FINITE MIXTURES OF DISTRIBUTIONS; THE PROBLEM OF ESTIMATING THE MIXING PROPORTIONS

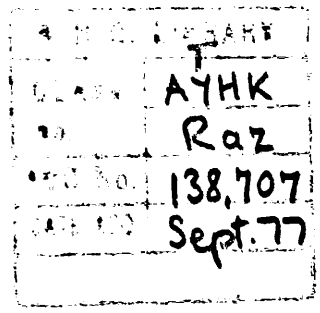
Mir-Mehdi Razzaghi-Kashani

A thesis submitted for the degree of

Doctor of Philosophy

in the

University of London



Department of Statistics and Computer Science

Royal Holloway College

London, February 1977

To Mehran

and my Parents

ABSTRACT

Constructing estimators for the parameters of a mixture of distributions has attracted many statisticians. Given that the distribution function $G_{\underset{\sim}{\theta}}(.)$ of a random variable X is a mixture of k ($1 < k < \infty$) known distribution functions $F_1(.),\ldots,F_k(.)$ with mixing proportions $\theta_1,\ldots,\theta_k$ respectively where $0 \leq \theta_j \leq 1$ for $j = 1,\ldots,k$ and $\sum_{j=1}^{k} \theta_j = 1$, and given that $G_{\underset{\sim}{\theta}}(.)$ determines $\theta_1,\ldots,\theta_k$ uniquely, estimation of the mixing proportions is considered. Different estimation techniques are studied in depth and the properties of the resulting estimators are discussed.

The necessary background to mixtures of distributions is first given and an extension of the method of moments for estimating $\theta_1,\ldots,\theta_k$ is then proposed. The generalized (weighted) least squares method, when the observations are grouped into (m+1) intervals, is considered and it is shown that the estimators possess certain desired asymptotic properties. The case when $m \to \infty$ is also investigated. Since the set of equations leading to the generalized least squares estimators are not in general solvable, an iteration process is proposed and is shown to produce satisfactory results after even one cycle. Finally, when $k = 2$, $\theta_1 = \theta$ and $\theta_2 = 1-\theta$, the problem of maximum likelihood estimation of θ is considered and the Fisher's scoring method is suggested to solve the likelihood equation. Properties of the first and second cycle solutions are derived.

ACKNOWLEDGEMENTS

I wish to express my gratitude and appreciation to my supervisor Dr. D. Mannion for his continued guidance and many invaluable comments in this research. I am grateful to Professor H. J. Godwin, the Head of the Department of Statistics and Computer Science, for his help and guidance throughout my postgraduate studies at Royal Holloway College. A debt of thanks is also owed to Professor J. F. Scott of the University of Sussex for initiating my interests in statistics whilst I was an undergraduate at that University. The financial support of Royal Holloway College and the Free University of Iran is gratefully acknowledged.

TABLE OF CONTENTS

Table of Contents (cont'd.)

CHAPTER 1

INTRODUCTION

1.1 General

In recent years, mixtures of distributions have received an increasing amount of attention in statistical literature, partly because of interest in their mathematical aspects and partly because of a considerable number of applied problems in which mixtures of distributions are encountered. In this thesis we consider a mixture of k ($1<k<\infty$) distinct distribution functions $F_1(.),\ldots,F_k(.)$ defined as

$$G_{\underset{\sim}{\theta}}(x) = \sum_{j=1}^{k} \theta_j F_j(x)$$

for every x belonging to some measurable subset of the real line with the condition that $0 \leq \theta_j \leq 1$ for $j = 1,\ldots,k$ and $\sum_{j=1}^{k} \theta_j = 1$. We assume that the distribution functions $F_1(.),\ldots,F_k(.)$ involve no unknown parameters and further that $G_{\underset{\sim}{\theta}}(x)$ determines uniquely $\theta_1,\ldots,\theta_k$ and $F_1(.),\ldots,F_k(.)$, i.e. $G_{\underset{\sim}{\theta}}(x)$ is "identifiable". We pose the problem of estimating the unknown parameters $\theta_1,\ldots,\theta_k$.

In this chapter, however, we give the preliminaries and the background to mixtures of distributions. In Section 1.2, we state our definitions and notations along with some of the elementary properties of mixtures of distributions. In 1.3 we give a formal statement of our problem together with a brief summary of the subsequent chapters of the thesis. In Sections 1.4, 1.5 and 1.6, we outline some of the problems arising in mixtures of distributions together with a summary of the work of previous authors. In 1.7 we look at some applications.

1.2 Definitions and Terminologies

Let $\mathcal{F} = \{F(x;\alpha);\ x \in \mathcal{X},\ \alpha \in \mathcal{A}\}$ be a family of one-dimensional cumulative distribution functions $F(x;\alpha)$ in the variable $x \in \mathcal{X}$ where $\mathcal{X}$ is a measurable subset of the real line to which every member of $\mathcal{F}$ assigns probability one. Each member of $\mathcal{F}$ is indexed by a finite dimensional parameter $\alpha = (\alpha^{(1)},\ldots,\alpha^{(s)})$, belonging to some measurable subset $\mathcal{A}$ of B^s, the σ-field of the Borel sets in R^s. Suppose that for each $\alpha \in \mathcal{A}$, the set of points to which $F(.;\alpha)$ assigns positive probability is independent of $\alpha^{(1)},\ldots,\alpha^{(s)}$ and that $F(x;\alpha)$ is measurable on the product space $\mathcal{X} \times \mathcal{A}$, a measurable subset of the $(s+1)$-dimensional Euclidean space R^{s+1}. For this, it suffices to stipulate that $F(x;\alpha)$ be measurable in α for all $x \in \mathcal{X}$ (Teicher [55]).

Denote by $\mathcal{Q}$ the class of non-degenerate s-dimensional cumulative distribution functions $Q(\alpha)$ whose induced Lebesgue-Stieltjes measure μ_Q assigns measure one to $\mathcal{A}$. Then

$$G(x) = G_Q(x) = \int_{\alpha \in \mathcal{A}} F(x;\alpha)\, dQ(\alpha) \qquad x \in \mathcal{X} \qquad (1.2.1)$$

is a one-dimensional cumulative distribution function (Robbins [47]) called a "Q-mixture" or more briefly a "mixture" of $\mathcal{F}$. The family $\mathcal{F}$ is called the "kernel" of the mixture while $Q(.)$ is referred to as the "mixing distribution". Following Teicher [54], the family $\mathcal{G} = \mathcal{G}(\mathcal{F})$ of mixtures $G(.)$ of $\mathcal{F}$ resulting as $Q(.)$ ranges over $\mathcal{Q}$ is called the class of mixtures of $\mathcal{F}$.

Now, in particular, if each $Q \in \mathcal{Q}$ is a step function with steps at $\alpha_1,\alpha_2,\ldots$ say, or equivalently if for each $Q \in \mathcal{Q}$, μ_Q is discrete assigning positive measure only to $\alpha_1,\alpha_2,\ldots$ (a countable number of points in R^s), then (1.2.1) reduces to

$$G_{\underset{\sim}{\theta}}(x) = \sum_{j=1}^{\infty} \theta_j F(x;\alpha_j) \qquad x \in \mathcal{X} \qquad (1.2.2)$$

where θ_j is the mass assigned by $Q(.)$ to α_j for $j = 1,2,\ldots$ called the "mixing proportions" and $\underset{\sim}{\theta} = (\theta_1,\theta_2,\ldots)'$. It is clear that $0 \leq \theta_j \leq 1$ for $j = 1,2,\ldots$ and $\sum_{j=1}^{\infty} \theta_j = 1$. Distributions of the type (1.2.2) are called "countable" mixtures of distributions. Moreover if for each $Q \in \mathcal{Q}$, the set $\{\alpha_1,\alpha_2,\ldots\} \subset R^s$ contains only a finite number of elements $\alpha_1,\ldots,\alpha_k$ then the resulting mixture of distributions is

$$G_{\underset{\sim}{\theta}}(x) = \sum_{j=1}^{k} \theta_j F(x;\alpha_j) \qquad x \in \mathcal{X} \qquad (1.2.3)$$

where $0 \leq \theta_j \leq 1$ for $j = 1,\ldots,k$, $\sum_{j=1}^{k} \theta_j = 1$ and $\underset{\sim}{\theta} = (\theta_1,\ldots,\theta_k)' = \sum_{j=1}^{k} \theta_j \underset{\sim}{e}_j$ with $\underset{\sim}{e}_j$, $1 \leq j \leq k$ being the k-dimensional vector with 1 at the jth position and zero elsewhere.

Distributions of the type (1.2.3) are called "finite" mixtures of distributions. The individual distribution functions $F(.;\alpha_j)$; $j = 1,\ldots,k$ being mixed to produce a particular $G_{\underset{\sim}{\theta}}(.)$ will be called the "components" of $G_{\underset{\sim}{\theta}}(.)$. Finally, if in (1.2.3), the values of $\alpha_1,\ldots,\alpha_k$ are known, (1.2.3) takes the form

$$G_{\underset{\sim}{\theta}}(x) = \sum_{j=1}^{k} \theta_j F_j(x) \qquad x \in \mathcal{X} \qquad (1.2.4)$$

where $F_j(.) = F(.;\alpha_j)$ for $j = 1,\ldots,k$.

The following two special cases are noted:

(i) Let $s=1$, $F(x;\alpha) = F(x-\alpha)$ in (1.2.1), then

$$G_Q(x) = \int_{\alpha \in \mathcal{A}} F(x-\alpha)\, d\, Q(\alpha) \qquad x \in \mathcal{X}$$

and it is well-known that in this case $G_Q(.)$ is called the convolution

of F(.) and Q(.) written as $G_Q(x) = F*Q(x)$ for $x \in \mathfrak{X}$. If X and Y are two independent random variables with respective distribution functions F(.) and Q(.) in R, then (Robbins [47]) the distribution $G_Q(.)$ of Z = X+Y is

$$G_Q(x) = \text{Prob}[X+Y \leq x] = F*Q(x) \qquad x \in \mathfrak{X} \quad . \quad (1.2.5)$$

However (1.2.5) is only necessary and not sufficient condition for independence of X and Y.

Further, if we denote by $\phi_1(t)$, $\phi_2(t)$ and $\phi(t)$ the characteristic functions corresponding to the distribution functions F(x), Q(x) and $G_Q(x)$ respectively then $G_Q(x) = F*Q(x)$ if and only if

$$\phi(t) = \phi_1(t) \cdot \phi_2(t) \ .$$

(ii) If in (1.2.1), $F(x;\alpha)$ is defined for non-negative integers $\alpha = 0,1,2,\ldots$ and it is the α-fold convolution of a given distribution function F(x) with itself, i.e. $F(x;\alpha) = F^{*\alpha}(x)$ and $Q(\alpha)$ is the univariate Poisson distribution with mean λ, then the resulting mixture of distributions

$$\sum_{\alpha=0}^{\infty} \frac{e^{-\lambda}\lambda^{\alpha}}{\alpha!} F^{*\alpha}(x) \qquad x \in \mathfrak{X}$$

is called a generalized Poisson distribution.

1.3 Statement of the Problem and Outline of the Thesis

In this thesis we shall be concerned with finite mixtures of distributions of the type (1.2.4), i.e. a mixture of the one-dimensional distribution functions $F_1(.),\ldots,F_k(.)$. It will be assumed that these distribution functions are all continuous to the right so that

$$F_j(x) = F_j(x+0) \qquad x \in \mathcal{X}$$

for $j = 1,\ldots,k$ and therefore $G_{\underset{\sim}{\theta}}(x)$ given by (1.2.4) is also continuous to the right. Further, we will assume that the component distribution functions $F_1(.),\ldots,F_k(.)$ of $G_{\underset{\sim}{\theta}}(.)$ are completely known and involve no unknown parameters, but no knowledge about the mixing proportions $\theta_1,\ldots,\theta_k$ is available except that they are k non-negative parameters adding up to unity. The number of components k giving rise to the finite mixture of distributions will always be assumed to be known.

The problem that we deal with in this thesis is the problem of estimating the mixing proportions $\theta_1,\ldots,\theta_k$ on the basis of n observations from a finite mixture of distributions. However, before the problem of estimation can meaningfully be considered, the identifiability that is the question of unique characterization of the mixture of distribution has to be established. Identifiability of mixtures of distributions will be dealt with in Section 1.4 and necessary and sufficient conditions will be given for the identifiability of finite mixtures of distributions due to Teicher [56] and Yakowitz and Spragins [60]. We assume throughout this thesis that the mixture of distributions, whose mixing proportions are to be estimated, is known *a priori* to be identifiable. We state the problem formally as follows:

> "Given a set of n independently and identically distributed random variables $X_1,\ldots,X_n$ with a common distribution function $G_{\underset{\sim}{\theta}}(.)$ given by (1.2.4), and with observed values $x_1,\ldots,x_n$ and given that $G_{\underset{\sim}{\theta}}(.)$ is identifiable, it is required to estimate the vector of the unknown mixing proportions $\underset{\sim}{\theta} = (\theta_1,\ldots,\theta_k)'$."

Our estimate of $\underset{\sim}{\theta}$ will be based on the empirical distribution function $G_n(.)$ defined as

$$G_n(x) = \frac{1}{n}[\text{no. of } x_1,\ldots,x_n \leq x] \qquad x \in \mathcal{X}, \quad (1.3.1)$$

i.e. $G_n(x)$ is the proportion of the observations which do not exceed x. We denote by $\Gamma_n(x)$ a random function whose realization is $G_n(x)$ for all $x \in \mathcal{X}$, so that

$$\Gamma_n(x) = \frac{1}{n}[\text{no. of } X_1,\ldots,X_n \leq x] \qquad x \in \mathcal{X}. \quad (1.3.2)$$

The functions $G_n(.)$ and $\Gamma_n(.)$ can also be represented in the following forms: Let $\eta(x)$ be the well-known Heaviside function defined as

$$\eta(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}$$

then

$$G_n(x) = \frac{1}{n}\sum_{j=1}^{n} \eta(x - x_j) \qquad x \in \mathcal{X} \quad (1.3.3)$$

and similarly

$$\Gamma_n(x) = \frac{1}{n}\sum_{j=1}^{n} \eta(x - X_j) \qquad x \in \mathcal{X} \quad (1.3.4)$$

The statistical properties of $\Gamma_n(x)$ are well-known (Darling [16]) and we state here (without proof) some of its more important properties.

(i) The expected value of $\Gamma_n(x)$ is $G_\theta(x)$ for every $x \in \mathcal{X}$ and the covariance of $\Gamma_n(x)$ and $\Gamma_n(y)$ is $\frac{1}{n}\, c\,(G_\theta(x), G_\theta(y))$ for every $x, y \in \mathcal{X}$, where

$$c(s,t) = \min(s,t) - st = \begin{cases} s(1-t) & s \leq t \\ t(1-s) & s \geq t \end{cases} \quad (1.3.5)$$

$$\text{for } 0 \leq s,t \leq 1$$

(ii) By the strong law of the large numbers

$$\Gamma_n(x) \to G_{\underset{\sim}{\theta}}(x)$$

with probability 1 as $n \to \infty$ for each $x \in \mathfrak{X}$

(iii) By the law of iterated logarithm,

$$\limsup_{n\to\infty} \sqrt{n}\, \frac{|\Gamma_n(x) - G_{\underset{\sim}{\theta}}(x)|}{\sqrt{2 \log \log n}} = \sqrt{G_{\underset{\sim}{\theta}}(x)\,(1 - G_{\underset{\sim}{\theta}}(x))}$$

with probability 1 for each $x \in \mathfrak{X}$.

(iv) By the multidimensional central limit theorem, for any set of values $\{t_i\}_{i=1}^{m}$ such that $t_i \in \mathfrak{X}$ for $i = 1,\ldots,m$, the random variables

$$\sqrt{n}\,(\Gamma_n(t_i) - G_{\underset{\sim}{\theta}}(t_i)) \qquad i = 1,\ldots,m$$

have a joint asymptotic m-dimensional normal distribution with mean vector $\underset{\sim}{0}_{m\times 1} = (0,0,\ldots,0)'$ and an $m\times m$ covariance matrix having $c(G_{\underset{\sim}{\theta}}(t_i), G_{\underset{\sim}{\theta}}(t_j))$ with c given by (1.3.5) as its (i,j)th element for $i,j = 1,\ldots,m$.

(v) By Glivenko-Cantelli lemma

$$\sup_{x\in\mathfrak{X}} |\Gamma_n(x) - G_{\underset{\sim}{\theta}}(x)| \to 0$$

with probability 1 as $n \to \infty$.

In Chapters 2 to 5, estimators of $\underset{\sim}{\theta}$ will be derived and their properties will be analysed. The results of some numerical studies will also be used to provide further illustrations. In Chapter 2, we consider estimating $\underset{\sim}{\theta}$ by using the method of moments. This method which has attracted many statisticians dealing with mixtures of distributions (c.f. Section 1.5), consists of equating as many sample

moments to their corresponding expected values as there are unknown parameters. We consider somewhat a generalization of this method. Assuming that the distribution of the random variable X is given by (1.2.4), the conventional method of moments is based upon solving the set of equations resulted from equating $\frac{1}{n}\sum_{i=1}^{n} x_i^t$ for t = 1,...,k to their corresponding expected values, i.e. the expectation of X^t for t = 1,...,k. Instead of using the function X^t, we define a real-valued function h(X,t) of X and $t \in R$ with the property that h(x,t) be a right-continuous function of $t \in R$ for each $x \in \mathfrak{X}$. Also h(x,t) has to satisfy some further mild restrictions (see Lemma 2.2.1). Instead of choosing t = 1,...,k, as in the method of moments, we choose a finite set of real values $t_1,\ldots,t_m$ where $m \geq k$. Since m may be greater than k, the set of equations resulted from equating the sample moments of $h(X,t_r)$; r = 1,...,m to their corresponding expected values, will have no solution and so we use the method of least squares to find an estimate of $\underset{\sim}{\theta}$. The properties of our estimate will be investigated and it will be discussed that using a generalized least squares, that is taking the covariances between $h(X,t_r)$ and $h(X,t_s)$ for r,s = 1,...,m into consideration, will improve our estimate. This will, however, cause some difficulties unless the exact form of h(x,t) for $x \in \mathfrak{X}$ and $t \in R$ is known.

In Chapter 3 we consider a special form of h(x,t) namely

$$h(x,t) = 1 \qquad t \in \mathfrak{X}, \quad x \leq t$$
$$= 0 \qquad \text{otherwise}$$

and will see that for this special case we are led to fitting the values $G_n(t_1),\ldots,G_n(t_m)$ to $G_{\underset{\sim}{\theta}}(t_1),\ldots,G_{\underset{\sim}{\theta}}(t_m)$ respectively by the method of generalized least squares. The underlying equations whose root forms the generalized least squares estimator of $\underset{\sim}{\theta}$

give rise to equations whose solution constitutes the minimum χ^2 estimate of $\underset{\sim}{\theta}$ and we can thus establish the properties of our estimator. Unfortunately the resulting equations are very difficult to solve in general, and we consider a special case namely that the mixture of distributions $G_{\underset{\sim}{\theta}}(x)$ given by (1.2.4) consists only of two components and each component is the distribution function of a uniformly distributed random variable. We study the properties of the estimate of the mixing proportion in more detail. We finally let m, the number of chosen values of t's, become very large and consider the case when $m \to \infty$ and establish the properties of the generalized least squares estimator of $\underset{\sim}{\theta}$ in this case.

In view of the fact that the set of equations having the generalized least squares estimator of $\underset{\sim}{\theta}$ as their solution are very difficult to solve, even in simple situations, we propose an iteration procedure in Chapter 4. Starting with a consistent but inefficient estimator of $\underset{\sim}{\theta}$, the iteration process after one iteration yields an estimator which is consistent asymptotically normally distributed and asymptotically fully efficient with respect to a given set of values $t_1,\ldots,t_m$. Analogous to Chapter 3, we consider the situation when $m \to \infty$ and will see that the iteration process converges to the maximum likelihood estimator of $\underset{\sim}{\theta}$.

Finally in Chapter 5, we deal with the problem of maximum likelihood estimation of $\underset{\sim}{\theta}$ in the special situation when $G_{\underset{\sim}{\theta}}(x)$ given by (1.2.4) consists only of two components $F_1(x)$ and $F_2(x)$ with respective mixing proportions $\theta_1 = \theta$ and $\theta_2 = 1-\theta$. Thus we have

$$G_{\underset{\sim}{\theta}}(x) = \theta\, F_1(x) + (1-\theta)\, F_2(x) \qquad x \in \mathfrak{X}$$

for $0 \leq \theta \leq 1$. It is seen that the identifiability (uniqueness) of

$G_{\underset{\sim}{\theta}}(x)$ is evident in this case for if

$$G_{\underset{\sim}{\theta^*}}(x) = \theta^* F_1(x) + (1-\theta^*) F_2(x) \qquad x \in \mathcal{X}$$

for $0 \leq \theta^* \leq 1$, then $G_{\underset{\sim}{\theta}}(x) \equiv G_{\underset{\sim}{\theta^*}}(x)$, i.e.

$$\theta(F_1(x)-F_2(x)) + F_2(x) \equiv \theta^* (F_1(x)-F_2(x)) + F_2(x)$$

for all $x \in \mathcal{X}$ if and only if $\theta = \theta^*$ unless $F_1(.) \equiv F_2(.)$. We discuss the properties of the equation whose root constitutes the maximum likelihood estimator of θ (likelihood equation) and show that certain regularity conditions are satisfied for $0<\theta<1$ so that the well-known properties of maximum likelihood estimators are applicable to our estimate in such situations. We also discuss the properties of the Fisher's information function $I(\theta)$ and give sufficient conditions for the existence of a unique root of the likelihood equation in (0,1). It turns out that the root of the likelihood equation cannot, in general, be obtained directly. We propose the use of an iteration process commonly known as Fisher's scoring method (Rao [45]). We discover the properties of the solutions given by the first and second cycles of the iteration process when the process is started with an arbitrary value chosen in (0,1) (independent of the observations). We see that the properties of the solution provided by two cycles of the iteration process are also applicable to the solutions obtained by the subsequent cycles of the process.

Concluding our investigations in Chapter 6, we suggest a few areas, related to mixtures of distributions, in which further research could be carried out.

1.4 Identifiability of Mixtures of Distributions

The question of identifiability of mixtures of distributions concerns their unique characterization. Teicher [54] was the first to use this term and it has since been used by others.

Definition 1.4.1 (Teicher [55]) : The mixture $G_Q(x)$ of $\mathcal{F}$ given by (1.2.1) is said to be identifiable in $\mathcal{Q}$ if the relationship

$$G_Q(x) \equiv G_{Q^*}(x)$$

i.e.

$$\int_{\alpha \in \mathcal{A}} F(x;\alpha)\, d\, Q(\alpha) \equiv \int_{\alpha \in \mathcal{A}} F(x;\alpha)\, d\, Q^*(\alpha)$$

holds for all $x \in \mathcal{X}$ if and only if $Q(.) \equiv Q^*(.)$ for all Q and Q^* belonging to $\mathcal{Q} \cup \mathcal{G}$ where $\mathcal{G}$ denotes the family of degenerate distribution functions, i.e. $\mathcal{G}$ is the family of distribution functions whose corresponding Lebesgue-Stieltjes measures assign measure one to a single point in R^s. If every $F(x,\alpha) \in \mathcal{F}$; $x \in \mathcal{X}$, $\alpha \in \mathcal{A}$ induces an identifiable mixture of distributions in $\mathcal{Q}$, then the corresponding class of mixtures of distributions $\mathcal{Y}$ is called identifiable in $\mathcal{Q}$ (with respect to $\mathcal{F}$).

Note that the identifiability of countable and finite mixtures of distributions can be defined in a similar way. In particular a countable mixture of distributions given by (1.2.2) is said to be identifiable if the relationship

$$G_{\underset{\sim}{\theta}}(x) \equiv G_{\underset{\sim}{\theta}^*}(x)$$

i.e.

$$\sum_{i=1}^{\infty} \theta_i\, F(x;\alpha_i) \equiv \sum_{j=1}^{\infty} \theta_j^*\, F(x;\alpha_j^*)$$

holds for all $x \in \mathcal{X}$ if and only if for each positive integer i, there is another positive integer j such that $\theta_i = \theta_j^*$ and $\alpha_i = \alpha_j^*$

and similarly (1.2.3) is identifiable if the relation

$$\sum_{i=1}^{k} \theta_i F(x;\alpha_i) \equiv \sum_{j=1}^{k^*} \theta_j^* F(x;\alpha_j^*)$$

holds for all $x \in \mathfrak{X}$ if and only if $k = k^*$ and for each $1 \leq i \leq k$ there is some $1 \leq j \leq k$ such that $\theta_i = \theta_j^*$ and $\alpha_i = \alpha_j^*$.

The lack of identifiability of a mixture of distributions is not uncommon. Consider as an example the family of binomial distributions with density function

$$f(x;\alpha) = \binom{n}{x} \alpha^x (1-\alpha)^{n-x} \qquad x = 0,1,\ldots,n$$

where $0 \leq \alpha \leq 1$ is unknown and n is a fixed positive integer. Then

$$g_Q(x) = \int_0^1 f(x;\alpha)\, dQ(\alpha)$$

is a linear function of the first n moments of $Q(\alpha)$ given by

$$\mu_Q^{(r)} = \int_0^1 \alpha^r\, dQ(\alpha)$$

for $r = 1,\ldots,n$. Consequently, a necessary and sufficient condition for any other $g_{Q^*}(x)$ with mixing distribution $Q^*(\alpha)$ be identical to $g_Q(x)$ for $x = 0,1,\ldots,n$ is that the first n moments of $Q^*(\alpha)$ given by

$$\mu_{Q^*}^{(r)} = \int_0^1 \alpha^r\, dQ^*(\alpha)$$

for $r = 1,\ldots,n$, be identical to $\mu_Q^{(r)}$ for $r = 1,\ldots,n$.

The most thorough investigation of the problem of identifiability of a mixture of distributions has been undertaken by Teicher [54, 55, 56]. Some conditions of identifiability are given in [54, 55] as

well as a discussion of the question of identifiability for several specific classes of mixtures of distributions. In [56] he considers the problem of identifiability of finite mixtures of distributions and shows that a necessary and sufficient condition for the class

$$\mathcal{G} = \{ \sum_{j=1}^{k} \theta_j F_j(x) \; ; \; x \in \mathcal{X}, \; 0 \leq \theta_j \leq 1 \text{ for } j = 1,\ldots,k, \; \sum_{j=1}^{k} \theta_j = 1 \} \tag{1.4.1}$$

of all finite mixtures of the finite family of distribution functions $\mathcal{F} = \{F_1(x),\ldots,F_k(x); \; x \in \mathcal{X}\}$ be identifiable is that there exists k real values $x_1,\ldots,x_k$ with $x_j \in \mathcal{X}$ for $j = 1,\ldots,k$ for which the determinant of the $k \times k$ matrix with $F_i(x_j)$ as its (i,j)th element for $i,j = 1,\ldots,k$, is non-zero. Using this result, the author establishes the identifiability of the class of finite mixtures of normal distributions and finite mixtures of gamma distributions.

Yakowitz and Spragins [60] have shown that a finite mixture of distributions is identifiable if and only if the components are linearly independent cumulative distribution functions, i.e. the class (1.4.1) is identifiable if and only if

$$\sum_{j=1}^{k} c_j F_j(x) \equiv 0 \quad \text{for } c_j \in R \text{ and } x \in \mathcal{X} \Leftrightarrow c_1 = c_2 = \ldots = c_k = 0 \; .$$

In fact they proved a rather more general result than stated, by considering $\mathcal{X}$ to be a measurable subset of R^m, i.e. each component of the finite mixture of distributions is the distribution function of an m-dimensional random variable. From their important result, they obtained the identifiability of finite mixtures of distributions with each component being

(i) an m-dimensional normal distribution,

(ii) the product of m negative exponential distributions,

(iii) one dimensional Cauchy distribution,

(iv) the negative binomial distribution,

(v) either an m-dimensional normal distribution or the product of m negative exponential distributions.

Notable contributions have been made by other authors such as Blischke [8] who gives a necessary and sufficient condition for the identifiability of mixtures of binomial distributions.

1.5 Estimation for Mixtures of Distributions

When the identifiability of a family of mixtures of distributions has been established, one can discuss the problem of estimating the unknowns. In the mixture of distributions $G_Q(.)$ defined by (1.2.1), $Q(\alpha)$; $\alpha \in \mathcal{A}$ (and thus $G_Q(.)$) is not in general known exactly, although the form of $F(.;\alpha)$; $\alpha \in \mathcal{A}$ will usually be assumed known. Direct information on $G_Q(x)$ is supplied only by n observations $x_1,\ldots,x_n$ being the realizations of $X_1,\ldots,X_n$ respectively of the random variable X whose distribution is $G_Q(.)$. The observations are then used to construct an empirical distribution function say $G_n(.)$ being an estimate of $G_Q(.)$.

The problem of exact estimation of the mixing distribution $Q(.)$ when Q could be any continuous distribution is of course an impractical task and this problem is closely related to the empirical Bayes Procedures, proposed first by Robbins [48], where the mixing distribution $Q(.)$ corresponds to the *a priori* distribution. Robbins [48] suggests that if the *a priori* distribution function is known to the experimenter, he can perform a Bayesian analysis of his experiment, but if such information is not available then the *a priori* distribution function has to be estimated. This is equivalent to estimating the mixing distribution $Q(.)$ in (1.2.1). As an exact estimate of Q cannot be obtained, one constructs a sequence of random step functions $Q_n(.) = Q_n(X_1,\ldots,X_n ; .)$ and requires $Q_n(.)$ to converge weakly to $Q(.)$

with probability 1, i.e. that Prob[$\lim_{n\to\infty} Q_n(\alpha) = Q(\alpha)$ at every continuity point $\alpha \in \mathcal{A}$ of Q] = 1.

A common method of constructing a sequence of estimators for Q(.) is to determine $Q_n(.)$ such that a suitable measure of distance between $G_Q(.)$ and the empirical distribution function $G_n(.)$ is minimized. The motivation of such an approach is found in Deely and Kruse [18] who suggest the use of Kolmogorov-Smirnov distance defined as

$$\| G_Q - G_n \| = \sup_{x\in\mathcal{X}} | G_Q(x) - G_n(x)| \ . \qquad (1.5.1)$$

Apparently, the amount of publication on the general problem of estimating the mixing distribution is very few and in view of the fact that some important families of distributions, while not generating identifiable arbitrary mixtures of distributions (of the type (1.2.1)), generate identifiable finite mixtures of distributions (e.g. the family of normal distributions with mean and variance both considered as parameters), most of the publications on problems of estimation in mixtures of distributions are concerned with finite mixtures of distributions. The estimation problems in this case arise, for example, in the situation in which a finite set of experiments $\{E_1,\ldots,E_k\}$ gives rise to a sequence of random variables $\{X_i\}_{i=1}^n$ as follows: At each observation time $1 \le j \le k$ with probability θ_j, at the exclusion of the other experiments, experiment E_j is selected and an observation x_i, the realization of X_i with distribution function $F_j(.)$, is made. This value x_i is taken to be the observed value of the ith element of the sequence $\{X_i\}_{i=1}^n$. The statistician does not know the parameters $\theta_1,\ldots,\theta_k$. He may not know the distribution fuctions $F_1(.),\ldots,F_k(.)$, or even the value of k. He is told that the component distribution functions are distinct and are all members

of a specified family $\mathcal{F}$ as defined in Section 1.2. The problem is then to determine the unknowns solely on the available information. It is to be emphasized that the statistician knows that the mixture of distributions generated by $F_1,\ldots,F_k$ with mixing proportions $\theta_1,\ldots,\theta_k$ respectively is identifiable, but he is never told which of the experiments $E_1,\ldots,E_k$ was performed at any time. We devote the remainder of this section to an outline of some of the estimation problems in finite mixtures of distributions considered by previous authors. For this reason, the term "mixture of distributions" or simply "mixture" will refer to a finite mixture of distributions.

The estimation problems dealt with in the past, all assume that k is known (often taken equal to 2) and $F_1,\ldots,F_k$ may or may not depend on some unknown parameters. The method of minimizing some measure of distance between the true distribution $G_{\underset{\sim}{\theta}}(.)$ given by (1.2.3) and the empirical distribution function, has been considered by some authors. Choi and Bulgren [11] use the Wolfowitz's distance given by

$$W(G_{\underset{\sim}{\theta}},G_n) = \int (G_{\underset{\sim}{\theta}}(x) - G_n(x))^2 \, d\, G_n(x) \qquad (1.5.2)$$

and Bartlett and Macdonald [4] suggest the method of weighted least squares using

$$\int \frac{(dG_n(x) - dG_{\underset{\sim}{\theta}}(x))^2}{dW(x)} \qquad (1.5.3)$$

where W(.) is a suitable increasing function. Macdonald [34] uses the Cramér-Von Mises distance

$$\int (G_n(x) - G_{\underset{\sim}{\theta}}(x))^2 \, d\, G_{\underset{\sim}{\theta}}(x) \, . \qquad (1.5.4)$$

Deely and Kruse [18] use Kolmogorov-Smirnov distance given by (1.5.1) and their solution is based upon solving a two-person zero sum game after each observation. Macdonald [35] compares the method of estimation suggested in Macdonald [34] with that of Choi and Bulgren [11] with regard to some numerical studies based on a mixture of two normal distributions.

The earliest attempt to separate a mixture of distributions into its components was made by Karl Pearson [41] in 1894. Pearson attempted to estimate the means, the variances and the mixing proportions of a mixture of two normal densities

$$g_{\underset{\sim}{\theta}}(x) = \frac{1}{\sqrt{2\pi}} \left\{ \frac{\theta_1}{\sigma_1} \exp\left[-\tfrac{1}{2} \left(\frac{x-\mu_1}{\sigma_1} \right)^2 \right] + \frac{\theta_2}{\sigma_2} \exp\left[-\tfrac{1}{2} \left(\frac{x-\mu_2}{\sigma_2} \right)^2 \right] \right\} \qquad (1.5.5)$$

where $0 \leq \theta_1 \leq 1$ and $\theta_2 = 1-\theta_1$ for $-\infty < x < \infty$, by using the method of moments and equating the first five population moments to their corresponding sample values. Solving these five equations in the five unknowns μ_1, μ_2, σ_1, σ_2 and θ_1 leads to a ninth degree polynomial equation having at least one real root. Each real root of the nonic gives a set of estimates for the parameters. Pearson proposed that the set of solutions making the sixth population moment nearest to the sixth sample moment be used as the final estimate. Although the computations are not difficult on a modern computer, the results are only optimal in the sense of fitting the first six moments. Also the procedure does not generalize easily to the case of a mixture of more than two populations. Rao [45] applied the method of moments to a mixture of two normal distributions with equal variances and showed that Pearson's nonic reduces to a cubic and the computation is considerably simplified and Cohen [12] also showed how the computation of Pearson's method can be lightened to some extent.

The most exhaustive statistical approach to mixtures of binomial

distributions has been given by Blischke [8] who employed the first (2k-1) factorial moments of a mixture of k ($k \geq 2$) binomial distributions,

$$g_{\underset{\sim}{\theta}}(x) = \sum_{j=1}^{k} \theta_j \binom{n}{x} p_j^x (1 - p_j)^{n-x} \qquad (1.5.6)$$

for $x = 0,1,\ldots,n$, $0 < p_j, \theta_j < 1$; $j = 1,\ldots,k$ and $\sum_{j=1}^{k} \theta_j = 1$, to obtain estimates of $p_1,\ldots,p_k$ and $\theta_1,\ldots,\theta_k$. He then showed that the estimates have joint asymptotically normal distribution and also investigated the asymptotic efficiency relative to the Cramér-Rao lower bound. Blischke found that if the mixing proportions are unknown then the joint asymptotic relative efficiency of the estimates tends to unity as the binomial parameter $n \to \infty$. However, if the mixing proportions are known, the relative efficiency approaches zero. No intuitive explanation was offered for this apparent anomaly.

The method of moments has also been used by Falls [20] to estimate the five parameters of a mixture of two Weibull distributions,

$$g_{\underset{\sim}{\theta}}(x) = \theta_1 \gamma_1 \alpha_1^{-1} x^{\gamma_1 - 1} \exp(-x^{\gamma_1}/\alpha_1) + \theta_2 \gamma_2 \alpha_2^{-1} x^{\gamma_2 - 1} \exp(-x^{\gamma_2}/\alpha_2) \qquad (1.5.7)$$

for $x \geq 0$, $0 \leq \theta_1 \leq 1$, $\theta_2 = 1-\theta_1$, $\gamma_1, \gamma_2 > 0$. In 1968, Tallis and Light [52] suggested the use of the fractional moments. Considering a mixture of two exponential distributions, they showed, by some numerical studies, that if a so called "optimal combination of moments" is used, the efficiency of the estimates will increase.

A study of the bias and accuracy of the moment estimators with particular reference to mixtures of two normal distributions has been given by Robertson and Fryer [49]. Their results suggest that although the method of moments generally leads to estimators which are less efficient than the maximum likelihood estimates, their use

can be justified when the absolute maxima of the likelihood function is unobtainable.

The method of Maximum Likelihood has also played an important role in the estimation problems of the theory of mixtures of distributions. The likelihood function of a mixture of distributions based on the observations $x_1,\ldots,x_n$ from the mixture is of the form

$$L = \prod_{i=1}^{n} \sum_{j=1}^{k} \theta_j f(x_i ; \alpha_j) \qquad \alpha_j \in \mathcal{A} \quad x_i \in \mathcal{X} \quad (1.5.8)$$

where $f(. ; \alpha_j)$ is the density function corresponding to the distribution function $F(. ; \alpha_j)$ for $j = 1,\ldots,k$. Recall from Section 1.2 that each $\alpha_j \in \mathcal{A}$ is an s-dimensional vector. The maximum likelihood estimators of $\theta_1,\ldots,\theta_k$ and $\alpha_1,\ldots,\alpha_k$ are those values which maximize (1.5.8) for fixed $x_1,\ldots,x_n$. The equations for the estimates usually turn out to be non-linear and difficult to solve.

A mixture of two or more normal densities has again been the centre of attraction for many investigators. It is to be stressed that in such cases if means, variances and mixing proportions are unknown, the likelihood function becomes unbounded near many points in the parameter space (c.f. Section 5.1) and hence the method of maximum likelihood breaks down. This important fact has been observed by Day [17], Fryer and Robertson [21] and Behboodian [5]. (Incidently, a similar feature characterizes other mixtures of distributions such as mixtures of displaced exponential densities). By imposing sufficient restrictions on the parameter space (e.g. equality of variances of the component distributions), or using coarsely grouped observations, the method of maximum likelihood can meaningfully be used and it often leads to efficient estimates. The example given by Behboodian [5] is that if in a mixture of two

normal densities given by (1.5.5), $\mu = \mu_1 = \mu_2$, then as an estimate of μ, the efficiency of the sample mean against the sample median tends to zero as σ_1/σ_2 is made arbitrarily small or large. On the other hand for $\sigma_1 = \sigma_2$, the sample mean is fully efficient for the true mean.

Rao [42] was first to use maximum likelihood estimation in mixtures of distributions. He considered a mixture of two normal densities with equal variances and obtained grouped maximum likelihood estimates for the means and the common variances of the components and the mixing proportion of the mixture of densities. He applied the method to a sample of size 454, tested his results with a chi-squared goodness of fit test. The fit turned out to be reasonably good. Hasselblad [25] has proposed a general method of iteration to obtain the maximum likelihood estimates of the k means, k variances and k-1 mixing proportions of a mixture of k normal densities from grouped data. In view of the fact that the likelihood function for ungrouped data is unbounded for this problem, Behboodian [5] has proposed using the values corresponding to the largest stationary maximum of the likelihood function as the estimates. The method may well lead to 'reasonable' estimates in many cases, but the estimates will clearly not possess the optimal asymptotic properties of the maximum likelihood estimates. Fryer and Robertson [21] claim that Behboodian's method is what Hasselblad [25] has proposed in effect, since although Hasselblad starts by considering grouped maximum likelihood estimates, he then assumes the width of each group to be sufficiently small to allow us to replace each group probability divided by its length by the appropriate value of the density function. They believe that the effectiveness of Hasselblad's method will vary a great deal depending on the parameter values. They discuss and compare the estimates of the parameters of a mixture of k normal

distributions obtained by (i) the method of moments, (ii) the method of maximum likelihood from grouped observations and (iii) the method of minimum χ^2 again when the data is grouped. Their comparison is based upon the bias and the mean squared error of the estimates and they concluded, from some numerical studies, that as regards the bias, minimum χ^2 estimators seem to be slighly better than grouped maximum likelihood estimators, but the difference between the two is often very small. Moment estimators are sometimes better and sometimes worse than the grouped estimators, and furthermore the differences are often considerable. Comparing the mean squared errors, they found that the grouped estimators are usually markedly superior to the moment estimators. This superiority, however, is not completely uniform, since the performances of the moment estimators are often preferred to that of the grouped estimates when the components of a mixture of distributions are not well separated.

A comparison of the method of moments and the method of maximum likelihood in estimation of the parameters of a mixture of two normal densities has also been done by Tan and Chang [53]. To ensure the existence of the maximum likelihood estimators, they restrict themselves to the case when the components have equal variances. Their comparison is based upon the asymptotic efficiencies of the estimates and they concluded that maximum likelihood generally results in better estimators especially when $\Delta = |(\mu_1-\mu_2)/\sigma|$ is small. Here μ_1 and μ_2 denote the means of the components and σ denotes their common standard deviation.

Maximum likelihood estimators of the parameters of a mixture of two normal densities have also been used by Dick and Bowden [19] and Hosmer [28]. They use numerical techniques to derive their estimators when independent samples from one or both components are available.

Macdonald [34] uses an iteration technique to derive the maximum likelihood estimates of the parameters of a mixture of $k \geq 2$ normal densities when (i) the mixing proportions are the only unknown parameters, (ii) the means and the common variance of the components are also unknown. He provides FORTRAN computer programs for the calculations. Day [17] has considered a mixture of two multivariate normal densities with identical but unknown covariance matrices. He derives estimates for the mean vectors and the common covariance matrix and also for the mixing proportions. He uses both the method of moments and the method of maximum likelihood and finds that the maximum likelihood estimators are generally better.

Unfortunately, the Bayesian analysis of the estimation problems of finite mixtures of distributions has not been yet fully investigated. The only publication on the subject is due to Behboodian [6]. He considers a mixture of two densities $f_1(.)$ and $f_2(.)$ viz:

$$g_{\underset{\sim}{\theta}}(x) = \theta_1 f_1(x) + (1-\theta_1) f_2(x) \qquad x \in \mathcal{X} \qquad (1.5.9)$$

for $0 \leq \theta_1 \leq 1$, in which the mixing proportion θ_1 is the only unknown parameter. By taking a beta distribution as the prior distribution of the mixing proportion, he derives its posterior distribution based on a sample of size n from the mixture (1.5.9). He shows that the posterior distribution is a mixture of (n+1) beta distributions and derives its mean and variance. Generalization to mixtures of more than two components is considered. The following comment illustrates a somewhat unusual characteristic of the Bayes estimator for the mixing proportion in a mixture of two densities given by (1.5.9). Let $\gamma(\theta)$; $0 \leq \theta \leq 1$ be the prior distribution function of θ_1. Then

the Bayes estimator under the square error loss function (being the mean of the Posterior distribution) based on the observations $x_1,\ldots,x_n$ is

$$\frac{\int_0^1 \theta_1 \prod_{j=1}^n g_{\underset{\sim}{\theta}}(x_j)\, d\,\gamma(\theta_1)}{\int_0^1 \prod_{j=1}^n g_{\underset{\sim}{\theta}}(x_j)\, d\,\gamma(\theta_1)} . \qquad (1.5.10)$$

But

$$\prod_{j=1}^n g_{\underset{\sim}{\theta}}(x_i) = \prod_{j=1}^n [\theta_1(f_1(x_j) - f_2(x_j)) + f_2(x_j)]$$

$$= \sum_{r=0}^n \theta_1^{n-r} A_{r,n}$$

where

$$A_{r,n} = \sum_{i_1<\ldots<i_r} \prod_{j=1}^r f_2(x_j) \prod_{\substack{j=1 \\ j \neq i_1,\ldots,i_r}}^n (f_1(x_j) - f_2(x_j))$$

for $r = 0,\ldots,n$, so that (1.5.10) gives, as the Bayes estimator of θ_1,

$$\frac{\sum_{r=0}^n A_{r,n} \int_0^1 \theta^{n-r+1}\, d\,\gamma(\theta)}{\sum_{r=0}^n A_{r,n} \int_0^1 \theta^{n-r}\, d\,\gamma(\theta)}$$

and it is seen that the Bayes estimator depends only on the first (n+1) moments of the *a priori* distribution and not on the *a priori* distribution itself. Hence the class of all Bayes estimators can be represented by the class of all vectors $(\xi_1,\ldots,\xi_{n+1})'$ where ξ_i, $i = 1,\ldots,n+1$ denotes the ith central moment of some *a priori* distribution on $[0,1]$.

1.6 Other Aspects of Mixtures of Distributions

Apart from estimation and identifiability of mixtures of distributions, there have been comparatively very few publications about other problems related to mixtures of distributions. In the

area of hypothesis testing, Tiago De Oliveira [58] has proposed a procedure to test whether the distribution function of a given random sample is either of the given discrete distribution functions $F_1(x)$ and $F_2(x)$ or a mixture of them. His method, however, is not efficient but only provides a rapid procedure. Given three random samples from three distinct populations, Thomas [57] gives a distribution free procedure to test whether the distribution function of one of the populations is a mixture of the distribution functions of the remaining two populations. His test statistic is based upon a function of the ranks of the observations. Consistency and asymptotic normality of the test statistic is proved. The author also gives a test statistic for the hypothesis that the mixing proportion is a constant parameter against the alternative that it is a function whose domain is the sample space of the three populations.

Another area which has interested some statisticians is the problem of finding lower bounds for the variance of the estimate of the mixing proportion θ_1 in a mixture of two densities (1.5.10) and their generalizations to mixtures of k ($2 \leq k < \infty$) densities. It is known that under certain regularity conditions (Zacks [61]), the variance of any unbiased estimator of θ_1 based on a random sample of size n, cannot be less than the Cramér-Rao lower bound $\frac{1}{n\, I(\theta_1)}$ where $I(\theta_1)$ is the Fisher's information function in a single observation. Hill [26] showed that the Cramér-Rao lower bound is

$$\frac{\theta_1 (1 - \theta_1)}{n (1 - S(\theta_1))} \tag{1.6.1}$$

where

$$S(\theta_1) = \int \frac{f_1(x)\, f_2(x)}{g_{\underset{\sim}{\theta}}(x)}\, dx$$

and further derived series expansions for $S(\theta_1)$ when the densities $f_1(x)$ and $f_2(x)$ are density functions of (i) normal distributions with equal scale parameters, (ii) exponential distributions. Boes [10] derived necessary and sufficient conditions for the attainment of the bound (1.6.1). He generalized his results to mixtures of k ($2 \leq k < \infty$) densities.

When the Cramér-Rao bound is not attained, it is sometimes possible to derive greater lower bounds based on Bhattacharyya matrix (Zacks [61]). Denote by $L(\theta_1)$ the likelihood function of the mixing proportion θ_1 based on a sample of size n from the mixture of densities (1.5.13). Let T be an estimate of some function of θ_1 having expected value $\tau(\theta_1)$, then $\mathrm{Var}(T) \geq t'J^{-1}t$ where

$$t = \left(\frac{\partial}{\partial\theta_1}\tau(\theta_1), \ldots, \frac{\partial^m}{\partial\theta_1^m}\tau(\theta_1)\right)',$$

and J is the m×m matrix with its (r,s)th element being the expected value of

$$\frac{\left(\dfrac{\partial^r L(\theta_1)}{\partial\theta_1^r} \cdot \dfrac{\partial^s L(\theta_1)}{\partial\theta_1^s}\right)}{(L(\theta_1))^2}$$

for r,s = 1,...,m, provided, of course that the derivatives exist. Matrix J is called the Bhattacharyya matrix of order k. Whittaker [59] derived the Bhattacharyya matrix for a mixture of two distributions.

Behboodian [7] has given a numerical method for computation of the Fisher's information matrix about the five parameters (two means, two variances and the mixing proportion) of a mixture of two normal densities. He shows that the computation of the information matrix

leads to the numerical evaluation of an integral and uses various numerical techniques to solve the integral.

1.7 Applications of Mixtures of Distributions

Practical problems involving mixtures of distributions arise in many different fields of study. These include biology, engineering, fisheries, psychology and medicine. A useful account of some of these applications can be found in Blischke [9] and in this section we refer to some of the authors who have used mixtures of distributions in their investigations of various applied problems.

For example, length-frequency data from a fish population is known to be best approached as a mixture of distributions. The population is composed of a number of component age groups mixed together in some proportions; each age group has a distinct length-frequency curve and the length-frequency curve for the population is a mixture of these component distributions. Macdonald [34], Hosmer [28] and Dick and Bowden [19] discuss these situations with respect to sampling from a normal populations. Another area of biology in which mixtures of distributions are frequently encountered is genetics where one is concerned with the study of inheritance in both natural and man-made populations. A proper genetic analysis of such populations sometimes involves mixtures of distributions. Rushforth *et al* [50] have used a mixture of two normal distributions as a model for the blood glucose level of a sample of Pima Indians, a population known to have an extremely high prevalence of diabetes mellitus.

A number of non-biological applications of mixtures of distributions, mostly from the chemical industry, have been discussed by Medgyessy [37]. These include the use of finite mixtures of

normal distributions in the investigation of absorption spectra and of electrophoretical separation of proteins of different molecular weight contained in a solution. Medgyessy also gives some applications of finite mixtures of binomial distributions.

Mixtures of distributions are known to fit adequately many distributions arising in technological and physical applications, particularly in the field of life-testing. Kao [31], for example, discusses a problem in life-testing of the electron tubes subjected to a sudden and delayed failure. Mendenhall and Hader [38] and Cox [14] discuss a similar problem in life testing of radio equipments. Falls [20] mentions that a mixture of two Weibull distributions, as well as being an appropriate model in life-testing, is also commonly used in the analysis of atmospheric data and consequently is of interest to aerospace scientists.

Amongst important applications of mixtures of distributions, is its appropriateness as a model in various psychological experiments. Lord [33] has discussed such applications in relation to mental test theory and Thomas [57] uses a mixture of two distributions as a model for psychological reaction time experiments.

CHAPTER 2

THE METHOD OF MOMENTS

2.1 Introduction

The method of moments is probably the oldest method of estimating the unknown parameters in a distribution. It often leads to equations which are more tractable than those derived from other methods. It is mainly for this reason that the method is still being used although a main disadvantage is the fact that it often results in inefficient estimates.

