On The Estimation of the Mixing Density Function In The Mixture Of Exponentials

by

Mohamed ABD-EL-Moniem Mahmoud Khalil Soleha

A Thesis Submitted for the Degree of
Doctor of Philosophy

The City University
School of Mathematics, Acturial Science and Statistics

London
March 1988

To My Mother

Zeinab

and My Father

Mahmoud

ABSTRACT

Given a finite number of data points, simulated from a mixture of exponentials, we propose two nonparametric techniques and a kernel method for estimating the mixing density function.

Firstly, an estimation technique based on Laplace transform, is introduced. We suggest a set of assumptions on which an estimation procedure is based. Simulations are presented that demonstrate the behaviour of the estimated mixing density. In this numerical study, some ways of improving the shape of the estimated density have been explored. Recommendations are given for controlling this shape.

A second estimation technique has been proposed by introducing a set of assumptions placing our estimation problem in an optimization form. The generalized simulated annealing algorithm (G.S.A) has been modified to adapt with our estimation setting. A criterion for measuring the performance of the adaptive (G.S.A.) is suggested. A sensitivity analysis of this adaptive algorithm is made, upon which some recommendations for improving its performance have been given.

In both of the above techniques a similarity is found between one of their parameters and the usual smoothing parameter in density estimation context. This is demonstrated by a numerical example in the case of the optimization technique.

A kernel method for estimating the mixing density is introduced within a Bayesian framework. Some characteristics (such as the limiting behaviour and the moment properties) of the derived kernel-type estimator, are studied. A graphical representation of the estimator, under two different values of (r), has been given using different sets of real data.

ACKNOWLEDGEMENTS

The work contained in this thesis was supervised by Professor H.P. Wynn. I am deeply grateful for his guidance, and for the considerable time he spent in giving his advice, criticism and encouragement during the writing of this thesis. I would like to thank [redacted] for initially introducing me to this area of research. I wish to thank all the staff who are involved in the running of the computer center at Imperial College specially [redacted]. I would like to extend my thanks to all staff, research students and friends at I.C. and T.C.U.

Finally, I would like to express my warm gratitude to [redacted] for her invaluable help in typing this thesis.

TABLE OF CONTENTS

CHAPTER ONE

INTRODUCTION

1.1 Estimation of the Mixing Distribution:

The methods of estimating the mixing distribution could be broadly classified into three major kinds: (i) methods depend upon the maximum likelihood approach (either parametrically or nonparametrically) (ii) the Bayesian approach for the estimation of the mixing distribution, and (iii) methods depend upon some approaches other than the previously mentioned ones, such as, for example, the minimum distance method for estimating the mixing distribution (firstly introduced by Wolfwoitz [1957]).

Laird, A [1979] suggested a nonparametric maximum likelihood estimator for the mixing distribution, by assuming that, such an estimator, is a step function having k which is unknown-number of steps, associated with the amount of probability (height) at each step. Having considered the previous assumptions, he proposed an algorithm for computing a non-parametric maximum likelihood estimator of the mixing distribution in this case.

Lindsay, B [1983I] proved certain properties of the maximum likelihood estimators of the mixing distribution, such as the existence, uniqueness, the convergence to the true mixing distribution and the discreteness of these estimators.

Giammo, T [1984] considered the same approach of Lindsey [1983I], calling it as "the distribution space maximum likelihood estimation", and giving some numerical results based, essentially , on the discreteness of the estimator of the mixing distribution.

It is well known that mixtures of distributions occur in the empirical Bayes procedure,

proposed by Robbin, H. [1964], in which the mixing distribution corresponds to a "priori" distribution. Thus, Ralph,E. [1968] and Meeden, G. [1972] had proposed a Bayesian approach for estimating the mixing distribution by constructing a prior probability distribution over the class of all probability distributions on [0, ∞], and using a certain loss function.

Choi, K and Bulgren, W [1968] had used the minimum distance approach for the estimation of the mixing distribution, based upon the minimization of certain distance-
quadratic function.

1.2 Estimation of the Mixing Density Function

This thesis discusses the estimation of the mixing density function in the mixture model, with emphasis on the mixture of two(or more) exponential components. Thus, the problem, concerning us, will be, generally, viewed as estimating the density function - the mixing density in our case - from a finite number of observations.

