull. Inst. Internat. Stat. (Part 2)*, 33:225-232 (1961).
- [63] W. Schilling, A Frequency Distribution Represented as the Sum of Two Poisson Distributions, *J. Amer. Statist. Assoc.*, 42:407-424 (1947).
- [64] S.S. Shapiro, A.J. Gross, *Statistical Modeling Techniques*, Marcel Dekker, New York, 1981.
- [65] A.F.M. Smith, U.E. Makov, A Quasi-Bayes Sequential Procedure for Mixtures, *J. Royal Statist. Soc. B*, 40:106-112 (1978).
- [66] H.W. Sorenson, *Parameter Estimation: Principles and Problems*, Marcel Dekker, New York, 1980.
- [67] G.M. Tallis, R. Light, The Use of Fractional Moments for Estimating the Parameters of a Mixed Exponential Distribution, *Technometrics*, 10:161-175 (1968).
- [68] W.Y. Tan, W.C. Chang, 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).
- [69] H. Teicher, Identifiability of Mixtures, *Ann. Math. Statist.*, 32:244-248 (1961).
- [70] G.L. Tietjen, *A Topical Dictionary of Statistics*, Chapman and Hall, New York, 1986.

-
- [71] D.M. Titterton, A.F.M. Smith, U.E. Makov, *Statistical Analysis of Finite Mixture Distributions*, John Wiley and Sons, Chichester, 1985.
 - [72] R.S. Valiveti, *Learning Algorithms for Data Retrieval*, Ph.D. Thesis, School of Computer Science, Carleton University, Ottawa, 1991.
 - [73] J.H. Wolfe, Pattern Clustering by Multivariate Mixture Analysis, *Multivariate Behavioral Res.*, 5:329-350 (1970).
 - [74] S.J. Yakowitz, J.D. Spragins, On the Identifiability of Finite Mixtures, *Ann. Math. Statist.*, 39:209-214 (1968).
 - [75] S. Zacks, *Parametric Statistical Inference: Basic Theory and Modern Approaches*, Pergamon Press, Oxford, 1981.

END

16-12-93

FIN