In the context of estimation problems related to mixtures of distributions, K. Pearson [41] was first to use the method to estimate the five parameters of a mixture of two normal distributions. The method consists of equating as many sample moments to their corresponding expected values as there are unknown parameters and solving the resulting equations. Tallis and Light [52] considered a rather different version of the method by taking fractional moments. They showed that by using a so called "optimal combination of moments", the efficiency of the estimates would increase.

In this chapter, we take a somewhat more general approach. Denote by $G_{\underset{\sim}{\theta}}(.)$ the distribution function of a mixture of distributions with k components $F_1(.), \ldots, F_k(.)$ and with mixing proportions $\theta_1, \ldots, \theta_k$ respectively. Let the random variables $X_1, \ldots, X_n$ have common distribution function $G_{\underset{\sim}{\theta}}(.)$ with respective realizations $x_1, \ldots, x_n$. Then the conventional moment estimators of $\theta_1, \ldots, \theta_k$ are the solutions of

$$\int_{x \in \mathcal{X}} x^t dG_{\underset{\sim}{\theta}}(x) = \int_{x \in \mathcal{X}} x^t dG_n(x) \qquad t = 1, \ldots, k$$

where $G_n(.)$ is the empirical distribution function and $\mathcal{X}$ is that subset of the real line to which each $F_j(.)$; $1 \leq j \leq k$ (and hence

$G_{\underset{\sim}{\theta}}(.)$) assigns probability one. Instead of using the sample and population moments of X^t for $t = 1, \ldots, k$, we define a real-valued function $h(x,t)$ of $x \in \mathcal{X}$ and the real value $t \in R$ and evaluate the two quantities $\lambda_{\underset{\sim}{\theta}}(t) = \int h(x,t)\, dG_{\underset{\sim}{\theta}}(x)$ and $\lambda_n(t) = \int h(x,t)\, dG_n(x)$ at $t_1, \ldots, t_m$; $m \geq k$, chosen such that the rank of the matrix A formed with $\int h(x,t_i)\, dF_j(x)$ for $i = 1, \ldots, m$, $j = 1, \ldots, k$ as its (i,j)th entry is k. The estimators of $\theta_1, \ldots, \theta_k$ are then obtained by fitting $\lambda_n(t_1), \ldots, \lambda_n(t_m)$ to $\lambda_{\underset{\sim}{\theta}}(t_1), \ldots, \lambda_{\underset{\sim}{\theta}}(t_m)$ by the method of least squares.

It is shown that our estimates possess certain desired properties and special attention is given to the more amenable case $k = 2$. At the end of the chapter, we shall see how our estimators work in practice in the light of some Monte Carlo studies.

2.2 Method of Estimation

Let

$$G_{\underset{\sim}{\theta}}(x) \equiv \sum_{j=1}^{k} \theta_j F_j(x) \qquad x \in \mathcal{X}, \qquad (2.2.1)$$

where $0 \leq \theta_j \leq 1$ for $j = 1, \ldots, k$ and $\sum_{j=1}^{k} \theta_j = 1$, denote a mixture of k known distribution functions $F_1, \ldots, F_k$. Let $\underset{\sim}{e}_j$, $1 \leq j \leq k$ be the standard k-dimensional unit vector, i.e. a k-dimensional vector with 1 at the jth position and zero elsewhere and $\underset{\sim}{\theta} = \sum_{j=1}^{k} \theta_j \underset{\sim}{e}_j = (\theta_1, \ldots, \theta_k)'$ be the vector of the unknown mixing proportions $\theta_1, \ldots, \theta_k$. For a function $\alpha(.)$, integrable over $\mathcal{X}$ with respect to $F_1, \ldots, F_k$, we define

$$E_{\underset{\sim}{e}_j}(\alpha(X)) = \int \alpha(x)\, dF_j(x) \qquad j = 1, \ldots, k$$

and

$$E_{\underset{\sim}{\theta}}(\alpha(X)) = \int \alpha(x)\, dG_{\underset{\sim}{\theta}}(x)$$

where X is a random variable whose distribution function is given by

(2.2.1).

Let $h(x,t)$ be a real-valued function, right-continuous in $t \in R$ for each $x \in \mathcal{X}$ i.e.

$$h(x,t) = h(x, t + 0) \qquad t \in R$$

for each $x \in \mathcal{X}$. Let $\mathcal{T}_j = \{t : t \in R, \ E_{e_j}(|h(X,t)|) < \infty\}$ for $j = 1, \ldots, k$, so that $\mathcal{T} = \bigcap_{j=1}^{k} \mathcal{T}_j = \{t : t \in R, \ E_{\theta}(|h(X,t)|) < \infty\}$. Then

$$\lambda_{\theta}(t) = E_{\theta}(h(X,t)) = \int h(x,t)\, dG_{\theta}(x) \tag{2.2.2}$$

and

$$\lambda_{e_j}(t) = E_{e_j}(h(X,t)) = \int h(x,t)\, dF_j(x) \tag{2.2.3}$$

for $j = 1, \ldots, k$ are right-continuous functions of $t \in \mathcal{T}$. For the rest of this chapter, the variable t, defined in this way, will be confined to $t \in \mathcal{T}$ unless otherwise stated.

Now, from (2.2.2) and (2.2.3), we have

$$\lambda_{\theta}(t) = \sum_{j=1}^{k} \theta_j \lambda_{e_j}(t) \tag{2.2.4}$$

and we estimate $\lambda_{\theta}(t)$ by

$$\lambda_n(t) = \int h(x,t)\, dG_n(x) = \frac{1}{n} \sum_{i=1}^{n} h(x_i,t) \tag{2.2.5}$$

where x_i, $i = 1, \ldots, n$ is the realization of the random variable X_i whose distribution is given by (2.2.1) and as before $G_n(.)$ is the empirical distribution function being the realization of the random function $\Gamma_n(.)$. Denoting by $L_n(t)$ the random function whose realized value is $\lambda_n(t)$, we have

$$L_n(t) = \int h(x,t)\, d\Gamma_n(x) = \frac{1}{n} \sum_{i=1}^{n} h(X_i,t) \tag{2.2.6}$$

and

$$E_{\underset{\sim}{\theta}}(L_n(t)) = E_{\underset{\sim}{\theta}}(h(X,t)) = \lambda_{\underset{\sim}{\theta}}(t) = \sum_{j=1}^{k} \theta_j \lambda_{\underset{\sim}{e}_j}(t). \qquad (2.2.7)$$

We now write

$$\lambda_n(t) = \sum_{j=1}^{k} \theta_j \lambda_{\underset{\sim}{e}_j}(t) + \varepsilon(t) \qquad (2.2.8)$$

where $\varepsilon(t)$ is the realization of a random function $\&(t)$ such that

$$E_{\underset{\sim}{\theta}}(\&(t)) = 0 .$$

Note also that since $E_{\underset{\sim}{\theta}}\{|h(X_i,t)|\} < \infty$ for $i = 1, \ldots, n$, it follows from the strong law of large numbers that $L_n(t)$, being the sum of independently and identically distributed random variables, converges almost surely to $E_{\underset{\sim}{\theta}}(h(X_i,t))$ as $n \to \infty$ i.e.

$$L_n(t) \xrightarrow{a.s.} \lambda_{\underset{\sim}{\theta}}(t) \qquad \text{as} \qquad n \to \infty \quad .$$

Choose distinct values $t_1, \ldots, t_m \in \mathcal{T}$, $m \geq k$ in such a way that the rank of the matrix

$$A = \begin{bmatrix} \lambda_{\underset{\sim}{e}_1}(t_1) & \cdots & \lambda_{\underset{\sim}{e}_k}(t_1) \\ \vdots & & \vdots \\ \lambda_{\underset{\sim}{e}_1}(t_m) & \cdots & \lambda_{\underset{\sim}{e}_k}(t_m) \end{bmatrix} \qquad (2.2.9)$$

is k and therefore $\det(A'A) \neq 0$ and $A'A$ is invertible. It is shown in the following lemma that by suitably restricting $h(x,t)$, such choices of $t_1, \ldots, t_m$ are always possible. Evaluating (2.2.8) at $t_1, \ldots, t_m$, the linear model

$$\underset{\sim}{\lambda}_n = \underset{\sim}{\lambda}_{\underset{\sim}{\theta}} + \underset{\sim}{\varepsilon} = A\underset{\sim}{\theta} + \underset{\sim}{\varepsilon} \qquad (2.2.10)$$

where $\underset{\sim}{\lambda}_n = (\lambda_n(t_1), \ldots, \lambda_n(t_m))'$, $\underset{\sim}{\lambda}_{\underset{\sim}{\theta}} = (\lambda_{\underset{\sim}{\theta}}(t_1), \ldots, \lambda_{\underset{\sim}{\theta}}(t_m))'$ and $\underset{\sim}{\varepsilon} = (\varepsilon(t_1), \ldots, \varepsilon(t_m))'$ is of full rank and $\underset{\sim}{\theta}$ may be estimated by using the least squares theories. Upon minimizing $(\underset{\sim}{\lambda}_n - \underset{\sim}{\lambda}_{\underset{\sim}{\theta}})'(\underset{\sim}{\lambda}_n - \underset{\sim}{\lambda}_{\underset{\sim}{\theta}})$ with respect to $\underset{\sim}{\theta}$, we obtain

$$\hat{\underset{\sim}{\theta}}_n = (A'A)^{-1} A'\underset{\sim}{\lambda}_n \qquad (2.2.II)$$

as an estimator of $\underset{\sim}{\theta}$.

Lemma 2.2.1: If

(i) the mixture $G_{\underset{\sim}{\theta}}(x)$, given by (2.2.1), is identifiable,

(ii) for each x belonging to a compact subset S of $\mathfrak{X}$, the set

$$T = \{h(x): h(x) = \sum_{i=1}^{\ell} \alpha_i\, h(x,t_i);\ \ell \in Z^+,$$

$$\underset{\sim}{\alpha} = (\alpha_1, \ldots, \alpha_\ell)' \in R^\ell,\ t_i \in \mathcal{T},\ i = 1, \ldots, \ell\}$$

is everywhere dense in C(S), the space of all the continuous functions on S with the property that

$$|h(x)| \leq M_h(x)$$

where $M_h(x)$ is an integrable function with respect to $F_1(.), \ldots, F_k(.)$, i.e.

$$\int M_h(x)\, dF_j(x) \quad \text{exists for} \quad j = 1, \ldots, k,$$

then we can choose distinct values $t_1, \ldots, t_m \in \mathcal{T}$, $m \geq k$ such that the vectors

$$\underset{\sim}{\lambda}_{\underset{\sim}{e}_j} = (\lambda_{\underset{\sim}{e}_j}(t_1), \ldots, \lambda_{\underset{\sim}{e}_j}(t_m))' \qquad j = 1, \ldots, k$$

are linearly independent.

Here, Z^+ denotes the space of all positive integers. Note also that the condition (ii) is equivalent to saying that every element of C(S) is the limit as $\ell \to \infty$ of a member of T.

Proof: If suffices to show that there exists values $t_1, \ldots, t_m \in \mathcal{T}$ such that for $c_1, \ldots, c_k \in R$, the relation $\sum_{j=1}^{k} c_j\, \lambda_{\underset{\sim}{e}_j}(t_i) = 0$ holds

if and only if $c_1 = c_2 = \ldots = c_k = 0$ for $i = 1, \ldots, m$.

The essence of the proof is the use of a fundamental property of Stieltjes integrals. It is known (e.g. Riesz and Nagy [46]) that a necessary and sufficient conditon that the Stieltjes integral

$$\int_S \beta(x)\, d\rho(x)$$

where S is compact, formed with a fixed function of bounded variation $\rho(x)$, be zero for every continuous function $\beta(x)$ is that the function $\rho(x)$ be constant on a set everywhere dense in S.

Suppose now that on the contrary there exists real values $c_1, \ldots, c_k \in R$ not all zero such that

$$\sum_{j=1}^{k} c_j \lambda_{\underset{\sim}{e}_j}(t) = 0 \qquad \text{for every } t \in \mathcal{T} \qquad (2.2.12)$$

then from (2.2.3), we have

$$\int h(x,t)\, d\Big(\sum_{j=1}^{k} c_j F_j(x)\Big) = 0 \qquad \text{for every } t \in \mathcal{T}. \qquad (2.2.13)$$

Since $F_j(x)$ $j = 1, \ldots, k$ are bounded monotonic functions, $\sum_{j=1}^{k} c_j F_j(x)$ is of bounded variation. Let S be a compact subset of $\mathcal{X}$, then for any $\beta(.) \in C(S)$, there exists $h(.) \in T$ such that $\beta(x) = \lim_{\ell\to\infty} h(x)$ for every $x \in S$ and thus

$$\int_S \beta(x)\, d\Big(\sum_{j=1}^{k} c_j F_j(x)\Big) = \int_S \lim_{\ell\to\infty} h(x)\, d\Big(\sum_{j=1}^{k} c_j F_j(x)\Big).$$

Now since $|h(x)| \le M_h(x)$ i.e. $h(x)$ is bounded by an integrable function, it follows from the Lebesgue dominated convergence theorem that

$$\int_S \lim_{\ell\to\infty} h(x)\, d\Big(\sum_{j=1}^{k} c_j F_j(x)\Big) = \lim_{\ell\to\infty} \int_S h(x)\, d\Big(\sum_{j=1}^{k} c_j F_j(x)\Big)$$

$$= \lim_{\ell\to\infty} \int_S \sum_{i=1}^{\ell} \alpha_i h(x,t_i)\, d\Big(\sum_{j=1}^{k} c_j F_j(x)\Big)$$

for some $\alpha_1, \ldots, \alpha_\ell \in R$ and $t_1, \ldots, t_\ell \in \mathcal{T}$, and so

$$\int_S \beta(x)\, d\Big(\sum_{j=1}^{k} c_j F_j(x)\Big) = \lim_{\ell\to\infty} \Big[\sum_{i=1}^{\ell} \alpha_i \int_S h(x,t_i)\, d\Big(\sum_{j=1}^{k} c_j F_j(x)\Big)\Big]$$

$$= 0 \text{ using (2.2.13)} .$$

Thus by the (above mentioned) property of Stieltjes integrals, $\sum_{j=1}^{k} c_j F_j(x)$ is constant on a dense set of every compact subset of $\mathcal{X}$.

Now, given any $x_0 \in \mathcal{X}$, there is a compact subset K of $\mathcal{X}$ containing x_0 and therefore $\sum_{j=1}^{k} c_j F_j(x)$ is constant on a dense set D of K i.e.

$$\sum_{j=1}^{k} c_j F_j(x) = c$$

for every $x \in D$. If $x_0 \notin D$, we can construct a sequence $\{x_r\} \subset D$, $r = 1, 2, \ldots$ such that $\lim_{r\to\infty} x_r = x_0$ and by letting $r\to\infty$ in

$$\sum_{j=1}^{k} c_j F_j(x_r) = c \qquad r = 1, 2, \ldots$$

we have

$$c = \lim_{r\to\infty} \sum_{j=1}^{k} c_j F_j(x_r) = \sum_{j=1}^{k} c_j F_j(x_0) \qquad (2.2.14)$$

for every $x_0 \in \mathcal{X}$ that is a continuity point of $F_1, \ldots, F_k$. If, on the other hand, x_0 is a discontinuity point of at least one of the distribution functions $F_1, \ldots, F_k$, then from the sequence $\{x_r\} \subset D$, we can construct an increasing subsequence $\{x_{r_i}\} \subset D$ and a decreasing subsequence $\{x_{r_i'}\} \subset D$ such that

$$x_{r_i} \uparrow x_0 \quad \text{as} \quad r_i \to \infty$$

and

$$x_{r_i'} \downarrow x_0 \quad \text{as} \quad r_i' \to \infty$$

and by letting r_i and r_i' tend to infinity in $\sum_{j=1}^{k} c_j F_j(x_{r_i}) = c$ and $\sum_{j=1}^{k} c_j F_j(x_{r_i'}) = c$ respectively, we find that (2.2.14) also holds at the discontinuity points. In particular, by letting $x \to -\infty$ in $\sum_{j=1}^{k} c_j F_j(x) = c$, we get

$$\sum_{j=1}^{k} c_j F_j(x) = 0 \qquad x \in \mathfrak{X} \ . \tag{2.2.15}$$

But (2.2.15) violates the identifiability assumption of $G_{\underset{\sim}{\theta}}(.)$. We mentioned in Chapter 1 that it is shown by Yakowitz and Spragins [60] that a necessary and sufficient conditon for the mixture of distribution $G_{\underset{\sim}{\theta}}(.)$ given by (2.2.1) to be identifiable is that the components $F_1, \ldots, F_k$ are linearly independent. Thus (2.2.14) holds if and only if $c_1 = \ldots = c_k = 0$ which contradicts (2.2.12) and therefore there exists at least one point $t_1 \in \mathcal{T}$ such that $\sum_{j=1}^{k} c_j \lambda_{\underset{\sim}{e}_j}(t_1) = 0$ with $c_j \in R$ for $j = 1, \ldots, k$, if and only if $c_1 = c_2 = \ldots c_k = 0$.

Finally, by the right continuity of $\lambda_{\underset{\sim}{e}_j}(t)$ for $j = 1, \ldots, k$ on $\mathcal{T}$, it follows that the points neighbouring t_1 and to the right of it have the same property as t_1 and hence we can find $t_1, \ldots, t_m \in \mathcal{T}$ such that the relation

$$\sum_{j=1}^{k} c_j \lambda_{\underset{\sim}{e}_j}(t_i) = 0 \qquad c_j \in R,\ i = 1, \ldots, m$$

holds if and only if $c_1 = c_2 = \ldots = c_k = 0$ which completes the proof of the lemma.

2.3 Properties of the Estimates

Let $\underset{\sim}{Z}_n$ denote a random vector whose realization is given by (2.2.10). Then

$$\underset{\sim}{Z}_n = (A'A)^{-1} A' \underset{\sim}{L}_n \tag{2.3.1}$$

where $\underset{\sim}{L}_n = (L_n(t_1), \ldots, L_n(t_m))'$. It is clear that $\hat{\underset{\sim}{\theta}}$ is an unbiased estimator of $\underset{\sim}{\theta}$ for

$$E_{\underset{\sim}{\theta}}[\underset{\sim}{Z}_n] = (A'A)^{-1}A' \; E_{\underset{\sim}{\theta}}[\underset{\sim}{L}_n] = (A'A)^{-1}A' \; \underset{\sim}{\lambda}_{\underset{\sim}{\theta}} = (A'A)^{-1}A'A \; \underset{\sim}{\theta} = \underset{\sim}{\theta} \tag{2.3.2}$$

from (2.2.6).

Proposition 2.3.1: If

$$c_{\underset{\sim}{\theta}}(r,s) \equiv \mathrm{Cov}_{\underset{\sim}{\theta}}(h(X,t_r),\, h(X,t_s)) \qquad r,s = 1, \ldots, m$$
$$= E_{\underset{\sim}{\theta}}(h(X,t_r) - \underset{\sim}{\lambda}_{\theta}(t_r))(h(X,t_s) - \underset{\sim}{\lambda}_{\theta}(t_s))$$

then

$$\mathrm{Cov}_{\underset{\sim}{\theta}}(L_n(t_r),\, L_n(t_s)) = \frac{1}{n}\, c_{\underset{\sim}{\theta}}(r,s)$$

Proof: From (2.2.5),

$$\mathrm{Cov}_{\underset{\sim}{\theta}}(L_n(t_r),\, L_n(t_s)) = E_{\underset{\sim}{\theta}}\left[\left(\frac{1}{n}\sum_{i=1}^{n} h(X_i,t_r) - \underset{\sim}{\lambda}_{\theta}(t_r)\right)\left(\frac{1}{n}\sum_{i=1}^{n} h(X_i,t_s) - \underset{\sim}{\lambda}_{\theta}(t_s)\right)\right]$$

$$= \frac{1}{n^2} E_{\underset{\sim}{\theta}}\left[\sum_{i=1}^{n} (h(X_i,t_r) - \underset{\sim}{\lambda}_{\theta}(t_r)) \sum_{i=1}^{n} (h(X_i,t_s) - \underset{\sim}{\lambda}_{\theta}(t_s))\right]$$

$$= \frac{1}{n^2} E_{\underset{\sim}{\theta}}\left[\sum_{i=1}^{n} (h(X_i,t_r) - \underset{\sim}{\lambda}_{\theta}(t_r))(h(X_i,t_s) - \underset{\sim}{\lambda}_{\theta}(t_s)) + \sum_{\substack{i=1 \\ i\neq j}}^{n}\sum_{j=1}^{n} (h(X_i,t_r) - \underset{\sim}{\lambda}_{\theta}(t_r))(h(X_j,t_s) - \underset{\sim}{\lambda}_{\theta}(t_s))\right]$$

$$= \frac{1}{n^2}\left[\sum_{i=1}^{n} E_{\underset{\sim}{\theta}}(h(X_i,t_r) - \underset{\sim}{\lambda}_{\theta}(t_r))(h(X_i,t_s) - \underset{\sim}{\lambda}_{\theta}(t_s)) + \sum_{\substack{i=1 \\ i\neq j}}^{n}\sum_{j=1}^{n} E_{\underset{\sim}{\theta}}(h(X_i,t_r) - \underset{\sim}{\lambda}_{\theta}(t_r))\, E_{\underset{\sim}{\theta}}(h(X_j,t_s) - \underset{\sim}{\lambda}_{\theta}(t_s))\right]$$

since $X_1, \ldots, X_n$ are independent. The term in the double sum vanishes by (2.2.2) and the result follows.

Proposition 2.3.2: With the notation of the previous proposition,

$$c_{\underset{\sim}{\theta}}(r,s) = \sum_{j=1}^{k} \theta_j \, c_{\underset{\sim}{e}_j}(r,s) + \sum_{j=1}^{k} \theta_j(1-\theta_j) \, \lambda_{\underset{\sim}{e}_j}(t_r) \, \lambda_{\underset{\sim}{e}_j}(t_s)$$

$$- \sum_{i=1}^{k} \sum_{\substack{j=1 \\ i \neq j}}^{k} \theta_i \theta_j \, \lambda_{\underset{\sim}{e}_i}(t_r) \, \lambda_{\underset{\sim}{e}_j}(t_s) \qquad (2.3.3)$$

and

$$c_{\underset{\sim}{\theta}}(r,s) = \sum_{j=1}^{k} \theta_j \, c_{\underset{\sim}{e}_j}(r,s) + \sum_{i>j} \theta_i \theta_j \, a_{ij}(r) \, a_{ij}(s) \qquad (2.3.4)$$

where

$$c_{\underset{\sim}{e}_j}(r,s) = \mathrm{Cov}_{\underset{\sim}{e}_j}(h(X,t_r), h(X,t_s))$$

$$= E_{\underset{\sim}{e}_j}(h(X,t_r) - \lambda_{\underset{\sim}{e}_j}(t_r))(h(X,t_s) - \lambda_{\underset{\sim}{e}_j}(t_s)) \qquad (2.3.5)$$

and

$$a_{ij}(r) = \lambda_{\underset{\sim}{e}_i}(t_r) - \lambda_{\underset{\sim}{e}_j}(t_r) \qquad r = 1, \ldots, m$$

Proof:

$$c_{\underset{\sim}{\theta}}(r,s) = \int h(x,t_r) \, h(x,t_s) \sum_{j=1}^{k} \theta_j \, dF_j(x)$$

$$- \sum_{j=1}^{k} \theta_j \, \lambda_{\underset{\sim}{e}_j}(t_r) \sum_{j=1}^{k} \theta_j \, \lambda_{\underset{\sim}{e}_j}(t_s)$$

$$= \sum_{j=1}^{k} \theta_j (c_{\underset{\sim}{e}_j}(r,s) + \lambda_{\underset{\sim}{e}_j}(t_r) \, \lambda_{\underset{\sim}{e}_j}(t_s)) - \sum_{j=1}^{k} \theta_j^2 \, \lambda_{\underset{\sim}{e}_j}(t_r) \, \lambda_{\underset{\sim}{e}_j}(t_s)$$

$$- \sum_{i=1}^{k} \sum_{\substack{j=1 \\ i \neq j}}^{k} \theta_i \theta_j \, \lambda_{\underset{\sim}{e}_i}(t_r) \, \lambda_{\underset{\sim}{e}_j}(t_s)$$

$$= \sum_{j=1}^{k} \theta_j \, c_{\underset{\sim}{e}_j}(r,s) + \sum_{j=1}^{k} \theta_j(1-\theta_j) \, \lambda_{\underset{\sim}{e}_j}(t_r) \, \lambda_{\underset{\sim}{e}_j}(t_s)$$

$$- \sum_{i=1}^{k} \sum_{\substack{j=1 \\ i \neq j}}^{k} \theta_i \theta_j \, \lambda_{\underset{\sim}{e}_i}(t_r) \, \lambda_{\underset{\sim}{e}_j}(t_s)$$

which proves (2.3.3)

Upon substituting $\sum_{\substack{i=1 \\ i \neq j}}^{k} \theta_i = 1 - \theta_j$ in (2.3.3),

$$c_{\underset{\sim}{\theta}}(r,s) = \sum_{j=1}^{k} \theta_j \, c_{\underset{\sim}{e}_j}(r,s) + \sum_{i=1}^{k} \sum_{\substack{j=1 \\ i \neq j}}^{k} \theta_i \theta_j \, [\lambda_{\underset{\sim}{e}_i}(t_r) \, \lambda_{\underset{\sim}{e}_i}(t_s)$$

$$- \lambda_{\underset{\sim}{e}_i}(t_r) \, \lambda_{\underset{\sim}{e}_j}(t_s)]$$

$$= \sum_{j=1}^{k} \theta_j \, c_{\underset{\sim}{e}_j}(r,s) + \sum_{i>j}^{k} \theta_i \theta_j \, (\lambda_{\underset{\sim}{e}_i}(t_r) \, \lambda_{\underset{\sim}{e}_i}(t_s)$$

$$- \lambda_{\underset{\sim}{e}_i}(t_r) \, \lambda_{\underset{\sim}{e}_j}(t_s) + \lambda_{\underset{\sim}{e}_j}(t_r) \, \lambda_{\underset{\sim}{e}_j}(t_s) - \lambda_{\underset{\sim}{e}_j}(t_r) \, \lambda_{\underset{\sim}{e}_i}(t_s))$$

$$= \sum_{j=1}^{k} \theta_j \, c_{\underset{\sim}{e}_j}(r,s) + \sum_{i>j}^{k} \theta_i \theta_j \, a_{ij}(r) \, a_{ij}(s)$$

establishing (2.3.4).

Denoting by V the covariance matrix of $\underset{\sim}{Z}_n$, we have, by using the result of the Proposition 2.3.1,

$$V = \mathrm{Cov}_{\underset{\sim}{\theta}}(\underset{\sim}{Z}_n) = (A'A)^{-1}A' \, [\mathrm{Cov}_{\underset{\sim}{\theta}}(L_n(t_r), L_n(t_s))] A(A'A)^{-1}$$

$$= \frac{1}{n} (A'A)^{-1}A'CA(A'A)^{-1} \qquad (2.3.5)$$

where $[\mathrm{Cov}_{\underset{\sim}{\theta}}(L_n(t_r), L_n(t_s))]$ and C are $m \times m$ matrices with $\mathrm{Cov}_{\underset{\sim}{\theta}}(L_n(t_r), L_n(t_s))$ and $c_{\underset{\sim}{\theta}}(r,s)$ respectively as their (r,s)th entry for $r,s = 1, \ldots, m$. Thus $\hat{\underset{\sim}{\theta}}_n$ is also a consistent estimator of $\underset{\sim}{\theta}$.

Further, it follows from (2.2.5) that since $E_{\underset{\sim}{\theta}}(h(X,t))$ and $E_{\underset{\sim}{\theta}}(h^2(X,t))$ both exist and are finite, $L_n(t)$ is the sum of independently and identically distributed random variables with finite first and second moments and thus by the central limit theorem, the distribution of $L_n(t)$ approaches normality as $n \to \infty$. Hence it follows from (2.3.1) and by the standard properties of normal distributions that the asymptotic distribution of $\underset{\sim}{Z}_n$, being a linear function of independent normally distributed random variables, is a k-variate normal distribution with mean vector $\underset{\sim}{\theta}$ and covariance matrix V. Therefore $\underset{\sim}{Z}_n$ is a consistent asymptotically normal (CAN) estimator of $\underset{\sim}{\theta}$.

2.4 Adjustment of the Estimators

We can adjust the estimating procedure to take into account the linear constraint $\sum_{j=1}^{k} \theta_j = 1$. To incorporate this information the method of restricted least squares is used. We seek that value of $\underset{\sim}{\theta}$ which minimizes $(\underset{\sim}{\lambda}_n - \underset{\sim}{\lambda}_{\theta})'(\underset{\sim}{\lambda}_n - \underset{\sim}{\lambda}_{\theta})$ subject to the restriction $\underset{\sim}{1}'\underset{\sim}{\theta} = 1$, where $\underset{\sim}{1} = (1, 1, \ldots, 1)'$ is a k-dimensional vector of 1's, and $\underset{\sim}{\lambda}_n$ and $\underset{\sim}{\lambda}_{\theta}$ are as defined in Section 2.2. Therefore we minimize

$$\Phi(\underset{\sim}{\theta},\xi) = (\underset{\sim}{\lambda}_n - A\underset{\sim}{\theta})'(\underset{\sim}{\lambda}_n - A\underset{\sim}{\theta}) - 2\xi(\underset{\sim}{1}'\underset{\sim}{\theta} - 1),$$

where ξ is a Lagrange multiplier, with respect to $\underset{\sim}{\theta}$ and ξ. Setting the derivative of $\Phi(\underset{\sim}{\theta},\xi)$ with respect to $\underset{\sim}{\theta}$ equal to zero gives for the minimizing value $\hat{\underset{\sim}{\theta}}^*_n$

$$\frac{1}{2}\frac{\partial \Phi}{\partial \underset{\sim}{\theta}} = -A'\underset{\sim}{\lambda}_n + (A'A)\hat{\underset{\sim}{\theta}}^*_n - \xi^*\underset{\sim}{1} = 0$$

whence

$$\hat{\underset{\sim}{\theta}}^*_n = (A'A)^{-1}A'\underset{\sim}{\lambda}_n + \xi^*(A'A)^{-1}\underset{\sim}{1} = \hat{\underset{\sim}{\theta}}_n + \xi^*(A'A)^{-1}\underset{\sim}{1} \qquad (2.4.1)$$

where $\hat{\underset{\sim}{\theta}}_n$, given by (2.2.11) is the unrestricted least squares estimator. Premultiplying (2.4.1) by $\underset{\sim}{1}'$ gives

$$\underset{\sim}{1}'\hat{\underset{\sim}{\theta}}^*_n = \underset{\sim}{1}'\hat{\underset{\sim}{\theta}}_n + \xi^*\underset{\sim}{1}'(A'A)^{-1}\underset{\sim}{1} \; ;$$

imposing the restriction $\underset{\sim}{1}'\hat{\underset{\sim}{\theta}}^*_n = 1$ gives

$$\xi^* = \frac{1 - \underset{\sim}{1}'\hat{\underset{\sim}{\theta}}_n}{\underset{\sim}{1}'(A'A)^{-1}\underset{\sim}{1}}$$

where we note that $\underset{\sim}{1}'(A'A)^{-1}\underset{\sim}{1}$ is only a scalar factor. Inserting this back into (2.4.1) yields

$$\begin{aligned}\hat{\underset{\sim}{\theta}}^*_n &= \hat{\underset{\sim}{\theta}}_n + \frac{(A'A)^{-1}\underset{\sim}{1}(1-\underset{\sim}{1}'\hat{\underset{\sim}{\theta}}_n)}{\underset{\sim}{1}'(A'A)^{-1}\underset{\sim}{1}} \\ &= \left[I - \frac{(A'A)^{-1}\underset{\sim}{1}\underset{\sim}{1}'}{\underset{\sim}{1}'(A'A)^{-1}\underset{\sim}{1}}\right]\hat{\underset{\sim}{\theta}}_n + \frac{(A'A)^{-1}\underset{\sim}{1}}{\underset{\sim}{1}'(A'A)^{-1}\underset{\sim}{1}} \qquad (2.4.2)\end{aligned}$$

where I is the $k \times k$ identity matrix.

It is seen that the restricted estimator $\hat{\theta}^*_{\sim n}$ differs from unrestricted one $\hat{\theta}_{\sim n}$ by a linear function of the amount $(1 - \underset{\sim}{1}'\hat{\theta}_{\sim n})$ by which the unrestricted estimator fails to satisfy the restriction. The sampling properties of $\hat{\theta}^*_{\sim n}$ may be derived as follows. Denote by $Z^*_{\sim n}$ the random vector whose realization in $\hat{\theta}^*_{\sim n}$, then from (2.4.2)

$$E_{\underset{\sim}{\theta}}[Z^*_{\sim n}] = \left(I - \frac{(A'A)^{-1}\underset{\sim}{1}\underset{\sim}{1}'}{\underset{\sim}{1}'(A'A)^{-1}\underset{\sim}{1}}\right) E_{\underset{\sim}{\theta}}[Z_{\sim n}] + \frac{(A'A)^{-1}\underset{\sim}{1}}{\underset{\sim}{1}'(A'A)^{-1}\underset{\sim}{1}}$$

which from (2.3.2) gives

$$E_{\underset{\sim}{\theta}}[Z^*_{\sim n}] = \left(I - \frac{(A'A)^{-1}\underset{\sim}{1}\underset{\sim}{1}'}{\underset{\sim}{1}'(A'A)^{-1}\underset{\sim}{1}}\right) \underset{\sim}{\theta} + \frac{(A'A)^{-1}\underset{\sim}{1}}{\underset{\sim}{1}'(A'A)^{-1}\underset{\sim}{1}}$$

and by using $\underset{\sim}{1}'\underset{\sim}{\theta} = 1$, we have

$$E_{\underset{\sim}{\theta}}[Z^*_{\sim n}] = \underset{\sim}{\theta}$$

proving the unbiasedness of $\hat{\theta}^*_{\sim n}$. Further, the covariance matrix V^* of $Z^*_{\sim n}$ is given by

$$V^* = \mathrm{Cov}_{\underset{\sim}{\theta}}(Z^*_{\sim n}) = \left(I - \frac{(A'A)^{-1}\underset{\sim}{1}\underset{\sim}{1}'}{\underset{\sim}{1}'(A'A)^{-1}\underset{\sim}{1}}\right) V \left(I - \frac{(A'A)^{-1}\underset{\sim}{1}\underset{\sim}{1}'}{\underset{\sim}{1}'(A'A)^{-1}\underset{\sim}{1}}\right)'$$

where V is the covariance matrix of $Z_{\sim n}$ given by (2.3.5) and therefore

$$V^* = \frac{1}{n}\left(I - \frac{(A'A)^{-1}\underset{\sim}{1}\underset{\sim}{1}'}{\underset{\sim}{1}'(A'A)^{-1}\underset{\sim}{1}}\right)(A'A)^{-1}A'\,C\,A(A'A)^{-1}\left(I - \frac{\underset{\sim}{1}\underset{\sim}{1}'(A'A)^{-1}}{\underset{\sim}{1}'(A'A)^{-1}\underset{\sim}{1}}\right) \tag{2.4.3}$$

The asymptotic normality of $Z^*_{\sim n}$ follows almost immediately from the fact that $Z^*_{\sim n}$ is a linear function of $Z_{\sim n}$ which was shown to have a k-variate normal distribution as its limiting distribution.

In practice it may be more convenient to compute $\hat{\theta}^*_{\sim n}$ by first using the restriction $\sum_{j=1}^{k} \theta_j = 1$ and substituting $\theta_k = 1 - \sum_{j=1}^{k-1} \theta_j$ in (2.2.7)

to obtain

$$\lambda_n(t) - \lambda_{\underset{\sim}{e}_k}(t) = \sum_{j=1}^{k-1} \theta_j \, (\lambda_{\underset{\sim}{e}_j}(t) - \lambda_{\underset{\sim}{e}_k}(t)) + \varepsilon(t) \quad . \quad (2.4.4)$$

Evaluating (2.4.4) at distinct values $t_1, \ldots, t_m \in \mathcal{T}$ $m \geq k$ chosen, as before, so that the rank of matrix A given by (2.3.8) is k, we can write

$$\underset{\sim}{\lambda}^*_n = A^* \, \underset{\sim}{\theta}^+ + \underset{\sim}{\varepsilon} \quad (2.4.5)$$

where $\underset{\sim}{\lambda}^*_n = \underset{\sim}{\lambda}_n - \lambda_{\underset{\sim}{e}_k}$, $\underset{\sim}{\theta}^+ = (\theta_1, \ldots, \theta_{k-1})'$ and A^* is the $m \times (k-1)$ matrix given by

$$A^* = \begin{bmatrix} \lambda_{\underset{\sim}{e}_1}(t_1) & \cdots & \lambda_{\underset{\sim}{e}_{k-1}}(t_1) \\ \vdots & & \vdots \\ \lambda_{\underset{\sim}{e}_1}(t_m) & \cdots & \lambda_{\underset{\sim}{e}_{k-1}}(t_m) \end{bmatrix} - \begin{bmatrix} \lambda_{\underset{\sim}{e}_k}(t_1) & \cdots & \lambda_{\underset{\sim}{e}_k}(t_1) \\ \vdots & & \vdots \\ \lambda_{\underset{\sim}{e}_k}(t_m) & \cdots & \lambda_{\underset{\sim}{e}_k}(t_m) \end{bmatrix} . \quad (2.4.6)$$

We show in the following lemma that A^* has rank $k - 1$, so that $(A^*)'(A^*)$ is invertible. Thus from (2.4.5), the least squares estimator of $\underset{\sim}{\theta}^+$ is

$$\hat{\underset{\sim}{\theta}}^+ = [(A^*)'(A^*)]^{-1}(A^*)' \, \underset{\sim}{\lambda}^*_n$$

giving estimates $\hat{\theta}^*_1, \hat{\theta}^*_2, \ldots, \hat{\theta}^*_{k-1}$ of $\theta_1, \ldots, \theta_{k-1}$. We finally estimate θ_k by

$$\hat{\theta}^*_k = 1 - \sum_{j=1}^{k-1} \hat{\theta}^*_j \, .$$

Lemma 2.4.1: Given that the rank of A in (2.3.8) is k, then the rank of A^* in (2.4.6) is $k - 1$.

Proof: We prove that the columns of A* are linearly independent. Suppose that on the contrary there is a linear relationship between the columns of A*, i.e. there are real values $c_1, \ldots, c_{k-1} \in R$, not all zero, such that

$$\sum_{j=1}^{k-1} c_j(\lambda_{\underset{\sim}{e}_j}(t_i) - \lambda_{\underset{\sim}{e}_k}(t_i)) = 0 \qquad i = 1, \ldots, m . \qquad (2.4.7)$$

Choose $c_k = -(c_1 + \ldots + c_{k-1})$, then (2.4.7) can be written as

$$\sum_{j=1}^{k} c_j \lambda_{\underset{\sim}{e}_j}(t_i) = 0 \qquad i = 1, \ldots, m . \qquad (2.4.8)$$

But since the rank of A is k, (2.4.8) holds if and only if $c_1 = c_2 = \ldots = c_{k-1} = c_k = 0$ contradicting (2.4.7).

2.5 Relation to Multinomial Distribution

Let $S_j = \{x : x \text{ is a point of increase of } F_j(.)\}$ for $j = 1, \ldots, k$. Assume that each $x \in \mathfrak{X}$ is the point of increase of at most one of $F_1(.), \ldots, F_k(.)$ so that $S_1, \ldots, S_k$ are disjoint i.e.

$$S_j \cap S_r = \phi \qquad r \neq j$$
$$= S_r \qquad r = j$$

and

$$\bigcup_{j=1}^{k} S_j \subseteq \mathfrak{X}$$

Let n_j be the number of observations contained in S_j and denote by N_j the random variable whose realized value is n_j.

Now

$$G_{\underset{\sim}{\theta}}(x) = \theta_r F_r(x) \quad \text{for} \quad x \in S_r \qquad r = 1, \ldots, k$$

and

$$P_{\underset{\sim}{\theta}}(X \in S_r) = \int_{S_r} dG_{\underset{\sim}{\theta}}(x) = \int_{S_r} \theta_r \, dF_r(x) = \theta_r$$

where $P_{\underset{\sim}{\theta}}$ is the probability measure corresponding to the distribution function $G_{\underset{\sim}{\theta}}(x)$. Then

$$P_{\underset{\sim}{\theta}}(N_1 = n_1, \ldots, N_k = n_k) = \frac{n!}{n_1!n_2!\ldots n_k!}\,\theta_1^{n_1}\,\theta_2^{n_2}\ldots\theta_k^{n_k}$$

which is the familiar multinomial distribution. In this case it is known that the minimum attainable covariance matrix of an unbiased estimator of $\underset{\sim}{\theta}$ is

$$\Sigma = \frac{1}{n}\begin{pmatrix} \theta_1(1-\theta_1) & -\theta_1\theta_2 & \cdots & -\theta_1\theta_k \\ -\theta_1\theta_2 & \theta_2(1-\theta_2) & \cdots & -\theta_2\theta_k \\ \vdots & \vdots & \ddots & \vdots \\ -\theta_1\theta_k & -\theta_2\theta_k & \cdots & \theta_k(1-\theta_k) \end{pmatrix} \qquad (2.5.1)$$

in the sense that if Σ^* is the covariance matrix of any other unbiased estimator of $\underset{\sim}{\theta}$, then $(\Sigma^* - \Sigma)$ is a positive semidefinite matrix. This provides a practical check on (2.4.3).

Let m = k and define h(x,t) at $t_1, \ldots, t_k$ as follows

$$\begin{aligned} h(x,t_j) &= 1 \quad \text{if} \quad x \in S_j \\ &= 0 \qquad\quad x \notin S_j \end{aligned} \qquad j = 1, \ldots, k$$

then from (2.2.5),

$$\lambda_n(t_j) = \frac{1}{n}\,[\text{number of } x_i\text{'s in } S_j] \qquad \begin{matrix} j = 1, \ldots, k \\ i = 1, \ldots, n \end{matrix}$$

and from (2.2.3),

$$\lambda_{\underset{\sim}{e}_j}(t_r) = \int h(x,t_r)\,dF_j(x) = \begin{cases} 1 & \text{if } r = j \\ 0 & \text{otherwise} \end{cases}$$

so that from (2.2.9), A = I and thus from (2.4.3),

$$V^* = \frac{1}{n}\left(I - \frac{\underset{\sim}{1}\underset{\sim}{1}'}{k}\right) C \left(I - \frac{\underset{\sim}{1}\underset{\sim}{1}'}{k}\right)' \qquad (2.5.2)$$

Now, from (2.3.5),

$$c_{\underset{\sim}{e}_j}(r,s) = 1 - 1 = 0 \qquad r = s$$
$$= 0 \qquad r \neq s$$

and therefore from (2.3.3),

$$c_{\underset{\sim}{\theta}}(r,s) = \theta_r(1-\theta_r) \qquad r = s$$
$$= -\theta_r\theta_s \qquad r \neq s$$

and so the matrix $\frac{1}{n}$ C whose (r,s)th element is $\frac{1}{n}\, c_{\underset{\sim}{\theta}}(r,s)$ reduces to Σ in (2.5.1). But $\underset{\sim}{1}'\,\Sigma = \underset{\sim}{0}'$ and $\Sigma\,\underset{\sim}{1} = \underset{\sim}{0}$ where $\underset{\sim}{0}$ is a k-dimensional vector of 0's. Hence (2.5.2) is identical to (2.5.1) and V* is the minimum attainable covariance matrix in the sense defined above.

2.6 The Case k = 2

We now focus our attention on the case where the distribution function of the mixture $G_{\underset{\sim}{\theta}}(x)$ consists of two components $F_1(x)$ and $F_2(x)$. We have

$$G_{\underset{\sim}{\theta}}(x) = \theta_1 F_1(x) + \theta_2 F_2(x)$$

where $\underset{\sim}{\theta} = (\theta_1,\theta_2)'$, $0 \leq \theta_j \leq 1$, $j = 1, 2$ and $\theta_1 + \theta_2 = 1$ and further

$$\lambda_{\underset{\sim}{\theta}}(t) = \theta_1 \lambda_{\underset{\sim}{e}_1}(t) + \theta_2 \lambda_{\underset{\sim}{e}_2}(t).$$

Let $a_{12}(t_r) = \lambda_{\underset{\sim}{e}_1}(t_r) - \lambda_{\underset{\sim}{e}_2}(t_r)$ $r = 1, \ldots, m$ and $\underset{\sim}{a}_{12} = (a_{12}(t_1), \ldots, a_{12}(t_m))'$. Also denote by C_j $j = 1, 2$, the $m \times m$ matrix with $c_{\underset{\sim}{e}_j}(r,s)$ as its (r,s)th $r,s = 1, \ldots, m$ entry.

From (2.4.2), we can write

$$\hat{\underset{\sim}{\theta}}^*_n = \left[I - \frac{(A'A)^{-1}\underset{\sim}{1}\underset{\sim}{1}'}{\underset{\sim}{1}'(A'A)^{-1}\underset{\sim}{1}}\right](A'A)^{-1}A'\underset{\sim}{\lambda}_n + \frac{(A'A)^{-1}\underset{\sim}{1}}{\underset{\sim}{1}'(A'A)^{-1}\underset{\sim}{1}}$$

$$= HA'\underset{\sim}{\lambda}_n + \underset{\sim}{b} \qquad (2.6.1)$$

where

$$H = \left[(A'A)^{-1} - \frac{(A'A)^{-1}\underset{\sim}{1}\underset{\sim}{1}'(A'A)^{-1}}{\underset{\sim}{1}'(A'A)^{-1}\underset{\sim}{1}}\right]$$

and

$$\underset{\sim}{b} = \frac{(A'A)^{-1}\underset{\sim}{1}}{\underset{\sim}{1}'(A'A)^{-1}\underset{\sim}{1}} \quad .$$

We note that H is a 2×2 matrix with the property that

$$\underset{\sim}{1}'H = (0,0) \text{ and } H\underset{\sim}{1} = \begin{pmatrix}0\\0\end{pmatrix} \text{ where } \underset{\sim}{1} = (1,1)'$$

and therefore H must take the form

$$H = \alpha \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

where α is a constant (possibly zero). If we write

$$(A'A)^{-1} = \frac{\text{adj}(A'A)}{\det(A'A)} \quad ,$$

we can express H and $\underset{\sim}{b}$ as

$$H = \frac{1}{\det(A'A)}\left[\text{adj}(A'A) - \frac{\text{adj}(A'A)\underset{\sim}{1}\underset{\sim}{1}'\,\text{adj}(A'A)}{\underset{\sim}{1}'\text{adj}(A'A)\underset{\sim}{1}}\right]$$

and

$$\underset{\sim}{b} = \frac{\text{adj}(A'A)\underset{\sim}{1}}{\underset{\sim}{1}'\text{adj}(A'A)\underset{\sim}{1}} \quad .$$

But from (2.2.9), since $k = 2$, we have

$$(A'A) = \begin{bmatrix} \underset{\sim}{\lambda}'\underset{\sim}{e}_1\,\underset{\sim}{\lambda}\underset{\sim}{e}_1 & \underset{\sim}{\lambda}'\underset{\sim}{e}_1\,\underset{\sim}{\lambda}\underset{\sim}{e}_2 \\ \underset{\sim}{\lambda}'\underset{\sim}{e}_1\,\underset{\sim}{\lambda}\underset{\sim}{e}_2 & \underset{\sim}{\lambda}'\underset{\sim}{e}_2\,\underset{\sim}{\lambda}\underset{\sim}{e}_2 \end{bmatrix}$$

and so

$$\text{adj}(A'A) = \begin{bmatrix} \underset{\sim}{\lambda}'\underset{\sim}{e}_2\,\underset{\sim}{\lambda}\underset{\sim}{e}_2 & -\underset{\sim}{\lambda}'\underset{\sim}{e}_1\,\underset{\sim}{\lambda}\underset{\sim}{e}_2 \\ -\underset{\sim}{\lambda}'\underset{\sim}{e}_1\,\underset{\sim}{\lambda}\underset{\sim}{e}_2 & \underset{\sim}{\lambda}'\underset{\sim}{e}_1\,\underset{\sim}{\lambda}\underset{\sim}{e}_1 \end{bmatrix}$$

which gives

$$\text{adj}(A'A)\underset{\sim}{1} = (\underset{\sim}{1}'\text{adj}(A'A))' = \begin{bmatrix} \underset{\sim}{\lambda}'_{e_2}\underset{\sim}{\lambda}_{e_2} & - & \underset{\sim}{\lambda}'_{e_1}\underset{\sim}{\lambda}_{e_2} \\ \\ \underset{\sim}{\lambda}'_{e_1}\underset{\sim}{\lambda}_{e_1} & - & \underset{\sim}{\lambda}'_{e_1}\underset{\sim}{\lambda}_{e_2} \end{bmatrix}$$

$$= \begin{bmatrix} -(\underset{\sim}{\lambda}'_{e1} - \underset{\sim}{\lambda}'_{e_2})\underset{\sim}{\lambda}_{e_2} \\ \\ (\underset{\sim}{\lambda}'_{e_1} - \underset{\sim}{\lambda}'_{e_2})\underset{\sim}{\lambda}_{e_1} \end{bmatrix} = \begin{bmatrix} -\underset{\sim}{a}'_{12}\underset{\sim}{\lambda}_{e_2} \\ \\ \underset{\sim}{a}'_{12}\underset{\sim}{\lambda}_{e_1} \end{bmatrix}$$

and $\underset{\sim}{1}'\text{adj}(A'A)\underset{\sim}{1} = \underset{\sim}{a}'_{12}\underset{\sim}{\lambda}_{e_1} - \underset{\sim}{a}'_{12}\underset{\sim}{\lambda}_{e_2} = \underset{\sim}{a}'_{12}\underset{\sim}{a}_{12}$.