It is known (see Wegman [1972]) that, current nonparametric density estimates may be found in three basic types, namely the orthogonal series estimators, kernel estimators and the maximum likelihood estimators of the density function.

In fact, the orthogonal series estimates are not densities, they may actually take on negative values. Also, from a philosophical point of view, the class of orthogonal series estimates has at least one major drawback, that is the choice of the series is not made by the data but it is left as an arbitrary choice of the user. This may lead to distorted estimates or, at least, to estimates whose appearence is dictated more by the arbitrary choice of the series than by the data.

The thesis is concerned with discussing the two basic types of density estimators, the kernel

and the maximum likelihood estimators. This thesis also makes three contributions in the context of estimating the mixing density function in the mixture of exponentials setting. In other words, the body of this thesis consists of four parts, which will be seen to be closely connected.

Pointing out the relations between these parts will be helpful in the better understanding of well-known methods, as well as the new proposed techniques, and detecting some fruitful features and characteristics of these suggested techniques.

The second chapter of this thesis reviews the idea of deriving the maximum likelihood and the kernel estimates of the probability density function.

In the third chapter we introduce a Laplace-based technique for estimating the mixing desnity in the mixture of exponential setting. This technique uses an approximation formula for the Laplace integral equation, which is suggested as being compatible with our mixture of exponentials case, and then employs the inversion operator to estimate the mixing density. An estimation procedure, summing up the previous steps, is constructed. A limiting property for our proposed formula has been given, which generalizes the discrete case of Lindsay [1983I].

A demonstration consists of a simulation study is given, in which we clarify the behaviour of the estimated mixing dsensity under changes of its parameters. This numerical study explores some ways of improving the characteristics of the resulting mixing density.

Finally, an extensive set of graphs has been given, representing the estimated mixing density under various changes of its parameters. This graphical representation of the estimated mixing density can be considered as guide line for controlling the shape of this mixing density in the mixture of exponentials case.

The review of the Bayesian interpretation of some nonparametric methods of density estimation, which will be made in the second chapter, expresses the need to considering some Bayesian-based approaches for density estimation. This has been our justification for proposing a theoretical Bayesian-based kernel estimator for the mixing density in our mixed exponentials problem. This estimator has been introduced in the fourth chapter, in an empirical Bayes framework.

In the fourth chapter, an unbiased estimator for the above-mentioned mixing density , $\pi(\lambda)$, is derived. This is done by taking the kernel density function, denoted by $\pi(\lambda,x)$, to be equal to the conditional distribution of λ given a single observation x.

In order to study the limiting features of this theoretical kernel estimator, a lemma will be introduced which will be useful in investigating how well the assumed kernel function (on which our theoretical estimator $\hat{\pi}(\lambda)$ has been based) behaves when one of its parameters (referred to as r) tends to infinity. There are links between the limiting behaviour of $\hat{\pi}(\lambda)$ as $r \rightarrow \infty$, and the behaviour of any other nonparametric density estimator (kernel, M.P.L., etc.) when the smoothing parameter approaches zero. This suggests that (r^{-1}) is the smoothing parameter in our theoretical kernel estimator $\hat{\pi}(\lambda)$.

In this chapter, an investigation of the bias of the proposed estimator $\hat{\pi}(\lambda)$ will be carried out. This aims at judging how far the moments of the derived estimator $\hat{\pi}(\lambda)$ imitate the moments of the underlying true density function $\pi(\lambda)$. At the end of this chapter a graphical representation of our estimator $\hat{\pi}(\lambda)$ is given using different sets of real data. This is done by presenting the estimator for two different values of r, namely, $r = 1$ and $r = 2$. The first case demonstrates the mixed exponential density as a special case of our estimator $\hat{\pi}(\lambda)$.

In the fifth chapter we propose a technique for estimating the mixing density in our exponential case. This is called "the optimization tecnique" of mixing density estimation.