Hence the leading diagonal element of H is

$$\frac{1}{\det(A'A)}\left[\underset{\sim}{\lambda}'_{e_2}\underset{\sim}{\lambda}_{e_2} - \frac{(\underset{\sim}{a}'_{12}\underset{\sim}{\lambda}_{e_2})^2}{\underset{\sim}{a}'_{12}\underset{\sim}{a}_{12}}\right]$$

which after some elementary algebra leads to $\frac{1}{\underset{\sim}{a}'_{12}\underset{\sim}{a}_{12}}$. Therefore $\alpha = \frac{1}{\underset{\sim}{a}'_{12}\underset{\sim}{a}_{12}}$ and we have

$$H = \frac{1}{\underset{\sim}{a}'_{12}\underset{\sim}{a}_{12}}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \text{ and } \underset{\sim}{b} = \frac{1}{\underset{\sim}{a}'_{12}\underset{\sim}{a}_{12}}\begin{bmatrix} -\underset{\sim}{a}'_{12}\underset{\sim}{\lambda}_{e_2} \\ \underset{\sim}{a}'_{12}\underset{\sim}{\lambda}_{e_1} \end{bmatrix}$$

Substituting these values into (2.6.1) and writing $A' = \begin{bmatrix} \underset{\sim}{\lambda}'_{e_1} \\ \underset{\sim}{\lambda}'_{e_2} \end{bmatrix}$, we get for $\underset{\sim}{\hat{\theta}}^*_n = (\hat{\theta}^*_1, \hat{\theta}^*_2)'$ the following form,

$$\underset{\sim}{\hat{\theta}}^*_n = \begin{bmatrix} \hat{\theta}^*_1 \\ \hat{\theta}^*_2 \end{bmatrix} = \frac{1}{\underset{\sim}{a}'_{12}\underset{\sim}{a}_{12}}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} \underset{\sim}{\lambda}'_{e_1}\underset{\sim}{\lambda}_n \\ \underset{\sim}{\lambda}'_{e_2}\underset{\sim}{\lambda}_n \end{bmatrix} + \frac{1}{\underset{\sim}{a}'_{12}\underset{\sim}{a}_{12}}\begin{bmatrix} -\underset{\sim}{a}'_{12}\underset{\sim}{\lambda}_{e_2} \\ \underset{\sim}{a}'_{12}\underset{\sim}{\lambda}_{e_1} \end{bmatrix}$$

$$= \frac{1}{\underset{\sim}{a}'_{12}\underset{\sim}{a}_{12}}\begin{bmatrix} \underset{\sim}{a}'_{12}(\underset{\sim}{\lambda}_n - \underset{\sim}{\lambda}_{e_2}) \\ \\ -\underset{\sim}{a}'_{12}(\underset{\sim}{\lambda}_n - \underset{\sim}{\lambda}_{e_1}) \end{bmatrix} \qquad (2.6.2)$$

where it is seen that $\hat{\theta}^*_1 + \hat{\theta}^*_2 = 1$. The estimates $\hat{\theta}^*_1$ and $\hat{\theta}^*_2$ are clearly unbiased for if Z^*_1 and Z^*_2 denote random variables with realizations

$\hat{\theta}_1^*$ and $\hat{\theta}_2^*$ respectively, then

$$E_{\underset{\sim}{\theta}}(Z_1^*) = \frac{1}{\underset{\sim}{a}_{12}' \underset{\sim}{a}_{12}} \left[\underset{\sim}{a}_{12}'(E_{\underset{\sim}{\theta}}(\underset{\sim}{L}_n) - \underset{\sim}{\lambda}_{e_2})\right]$$

$$= \frac{1}{\underset{\sim}{a}_{12}' \underset{\sim}{a}_{12}} \left[\underset{\sim}{a}_{12}'(\underset{\sim}{\lambda}_{\theta} - \underset{\sim}{\lambda}_{e_2})\right] = \theta_1$$

and similarly

$$E_{\underset{\sim}{\theta}}(Z_2^*) = \frac{1}{\underset{\sim}{a}_{12}' \underset{\sim}{a}_{12}} \left[-\underset{\sim}{a}_{12}'(E_{\underset{\sim}{\theta}}(\underset{\sim}{L}_n) - \underset{\sim}{\lambda}_{e_1})\right]$$

$$= \frac{1}{\underset{\sim}{a}_{12}' \underset{\sim}{a}_{12}} \left[\underset{\sim}{a}_{12}'(\underset{\sim}{\lambda}_{e_1} - \underset{\sim}{\lambda}_{\theta})\right] = \theta_2 .$$

Further, by the use of proposition (2.3.1), we have $\mathrm{Var}_{\underset{\sim}{\theta}}(Z_1^*) = \dfrac{\underset{\sim}{a}_{12}' C \underset{\sim}{a}_{12}}{n(\underset{\sim}{a}_{12}' \underset{\sim}{a}_{12})^2}$ where, as before, C is the $m \times m$ matrix with $c_{\underset{\sim}{\theta}}(r,s)$ as its (r,s)th entry. Now, from (2.3.4), we can write

$$c_{\underset{\sim}{\theta}}(r,s) = \theta_1 c_{\underset{\sim}{e}_1}(r,s) + \theta_2 c_{\underset{\sim}{e}_2}(r,s) + \theta_1\theta_2 a_{12}(t_r) a_{12}(t_s)$$
$$r,s = 1, \ldots, m$$

so that C becomes

$$C = \theta_1 C_1 + \theta_2 C_2 + \theta_1\theta_2 \underset{\sim}{a}_{12} \underset{\sim}{a}_{12}' ,$$

and therefore

$$\mathrm{Var}_{\underset{\sim}{\theta}}(Z_1^*) = \frac{1}{n} \left\{\theta_1(1-\theta_1) + \frac{\theta_1 \underset{\sim}{a}_{12}' C_1 \underset{\sim}{a}_{12} + (1-\theta_1) \underset{\sim}{a}_{12}' C_2 \underset{\sim}{a}_{12}}{(\underset{\sim}{a}_{12}' \underset{\sim}{a}_{12})^2}\right\} \tag{2.6.3}$$

It is interesting to note that the variance of the estimator of θ_1 depends only on the two quantities $(\underset{\sim}{a}_{12}' C_j \underset{\sim}{a}_{12})/(\underset{\sim}{a}_{12}' \underset{\sim}{a}_{12})^2$ $j = 1, 2$. Thus in practice, to increase the efficiency of the estimator, $h(x,t)$ and the values $t_1, \ldots, t_m$ should be chosen so that these two quantities become as small as possible.

2.7 Monte Carlo Studies

In order to investigate the practical value of the method of estimation discussed in this chapter, a small Monte Carlo experiment was carried out. A mixture of two normal distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, i.e. normal distributions with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 respectively, was considered. The function $h(x,t)$ was chosen to be e^{xt}, so that $\lambda_{\underset{\sim}{e}_j}(t)$, $j = 1, 2$ is the conventional moment generating function of a random variable whose distribution is $N(\mu_j, \sigma_j^2)$, i.e.

$$\lambda_{\underset{\sim}{e}_j}(t) = \exp\{\mu_j t + \frac{\sigma_j^2 t^2}{2}\} \qquad j = 1, 2 \qquad (2.7.1)$$

existing for all real values of t. Consequently

$$\lambda_{\underset{\sim}{\theta}}(t) = \theta_1 \lambda_{\underset{\sim}{e}_1}(t) + \theta_2 \lambda_{\underset{\sim}{e}_2}(t)$$

where $\theta_1 + \theta_2 = 1$, $t \in R$ and $\lambda_{\underset{\sim}{e}_j}(t)$ $j = 1, 2$ is given by (2.7.1) is the moment generating function of the random variable X whose distribution is the mixture of the two components $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$.

Recall that the distinct values $t_1, \ldots, t_m \in R, m \geq 2$ have to be chosen in such a way that the rank of the matrix A given by

$$A' = \begin{bmatrix} \lambda_{\underset{\sim}{e}_1}(t_1) & \ldots\ldots & \lambda_{\underset{\sim}{e}_1}(t_m) \\ \\ \lambda_{\underset{\sim}{e}_2}(t_1) & \ldots\ldots & \lambda_{\underset{\sim}{e}_2}(t_m) \end{bmatrix}$$

is exactly 2. Thus it suffices to ensure that there are at least two values t_r and t_s, say, such that

$$\frac{\lambda_{\underset{\sim}{e}_1}(t_r)}{\lambda_{\underset{\sim}{e}_2}(t_r)} \neq \frac{\lambda_{\underset{\sim}{e}_1}(t_s)}{\lambda_{\underset{\sim}{e}_2}(t_s)} \qquad t_r \neq t_s \qquad t_r, t_s \in R \qquad 1 \leq r, s \leq m$$

which after substituting for $\lambda_{\underset{\sim}{e}_j}(t_r)$ and $\lambda_{\underset{\sim}{e}_j}(t_s)$ $j = 1, 2$ from (2.7.1) gives

$$(\mu_2 - \mu_1) + \frac{(\sigma_2^2 - \sigma_1^2)}{2} (t_r + t_s) \neq 0 \qquad (2.7.2)$$

for some $t_r \neq t_s$, $t_r, t_s \in R$ and $1 \leq r,s \leq m$.

Further, we can assume without loss of generality that one of the component distributions in the mixture, say $N(\mu_1, \sigma_1^2)$, is the standard normal distribution, i.e. $\mu_1 = 0$ and $\sigma_1^2 = 1$. This condition can always be maintained by the use of the transformation $Y = \frac{X - \mu_1}{\sigma_1}$. That is, if the distribution of X is the mixture of $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, then the distribution of $Y = \frac{X - \mu_1}{\sigma_1}$ is the mixture of $N(0,1)$ and $N\left(\frac{\mu_2 - \mu_1}{\sigma_1}, \frac{\sigma_2^2}{\sigma_1^2}\right)$. Letting $\mu = \frac{\mu_2 - \mu_1}{\sigma_1}$ and $\sigma^2 = \frac{\sigma_2^2}{\sigma_1^2}$, the required condition is maintained.

A sample of size n was generated from the mixture by sampling with probability θ_1 from $N(0,1)$ and with probability $\theta_2 = 1 - \theta_1$ from $N(\mu, \sigma^2)$. Choosing the values $t_1, \ldots, t_m \in R$ satisfying (2.7.2), the estimates $\hat{\theta}_1^*$ and $\hat{\theta}_2^*$ were found according to (2.6.2) and the experiment was repeated n_1 times with $n \cdot n_1 = N$ being a fixed number. Note also that from (2.3.5),

$$\begin{aligned} c_{\underset{\sim}{e}_j}(r,s) &= E_{\underset{\sim}{e}_j}(e^{Xt_r} \cdot e^{Xt_s}) - E_{\underset{\sim}{e}_j}(e^{Xt_r})\, E_{\underset{\sim}{e}_j}(e^{Xt_s}) \\ &= \lambda_{\underset{\sim}{e}_j}(t_r + t_s) - \lambda_{\underset{\sim}{e}_j}(t_r)\, \lambda_{\underset{\sim}{e}_j}(t_s) \end{aligned} \qquad (2.7.3)$$

for $r,s = 1, \ldots, m$ and $j = 1, 2$. So, $c_{\underset{\sim}{e}_j}(r,s)$ may be estimated by substituting the estimates $\hat{\theta}_1^*$ and $\hat{\theta}_2^*$ in (2.7.3). Having estimated $c_{\underset{\sim}{e}_j}(r,s)$, we can then estimate $\mathrm{Var}_{\underset{\sim}{\theta}}(Z_1^*)$ by using (2.6.3).

Throughout the thesis, our numerical results will be compared with Table 1 of Macdonald [35] where, as mentioned in Chapter 1, a Monte Carlo comparison of two methods of estimation are given. The first method is due to Macdonald [34] who minimizes

$$\int (G_{\underset{\sim}{\theta}}(x) - G_n(x))^2 \, dG_{\underset{\sim}{\theta}}(x)$$

and the second method is due to Choi and Bulgren [11] where

$$\int (G_{\underset{\sim}{\theta}}(x) - G_n(x))^2 \, dG_n(x)$$

is minimized. For completeness, we produce Table 1 of Macdonald [35] here (Table 1.1) in which the standard error and the mean-square-error

Table 1.1

				Estimate of θ_1			
				Macdonald		Choi & Bulgren	
n	$(\mu_2-\mu_1)/\sigma_1$	σ_2/σ_1	θ_1	Mean	MSE	Mean	MSE
50	0.25	1	0.5	0.438 ± 0.038	0.392	0.559 ± 0.038	0.392
10	0.5	1	0.5	0.536 ± 0.040	0.473	0.852 ± 0.040	0.596
10	1	1	0.5	0.495 ± 0.017	0.137	0.660 ± 0.017	0.163
10	1	1	0.8	0.804 ± 0.019	0.112	0.968 ± 0.020	0.142
20	5	1	0.5	0.493 ± 0.011	0.017	0.530 ± 0.011	0.018
10	0	2	0.5	0.372 ± 0.030	0.290	0.367 ± 0.032	0.324
50	0	2	0.5	0.484 ± 0.021	0.044	0.485 ± 0.020	0.041
10	0.5	2	0.5	0.504 ± 0.041	0.262	0.695 ± 0.040	0.275

Each case is based on 100 to 500 samples of size n.

of each estimate is also given. These quantities are calculated by letting $\hat{\theta}_1^{(i)}$, i = 1, ..., n_1 be the estimator of θ_1 obtained by using the ith sample of size n and

$$\bar{\hat{\theta}} = \text{Mean} = \frac{1}{n_1} \sum_{i=1}^{n_1} \hat{\theta}_1^{(i)}$$

where n_1 is the number of times a particular experiment is repeated. Then if

$$S = \left\{ \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (\hat{\theta}_1^{(i)} - \bar{\hat{\theta}})^2 \right\}^{\frac{1}{2}},$$

Table 2.2

Estimator of the mixing proportion in various mixtures of two normal distributions

				Estimate of θ_1							
				m = 2		m = 10		m = 20		m = 80	
n	$(\mu_2-\mu_1)/\sigma_1$	σ_2/σ_1	θ_1	Mean	MSE	Mean	MSE	Mean	MSE	Mean	MSE
50	0.25	1	0.5	0.419 ± 0.052	0.283	0.419 ± 0.052	0.282	0.420 ± 0.052	0.282	0.422 ± 0.052	0.278
10	0.5	1	0.5	0.456 ± 0.029	0.417	0.456 ± 0.028	0.416	0.456 ± 0.028	0.415	0.457 ± 0.028	0.415
10	1	1	0.5	0.474 ± 0.015	0.121	0.474 ± 0.015	0.120	0.475 ± 0.015	0.120	0.475 ± 0.015	0.120
10	1	1	0.8	0.789 ± 0.015	0.115	0.789 ± 0.015	0.115	0.789 ± 0.015	0.114	0.789 ± 0.015	0.114
20	5	1	0.5	0.489 ± 0.007	0.014	0.489 ± 0.007	0.014	0.489 ± 0.007	0.014	0.489 ± 0.007	0.014
10	0	2	0.5	0.476 ± 0.027	0.414	0.493 ± 0.020	0.198	0.505 ± 0.019	0.198	0.501 ± 0.019	0.181
50	0	2	0.5	0.520 ± 0.068	0.470	0.494 ± 0.021	0.044	0.494 ± 0.021	0.044	0.499 ± 0.020	0.049
10	0.5	2	0.5	0.443 ± 0.031	0.483	0.443 ± 0.027	0.372	0.445 ± 0.026	0.359	0.469 ± 0.025	0.311

Each case is based on $n_1 = \frac{5000}{n}$ samples of size n and a standard error is given for each mean to indicate the precision of the Monte Carlo computation.

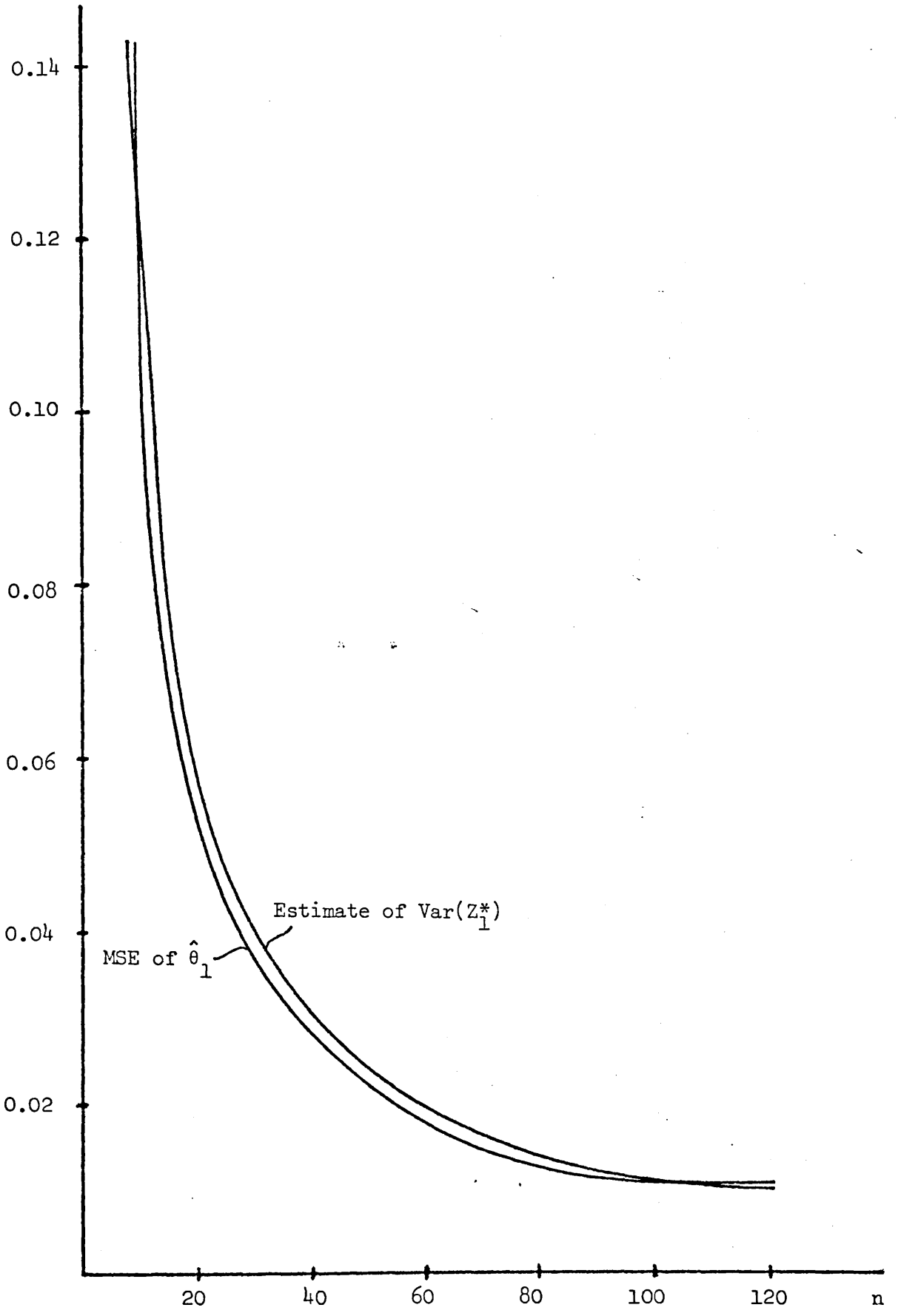


Fig. 2.1 - Mean-Square-Error and Estimate of variance of the estimator of θ_1, the mixing proportion in a mixture of two normal distributions N(0,1) and N(1,1), for varying n.

the standard error of $\bar{\hat{\theta}}$ is defined by $S/\sqrt{n_1}$ and

$$\text{Mean-Square-Error} = \left\{ \frac{1}{n_1} \sum_{i=1}^{n_1} (\hat{\theta}_1^{(i)} - \theta)^2 \right\} .$$

In our Monte Carlo studies, we took m to be an even integer and $t_1, \ldots, t_m$ were chosen as follows;

$$t_j = -\frac{(m/2) - (j-1)}{100} \qquad j = 1, \ldots, (m/2)$$

$$= \frac{j - (m/2)}{100} \qquad j = (m/2) + 1, \ldots, m$$

where it is clear that these choices are merely arbitrary and bear no optimal properties. Table 2.2 gives estimates of θ_1, corresponding to Table 2.1, using the method of moments, described in this chapter, for different values of m and for N = 5000.

Further, taking $\frac{\mu_2 - \mu_1}{\sigma_1} = 1$, $\frac{\sigma_2}{\sigma_1} = 1$ and $\theta = 0.5$, the mean-square-error of $\hat{\theta}_1^*$ and estimate of $\text{Var}_{\underset{\sim}{\theta}}(Z_1^*)$ were plotted against n when N = 5000 and m = 6. This is shown in Figure 2.1 where it is clearly seen that as n becomes larger than 20, both quantities $\text{Var}_{\underset{\sim}{\theta}}(Z_1^*)$ and mean-square-error of $\hat{\theta}_1^*$ decrease rapidly.

2.8 Conclusions

The asymptotic properties of the estimator of $\underset{\sim}{\theta}$ derived in this chapter, together with the Monte Carlo studies, indicate that the estimates (2.2.11) and (2.4.2) are reliable and the method of estimation works well in practice for moderate sample sizes. The first feature of the method is its simplicity. It provides unbiased CAN estimates in a very simple manner. However its rather general nature indicates that we cannot hope for very efficient estimates in small samples.

The efficiency will, however, be improved if the generalized sum of squares

$$(\underset{\sim}{\lambda}_n - A\underset{\sim}{\theta})' \, C^{-1} \, (\underset{\sim}{\lambda}_n - A\underset{\sim}{\theta}) \qquad (2.8.1)$$

is minimized with respect to $\underset{\sim}{\theta}$. Indeed the value of $\underset{\sim}{\theta}$ minimizing (2.8.1) is, by the Gauss Markoff theorem, the minimum variance linear unbiased estimator of $\underset{\sim}{\theta}$ for a given set of values $t_1, \ldots, t_m$. But minimizing (2.8.1) requires full knowledge (up to a constant multiple) of C which is not available, since the elements $c_{\underset{\sim}{\theta}}(r,s)$ $r,s = 1, \ldots, m$ of C depend on the unknown vector $\underset{\sim}{\theta}$. Thus some problems are caused which will be discussed in the next chapter.

CHAPTER 3

LEAST SQUARES ESTIMATION

3.1 Introduction

It was indicated in the last chapter that to increase the efficiency of the estimate of $\underset{\sim}{\theta}$, the vector of the mixing proportions in the mixture of distributions $G_{\underset{\sim}{\theta}}(.)$ given by (2.2.1), the generalized sum of squares given by (2.8.1) should be minimized. But since the covariance matrix C with $c_{\underset{\sim}{\theta}}(r,s)$, given by (2.3.3), as its (r,s)th element depends on $\underset{\sim}{\theta}$, the minimization procedure becomes more complicated than usual. In practice, however, it is suggested by some authors that if the covariance matrix is not available, one should use an estimate of it. Rao [44] showed that by merely substituting an estimate of the covariance matrix in the least squares estimate, the optimal properties are not necessarily preserved and improvements, depending on the structure of the covariance matrix, can be made. Gleser and Olkin [23] discuss the problem of maximum likelihood estimation of the covariance matrix in a linear model when the residual error vector has a multivariate normal distribution.

In this chapter, we consider a special case of the method of estimation suggested in the last chapter which is of particular interest in itself. We define

$$h(x,t) = 1 \qquad t \in \mathfrak{X}, \quad x \leq t$$

$$= 0 \qquad \text{otherwise}$$

so that the $\lambda_{\underset{\sim}{e}_j}(t)$ $j = 1,\ldots,k$ given by (2.2.3) become identical to the component distribution functions $F_j(t)$ for $t \in \mathfrak{X}$ and also $\lambda_n(t)$ given by (2.2.5) becomes the empirical distribution function $G_n(t)$ for $t \in \mathfrak{X}$. We shall see that in this case the elements $c_{\underset{\sim}{\theta}}(r,s)$; $r,s = 1,\ldots,m$ of the covariance matrix C become quadratic functions of $\underset{\sim}{\theta}$. Using this property, C is inverted and the generalized Least Squares (GLS) estimation of $\underset{\sim}{\theta}$ is discussed.

After the explanation of the GLS estimation of $\underset{\sim}{\theta}$ in section 3.2, the properties of the estimate are investigated in section 3.3. In section 3.4, the GLS estimator is derived for the class of mixtures of two rectangular distributions and section 3.5 compares the method of estimation discussed in this chapter with that of the previous chapter in the light of some Monte Carlo studies. Section 3.7 is devoted to the discussion of the GLS estimator of $\underset{\sim}{\theta}$ from ungrouped data.

3.2 Method of Estimation

Define $h(x, t) = 1 \quad t \in \mathcal{X}$ and $x \leq t$

$= 0$ otherwise ,

then from the definition of $\lambda_{\underset{\sim}{e}_j}(t)$ $j = 1,\ldots,k$ given in (2.2.3), we have

$$\lambda_{\underset{\sim}{e}_j}(t) = E_{\underset{\sim}{e}_j}(h(X,t)) = P_{\underset{\sim}{e}_j}(X \leq t) = F_j(t) \quad t \in \mathcal{X} \quad j = 1,\ldots,k$$

where $P_{\underset{\sim}{e}_j}$ is the probability measure corresponding to the distribution function F_j for $j = 1,\ldots,k$. Similarly from (2.2.2), $\lambda_{\underset{\sim}{\theta}}(t) = G_{\underset{\sim}{\theta}}(t)$ and further from (2.2.5),

$$\lambda_n(t) = \frac{1}{n}\sum_{i=1}^{n} h(x_i,t) = \frac{1}{n}(\text{no. of } x_i \leq t) = G_n(t) \quad t \in \mathcal{X}$$

where x_i is the realization of the random variable X_i with $G_{\underset{\sim}{\theta}}(.)$ as its distribution function. Thus we can write (2.2.8) as

$$G_n(x) = \sum_{j=1}^{k} \theta_j F_j(x) + \varepsilon(x) \qquad x \in \mathcal{X} \quad . \tag{3.2.1}$$

Choose distinct values $t_1,\ldots,t_m$; $t_j \in \mathcal{X}$ for $j = 1,\ldots,m$ and $m \geq k$ such that the rank of the matrix A in (2.2.9) which now becomes

$$A = \begin{bmatrix} F_1(t_1) & \dots & F_k(t_1) \\ F_1(t_2) & \dots & F_k(t_2) \\ \vdots & & \vdots \\ F_1(t_m) & \dots & F_k(t_m) \end{bmatrix} \tag{3.2.2}$$

is exactly k. For this condition to hold, it suffices to ensure that $t_1,\dots,t_m$ are chosen so that for each pair of distribution functions F_i and F_j $i,j = 1,\dots,k$ $i \neq j$, there is at least one value t_r; $1 \leq r \leq m$ such that $F_i(t_r) \neq F_j(t_r)$. Since the choice of the values $t_1,\dots,t_m$, $m \geq k$ is not in general unique, the term "partition" or "grouping" will be used to refer to any particular choice of these values.

Evaluating (3.2.1) at $t_1,\dots,t_m$, we obtain the linear model

$$\underset{\sim}{G}_n = A\underset{\sim}{\theta} + \underset{\sim}{\varepsilon} \tag{3.2.3}$$

where A is given by (3.2.2), $\underset{\sim}{G}_n = (G_n(t_1), \dots, G_n(t_m))'$ and as before $\underset{\sim}{\varepsilon} = (\varepsilon(t_1), \dots, \varepsilon(t_m))'$ and $\underset{\sim}{\theta} = (\theta_1,\dots,\theta_k)'$. Note that $A\underset{\sim}{\theta} = \underset{\sim}{G}_\theta = (G_\theta(t_1),\dots,G_\theta(t_m))'$.

The generalized least squares (GLS) estimator of $\underset{\sim}{\theta}$ is defined as that value of $\underset{\sim}{\theta}$ minimizing the generalized sum of squares (GSS) given by

$$\Phi(\underset{\sim}{\theta}) = \underset{\sim}{\varepsilon}'C^{-1}\underset{\sim}{\varepsilon} = (\underset{\sim}{G}_n - A\underset{\sim}{\theta})' \, C^{-1} \, (G_n - A\underset{\sim}{\theta}) \tag{3.2.4}$$

and thus it is clear that the covariance matrix C has to be inverted so that (3.2.4) can be minimized. But note that the (r,s)th element of C is

$$c_{\underset{\sim}{\theta}}(r,s) = \mathrm{Cov}_{\underset{\sim}{\theta}}(h(X,t_r), h(X,t_s)) \qquad r,s = 1,\dots,m$$

$$= E_{\underset{\sim}{\theta}}(h(X,t_r) \cdot h(X,t_s)) - E_{\underset{\sim}{\theta}}(h(X,t_r)) \, E_{\underset{\sim}{\theta}}(h(X,t_s))$$

$$= P_{\underset{\sim}{\theta}}(X \leq \min(t_r,t_s)) - P_{\underset{\sim}{\theta}}(X \leq t_r)\, P_{\underset{\sim}{\theta}}(X \leq t_s)$$

$$= G_{\underset{\sim}{\theta}}(\min(t_r,t_s)) - G_{\underset{\sim}{\theta}}(t_r)\, G_{\underset{\sim}{\theta}}(t_s) \qquad (3.2.5)$$

where $\min(x,y) = \begin{cases} x & x \leq y \\ y & y \leq x \end{cases}$. Therefore, we can see that since $G_{\underset{\sim}{\theta}}(.)$ is a linear function of $\underset{\sim}{\theta}$, $c_{\underset{\sim}{\theta}}(r,s)$ is a quadratic in $\underset{\sim}{\theta}$. We now discuss the inversion and some of the interesting properties of the covariance matrix C.

Without loss of generality, we can assume

$$t_1 < t_2 < \ldots < t_m ,$$

and hereafter we suppose that $t_1,\ldots,t_m$ are chosen in such a way that $F_j(.)\,;\, 1 \leq j \leq k$ attains two distinct values when evaluated at t_r and t_s $r \neq s$, $r,s = 1,\ldots,m$, i.e.

$$F_j(t_r) \neq F_j(t_s) \quad r \neq s \quad r,s = 1,\ldots,m \quad j = 1,\ldots,k .$$

Then by the monotonicity of $F_j(.)$, we have

$$F_j(t_1) < F_j(t_2) < \ldots < F_j(t_m) \quad j = 1,\ldots,k \qquad (3.2.6)$$

and consequently,

$$G_{\underset{\sim}{\theta}}(t_1) < G_{\underset{\sim}{\theta}}(t_2) < \ldots < G_{\underset{\sim}{\theta}}(t_m) . \qquad (3.2.7)$$

Thus from (3.2.5) ,

$$c_{\underset{\sim}{\theta}}(r,s) = \min(G_{\underset{\sim}{\theta}}(t_r),\, G_{\underset{\sim}{\theta}}(t_s)) - G_{\underset{\sim}{\theta}}(t_r)\, G_{\underset{\sim}{\theta}}(t_s) \qquad (3.2.8)$$

and so the matrix C may be written as

$$C = B - D \qquad (3.2.9)$$

where

$$B = \begin{bmatrix} \underset{\sim}{G}_\theta(t_1) & \underset{\sim}{G}_\theta(t_1) & \underset{\sim}{G}_\theta(t_1) & \cdots & \underset{\sim}{G}_\theta(t_1) & \underset{\sim}{G}_\theta(t_1) \\ \underset{\sim}{G}_\theta(t_1) & \underset{\sim}{G}_\theta(t_2) & \underset{\sim}{G}_\theta(t_2) & \cdots & \underset{\sim}{G}_\theta(t_2) & \underset{\sim}{G}_\theta(t_2) \\ \underset{\sim}{G}_\theta(t_1) & \underset{\sim}{G}_\theta(t_2) & \underset{\sim}{G}_\theta(t_3) & \cdots & \underset{\sim}{G}_\theta(t_3) & \underset{\sim}{G}_\theta(t_3) \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ \underset{\sim}{G}_\theta(t_1) & \underset{\sim}{G}_\theta(t_2) & \underset{\sim}{G}_\theta(t_3) & \cdots & \underset{\sim}{G}_\theta(t_{m-1}) & \underset{\sim}{G}_\theta(t_{m-1}) \\ \underset{\sim}{G}_\theta(t_1) & \underset{\sim}{G}_\theta(t_2) & \underset{\sim}{G}_\theta(t_3) & \cdots & \underset{\sim}{G}_\theta(t_{m-1}) & \underset{\sim}{G}_\theta(t_m) \end{bmatrix} \qquad (3.2.10)$$

and $D = \underset{\sim}{G}_\theta \cdot \underset{\sim}{G}'_\theta$

Lemma 3.2.1: If $\underset{\sim}{G}_\theta(t_1) > 0$, then

$$B^{-1} = \begin{bmatrix} \rho_1+\rho_2 & -\rho_2 & & & & \\ -\rho_2 & \rho_2+\rho_3 & -\rho_3 & & & \\ & -\rho_3 & \rho_3+\rho_4 & -\rho_4 & & \\ & & \ddots & \ddots & \ddots & \\ & & -\rho_{m-1} & \rho_{m-1}+\rho_m & -\rho_m & \\ & & & -\rho_m & \rho_m & \end{bmatrix} \qquad (3.2.11)$$

where $\rho_1 = [\underset{\sim}{G}_\theta(t_1)]^{-1}$ and $\rho_r = [\underset{\sim}{G}_\theta(t_r) - \underset{\sim}{G}_\theta(t_{r-1})]^{-1}$

$r = 2, \ldots, m$ (3.2.12)

and therefore by (3.2.7) $\rho_j > 0 \quad j = 1,\ldots,m$.

Proof: If $\underset{\sim}{G}_\theta(t_1) > 0$, then by (3.2.7) the rank of B is exactly m and thus B^{-1} exists. Define an m×m matrix R as follows

$$R = \begin{bmatrix} 1 & & & \\ -1 & 1 & & \\ & -1 & 1 & \\ & & \ddots & \ddots \\ & & & -1 \quad 1 \end{bmatrix} .$$

Note that the rank of R is m and R^{-1} exists. Consider

$$Q = RBR' = \begin{bmatrix} \underset{\sim}{G}_\theta(t_1) & & & \\ & \underset{\sim}{G}_\theta(t_2)-\underset{\sim}{G}_\theta(t_1) & & \\ & & \ddots & \\ & & & \underset{\sim}{G}_\theta(t_m)-\underset{\sim}{G}_\theta(t_{m-1}) \end{bmatrix}$$

i.e. Q is a diagonal matrix with $\underset{\sim}{G}_\theta(t_1)$ and $\underset{\sim}{G}_\theta(t_r) - \underset{\sim}{G}_\theta(t_{r-1})$ $r = 2,\ldots,m$ as its diagonal elements. Therefore we have $Q^{-1} = R'^{-1} B^{-1} R^{-1}$ and so $B^{-1} = R'Q^{-1} R$. But

$$Q^{-1} = \begin{bmatrix} \rho_1 & & & \\ & \rho_2 & & \\ & & \ddots & \\ & & & \rho_m \end{bmatrix}$$

with $\rho_1,\ldots,\rho_m$ given by (3.2.12). Premultiplying Q^{-1} by R' and post-multiplying by R, the result follows immediately.

Theorem 3.2.1: If $\underset{\sim}{G}_\theta(t_1) > 0$ and $\underset{\sim}{G}_\theta(t_m) < 1$, then the covariance matrix C possesses an inverse of the form

$$C^{-1} = B^{-1} + \frac{1}{1 - \underset{\sim}{G}_\theta(t_m)} L \qquad (3.2.13)$$

where L is an m×m matrix with zero in all the entries except the last element of the last column which is unity.

Proof: From (3.2.9), we can write

$$C = B\,[I - B^{-1}D]$$

so that

$$C^{-1} = [I - B^{-1} D]^{-1} B^{-1} \quad . \qquad (3.2.14)$$

Let $$H = B^{-1}D = B^{-1}(\underset{\sim}{G}_\theta . \underset{\sim}{G}'_\theta) = (B^{-1}\underset{\sim}{G}_\theta).\underset{\sim}{G}'_\theta$$

and observe that the elements of the last column of B are identical to the elements of $\underset{\sim}{G}_\theta$. Thus $B^{-1}\underset{\sim}{G}_\theta$ gives the last column of the identity matrix and hence

$$H = \begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 0 \\ \underset{\sim}{G}_\theta(t_1) & \underset{\sim}{G}_\theta(t_2) & \dots & \underset{\sim}{G}_\theta(t_m) \end{bmatrix} .$$

The m eigenvalues of H are the roots of the determinental equation $\det(\xi I - H) = 0$ i.e. $\xi^{m-1}(\xi - \underset{\sim}{G}_\theta(t_m)) = 0$. Thus the eigenvalues of H are $\xi_1 = \xi_2 = \dots = \xi_{m-1} = 0$ and $\xi_m = \underset{\sim}{G}_\theta(t_m) < 1$ i.e. all the eigenvalues of H have moduli < 1. Therefore

$$(I - H)^{-1} = I + H + H^2 + H^3 + \dots \quad .$$

It is easy to see that the higher powers of H are given by

$$H^r = [\underset{\sim}{G}_\theta(t_m)]^{r-1} H \qquad r = 1,2,\dots$$

so that

$$(I-H)^{-1} = I + H + (\underset{\sim}{G}_\theta(t_m))H + (\underset{\sim}{G}_\theta(t_m))^2 H + \dots$$

$$= I + \frac{1}{1 - \underset{\sim}{G}_\theta(t_m)} H$$

and substituting in (3.2.14),

$$C^{-1} = \left[I + \frac{1}{1 - \underset{\sim}{G}_\theta(t_m)} H \right] B^{-1} \quad . \qquad (3.2.15)$$

Finally, let $L = HB^{-1} = B^{-1} DB^{-1} = B^{-1} (\underset{\sim}{G}_\theta \cdot \underset{\sim}{G}'_\theta)\, B^{-1}$

i.e., $L = (B^{-1}\underset{\sim}{G}_\theta)(B^{-1}\underset{\sim}{G}_\theta)'$ since B^{-1} is a symmetric matrix. Further, since $B^{-1}\underset{\sim}{G}_\theta$ is the last column of the identity matrix, we have

$$HB^{-1} = L = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

and therefore from (3.2.15) we have

$$C^{-1} = \left[B^{-1} + \frac{1}{1 - G_\theta(t_m)}\, L\right] \qquad (3.2.16)$$

which is the required result.

Having obtained the inverse of C, we now substitute (3.2.16) into (3.2.4) to obtain the GSS. Thus

$$\Phi(\theta) = (\underset{\sim}{G}_n - \underset{\sim}{G}_\theta)'\, B^{-1}\, (\underset{\sim}{G}_n - \underset{\sim}{G}_\theta) + \frac{(G_n(t_m) - G_\theta(t_m))^2}{1 - G_\theta(t_m)}$$

and by substituting for B^{-1} from (3.2.11), after some algebraic manipulations, we get

$$\Phi(\theta) = \sum_{r=1}^{m+1} \frac{\left[\left(G_n(t_r) - G_n(t_{r-1})\right) - \left(G_\theta(t_r) - G_\theta(t_{r-1})\right)\right]^2}{G_\theta(t_r) - G_\theta(t_{r-1})} \qquad (3.2.17)$$

where t_o is very small (possibly $-\infty$) so that $F_j(t_o) = G_n(t_o) = 0$ for $j = 1,\ldots,k$ and t_{m+1} is very large (possibly $+\infty$) so that $F_j(t_{m+1}) = G_n(t_{m+1}) = 1$ for $j = 1,\ldots,k$. Therefore $G_\theta(t_o) = \sum_{j=1}^{k} \theta_j F_j(t_o) = 0$ and $G_\theta(t_{m+1}) = \sum_{j=1}^{k} \theta_j F_j(t_{m+1}) = \sum_{j=1}^{k} \theta_j = 1$. Note that also

$G_n(t_r) - G_n(t_{r-1}) \geq 0$ since $t_r > t_{r-1}$ for $r = 1,\ldots,m+1$.

Hence the GLS estimator of $\underset{\sim}{\theta}$ can be obtained by minimizing (3.2.17) with respect to $\underset{\sim}{\theta}$.

3.3 Properties of the GLS Estimator

Let Θ denote the k-fold Product space

$$\Theta = [0,1] \times [0,1] \times \ldots \times [0,1] \quad .$$

Then it is clear that $\Phi(\underset{\sim}{\theta})$ is a continuous function of $\underset{\sim}{\theta} \in \Theta$.

Theorem 3.3.1: $\Phi(\underset{\sim}{\theta})$ is a convex function of $\underset{\sim}{\theta} \in \Theta$.

Proof: It is known (Hardy, Littlewood and Pólya [24]) that a necessary and sufficient condition that $\Phi(\underset{\sim}{\theta})$ should be convex in Θ is that

(i) the second partial derivatives $\frac{\partial^2 \Phi}{\partial \theta_i \, \partial \theta_j}$ exists for $i,j = 1,\ldots,k$.

(ii) The matrix of the second partial derivatives is positive semidefinite.

Differentiating (3.2.17) with respect to θ_i and θ_j where $1 \leq i, j \leq k$, it is easy to see that

$$\frac{1}{2} \frac{\partial^2 \Phi}{\partial \theta_i \partial \theta_j} = \sum_{r=1}^{m+1} \frac{\left(F_i(t_r)-F_i(t_{r-1})\right)\left(G_n(t_r)-G_n(t_{r-1})\right)^2\left(F_j(t_r)-F_j(t_{r-1})\right)}{\left(G_{\underset{\sim}{\theta}}(t_r) - G_{\underset{\sim}{\theta}}(t_{r-1})\right)^3} \tag{3.3.1}$$

which exists for all $i,j = 1,\ldots,k$ and is finite for every $\underset{\sim}{\theta} \in \Theta$.

Denote by S a $k \times k$ matrix whose (i,j)th element is given by (3.3.1). Then for any $\underset{\sim}{b} = (b_1,\ldots,b_k)' \neq \underset{\sim}{0}$,

$$\underset{\sim}{b}' S \, \underset{\sim}{b} = \sum_{j=1}^{k} \sum_{r=1}^{m+1} \frac{b_j^2 \left(F_j(t_r) - F_j(t_{r-1})\right)^2 \left(G_n(t_r)-G_n(t_{r-1})\right)^2}{\left(G_{\underset{\sim}{\theta}}(t_r) - G_{\underset{\sim}{\theta}}(t_{r-1})\right)^3}$$

$$+ \sum_{\substack{i,j=1 \\ i\neq j}}^{k} \sum_{r=1}^{m+1} \frac{b_i \{F_i(t_r)-F_i(t_{r-1})\}\{G_n(t_r)-G_n(t_{r-1})\}^2\{F_j(t_r)-F_j(t_{r-1})\}b_j}{\{G_{\underset{\sim}{\theta}}(t_r) - G_{\underset{\sim}{\theta}}(t_{r-1})\}^3}$$

$$= \sum_{r=1}^{m+1} \left\{ \frac{\{G_n(t_r) - G_n(t_{r-1})\}^2}{\{G_{\underset{\sim}{\theta}}(t_r) - G_{\underset{\sim}{\theta}}(t_{r-1})\}^3} \left\{ \sum_{j=1}^{k} b_j \{F_j(t_r) - F_j(t_{r-1})\} \right\}^2 \right\} . \tag{3.3.2}$$

which by using (3.2.7) shows that $\underset{\sim}{b}' S \underset{\sim}{b} \geq 0$, proving the positive semi-definiteness of S. Hence the theorem is proved.

Corollary 3.3.1: If $t_1,\ldots,t_m$ are chosen so that $G_n(t_r) \neq G_n(t_{r-1})$ for every $1 \leq r \leq m+1$, i.e. if each of the intervals $\Delta_r = t_r - t_{r-1}$ $r = 1,\ldots,m+1$ contains at least one observation, then $\Phi(\underset{\sim}{\theta})$ is a strictly convex function of $\underset{\sim}{\theta} \in \Theta$.

Proof: $\Phi(\underset{\sim}{\theta})$ becomes a strictly convex function of $\underset{\sim}{\theta} \in \Theta$ whenever (3.3.2) is strictly positive, i.e. when S, the matrix of the second partial derivatives of $\Phi(\underset{\sim}{\theta})$, is positive definite. Thus assuming $G_n(t_r) \neq G_n(t_{r-1})$ for every $r = 1,\ldots,m+1$, it suffices to show that

$$\sum_{j=1}^{k} b_j (F_j(t_r) - F_j(t_{r-1})) \neq 0$$

for at least one $1 \leq r \leq m+1$. Suppose on the contrary that

$$\sum_{j=1}^{k} b_j (F_j(t_r) - F_j(t_{r-1})) = 0$$

for every $r = 1,\ldots,m+1$. Then

$$\sum_{j=1}^{k} b_j = \sum_{j=1}^{k} b_j F_j(t_m) = \sum_{j=1}^{k} b_j F_j(t_{m-1}) = \ldots = \sum_{j=1}^{k} b_j F_j(t_1)$$

$$= \sum_{j=1}^{k} b_j F_j(t_o) = 0 .$$

But since the rank of the matrix A given by (3.2.2) is exactly k, $\sum_{j=1}^{k} b_j F_j(t_r) = 0$ for $r = 1,\ldots,m+1$ if and only if $b_1 = b_2 = \ldots = b_k = 0$, contradicting the fact that $\underset{\sim}{b} = (b_1,\ldots,b_k)' \neq \underset{\sim}{0}$. Hence there is at least one r, $1 \leq r \leq m+1$, such that

$$\sum_{j=1}^{k} b_j \left(F_j(t_r) - F_j(t_{r-1})\right) \neq 0$$

proving the positive definiteness of S.

We conclude therefore, from the standard properties of convex functions, that if $\Phi(\underset{\sim}{\theta})$ has any stationary points then they must be minimum points. Further any local minimum of $\Phi(\underset{\sim}{\theta})$ is an absolute minimum for if $\Phi(\underset{\sim}{\theta})$ has a local minimum at $\underset{\sim}{\theta}_o \in \Theta$, then given any $\underset{\sim}{\theta} \in \Theta$, $\Phi(\underset{\sim}{\theta}_o) \leq \Phi\left((1-\alpha)\underset{\sim}{\theta}_o + \alpha\underset{\sim}{\theta}\right)$ for sufficiently small $0 < \alpha < 1$. By using the convexity of $\Phi(\underset{\sim}{\theta})$, $\Phi(\underset{\sim}{\theta}_o) \leq (1-\alpha)\,\Phi(\underset{\sim}{\theta}_o) + \alpha\,\Phi(\underset{\sim}{\theta})$ for all $\underset{\sim}{\theta} \in \Theta$. Thus $\Phi(\underset{\sim}{\theta}_o) \leq \Phi(\underset{\sim}{\theta})$ for all $\underset{\sim}{\theta} \in \Theta$ showing that θ_o is an absolute minimum point.

It is also worth noting that the condition of the corollary 3.3.1 for $\Phi(\underset{\sim}{\theta})$ becoming a strictly convex function of $\underset{\sim}{\theta} \in \Theta$ is sufficient and not in general necessary. However if $\Phi(\underset{\sim}{\theta})$ is a strictly convex function of $\underset{\sim}{\theta} \in \Theta$, then $\Phi(\underset{\sim}{\theta})$ has at most one unique minimum in Θ. This is easy to see for if there are two points $\underset{\sim}{\theta}_o$ and $\underset{\sim}{\theta}^*_o$ in Θ at which the derivative with respect to $\underset{\sim}{\theta}$ of $\Phi(\underset{\sim}{\theta})$ vanishes, then $\underset{\sim}{\theta}_o$ and $\underset{\sim}{\theta}^*_o$ are both absolute minimum points and thus $\Phi(\underset{\sim}{\theta}_o) < \Phi(\alpha\,\underset{\sim}{\theta}_o + (1-\alpha)\,\underset{\sim}{\theta}^*_o)$ and $\Phi(\underset{\sim}{\theta}^*_o) < \Phi(\alpha\,\underset{\sim}{\theta}_o + (1-\alpha)\underset{\sim}{\theta}^*_o)$ for every $0 < \alpha < 1$. Now

$$\alpha\,\Phi(\underset{\sim}{\theta}_o) + (1-\alpha)\Phi(\underset{\sim}{\theta}^*_o) < \alpha\,\Phi(\alpha\,\underset{\sim}{\theta}_o + (1-\alpha)\underset{\sim}{\theta}^*_o) + (1-\alpha)\,\Phi(\alpha\,\theta_o + (1-\alpha)\theta^*_o)$$

$$= \Phi(\alpha\,\underset{\sim}{\theta}_o + (1-\alpha)\,\underset{\sim}{\theta}^*_o)$$

for every $0 < \alpha < 1$. This contradicts the strict convexity of $\Phi(\underset{\sim}{\theta})$.

The asymptotic properties of the GLS estimator, obtained by

minimizing $\phi(\underset{\sim}{\theta})$ in (3.2.17), can readily be established. Let $\pi_i(\underset{\sim}{\theta})$ be the probability of having an observation in the interval $\Delta_i = t_i - t_{i-1}$; $i = 1,\ldots,m+1$ and let n_i be the corresponding group frequencies. Note that Δ_i carries the mass $p_i = \frac{n_i}{n}$ in the sample distribution and the mass $\pi_i(\underset{\sim}{\theta})$ in the hypothetical distribution. Then it is immediately seen from (3.2.17) that

$$\Phi(\underset{\sim}{\theta}) = \sum_{i=1}^{m+1} \frac{(\pi_i(\underset{\sim}{\theta}) - p_i)^2}{\pi_i(\underset{\sim}{\theta})} = \sum_{i=1}^{m+1} \frac{p_i^2}{\pi_i(\underset{\sim}{\theta})} - 1 \qquad (3.3.3)$$

and it is well-known that $n\left\{\sum_{i=1}^{m+1} \frac{p_i^2}{\pi_i(\underset{\sim}{\theta})} - 1\right\}$ has, for large n, a central χ^2 distribution with m degrees of freedom (Cramér [15]). Hence minimizing (3.3.3) constitutes the celebrated minimum χ^2 method of estimation. The estimating equations are then obtained, by differentiating (3.3.3) with respect to θ_j for $j = 1,\ldots,k$, to be

$$\sum_{i=1}^{m+1} \frac{p_i^2}{\pi_i^2(\underset{\sim}{\theta})} \frac{\partial \pi_i(\underset{\sim}{\theta})}{\partial \theta_j} = 0 \qquad j = 1,\ldots,k \; . \qquad (3.3.4)$$

Rao [43] has proved that the estimating equations (3.3.4) have the following properties;

(i) With probability approaching unity, there is a root $\hat{\underset{\sim}{\theta}}_n = (\hat{\theta}_1,\ldots,\hat{\theta}_k)'$ of the set of equations (3.3.4) with the property that if we denote by $\underset{\sim}{Z}_n = (Z_1,\ldots,Z_k)'$, the random vector whose realization is $\hat{\underset{\sim}{\theta}}_n$, then $\underset{\sim}{Z}_n$ converges in probability to the true parameter value $\underset{\sim}{\theta}^* = (\theta_1^*,\ldots,\theta_k^*)'$, i.e. $\hat{\underset{\sim}{\theta}}_n$ is a consistent estimator of $\underset{\sim}{\theta}^*$.

(ii) This consistent estimator is unique in the sense that if there is another root $\hat{\underset{\sim}{\theta}}_n^+ = (\hat{\theta}_1^+, \ldots,\hat{\theta}_k^+)' \neq \hat{\underset{\sim}{\theta}}_n$ of the set of equations (3.3.4) which is also a consistent estimator of $\underset{\sim}{\theta}^*$ and if $\hat{\underset{\sim}{\theta}}_n^+$ is the realization of the random vector $\underset{\sim}{Z}_n^+ = (Z_1^+,\ldots,Z_k^+)'$, then $\sqrt{n}\,(\underset{\sim}{Z}_n - \underset{\sim}{Z}_n^+) \to 0$ in probability as $n \to \infty$.

(iii) The distribution of the random vector $\sqrt{n}\,(\underset{\sim}{Z}_n - \underset{\sim}{\theta}^*)$ is asymptotically normal with mean vector $\underset{\sim}{0}$ and covariance matrix equal to the inverse of the Fisher's information matrix (Cramér-Rao lower bound).

(iv) With probability approaching unity, the absolute minimum of (3.3.3) is attained at a root of (3.3.4).

3.4 An Example

The estimating equations given by (3.3.4) cannot in general be solved analytically. However, for the class of mixtures of rectangular distributions analytic solutions may be obtained. In this section, we consider a mixture of two rectangular distributions and obtain the GLS estimator of $\underset{\sim}{\theta} = (\theta_1, \theta_2)'$, the vector of the mixing proportions.

Let $F_1(x)$ and $F_2(x)$ be the distribution functions of two rectangular distributions. Then the mixture of $F_1(x)$ and $F_2(x)$ with mixing proportions θ_1 and $\theta_2 = 1-\theta_1$ with $0 \leq \theta_1 \leq 1$ respectively, will depend on the scalar parameter θ_1 only and may be written as

$$G_{\underset{\sim}{\theta}}(x) = \theta_1 F_1(x) + (1-\theta_1) F_2(x) \qquad 0 \leq \theta_1 \leq 1$$

where we assume that the ranges of $F_1(x)$ and $F_2(x)$ are independent of θ_1. We further assert that, without loss of generality, one of the component distributions, say $F_1(x)$, can be assumed to be the distribution function of a random variable, uniformly distributed over the interval [0,1]. This condition can always be maintained by the use of the transformation $Y = F_1(X)$. That is if the distribution function of a random variable X is the mixture of $F_1(x)$ and $F_2(x)$ with ranges $[a_1,b_1]$ and $[a_2,b_2]$ respectively with a_j and b_j $j = 1,2$ being finite and independent of θ_1 so that

$$F_1(x) = \begin{cases} 0 & x \leq a_1 \\ \dfrac{x-a_1}{b_1-a_1} & a_1 \leq x \leq b_1 \\ 1 & x \geq b_1 \end{cases}$$

and

$$F_2(x) = \begin{cases} 0 & x \leq a_2 \\ \dfrac{x-a_2}{b_2-a_2} & a_2 \leq x \leq b_2 \\ 1 & x \geq b_2 \end{cases}$$

then the distribution function of the random variable $Y = F_1(X)$ is the mixture of two rectangular distribution functions with ranges $[0,1]$ and $\left[\dfrac{a_2-a_1}{b_1-a_1}, \dfrac{b_2-a_1}{b_1-a_1}\right]$ respectively.

Letting $a = \dfrac{a_2-a_1}{b_1-a_1}$ and $b = \dfrac{b_2-a_1}{b_1-a_1}$, the required condition is maintained.

So let

$$F_1(x) = \begin{cases} 0 & x \leq 0 \\ x & 0 \leq x \leq 1 \\ 1 & x \geq 1 \end{cases}$$

and

$$F_2(x) = \begin{cases} 0 & x \leq a \\ \dfrac{x-a}{b-a} & a \leq x \leq b \\ 1 & x \geq b \end{cases}$$

where a and b are finite and independent of θ_1. The following three distinct possible situations are considered.

(i) $b > a \geq 1$ (see Fig. (i))

then

$$G_{\underset{\sim}{\theta}}(x) = \begin{cases} 0 & x \leq 0 \\ \theta_1 x & 0 \leq x \leq 1 \\ \theta_1 & 1 \leq x \leq a \\ \theta_1 + (1-\theta_1)\frac{x-a}{b-a} & a \leq x \leq b \\ 1 & x \geq b \end{cases}$$

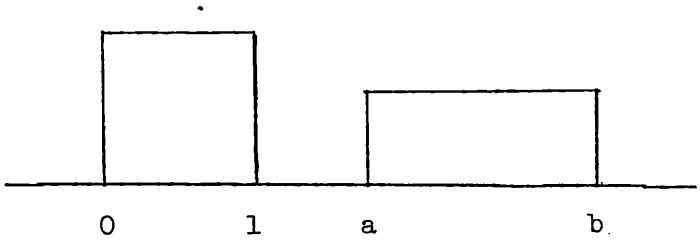


Fig. (i)

Choose $0 = t_o < t_1 < \ldots < t_\alpha = 1 \qquad 1 \leq \alpha \leq m$

and $a < t_{\alpha+1} < t_{\alpha+2} < \ldots < t_{m+1} = b$.

Note that no values of $\{t_i\}_{i=1}^{m}$ can be chosen in the interval $(1,a]$ since in such a case the condition (3.2.7) will not be satisfied. Also since the probability of having an observation in that interval is zero, the number of observations not exceeding 1 in magnitude is equal to the number of observations not exceeding a. Hence we have $G_n(1) = G_n(a)$

Now

$$\pi_{i\,\underset{\sim}{\theta}}(\theta) = G_{\underset{\sim}{\theta}}(t_i) - G_{\underset{\sim}{\theta}}(t_{i-1}) = \theta_1(t_i - t_{i-1}) \qquad i = 1,\ldots,\alpha$$

$$\pi_{\alpha+1\,\underset{\sim}{\theta}}(\theta) = G_{\underset{\sim}{\theta}}(t_{\alpha+1}) - G_{\underset{\sim}{\theta}}(t_\alpha) = \theta_1 + (1-\theta_1)\frac{t_{\alpha+1}-a}{b-a} - \theta_1$$

$$= (1-\theta_1)\frac{t_{\alpha+1}-a}{b-a}$$

$$\pi_i(\underset{\sim}{\theta}) = G_{\underset{\sim}{\theta}}(t_i) - G_{\underset{\sim}{\theta}}(t_{i-1}) = (1-\theta_1)\frac{t_i - t_{i-1}}{b-a}$$

$$i = \alpha+2, \ldots, m+1 .$$

Substituting the above values in (3.3.4), we get

$$\frac{1}{\hat{\theta}_1^2}\sum_{i=1}^{\alpha}\frac{p_i^2}{t_i - t_{i-1}} - \frac{b-a}{(1-\hat{\theta}_1)^2}\left\{\frac{p_{\alpha+1}^2}{t_{\alpha+1}-a} + \sum_{i=\alpha+2}^{m+1}\frac{p_i^2}{t_i - t_{i-1}}\right\} = 0 \qquad (3.4.1)$$

where $\hat{\theta}_1$ denotes the least squares estimate of θ_1, and θ_2 is estimated by $1-\hat{\theta}_1$. The solution of (3.4.1) is

$$\hat{\theta}_1 = \frac{\left[\sum_{i=1}^{\alpha}\frac{p_i^2}{t_i - t_{i-1}}\right]^{\frac{1}{2}}}{\left[\sum_{i=1}^{\alpha}\frac{p_i^2}{t_i - t_{i-1}}\right]^{1/2} + \left[(b-a)\left\{\frac{p_{\alpha+1}^2}{t_{\alpha+1}-a} + \sum_{i=\alpha+2}^{m+1}\frac{p_i^2}{t_i - t_{i-1}}\right\}\right]^{1/2}} \qquad (3.4.2)$$

In particular if we choose $\alpha = 1$, i.e. $t_o = 0$, $t_1 = 1$ and $t_2 = b$, we have

$$p_1 = G_n(1) - G_n(0) = G_n(1)$$

$$p_2 = G_n(b) - G_n(1) = G_n(b) - G_n(a)$$

so that (3.4.2) gives

$$\hat{\theta}_1 = \frac{G_n(1)}{G_n(1) + G_n(b) - G_n(a)} = \frac{G_n(1)}{G_n(b)} \qquad (3.4.3)$$

which we now show is the maximum likelihood estimator of θ_1.

The logarithm of the likelihood function is

$$L = n\, G_n(1)\, \ell n\, \theta_1 + n(G_n(b) - G_n(a))(\ell n(1-\theta_1) - \ell n\,(b-a))$$

and upon differentiating L with respect to θ_1 and solving $\frac{dL}{d\theta_1} = 0$,

we obtain the root

$$\hat{\theta}_1 = \frac{G_n(1)}{G_n(1) + G_n(b) - G_n(a)} = \frac{G_n(1)}{G_n(b)}$$

which is identical to (3.4.3). Hence the estimator (3.4.3) has all the well-known asymptotic properties of the maximum likelihood estimator (Cramér [15])

(ii) $a > 0 \quad b < 1$ (See Fig. (ii))

then

$$G_{\underset{\sim}{\theta}}(x) = \begin{cases} 0 & x \leq 0 \\ \theta_1 x & 0 \leq x \leq a \\ \theta_1 x + (1-\theta_1)\frac{x-a}{b-a} & a \leq x \leq b \\ \theta_1 x + (1-\theta_1) & b \leq x \leq 1 \\ 1 & x \geq 1 \end{cases}$$

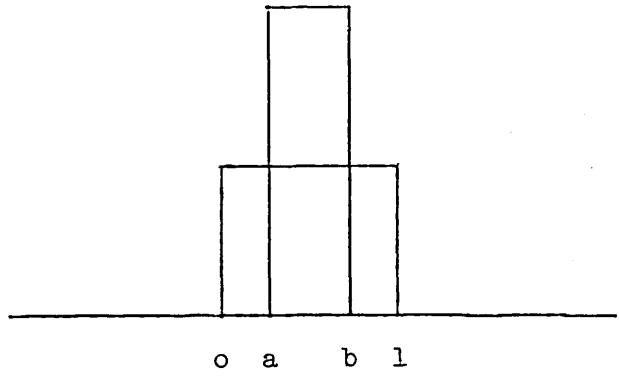


Fig. (ii)

Choose $0 = t_o < t_1 < \ldots < t_\alpha = a < t_{\alpha+1} < \ldots < t_\beta = b < t_{\beta+1} < \ldots < t_{m+1} = 1$

$0 < \alpha < \beta \leq m$

If either $a = 0$ or $b = 1$ (but not both), obvious adjustments are made in the choice of $\{t_i\}_{i=1}^{m}$.

We therefore have

$$\pi_i(\underset{\sim}{\theta}) = \theta_1 (t_i - t_{i-1}) = \theta_1 \Delta_i \qquad i = 1,\ldots,\alpha$$

$$= \theta_1 \Delta_i + (1-\theta_1)\frac{\Delta_i}{b-a} \qquad i = \alpha+1,\ldots,\beta$$

$$= \theta_1 \Delta_i \qquad i = \beta+1,\ldots,m+1 \ .$$

Substituting in (3.3.4), we get

$$\frac{1}{\hat{\theta}_1^2}\sum_{i=1}^{\alpha}\frac{p_i^2}{\Delta_i} - \frac{\left[\frac{1}{b-a} - 1\right]}{\left[\hat{\theta}_1 + \frac{1-\hat{\theta}_1}{b-a}\right]^2}\sum_{i=\alpha+1}^{\beta}\frac{p_i^2}{\Delta_i} + \frac{1}{\hat{\theta}_1^2}\sum_{i=\beta+1}^{m+1}\frac{p_i^2}{\Delta_i} = 0$$

giving the least squares estimator of θ_1 as

$$\hat{\theta}_1 = \frac{\left[\sum_{i=1}^{\alpha}\frac{p_i^2}{\Delta_i} + \sum_{i=\beta+1}^{m+1}\frac{p_i^2}{\Delta_i}\right]^{1/2}}{[1-(b-a)]\left[\sum_{i=1}^{\alpha}\frac{p_i^2}{\Delta_i} + \sum_{i=\beta+1}^{m+1}\frac{p_i^2}{\Delta_i}\right]^{1/2} + (b-a)[1-(b-a)]^{1/2}\left[\sum_{i=\alpha+1}^{\beta}\frac{p_i^2}{\Delta_i}\right]^{1/2}} \ .$$

In particular if $m = 2$, so that $t_o = 0$, $t_1 = a$, $t_2 = b$ and $t_3 = 1$, we get

$$\hat{\theta}_1 = \frac{\left[\frac{G_n^2(a)}{a} + \frac{(1-G_n(b))^2}{1-b}\right]^{1/2}}{[1-(b-a)]\left[\frac{G_n^2(a)}{a} + \frac{(1-G_n(b))^2}{1-b}\right]^{1/2} + [1-(b-a)]^{1/2}(G_n(b)-G_n(a))} \ . \qquad (3.4.4)$$

Let Z_n be a random variable whose realization $\hat{\theta}_1$ is given by (3.4.4). We prove that Z_n has the same limiting distribution as the maximum likelihood estimator of θ_1 as $n \to \infty$. Recall that $\Gamma_n(x)$, being the random function with realization $G_n(x)$, converges in probability to $G_{\underset{\sim}{\theta}}(x)$.

Now, the logarithm of the likelihood function of θ_1 is

$$L = n\,G_n(a)\,\ell n\,(\theta_1) + n(G_n(b) - G_n(a))\,\ell n\left[\theta_1 + \frac{1-\theta_1}{b-a}\right] + n(1-G_n(b))\,\ell n\,(\theta_1).$$

Differentiating with respect to θ_1,

$$\frac{dL}{d\theta_1} = \frac{n\, G_n(a)}{\theta_1} + n(G_n(b) - G_n(a)) \frac{1 - \frac{1}{b-a}}{\theta_1 + \frac{1-\theta_1}{b-a}} + \frac{n(1 - G_n(b))}{\theta_1} \qquad (3.4.5)$$

from which θ_1^+, the maximum likelihood estimator of θ_1, being the root of (3.4.5) is given by

$$\theta_1^+ = \frac{G_n(a)+1 - G_n(b)}{a+1-b} \quad .$$

Let Z_n^+ be a random variable with realization θ_1^+ and consider

$$Z_n - Z_n^+ = \frac{\left[\frac{\Gamma_n^2(a)}{a} + \frac{(1 - \Gamma_n(b))^2}{1-b}\right]^{1/2}}{(a+1-b)\left[\frac{\Gamma_n^2(a)}{a} + \frac{(1-\Gamma_n^2(b))^2}{1-b}\right]^{1/2} + (a+1-b)(\Gamma_n(b)-\Gamma_n(a))} - \frac{\Gamma_n(a) + 1 - \Gamma_n(b)}{a+1-b}$$

$$= \frac{(\Gamma_n(b)-\Gamma_n(a))\{(a+1-b)^{1/2}\left[\frac{\Gamma_n^2(a)}{a} + \frac{(1-\Gamma_n(b))^2}{1-b}\right]^{1/2} - (\Gamma_n(a)+1-\Gamma_n(b))\}}{(a+1-b)^{1/2}\{(a+1-b)\left[\frac{\Gamma_n^2(a)}{a} + \frac{(1-\Gamma_n(b))^2}{1-b}\right]^{1/2} + (a+1-b)^{1/2}(\Gamma_n(b) - \Gamma_n(a))\}}$$

Now, as $n \to \infty$ $\Gamma_n(a) \xrightarrow{p} G_\theta(a) = \theta_1 a$ and $\Gamma_n(b) \xrightarrow{p} G_\theta(b) = \theta_1 b + (1-\theta_1)$ and since $(Z_n - Z_n^+)$ is a continuous function of $\Gamma_n(a)$ and $\Gamma_n(b)$, it is seen that as $n \to \infty$, $Z_n - Z_n^+ \xrightarrow{p} 0$. Therefore Z_n and Z_n^+ have the same asymptotic distributions, i.e. Z_n is asymptotically normally distributed with mean θ_1 and variance equal to the Cramér-Rao lower bound which is easily found to be

$$\frac{\theta_1 [\theta_1(b-a) + (1-\theta_1)]}{n[1 - (b-a)]} \quad .$$

(iii) $\quad 0 < a < 1 \,, \quad b > 1 \qquad$ (see Fig. (iii))

Then

$$G_{\underset{\sim}{\theta}}(x) = \begin{cases} \theta_1 x & 0 \le x \le a \\ \theta_1 x + (1-\theta_1)\frac{x-a}{b-a} & a \le x \le 1 \\ \theta_1 + (1-\theta_1)\frac{x-a}{b-a} & 1 \le x \le b \\ 1 & x \ge b \end{cases}$$

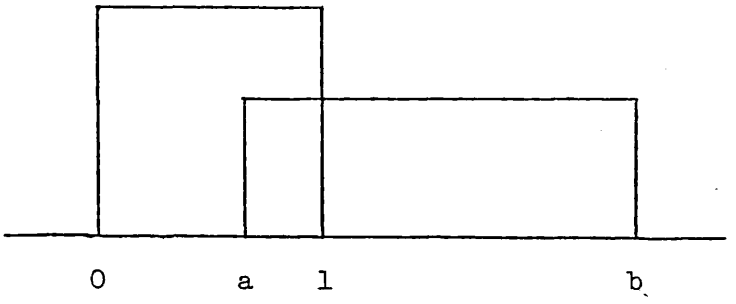


Fig. (iii)

Choose $0 = t_o < t_1 < \ldots < t_\alpha = a < t_{\alpha+1} < \ldots < t_\beta = 1 < t_{\beta+1} < \ldots < t_{m+1} = b \qquad 0 < \alpha < \beta \le m$

If $a = 0$, then we choose $\alpha = 0$.

We therefore have

$$\begin{aligned} \pi_{\underset{\sim}{i}}(\theta) &= \theta_1 \Delta_i && i = 1,\ldots,\alpha \\ &= \theta_1 \Delta_i + (1-\theta_1)\frac{\Delta_i}{b-a} && i = \alpha+1,\ldots,\beta \\ &= (1-\theta_1)\frac{\Delta_i}{b-a} && i = \beta+1,\ldots,m+1 \end{aligned}$$

and substituting in (3.3.4) we get

$$\frac{1}{\hat{\theta}_1^2}\sum_{i=1}^{\alpha}\frac{p_i^2}{\Delta_i} + \frac{\left[1-\frac{1}{b-a}\right]}{\left[\hat{\theta}_1 + \frac{1-\hat{\theta}_1}{b-a}\right]}\sum_{i=\alpha+1}^{\beta}\frac{p_i^2}{\Delta_i} - \frac{(b-a)}{(1-\hat{\theta}_1)^2}\sum_{i=\beta+1}^{m+1}\frac{p_i^2}{\Delta_i} = 0$$

resulting in the following quartic equation in $\hat{\theta}_1$, the estimate of θ_1;

$$c_o\hat{\theta}_1^4 + c_1\hat{\theta}_1^3 + c_2\hat{\theta}_1^2 + c_3\hat{\theta}_1 + c_4 = 0 \qquad (3.4.6).$$

where

$$c_o = (1-\omega)\left[(1-\omega)\, S_{1,\alpha} + S_{\alpha+1,\beta} - \frac{1-\omega}{\omega}\, S_{\beta+1,m+1}\right]$$

$$c_1 = -2(1-\omega)[(1-4\omega)\, S_{1,\alpha} + S_{\alpha+1,m+1}]$$

$$c_2 = [(1-\omega)^2 + \omega(5\omega-4)]\, S_{1,\alpha} + (1-\omega)\, S_{\alpha+1,\beta} - \omega S_{\beta+1,m+1}]$$

$$c_3 = 2\,\omega\,(1-2\omega)\, S_{1,\alpha}$$

$$c_4 = \omega^2\, S_{1,\alpha}$$

with $\omega = \frac{1}{b-a}$ and $S_{r,\ell} = \sum_{i=r}^{\ell} \frac{p_i^2}{\Delta_i}$ $\qquad 1 \le r < \ell \le m+1$.

It is known that the polynomials of degree up to and including 4 are solvable. A quartic is first reduced to a cubic and is then solved. We can also solve the quartic (3.4.6) by Ferrari's method based on dissecting the quartic into the difference of the squares of a quadratic and a linear function of $\hat{\theta}_1$ (Archbold [2]) .

Thus it is possible to obtain $\hat{\theta}_1$ from (3.4.5), but unfortunately the underlying solution becomes intractable. However with some numerical values, the roots can easily be found.

3.5 Monte Carlo Studies

In order to compare the method of estimation of $\underset{\sim}{\theta}$ suggested in this chapter with that of chapter 2, we consider a mixture of two normal distributions and derive the estimate of $\underset{\sim}{\theta} = (\theta_1, \theta_2)'$ where $\theta_2 = 1-\theta_1$ and $0 \le \theta_1 \le 1$.

We mentioned in section 2.7 that one of the component distributions can, without loss of generality, be assumed to be the distribution function of the standard normal distribution N(0,1), i.e. a normal

distribution with mean zero and variance unity. So let

$$G_{\underset{\sim}{\theta}}(x) = \theta_1 F_1(x) + (1-\theta_1) F_2(x) \qquad 0 \leq \theta_1 \leq 1 \qquad -\infty < x < \infty \qquad (3.5.1)$$

where

$$F_1(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\{-\tfrac{1}{2}y^2\}\, dy$$

and

$$F_2(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \int_{-\infty}^{x} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\} dy \ .$$

The values $t_1,\ldots,t_m$ can be chosen to be any m distinct ordered finite real values such that $t_1 < t_2 < \ldots < t_m$. This gives

$$F_j(t_1) < F_j(t_2) < \ldots < F_j(t_m) \qquad j = 1,2$$

and consequently

$$G_{\underset{\sim}{\theta}}(t_1) < G_{\underset{\sim}{\theta}}(t_2) < \ldots < G_{\underset{\sim}{\theta}}(t_m) \ .$$

Hence the conditions of the lemma 3.2.1 and the theorem 3.2.1 are satisfied and the covariance matrix given by (3.2.9) possesses an inverse. Note that

$$\pi_i(\underset{\sim}{\theta}) = G_{\underset{\sim}{\theta}}(t_i) - G_{\underset{\sim}{\theta}}(t_{i-1}) = \theta_1 \left(F_1(t_i) - F_1(t_{i-1})\right) + (1-\theta_1)\left(F_2(t_i) - F_2(t_{i-1})\right) \qquad (3.5.2)$$

with $t_o = -\infty$ and $t_{m+1} = +\infty$. So by differentiating (3.5.2) with respect to θ_1,

$$\frac{d\,\pi_i(\underset{\sim}{\theta})}{d\theta_1} = \left(F_1(t_i) - F_1(t_{i-1})\right) - \left(F_2(t_i) - F_2(t_{i-1})\right) \qquad (3.5.3)$$

and therefore by substituting (3.5.2) and (3.5.3) in (3.3.4), the

estimating equation is obtained as

$$\sum_{i=1}^{m+1} \frac{p_i^2(\beta_{1i} - \beta_{2i})}{\theta_1\beta_{1i}+(1-\theta_1)\beta_{2i}} = 0 \qquad (3.5.4)$$

where $\beta_{ji} = F_j(t_i) - F_j(t_{i-1})$ for $j = 1,2$ and for $i = 1,\ldots,m+1$. The root of (3.5.4) constitutes the GLS estimate of θ_1.

A sample of size n was generated from the mixture (3.5.1) by sampling with probability $0 \leq \theta_1 \leq 1$ from $N(0,1)$ and with probability $\theta_2 = 1-\theta_1$ from $N(\mu,\sigma^2)$. Choosing the real values $t_1 < t_2 < \ldots < t_m$, the root of (3.5.4) was obtained by using standard numerical techniques on the computer. The experiment was repeated n_1 times where $n.n_1 = N$ being a fixed number.

In our Monte Carlo studies, the values t_1 and t_m were chosen as

$$t_1 = \min (\mu_1 - 2\sigma_1, \ \mu_2 - 2\sigma_2) \qquad (3.5.5)$$

and

$$t_m = \max (\mu_1 + 2\sigma_1, \ \mu_2 + 2\sigma_2) \qquad (3.5.6)$$

and the distance $t_m - t_1$ was divided into (m-1) equal intervals. The values $t_2,\ldots,t_{m-1}$ were then chosen as the division points, so that

$$t_i = t_1 + (i-1)\frac{t_m - t_1}{m-1} \qquad i = 1,\ldots,m \ .$$

The choices of t_1 and t_m given by (3.5.5) and (3.5.6) respectively, seem reasonable since over 95% of the observations from each component of $G_{\underset{\sim}{\theta}}(.)$ fall in the interval $t_m - t_1$. Analogous to table 2.2, table 3.1 gives the GLS estimator of θ_1 for different values of m when N = 5000. The mean-square-error of each estimate and the standard error of each mean are calculated as explained in section 2.7. Comparing the two tables, it is seen that for large m, the estimate of θ_1 is improved.

To investigate the dependence of the GLS estimator of θ_1 upon the number of division points m, we picked the case when $\frac{\mu_2 - \mu_1}{\sigma_1} = 1$

Table 3.1

Generalized Least Squares estimator of the mixing proportion in various mixtures of two normal distributions

n	$(\mu_2-\mu_1)/\sigma_1$	σ_2/σ_1	θ_1	Estimate of θ_1							
				m = 3		m = 10		m = 20		m = 80	
				Mean	MSE	Mean	MSE	Mean	MSE	Mean	MSE
50	0.25	1	0.5	0.462 ± 0.066	0.431	0.504 ± 0.049	0.242	0.458 ± 0.046	0.212	0.475 ± 0.038	0.143
10	0.5	1	0.5	0.484 ± 0.038	0.648	0.489 ± 0.030	0.495	0.500 ± 0.030	0.466	0.500 ± 0.028	0.382
10	1	1	0.5	0.505 ± 0.019	0.190	0.521 ± 0.018	0.163	0.507 ± 0.017	0.152	0.492 ± 0.016	0.136
10	1	1	0.8	0.866 ± 0.021	0.222	0.847 ± 0.020	0.162	0.859 ± 0.017	0.131	0.864 ± 0.016	0.128
20	5	1	0.5	0.501 ± 0.007	0.013	0.500 ± 0.006	0.011	0.501 ± 0.006	0.009	0.495 ± 0.005	0.007
10	0	2	0.5	0.379 ± 0.028	0.380	0.459 ± 0.021	0.224	0.461 ± 0.020	0.211	0.551 ± 0.020	0.206
50	0	2	0.5	0.573 ± 0.051	0.260	0.390 ± 0.023	0.055	0.413 ± 0.022	0.049	0.417 ± 0.018	0.038
10	0.5	2	0.5	0.440 ± 0.026	0.337	0.547 ± 0.025	0.316	0.465 ± 0.025	0.305	0.478 ± 0.024	0.298

Each case is based on $n_1 = \frac{5000}{n}$ samples of size n. The standard error of each mean is given to indicate the accuracy of the Monte Carlo computation.

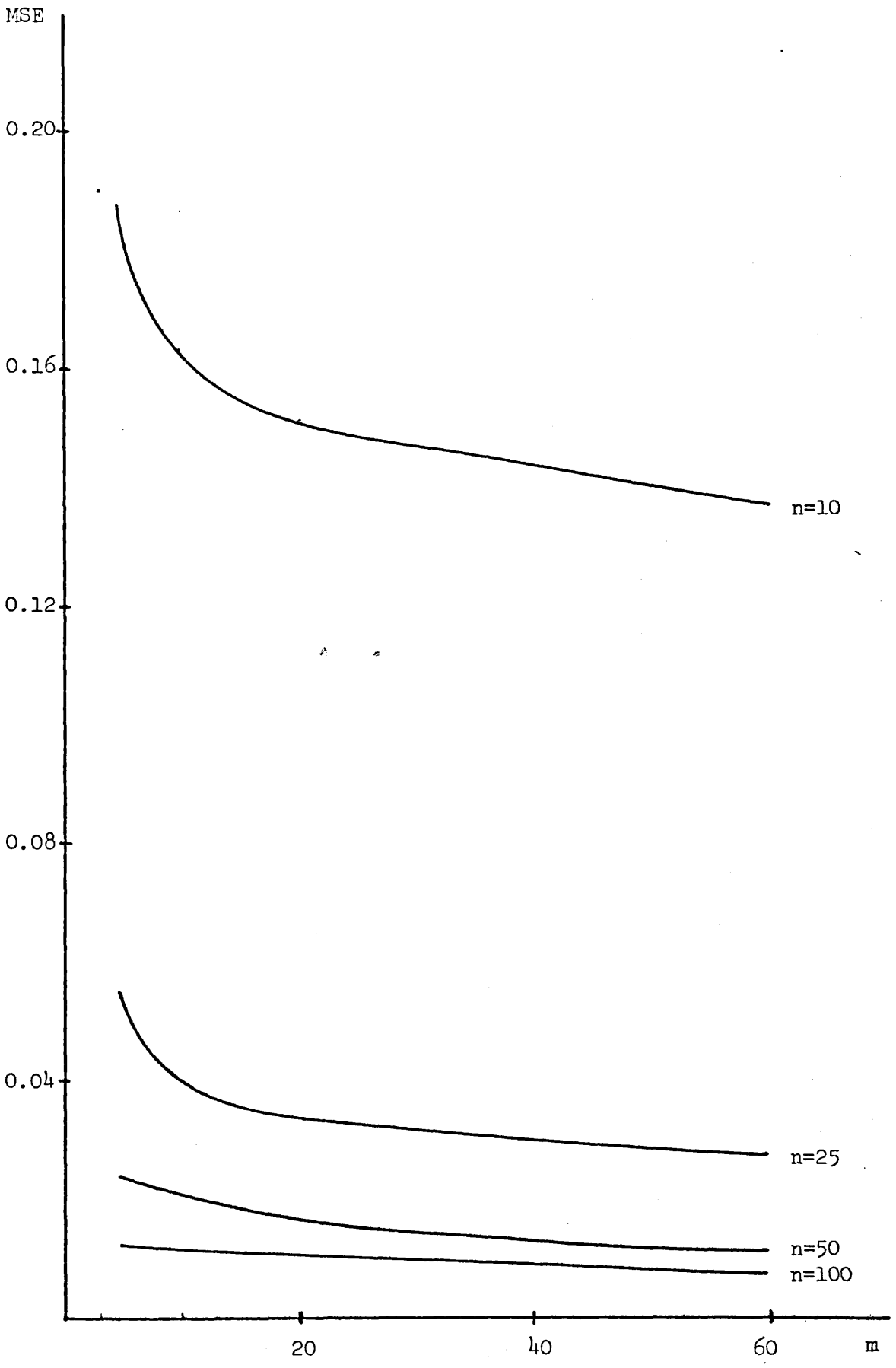


Fig. 3.1 - The Mean-Square-Error of the generalized Least Squares estimator of the mixing proportion in a mixture of two normal distributions N(0,1) and N(1,1), against m, the number of division points of the sample space, for different sample sizes.

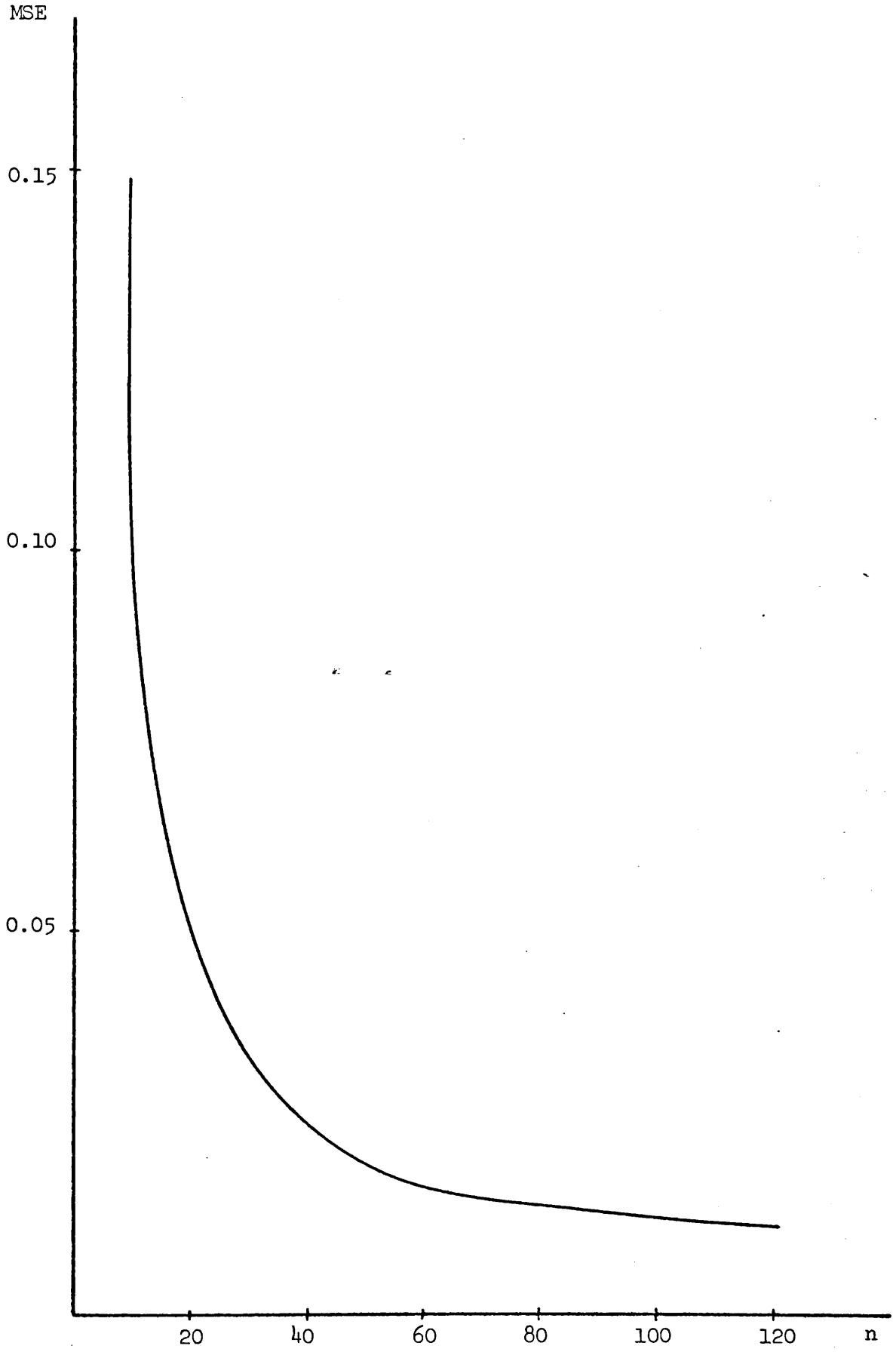


Fig. 3.2 - The Mean-Square-Error of the generalized Least Squares estimate of the mixing proportion in a mixture of two normal distributions N(0,1) and N(1,1), for varying sample sizes.

and $\frac{\sigma_2}{\sigma_1} = 1$ and plotted the mean-square error of our GLS estimator of θ_1 against m for different sample sizes. The value of N was taken to be 5000. Figure 3.1 shows that for relatively large n, the effect of choosing m greater than 20 is very little and perhaps, in some cases, not worth the computational effort.

Finally, for the same values of $(\mu_2-\mu_1)/\sigma_1$, σ_2/σ_1 and N and with m = 10, we plotted the mean-square-error of the GLS estimator of θ against n in figure 3.2. Again, as in section 2.7, the sharp fall of the curve as n increases, is noted.

3.6 Discussion

It can be argued that if the data is available ungrouped, then both the number of classes m and the division points $\{t_i\}_{i=1}^{m}$ may be chosen in many different ways and we always run the risk of influencing our results by these arbitrary choices of the class intervals. As R.A. Fisher points out "grouping causes a loss of information. By grouping we sacrifice knowledge of the exact size of the single observation, and hope to get compensation by an easier collection of our data".

The problem of finding the "best" choice of partitions of the sample space is of crucial concern in statistics and some authors have considered the problem in specific cases. Here, we mention just some of the more important publications in this subject. Gjeddebaek [22] considers different problems concerned with the estimation of the mean and variance of a normal population. He compares their maximum likelihood estimators when the observations are grouped with the corresponding estimates obtained from ungrouped data. The comparison is based on the asymptotic efficiencies of the estimates and he concludes that the loss of information due to grouping is not asymptotically significant when the group intervals are about twice the standard deviation.

The author gives examples of applications to some other statistical problems. Also in the context of maximum likelihood estimation, Kulldorf [32] makes a thorough study of some of the specialized problems of maximum likelihood estimation from grouped data.

Cox [13] has considered the problem of grouping in a more general context. He associates a value ξ_i to the ith group and this value is given to an individual falling in that group. Then the random variable $\xi(X)$ being a function of the random variable X, whose range is to be partitioned, is defined by $\xi(x) = \xi_i$ where x is in the range corresponding to the ith group. The author defines the loss due to grouping an individual in the ith group as $(x-\xi_i)^2/\sigma^2$ where σ^2 is the variance of X. He then considers the problem of minimizing the expected loss given by $E[X - \xi(X)]^2/\sigma^2$. In the theory of χ^2-test also, the problem of optimum grouping is discussed by some authors. Mann and Wald [36] suggest that the width of the class intervals be determined so that under the null hypothesis, specifying the distribution completely, the probability content of the classes are equal.

It is now clear that there is no general theory of the choice of partition points of the sample space. By the nature of the difficulties of the problem, outlined above, a complete solution to the problem is unlikely to be forthcoming soon.

Now, as far as our estimation problem is concerned, we are, to begin with, required to define what is meant by the "best" choices of m and $\{t_i\}_{i=1}^{m}$. When the grouping is done for convenience of exposition, any mathematical condition set up to define the "best" system of grouping is bound to be somewhat artificial. It seems reasonable to try to minimize the variance of the estimate, but unfortunately obtaining an explicit formula for the variance of the estimate seems impractical.

In practice, we may form our class intervals with full knowledge of the data and then proceed as though these intervals were known

a priori. This seems intuitively plausible since any fixed set of class intervals leads to the same asymptotic distribution. But, in small samples, to increase the accuracy, the general feeling of the statisticians is to put more computational work by increasing the number of intervals and choosing the intervals small enough to avoid classes with high expected frequencies. Thus the necessity to consider the case when $m \to \infty$ becomes evident and this is done in the next section.

3.7 The Generalized Least Squares Estimation of $\underset{\sim}{\theta}$ from Ungrouped Data

Let $s = G_{\underset{\sim}{\theta}}(x)$ for $x \in \mathcal{X}$ and define $W_n(s) = G_n(x)$, $\Omega_n(s) = \Gamma_n(x)$. Put $y_n(s) = s - W_n(s)$ and $Y_n(s) = s - \Omega_n(s)$. Then we have the linear model

$$s = W_n(s) + \varepsilon(s) \qquad (3.7.1)$$

where $0 \leq s \leq 1$ and $\varepsilon(s)$ is the realization of a random function $\&(s)$ with $E_{\underset{\sim}{\theta}}(\&(s)) = 0$ for every $0 \leq s \leq 1$. Similarly, let $s' = G_{\underset{\sim}{\theta}}(x')$ for $x' \in \mathcal{X}$ and define $W_n(s')$, $\Omega_n(s')$ accordingly. Then

$$\begin{aligned} K_{\underset{\sim}{\theta}}(s,s') &= \mathrm{Cov}_{\underset{\sim}{\theta}}(Y_n(s),Y_n(s')) = \mathrm{Cov}_{\underset{\sim}{\theta}}(\Omega_n(s),\Omega_n(s')) \\ &= \mathrm{Cov}_{\underset{\sim}{\theta}}(\Gamma_n(x),\Gamma_n(x')) = \min(G_{\underset{\sim}{\theta}}(x),G_{\underset{\sim}{\theta}}(x')) - G_{\underset{\sim}{\theta}}(x)G_{\underset{\sim}{\theta}}(x') \\ &= \min(s,s') - ss' \,. \end{aligned} \qquad (3.7.2)$$

It is well-known that $K_{\underset{\sim}{\theta}}(s,s')$, being a positive definite symmetric kernel, can be expressed uniquely in terms of its eigenvalues $j\pi$, $j = 1,2,\ldots$ and the corresponding eigenvectors $\sin(j\pi s)$ as

$$K_{\underset{\sim}{\theta}}(s,s') = \frac{2}{\pi^2} \sum_{j=1}^{\infty} \frac{\sin(j\pi s)\, \sin(j\pi s')}{j^2}$$

where the series is uniformly and absolutely convergent (Anderson and Darling [1]).

To begin with, assume that $s = G_{\underset{\sim}{\theta}}(x)$ is a simple function of $x \in \mathcal{X}$ taking only distinct values $s_1,\ldots,s_m$ such that $0 < s_1 < s_2 < \ldots < s_m < 1$. Then, denoting by Σ, the m×m matrix with (i,j)th element given by $K_{\underset{\sim}{\theta}}(s_i,s_j)$ for i,j = 1,...,m, from (3.7.2), we have

$$\Sigma = \begin{bmatrix} s_1 & s_1 & \cdots & s_1 \\ s_1 & s_2 & \cdots & s_2 \\ \vdots & \vdots & & \vdots \\ s_1 & s_2 & \cdots & s_m \end{bmatrix} - \begin{bmatrix} s_1^2 & s_1s_2 & \cdots & s_1s_m \\ s_1s_2 & s_2^2 & \cdots & s_2s_m \\ \vdots & \vdots & & \vdots \\ s_1s_m & s_2s_m & \cdots & s_m^2 \end{bmatrix}$$

Note that if in (3.2.9), we put $s_j = G_{\underset{\sim}{\theta}}(t_j)$ for j = 1,...,m, then Σ becomes identical to C. Hence analogous to the theorem 3.2.1, we have

$$\Sigma^{-1} = \begin{bmatrix} \frac{1}{\delta s_1} + \frac{1}{\delta s_2} & -\frac{1}{\delta s_2} & & & \\ -\frac{1}{\delta s_2} & \frac{1}{\delta s_2} + \frac{1}{\delta s_3} & -\frac{1}{\delta s_3} & & \\ & -\frac{1}{\delta s_3} & \frac{1}{\delta s_3} + \frac{1}{\delta s_4} & \ddots & \\ & & \ddots & \ddots & -\frac{1}{\delta s_m} \\ & & & -\frac{1}{\delta s_m} & \frac{1}{\delta s_m} + \frac{1}{\delta s_{m+1}} \end{bmatrix}$$

where $\delta s_j = s_j - s_{j-1}$ j = 1,...,m+1 and $s_o = 0$, $s_{m+1} = 1$.

Denote by $\underset{\sim}{y}$, the m×1 vector $\underset{\sim}{y} = (y_n(s_1),\ldots,y_n(s_m))'$ and let $\Delta(s_j) = y_n(s_j) - y_n(s_{j-1})$, $\gamma_n(s_j) = W_n(s_j) - W_n(s_{j-1})$ for j = 1,...,m. Then the GSS in the linear model (3.7.1) is given by

$$\underset{\sim}{y}'\Sigma^{-1}\underset{\sim}{y} = \sum_{j=1}^{m+1} \frac{(\Delta(s_j))^2}{\delta s_j} = \sum_{j=1}^{m+1} \left(\frac{\delta s_j - \gamma_n(s_j)}{\delta s_j}\right)^2 \delta s_j \quad . \tag{3.7.3}$$

For an appropriate sequence of progressively finer intervals δs_j, $j = 1,\ldots,m+1$, it is seen that as $m \to \infty$, the limit of (3.7.3) is

$$\Phi(\underset{\sim}{\theta}) = \int \left(\frac{dG_n(x) - dG_{\underset{\sim}{\theta}}(x)}{dG_{\underset{\sim}{\theta}}(x)}\right)^2 dG_{\underset{\sim}{\theta}}(x) = \int \left(\frac{dG_n(x)}{dG_{\underset{\sim}{\theta}}(x)}\right)^2 dG_{\underset{\sim}{\theta}}(x) - 1 \, . \tag{3.7.4}$$

Thus the GLS estimator of $\underset{\sim}{\theta}$ is that value $\hat{\underset{\sim}{\theta}}_n$ which minimizes (3.7.4).

It is interesting to note the similarity of (3.7.4) with the measure of distance defined by Bartlett and Macdonald [4]. They define the least squares estimator of $\underset{\sim}{\theta}$ as the value minimizing

$$\int \frac{(dG_n(x) - dG_{\underset{\sim}{\theta}}(x))^2}{dW(x)}$$

where W is a suitable increasing function of x and conclude that the best choice of W(x) is indeed $G_{\underset{\sim}{\theta}}(x)$ since with this choice, the estimate of $\underset{\sim}{\theta}$ will be asymptotically efficient.

In the following, we establish the asymptotic properties of the GLS estimator of $\underset{\sim}{\theta}$. For simplicity, we deal with the case $k = 2$, although there is an immediate generalization. Let $\underset{\sim}{\theta} = (\theta_1,\theta_2)'$ denote the vector of the mixing proportions θ_1; $0 \le \theta_1 \le 1$ and $\theta_2 = 1-\theta_1$ in the mixture

$$G_{\underset{\sim}{\theta}}(x) = \theta_1 F_1(x) + \theta_2 F_2(x) \qquad x \in \mathfrak{X}$$

and suppose $\underset{\sim}{\theta}^* = (\theta_1^*,\theta_2^*)'$ with $0 \le \theta_1^* \le 1$ and $\theta_2^* = 1-\theta_1^*$ is the vector of the true parameter values.

Notation: In the sequel, the well-known symbols $o_p(.)$ and $O_p(.)$ will be used to denote the rate of convergence in probability. Thus if $U_n = o_p(W_n)$ as $n \to \infty$, then for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \text{Prob}\,(|U_n/W_n| > \varepsilon) = 0$$

and similarly $U_n = O_p(W_n)$ if there exists a constant C, $0 < C < \infty$, such that

$$\lim_{n\to\infty} \text{Prob}\,(|U_n/W_n| \leq C) = 1 \ .$$

Lemma 3.7.1: The function $\Phi(\underset{\sim}{\theta})$ given by (3.7.4) is infinitely differentiable with respect to θ_1 under the integral sign.

Proof: From (3.7.4),

$$\Phi(\underset{\sim}{\theta}) = \theta_1 \int \left[\frac{dG_n(x)}{dG_{\underset{\sim}{\theta}}(x)}\right]^2 dF_1(x) + (1-\theta_1) \int \left[\frac{dG_n(x)}{dG_{\underset{\sim}{\theta}}(x)}\right]^2 dF_2(x) - 1$$

and using the Lebesgue dominated convergence theorem, $\Phi(\underset{\sim}{\theta})$ is infinitely differentiable with respect to θ_1 whenever

$\frac{\partial^r}{\partial\theta_1^r}\left[\frac{dG_n(x)}{dG_{\underset{\sim}{\theta}}(x)}\right]^2$ exists for $r = 1,2,\ldots$ and is bounded by a function of $x \in \mathfrak{X}$ only (except possibly for a set of points to which $F_1(.)$ and $F_2(.)$ assign zero probability) which is integrable with respect to $F_1(.)$ and $F_2(.)$. Now

$$\left| \frac{\partial^r}{\partial\theta} \left[\frac{dG_n(x)}{dG_{\underset{\sim}{\theta}}(x)}\right]^2 \right| = (r+1)! \left| \frac{(dF_1(x) - dF_2(x))^r (dG_n(x))^2}{(dG_{\underset{\sim}{\theta}}(x))^{r+2}} \right|$$

$$= \begin{cases} (r+1)! \dfrac{\{dF_1(x)-dF_2(x)\}^r \{dG_n(x)\}^2}{\{\theta_1(dF_1(x)-dF_2(x))+dF_2(x)\}^{r+2}} & \text{for } x \in \{y: y\in\mathfrak{X},\ dF_1(y)\geq dF_2(y)\} \\ (r+1)! \dfrac{\{dF_2(x)-dF_1(x)\}^r \{dG_n(x)\}^2}{\{dF_1(x)+(1-\theta_1)(dF_2(x)-dF_1(x))\}^{r+2}} & \text{for } x\in\{y: y\in\mathfrak{X},\ dF_1(y)\leq dF_2(y)\} \end{cases}$$

$$\leq \begin{cases} (r+1)! \dfrac{\{dF_1(x)-dF_2(x)\}^r \{dG_n(x)\}^2}{\{dF_2(x)\}^{r+2}} & \text{for } x \in \{y: y\in\mathfrak{X},\ dF_1(y) \geq dF_2(y)\} \\ (r+1)! \dfrac{\{dF_2(x)-dF_1(x)\}^r \{dG_n(x)\}^2}{\{dF_1(x)\}^{r+2}} & \text{for } x \in \{y: y\in\mathfrak{X},\ dF_1(y) \leq dF_2(y)\} \end{cases}$$

for $r = 1,2,\ldots$. Thus the conditions of the dominated convergence theorem are satisfied and the lemma is proved.

Suppose now that the distribution functions $F_1(.)$ and $F_2(.)$ and hence $G_{\underset{\sim}{\theta}}(.)$ are absolutely continuous and there exists densities

$$f_1(x) = \frac{dF_1(x)}{dx}, f_2(x) = \frac{dF_2(x)}{dx} \quad \text{and} \quad g_{\underset{\sim}{\theta}}(x) = \frac{dG_{\underset{\sim}{\theta}}(x)}{dx} \quad \text{such that}$$

$$g_{\underset{\sim}{\theta}}(x) = \theta_1 f_1(x) + (1-\theta_1) f_2(x) \qquad x \in \mathfrak{X} \quad .$$

Theorem 3.7.1: If we denote by $\underset{\sim}{Z}_n = (Z_1,Z_2)'$ the random vector whose realization $\underset{\sim}{\hat{\theta}}_n = (\hat{\theta}_1,\hat{\theta}_2)'$ is the GLS estimator of $\underset{\sim}{\theta}$, then $\underset{\sim}{Z}_n$ is CAN with asymptotic variance reaching the Cramér-Rao lower bound.