The generalized simulated annealing algorithm (abbreviated by G.S.A) for function optimization [1986], will be employed to perform the optimization process. Some modifications of the (G.S.A) algorithm will be made to adapt it for dealing with our formulation of the optimization problem. By these adaptations the (G.S.A) algorithm will be capable of not only calculating the optimal value of the objective function, associated with the optimal set of estimated probabilities, but also of exploring some new criteria for judging its performance in this estimation case. For example, to assess the performance of the adaptive (G.S.A) algorithm - being applied to our estimation problem - we suggest a certain criterion for measuring the convergence of the algorithm.

Some numerical examples have been given to measure the sensitivity of the adaptive (G.S.A) algorithm to changes in the parameters which specify and affect our estimation problem. In this context, attempts will be made, for pursuing possible ways of improving the algorithm's performance, measured by how far we are successful in achieving the desired characteristics of the algorithm namely, the convergence and the optimality of the results.

Finally we finish this chapter with some comparisons between our optimization technique of estimation and two other density estimation methods, namely, the kernel method and the maximum likelihood method.

CHAPTER TWO

REVIEW OF SOME APPROACHES TO NONPARAMETRIC DENSITY ESTIMATION

CHAPTER TWO

REVIEW OF SOME APPROACHES TO NONPARAMETRIC DENSITY ESTIMATION

2.1 INTRODUCTION

In this chapter, we review two nonparametric approaches to probability density estimation, namely, the maximum penalized likelihood method and the kernel density approach.

The (M.P.L) method is presented. The idea of penalization is discussed, with the concept of the discretized maximum penalized likelihood estimator (D.M.P.L.E). A limiting property of this (D.M.P.L.E) is mentioned. A Bayesian approach to the (M.P.L) method is reviewed, together with a Bayesian rationale for choosing the correct degree of smoothing. Some general features of the (M.P.L) approach are briefly discussed.

The kernel density estimation approach is defined, with emphasis on the technique of optimally choosing the kernel function. Firstly, a Bayesian-based approach for such choice is reviewed. Secondly, an optimization techniques are discussed. The problem of optimally choosing the smoothing parameter is reviewed with reference to some data-based procedures for solving this problem.

A theoretical comparisons between the above approaches are presented. We suggest a compromise between the Bayesian approach and the other two approaches, namely, the (M.P.L) and the kernel approach.

Finally, recommendations are given concerning some other techniques of density estimation.

2.2 MAXIMUM PENALIZED LIKELIHOOD ESTIMATION APPROACH

2.2.1 Penalizing the Likelihood Function

We start by some basic definitions, discuss the main difficulty in finding a nonparametric generalization of the maximum likelihood procedure and the proposed method of Good and Gaskins [1972] to avoid this difficulty.

Definition 2.2.1:

Given a random sample $x_1,\ldots,x_n$ from a density function f defined on the set $\Omega = (a,b)$, we let $H(a,b) = H(\Omega)$ be a manifold in $L_1(\Omega)$. We define the likelihood that a function $f \in L_1(\Omega)$, which gives rise to the random sample, as

$$L(f) = \prod_{j=1}^{n} f(x_j) \tag{2.1}$$

Definition 2.2.2:

Consider the following constrained optimization problem

$$\left.\begin{array}{ll} \max L(f) & \\ \text{subject to } f \in H(\Omega)\,, & \int f(t)dt = 1 \\ \text{and } f(t) > 0 & \forall\ t \in \Omega \end{array}\right] \tag{2.2}$$

Any solution to problem (2.2) is defined to be a maximum likelihood estimate based on the sample $x_1,\ldots,x_n$.

The main difficulty with problem (2.2) is that the likelihood considered as a functional is unbounded. That is, a linear combination of Dirac delta function at the sample points results in a value of infinity for the objective likelihood functional. This is not an acceptable

estimate for the probability density function.

To avoid the delta function candidates in problem 2.2, Good and Gaskins (1977) suggested a penalty functional

$$\Phi : H(\Omega) \rightarrow R_{+}, \tag{2.3}$$

which would evaluate the smoothness of a particular density estimate on an interval scale. Thus, by replacing the likelihood by a penalized likelihood, they defined the the $\Phi-$penalized likelihood of $f \in H(\Omega)$ by

$$\mathcal{L}(f) = \prod_{j=1}^{n} f(x_j)\, e^{-\Phi(f)}, \tag{2.4}$$

for a given sample $x_1,\ldots,x_n$.