Proof: Using the lemma 3.7.1 and differentiating $\Phi(\underset{\sim}{\theta})$, given by (3.7.4), with respect to θ_1, we see that Z_1 is the root of

$$\frac{d\Phi(\underset{\sim}{\theta})}{d\theta_1} = -\int \frac{(dF_n(x)-dG_{\underset{\sim}{\theta}}(x))(dF_1(x)-dF_2(x))}{dG_{\underset{\sim}{\theta}}(x)} \left\{1 + \frac{dF_n(x)}{dG_{\underset{\sim}{\theta}}(x)}\right\} . \qquad (3.7.5)$$

Now $d\Gamma_n(x) \equiv \Gamma_n(x+\Delta x) - \Gamma_n(x)$ where $\Delta x > 0$ is a very small quantity and recall (c.f. equation (1.3.4)) that

$$\Gamma_n(x) = \frac{1}{n} \sum_{j=1}^{n} \eta(x-X_j) \qquad x \in \mathfrak{X}$$

where $X_1,\ldots,X_n$ are independent random variables with common distribution function $G_{\underset{\sim}{\theta}}(.)$ and $\eta(.)$ is the well-known Heaviside function. Since

$$E_{\underset{\sim}{\theta}}[\eta(x-X_j)] = \int \eta(x-t)\, dG_{\underset{\sim}{\theta}}(t) = G_{\underset{\sim}{\theta}}(x) \qquad x \in \mathfrak{X},$$

then $\Gamma_n(x) \xrightarrow{p} G_{\underset{\sim}{\theta}}(x)$ as $n \to \infty$ for every $x \in \mathfrak{X}$ and therefore

$$\Gamma_n(x+\Delta x) - \Gamma_n(x) \xrightarrow{p} \int_x^{x+\Delta x} g_{\underset{\sim}{\theta}}(t)\, dt = G_{\underset{\sim}{\theta}}(x+\Delta x) - G_{\underset{\sim}{\theta}}(x)$$

$$\equiv dG_{\underset{\sim}{\theta}}(x) \text{ as } n \to \infty .$$

Hence for sufficiently large n, $\dfrac{d\Gamma_n(x)}{dG_{\underset{\sim}{\theta}}(x)} = 1 + o_p(1)$ and thus (3.7.5) gives

$$\frac{d\Phi(\underset{\sim}{\theta})}{d\theta_1} = -\int \frac{(d\Gamma_n(x)-dG_{\underset{\sim}{\theta}}(x))(dF_1(x)-dF_2(x))}{dG_{\underset{\sim}{\theta}}(x)} \,(2 + o_p(1))$$

which shows that for sufficiently large n, the influence of the term involving $o_p(1)$ is negligible and Z_1 is the root of

$$\int \frac{dF_1(x)-dF_2(x)}{dG_{\underset{\sim}{\theta}}(x)} (d\Gamma_n(x) - dG_{\underset{\sim}{\theta}}(x)) = \int \frac{f_1(x)-f_2(x)}{g_{\underset{\sim}{\theta}}(x)}\, d\Gamma_n(x)$$

$$= \frac{1}{n} \sum_{j=1}^{n} \frac{f_1(X_j)-f_2(X_j)}{g_{\underset{\sim}{\theta}}(X_j)} \qquad (3.7.6)$$

and consequently θ_1 is the root of

$$\frac{1}{n}\sum_{j=1}^{n}\frac{f_1(x_j)-f_2(x_j)}{g_{\underset{\sim}{\theta}}(x_j)} = 0 \qquad (3.7.7)$$

where $x_1,\ldots,x_n$ are the realizations of $X_1,\ldots,X_n$ respectively.

But it will be shown later in section 5.2 that the root of (3.7.7) corresponds to the maximum likelihood estimator of θ_1^* and therefore by the well-known properties of the maximum likelihood estimators, Z_1 (being the root of (3.7.6) for sufficiently large n) possesses the asymptotic properties stated in the theorem.

3.8 Conclusions

The claim that the consideration of the covariance matrix in the minimization of the sum of squares, will improve the efficiency of the estimate of the vector of the mixing proportions $\underset{\sim}{\theta}$, is now justified. It is seen that by defining h(x,t) as in section 3.2, and minimizing the generalized sum of squares, a fully efficient estimator of $\underset{\sim}{\theta}$ is obtained.

The difficulties, however, are clearly in solving the resulting equations. To solve (3.3.4) analytically is, of course, an impractical task and unless some numerical approach, e.g. successive substitution is taken, the estimation procedure cannot be usefully employed. This approach is discussed in the next chapter.

CHAPTER 4

LEAST SQUARES ESTIMATION - USE OF ITERATION

4.1 Introduction

It is clear that since the covariance matrix C given by (3.2.9) depends on $\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_k)'$, the estimating equations (3.3.4) become non-linear functions of $\theta_1, \ldots, \theta_k$ and thus the GLS estimator of $\underset{\sim}{\theta}$, i.e. the value of $\underset{\sim}{\theta}$ minimizing (3.2.17), becomes very cumbersome to calculate. In fact, in most instances a solution cannot be obtained directly. In this chapter, we propose an iterative procedure whereby the covariance matrix is calculated in each step and is used to find the GLS estimate of $\underset{\sim}{\theta}$ in the following step. It turns out that the sequence of estimators obtained in this way has special characteristics and indeed when m in (3.3.4) is large, the process converges to the maximum likelihood estimate of $\underset{\sim}{\theta}$.

In Section 4.2, the iterative process is introduced and it is shown that when

(i) the partitioning of the sample space $\mathfrak{X}$ is done by fixing the division points $t_1 < t_2 < \ldots < t_m$ satisfying (3.2.7),

(ii) the random sample $X_1, \ldots, X_n$ with realizations $x_1, \ldots, x_n$ respectively, from the mixture of distributions $G_{\underset{\sim}{\theta}}(x)$ given by (2.2.1), is grouped accordingly,

(iii) the iteration process is started with a consistent estimate of $\underset{\sim}{\theta}$,

then the estimator of $\underset{\sim}{\theta}$ obtained after a 1-cycle iteration is CAN with minimum attainable variance with respect to such a grouping. The results of a small Monte Carlo study are presented in Section 4.3. Discussing the iteration process for the ungrouped data in Section 4.4, we prove that the process will converge to the maximum likelihood estimator of $\underset{\sim}{\theta}$. It is assumed in Section 4.4 that the densities $f_1(x), \ldots, f_k(x)$ and

hence $g_{\underset{\sim}{\theta}}(x) = \sum_{j=1}^{k} \theta_j f_j(x)$, where $\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_k)' \in \Theta$, of the distribution functions $F_1(x), \ldots, F_k(x)$ and $G_{\underset{\sim}{\theta}}(x) = \sum_{j=1}^{k} \theta_j F_j(x)$ respectively exist and are differentiable with respect to $x \in \mathfrak{X}$. Recall from Section 3.3 that Θ denotes the k-fold product space $[0,1] \times [0,1] \times \ldots \times [0,1]$.

4.2 The Iteration Process

An arbitrary value of $\underset{\sim}{\theta}$, say $\hat{\underset{\sim}{\theta}}_n^{(0)} = (\hat{\theta}_1^{(0)}, \ldots, \hat{\theta}_k^{(0)})'$ is chosen in Θ such that $\sum_{j=1}^{k} \hat{\theta}_j^{(0)} = 1$. This value is substituted in (3.2.13) for $\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_k)'$ to obtain C_0^{-1} which is then inserted for C^{-1} in (3.2.4) to give

$$\Phi_0(\underset{\sim}{\theta}) = (\underset{\sim}{G}_n - A\underset{\sim}{\theta})' \, C_0^{-1} \, (\underset{\sim}{G}_n - A\underset{\sim}{\theta}). \qquad (4.2.1)$$

To impose the constraint $\sum_{j=1}^{k} \theta_j = 1$, put $\theta_\ell = 1 - \sum_{\substack{j=1 \\ j \neq \ell}}^{k} \theta_j$, for some integer ℓ, $1 \leq \ell \leq k$, in (4.2.1) and minimize it with respect to $\theta_1, \ldots, \theta_{\ell-1}, \theta_{\ell+1}, \ldots, \theta_k$. (For convenience, we may choose $\ell = k$ and put $\theta_k = 1 - \sum_{j=1}^{k-1} \theta_j$ in (4.2.1) and minimize it with respect to $\theta_1, \ldots, \theta_{k-1}$. Let $\hat{\theta}_1^{(1)}, \ldots, \hat{\theta}_{\ell-1}^{(1)}, \hat{\theta}_{\ell+1}^{(1)}, \ldots, \hat{\theta}_k^{(1)}$ be the minimizing values and set $\hat{\theta}_\ell^{(1)} = 1 - \sum_{\substack{j=1 \\ j \neq \ell}}^{k} \hat{\theta}_j^{(1)}$. Then $\hat{\underset{\sim}{\theta}}_n^{(1)} = (\hat{\theta}_1^{(1)}, \ldots, \hat{\theta}_k^{(1)})'$ forms the GLS estimator of $\underset{\sim}{\theta}$ after a 1-cycle iteration. We now substitute $\hat{\underset{\sim}{\theta}}_n^{(1)}$ in (3.2.13) for $\underset{\sim}{\theta}$ to obtain C_1^{-1} and the process is repeated so that after the rth cycle of the process r = 0, 1, 2, ..., the following steps are taken;

(i) $\hat{\underset{\sim}{\theta}}_n^{(r)} = (\hat{\theta}_1^{(r)}, \ldots, \hat{\theta}_k^{(r)})'$ is substituted for $\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_k)'$ in $G_{\underset{\sim}{\theta}}(t_i) = \sum_{j=1}^{k} \theta_j F_j(t_i)$ for i = 1, ..., m

(ii) Analogous to (3.2.12), $\rho_1^{(r)}, \ldots, \rho_m^{(r)}$ are calculated from

$$\rho_1^{(r)} = \left[G_{\hat{\underset{\sim}{\theta}}_n^{(r)}}(t_1) \right]^{-1} \text{ and } \rho_i^{(r)} = \left[G_{\hat{\underset{\sim}{\theta}}_n^{(r)}}(t_i) - G_{\hat{\underset{\sim}{\theta}}_n^{(r)}}(t_{i-1}) \right]^{-1}$$

for i = 2, 3, ..., m

(iii) $\rho_1^{(r)}, \ldots, \rho_m^{(r)}$ are substituted for $\rho_1, \ldots, \rho_m$ in (3.2.11) to obtain B_r^{-1}.

(iv) C_r^{-1} is calculated, analogous to (3.2.13), from

$$C_r^{-1} = B_r^{-1} + \frac{1}{1 - G_{\hat{\theta}_n^{(r)}}(t_m)} L$$

where L, as before, is an m × m matrix with zero in all the entries except the last element of the last column which is unity.

(v) C_r^{-1}, defined in this way, is inserted for C^{-1} in (3.2.4) to give

$$\Phi_r(\underset{\sim}{\theta}) = (\underset{\sim}{G}_n - A\underset{\sim}{\theta})' \, C_r^{-1} \, (\underset{\sim}{G}_n - A\underset{\sim}{\theta}) \qquad (4.2.2)$$

where $\underset{\sim}{G}_n = (G_n(t_1), \ldots, G_n(t_m))'$ and A is as given by (3.2.2).

(vi) To impose the constraint $\sum_{j=1}^{k} \theta_j = 1$, pick θ_ℓ amongst $\theta_1, \ldots, \theta_k$ for some integer $1 \leq \ell \leq k$ (possibily $\ell = k$ for convenience) and substitute $\theta_\ell = 1 - \sum_{\substack{j=1 \\ j \neq \ell}}^{k} \theta_j$ into $\Phi_r(\underset{\sim}{\theta})$, given by (4.2.2), and minimize $\Phi_r(\underset{\sim}{\theta})$ with respect to $\theta_1, \ldots, \theta_{\ell-1}, \theta_{\ell+1}, \ldots, \theta_k$. Call the minimizing values $\hat{\theta}_1^{(r+1)}, \ldots, \hat{\theta}_{\ell-1}^{(r+1)}, \hat{\theta}_{\ell+1}^{(r+1)}, \ldots, \hat{\theta}_k^{(r+1)}$ and set $\hat{\theta}_\ell^{(r+1)} = 1 - \sum_{\substack{j=1 \\ j \neq \ell}}^{k} \hat{\theta}_j^{(r+1)}$. Then $\hat{\underset{\sim}{\theta}}_n^{(r+1)} = (\hat{\theta}_1^{(r+1)}, \ldots, \hat{\theta}_k^{(r+1)})'$ forms the GLS estimator of $\underset{\sim}{\theta}$ after r cycles of the iteration process.

The iteration process is continued until $d(\hat{\underset{\sim}{\theta}}_n^{(r+1)}, \hat{\underset{\sim}{\theta}}_n^{(r)})$, where d is some suitable distance function defined over Θ, becomes negligible. A practical choice of d may be

$$d(\underset{\sim}{\alpha}, \underset{\sim}{\beta}) = \left[\sum_{i=1}^{k} (\alpha_i - \beta_i)^2 \right]^{\frac{1}{2}}$$

for any $\underset{\sim}{\alpha}$, $\underset{\sim}{\beta} \in R^k$ where R^k is the space of k-dimensional real vectors. But this choice of d is clearly by no means unique.

We establish in theroem 4.2.1 (below)the properties of $\hat{\underset{\sim}{\theta}}_n^{(1)}$, the estimate of $\underset{\sim}{\theta}$ provided by a 1-cycle iteration in the above process. To avoid unnecessary algebraic manipulations, the theorem will be proved for the case k = 2,i.e. when the mixture of distributions $G_{\underset{\sim}{\theta}}(x)$ given by (2.2.1) consists only of two components $F_1(x)$ and $F_2(x)$. The generalization of the theorem to the case k > 2,i.e. when $G_{\underset{\sim}{\theta}}(x)$ consists of more than two components is laborious and lengthy and the details are given in Appendix B.

The mixture $G_{\underset{\sim}{\theta}}(x)$ of two distribution functions $F_1(x)$ and $F_2(x)$ with mixing proportions $0 \leq \theta_1 \leq 1$ and $\theta_2 = 1 - \theta_1$ respectively and with the vector of the mixing proportions $\underset{\sim}{\theta} = (\theta_1, \theta_2)'$, will depend on the scalar parameter θ_1 only and is written as

$$G_{\underset{\sim}{\theta}}(x) = \theta_1 F_1(x) + (1-\theta_1) F_2(x) \qquad \begin{matrix} 0 \leq \theta_1 \leq 1 \\ x \in \mathfrak{X} \end{matrix} \qquad (4.2.3)$$

Also, we can show that $\Phi_r(\underset{\sim}{\theta})$ r = 0, 1, ... given by (4.2.2) can be written as

$$\Phi_r(\underset{\sim}{\theta}) = \sum_{i=1}^{m+1} \frac{(p_i - \pi_i(\underset{\sim}{\theta}))^2}{\pi_i(\hat{\underset{\sim}{\theta}}_n^{(r)})} \qquad r = 0,1,2, \ldots \quad (4.2.4)$$

where, as before, $p_i = G_n(t_i) - G_n(t_{i-1})$, $\pi_i(\underset{\sim}{\theta}) = G_{\underset{\sim}{\theta}}(t_i) - G_{\underset{\sim}{\theta}}(t_{i-1})$ and $\min_i \pi_i(\underset{\sim}{\theta}) > 0$ for i = 1, ..., m + 1 and for every $\underset{\sim}{\theta} \in \Theta$. Substituting for $G_{\underset{\sim}{\theta}}(x)$ from (4.2.3) and setting the derivative with respect to θ_1 of $\Phi_r(\underset{\sim}{\theta})$ equal to zero, we find that the estimate of θ_1, after r + 1 cycles of the iteration process is given by

$$\hat{\theta}_1^{(r+1)} = \frac{\sum_{i=1}^{m+1} \frac{(p_i-\beta_{2i})(\beta_{1i}-\beta_{2i})}{\pi_i(\hat{\theta}_{\sim n}^{(r)})}}{\sum_{i=1}^{m+1} \frac{(\beta_{1i}-\beta_{2i})^2}{\pi_i(\hat{\theta}_{\sim n}^{(r)})}} \qquad r = 0, 1, \ldots \quad (4.2.5)$$

where $\beta_{ji} = F_j(t_i) - F_j(t_{i-1})$ for $j = 1, 2$ and for $i = 1, \ldots, m+1$ with $F_j(t_0) = G_n(t_0) = 0$ and $F_j(t_{m+1}) = G_n(t_{m+1}) = 1$. Further, the numerator of (4.2.5) can be written as

$$\sum_{i=1}^{m+1} \frac{(p_i-\beta_{2i})(\beta_{1i}-\beta_{2i})}{\pi_i(\hat{\theta}_{\sim n}^{(r)})} = \sum_{i=1}^{m+1} \frac{(p_i-\pi_i(\hat{\theta}_{\sim n}^{(r)}))(\beta_{1i}-\beta_{2i}) + (\pi_i(\hat{\theta}_{\sim n}^{(r)})-\beta_{2i})(\beta_{1i}-\beta_{2i})}{\pi_i(\hat{\theta}_{\sim n}^{(r)})}$$

and since $\pi_i(\underset{\sim}{\theta}) = \theta_1 \beta_{1i} + (1 - \theta_1)\beta_{2i}$ for $i = 1, \ldots, m+1$, we have $\pi_i(\underset{\sim}{\theta}) - \beta_{2i} = \theta_1 (\beta_{1i} - \beta_{2i})$ for $i = 1, \ldots, m+1$. Thus

$$\sum_{i=1}^{m+1} \frac{(p_i-\beta_{2i})(\beta_{1i}-\beta_{2i})}{\pi_i(\hat{\theta}_{\sim n}^{(r)})} = \sum_{i=1}^{m+1} \frac{(p_i-\pi_i(\hat{\theta}_{\sim n}^{(r)}))(\beta_{1i}-\beta_{2i})}{\pi_i(\hat{\theta}_{\sim n}^{(r)})}$$

$$+ \hat{\theta}_1^{(r)} \sum_{i=1}^{m+1} \frac{(\beta_{1i}-\beta_{2i})^2}{\pi_i(\hat{\theta}_{\sim n}^{(r)})}$$

$$= \sum_{i=1}^{m+1} \frac{p_i(\beta_{1i}-\beta_{2i})}{\pi_i(\hat{\theta}_{\sim n}^{(r)})} - \sum_{i=1}^{m+1} (\beta_{1i}-\beta_{2i}) + \hat{\theta}_1^{(r)} \sum_{i=1}^{m+1} \frac{(\beta_{1i}-\beta_{2i})^2}{\pi_i(\hat{\theta}_{\sim n}^{(r)})}$$

$$= \sum_{i=1}^{m+1} \frac{p_i(\beta_{1i}-\beta_{2i})}{\pi_i(\hat{\theta}_{\sim n}^{(r)})} + \hat{\theta}_1^{(r)} \sum_{i=1}^{m+1} \frac{(\beta_{1i}-\beta_{2i})^2}{\pi_i(\hat{\theta}_{\sim n}^{(r)})} \qquad (4.2.6)$$

since $\sum_{i=1}^{m+1} (\beta_{1i}-\beta_{2i}) = F_1(t_{m+1}) - F_2(t_{m+1}) = 0$. Hence by substituting (4.2.6) in (4.2.5), we obtain

$$\hat{\theta}_1^{(r+1)} = \hat{\theta}_1^{(r)} + \frac{\sum_{i=1}^{m+1} \frac{p_i(\beta_{1i}-\beta_{2i})}{\pi_i(\hat{\theta}_{\sim n}^{(r)})}}{\sum_{i=1}^{m+1} \frac{(\beta_{1i}-\beta_{2i})^2}{\pi_i(\hat{\theta}_{\sim n}^{(r)})}} \qquad r = 0,1, \ldots \qquad (4.2.7)$$

We finally denote the true value of the parameter $\underset{\sim}{\theta} = (\theta_1, \theta_2)'$ by $\underset{\sim}{\theta}^* = (\theta_1^*, \theta_2^*)'$ where $0 \leq \theta_1^* \leq 1$ and $\theta_2^* = 1 - \theta_1^*$ and also use the symbol $o_p(.)$ introduced in Section 3.7.

Lemma 4.2.1: Let $\xi_1, \xi_2, \ldots$ be a sequence of random variables with the distribution functions $F_1, F_2, \ldots$. Suppose that $F_n(.)$ tends to a distribution function $F(.)$ as $n \to \infty$. Let $\gamma_1, \gamma_2, \ldots$ be another sequence of random variables, and suppose that γ_n converges in probability to a constant c. Then the distribution function of $Y_n = \xi_n + \gamma_n$ tends to $F(x-c)$ as $n \to \infty$.

Proof: Cramer [15].

Theorem 4.2.1: Let $\hat{\underset{\sim}{\theta}}_n^{(r)} = (\hat{\theta}_1^{(r)}, 1 - \hat{\theta}_1^{(r)})'$ be the realization of the random vector $\underset{\sim}{Z}_n^{(r)} = (Z_1^{(r)}, 1 - Z_1^{(r)})'$ for $r = 0, 1, \ldots$. If $\hat{\theta}_1^{(0)}$ is a consistent estimator of θ_1^* such that $Z_1^{(0)} - \theta_1^* = o_p(n^{-\frac{1}{4}})$ as $n \to \infty$, then $\hat{\theta}_1^{(1)}$ has the property that $Z_1^{(1)}$ is consistent asymptotically normal with asymptotic variance given by

$$\left[n \sum_{i=1}^{m+1} \frac{(\beta_{1i} - \beta_{2i})^2}{\pi_i(\underset{\sim}{\theta}^*)} \right]^{-1}$$

Proof: Denote by P_i the random variable whose realization is $p_i = G_n(t_i) - G_n(t_{i-1})$ so that $P_i = \Gamma_n(t_i) - \Gamma_n(t_{i-1})$ for $i = 1, \ldots, m+1$. Write $Z_1^{(0)} = \theta_1^* + \varepsilon$ and $P_i = \pi_i(\underset{\sim}{\theta}^*) + \eta_i$ for $i = 1, \ldots, m+1$. Then $\varepsilon = o_p(n^{-\frac{1}{4}})$ as $n \to \infty$ and since P_i is a random variable admitting first and second order moments, $\eta_i = o_p(n^{-\alpha})$ as $n \to \infty$ for all $\alpha < 1$. Thus from (4.2.7), we have

$$Z_1^{(1)} = Z_1^{(0)} + \left[\sum_{i=1}^{m+1} \frac{(\pi_i(\underset{\sim}{\theta^*})+\eta_i)(\beta_{1i}-\beta_{2i})}{\pi_i(\underset{\sim}{\theta^*})\left[1 + \frac{\varepsilon(\beta_{1i}-\beta_{2i})}{\pi_i(\underset{\sim}{\theta^*})}\right]}\right]\left[\sum_{i=1}^{m+1} \frac{(\beta_{1i}-\beta_{2i})^2}{\pi_i(\underset{\sim}{\theta^*})\left[1 + \frac{\varepsilon(\beta_{1i}-\beta_{2i})}{\pi_i(\underset{\sim}{\theta^*})}\right]}\right]^{-1}$$

$$= Z_1^{(0)} + \left[\sum_{i=1}^{m+1} \frac{(\pi_i(\underset{\sim}{\theta^*})+\eta_i)(\beta_{1i}-\beta_{2i})}{\pi_i(\underset{\sim}{\theta^*})}\left[1 - \frac{\varepsilon(\beta_{1i}-\beta_{2i})}{\pi_i(\underset{\sim}{\theta^*})} + O(\varepsilon^2)\right]\right]$$

$$\left[\left[\sum_{i=1}^{m+1} \frac{(\beta_{1i}-\beta_{2i})^2}{\pi_i(\underset{\sim}{\theta^*})}\right]^{-1} + O(\varepsilon)\right]$$

$$= Z_1^{(0)} + \frac{\sum_{i=1}^{m+1} \frac{\eta_i(\beta_{1i}-\beta_{2i})}{\pi_i(\underset{\sim}{\theta^*})}}{\sum_{i=1}^{m+1} \frac{(\beta_{1i}-\beta_{2i})^2}{\pi_i(\underset{\sim}{\theta^*})}} - \varepsilon - \varepsilon \frac{\sum_{i=1}^{m+1} \frac{\eta_i(\beta_{1i}-\beta_{2i})^2}{(\pi_i(\underset{\sim}{\theta^*}))^2}}{\sum_{i=1}^{m+1} \frac{(\beta_{1i}-\beta_{2i})^2}{\pi_i(\underset{\sim}{\theta^*})}} + O(\varepsilon^2) \quad .$$

Hence

$$\sqrt{n}(Z_1^{(1)} - \theta_1^*) = \frac{\sqrt{n}\sum_{i=1}^{m+1} \frac{\eta_i(\beta_{1i}-\beta_{2i})}{\pi_i(\underset{\sim}{\theta^*})}}{\sum_{i=1}^{m+1} \frac{(\beta_{1i}-\beta_{2i})^2}{\pi_i(\underset{\sim}{\theta^*})}} + o_p(1) \qquad (4.2.8)$$

and by substituting $\eta_i = P_i - \pi_i(\underset{\sim}{\theta^*})$ in (4.2.8), and using the lemma 4.2.1, we see that the random variables $\sqrt{n}(Z_1^{(1)} - \theta_1^*)$ and

$$Y = \frac{\sqrt{n}\sum_{i=1}^{m+1} \frac{P_i(\beta_{1i}-\beta_{2i})}{\pi_i(\underset{\sim}{\theta^*})}}{\sum_{i=1}^{m+1} \frac{(\beta_{1i}-\beta_{2i})^2}{\pi_i(\underset{\sim}{\theta^*})}}$$

have identical asymptotic distributions.

Now, it is shown in Appendix A that the joint asymptotic distribution of $P_1, \ldots, P_{m+1}$ is a (m+1)-variate normal distribution with mean vector

$$\underset{\sim}{\pi}(\underset{\sim}{\theta^*}) = (\pi_1(\underset{\sim}{\theta^*}), \ldots, \pi_{m+1}(\underset{\sim}{\theta^*}))'$$

and (m+1) × (m+1) covariance matrix $\frac{1}{n}\Sigma$, where the (i,j)th element of Σ is given by

$$\sigma_{ij} = (\pi_i(\underset{\sim}{\theta^*}) - \pi_i^2(\underset{\sim}{\theta^*})) \qquad i = j$$

$$= -\pi_i(\underset{\sim}{\theta^*})\,\pi_j(\underset{\sim}{\theta^*}) \qquad i \neq j$$

for $i, j = 1, \ldots, m+1$. Let $\underset{\sim}{b} = (b_1, \ldots, b_{m+1})'$ with

$$b_i = \frac{\sqrt{n}\,\frac{\beta_{1i}-\beta_{2i}}{\pi_i(\underset{\sim}{\theta^*})}}{\sum_{i=1}^{m+1} \frac{(\beta_{1i}-\beta_{2i})^2}{\pi_i(\underset{\sim}{\theta^*})}} \qquad i = 1, \ldots, m+1$$

and $\underset{\sim}{P} = (P_1, \ldots, P_{m+1})'$. Then we can write $Y = \underset{\sim}{b}'\underset{\sim}{P}$ and hence the asymptotic distribution of Y is a univariate normal distribution with mean

$$E_{\underset{\sim}{\theta^*}}(Y) = \underset{\sim}{b}'\underset{\sim}{\pi}(\underset{\sim}{\theta^*})$$

and variance

$$\mathrm{Var}_{\underset{\sim}{\theta^*}}(Y) = \frac{1}{n}\,\underset{\sim}{b}'\,\Sigma\,\underset{\sim}{b}\,.$$

But

$$\underset{\sim}{b}'\underset{\sim}{\pi}(\underset{\sim}{\theta^*}) = \sum_{i=1}^{m+1} b_i\,\pi_i(\underset{\sim}{\theta^*}) = \frac{\sqrt{n}\sum_{i=1}^{m+1}(\beta_{1i}-\beta_{2i})}{\sum_{i=1}^{m+1}\frac{(\beta_{1i}-\beta_{2i})^2}{\pi_i(\underset{\sim}{\theta^*})}} = 0$$

since

$$\sum_{i=1}^{m+1}(\beta_{1i}-\beta_{2i}) = F_1(t_{m+1}) - F_2(t_{m+1}) = 0$$

and

$$\frac{1}{n}\,\underset{\sim}{b}'\,\Sigma\,\underset{\sim}{b} = \frac{n\sum_{i=1}^{m+1}\sum_{j=1}^{m+1}\frac{(\beta_{1i}-\beta_{2i})}{\pi_i(\underset{\sim}{\theta^*})}\,\sigma_{ij}\,\frac{(\beta_{1j}-\beta_{2j})}{\pi_j(\underset{\sim}{\theta^*})}}{\left(\sum_{i=1}^{m+1}\frac{(\beta_{1i}-\beta_{2i})^2}{\pi_i(\underset{\sim}{\theta^*})}\right)^2}$$

$$= \frac{\sum_{i=1}^{m+1} \frac{\pi_i(\underset{\sim}{\theta^*})(1-\pi_i(\underset{\sim}{\theta^*}))}{(\pi_1(\underset{\sim}{\theta^*}))^2} (\beta_{1i}-\beta_{2i})^2}{\left[\sum_{i=1}^{m+1} \frac{(\beta_{1i}-\beta_{2i})^2}{\pi_i(\underset{\sim}{\theta^*})}\right]^2}$$

$$- \frac{\sum_{i\neq j} \pi_i(\underset{\sim}{\theta^*})\, \pi_j(\underset{\sim}{\theta^*}) \frac{(\beta_{1i}-\beta_{2i})}{\pi_i(\underset{\sim}{\theta^*})} \cdot \frac{(\beta_{1j}-\beta_{2j})}{\pi_j(\underset{\sim}{\theta^*})}}{\left[\sum_{i=1}^{m+1} \frac{(\beta_{1i}-\beta_{2i})^2}{\pi_i(\underset{\sim}{\theta^*})}\right]^2}$$

$$= \frac{1}{\sum_{i=1}^{m+1} \frac{(\beta_{1i}-\beta_{2i})^2}{\pi_i(\underset{\sim}{\theta^*})}} - \sum_{i=1}^{m+1}\sum_{j=1}^{m+1} (\beta_{1i}-\beta_{2i})(\beta_{1j}-\beta_{2j})$$

$$= \frac{1}{\sum_{i=1}^{m+1} \frac{(\beta_{1i}-\beta_{2i})^2}{\pi_i(\underset{\sim}{\theta^*})}} \qquad (4.2.9)$$

since

$$\sum_{i=1}^{m+1}\sum_{j=1}^{m+1} (\beta_{1i}-\beta_{2i})(\beta_{1j}-\beta_{2j})$$

$$= \left[\sum_{i=1}^{m+1} (\beta_{1i}-\beta_{2i})\right]^2 = 0 \quad .$$

Hence the asymptotic distribution of $\sqrt{n}\ (Z_1^{(1)} - \theta_1^*)$ is normal with mean zero and variance given by (4.2.10). Thus the theorem is proved.

The following remark is of interest;

Remark: We can show that if the given random sample $X_1, \ldots, X_n$ with common distribution $G_{\underset{\sim}{\theta}}(x) = \theta_1 F_1(x) + (1-\theta_1) F_2(x)$; $0 \leq \theta_1 \leq 1$, is grouped into $m + 1$ intervals with the division points

$$t_0 < t_1 < \ldots < t_m < t_{m+1}$$

where t_0 and t_{m+1} are as defined in Section 3.2, then the Fisher's information in a single observation with respect to such a grouping coincides with the denominator in (4.2.10). This is easy to see since

$$\pi_i(\underset{\sim}{\theta}^*) = G_{\underset{\sim}{\theta}^*}(t_i) - G_{\underset{\sim}{\theta}^*}(t_{i-1}) = \theta_1^* (\beta_{1i}-\beta_{2i}) + \beta_{2i}$$

where, as before, $\beta_{ji} = F_j(t_i) - F_j(t_{i-1})$ for $i = 1, \ldots, m+1$ and $j = 1, 2$.

This interesting coincidence shows that, under the condition of Theorem 4.2.1, $Z_1^{(1)}$ is also asymptotically fully efficient in the sense that its asymptotic variance is minimum with respect to a given set of division points $\{t_i\}_{i=1}^m$.

4.3 Monte Carlo Studies

In this section, we continue our Monte Carlo study concerning the problem of estimating the mixing proportion in a mixture of two normal distributions. This problem was considered in Sections 2.7 and 3.5 of Chapters 2 and 3 respectively.

Consider a mixture $G_{\underset{\sim}{\theta}}(x)$ of two normal distributions $N(0,1)$ and $N(\mu,\sigma^2)$ that is (c.f. (3.5.1))

$$G_{\underset{\sim}{\theta}}(x) = \theta_1 F_1(x) + (1-\theta_1) F_2(x) \qquad 0 \le \theta_1 \le 1 \quad (4.3.1)$$
$$- \infty < x < \infty$$

where

$$F_1(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\{-\tfrac{1}{2}y^2\}dy \qquad - \infty < x < \infty$$

and

$$F_2(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\{-\frac{1}{2\sigma^2}(y-\mu)^2\}dy \qquad - \infty < x < \infty \quad .$$

Choosing the values $t_1, \ldots, t_m$ as in Section 3.5, an initial estimate $\theta_1^{(0)}$ of θ_1 is chosen and by using (4.2.7), successive estimates of θ_1 may be obtained. The value $\theta_1^{(0)}$ can be chosen in two ways. We

can choose an arbitrary value in the interval [0,1] as this initial value. Like most other iteration processes, this arbitrary initial choice should not affect the final result if the number of iterations is sufficiently large. But if only a small number of iterations are taken, the estimate $\hat{\theta}_1^{(0)}$ is chosen to be a consistent (but inefficient) estimator of θ_1. Here, we use the method of moments to find this initial choice. Thus if the distribution of the random variable X is $G_{\underset{\sim}{\theta}}(.)$ given by (4.3.1), then

$$E_{\underset{\sim}{\theta}}(X) = (1-\theta_1)\mu$$

and

$$E_{\underset{\sim}{\theta}}(X^2) = \theta_1 + (1-\theta_1)(\sigma^2+\mu^2)$$

and in any random sample of size n; $X_1, \ldots, X_n$ with common distribution function $G_{\underset{\sim}{\theta}}(.)$ and with realizations $x_1, \ldots, x_n$ respectively, we equate the first sample moment

$$\bar{x} = \frac{1}{n} \sum_{j=1}^{n} x_j$$

to $E_{\underset{\sim}{\theta}}(X)$ to get

$$\hat{\theta}_1^{(0)} = 1 - \frac{\bar{x}}{\mu}$$

whenever $\mu \neq 0$. If, on the other hand $\mu = 0$, then the second sample moment

$$\text{m.s.} = \frac{1}{n} \sum_{j=1}^{n} x_j^2$$

is equated to $E_{\underset{\sim}{\theta}}(X^2)$ to obtain $\hat{\theta}_1^{(0)}$. Thus $\hat{\theta}_1^{(0)}$ is defined as

$$\hat{\theta}_1^{(0)} = 1 - \frac{\bar{x}}{\mu} \qquad \mu \neq 0$$

$$= \frac{\text{m.s.} - \sigma^2}{1 - \sigma^2} \qquad \mu = 0 .$$

Table 4.1

Generalized Least Squares estimator of the mixing proportion, based on the 1-cycle solution of the iteration process, for various mixtures of two normal distributions

				Estimate of θ_1							
				m = 3		m = 10		m = 20		m = 80	
n	$(\mu_2-\mu_1)/\sigma_1$	σ_2/σ_1	θ_1	Mean	MSE	Mean	MSE	Mean	MSE	Mean	MSE
50	0.25	1	0.5	0.651 ± 0.057	0.344	0.599 ± 0.052	0.286	0.575 ± 0.053	0.282	0.578 ± 0.052	0.279
10	0.5	1	0.5	0.535 ± 0.032	0.527	0.505 ± 0.031	0.504	0.537 ± 0.031	0.498	0.549 ± 0.029	0.393
10	1	1	0.5	0.498 ± 0.021	0.216	0.499 ± 0.019	0.198	0.524 ± 0.019	0.179	0.540 ± 0.018	0.152
10	1	1	0.8	0.781 ± 0.015	0.115	0.785 ± 0.015	0.115	0.781 ± 0.015	0.115	0.783 ± 0.015	0.113
20	5	1	0.5	0.496 ± 0.008	0.015	0.498 ± 0.007	0.013	0.498 ± 0.007	0.013	0.489 ± 0.007	0.013
10	0	2	0.5	0.546 ± 0.030	0.453	0.539 ± 0.020	0.218	0.506 ± 0.019	0.197	0.458 ± 0.050	0.179
50	0	2	0.5	0.528 ± 0.023	0.053	0.530 ± 0.020	0.041	0.516 ± 0.019	0.038	0.517 ± 0.019	0.038
10	0.5	2	0.5	0.537 ± 0.031	0.0471	0.533 ± 0.027	0.359	0.527 ± 0.026	0.332	0.525 ± 0.025	0.319

Each case is based on $n_1 = \frac{5000}{n}$ samples of size n. The standard error of each mean is given to indicate the accuracy of the Monte Carlo computation.

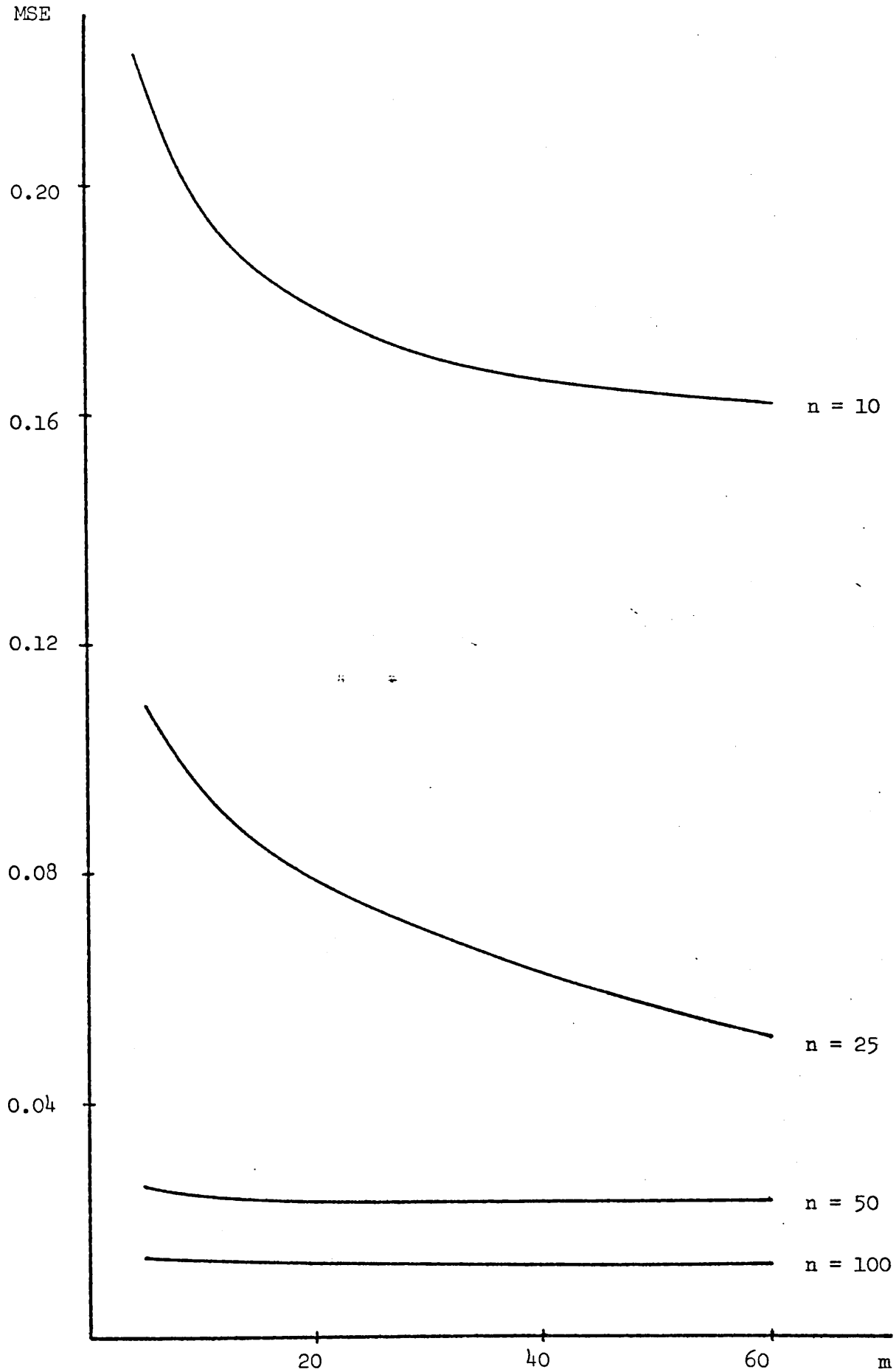


Fig. 4.1 - The Mean-Square-Error of the generalized Least Squares Estimator of the mixing proportion, based on the 1-cycle solution of the iteration process, in a mixture of two normal distributions N(0,1) and N(1,1), against m, the number of division points of the sample space, for different sample sizes.

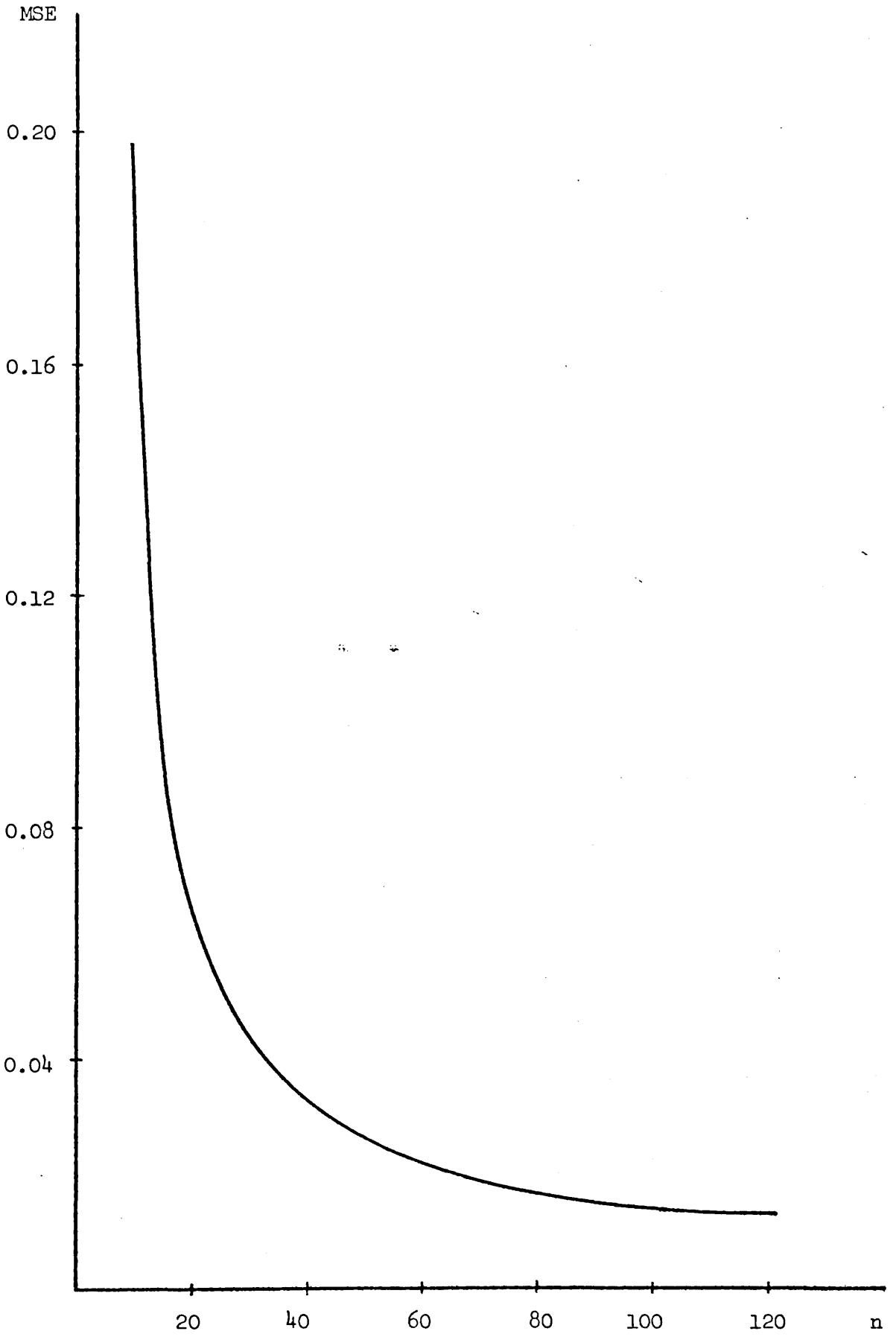


Fig. 4.2 - The Mean-Square-Error of the generalized Least Squares estimator of the mixing proportion, based on the 1-cycle solution of the iteration process, in a mixture of two normal distributions N(0,1) and N(1,1), for varying sample sizes.

Note that when $\mu = 0$, then $\sigma^2 \neq 1$, for otherwise both components of $G_{\underset{\sim}{\theta}}(.)$ become identical which is a contradiction.

In order to see the accuracy of the iteration process suggested in this chapter, a table analogous to that of Table 3.1 was constructed (Table 4.1). Each estimate of θ_1 is based on the solution of a 1-cycle iteration and, as in Sections 2.7 and 3.5, on $n_1 = \frac{5000}{n}$ samples of size n. The mean-square-error of each estimate and the standard error of each mean are calculated as explained in Section 2.7. By comparison, it is seen that although the mean-square-errors in Table 4.1 are generally higher in most cases, a good approximation to the root of (3.3.4) can be obtained by using a 1-cycle solution of the iteration process. The estimates improve as m increases and Figure 4.1 shows how the mean-square-error of $\hat{\theta}_1^{(1)}$ depends on m for different values of n when $\frac{\mu_2-\mu_1}{\sigma_1} = 1$ and $\frac{\sigma_2}{\sigma_1} = 1$. Comparing Figures 4.1 and 3.1, we find that for large sample sizes there is not a substantial difference in the accuracies of both methods. Finally, for the same values of $\frac{\mu_2-\mu_1}{\sigma_1}$ and $\frac{\sigma_2}{\sigma_1}$ and for m = 10, the mean-square-error of $\hat{\theta}_1^{(1)}$ is plotted against n in Figure 4.2 where it is observed that for sample sizes of more than 50, the mean-square-error is very small.

4.4 The Iteration Process in Ungrouped Data

Analogous to Section 3.7, in this section, we let the widths of the intervals $\Delta_i = t_i - t_{i-1}$ $i = 1, \ldots, m + 1$ become progressively finer and eventually consider the situation when $m \to \infty$. We prove that in this manner, the iteration process defined in Section 4.2, converges to the maximum likelihood estimator (MLE) of $\underset{\sim}{\theta}$ as r, the number of iterations, tends to infinity for any fixed sample size. First, we derive the set of equations which yield the MLE of $\underset{\sim}{\theta}$.

Assume that in the mixture of distributions $G_{\underset{\sim}{\theta}}(x)$, given by

(2.2.1), the component distribution functions $F_1(x)$, ..., $F_k(x)$ and hence $G_{\underset{\sim}{\theta}}(x)$ are twice differentiable functions of $x \in \mathcal{X}$ so that the densities $f_1(x)$, ..., $f_k(x)$ and hence $g_{\underset{\sim}{\theta}}(x) = \sum_{j=1}^{k} \theta_j f_j(x)$ of $F_1(x), \ldots, F_k(x)$ and $G_{\underset{\sim}{\theta}}(x) = \sum_{j=1}^{k} \theta_j F_j(x)$ respectively exist and are differentiable with respect to x. Given the observations x_1, ..., x_n, the realizations of random variables X_1, ..., X_n respectively, from the mixture of distributions $G_{\underset{\sim}{\theta}}(x)$, the MLE of $\underset{\sim}{\theta}^*$ is that value of $\underset{\sim}{\theta}$ which maximizes

$$\prod_{i=1}^{n} g_{\underset{\sim}{\theta}}(x_i) = \prod_{i=1}^{n} (\theta_1 f_1(x_i) + \ldots + \theta_k f_k(x_i))$$

or equivalently $\ell n \left(\prod_{i=1}^{n} g_{\underset{\sim}{\theta}}(x_i) \right) = \sum_{i=1}^{n} \ell n \, (\theta_1 f_1(x_i) + \ldots + \theta_k f_k(x_i))$ subject to $\sum_{j=1}^{k} \theta_j = 1$. Here $\ell n(y)$ denotes the natural logarithm of the positive real argument y. Therefore we maximize

$$\Phi(\underset{\sim}{\theta}, \xi) = \sum_{i=1}^{n} \ell n \, (\theta_1 f_1(x_i) + \ldots + \theta_k f_k(x_i)) + \xi \left(\sum_{j=1}^{k} \theta_j - 1 \right)$$

where ξ is the Lagrange multiplier, with respect to θ_1, ..., θ_k and ξ. Setting the derivative of $\Phi(\underset{\sim}{\theta}, \xi)$ with respect to θ_j; $1 \leq j \leq k$ equal to zero gives

$$\sum_{i=1}^{n} \frac{f_j(x_i)}{\theta_1 f_1(x_i) + \ldots + \theta_k f_k(x_i)} + \xi = 0 \qquad (4.4.1)$$

for $j = 1, \ldots, k$.

Multiplying (4.4.1) by θ_j and summing over $j = 1, \ldots, k$, we get

$$n + \sum_{j=1}^{k} \theta_j \xi = 0 \; ;$$

imposing the restriction $\sum_{j=1}^{k} \theta_j = 1$ yields $\xi = -n$ and upon inserting this back into (4.4.1), we have

$$\frac{1}{n} \sum_{i=1}^{n} \frac{f_j(x_i)}{g_{\underset{\sim}{\theta}}(x_i)} = 1 \qquad j = 1, \ldots, k \qquad (4.4.2)$$

whose root gives the MLE of $\underset{\sim}{\theta}^*$. Hereafter, the set of equations (4.4.2) for j =1, ..., k will be referred to as the "likelihood equations".

Theorem 4.4.1: If $\hat{\underset{\sim}{\theta}}_n^{(r)}$ denotes the estimate of $\underset{\sim}{\theta}$ obtained after rth cycle in the iteration process defined in Section 4.2, then for any fixed sample size n, and sufficiently large m

$$\hat{\underset{\sim}{\theta}}_n = (\hat{\theta}_1, \ldots, \hat{\theta}_k)' = \lim_{r\to\infty} \hat{\underset{\sim}{\theta}}_n^{(r)}$$

satisfies the likelihood equations.

Proof: Write

$$\theta_\ell = 1 - (\theta_1 + \theta_2 + \ldots + \theta_{\ell-1} + \theta_{\ell+1} + \ldots + \theta_k) \quad 1 \leq \ell \leq k .$$

Then we have

$$G_{\underset{\sim}{\theta}}(x) = \sum_{j=1}^{k} \theta_j (F_j(x) - F_\ell(x)) + F_\ell(x)$$

and for a given set of partition points $t_1 < t_2 < \ldots < t_m$,

$$\pi_i(\underset{\sim}{\theta}) = G_{\underset{\sim}{\theta}}(t_i) - G_{\underset{\sim}{\theta}}(t_{i-1}) = \sum_{j=1}^{k} \theta_j (\beta_{ji} - \beta_{\ell i}) + \beta_{\ell i} \qquad (4.4.3)$$

where $\beta_{ji} = F_j(t_i) - F_j(t_{i-1})$ for j = 1, ..., k and i = 1, ..., m + 1 with t_0 and t_{m+1} defined as before. According to step (vi) of the iteration process defined in Section 4.2, we minimize (4.2.4) with respect to θ_j; j = 1, ..., k, j $\neq \ell$ to obtain the estimate of $\underset{\sim}{\theta}$ after (r+1) cycles of the iteration. Thus substituting (4.4.3) into (4.2.4) and setting the derivative with respect to θ_j $1 \leq j \leq k$ $j \neq \ell$ equal to zero, we get

$$\begin{cases} \theta_j^{(r+1)} = \dfrac{\sum_{i=1}^{m+1} \dfrac{(p_i-\beta_{\ell i})(\beta_{ji}-\beta_{\ell i})}{\pi_i(\hat{\theta}_{\sim n}^{(r)})}}{\sum_{i=1}^{m+1} \dfrac{(\beta_{ji}-\beta_{\ell i})^2}{\pi_i(\hat{\theta}_{\sim n}^{(r)})}} & j = 1, \ldots, k \quad (4.4.4) \\ & j \neq \ell \\ \hat{\theta}_\ell^{(r+1)} = 1 - \sum_{\substack{j=1 \\ j\neq\ell}}^{k} \hat{\theta}_j^{(r+1)} & r = 0, 1, \ldots \end{cases}$$

Analogous to the steps taken from the equation (4.2.5) to the equation (4.2.7) in the case k = 2, here we can reduce (4.4.4) to

$$\begin{cases} \hat{\theta}_j^{(r+1)} - \hat{\theta}_j^{(r)} = \dfrac{\sum_{i=1}^{m+1} \dfrac{p_i(\beta_{ji}-\beta_{\ell i})}{\pi_i(\hat{\theta}_{\sim n}^{(r)})}}{\sum_{i=1}^{m+1} \dfrac{(\beta_{ji}-\beta_{\ell i})^2}{\pi_i(\hat{\theta}_{\sim n}^{(r)})}} & j = 1, \ldots, k \quad (4.4.5) \\ & j \neq \ell \\ \hat{\theta}_\ell^{(r+1)} = 1 - \sum_{\substack{j=1 \\ j\neq\ell}}^{k} \hat{\theta}_j^{(r+1)} \end{cases}$$

Now, suppose that m is large so that the intervals $\Delta_i = t_i - t_{i-1}$ $i = 1, \ldots, m+1$ are uniformly small. Then since $F_1(.), \ldots, F_k(.)$ and hence $G_{\underset{\sim}{\theta}}(.)$ are twice differentiable,

$$\frac{\beta_{ji}}{\Delta_i} = \frac{F_j(t_i) - F_j(t_{i-1})}{t_i - t_{i-1}} = f_j(t_i) + O(\Delta_i) \quad (4.4.6)$$

for i = 1, ..., m+1

and j = 1, ..., k

and

$$\frac{\pi_i(\underset{\sim}{\theta})}{\Delta_i} = \frac{G_{\underset{\sim}{\theta}}(t_i) - G_{\underset{\sim}{\theta}}(t_{i-1})}{t_i - t_{i-1}} = g_{\underset{\sim}{\theta}}(t_i) + O(\Delta_i)$$

for i = 1, ..., m+1

and hence

$$\sum_{i=1}^{m+1} \frac{p_i(\beta_{ji}-\beta_{\ell i})}{\pi_i(\hat{\theta}_{\sim n}^{(r)})} = \sum_{i=1}^{m+1} \frac{p_i(f_j(t_i) - f_\ell(t_i)) + O(\Delta_i)}{g_{\hat{\underset{\sim}{\theta}}_n^{(r)}}(t_i) + O(\Delta_i)}$$

$$= \sum_{i=1}^{m+1} \left[\frac{p_i (f_j(t_i) - f_\ell(t_i))}{g_{\hat{\theta}_{\sim n}}(r)(t_i)} + O(\Delta_i) \right] . \qquad (4.4.7)$$

Further, if $t_{\alpha_1}, t_{\alpha_2}, \ldots, t_{\alpha_n}$ for some $1 \leq \alpha_1 < \ldots < \alpha_n \leq m$ coincide with the observations $x_1, \ldots, x_n$, we have

$$p_i = G_n(t_i) - G_n(t_{i-1}) = \begin{cases} \frac{1}{n} & \text{for } i = \alpha_1, \ldots, \alpha_n \\ 0 & \text{otherwise.} \end{cases} \qquad (4.4.8)$$

Finally by letting $r \to \infty$ in (4.4.5) and using (4.4.7) and (4.4.8), we see that the left hand side gives zero and for sufficiently large m, $\hat{\theta}_{\sim n}$ is the root of

$$\frac{1}{n} \sum_{i=1}^{n} \frac{f_j(x_i) - f_\ell(x_i)}{g_{\hat{\theta}_{\sim n}}(x_i)} = 0 \qquad j = 1, \ldots, k . \qquad (4.4.9)$$

Multiplying (4.4.9) by θ_j and summing over $j = 1, \ldots, k$, by using $\sum_{j=1}^{k} \theta_j = 1$ we obtain,

$$\frac{1}{n} \sum_{i=1}^{n} \frac{f_\ell(x_i)}{g_{\hat{\theta}_{\sim n}}(x_i)} = 1 \qquad 1 \leq \ell \leq k \qquad (4.4.10)$$

which is the ℓth equation in the system of the likelihood equations with root $\hat{\theta}_{\sim n} = (\hat{\theta}_1, \ldots, \hat{\theta}_k)'$ given by (4.4.2). As ℓ takes integers between one through to k, the whole system of the likelihood equations is obtained. This completes the proof of the theorem.

It was stated in step (vi) of the iteration process suggested in Section 4.2 that, after the rth cycle of the iteration process $r = 0, 1, \ldots,$ to impose the constraint $\sum_{j=1}^{k} \theta_j = 1$, we pick θ_ℓ amongst $\theta_1, \ldots, \theta_k$ for some $1 \leq \ell \leq k$ and substitute $\theta_\ell = 1 - \sum_{\substack{j=1 \\ j \neq \ell}}^{k} \theta_j$ in $\Phi_r(\underset{\sim}{\theta})$ given by (4.2.2) or equivalently (4.2.4). We then minimize

$\Phi_r(\underset{\sim}{\theta})$ with respect to $\theta_1, \ldots, \theta_{\ell-1}, \theta_{\ell+1}, \ldots, \theta_k$ and call the minimizing values $\hat{\theta}_1^{(r+1)}, \ldots, \hat{\theta}_{\ell-1}^{(r+1)}, \hat{\theta}_{\ell+1}^{(r+1)}, \ldots, \hat{\theta}_k^{(r+1)}$ so that with $\hat{\theta}_\ell^{(r+1)} = 1 - \sum_{\substack{j=1 \\ j\neq\ell}}^{k} \hat{\theta}_j^{(r+1)}$, the estimate of $\underset{\sim}{\theta}$ after r + 1 cycles of the iteration is formed as $\hat{\underset{\sim}{\theta}}_n^{(r+1)} = (\hat{\theta}_1^{(r+1)}, \ldots, \hat{\theta}_k^{(r+1)})'$. This choice of θ_ℓ is of course arbitrary and is made according to the convenience of the experimenter. It may, however, be argued that by merely choosing an arbitrary θ_ℓ for some $1 \leq \ell \leq k$, and substituting $\theta_\ell = 1 - \sum_{\substack{j=1 \\ j\neq\ell}}^{k} \theta_j$, we are taking a non-symmetric approach even though the constraint $\sum_{j=1}^{k} \theta_j = 1$ is imposed. For this reason, we use the Lagrange multiplier technique in the minimization of $\Phi_r(\underset{\sim}{\theta})$. Therefore we minimize

$$\Phi_r(\underset{\sim}{\theta},\xi) = (\underset{\sim}{G}_n - A\underset{\sim}{\theta})' \, C_r^{-1} \, (\underset{\sim}{G}_n - A\underset{\sim}{\theta}) - 2\xi(\underset{\sim}{1}'\underset{\sim}{\theta} - 1) \qquad (4.4.11)$$

where ξ is the Lagrange multiplier and $\underset{\sim}{1}$ is the k-dimensional vector of 1's i.e. $\underset{\sim}{1} = (1, \ldots, 1)'$, with respect to $\underset{\sim}{\theta}$ and ξ. Setting the derivative with respect to $\underset{\sim}{\theta}$, of (4.4.11) equal to zero gives for the minimizing values $\hat{\underset{\sim}{\theta}}_n^{(r+1)}$ and ξ^*,

$$(A'C_r^{-1}A) \, \hat{\underset{\sim}{\theta}}_n^{(r+1)} - (A'C_r^{-1}\underset{\sim}{G}_n) - \xi^*\underset{\sim}{1} = 0 \, .$$

Since A has rank k, $(A'C_r^{-1}A)$ is of rank k and thus invertible. Therefore

$$\hat{\underset{\sim}{\theta}}_n^{(r+1)} = (A'C_r^{-1}A)^{-1} \, A'C_r^{-1}\underset{\sim}{G}_n + \xi^* \, (A'C_r^{-1}A)^{-1}\underset{\sim}{1} \qquad (4.4.12)$$

and since C_r is independent of $\hat{\underset{\sim}{\theta}}_n^{(r+1)}$, imposing the constraint $\underset{\sim}{1}'\hat{\underset{\sim}{\theta}}_n^{(r+1)} = 1$, we have

$$1 = \underset{\sim}{1}'\hat{\underset{\sim}{\theta}}_n^{(r+1)} = \underset{\sim}{1}'(A'C_r^{-1}A)^{-1} \, A'C_r^{-1}\underset{\sim}{G}_n + \xi^*\underset{\sim}{1}'(A'C_r^{-1}A)^{-1}\underset{\sim}{1}$$

so that

$$\xi^* = \frac{1 - \underset{\sim}{1}'(A'C_r^{-1}A)^{-1} \, A'C_r^{-1}\underset{\sim}{G}_n}{\underset{\sim}{1}'(A'C_r^{-1}A)^{-1}\underset{\sim}{1}}$$

where we note that $\underset{\sim}{1}'(A'C_r^{-1}A)^{-1}\underset{\sim}{1}$ is only a scalar factor. Inserting ξ^* back into (4.4.12) yields

$$\hat{\underset{\sim}{\theta}}_n^{(r+1)} = (A'C_r^{-1}A)^{-1}A'C_r^{-1}\underset{\sim}{G}_n + \left\{\frac{1 - \underset{\sim}{1}'(A'C_r^{-1}A)^{-1}A'C_r^{-1}\underset{\sim}{G}_n}{\underset{\sim}{1}'(A'C_r^{-1}A)^{-1}\underset{\sim}{1}}\right\}(A'C_r^{-1}A)^{-1}\underset{\sim}{1}$$

for $r = 0, 1, 2, \ldots$ (4.4.13)

It is to be stressed that both approaches yield the same result i.e. (4.4.13) gives the MLE of $\underset{\sim}{\theta}$ for large m as $r \to \infty$. It is known that if a function is to be minimized subject to certain restrictions, one can use the Lagrange multiplier technique or equivalently, one can solve the restrictions and substitute in the function and proceed with the minimization of the function. The final result is not altered whichever method is used. Here, we verify this fact for the case k = 2.

Proposition 4.4.1: If in the mixture of distributions $G_{\underset{\sim}{\theta}}(x)$, given by (2.2.1), k = 2, then for any fixed number n of observations $x_1, \ldots, x_n$ from this distribution and for sufficiently large m,

$$\hat{\underset{\sim}{\theta}}_n = (\hat{\theta}_1, \hat{\theta}_2)' = \lim_{r\to\infty} \hat{\underset{\sim}{\theta}}_n^{(r)}$$

is the maximum likelihood estimator of $\underset{\sim}{\theta} = (\theta_1, \theta_2)'$, the unknown vector of the mixing proportions. Here $\hat{\underset{\sim}{\theta}}_n^{(r)} = (\hat{\theta}_1^{(r)}, \hat{\theta}_2^{(r)})'$ $r = 1, 2, \ldots$ is given by (4.4.13) and $\hat{\underset{\sim}{\theta}}_n^{(0)} = (\hat{\theta}_1^{(0)}, \hat{\theta}_2^{(0)})'$ is chosen so that $0 \leq \hat{\theta}_1^{(0)} \leq 1$ and $\hat{\theta}_2^{(0)} = 1 - \hat{\theta}_1^{(0)}$.

Proof: Put $R = (A'C_r^{-1}A)$. Then R is a symmetric 2×2 matrix and by substituting A from (3.2.2) with k = 2 and C_r^{-1} as defined in the step (iv) of the iteration process given in Section 4.2, we see that if

$$R = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}$$

then

$$R_{11} = \sum_{i=1}^{m+1} \frac{\beta_{1i}^2}{\pi_i(\hat{\underset{\sim}{\theta}}_n^{(r)})}$$

$$R_{12} = R_{21} = \sum_{i=1}^{m+1} \frac{\beta_{1i}\,\beta_{2i}}{\pi_i(\hat{\underset{\sim}{\theta}}_n^{(r)})}$$

$$R_{22} = \sum_{i=1}^{m+1} \frac{\beta_{2i}^2}{\pi_i(\hat{\underset{\sim}{\theta}}_n^{(r)})}$$

where, as before, $\beta_{ji} = F_j(t_i) - F_j(t_{i-1})$ for $i = 1, \ldots, m+1$ and $j = 1, 2$ and $\pi_i(\underset{\sim}{\theta}) = G_{\underset{\sim}{\theta}}(t_i) - G_{\underset{\sim}{\theta}}(t_{i-1})$ for $i = 1, \ldots, m+1$. Now,

$$D = \det(R) = R_{11}\,R_{22} - R_{12}^2$$

$$= \left[\sum_{i=1}^{m+1} \frac{\beta_{1i}^2}{\pi_i(\hat{\underset{\sim}{\theta}}_n^{(r)})}\right]\left[\sum_{i=1}^{m+1} \frac{\beta_{2i}^2}{\pi_i(\hat{\underset{\sim}{\theta}}_n^{(r)})}\right] - \left[\sum_{i=1}^{m+1} \frac{\beta_{1i}\,\beta_{2i}}{\pi_i(\hat{\underset{\sim}{\theta}}_n^{(r)})}\right]^2$$

and by using the Cauchy-Schwartz inequality, $D > 0$. Hence R^{-1} exists and

$$R^{-1} = (A'C_r^{-1}A)^{-1} = \frac{1}{D}\begin{bmatrix} R_{22} & -R_{12} \\ -R_{12} & R_{11} \end{bmatrix}.$$

Let $Q = A'C_r^{-1}G_{\underset{\sim}{n}} = \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix}$, then in (4.4.13), by substituting for R^{-1} and Q, we have

$$\begin{bmatrix} \hat{\theta}_1^{(r+1)} \\ \hat{\theta}_2^{(r+1)} \end{bmatrix} = \frac{1}{D}\begin{bmatrix} R_{22} & -R_{12} \\ -R_{12} & R_{11} \end{bmatrix}\left\{\begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} + \frac{R_{11}R_{22}-R_{12}^2-(R_{22}Q_1-R_{12}Q_2+R_{11}Q_2-R_{12}Q_1)}{R_{11}-2R_{12}+R_{22}}\begin{bmatrix} 1 \\ 1 \end{bmatrix}\right\}$$

$$r = 0, 1, 2, \ldots$$

which after some algebra leads to

$$\hat{\theta}_1^{(r+1)} = \frac{(Q_1-Q_2) + (R_{22}-R_{12})}{R_{11} - 2R_{12} + R_{22}} = \frac{\sum_{i=1}^{m+1} \frac{(p_i-\beta_{2i})(\beta_{1i}-\beta_{2i})}{\pi_i(\hat{\underset{\sim}{\theta}}_n^{(r)})}}{\sum_{i=1}^{m+1} \frac{(\beta_{1i}-\beta_{2i})^2}{\pi_i(\hat{\underset{\sim}{\theta}}_n^{(r)})}} \qquad (4.4.14)$$

$$r = 0, 1, \ldots$$

and

$$\hat{\theta}_2^{(r+1)} = \frac{(Q_2-Q_1) + (R_{11}-R_{12})}{R_{11} - 2R_{12} + R_{22}} = 1 - \hat{\theta}_1^{(r+1)} \quad \text{for } r = 0, 1, \ldots \qquad (4.4.15)$$

We now note that (4.4.14) is identical to (4.2.5) and therefore by a proof parallel to that of Theorem 4.4.1, the proposition can be established.

4.5 Conclusions

The iteration process introduced in this chapter shows that in practice, a reliable estimate of the vector of the unknown mixing proportions $\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_k)' \in \Theta$ may be obtained in a very simple way. It is believed that to obtain a relatively efficient estimate, even a few iterations are sufficient in moderate sample sizes, provided that the iteration is started with a consistent, but inefficient estimator of $\underset{\sim}{\theta}$. The interesting result of Theorem 4.2.1 supports this belief.

Further, we have seen that as we increase the number of intervals in the grouping, our estimate, obtained by the iteration process, approaches the maximum likelihood estimator of $\underset{\sim}{\theta}$, as the number of iterations are increased. The convergence of the iteration process to the maximum likelihood estimator of $\underset{\sim}{\theta}$ when the widths of the intervals become very small is of particular interest. The latter class of estimates, in general, play an important role in mathematical statistics. Applications and special properties of this class in the framework of problem of mixtures of distributions form the subject of the next chapter.

CHAPTER 5

MAXIMUM LIKELIHOOD ESTIMATION

5.1 Introduction

Maximum Likelihood estimates, in general, form an important class of estimates in the theory of point estimation. It is well-known (Cramér [15]) that under very mild regularity conditions, maximum likelihood estimators of the unknown parameters in a distribution are CAN and asymptotically fully efficient. Given a random sample $X_1,\ldots,X_n$ with realizations $x_1,\ldots,x_n$ respectively, from a population with probability distribution indexed by an unknown vector of parameters $\underset{\sim}{\alpha}$, the likelihood function L is defined by

$$L(\underset{\sim}{\alpha}: x_1,\ldots,x_n) = \prod_{i=1}^{n} f_{\underset{\sim}{\alpha}}(x_i)$$

where $f_{\underset{\sim}{\alpha}}(x)$ is the common probability density function of the random variables $X_1,\ldots,X_n$. The method of maximum likelihood consists in choosing, as an estimate of the unknown population value of $\underset{\sim}{\alpha}$, the particular value that renders L, or equivalently $\ell n\,(L)$, as great as possible.

Unfortunately, maximization of the likelihood function often leads to some intractable set of equations and indeed estimation problems concerned with mixtures of distributions are no exception. The problem of maximum likelihood estimation of the unknown parameters in a mixture of distributions has been considered by several authors and the reader is referred to Chapter 1 for a survey of the relevant literature. The results of the previous papers, although interesting, are mostly empirical and based on numerical studies and thus lack a theoretical justification. Further, they consider specific cases and in particular mixtures of distributions

consisting of components which are the distribution functions of normally distributed random variables with unknown means and unknown variances. In such cases, as noted by Day [17], Behboodian [5] and Fryer and Robertson [21], unless sufficient conditions are imposed on the variances of the components (e.g. equality), each sample point generates a singularity in the likelihood function. This can be seen by considering the likelihood function generated by observations $x_1,\ldots,x_n$ from a population with density function $g_{\underset{\sim}{\theta}}(x)$ given by

$$g_{\underset{\sim}{\theta}}(x) = \sum_{i=1}^{2} \theta_i \, (2\pi\sigma_i^2)^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}\left[\frac{x-\mu_i}{\sigma_i}\right]^2\right\}$$

for $-\infty < x < +\infty$ with $\theta_1 + \theta_2 = 1$, and where $\underset{\sim}{\theta} = (\theta_1, \theta_2)'$, μ_i is the mean, σ_i^2 is the variance and θ_i is the mixing proportion of the ith component in the mixture for $i = 1,2$. Denote the likelihood function by $L(\theta_1, \mu_1, \mu_2, \sigma_1, \sigma_2)$; then

$$L(\theta, x_i, \mu_2, 0, \sigma_2) = L(\theta, \mu_1, x_i, \sigma_1, 0) = \infty$$

for $i = 1,..,n$. Hence the method of maximum likelihood estimation clearly breaks down in this case. In view of the fact that the likelihood function is unbounded for this problem, Behboodian [5] has proposed using the value corresponding to the largest stationary maximum of the likelihood function as an estimate. However, when the mixing proportions are the only unknown parameters, as mentioned by Hill [26] and Macdonald [35], the likelihood function is a concave function of the parameters and therefore it has at most one relative maxima.

In this chapter, we consider a mixture $G_{\underset{\sim}{\theta}}(.)$ of two known distribution functions $F_1(.)$ and $F_2(.)$ with mixing proportions $\theta_1 = \theta$,

$0 \leq \theta \leq 1$ and $\theta_2 = 1-\theta$ respectively. Thus

$$G_{\underset{\sim}{\theta}}(x) = \theta F_1(x) + (1-\theta) F_2(x) \qquad 0 \leq \theta \leq 1 \qquad x \in \mathfrak{X}$$

where $\underset{\sim}{\theta} = (\theta_1, \theta_2)'$ and we note that the mixing proportion θ is the only unknown parameter. It is assumed that there exists a σ-finite measure μ on the Borel sets of $\mathfrak{X}$ dominating $F_1(.)$ and $F_2(.)$ and hence also $G_{\underset{\sim}{\theta}}(.)$ so that by the Radon-Nikodym theorem there exists densities $f_j(.)$ $j = 1,2$ and $g_{\underset{\sim}{\theta}}(.)$ so that

$$g_{\underset{\sim}{\theta}}(x) = \theta f_1(x) + (1-\theta) f_2(x) \qquad 0 \leq \theta \leq 1 \qquad x \in \mathfrak{X} . \tag{5.1.1}$$

Thus if A is any Borel subset of $\mathfrak{X}$,

$$P_{\underset{\sim}{\theta}}(X \in A) = \int_A g_{\underset{\sim}{\theta}}(x)\, d\mu$$

and

$$P_{\underset{\sim}{e}_j}(X \in A) = \int_A f_j(x)\, d\mu \qquad j = 1,2,$$

where $P_{\underset{\sim}{\theta}}$ and $P_{\underset{\sim}{e}_j}$ $j = 1,2$ denote the probability measures corresponding to the distribution functions $G_{\underset{\sim}{\theta}}(.)$ and $F_j(.)$ $j = 1,2$ respectively.

We shall see that the regularity conditions under which a maximum likelihood estimator possesses the well-known asymptotic properties (Cramér [15]), are satisfied by $g_{\underset{\sim}{\theta}}(.)$ for every $\theta \in (0,1)$. Thus confining θ to the interval (0,1), we examine the likelihood equation from a somewhat more theoretical point of view and give sufficient conditions for the existence of a unique root of the likelihood equation in the interval (0,1) with probability approaching unity as the sample size increases. As mentioned above, an analytic solution of the likelihood equation seems unobtainable and we are naturally led to the consideration of the iterative solutions of the

likelihood equation. We shall use the well-known Fisher's scoring method of iteration and establish the properties of the solutions given by the first and the subsequent iterations.

In the sequel, the following partition of the sample space $\mathfrak{X}$ will be used; let S_1 and S_2 be those subsets of $\mathfrak{X}$ such that for every $x \in S_1$, $f_1(x)$ exceeds $f_2(x)$ and similarly for every $x \in S_2$, $f_2(x)$ exceeds $f_1(x)$. Therefore

$$S_1 = \{x : x \in \mathfrak{X} , f_1(x) > f_2(x)\} , \qquad (5.1.2)$$

$$S_2 = \{x : x \in \mathfrak{X} , f_1(x) < f_2(x)\} \qquad (5.1.3)$$

and hence S_1 and S_2 are disjoint subsets of $\mathfrak{X}$, i.e.

$$S_1 \cap S_2 = \phi \qquad \text{(empty set)}$$

and further

$$S_1 \cup S_2 = \mathfrak{X} - \{x : x \in \mathfrak{X}, f_1(x) = f_2(x)\} . \qquad (5.1.4)$$

5.2 Statement of the Problem and Existence of a Solution

Given n independently and identically distributed random variables $X_1,\ldots,X_n$ with realizations $x_1,\ldots,x_n$ respectively and with common density function

$$g_{\underset{\sim}{\theta}}(x) = \theta\, f_1(x) + (1-\theta)\, f_2(x) \qquad 0 \le \theta \le 1$$

the maximum likelihood estimator (MLE) of θ^* (the true value of the parameter θ) is desired. The likelihood function based on the given sample

$$L(\theta: x_1,\ldots,x_n) = \prod_{i=1}^{n} g_{\underset{\sim}{\theta}}(x_i) = \prod_{i=1}^{n} [\theta\, f_1(x_i) + (1-\theta)\, f_2(x_i)]$$

is a function of θ only and it is positive differentiable for all $0 \leq \theta \leq 1$. Thus the MLE of θ^* is the solution of

$$\frac{\partial L}{\partial \theta} = 0$$

or equivalently

$$\frac{\partial \ \ell n(L)}{\partial \theta} = 0$$

where $\ell n(y)$ denotes the natural logarithm of the real positive argument y (Recall that logarithm is a monotone function). Then

$$\ell n(L(\theta: x_1,\ldots,x_n)) = \sum_{i=1}^{n} \ell n \ [\theta \ f_1(x_i) + (1-\theta) \ f_2(x_i)]$$

and upon setting the derivative with respect to θ equal to zero, we obtain

$$\psi(\theta) = \sum_{i=1}^{n} \frac{f_1(x_i) - f_2(x_i)}{\theta f_1(x_i) + (1-\theta) \ f_2(x_i)} = 0 \qquad (5.2.1)$$

whose root constitutes the MLE of θ^*. Hereafter, the equation (5.2.1) will be referred to as the "likelihood equation". We further denote by $\Psi(\theta)$ the random function of θ whose realized value is $\psi(\theta)$ and therefore

$$\Psi(\theta) = \frac{1}{n} \sum_{i=1}^{n} \frac{f_1(X_i) - f_2(X_i)}{\theta f_1(X_i) + (1-\theta) \ f_2(X_i)} \ . \qquad (5.2.2)$$

Note also that the Fisher's information function in a single observation is given by

$$I(\theta) = E_{\underset{\sim}{\theta}}\left[\frac{\partial}{\partial \theta} \ \ell n \ (g_{\underset{\sim}{\theta}}(X))\right]^2 = E_{\underset{\sim}{\theta}}\left[\frac{f_1(X) - f_2(X)}{g_{\underset{\sim}{\theta}}(X)}\right]^2$$

$$= \int \frac{(f_1(x) - f_2(x))^2}{g_{\underset{\sim}{\theta}}(x)} \ d\mu(x) \qquad (5.2.3)$$

so that $\frac{1}{n\ I(\theta^*)}$ is the conventional Cramér-Rao lower bound for a sample of size n and thus the variance of any unbiased estimator of θ^* on this sample is at least $\frac{1}{n\ I(\theta^*)}$.

It is shown in Cramér [15] that if the following conditions are satisfied:

(i) For almost all $x \in \mathfrak{X}$, the derivatives $\frac{\partial}{\partial\theta} \ell n(g_{\underset{\sim}{\theta}}(x))$, $\frac{\partial^2}{\partial\theta^2} \ell n(g_{\underset{\sim}{\theta}}(x))$ and $\frac{\partial^3}{\partial\theta^3} \ell n(g_{\underset{\sim}{\theta}}(x))$ exist for every $\underset{\sim}{\theta}$ belonging to a non-degenerate interval of R.

(ii) For every θ for which (i) is satisfied, we have

$$\left|\frac{\partial}{\partial\theta} g_{\underset{\sim}{\theta}}(x)\right| < A_1(x) , \qquad \left|\frac{\partial^2}{\partial\theta^2} g_{\underset{\sim}{\theta}}(x)\right| < A_2(x) \text{ and}$$

$\left|\frac{\partial^3}{\partial\theta^3} \ell n\ (g_{\underset{\sim}{\theta}}(x))\right| < A_3(x)$, the functions $A_1(x)$ and $A_2(x)$ being integrable over $\mathfrak{X}$ while $E_{\underset{\sim}{\theta}}[A_3(X)] < M$, where M is independent of θ,

(iii) For every θ for which (i) is satisfied,

$I(\theta) = E_{\underset{\sim}{\theta}}[\frac{\partial}{\partial\theta} \ell n\ (g_{\underset{\sim}{\theta}}(X))]^2$ is finite and positive,

then the likelihood equation has a solution which converges in probability to θ^* as $n \to \infty$. This solution is an asymptotically normal and asymptotically efficient estimate of θ^*.

Proposition 5.2.1. The density function $g_{\underset{\sim}{\theta}}(x)$ given by (5.1.1) satisfies the regularity conditions (i), (ii) and (iii) for all values of θ satisfying $0 < \theta < 1$.

Proof. From (5.1.1),

$$\ell n\ (g_{\underset{\sim}{\theta}}(x)) = \ell n\ (\theta f_1(x) + (1-\theta)\ f_2(x))$$

where $x \in \mathfrak{X}$ and $0 < \theta < 1$. Thus we have

$$\frac{\partial}{\partial\theta} \ell n\ (g_{\underset{\sim}{\theta}}(x)) = \frac{f_1(x) - f_2(x)}{\theta f_1(x) + (1-\theta)\ f_2(x)}\ ,$$

$$\frac{\partial^2}{\partial\theta^2} \ell n\ (g_{\underset{\sim}{\theta}}(x)) = -\left(\frac{f_1(x) - f_2(x)}{\theta f_1(x) + (1-\theta)\ f_2(x)}\right)^2\ ,$$

and

$$\frac{\partial^3}{\partial\theta^3} \ell n\ (g_{\underset{\sim}{\theta}}(x)) = 2\left(\frac{f_1(x) - f_2(x)}{\theta f_1(x) + (1-\theta)\ f_2(x)}\right)^3$$

which clearly show that the first three partial derivatives of $\ell n\ (g_{\underset{\sim}{\theta}}(x))$ exist for every $\theta \in (0,1)$. Further,

$$\left|\frac{\partial}{\partial\theta}\ g_{\underset{\sim}{\theta}}(x)\right| = |f_1(x) - f_2(x)| < f_1(x) + f_2(x)$$

and

$$\left|\frac{\partial^2}{\partial\theta^2}\ g_{\underset{\sim}{\theta}}(x)\right| = 0$$

and therefore the first and the second partial derivatives of $g_{\underset{\sim}{\theta}}(x)$ are bounded by integrable functions for every $\theta \in (0,1)$. Now

$$\left|\frac{\partial^3}{\partial\theta^3} \ell n\ (g_{\underset{\sim}{\theta}}(x))\right| = 2\left|\frac{f_1(x) - f_2(x)}{\theta f_1(x) + (1-\theta)\ f_2(x)}\right|^3 <$$

$$\begin{cases} \left(\dfrac{f_1(x) - f_2(x)}{\theta f_1(x) + (1-\theta)\ f_2(x)}\right)^3 & \text{for every } x \in S_1 \\ \left(\dfrac{f_2(x) - f_1(x)}{\theta f_1(x) + (1-\theta)\ f_2(x)}\right)^3 & \text{for every } x \in S_2 \\ 0 & \text{for every } x \in \mathfrak{X} - S_1 \cup S_2 \end{cases}$$

$$= \begin{cases} \frac{1}{\theta^2} \dfrac{(f_1(x) - f_2(x))^3}{(\theta f_1(x) + (1-\theta)f_2(x))\,(f_1(x) + \frac{(1-\theta)}{\theta} f_2(x))^2} & \text{for every } x \in S_1 \\ \frac{1}{(1-\theta)^2} \dfrac{(f_2(x) - f_1(x))^3}{(\theta f_1(x) + (1-\theta)f_2(x))\,(\frac{\theta}{(1-\theta)} f_1(x) + f_2(x))^2} & \text{for every } x \in S_2 \\ 0 & \text{for every } x \in \mathfrak{X} - S_1 \cup S_2 \end{cases}$$

$$\leq A_3(x) = \begin{cases} \frac{1}{\theta^2} \dfrac{(f_1(x) - f_2(x))^3}{(\theta f_1(x) + (1-\theta)f_2(x))\,(f_1(x))^2} & \text{for every } x \in S_1 \\ \frac{1}{(1-\theta)^2} \dfrac{(f_2(x) - f_1(x))^3}{(\theta f_1(x) + (1-\theta)f_2(x))\,(f_2(x))^2} & \text{for every } x \in S_2 \\ 0 & \text{for every } x \in \mathfrak{X} - S_1 \cup S_2 \end{cases}$$

and so

$$E_{\underset{\sim}{\theta}}[A_3(x)] = \int A_3(x) g_{\underset{\sim}{\theta}}(x)d\mu = \int_{S_1} A_3(x)\, g_{\underset{\sim}{\theta}}(x)d\mu + \int_{S_2} A_3(x)\, g_{\underset{\sim}{\theta}}(x)d\mu$$

$$= \frac{1}{\theta^2} \int_{S_1} \frac{(f_1(x) - f_2(x))^3}{f_1^2(x)}\, d\mu + \frac{1}{(1-\theta)^2} \int_{S_2} \frac{(f_2(x) - f_1(x))^3}{f_2^2(x)}\, d\mu$$

$$< \frac{1}{\theta^2} \int_{S_1} f_1(x)\, d\mu + \frac{1}{(1-\theta)^2} \int_{S_2} f_2(x)\, d\mu$$

$$= \frac{1}{\theta^2} P_{\underset{\sim}{e}_1}(X \in S_1) + \frac{1}{(1-\theta)^2} P_{\underset{\sim}{e}_2}(X \in S_2)$$

where $P_{\underset{\sim}{e}_1}$ and $P_{\underset{\sim}{e}_2}$ are probability measures corresponding to the distribution functions $F_1(x)$ and $F_2(x)$ respectively. Now, for any θ contained in the interval $(0,1)$, the coefficients of $P_{\underset{\sim}{e}_1}(X \in S_1)$ and $P_{\underset{\sim}{e}_2}(X \in S_2)$ could be held bounded and thus (ii) is also verified.

To establish (iii), note that $I(\theta) = \int \frac{(f_1(x) - f_2(x))^2}{g_\theta(x)} d\mu$ and thus $I(\theta) = 0$ implies $f_1(x) = f_2(x)$ almost everywhere μ. Further,

$$I(\theta) = \int \frac{f_1^2(x)}{g_\theta(x)} d\mu - 2\int \frac{f_1(x)f_2(x)}{g_\theta(x)} d\mu + \int \frac{f_2^2(x)}{g_\theta(x)} d\mu$$

$$= \int (f_1(x) - f_2(x)) \frac{f_1(x)}{g_\theta(x)} d\mu + \int (f_2(x) - f_1(x)) \frac{f_2(x)}{g_\theta(x)} d\mu$$

$$\leq \int_{S_1} (f_1(x) - f_2(x)) \frac{f_1(x)}{g_\theta(x)} d\mu + \int_{S_2} (f_2(x) - f_1(x)) \frac{f_2(x)}{g_\theta(x)} d\mu$$

$$= \frac{1}{\theta} \int_{S_1} (f_1(x) - f_2(x)) \frac{f_1(x)}{f_1(x) + \frac{(1-\theta)}{\theta} f_2(x)} d\mu$$

$$+ \frac{1}{1-\theta} \int_{S_2} (f_2(x) - f_1(x)) \frac{f_2(x)}{\frac{\theta}{1-\theta} f_1(x) + f_2(x)} d\mu$$

$$\leq \frac{1}{\theta} \int_{S_1} (f_1(x) - f_2(x)) d\mu + \frac{1}{1-\theta} \int_{S_2} (f_2(x) - f_1(x)) d\mu$$

$$= \frac{1}{\theta} [P_{e_1}(X \in S_1) - P_{e_2}(X \in S_1)] + \frac{1}{1-\theta} [P_{e_2}(X \in S_2) - P_{e_1}(X \in S_2)]$$

$$< \infty \text{ for every } 0 < \theta < 1$$

completing the proof of the proposition.

In the light of the Proposition 5.2.1, in the rest of this chapter, we shall assume that the parameter θ is confined to the interval $(0,1)$ unless otherwise stated. It is noted, however, that if $\theta = 0$ or $\theta = 1$, then $I(\theta)$ can become infinite and similarly $E_\theta[\frac{\partial^3}{\partial\theta^3} \ln (g_\theta(x))]$ can become unbounded.

5.3 Properties of the Information Function

The information function $I(\theta)$, given by (5.2.3), has certain interesting properties. These properties will be useful in the analysis of the likelihood equation and will be discussed here. Hill

[26] has obtained Power series expansions for $I(\theta)$ when $f_1(x)$ and $f_2(x)$ are both density functions of (i) negative exponential distributions and (ii) normal distributions.

Now,

$$I(\theta) = \int \frac{(f_1(x) - f_2(x))^2}{g_{\underset{\sim}{\theta}}(x)} d\mu = -\frac{1}{\theta(1-\theta)} \int \frac{(g_{\underset{\sim}{\theta}}(x) - f_1(x))\,(g_{\underset{\sim}{\theta}}(x) - f_2(x))}{g_{\underset{\sim}{\theta}}(x)} d\mu$$

$$= \frac{1}{\theta(1-\theta)} \left[1 - \int \frac{f_1(x) f_2(x)}{g_{\underset{\sim}{\theta}}(x)} d\mu \right]$$

and due to positivity of $I(\theta)$, proved in proposition (5.2.1), we have

$$0 \leq \int \frac{f_1(x) f_2(x)}{g_{\underset{\sim}{\theta}}(x)} d\mu \leq 1 .$$

Since $\frac{1}{\theta(1-\theta)}$ is the information function for θ in a pure binomial situation, we see that the additional uncertainty as to which of the two populations, in the mixture of distributions $g_{\underset{\sim}{\theta}}(x)$, an observation comes from, is reflected in the factor

$$\left[1 - \int \frac{f_1(x) f_2(x)}{g_{\underset{\sim}{\theta}}(x)} d\mu \right] .$$

If the densities $f_1(x)$ and $f_2(x)$ do not overlap, then we obtain the full binomial information, while if they are identical, the information is zero. This clearly indicates that unless the densities are rather well-separated, it will take an extremely large sample size to get a reasonably precise estimate of θ.

Proposition 5.3.1. The information function $I(\theta)$ is infinitely differentiable under the integral sign.

Proof. Recall from the Lebesgue dominated convergence theorem that for any function $R_\theta(x)$ depending on x and the parameter θ,

$$\frac{\partial^r}{\partial\theta^r}\int R_\theta(x)\,d\mu\Big|_{\theta=\theta_0} = \int\frac{\partial^r}{\partial\theta^r}R_\theta(x)\Big|_{\theta=\theta_0} d\mu$$

for $r = 1,2,\ldots$ if $\frac{\partial^r}{\partial\theta^r} R_\theta(x)$ exists at $\theta = \theta_0$ and if there exists an integrable function $H(x)$ such that

$$\left|\frac{\partial^r}{\partial\theta^r} R_\theta(x)\right| \leq H(x)$$

almost everywhere μ, for every θ belonging to some neighbourhood of θ_0.

Now clearly, $\frac{\partial^r}{\partial\theta^r}\left[\frac{(f_1(x) - f_2(x))^2}{g(x)}\right]$ exists and

$$\left|\frac{\partial^r}{\partial\theta^r}\left[\frac{(f_1(x) - f_2(x))^2}{g_\theta(x)}\right]\Big|_{\theta=\theta_0}\right| = \left|\frac{(-1)^r\, r!\,(f_1(x) - f_2(x))^{r+2}}{g^{r+1}_{\underset{\sim}{\theta}_0}(x)}\right|$$

where $\underset{\sim}{\theta}_0 = (\theta_0, 1-\theta_0)'$ for $r = 1,2,\ldots$

$$= \begin{cases} r!\,\dfrac{(f_1(x) - f_2(x))^{r+2}}{g^{r+1}_{\underset{\sim}{\theta}_0}(x)} & \text{for every } x \in S_1 \\[2ex] r!\,\dfrac{(f_2(x) - f_1(x))^{r+2}}{g^{r+1}_{\underset{\sim}{\theta}_0}(x)} & \text{for every } x \in S_2 \\[2ex] 0 & \text{for every } x \in \mathfrak{X} - S_1 \cup S_2 \end{cases}$$

$$= \begin{cases} \dfrac{r!}{\theta_0^{r+1}} \dfrac{(f_1(x) - f_2(x))^{r+2}}{\left(f_1(x) + \dfrac{(1-\theta_0)}{\theta_0} f_2(x)\right)^{r+1}} & \text{for every } x \in S_1 \\ \dfrac{r!}{(1-\theta_0)^{r+1}} \dfrac{(f_2(x) - f_1(x))^{r+2}}{\left(\dfrac{\theta_0}{1-\theta_0} f_1(x) + f_2(x)\right)^{r+1}} & \text{for every } x \in S_2 \\ 0 & \text{for every } x \in \mathfrak{X} - S_1 \cup S_2 \end{cases}$$

$$\leq \begin{cases} \dfrac{r!}{\theta_0^{r+1}} f_1(x) & \text{for every } x \in S_1 \\ \dfrac{r!}{(1-\theta_0)^{r+1}} f_2(x) & \text{for every } x \in S_2 \\ 0 & \text{for every } x \in \mathfrak{X} - S_1 \cup S_2 \end{cases}$$

for $r = 1,2,\ldots$. (5.3.1)

For any neighbourhood of θ_0 contained in $(0,1)$, the coefficients of $f_1(x)$ and $f_2(x)$ in (5.3.1) can be held bounded and hence

$$\left| \frac{\partial^r}{\partial \theta^r} \left[\frac{(f_1(x) - f_2(x))^2}{g_\theta(x)} \right] \right| \qquad r = 1,2,\ldots$$

is dominated by an integrable function. Therefore the conditions of the Lebesgue dominated convergence theorem are satisfied by the function $\dfrac{(f_1(x) - f_2(x))^2}{g_\theta(x)}$, completing the proof.

Corollary 5.3.1. The Cramér-Rao lower bound $\dfrac{1}{n\, I(\theta)}$ is an infinitely differentiable function of $\theta \in (0,1)$.

Proof. From the Proposition 5.3.1, $I(\theta)$ is an infinitely differentiable function of $\theta \in (0,1)$. Thus it clearly follows that $\dfrac{1}{n\, I(\theta)}$ is also an infinitely differentiable function of $\theta \in (0,1)$.

Proposition 5.3.2. The Cramér-Rao lower bound $\frac{1}{n\ I(\theta)}$ is a concave function of $\theta \in (0,1)$.

Proof. Differentiating $\frac{1}{n\ I(\theta)}$ twice with respect to θ,

$$\frac{d^2}{d\theta^2}\left(\frac{1}{nI(\theta)}\right) = \frac{-2\left\{\int \frac{(f_1(x)-f_2(x))^2}{g_\theta(x)}\,d\mu \int \frac{(f_1(x)-f_2(x))^4}{g_\theta^3(x)}\,d\mu - \left[\int \frac{(f_1(x)-f_2(x))^3}{g_\theta^2(x)}\,d\mu\right]^2\right\}}{n\left[\int \frac{(f_1(x)-f_2(x))^2}{g_\theta(x)}\,d\mu\right]^3} \qquad (5.3.2)$$

Now, by the Cauchy-Schwartz inequality, we have

$$\left[\int \frac{(f_1(x)-f_2(x))^3}{g_\theta^2(x)}\,d\mu\right]^2 \le \int \frac{(f_1(x)-f_2(x))^2}{g_\theta(x)}\,d\mu \int \frac{(f_1(x)-f_2(x))^4}{g_\theta^3(x)}\,d\mu$$

which shows that the numerator of (5.3.2) is non-positive. Thus the second derivative of $\frac{1}{n\ I(\theta)}$ is non-positive for every $\theta \in (0,1)$ since the denominator in (5.3.2) is positive for such values of θ. Hence $\frac{1}{n\ I(\theta)}$ is a concave function of $\theta \in (0,1)$.

Corollary 5.3.2. The Cramér-Rao lower bound $\frac{1}{n\ I(\theta)}$ has a finite unique relative maximum in $(0,1)$.

Proof. In view of the fact that $\frac{1}{n\ I(\theta)}$ is a continuous function of $\theta \in (0,1)$, its concavity implies that the function has a unique maximum in $(0,1)$.

5.4 Properties of the Likelihood Equation

Given the observations $x_1,\ldots,x_n$ from a distribution with density function $g_\theta(x)$ given by (5.1.1), we assume that $f_1(x_i) \neq f_2(x_i)$ for $i = 1,\ldots,n$. This assumption is plausible for if $f_1(x_\ell) = f_2(x_\ell)$ for some $1 \le \ell \le n$, then the information contained in x_ℓ is zero and

can be thus dismissed. Now, the likelihood equation $\psi(\theta)$ given by (5.2.1), is a decreasing function of θ since

$$\frac{d}{d\theta}\psi(\theta) = -\sum_{i=1}^{n}\left[\frac{f_1(x_i) - f_2(x_i)}{\theta f_1(x_i) + (1-\theta) f_2(x_i)}\right]^2$$

is strictly negative. Let

$$q_j = \frac{f_2(x_j)}{f_2(x_j) - f_1(x_j)} \qquad \text{for } j = 1,\ldots,n \qquad (5.4.1)$$

be the realization of the random variables Q_j given by

$$Q_j = \frac{f_2(X_j)}{f_2(X_j) - f_1(X_j)} \qquad \text{for } j = 1,\ldots,n \qquad (5.4.2)$$

where, as before, X_j is a random variable whose realized value is x_j for $j = 1,\ldots,n$. We can assume without loss of generality that $q_1 < q_2 < \ldots < q_n$ for if not, then q_j's can be rearranged to satisfy this condition. Denote by p, the probability that each $q_j, 1 \leq j \leq n$, is non-positive, i.e.

$$p = \text{prob.}\,[Q_j \leq 0 \text{ for some } 1\leq j\leq n] = \text{prob.}\,[f_1(X_j) > f_2(X_j) \text{ for some } 1\leq j\leq n]$$

$$= \text{prob.}\,[X_j \in S_1 \text{ for some } 1\leq j\leq n] = \int_{S_1} g_\theta(x)\, d\mu \,.$$

Then

$$1-p = \text{prob.}\,[Q_j > 0 \text{ for some } 1\leq j\leq n] = \text{prob.}\,[X_j \in S_2 \text{ for some } 1\leq j\leq n]$$

$$= \int_{S_2} g_\theta(x)\, d\mu \,.$$

Now, using (5.4.1), we can write $\psi(\theta)$ as

$$\psi(\theta) = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{\theta - q_j}$$

which shows that $\psi(\theta)$ is a continuous function of θ for $q_j < \theta < q_{j+1}$ where $j = 1,\ldots,n-1$ and finding the roots of $\psi(\theta) = 0$, leads to

solving a polynomial of degree n-1 with leading coefficient unity. Further for each $1 \leq j \leq n-1$,

$$\psi(q_j + 0) \to + \infty$$

and

$$\psi(q_{j+1} - 0) \to - \infty$$

and therefore in view of the fact that $\psi(\theta)$ is decreasing in the intervals (q_j, q_{j+1}) $j = 1,\ldots,n-1$, it follows that it has a unique root in each of these intervals. The function has simple poles at $q_1,\ldots,q_n$.

Further, if $q_1 < q_2 < \ldots < q_n \leq 0$, the probability of which is p^n, then the roots of $\psi(\theta)$ are negative and the value $\hat{\theta}_n = 0$ is chosen as the estimate of θ^*. On the other hand, if $0 < q_1 < q_2 < \ldots < q_n$, which happends with probability $(1-p)^n$, then with the same probability, $1 \leq q_1 < q_2 < \ldots < q_n$. In this case, the roots of $\psi(\theta)$ exceed unity and we choose $\hat{\theta}_n = 1$ as the estimate of θ^*. In particular if q_j's change sign at q_r for some $1 \leq r \leq n-1$ so that $q_r \leq 0$ and $q_{r+1} > 0$, then there exists a unique value of θ satisfying $\psi(\theta) = 0$ such that $q_r < \theta < q_{r+1}$ which forms the estimate of θ^* (Figure 5.1). Hence the maximum likelihood estimator $\hat{\theta}_n$ of θ^* is defined as

$$\begin{cases} \hat{\theta}_n = 0 & \text{if } q_1 < q_2 < \ldots < q_n \leq 0 \\ q_r < \hat{\theta}_n < q_{r+1} \text{ satisfying } \psi(\hat{\theta}_n) = 0 & \text{if } q_1 < q_2 < \ldots < q_r \leq 0 \\ & \text{and } 0 < q_{r+1} < \ldots < q_n \\ & \text{for some } 1 \leq r \leq n-1 \\ \hat{\theta}_n = 1 & \text{if } 0 < q_1 < \ldots < q_n \end{cases} \tag{5.4.3}$$

and if $\hat{\theta}_n$ is the realization of the random variable Z_n, by using

(5.2.2), we have

$$\left\{\begin{array}{ll} Z_n = 0 & \text{with probability } p^n \\ Q_r < Z_n < Q_{r+1} \text{ for some } 1 \le r \le n-1 \text{ and satisfies } \Psi(Z_n) = 0 & \text{with probability } [1-p^n-(1-p)^n] \\ Z_n = 1 & \text{with probability } (1-p)^n \end{array}\right\}$$

It should be noted that the estimator of θ^* defined by (5.4.3) can take values outside the interval (0,1). Indeed the value $\hat{\theta}_n$, $q_r < \hat{\theta}_n < q_{r+1}$ with $q_r \le 0$ and $q_{r+1} > 0$ defined by (5.4.3) satisfies $q_r \le 0 < \hat{\theta}_n < 1 \le q_{r+1}$ if and only if

$$\psi(0) = \frac{1}{n} \sum_{i=1}^{n} \frac{f_1(x_i) - f_2(x_i)}{f_2(x_i)} = \frac{1}{n} \sum_{i=1}^{n} \frac{f_1(x_i)}{f_2(x_i)} - 1$$

is positive and

$$\psi(1) = \frac{1}{n} \sum_{i=1}^{n} \frac{f_1(x_i) - f_2(x_i)}{f_1(x_i)} = 1 - \frac{1}{n} \sum_{i=1}^{n} \frac{f_2(x_i)}{f_1(x_i)}$$

is negative. Thus defining

$$a_{12}(x) = \frac{f_1(x)}{f_2(x)} \quad \text{and} \quad a_{21} = \frac{f_2(x)}{f_1(x)} \quad ,$$

we have $0 < \hat{\theta}_n < 1$ if and only if the sample means of $a_{12}(x)$ and $a_{21}(x)$ given by

$$\bar{a}_{12} = \frac{1}{n} \sum_{i=1}^{n} \frac{f_1(x_i)}{f_2(x_i)} \quad \text{and} \quad \bar{a}_{21} = \frac{1}{n} \sum_{i=1}^{n} \frac{f_2(x_i)}{f_1(x_i)}$$

both exceed unity. Hence denoting by $\bar{A}_{12}$ and $\bar{A}_{21}$ the random variables whose realizations are $\bar{a}_{12}$ and $\bar{a}_{21}$ respectively, we have

$$\text{Prob}(0<\hat{\theta}_n<1) = 1 - \text{Prob}(\bar{A}_{12}\leq 1) - \text{Prob}(\bar{A}_{21}\leq 1).$$

In the following theorem, we give sufficient conditions for the existence of a unique root of $\psi(\theta) = 0$ in the interval $(0,1)$ in large sample sizes.

Theorem 5.4.1. If $I(0)$ and $I(1)$, where $I(\theta)$ is given by (5.2.3), both exist and are finite, then with probability approaching unity $\psi(\theta) = 0$ has a unique root in the interval $(0,1)$.

Proof. From (5.2.2), we have

$$\Psi(0) = \frac{1}{n}\sum_{i=1}^{n}\frac{f_1(X_i)-f_2(X_i)}{f_2(X_i)} \quad \text{and} \quad \Psi(1) = \frac{1}{n}\sum_{i=1}^{n}\frac{f_1(X_i)-f_2(X_i)}{f_1(X_i)}.$$

Thus $\Psi(0)$ and $\Psi(1)$ are both sums of independently and identically distributed random variables with

$$E_\theta\left[\frac{f_1(X)-f_2(X)}{f_2(X)}\right] = \int\frac{f_1(x)-f_2(x)}{f_2(x)}\,[\theta f_1(x) + (1-\theta)f_2(x)]\,d\mu$$

$$= \theta\int\frac{(f_1(x)-f_2(x))^2}{f_2(x)}\,d\mu + \int\frac{(f_1(x)-f_2(x))}{f_2(x)}\,f_2(x)\,d\mu$$

$$= \theta\, I(0)$$

and similarly

$$E_\theta\left[\frac{f_1(X)-f_2(X)}{f_1(X)}\right] = \int\frac{f_1(x)-f_2(x)}{f_1(x)}\,[\theta f_1(x) + (1-\theta)f_2(x)]\,d\mu$$

$$= -(1-\theta)\int\frac{(f_1(x)-f_2(x))^2}{f_1(x)}\,d\mu + \int\frac{(f_1(x)-f_2(x))}{f_1(x)}\,f_1(x)\,d\mu$$

$$= -(1-\theta)\, I(1).$$

So, if I(0) and I(1) both exist and are finite, then by the weak law of large numbers $\Psi(0)$ and $\Psi(1)$ converge in probability to $\theta I(0)$ and $-(1-\theta)\, I(1)$ respectively, i.e.

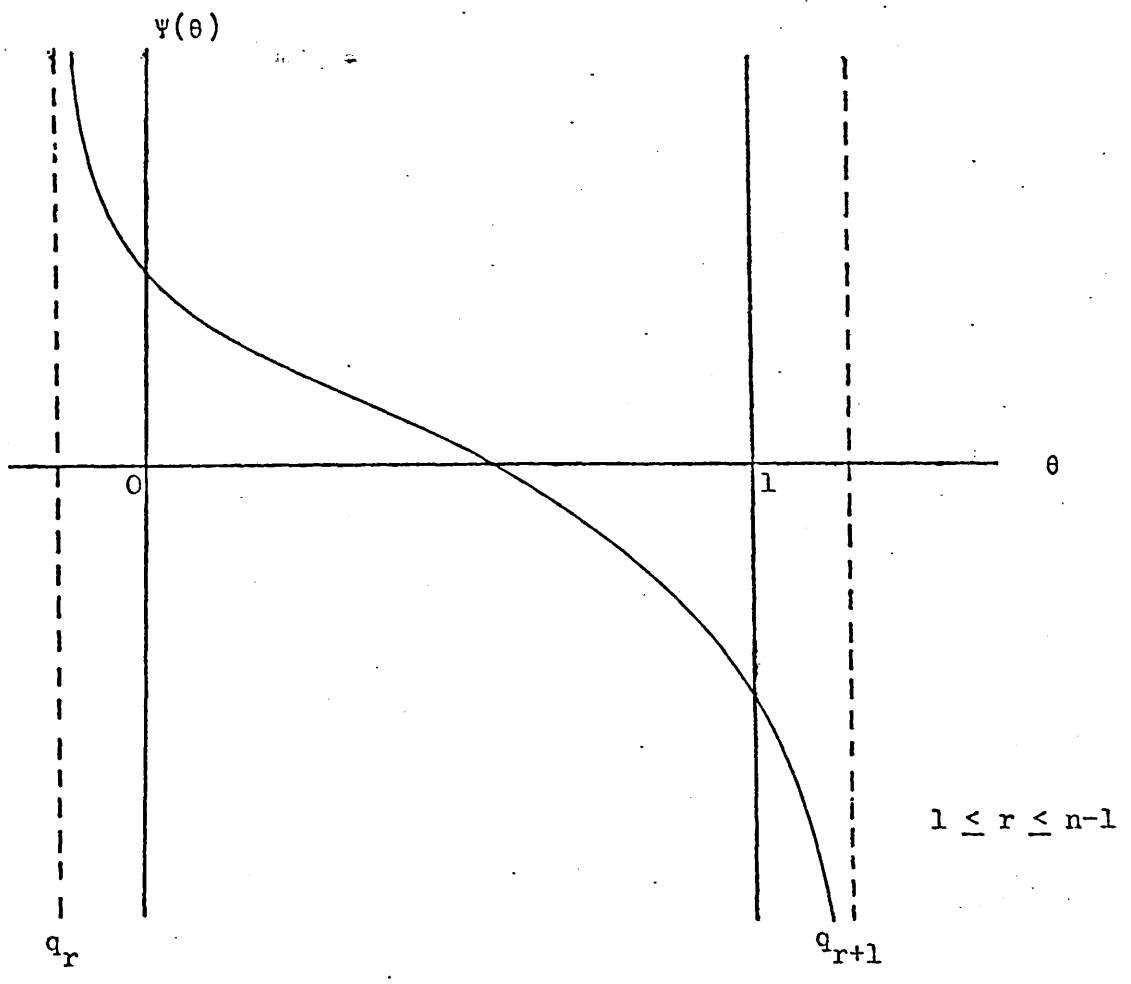
$$\Psi(0) = \theta I(0) + o_p(1) \qquad \text{as } n \to \infty$$

and

$$\Psi(1) = -(1-\theta)\, I(1) + o_p(1) \qquad \text{as } n \to \infty .$$

Now, since I(0) and I(1) are both positive, with probability approaching unity $\Psi(0) > 0$ and $\Psi(1) < 0$ as $n \to \infty$. Hence with such a probability there exists a unique root of $\psi(\theta) = 0$ in (0,1) as $n \to \infty$.

Figure 5.1



5.5 Iterative Solutions of the Likelihood Equation

The existence of a root of $\psi(\theta) = 0$ forming the maximum likelihood estimator of θ^* was discussed in previous sections. To obtain this root is yet another problem since solving $\psi(\theta) = 0$, in general, is of course an impractical task. We can, however, use numerical techniques to obtain the solution of $\psi(\theta) = 0$. The computational routines for finding the roots of a likelihood equation has been the subject of many papers. Barnett [3] gives an analysis of the various numerical techniques used to approximate the roots of the likelihood equation and Kale [29, 30] investigated the large sample properties of iterative processes. R.A. Fisher was the first to discuss and advocate the use of successive iterations to solve the likelihood equation. Fisher argued that in many cases where the regularity conditions listed in Section 4.2 are satisfied, it would be sufficient to execute only one cycle iteration in order to arrive at a good approximation. Northan [39] has shown, however, that several cycles of iteration may be required to obtain a reasonable convergence. Barnett [3] illustrates the properties of several successive approximation techniques for small samples, when the random variables have a Cauchy distribution depending on a location parameter.

Perhaps the most commonly used numerical method for locating the relative maxima of the likelihood equation is the celebrated Newton-Raphson method and other well-known techniques are variants of it. The Newton-Raphson method is based on the expansion of the likelihood equation in Taylor's series around its root. Thus if $\hat{\theta}_n$ denotes the root of $\psi(\theta) = 0$,

$$0 = \psi(\hat{\theta}_n) = \psi(\hat{\theta}_n^{(0)}) + (\hat{\theta}_n - \hat{\theta}_n^{(0)}) \frac{\partial}{\partial\theta} \psi(\hat{\theta}_n + \nu(\hat{\theta}_n - \hat{\theta}_n^{(0)})) \quad (5.5.1)$$

for some $0 \leq \nu \leq 1$, where $\hat{\theta}_n^{(0)}$ is an initial solution. If we take

$\nu = 0$ in (5.5.1), we obtain an approximation for $\hat{\theta}_n$ in (5.5.1), namely

$$\hat{\theta}_n^{(1)} = \hat{\theta}_n^{(0)} - \left[\frac{\psi(\theta)}{\frac{\partial}{\partial\theta}\psi(\theta)}\right]_{\theta = \hat{\theta}_n^{(0)}} \tag{5.5.2}$$

and by using (5.2.1), we obtain

$$\hat{\theta}_n^{(1)} = \hat{\theta}_n^{(0)} + \frac{\frac{1}{n}\sum_{j=1}^{n} \frac{f_1(x_j) - f_2(x_j)}{g_{\hat{\theta}_{\sim n}^{(0)}}(x_j)}}{\frac{1}{n}\sum_{j=1}^{n}\left[\frac{f_1(x_j) - f_2(x_j)}{g_{\hat{\theta}_{\sim n}^{(0)}}(x_j)}\right]^2} \tag{5.5.3}$$

where $\hat{\theta}_{\sim n}^{(0)} = (\hat{\theta}_n^{(0)}, 1 - \hat{\theta}_n^{(0)})'$. The value $\hat{\theta}_n^{(1)}$ can be substituted in (5.5.3) for $\hat{\theta}_n^{(0)}$ to obtain another value $\hat{\theta}_n^{(2)}$, and so on. Generally, starting from an initial solution $\hat{\theta}_n^{(0)}$, we generate a sequence $\{\hat{\theta}_n^{(r)} ; r = 0,1,2,\ldots\}$, which is determined successively by the formula

$$\hat{\theta}_n^{(r+1)} = \hat{\theta}_n^{(r)} + \frac{\frac{1}{n}\sum_{j=1}^{n} \frac{f_1(x_j) - f_2(x_j)}{g_{\hat{\theta}_{\sim n}^{(r)}}(x_j)}}{\frac{1}{n}\sum_{j=1}^{n}\left[\frac{f_1(x_j) - f_2(x_j)}{g_{\hat{\theta}_{\sim n}^{(r)}}(x_j)}\right]^2} \qquad r = 0,1,\ldots \tag{5.5.4}$$

where $\hat{\theta}_{\sim n}^{(r)} = (\hat{\theta}_n^{(r)}, 1 - \hat{\theta}_n^{(r)})'$ for $r = 1,2,\ldots$. If the initial solution $\hat{\theta}_n^{(0)}$ was chosen close to the root of the likelihood equation $\hat{\theta}_n$, there is a good chance that the sequence generated by (5.5.4) will converge to the root $\hat{\theta}_n$.

It is shown in Zacks [61], that the Newton-Raphson method of iteration generally leads to a CAN and asymptotically efficient estimator after the first cycle of the iteration process is completed, provided

that the initial solution is a consistent estimator. Thus if $\hat{\theta}_n^{(0)}$ is chosen to be a consistent estimator of θ^*, and if $\hat{\theta}_n^{(0)}$ is the realization of a random variable $Z_n^{(0)}$, then using (5.5.3), by Zack's theorem we can say that the random variable $Z_n^{(1)}$ given by

$$Z_n^{(1)} = Z_n^{(0)} + \frac{\frac{1}{n}\sum_{j=1}^{n} \frac{f_1(X_j) - f_2(X_j)}{Z_n^{(0)} f_1(X_j) + (1-Z_n^{(0)}) f_2(X_j)}}{\frac{1}{n}\sum_{j=1}^{n} \left[\frac{f_1(X_j) - f_2(X_j)}{Z_n^{(0)} f_1(X_j) + (1-Z_n^{(0)}) f_2(X_j)}\right]^2} \tag{5.5.5}$$

is asymptotically normally distributed with mean θ^* and variance $\frac{1}{n\, I(\theta^*)}$.

The method of iteration that we adopt in here is the method of "scoring for parameter" which is derived from the Newton-Raphson method by replacing the denominator of the correction term in (5.5.2), namely $\frac{\partial}{\partial\theta}\Psi(\theta)$, by its expected value given by

$$E_{\underset{\sim}{\theta}}\left[\frac{\partial}{\partial\theta}\Psi(\theta)\right] = -E_{\underset{\sim}{\theta}}\left[\frac{f_1(X_j) - f_2(X_j)}{g_{\underset{\sim}{\theta}}(X_j)}\right]^2 = -\int \frac{(f_1(x) - f_2(x))^2}{g_{\underset{\sim}{\theta}}(x)}\, d\mu = -I(\theta)\,.$$

Thus choosing the initial solution $\hat{\theta}_n^{(0)}$, the sequence $\{\hat{\theta}_n^{(r)}\ ;\ r = 0,1,\ldots\}$ is generated by successive substitution in the formula

$$\hat{\theta}_n^{(r+1)} = \hat{\theta}_n^{(r)} + \frac{1}{n\, I(\hat{\theta}_n^{(r)})} \sum_{j=1}^{n} \frac{f_1(x_j) - f_2(x_j)}{g_{\hat{\underset{\sim}{\theta}}_n^{(r)}}(x_j)} \qquad r = 0,1,\ldots \tag{5.5.6}$$

where $\hat{\underset{\sim}{\theta}}_n^{(r)} = (\hat{\theta}_n^{(r)},\ 1-\hat{\theta}_n^{(r)})'$ for $r = 0,1,2,\ldots$ and $I(\theta)$ is the Fisher's information function.

The method of scoring for parameter was first introduced by Fisher and it is argued by various authors (e.g. Kale [29]) that this method is often more appropriate from the computational point of view in certain cases specially for large sample sizes. Using the method of scoring for parameter to find the root of $\psi(\theta) = 0$, it turns out that the

solution obtained after a 1-cycle iteration has certain interesting properties when the initial solution is chosen to be any arbitrary value in the interval (0,1). In the followings, we investigate these properties together with the properties of the solution obtained after the second cycle of the iteration.

Suppose that the value $\hat{\theta}^{(0)}$ is chosen (independent of the observations) in the interval (0,1). Substituting $\hat{\theta}^{(0)}$ for $\hat{\theta}_n^{(0)}$ in (5.5.6) with r = 0, we get

$$\hat{\theta}_n^{(1)} = \hat{\theta}^{(0)} + \frac{1}{n\, I(\hat{\theta}^{(0)})} \sum_{j=1}^{n} \frac{f_1(x_j) - f_2(x_j)}{g_{\hat{\underset{\sim}{\theta}}^{(0)}}(x_j)} \qquad (5.5.7)$$

where $\hat{\underset{\sim}{\theta}}^{(0)} = (\hat{\theta}^{(0)}, 1-\hat{\theta}^{(0)})'$, as the estimate of θ^* obtained after the first cycle of the iteration. Let $Z_n^{(1)}$ be the random variable whose realization $\hat{\theta}_n^{(1)}$ is given by (5.5.7).

Theorem 5.5.1. The random variable $Z_n^{(1)}$ converges almost surely to θ^* as $n \to \infty$.

Proof. Consider

$$\Psi(\hat{\theta}^{(0)}) = \frac{1}{n} \sum_{j=1}^{n} \frac{f_1(X_j) - f_2(X_j)}{g_{\hat{\underset{\sim}{\theta}}^{(0)}}(X_j)},$$

which is the sum of independently and identically distributed random variables. Then a necessary and sufficient condition for $\Psi(\hat{\theta}^{(0)})$ to obey the strong law of large numbers is that

$$E_{\underset{\sim}{\theta}^*}\left|\frac{f_1(X) - f_2(X)}{g_{\hat{\underset{\sim}{\theta}}^{(0)}}(X)}\right| < \infty. \qquad (5.5.8)$$

Now,

$$\left|\frac{f_1(x) - f_2(x)}{g_{\hat{\underset{\sim}{\theta}}(0)}(x)}\right| = \begin{cases} \dfrac{f_1(x) - f_2(x)}{g_{\hat{\underset{\sim}{\theta}}(0)}(x)} & \text{for every } x \in S_1 \\ \dfrac{f_2(x) - f_1(x)}{g_{\hat{\underset{\sim}{\theta}}(0)}(x)} & \text{for every } x \in S_2 \\ 0 & \text{for every } x \in \mathfrak{X} - S_1 \cup S_2 \end{cases}$$

where S_1 and S_2 are, as before, defined by

$$S_1 = \{x \in \mathfrak{X} : f_1(x) > f_2(x)\} \text{ and } S_2 = \{x \in \mathfrak{X} : f_2(x) > f_1(x)\}.$$

Then

$$\left|\frac{f_1(x) - f_2(x)}{g_{\hat{\underset{\sim}{\theta}}(0)}(x)}\right| = \begin{cases} \dfrac{1}{\hat{\theta}^{(0)}} \dfrac{f_1(x) - f_2(x)}{(f_1(x)-f_2(x)) + \frac{1}{\hat{\theta}^{(0)}} f_2(x)} & \text{for every } x \in S_1 \\ \dfrac{1}{1-\hat{\theta}^{(0)}} \dfrac{f_2(x) - f_1(x)}{(f_2(x)-f_1(x)) + \frac{1}{1-\hat{\theta}^{(0)}} f_1(x)} & \text{for every } x \in S_2 \\ 0 & \text{for every } x \in \mathfrak{X} - S_1 \cup S_2 \end{cases}$$

$$\leq \begin{cases} \dfrac{1}{\hat{\theta}^{(0)}} & \text{for every } x \in S_1 \\ \dfrac{1}{1-\hat{\theta}^{(0)}} & \text{for every } x \in S_2 \\ 0 & \text{for every } x \in \mathfrak{X} - S_1 \cup S_2 \end{cases}$$

$$\leq \frac{1}{\min(\hat{\theta}^{(0)}, 1 - \hat{\theta}^{(0)})} \quad \text{for every } x \in \mathfrak{X}.$$

Hence (5.5.8) holds since $0 < \hat{\theta}^{(0)} < 1$ and therefore by the strong law of large numbers, $\Psi(\hat{\theta}^{(0)})$ converges almost surely to

$$E_{\underset{\sim}{\theta^*}}\left[\frac{f_1(X)-f_2(X)}{g_{\underset{\sim}{\hat{\theta}}^{(0)}}(X)}\right] = \int \frac{f_1(x)-f_2(x)}{g_{\underset{\sim}{\hat{\theta}}^{(0)}}(x)} \left[\theta^*(f_1(x)-f_2(x)) + f_2(x)\right] d\mu$$

$$= \theta^* \int \frac{(f_1(x)-f_2(x))^2}{g_{\underset{\sim}{\hat{\theta}}^{(0)}}(x)} d\mu + \int \frac{(f_1(x)-f_2(x))}{g_{\underset{\sim}{\hat{\theta}}^{(0)}}(x)} f_2(x)\, d\mu$$

$$= (\theta^* - \hat{\theta}^{(0)}) \int \frac{(f_1(x)-f_2(x))^2}{g_{\underset{\sim}{\hat{\theta}}^{(0)}}(x)} d\mu$$

$$+ \int \frac{(f_1(x)-f_2(x))}{g_{\underset{\sim}{\hat{\theta}}^{(0)}}(x)} \left[\hat{\theta}^{(0)}(f_1(x)-f_2(x)) + f_2(x)\right] d\mu$$

$$= (\theta^* - \hat{\theta}^{(0)})\, I(\hat{\theta}^{(0)}) + \int (f_1(x)-f_2(x))\, d\mu$$

$$= (\theta^* - \hat{\theta}^{(0)})\, I(\hat{\theta}^{(0)}) \,. \qquad (5.5.9)$$

Now, from (5.5.7),

$$Z_n^{(1)} = \hat{\theta}^{(0)} + \frac{\psi(\hat{\theta}^{(0)})}{I(\hat{\theta}^{(0)})} \qquad (5.5.10)$$

which by using (5.5.9), we see that as $n \to \infty$, $Z_n^{(1)}$ converges almost surely to

$$\hat{\theta}^{(0)} + \frac{(\theta^* - \hat{\theta}^{(0)})\, I(\hat{\theta}^{(0)})}{I(\hat{\theta}^{(0)})} = \theta^*$$

completing the proof of the theorem.

Theorem 5.5.2. The estimate $\hat{\theta}_n^{(1)}$ given by (5.5.7) is unbiased and CAN with

$$\mathrm{Var}_{\underset{\sim}{\theta^*}}[Z_n^{(1)}] = \frac{1}{n}\left\{\frac{1}{I^2(\hat{\theta}^{(0)})} \int \frac{(f_1(x)-f_2(x))^2}{g^2_{\underset{\sim}{\hat{\theta}}^{(0)}}(x)} g_{\underset{\sim}{\theta^*}}(x)\, d\mu - (\theta^*-\hat{\theta}^{(0)})^2\right\}. \qquad (5.5.11)$$

Proof. From (5.5.10),

$$E_{\underset{\sim}{\theta}*}[Z_n^{(1)}] = \hat{\theta}^{(0)} + \frac{1}{I(\hat{\theta}^{(0)})} E_{\underset{\sim}{\theta}*}\left[\frac{f_1(X)-f_2(X)}{g_{\underset{\sim}{\hat{\theta}}^{(0)}}(X)}\right] = \theta^*$$

which proves the unbiasedness. Further,

$$\mathrm{Var}_{\underset{\sim}{\theta}*}[Z_n^{(1)}] = \frac{1}{n\, I^2(\hat{\theta}^{(0)})} \mathrm{Var}_{\underset{\sim}{\theta}*}\left[\frac{f_1(X)-f_2(X)}{g_{\underset{\sim}{\hat{\theta}}^{(0)}}(X)}\right]$$

$$= \frac{1}{n\, I^2(\hat{\theta}^{(0)})} \left\{E_{\underset{\sim}{\theta}*}\left[\frac{f_1(X)-f_2(X)}{g_{\underset{\sim}{\hat{\theta}}^{(0)}}(X)}\right]^2 - \left[E_{\underset{\sim}{\theta}*}\left\{\frac{f_1(X)-f_2(X)}{g_{\underset{\sim}{\hat{\theta}}^{(0)}}(X)}\right\}\right]^2\right\}$$

$$= \frac{1}{n\, I^2(\hat{\theta}^{(0)})} \left\{\int \frac{(f_1(x)-f_2(x))^2}{g^2_{\underset{\sim}{\hat{\theta}}^{(0)}}(x)} g_{\underset{\sim}{\theta}*}(x)\,d\mu - (\theta^*-\hat{\theta}^{(0)})^2 I^2(\hat{\theta}^{(0)})\right\}$$

which establishes (5.5.11).

Finally, since $\Psi(\hat{\theta}^{(0)})$ is the sum of independently and identically distributed random variables with finite first and second moments, by the central limit theorem, the asymptotic distribution of $\Psi(\hat{\theta}^{(0)})$ is normal. Thus it follows from (5.5.10) that the asymptotic distribution of $Z_n^{(1)}$ is also normal with mean θ^* and variance given by (5.5.11).

Hence by using theorem 5.5.2, the variance of $Z_n^{(1)}$ can be found for any finite samples from the formulae (5.5.11). We shall now prove that although the estimator $\hat{\theta}_n^{(1)}$ is not uniformly minimum variance unbiased, it has, however, the optimal property of being locally minimum variance unbiased estimator at $\hat{\theta}^{(0)}$. First we define these terms and in doing so, we denote by $\mathcal{M}$ the class of all unbiased estimates of θ^*.

Definition 5.5.1. (Zacks [61]): An unbiased estimator of θ^* say $\hat{\theta} \in \mathcal{M}$, which is the realization of a random variable Z, is said to be uniformly minimum variance unbiased (UMVU) estimator if given any other unbiased estimator, say $\bar{\theta} \in \mathcal{M}$ which is the realization of the random variable $\bar{Z}$, we have

$$\mathrm{Var}_{\underset{\sim}{\theta}}(Z) \leq \mathrm{Var}_{\underset{\sim}{\theta}}(\bar{Z})$$

for every $\underset{\sim}{\theta} = (\theta, 1-\theta)'$ with $\theta \in (0,1)$. The UMVU estimator is often called the best unbiased estimator in statistical literature.

Definition 5.5.2. (Zacks [61]): The estimator $\hat{\theta}$ defined above is said to be locally minimum variance unbiased (LMVU) estimator at $\theta_0 \in (0,1)$ if

$$\mathrm{Var}_{\underset{\sim}{\theta}_0}(Z) \leq \mathrm{Var}_{\underset{\sim}{\theta}_0}(\bar{Z})$$

where $\underset{\sim}{\theta}_0 = (\theta_0, 1-\theta_0)'$ and $\bar{Z}$ is as defined in Definition 5.5.1.

Theorem 5.5.3. The estimate $\hat{\theta}_n^{(1)}$ given by (5.5.7) is LMVU at $\hat{\theta}^{(0)}$.

Proof. From Definition 5.5.1, it is clear that $\hat{\theta}_n^{(1)}$ is not UMVU since $n \mathrm{Var}_{\underset{\sim}{\theta}}(Z_n^{(1)})$ obtained from (5.5.11) is not equivalent to the inverse of the information function $I(\theta)$. However, we can see from (5.5.11) that

$$\mathrm{Var}_{\underset{\sim}{\hat{\theta}}^{(0)}}(Z_n^{(1)}) = \frac{1}{n\, I(\hat{\theta}^{(0)})}$$

which is the minimum attainable variance for an unbiased estimator according to the Cramér-Rao inequality. Hence by Definition 5.5.2, $\hat{\theta}_n^{(1)}$ is LMVU at $\hat{\theta}^{(0)}$.

The LMVU estimates, apart from being of interest on their own, are often used in statistical estimation problems when UMVU estimates of the unknown parameters in a distribution do not exist. Since the essential element for the existence of a UMVU estimator is the completeness of the family of distributions in the given statistical model (Zacks [61]), a UMVU estimate will not exist in the absence of completeness.

In such situations the LMVU estimators may, however, exist and serve as a practical alternative.

Zacks [61] gives a useful general account of LMVU estimators and proves that a necessary and sufficient condition for an unbiased estimator to be LMVU at a certain point of the parameter space is that the estimator should be uncorrelated with any unbiased estimator of 0 with a finite variance. Thus if $\eta(X)$ is an unbiased estimator of 0 such that

$$E_{\underset{\sim}{\theta}}[\eta(X)] = 0$$

for every $\underset{\sim}{\theta} = (\theta, 1-\theta)'$ with $\theta \in (0,1)$, by using theorem 5.5.3, we have

$$\mathrm{Cov}_{\underset{\sim}{\theta}^{(0)}}(Z_n^{(1)}, \eta(X)) = 0$$

whenever

$$\mathrm{Var}_{\underset{\sim}{\theta}^{(0)}}(\eta(X)) < \infty .$$

We have therefore established certain properties of $\hat{\theta}_n^{(1)}$, the estimator of θ^* given by a 1-cycle iteration of the Fisher's scoring method. But the choice of the starting solution $\hat{\theta}^{(0)}$ which is an arbitrary point (independent of the observations) taken from the interval (0,1) can, of course, influence $\hat{\theta}_n^{(1)}$ and thus $\hat{\theta}^{(0)}$ should be chosen so that the anomaly associated with this choice is minimized in some sense. We note from (5.5.11) that $\mathrm{Var}_{\underset{\sim}{\theta}^*}[Z_n^{(1)}]$ is a continuous and differentiable function of θ^*. Upon differentiating it twice with respect to θ^*, we obtain

$$\frac{d}{d\theta^*}\left(\mathrm{Var}_{\underset{\sim}{\theta}^*}[Z_n^{(1)}]\right) = \frac{1}{n}\left\{\frac{1}{I^2(\hat{\theta}^{(0)})}\int \frac{(f_1(x)-f_2(x))^3}{g^2_{\underset{\sim}{\hat{\theta}}^{(0)}}(x)}\,d\mu - 2(\theta^*-\hat{\theta}^{(0)})\right\} \tag{5.5.12}$$

and

$$\frac{d^2}{d\theta^{*2}} \left(\mathrm{Var}_{\underset{\sim}{\theta^*}} [Z_n^{(1)}]\right) = -2 \qquad (5.5.13)$$

which shows that $\mathrm{Var}_{\underset{\sim}{\theta^*}} [Z_n^{(1)}]$ is a concave function of θ^* achieving its maximum at the root of (5.5.12) given by

$$\underset{\sim}{\theta^*} = \hat{\theta}^{(0)} + \frac{1}{2\, I^2(\hat{\theta}^{(0)})} \int \frac{(f_1(x) - f_2(x))^3}{g^2_{\underset{\sim}{\hat{\theta}}^{(0)}}(x)}\, d\mu \,. \qquad (5.5.14)$$

Substituting (5.5.14) into (5.5.11), we get

$$\sup_{0<\theta^*<1} \mathrm{Var}_{\underset{\sim}{\theta^*}} [Z_n^{(1)}] = \frac{1}{n\, I(\hat{\theta}^{(0)})} \left\{ 1 + \frac{1}{4\, I^3(\hat{\theta}^{(0)})} \left[\int \frac{(f_1(x)-f_2(x))^3}{g^2_{\underset{\sim}{\hat{\theta}}^{(0)}}(x)} \right]^2 \right\} \qquad (5.5.15)$$

and by minimizing (5.5.15) with respect to $\hat{\theta}^{(0)}$, we can determine the arbitrary starting point so that the maximum variance after one iteration is minimum (i.e. $\hat{\theta}_n^{(1)}$ is the minimax estimator under the squared error loss function).

Let

$$J_r(\theta) = \int \frac{(f_1(x) - f_2(x))^{r+1}}{g^r_{\underset{\sim}{\theta}}(x)}\, d\mu$$

for $r = 0,1,\ldots$ and for every $\theta \in (0,1)$ where $\underset{\sim}{\theta} = (\theta, 1-\theta)'$. Note that in particular we have $J_0(\theta) = 0$ for every $\theta \in (0,1)$ and $J_1(\theta) = I(\theta)$. Since $J_r(\theta)$ has the property that

$$\frac{d}{d\theta} J_r(\theta) = -r\, J_{r+1}$$

for $r = 0,1,\ldots$, we can write (5.5.15) as

$$\sup_{0<\theta^*<1} \operatorname{Var}_{\theta^*}\left[Z_n^{(1)}\right] = \frac{1}{n}\left\{\frac{1}{4}\left[\frac{J_2(\hat{\theta}^{(0)})}{J_1^2(\hat{\theta}^{(0)})}\right]^2 + \frac{1}{J_1(\hat{\theta}^{(0)})}\right\}$$

$$= \frac{1}{n}\left\{\frac{1}{4}\left[-\frac{d}{d\hat{\theta}^{(0)}}\left[\frac{1}{J_1(\hat{\theta}^{(0)})}\right]\right]^2 + \frac{1}{J_1(\hat{\theta}^{(0)})}\right\} \qquad (5.5.16)$$

To minimize (5.5.16), we set its derivative, with respect to $\hat{\theta}^{(0)}$, equal to zero. We find that the stationary points of (5.5.16) are the roots of

$$\frac{1}{2}\frac{d^2}{d\theta^2}\left[\frac{1}{J_1(\theta)}\right] \cdot \frac{d}{d\theta}\left[\frac{1}{J_1(\theta)}\right] + \frac{d}{d\theta}\left[\frac{1}{J_1(\theta)}\right] = 0$$

or

$$\frac{d}{d\theta}\left[\frac{1}{J_1(\theta)}\right]\left\{\frac{1}{2}\frac{d^2}{d\theta^2}\left[\frac{1}{J_1(\theta)}\right] + 1\right\} = 0$$

from which

$$\frac{d}{d\theta}\left[\frac{1}{J_1(\theta)}\right] = 0$$

yields

$$J_2 = 0$$

and

$$\frac{1}{2}\frac{d^2}{d\theta^2}\left[\frac{1}{J_1(\theta)}\right] + 1 = 0$$

has a solution of the form

$$J_1(\theta) = \frac{1}{-\theta^2 + A\theta + B} \qquad (5.5.17)$$

where A and B are real constants determined such that $0 < J_1(\theta) < \infty$ for every $0 < \theta < 1$.

By differentiating (5.5.16) twice, it will be seen that the solution $J_2(\theta) = 0$ corresponds to the minimum of (5.5.16) while (5.5.17) corresponds to its maximum. But according to corollary 5.3.2, the

equation

$$J_2(\theta) = \frac{d}{d\theta}\,(-\,I(\theta)) = \int \frac{(f_1(x) - f_2(x))^3}{g^2_{\underset{\sim}{\theta}}(x)}\,d\mu = 0 \qquad (5.5.18)$$

has necessarily a unique root in (0,1) and hence if this root is used as the starting solution of the iteration process, then $\hat{\theta}_n^{(1)}$, the estimate of θ^* given by a 1-cycle iteration, is minimax under the squared error loss function.

This result is intuitively plausible for if we denote the unique root of (5.5.18) by θ_m and substitute θ_m in (5.5.15) for $\hat{\theta}^{(0)}$, we find that

$$\sup_{0<\theta^*<1} \operatorname{Var}_{\underset{\sim}{\theta^*}}\left[Z_n^{(1)}\right]\Bigg|_{\hat{\theta}^{(0)} = \theta_m} = \frac{1}{n\,I(\theta_m)}$$

which is minimum according to the Cramér-Rao inequality.

So far, we have established the properties of the estimate of θ^* obtained after the first cycle of the iteration. The question naturally arising now is the behaviour of the subsequent iterations. In the following theorem, we establish the properties of the estimate of θ^* obtained after the second cycle of the iteration is completed. Thus suppose that $\hat{\theta}_n^{(1)}$ given by (5.5.7) is substituted in (5.5.6) with r = 1, to obtain $\hat{\theta}_n^{(2)}$ and denote by $Z_n^{(2)}$ the random variable whose realization is $\hat{\theta}_n^{(2)}$. Hence

$$\hat{\theta}_n^{(2)} = \hat{\theta}_n^{(1)} + \frac{1}{n\,I(\hat{\theta}_n^{(1)})} \sum_{j=1}^{n} \frac{f_1(x_j) - f_2(x_j)}{g_{\hat{\underset{\sim}{\theta}}_n^{(1)}}(x_j)} \qquad (5.5.19)$$

and

$$Z_n^{(2)} = Z_n^{(1)} + \frac{1}{n\,I(Z_n^{(1)})} \sum_{j=1}^{n} \frac{f_1(X_j) - f_2(X_j)}{g_{Z_{\underset{\sim}{n}}^{(1)}}(X_j)} \qquad (5.5.20)$$

where $\underset{\sim}{Z}_n^{(1)} = (Z_n^{(1)}, 1-Z_n^{(1)})'$ and as before $\underset{\sim}{\hat{\theta}}_n^{(1)} = (\hat{\theta}_n^{(1)}, 1-\hat{\theta}_n^{(1)})'$. (In the following theorem, we use the symbols $O_p(.)$ and $o_p(.)$ introduced in Section 3.7).

Theorem 5.5.4. The estimator of θ^* given by the second cycle of the iteration process is CAN with asymptotic variance given by $1/(n\, I(\theta^*))$, proving that it is asymptotically fully efficient.

Proof. Expanding (5.5.20) by Taylor's series about θ^*, we get

$$Z_n^{(2)} = Z_n^{(1)} + \frac{1}{n}\left\{\sum_{j=1}^{n} \frac{f_1(X_j)-f_2(X_j)}{g_{\underset{\sim}{\theta}^*}(X_j)} - (Z_n^{(1)}-\theta^*) \sum_{j=1}^{n} \frac{(f_1(X_j)-f_2(X_j))^2}{g^2_{\underset{\sim}{\theta}^*}(X_j)} + O_p(Z_n^{(1)}-\theta^*)^2\right\}\left\{I^{-1}(\theta^*) + o_p(Z_n^{(1)}-\theta^*)\right\} \quad (5.5.21)$$

Now $\dfrac{f_1(X) - f_2(X)}{g_{\theta^*}(X)}$ is a random variable whose first and second moments exist and thus

$$\frac{1}{n}\sum_{j=1}^{n} \frac{f_1(X_j) - f_2(X_j)}{g_{\underset{\sim}{\theta}^*}(X_j)} = o_p\,(n^{-\alpha}) \qquad \text{as } n \to \infty$$

for every $\alpha < 1$ and similarly

$$\frac{1}{n}\sum_{j=1}^{n} \frac{(f_1(X_j) - f_2(X_j))^2}{g^2_{\underset{\sim}{\theta}^*}(X_j)} = I(\theta^*) + o_p\,(n^{-\alpha})$$

for every $\alpha < 1$. So

$$Z_n^{(1)} = \hat{\theta}^{(0)} + \frac{1}{n\, I(\hat{\theta}^{(0)})} \sum_{j=1}^{n} \frac{f_1(X_j) - f_2(X_j)}{g_{\underset{\sim}{\hat{\theta}}^{(0)}}(X_j)}$$

$$= \hat{\theta}^{(0)} + \frac{1}{n\, I(\hat{\theta}^{(0)})}\{n\,(\theta^*-\hat{\theta}^{(0)})\, I(\hat{\theta}^{(0)}) + o_p\,(n^{-\alpha})\}$$

$$= \theta^* + o_p\,(n^{-\alpha}) \qquad \text{for every } \alpha < 1.$$

Hence from (5.5.21),

$$Z_n^{(2)} = Z_n^{(1)} + \frac{\frac{1}{n}\sum_{j=1}^{n} \frac{f_1(X_j)-f_2(X_j)}{g_{\theta^*}(X_j)}}{I(\theta^*)} - (Z_n^{(1)} - \theta^*) + o_p(n^{-\alpha}) \quad \alpha < 1$$

and therefore

$$\sqrt{n}\,(Z_n^{(2)} - \theta^*) = \frac{\frac{1}{\sqrt{n}}\sum_{j=1}^{n} \frac{f_1(X_j)-f_2(X_j)}{g_{\theta^*}(X_j)}}{I(\theta^*)} + o_p(n^{-(\alpha-\frac{1}{2})}) \quad (5.5.22)$$

for every $\alpha < 1$. Thus for every α such that $\frac{1}{2} < \alpha < 1$, and by using the lemma 4.2.1, we see that $\sqrt{n}(Z_n^{(2)}-\theta^*)$ and $\frac{\frac{1}{\sqrt{n}}\sum_{j=1}^{n} \frac{f_1(X_j)-f_2(X_j)}{g_{\theta^*}(X_j)}}{I(\theta^*)}$ have the same asymptotic distributions. But

$$\frac{\frac{1}{\sqrt{n}}\sum_{j=1}^{n} \frac{f_1(X_j) - f_2(X_j)}{g_{\theta^*}(X_j)}}{I(\theta^*)}$$

is the sum of independently and identically distributed random variables admitting first and second order moments and hence by the central limit theorem its asymptotic distribution is normal with mean

$$E_{\theta^*}\left[\frac{\frac{1}{\sqrt{n}}\sum_{j=1}^{n} \frac{f_1(X_j) - f_2(X_j)}{g_{\theta^*}(X_j)}}{I(\theta^*)}\right] = 0$$

and variance

$$\mathrm{Var}_{\theta^*}\left[\frac{\frac{1}{\sqrt{n}}\sum_{j=1}^{n} \frac{f_1(X_j) - f_2(X_j)}{g_{\theta^*}(X_j)}}{I(\theta^*)}\right] = \frac{1}{I(\theta^*)}$$

which completes the proof of the theorem.

5.6 Monte Carlo Studies

This section is devoted to a small numerical study of the Fisher's scoring method of iteration for finding the solution of the likelihood equation. Similar to numerical studies performed in previous chapters, we shall be concerned with a mixture of two normal distributions. As indicated at the beginning of this chapter, the estimation problems concerned with a mixture of two normal distributions has created considerable difficulties in the past, but since we are only estimating the mixing proportion with other parameters known, the maximum likelihood estimator of the mixing proportion exists.

Consider a mixture $g_{\underset{\sim}{\theta}}(.)$ of two normal density functions

$$f_1(x) = \frac{1}{\sqrt{2\pi}} \exp\{-\tfrac{1}{2} x^2\} \qquad -\infty < x < +\infty$$

and

$$f_2(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\} \qquad -\infty < x < +\infty$$

so that

$$g_{\underset{\sim}{\theta}}(x) = \theta\, f_1(x) + (1-\theta)\, f_2(x) \qquad -\infty < x < +\infty$$

where $\underset{\sim}{\theta} = (\theta, 1-\theta)'$ with $\theta \in (0,1)$ and where $\sigma > 0$ and μ are known parameters.

Given a random sample $X_1,\ldots,X_n$ with observed values $x_1,\ldots,x_n$ respectively, (5.5.6) was used for $r = 0,1$ to obtain the estimate of θ after the first and second cycle iterations. The initial solution $\hat{\theta}^{(0)}$ can be chosen to be any arbitrary point in the interval $(0,1)$ (independent of the observations). Table 5.1 gives $\hat{\theta}_n^{(1)}$ and $\hat{\theta}_n^{(2)}$ when $\hat{\theta}^{(0)} = 0.3$ for mixtures of distributions considered in sections 2.7, 3.5 and 4.3. The mean-square-error of each estimate and the

standard error of each mean are calculated as explained in Section 2.7. When n is large, $\hat{\theta}_n^{(2)}$ has a smaller mean-square-error than $\hat{\theta}_n^{(1)}$. This is in agreement with the theorem 5.5.4 which asserts that $\hat{\theta}_n^{(2)}$ possesses the asymptotic properties of the maximum likelihood estimator of θ. However, for small n, $\hat{\theta}_n^{(1)}$ seems to be preferable.

In order to investigate the dependence of $\hat{\theta}_n^{(1)}$ on the choice of $\hat{\theta}^{(0)}$, we picked three following cases:

(i) $\frac{\mu_2 - \mu_1}{\sigma_1} = 0.5$ $\frac{\sigma_2}{\sigma_1} = 1$ n = 50 N = 5000

(ii) $\frac{\mu_2 - \mu_1}{\sigma_1} = 1$ $\frac{\sigma_2}{\sigma_1} = 1$ n = 10 N = 5000

(iii) $\frac{\mu_2 - \mu_1}{\sigma_1} = 5$ $\frac{\sigma_2}{\sigma_1} = 1$ n = 20 N = 5000

and plotted the mean-square-error of $\hat{\theta}_n^{(1)}$ against different values of $\hat{\theta}^{(0)}$ (Figure 5.2). The dependence of $\hat{\theta}_n^{(1)}$ on $\hat{\theta}^{(0)}$ in cases (i) and (iii) seems to be negligible and in (ii) very small. This is believed to be due to relatively large sample sizes in cases (i) and (iii). The mean-square-error reduces substantially as $\frac{\mu_2 - \mu_1}{\sigma_1}$ increases which stresses the point, already discussed in Section 5.3, that the more the components of a mixture of two distributions are separated, the easier is the estimation of the mixing proportion based on a given sample.

Finally, in Figure 5.3, the mean-square-error of $\hat{\theta}_n^{(1)}$ and $\hat{\theta}_n^{(2)}$ are plotted against n when $\frac{\mu_2 - \mu_1}{\sigma_1}$ and $\frac{\sigma_2}{\sigma_1}$ are as in (ii). Again it is observed that $\hat{\theta}_n^{(2)}$ is to be preferred to $\hat{\theta}_n^{(1)}$ for large n, whereas in small samples the latter has a much lower mean-square-error in comparison with the former.

Table 5.1

Maximum likelihood estimator of the mixing proportion, based on the solution of the first and second cycles of the Fisher's scoring method of iteration, for various mixtures of two normal distributions

n	$(\mu_2-\mu_1)/\sigma_1$	σ_2/σ_1	θ	Estimate of θ after the first and second Iterations			
				$\hat{\theta}_n^{(1)}$		$\hat{\theta}_n^{(2)}$	
				Mean	MSE	Mean	MSE
50	0.25	1	0.5	0.419 ± 0.053	0.285	0.421 ± 0.047	0.232
10	0.5	1	0.5	0.456 ± 0.029	0.425	0.467 ± 0.032	0.517
10	1	1	0.5	0.476 ± 0.016	0.126	0.490 ± 0.025	0.323
10	1	1	0.8	0.788 ± 0.016	0.130	0.769 ± 0.021	0.231
20	5	1	0.5	0.493 ± 0.007	0.013	0.493 ± 0.007	0.013
10	0	2	0.5	0.476 ± 0.018	0.168	0.457 ± 0.020	0.202
50	0	2	0.5	0.476 ± 0.019	0.037	0.476 ± 0.018	0.037
10	0.5	2	0.5	0.468 ± 0.017	0.155	0.467 ± 0.026	0.346

Each case is based on $n_1 = \frac{5000}{n}$ samples of size n. The standard error of each mean is given to indicate the accuracy of the Monte Carlo computation.

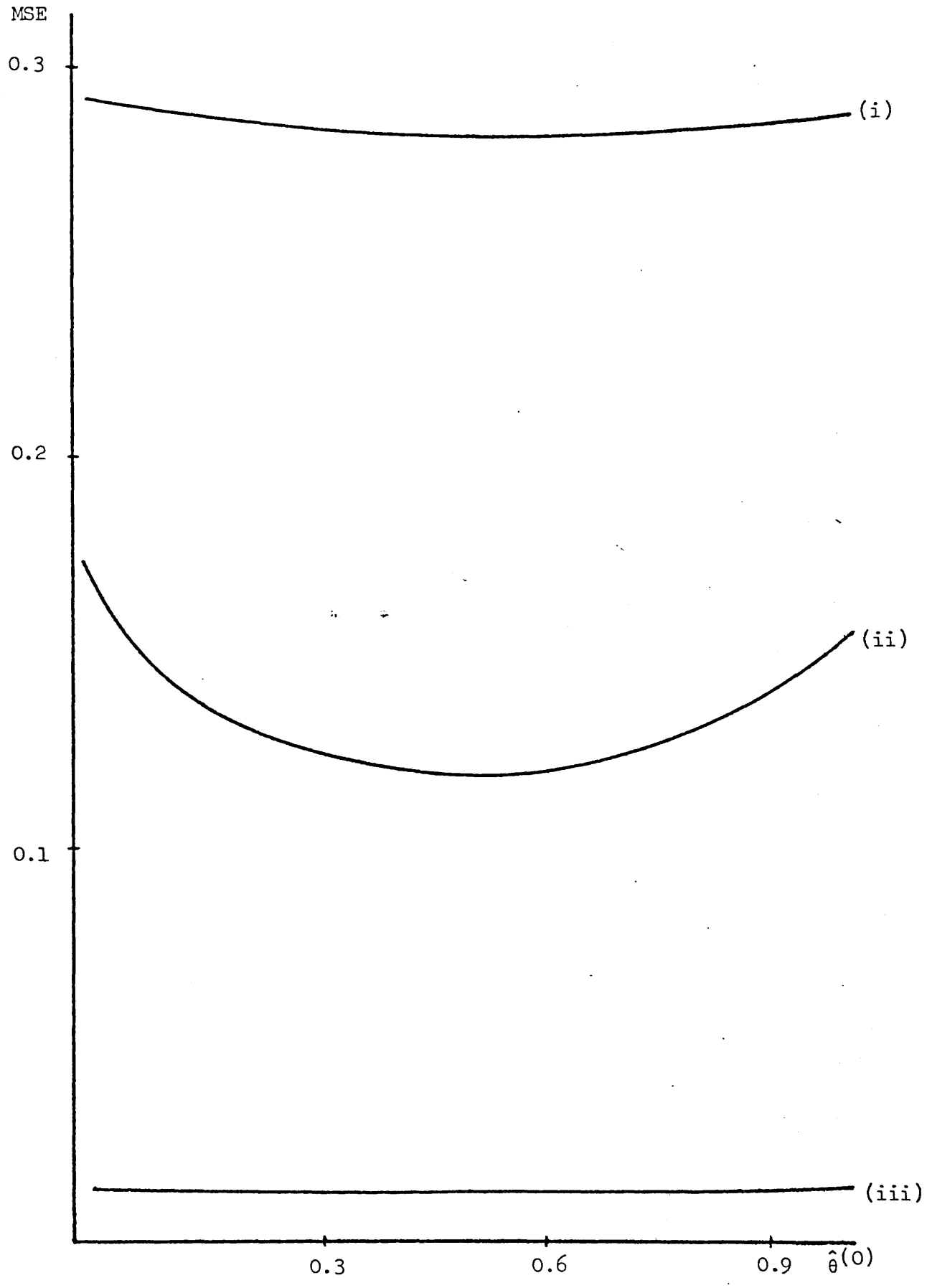


Fig. 5.2 - The mean-square-error of $\theta_n^{(1)}$, the maximum likelihood estimator of the mixing proportion, based on the 1-cycle solution of the Fisher's scoring method of iteration, in various mixtures of two normal distributions, against the arbitrary starting point $\hat{\theta}^{(0)}$.

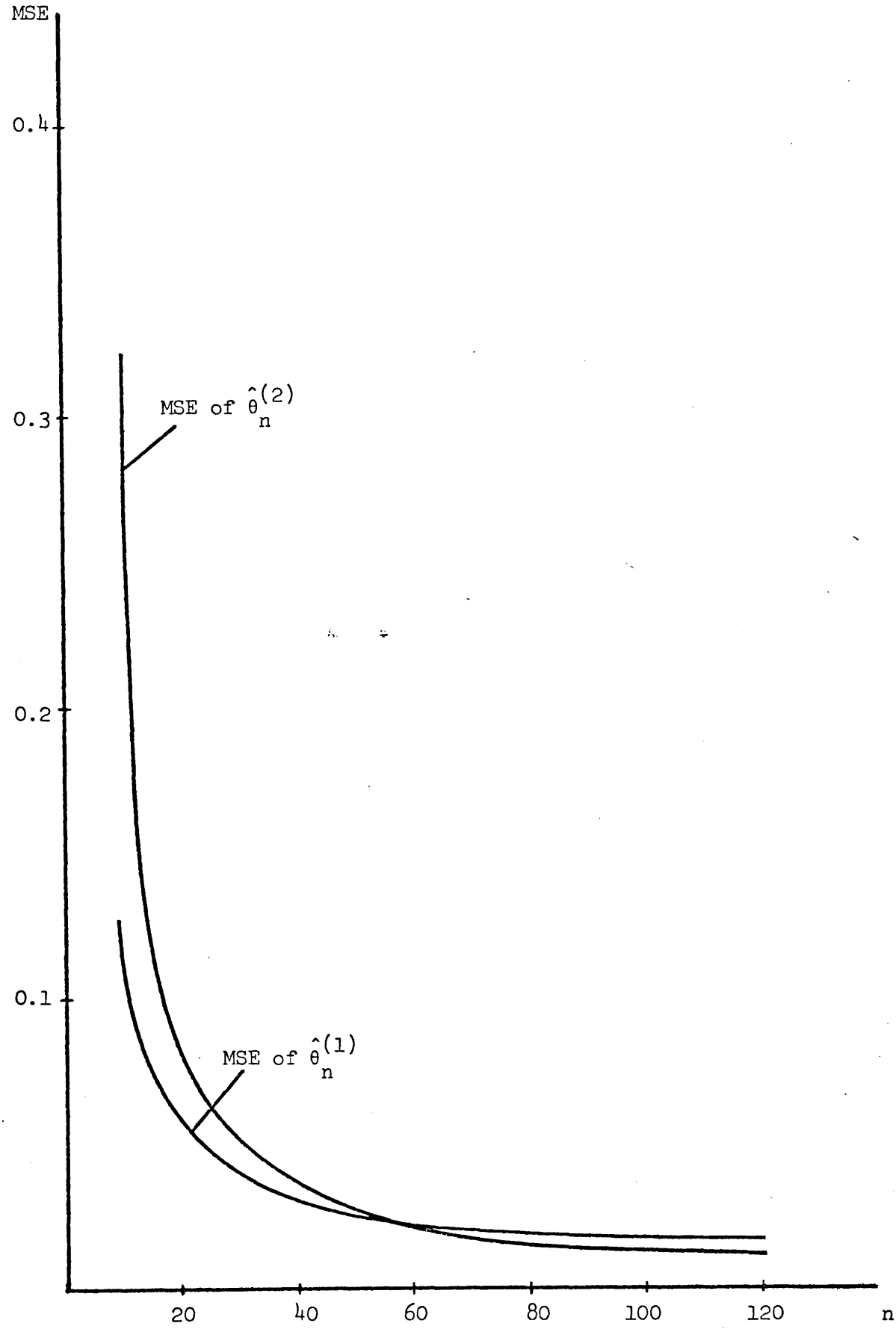


Fig. 5.3 - The mean-square-error of $\theta_n^{(1)}$ and $\theta_n^{(2)}$, the maximum likelihood estimator of the mixing proportion θ, based on the first and second cycles of the Fisher's scoring method of iteration, in a mixture of two normal distributions N(0,1) and N(1,1), for varying sample sizes.

5.7 Conclusions

The maximum likelihood estimator of the mixing proportion in a mixture of two distributions is the root of an equation which cannot be solved directly. Due to the interesting properties of maximum likelihood estimates, we use iteration to obtain a close approximation to the root of the likelihood equation. It is seen that a 1-cycle iteration of the Fisher's scoring method yields a CAN and locally minimum variance unbiased estimate whilst the solution obtained after the completion of the second cycle of the iteration process is CAN and asymptotically fully efficient.

CHAPTER 6

CONCLUSIONS

This chapter contains a brief summary of the important points raised in Chapters 2 to 5 and a short discussion of some topics for further studies. In our investigations, we have attempted to throw some light on the problem of estimating the mixing proportions in a finite mixture of distributions by simple adaptation and utilization of various well-known estimation techniques. The aim throughout the thesis has been to construct estimators which are of value both in theory and practice.

The extension of the method of moments in Chapter 2 has an interesting feature and gives way to new methods of obtaining reliable estimates. Although the method has some desirable asymptotic properties and works well in practice, it relies very much on the trial and error procedure. On the other hand, when the observations are grouped, the results of Chapter 3 show that the generalized least squares estimators are asymptotically efficient with respect to a given set of division points. The main problems, however, are to choose the division points of the sample space and secondly to find the solution of the underlying equations. With the advent of modern computers, the latter is not an obstacle whilst there is no unified theory of choosing the best set of class intervals and it is generally believed that the greater the number of division points, the better the results. To this end, we have proved that as the number of division points become infinite, the resultant estimators are asymptotically fully efficient.

In order to obtain simple approximations to the root of the set of equations whose root constitutes the generalized least squares estimators of the mixing proportions, we have seen that the iteration process proposed in Chapter 4 gives, after even one cycle, estimates

which are asymptotically efficient with respect to a given set of division points. When the lengths of the group intervals become small, the solutions given by successive iterations approach the maximum likelihood estimates.

The interesting features of the maximum likelihood estimates are their asymptotic properties. Dealing with a mixture of two distributions in Chapter 5, it is seen that the maximum likelihood estimator of the mixing proportion always exists and possesses the well-known asymptotic properties, provided that the mixing proportion is strictly between zero and one. If the Fisher's information function is defined and is finite at zero and one, then with probability approaching unity the likelihood equation has a unique root in the interval (0,1). Similar to Chapter 3, the main difficulty is to find the solution of the likelihood equation and by appealing to an iteration process commonly known as the Fisher's scoring method, approximate solutions can be found. A deep study of the first and second cycle solutions together with the results of the Monte Carlo studies reveal the fact that even one or two cycles are sufficient to produce close approximations to the solution of the likelihood equation. This, we believe, has an important practical implication since by simple manipulations, an efficient estimator of the mixing proportion can be obtained.

It is, nevertheless, clear that there are many interesting questions, concerning finite mixtures of distributions, which require further investigations. Firstly, in the broad sense, the problem of hypothesis testing is an area which needs further research. A complete Bayesian analysis of mixtures of distributions is also still to be undertaken.

The problem of identifiability of finite mixtures of distributions can raise many interesting problems. Although the results of Teicher

[56] and Yakowitz and Spragins [60] provide useful tools for checking the identifiability of a given finite mixture of distributions, it would be interesting to know that if a finite mixture of distributions $G_{\underset{\sim}{\theta}}(.)$ is not identifiable, then what set of values of the mixing proportions give rise to the same value of $G_{\underset{\sim}{\theta}}(.)$.

Another area which has attracted some statisticians and could have interesting implications when applied to mixtures of distributions is the problem of inference about a change-point in a sequence of random variables. A sequence of random variables $X_1,\ldots,X_n$ is said to have a change-point at r ($1 \leq r \leq n$) if the common distribution function of $X_1,\ldots,X_r$ is $G_1(.,\underset{\sim}{\theta}_1)$ whereas $X_{r+1},\ldots,X_n$ have a common distribution function $G_2(.,\underset{\sim}{\theta}_2)$ where $G_1(.,\underset{\sim}{\theta}_1) \neq G_2(.,\underset{\sim}{\theta}_2)$. Page [40] used a cumulative sum technique to detect the existence of a distributional change in the sequence $X_1,\ldots,X_n$. To make inference about the change-point r, Hinkley [27] used arguments based on maximum likelihood estimates, likelihood ratio tests and cumulative sum tests and recently Smith [51] has treated the problem from a Bayesian view-point. Now, in the context of mixtures of distributions, the distribution functions $G_1(.,\theta_1)$ and $G_2(.,\theta_2)$ may be taken to be finite mixtures of distributions with different mixing proportions (possibly involving the same components). The problem would then be to estimate the unknown mixing proportions and the change-point r.

We finally close this thesis by bearing in mind the following remark due to K. Pearson:

"No scientific investigation can be final; it merely represents the most probable conclusion which can be drawn from the data at the disposal of the writer. A wider range of facts, or more refined analysis, experiment, and observation will lead to new formulae and theories. This is the essence of scientific progress."

APPENDIX A

ON THE JOINT ASYMPTOTIC DISTRIBUTION OF $P_1, \ldots, P_{m+1}$

Let X be a random variable whose distribution function $G_{\underset{\sim}{\theta}}(x)$, a mixture of distribution functions $F_1(x),\ldots,F_k(x)$, is given by (2.2.1). Given the random variables $X_1,\ldots,X_n$ with common distribution $G_{\underset{\sim}{\theta}}(x)$ and with realizations $x_1,\ldots,x_n$ respectively, denote by $G_n(x)$ the empirical distribution function based on this sample, i.e. the proportion of the observations which are not greater than x. Let $G_n(x)$ be the realization of the random function $\Gamma_n(x)$ and assume that the sample space $\mathcal{X}$ is partitioned into m+1 intervals at the points

$$t_o < t_1 < \ldots < t_m < t_{m+1}$$

where $G_n(t_o) = G_{\underset{\sim}{\theta}}(t_o) = 0$ and $G_n(t_{m+1}) = G_{\underset{\sim}{\theta}}(t_{m+1}) = 1$.

Put $\pi_i(\underset{\sim}{\theta}) = G_{\underset{\sim}{\theta}}(t_i) - G_{\underset{\sim}{\theta}}(t_{i-1})$, $p_i = G_n(t_i) - G_n(t_{i-1})$ and $P_i = \Gamma_n(t_i) - \Gamma_n(t_{i-1})$ for $i = 1,\ldots,m+1$. In this appendix, we establish the joint asymptotic distribution of $P_1,\ldots,P_{m+1}$.

Proposition A1: The joint asymptotic distribution of $P_1,\ldots,P_{m+1}$ is a (m+1)-variate normal distribution with mean vector

$$\underset{\sim}{\pi}(\underset{\sim}{\theta}^*) = (\pi_1(\underset{\sim}{\theta}^*), \ldots, \pi_{m+1}(\underset{\sim}{\theta}^*))'$$

and (m+1)×(m+1) covariance matrix $\frac{1}{n}\Sigma$ where the $(i,j)^{th}$ element of Σ is given by

$$\begin{aligned} \sigma_{ij} &= \pi_i(\underset{\sim}{\theta}^*)(1 - \pi_i(\underset{\sim}{\theta}^*)) && i = j \\ &= -\pi_i(\underset{\sim}{\theta}^*)\,\pi_j(\underset{\sim}{\theta}^*) && i \neq j \end{aligned}$$

for $i,j = 1,\ldots,m+1$. (A.1)

Here, $\underset{\sim}{\theta}^* = (\theta_1^*,\ldots,\theta_k^*)'$ denotes the true value of the unknown vector of parameters $\underset{\sim}{\theta} = (\theta_1,\ldots,\theta_k)'$.

Proof: As in (1.3.3), we can write

$$G_n(x) = \frac{1}{n}\sum_{j=1}^{n} \eta(x-x_j)$$

where $$\eta(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}$$

and therefore

$$\Gamma_n(x) = \frac{1}{n}\sum_{j=1}^{n} \eta(x-X_j) \ .$$

Define $U_{ij} = \eta(t_i-X_j) - \eta(t_{i-1}-X_j)$

for $i = 1,\ldots,m+1$ and $j = 1,\ldots,n$. Then

$$E_{\underset{\sim}{\theta}^*}(U_{ij}) = E_{\underset{\sim}{\theta}^*}(\eta(t_i-X_j)) - E_{\underset{\sim}{\theta}^*}(\eta(t_{i-1}-X_j))$$

$$= G_{\underset{\sim}{\theta}^*}(t_i) - G_{\underset{\sim}{\theta}^*}(t_{i-1}) = \pi_i(\underset{\sim}{\theta}^*) \qquad (A.2)$$

for $i = 1,\ldots,m+1$ and $j = 1,\ldots,n$. Further,

$$\mathrm{Var}_{\underset{\sim}{\theta}^*}(U_{ij}) = \mathrm{Var}_{\underset{\sim}{\theta}^*}[\eta(t_i-X_j)] + \mathrm{Var}_{\underset{\sim}{\theta}^*}[\eta(t_{i-1}-X_j)]$$

$$- 2\,\mathrm{Cov}_{\underset{\sim}{\theta}^*}[\eta(t_i-X_j),\ \eta(t_{i-1}-X_j)] \qquad (A.3)$$

for $i = 1,\ldots,m+1$ and $j = 1,\ldots,n$.

But $\mathrm{Var}_{\underset{\sim}{\theta}^*}(\eta(t_i-X_j)) = G_{\underset{\sim}{\theta}^*}(t_i)(1 - G_{\underset{\sim}{\theta}^*}(t_i))$

and $\mathrm{Cov}_{\underset{\sim}{\theta}^*}(\eta(t_i-X_j),\eta(t_r-X_j)) = \min(G_{\underset{\sim}{\theta}^*}(t_i),\ G_{\underset{\sim}{\theta}^*}(t_r))$

$$- G_{\underset{\sim}{\theta}^*}(t_i)\ G_{\underset{\sim}{\theta}^*}(t_r)$$

for $i,r = 1,\ldots,m+1$ and $j = 1,\ldots,n$ where

$$\min(x,y) = \begin{cases} x & x \le y \\ y & y \le x \end{cases} .$$

Therefore, (A.3) gives

$$\begin{aligned} \underset{\sim}{\mathrm{Var}}_{\theta^*}(U_{ij}) &= \underset{\sim}{G}_{\theta^*}(t_i) - \underset{\sim}{G}^2_{\theta^*}(t_i) + \underset{\sim}{G}_{\theta^*}(t_{i-1}) - \underset{\sim}{G}^2_{\theta^*}(t_{i-1}) \\ &\quad - 2\,\underset{\sim}{G}_{\theta^*}(t_{i-1}) + 2\,\underset{\sim}{G}_{\theta^*}(t_i)\,\underset{\sim}{G}_{\theta^*}(t_{i-1}) \\ &= (\underset{\sim}{G}_{\theta^*}(t_i) - \underset{\sim}{G}_{\theta^*}(t_{i-1})) - (\underset{\sim}{G}_{\theta^*}(t_i) - \underset{\sim}{G}_{\theta^*}(t_{i-1}))^2 \\ &= \underset{\sim}{\pi}_i(\theta^*) - \underset{\sim}{\pi}^2_i(\theta^*) \end{aligned} \qquad (A.4)$$

for $i = 1,\ldots,m+1$ and $j = 1,\ldots,n$.

Also for $r \neq i$, we have

$$\begin{aligned} \underset{\sim}{\mathrm{Cov}}_{\theta^*}(U_{ij},U_{rj}) &= \underset{\sim}{\mathrm{Cov}}_{\theta^*}[\,(\eta(t_i-X_j) - \eta(t_{i-1}-X_j)),\ (\eta(t_r-X_j) - \eta(t_{r-1}-X_j))] \\ &= \underset{\sim}{\mathrm{Cov}}_{\theta^*}(\eta(t_i-X_j),\ \eta(t_r-X_j)) - \underset{\sim}{\mathrm{Cov}}_{\theta^*}(\eta(t_{i-1}-X_j),\ \eta(t_r-X_j)) \\ &\quad - \underset{\sim}{\mathrm{Cov}}_{\theta^*}(\eta(t_i-X_j),\ \eta(t_{r-1}-X_j)) + \underset{\sim}{\mathrm{Cov}}_{\theta^*}(\eta(t_{i-1}-X_j),\ \eta(t_{r-1}-X_j)) \\ &= \min(\underset{\sim}{G}_{\theta^*}(t_i),\ \underset{\sim}{G}_{\theta^*}(t_r)) - \underset{\sim}{G}_{\theta^*}(t_i)\,\underset{\sim}{G}_{\theta^*}(t_r) \\ &\quad - \min(\underset{\sim}{G}_{\theta^*}(t_{i-1}),\ \underset{\sim}{G}_{\theta^*}(t_r)) + \underset{\sim}{G}_{\theta^*}(t_{i-1})\,\underset{\sim}{G}_{\theta^*}(t_r) \\ &\quad - \min(\underset{\sim}{G}_{\theta^*}(t_i),\ \underset{\sim}{G}_{\theta^*}(t_{r-1})) + \underset{\sim}{G}_{\theta^*}(t_i)\,\underset{\sim}{G}_{\theta^*}(t_{r-1}) \\ &\quad + \min(\underset{\sim}{G}_{\theta^*}(t_{i-1}),\ \underset{\sim}{G}_{\theta^*}(t_{r-1})) - \underset{\sim}{G}_{\theta^*}(t_{i-1})\,\underset{\sim}{G}_{\theta^*}(t_{r-1}) \end{aligned} \qquad (A.5)$$

for $i,r = 1,\ldots,m+1$ and $j = 1,\ldots,n$. Let without loss of generality $i < r$ in (A.5), then either $i = r-1$ or $i = 1,2,\ldots,r-2$. If $i = r-1$, (A.5) gives

$$\mathrm{Cov}_{\underset{\sim}{\theta}^*}\{U_{(r-1)j}, U_{rj}\} = -\{G_{\underset{\sim}{\theta}^*}(t_{r-1}) - G_{\underset{\sim}{\theta}^*}(t_{r-2})\}\{G_{\underset{\sim}{\theta}^*}(t_r) - G_{\underset{\sim}{\theta}^*}(t_{r-1})\}$$

$$= -\pi_{r-1}(\underset{\sim}{\theta}^*)\,\pi_r(\underset{\sim}{\theta}^*) \qquad r = 2,\ldots,m+1$$

while for $i = 1,\ldots,(r-2)$, (A.5) yields

$$\mathrm{Cov}_{\underset{\sim}{\theta}^*}(U_{ij}, U_{rj}) = -\{G_{\underset{\sim}{\theta}^*}(t_i) - G_{\underset{\sim}{\theta}^*}(t_{i-1})\}\{G_{\underset{\sim}{\theta}^*}(t_r) - G_{\underset{\sim}{\theta}^*}(t_{r-1})\}$$

$$= -\pi_i(\underset{\sim}{\theta}^*)\,\pi_r(\underset{\sim}{\theta}^*) \qquad i = 1,\ldots,r-2$$

$$r = 2,\ldots,m+1 \quad .$$

Hence, if we define

$$\underset{\sim}{U}_j = (U_{1j}, \ldots, U_{m+1j})' \qquad 1 \le j \le n$$

then $\underset{\sim}{U}_j$ is a $(m+1)$-dimensional random vector with

$$E_{\underset{\sim}{\theta}^*}(\underset{\sim}{U}_j) = (\pi_1(\underset{\sim}{\theta}^*), \ldots, \pi_{m+1}(\underset{\sim}{\theta}^*))' = \underset{\sim}{\pi}(\underset{\sim}{\theta}^*)$$

and $(m+1)\times(m+1)$ covariance matrix Σ with σ_{ij}, given by (A.1), as its (i,j)th element.

Note that, since $X_1,\ldots,X_n$ are independent random variables with a common distribution, it follows that $\underset{\sim}{U}_1, \underset{\sim}{U}_2, \ldots, \underset{\sim}{U}_n$ form a sequence of independently and identically distributed random vectors admitting first - and second order moments.

Now,

$$P_i = \Gamma_n(t_i) - \Gamma_n(t_{i-1}) = \frac{1}{n}\sum_{j=1}^{n} U_{ij}$$

for $i = 1,\ldots,m+1$. Let $\underset{\sim}{P} = (P_1, \ldots, P_{m+1})'$, then by the multivariate central limit theorem, the asymptotic distribution of $\sqrt{n}\,(\underset{\sim}{P} - \underset{\sim}{\pi}(\underset{\sim}{\theta}^*))$ is $N_{m+1}(\underset{\sim}{0}, \Sigma)$, that is an $(m+1)$-variate normal with zero mean and $(m+1)\times(m+1)$ covariance matrix Σ with the density function

$$N_{m+1}(\underset{\sim}{u}|\underset{\sim}{0}, \Sigma) = (2\pi)^{-(m+1)/2}\,|\Sigma|^{-\frac{1}{2}}\,\exp\{-\tfrac{1}{2}(\underset{\sim}{u}-\underset{\sim}{\pi}(\underset{\sim}{\theta}^*))'\,\Sigma^{-1}(\underset{\sim}{u}-\underset{\sim}{\pi}(\underset{\sim}{\theta}^*))\} \quad .$$

Here $\underset{\sim}{0}$ denotes the (m+1)-dimensional vector of 0's. Hence the proof is completed.

APPENDIX B

GENERALIZATION OF THE THEOREM 4.2.1

In Chapter 4, an iteration process was introduced to find the GLS estimator of $\underset{\sim}{\theta}^* = (\theta_1^*,\ldots,\theta_k^*)'$, the true value of the unknown vector $\underset{\sim}{\theta} = (\theta_1,\ldots,\theta_k)'$. Here $\underset{\sim}{\theta}$ is the vector of the mixing proportions in the mixture of distributions $G_{\underset{\sim}{\theta}}(x)$ given by (2.2.1). We proved in theorem 4.2.1 that for $k = 2$, if the iteration is started with a consistent estimator of $\underset{\sim}{\theta}^*$, then the solution obtained after a 1-cycle iteration is CAN with asymptotic variance being minimum with respect to a fixed set of division points $t_1 < t_2 < \ldots < t_m$ of the sample space $\mathfrak{X}$. In this appendix, the result of the theorem 4.2.1 is generalized for the case $k > 2$. Recall that $\{t_i\}_{i=1}^m$ are chosen so that the rank of the matrix A given by (3.2.2) is exactly k and that $0 < G_{\underset{\sim}{\theta}}(t_1) < \ldots < G_{\underset{\sim}{\theta}}(t_m) < 1$ for every $\underset{\sim}{\theta} \in \Theta$.

Since $\sum_{j=1}^{k} \theta_j = 1$, we put $\theta_k = 1 - \sum_{j=1}^{k-1} \theta_j$ in $G_{\underset{\sim}{\theta}}(x) = \sum_{j=1}^{k} \theta_j F_j(x)$ to get

$$G_{\underset{\sim}{\theta}}(x) = \theta_1 (F_1(x) - F_k(x)) + \ldots + \theta_{k-1} (F_{k-1}(x) - F_k(x)) + F_k(x)$$

where $0 \leq \theta_j \leq 1$ for $j = 1,\ldots,k-1$ and $x \in \mathfrak{X}$. Let $\beta_{ji} = F_j(t_i) - F_j(t_{i-1})$ for $j = 1,\ldots,k$ and for $i = 1,\ldots,m+1$ where as before $F_j(t_{m+1}) = 1$ and $F_j(t_0) = 0$ for $j = 1,\ldots,k$, so that

$$\pi_i(\underset{\sim}{\theta}) = G_{\underset{\sim}{\theta}}(t_i) - G_{\underset{\sim}{\theta}}(t_{i-1}) = \sum_{j=1}^{k-1} \theta_j (\beta_{ji} - \beta_{ki}) + \beta_{ki} \qquad \text{(B.1)}$$

for $i = 1,\ldots,m+1$. Upon substituting $\pi_i(\underset{\sim}{\theta})$ in $\Phi_r(\underset{\sim}{\theta})$ given by (4.2.4) and setting the derivative of $\Phi_r(\underset{\sim}{\theta})$ with respect to θ_j; $j = 1,\ldots,k-1$, equal to zero, a system of k-1 equations in k-1 unknowns is obtained as follows,

$$\frac{\partial}{\partial \theta_j} \Phi_r(\underset{\sim}{\theta}) = 2 \sum_{i=1}^{k-1} \frac{(p_i - \pi_i(\underset{\sim}{\theta}))(\beta_{ji} - \beta_{ki})}{\pi_i(\hat{\underset{\sim}{\theta}}_n^{(r)})} = 0 \qquad \text{(B.2)}$$

where $j = 1,\ldots,k-1$ and $\hat{\underset{\sim}{\theta}}_n^{(r)} = (\hat{\theta}_1^{(r)},\ldots,\hat{\theta}_k^{(r)})$; $r = 0,1,\ldots,$ is the estimate of $\underset{\sim}{\theta}^*$ obtained after the rth cycle of the iteration process. Using (B.1), (B.2) yields

$$\theta_1 \sum_{i=1}^{m+1} \frac{(\beta_{1i} - \beta_{ki})(\beta_{ji} - \beta_{ki})}{\pi_i(\hat{\underset{\sim}{\theta}}_n^{(r)})} + \theta_2 \sum_{i=1}^{m+1} \frac{(\beta_{2i} - \beta_{ki})(\beta_{ji} - \beta_{ki})}{\pi_i(\hat{\underset{\sim}{\theta}}_n^{(r)})} + \ldots$$

$$+ \theta_{(k-1)} \sum_{i=1}^{m+1} \frac{(\beta_{(k-1)i} - \beta_{ki})(\beta_{ji} - \beta_{ki})}{\pi_i(\hat{\underset{\sim}{\theta}}_n^{(r)})} = \sum_{i=1}^{m+1} \frac{(p_i - \beta_{ki})(\beta_{ji} - \beta_{ki})}{\pi_i(\hat{\underset{\sim}{\theta}}_n^{(r)})} \qquad \text{(B.3)}$$

for $j = 1,\ldots,k-1$.

Define by $R(\underset{\sim}{\theta})$ a symmetric $(k-1) \times (k-1)$ matrix whose (j,ℓ)th element is given by

$$R_{j\ell}(\underset{\sim}{\theta}) = \sum_{i=1}^{m+1} \frac{(\beta_{ji} - \beta_{ki})(\beta_{\ell i} - \beta_{ki})}{\pi_i(\underset{\sim}{\theta})} \qquad j,\ell = 1,\ldots,k-1$$

and let $\underset{\sim}{\tau} = (\theta_1,\ldots,\theta_{k-1})'$, then we can write (B.3) for $j = 1,\ldots,k-1$ as

$$R(\hat{\underset{\sim}{\theta}}_n^{(r)})\ \underset{\sim}{\tau} = \underset{\sim}{b} \qquad r = 0,1,\ldots \qquad \text{(B.4)}$$

where $\underset{\sim}{b}$ is a $(k-1)$-dimensional vector with its jth element defined as

$$b_j = \sum_{i=1}^{m+1} \frac{(p_i - \beta_{ki})(\beta_{ji} - \beta_{ki})}{\pi_i(\hat{\underset{\sim}{\theta}}_n^{(r)})} \qquad r = 0,1,\ldots$$

for $j = 1,\ldots,k-1$. Further

$$b_j = \sum_{i=1}^{m+1} \frac{(p_i - \pi_i(\hat{\underset{\sim}{\theta}}_n^{(r)}))(\beta_{ji} - \beta_{ki})}{\pi_i(\hat{\underset{\sim}{\theta}}_n^{(r)})} + \sum_{i=1}^{m+1} \frac{(\pi_i(\hat{\underset{\sim}{\theta}}_n^{(r)}) - \beta_{ki})(\beta_{ji} - \beta_{ki})}{\pi_i(\hat{\underset{\sim}{\theta}}_n^{(r)})}$$

$$= \sum_{i=1}^{m+1} \frac{p_i(\beta_{ji}-\beta_{ki})}{\pi_i(\hat{\theta}_{\sim n}^{(r)})} - \sum_{i=1}^{m+1} (\beta_{ji}-\beta_{ki}) + \hat{\theta}_1^{(r)} \sum_{i=1}^{m+1} \frac{(\beta_{1i}-\beta_{ki})(\beta_{ji}-\beta_{ki})}{\pi_i(\hat{\theta}_{\sim n}^{(r)})}$$

$$+ \hat{\theta}_2^{(r)} \sum_{i=1}^{m+1} \frac{(\beta_{2i}-\beta_{ki})(\beta_{ji}-\beta_{ki})}{\pi_i(\hat{\theta}_{\sim n}^{(r)})} + \ldots + \hat{\theta}_{(k-1)}^{(r)} \sum_{i=1}^{m+1} \frac{(\beta_{(k-1)i}-\beta_{ki})(\beta_{ji}-\beta_{ki})}{\pi_i(\hat{\theta}_{\sim n}^{(r)})}$$

for $j = 1,\ldots,k-1$. Then by using $\sum_{i=1}^{m+1} (\beta_{ji}-\beta_{ki}) = 0$, we can write $\underset{\sim}{b}$ as

$$\underset{\sim}{b} = \underset{\sim}{q}(\hat{\theta}_{\sim n}^{(r)}) + R(\hat{\theta}_{\sim n}^{(r)})\ \hat{\tau}_{\sim n}^{(r)} \qquad r = 0,1,\ldots$$

where $\hat{\tau}_{\sim n}^{(r)} = (\hat{\theta}_1^{(r)},\ldots,\hat{\theta}_{(k-1)}^{(r)})'$ and $\underset{\sim}{q}(\hat{\theta}_{\sim n}^{(r)})$ is a $(k-1)$-dimensional vector whose jth element is defined as

$$q_j(\underset{\sim}{\theta}) = \sum_{i=1}^{m+1} \frac{p_i(\beta_{ji}-\beta_{ki})}{\pi_i(\underset{\sim}{\theta})} \tag{B.6}$$

for $j = 1,\ldots,k-1$. Hence by substituting (B.5) in (B.4), we have

$$R(\hat{\theta}_{\sim n}^{(r)})\ \underset{\sim}{\tau} = \underset{\sim}{q}(\hat{\theta}_{\sim n}^{(r)}) + R(\hat{\theta}_{\sim n}^{(r)})\ \hat{\tau}_{\sim n}^{(r)} \qquad r = 0,1,\ldots \tag{B.7}$$

whose root $\hat{\tau}_{\sim n}^{(r+1)} = (\hat{\theta}_1^{(r+1)},\ldots,\hat{\theta}_{(k-1)}^{(r+1)})'$ constitutes the estimates $\hat{\theta}_1^{(r+1)},\ldots,\hat{\theta}_{(k-1)}^{(r+1)}$ of $\theta_1^*,\ldots,\theta_{k-1}^*$ respectively which together with the estimate of θ_k^* given by $\hat{\theta}_k^{(r+1)} = 1 - \sum_{j=1}^{k-1} \hat{\theta}_j^{(r+1)}$ form $\hat{\underset{\sim}{\theta}}^{(r+1)} = (\hat{\theta}_1^{(r+1)},\ldots,\hat{\theta}_k^{(r+1)})'$ being the estimate of $\underset{\sim}{\theta}^*$ obtained after the $(r+1)$th cycle of the iteration process for $r = 0,1,\ldots$.

Theorem B.1. Let $\hat{\tau}_{\sim n}^{(r)} = (\hat{\theta}_1^{(r)},\ldots,\hat{\theta}_{(k-1)}^{(r)})'$ be the realization of a random vector $T_{\sim n}^{(r)} = (Z_1^{(r)},\ldots,Z_{(k-1)}^{(r)})'$ for $r = 0,1,\ldots$ and let $\underset{\sim}{\tau}^* = (\theta_1^*,\ldots,\theta_{k-1}^*)'$. If $\hat{\tau}_{\sim n}^{(0)}$ is chosen so that $T_{\sim n}^{(0)} - \underset{\sim}{\tau}^* = o_p(n^{-\frac{1}{4}})$ as $n \to \infty$, then $\hat{\tau}_{\sim n}^{(1)}$ has the property that $T_{\sim n}^{(1)}$ is consistent and its asymptotic distribution is a $(k-1)$-variate normal distribution with mean vector $\underset{\sim}{\tau}^*$ and covariance matrix given by $\frac{1}{n} R^{-1}(\underset{\sim}{\theta}^*)$.

Proof. Let $q_j(\underset{\sim}{\theta})$ given by (B.6) be the realization of a random function $Q_j(\underset{\sim}{\theta})$ so that

$$Q_j(\underset{\sim}{\theta}) = \sum_{i=1}^{m+1} \frac{P_i(\beta_{ji} - \beta_{ki})}{\pi_i(\underset{\sim}{\theta})} \tag{B.8}$$

for $j = 1,\ldots,k-1$, where P_i, $i = 1,\ldots,m+1$ are as in Theorem 4.2.1. Putting $\underset{\sim}{Q}(\underset{\sim}{\theta}) = (Q_1(\underset{\sim}{\theta}),\ldots,Q_{(k-1)}(\underset{\sim}{\theta}))'$, from (B.7) we have

$$\underset{\sim}{T}_n^{(1)} = \underset{\sim}{T}_n^{(0)} + \underset{\sim}{R}^{-1}(\underset{\sim}{Z}_n^{(0)})\ \underset{\sim}{Q}(\underset{\sim}{Z}_n^{(0)}) \tag{B.9}$$

where $\underset{\sim}{Z}_n^{(r)}$ $r = 0,1,\ldots$ is a random vector whose realization is $\hat{\underset{\sim}{\theta}}_n^{(r)}$. Write $Z_j^{(0)} = \theta_j^* + \varepsilon_j$ for $j = 1,\ldots,k-1$ and $P_i = \pi_i(\underset{\sim}{\theta}^*) + \eta_i$ for $i = 1,\ldots,m+1$. Then $\varepsilon_j = o_p(n^{-\frac{1}{4}})$ as $n \to \infty$ and since P_i is a random variable admitting first and second moments, $\eta_i = o_p(n^{-\alpha})$ as $n \to \infty$ for all $\alpha < 1$. Now from (B.1)

$$\pi_i(\underset{\sim}{Z}_n^{(0)}) = \sum_{j=1}^{k-1} Z_j^{(0)}(\beta_{ji}-\beta_{ki}) + \beta_{ki} = \pi_i(\underset{\sim}{\theta}^*) + \sum_{j=1}^{k-1} \varepsilon_j(\beta_{ji}-\beta_{ki})$$

and thus from (B.8), for $\ell = 1,\ldots,k-1$,

$$Q_\ell(\underset{\sim}{Z}_n^{(0)}) = \sum_{i=1}^{m+1} \frac{(\pi_i(\underset{\sim}{\theta}^*) + \eta_i)(\beta_{\ell i}-\beta_{ki})}{\pi_i(\underset{\sim}{\theta}^*)\left[1 + \dfrac{\sum_{j=1}^{k-1}\varepsilon_j(\beta_{ji}-\beta_{ki})}{\pi_i(\underset{\sim}{\theta}^*)}\right]}$$

$$= \sum_{i=1}^{m+1} \frac{(\pi_i(\underset{\sim}{\theta}^*)+\eta_i)(\beta_{\ell i}-\beta_{ki})}{\pi_i(\underset{\sim}{\theta}^*)}\left[1 - \frac{\sum_{j=1}^{k-1}\varepsilon_j(\beta_{ji}-\beta_{ki})}{\pi_i(\underset{\sim}{\theta}^*)} + O(\underset{\sim\sim}{\varepsilon\varepsilon}')\right]$$

where $\underset{\sim}{\varepsilon} = (\varepsilon_1,\ldots,\varepsilon_{k-1})'$ and by using $\sum_{i=1}^{m+1}(\beta_{ji}-\beta_{ki}) = 0$ for $j = 1,\ldots,k-1$, we have

$$Q_\ell(Z_{\sim n}^{(0)}) = \sum_{i=1}^{m+1} \frac{P_i(\beta_{\ell i}-\beta_{ki})}{\pi_i(\theta^*)} - \sum_{j=1}^{k-1} \varepsilon_j \left[\sum_{i=1}^{m+1} \frac{(\beta_{ji}-\beta_{ki})(\beta_{\ell i}-\beta_{ki})}{\pi_i(\theta^*)} \right]$$

$$- S_\ell + O(\varepsilon\varepsilon') \qquad \ell = 1,\ldots,k-1$$

where $S_\ell = \sum_{i=1}^{m+1} \sum_{j=1}^{k-1} \frac{\eta_i \varepsilon_j (\beta_{ji}-\beta_{ki})(\beta_{\ell i}-\beta_{ki})}{\pi_i^2(\theta^*)}$ for $\ell = 1,\ldots,k-1$ and therefore

$$Q(Z_{\sim n}^{(0)}) = Q(\theta^*) - R(\theta^*)\varepsilon - S + O(\varepsilon\varepsilon') \tag{B.10}$$

where $S = (S_1,\ldots,S_{k-1})'$.

Further

$$R^{-1}(Z_{\sim n}^{(0)}) = R^{-1}(\theta^*) + O(\varepsilon) \tag{B.11}$$

and substituting (B.10) and (B.11) into (B.9) yields

$$T_{\sim n}^{(1)} = T_{\sim n}^{(0)} + R^{-1}(\theta^*)\, Q(\theta^*) - \varepsilon - R^{-1}(\theta^*)S + (U+S)\, O(\varepsilon) + O(\varepsilon\varepsilon')$$

where $U = (U_1,\ldots,U_{k-1})'$, $U_\ell = \sum_{i=1}^{m+1} \frac{\eta_i(\beta_{\ell i}-\beta_{ki})}{\pi_i(\theta^*)}$ for $\ell = 1,\ldots,k-1$.
Hence,

$$\sqrt{n}\,(T_{\sim n}^{(1)} - \tau^*) = \sqrt{n}\, R^{-1}(\theta^*)\, Q(\theta^*) + o_p(1) \tag{B.12}$$

where $o_p(1)$ denotes a $(k-1) \times 1$ random vector whose $(k-1)$ elements are all $o_p(1)$. Therefore, by an obvious extension of the lemma 4.2.1, the asymptotic distribution of $\sqrt{n}\,(T_{\sim n}^{(1)} - \tau^*)$ is the same as the limiting distribution of

$$Y = \sqrt{n}\, R^{-1}(\theta^*)\, Q(\theta^*) \quad . \tag{B.13}$$

Now, using the result of the Appendix A, the asymptotic distribution of $P = (P_1,\ldots,P_{m+1})'$ is a $(m+1)$-variate normal distribution with mean

vector $\underset{\sim}{\pi}(\underset{\sim}{\theta}^*) = (\pi_1(\underset{\sim}{\theta}^*),\ldots,\pi_{m+1}(\underset{\sim}{\theta}^*))'$ and covariance matrix $\frac{1}{n}\Sigma$ where Σ is an $(m+1) \times (m+1)$ matrix whose (i,j)th element σ_{ij} is given by the equation (A.1). Define a $(k-1) \times (m+1)$ matrix B with its (i,j)th element B_{ij} is given by

$$B_{ij} = \frac{\beta_{ij} - \beta_{kj}}{\pi_j(\underset{\sim}{\theta}^*)}$$

for $i = 1,\ldots,k-1$ and $j = 1,\ldots,m+1$. By using (B.8), we have $\underset{\sim}{Q}(\underset{\sim}{\theta}^*) = B\underset{\sim}{P}$ and hence by the standard properties of normal distributions, the asymptotic distribution of $\underset{\sim}{Q}(\underset{\sim}{\theta}^*)$ is a $(k-1)$-variate normal distribution with mean vector $B\underset{\sim}{\pi}(\underset{\sim}{\theta}^*)$ and covariance matrix $\frac{1}{n} B \Sigma B'$. Finally, (B.13) shows that the asymptotic distribution of Y is a $(k-1)$-variate normal distribution with mean vector $\sqrt{n}\, R^{-1}(\underset{\sim}{\theta}^*)\, B\, \underset{\sim}{\pi}(\underset{\sim}{\theta}^*)$ and covariance matrix $R^{-1}(\underset{\sim}{\theta}^*)\, B\Sigma B'\, R^{-1}(\underset{\sim}{\theta}^*)$.

Using $\sum_{j=1}^{m+1} (\beta_{ij}-\beta_{kj}) = 0$ for $i = 1,\ldots,k-1$, it is not difficult to see that $B\,\underset{\sim}{\pi}(\underset{\sim}{\theta}^*) = \underset{\sim}{0}$ and further that the (i,j)th element of $(B\,\Sigma\,B')$ is given by

$$(B\,\Sigma\,B')_{ij} = \sum_{\ell=1}^{m+1}\sum_{r=1}^{m+1} B_{i\ell}\sigma_{\ell r} B_{jr}$$

$$= \sum_{\ell=1}^{m+1} B_{i\ell}\sigma_{\ell\ell}B_{j\ell} + \sum_{\substack{\ell=1 \\ \ell \neq r}}^{m+1}\sum_{r=1}^{m+1} B_{i\ell}\sigma_{\ell r}B_{rj}$$

$$= \sum_{\ell=1}^{m+1} \frac{\beta_{i\ell}-\beta_{k\ell}}{\pi_\ell(\underset{\sim}{\theta}^*)}\, \pi_\ell(\underset{\sim}{\theta}^*)\, \frac{\beta_{j\ell}-\beta_{k\ell}}{\pi_\ell(\underset{\sim}{\theta}^*)} - \sum_{\ell=1}^{m+1} \frac{\beta_{i\ell}-\beta_{k\ell}}{\pi_\ell(\underset{\sim}{\theta}^*)}\, \pi_\ell^2(\underset{\sim}{\theta}^*)\, \frac{\beta_{j\ell}-\beta_{k\ell}}{\pi_\ell(\underset{\sim}{\theta}^*)}$$

$$- \sum_{\substack{\ell=1 \\ \ell \neq r}}^{m+1}\sum_{r=1}^{m+1} \frac{\beta_{i\ell}-\beta_{k\ell}}{\pi_\ell(\underset{\sim}{\theta}^*)}\, \pi_\ell(\underset{\sim}{\theta}^*)\, \pi_r(\underset{\sim}{\theta}^*)\, \frac{\beta_{jr}-\beta_{kr}}{\pi_r(\underset{\sim}{\theta}^*)}$$

$$= \sum_{\ell=1}^{m+1} \frac{(\beta_{i\ell}-\beta_{k\ell})(\beta_{j\ell}-\beta_{k\ell})}{\pi_\ell(\underset{\sim}{\theta}^*)} - \left[\sum_{\ell=1}^{m+1} (\beta_{i\ell}-\beta_{k\ell})\right]\left[\sum_{r=1}^{m+1} (\beta_{jr}-\beta_{kr})\right]$$

$$= R_{ij}(\underset{\sim}{\theta}^*).$$

Therefore $B \Sigma B' = R(\underset{\sim}{\theta}^*)$ and thus the covariance matrix of the asymptotic distribution of Y is $R^{-1}(\underset{\sim}{\theta}^*)$. Hence the limiting distribution of $\sqrt{n}\,(\underset{\sim}{T}_n^{(1)} - \underset{\sim}{\tau}^*)$, being the same as the asymptotic distribution of Y , is a (k-1)-variate normal distribution with mean vector $\underset{\sim}{0}$ and covariance matrix $R^{-1}(\underset{\sim}{\theta}^*)$. This completes the proof of the theorem.

We finally remark that by a close examination of the matrix $[n\, R(\underset{\sim}{\theta}^*)]$, we see that it is in fact the Fisher's information matrix for a grouped sample with division points being $t_1,\ldots,t_m$. This shows (analogous to the case k = 2) that under the condition of the theorem B.1, the estimate $\underset{\sim}{T}_n^{(1)}$ is also asymptotically fully efficient with respect to a fixed set of division points $t_1,\ldots,t_m$.

REFERENCES

1. Anderson, T.W. and Darling, D.A. (1952). Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. Ann. Math. Statist., 23: 193-212.

2. Archbold, J.W. (1964). Algebra. 3rd edition. London: Pitman Press.

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

4. Bartlett, M.S. and Macdonald, P.D.M. (1968). "Least-squares" estimation of distribution mixtures. Nature, 217: 195-196.

5. Behboodian, J. (1970). On a mixture of normal distributions. Biometrika, 57: 215-217.

6. Behboodian, J. (1972). Bayesian estimation for the proportions in a mixture of distributions. Sankhyā, B, 34: 15-20.

7. Behboodian, J. (1972). Information matrix for a mixture of two normal distributions. J. Statist. Comput. Simul., 1: 295-314.

8. Blischke, W.R. (1964). Estimating the parameters of mixtures of binomial distributions. J. Amer. Statist. Ass., 59: 510-528.

9. Blischke, W.R. (1965). Mixtures of discrete distributions. In "Classical and contagious discrete distributions" edited by G.P. Patil, Statistical Publishing Society, Calcutta: 351-373.

10. Boes, D.C. (1966). On the estimation of mixing distributions. Ann. Math. Statist., 37: 177-188.

11. Choi, K. and Bulgren, W.G. (1968). An estimation procedure for mixtures of distributions. J. R. Statist. Soc., B, 30: 444-460.

12. Cohen, A.C., Jr. (1967). Estimation in mixtures of two normal distributions. Technometrics, 9: 15-28.

13. Cox, D.R. (1957). Note on grouping. J. Amer. Statist. Ass., 52: 543-547.

14. Cox, D.R. (1959). The analysis of exponentially distributed life-time

with two types of failure. J. R. Statist. Soc., B, 21: 411-421.

15. Cramér, H. (1946). Mathematical Methods of Statistics. Princeton University Press.

16. Darling, D.A. (1957). The Kolmogorov-Smirnov, Cramér-Von Mises tests. Ann. Math. Statist., 28: 823-838.

17. Day, N.E. (1969). Estimating the components of a mixture of normal distributions. Biometrika, 56: 463-474.

18. Deely, J.J. and Kruse, R.L. (1968). Construction of sequences estimating the mixing distribution. Ann. Math. Statist., 39: 286-288.

19. Dick, N.P. and Bowden, D.C. (1973). Maximum likelihood estimation for mixtures of two normal distributions. Biometrics, 29: 781-790.

20. Falls, L.W. (1970). Estimation of parameters in compound Weibull distributions. Technometrics, 12: 399-407.

21. Fryer, J.G. and Robertson, C.A. (1972). A comparison of some methods for estimating mixed normal distributions. Biometrika, 59: 639-648.

22. Gjeddebaek, N.F. (1961). Contribution to the study of grouped observations. Skand. Aktuarietidiskr, 44: 55-73.

23. Gleser, L.J. and Olkin, I. (1972). Estimation for a regression model with an unknown covariance matrix. Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability. Berkeley and Los Angeles, University of California Press, Vol. 1: 541-568.

24. Hardy, G.H., Littlewood, J.E. and Pólya, G. (1967). Inequalities. Cambridge University Press.

25. Hasselblad, V. (1966). Estimation of parameters for a mixture of normal distributions. Technometrics, 8: 431-444.

26. Hill, B.M. (1963). Information for estimating the proportions in mixtures of exponential and normal distributions. J. Amer. Statist. Ass., 58: 918-932.

27. Hinkley, D.V. (1970). Inference about the change-point in a sequence of random variables. Biometrika, 57: 1-17.

28. Hosmer, D.W., Jr. (1973). A comparison of iterative maximum likelihood estimates of the parameters of a mixture of two normal distributions under three different types of samples. Biometrics, 29: 761-770.

29. Kale, B.K. (1961). On the solution of the likelihood equation by iteration processes. Biometrika, 48: 452-456.

30. Kale, B.K. (1962). On the solution of likelihood equations by iteration processes. The multiparameter case. Biometrika, 49: 479-486.

31. Kao, J.H.K. (1959). A graphical estimation of mixed Weibull parameters in life-testing of electron tubes. Technometrics, 1: 389-407.

32. Kulldorf, G. (1961). Contribution to the Theory of Estimation from Grouped and Partially Grouped Samples. Almqvist and Wiksells, Uppsala.

33. Lord, F.M. (1969). Estimating true-score distributions in psychological testing (an empirical Bayes estimation problem). Psychometrika, 34: 259-299.

34. Macdonald, P.D.M. (1969). FORTRAN programs for statistical estimation of distribution mixtures: some techniques for statistical analysis of length-frequency data. Fish. Res. Bd. Canada, Tech. Rept. No. 129.

35. Macdonald, P.D.M. (1971). Comment on "An estimation procedure for mixtures of distributions" by Choi and Bulgren. J.R. Statist. Soc., B, 33: 326-329.

36. Mann, H.B. and Wald, A. (1942). On the choice of the number of intervals in the application of the chi-square test. Ann. Math. Statist., 13: 306-317.

37. Medgyessy, P. (1961). Decomposition of superpositions of Distribution Functions. Publishing House of the Hungarian Academy of Science, Budapest.

38. Mendenhall, W. and Hader, R.J. (1958). Estimation of parameters of mixed exponentially distributed failure time distributions from censored life test data. Biometrika, 45: 504-520.

39. Northan, H.W. (1956). One likelihood adjustment may be inadequate. Biometrics, 12: 79-81.

40. Page, E.S. (1957). On problems in which a change in parameter occurs at an unknown point. Biometrika, 44: 248-252.

41. Pearson, K. (1894). Contributions to the mathematical theory of evolution. Phil. Trans. Roy. Soc., 185A: 71-110.

42. Rao, C.R. (1948). The utilization of multiple measurements in problems of biological classification. J.R. Statist. Soc., B, 10: 159-203.

43. Rao, C.R. (1957). Theory of the method of estimation by minimum chi-square. Bull. Int. Statist. Inst., 35: 25-32.

44. Rao, C.R. (1967). Least squares theory using an estimated dispersion matrix and its application to measurement of signals. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Berkeley and Los Angeles, University of California Press, Vol. 1: 355-372.

45. Rao, C.R. (1970). Advanced Statistical Methods in Biometric Research. Hafner Publishing Company, Darien, Conn.

46. Riesz, F. and Sz-Nagy, B. (1956). Functional Analysis. Blackie, London.

47. Robbins, H. (1948). Mixture of distributions. Ann. Math. Statist., 19: 360-369.

48. Robbins, H. (1956). An empirical Bayes approach to statistics. Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability. Berkeley and Los Angeles, University of California Press, Vol. 1: 157-163.

49. Robertson, C.A. and Fryer, J.G. (1970). The bias and accuracy of moment estimators. Biometrika, 57: 57-65.

50. Rushforth, N.B., Bennet, P.H., Steinburg, A.G., Burch, T.A. and Miller, M. (1971). Diabetes in the Pima Indians: Evidence of bimodality in glucose tolerance distributions. Diabetes, 20: 756-765.

51. Smith, A.F.M. (1975). A Bayesian approach to inference about a change-point in a sequence of random variables, Biometrika, 62: 407-416.

52. Tallis, G.M. and Light, R. (1968). The use of fractional moments for estimating the parameters of a mixed exponential distribution. Technometrics, 10: 161-175.

53. Tan, W.Y. and Chang, W.C. (1972). 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. Ass., 67: 702-708.

54. Teicher, H. (1960). On the mixture of distributions. Ann. Math. Statist., 31: 55-73.

55. Teicher, H. (1961). Identifiability of mixtures. Ann. Math. Statist., 32: 244-248.

56. Teicher, H. (1963). Identifiability of finite mixtures. Ann. Math. Statist., 34: 1265-1269.

57. Thomas, E.A.C. (1969). Distribution free tests for mixed probability distributions. Biometrika, 56: 475-484.

58. Tiago De Oliveira, J. (1965). Some elementary tests for mixtures of discrete distributions. In "Classical and Contagious discrete distributions" edited by G. P. Patil, Statistical Publishing Society Calcutta: 379-384.

59. Whittaker, J. (1973). The Bhattacharyya matrix for the mixture of two distributions. Biometrika, 60: 201-202.

60. Yakowitz, S.J. and Spragins, J.D. (1968). On the identifiability of finite mixtures. Ann. Math. Statist. 39: 209-214.

61. Zacks, S. (1971). The Theory of Statistical Inference. Wiley, New York.