Definition 2.2.3:

Consider the constrained optimization problem

$$\left.\begin{array}{lll} \text{maximize } \mathcal{L}(f) & & \\ \text{subject to} \quad f \in H(\Omega)\,, & \int_{\Omega} f(t)dt = 1 & \\ \text{and} \quad f(t) \geq 0\,, & \forall\ t \in \Omega & \end{array}\right] \tag{2.5}$$

Any solution of (2.5) is called a maximum penalized likelihood estimate (M.P.L.E.). This solution is a measurable function $f_n : R^n \rightarrow \Omega$ which maximizes the penalized likelihood (2.5), and where Ω is the class of all continuous probability densities on the real line.

A manifold $H(\Omega) \equiv H(a,b)$ is said to be reproducing kernel Hilbert space (R.K.H.S) if for every $T \in [a,b]$ the point evaluation functional $E_t : H(a,b) - R$ defined by

$$E_t(f) = f(t), \tag{2.6}$$

is continuous.

Denoting the inner product in H(a,b) by $<.,.>_H$, the uniqueness feature can be stated by the following theorem.

Theorem 2.1:

Suppose that H(a,b) is (R.K.H.S), then integration over [a,b] is continuous operation, and there exists at least one $f \in H(a,b)$ which integrates to one, is nonnegative and is positive at the sample points $x_1,...,x_n$. Then (2.5) with $H \equiv H(a,b)$ and $\Phi(f) = \alpha<f,f>_H$ for every $\alpha > 0$, has a unique solution.

It has been suggested, here, [1971] by Good and Gaskins, to choose a manifold and penalty function that lead to polynomial splines. The following definition is required:

Definition 2.5:

Let $H_0^k[a,b]$ be a Sobolev space of funcitons defined on a finite interval [a,b] whose first $(k-1)$ derivatives are absolutely continuous and vanish at "a" and at "b" and whose k^{th} derivative $\in L^2[a,b]$. It is well known that $H_0^k[a,b]$ is (R.K.H.S) with inner product defined as

$$<g,h>_{H_0^k} = \int_a^b g^{(k)}(t)\, h^{(k)}(t)dt. \tag{2.7}$$

Theorem 2.2:

If in theorem 2.1, mentioned above, we let $H[a,b] = H_0^k[a,b]$ and $\Phi(f) = \alpha<f,f>_{H_0^k}$ for

every $\alpha > 0$, then the solution of the maximum penalized likelihood exists, is unique and is a polynomial spline of degree (2k).

2.3: Penalizing the Logarithm of the Density

Given a set of observations $x_1,...,x_n$, the penalized log likelihood is defined as

$$v(f) = \sum_{i=1}^{n} \log f(X_i) - \alpha\Phi(f), \tag{2.8}$$

where $\Phi(f)$ is a certain functional such as $\int(f'')$ and the parameter α controls the amount by which the data are smoothed to give the estimate. The above estimate, introduced by Silverman B (1982), is an alternative for Good-Gaskins estimator, by which the logarithm of the density (rather than the density itself) will be penalized for roughness.

Silverman B (1982) illustrated his idea by considering a special case, in which the penalty,

$$\Phi_\ell(f) = \int_{-\infty}^{\infty} [(d/dx)^3 \log f(x)]^2 dx \tag{2.9}$$

is used. He showed that the limiting estimate as the parameter α – in equation (2.8) – tends to infinity will be the normal density with the same mean and variance as the data. A result of this case is that as α varies, the method will give a range of estimates from the infinitly rough sum of delta functions to the infinitely smooth maximum likelihood normal fit to the data.

An interesting feature is that if we substitute in (2.8) using (2.9) we arrive at the following

$$w(f) = \sum_{i=1}^{n} \log f(X_i) - \alpha\, \Phi_\ell(f), \tag{2.10}$$

in which the functional "w" depends only on the logarithm of the density. This guarantees

that any density estimate obtained will be automatically positive.

Leonard (1978) deals with the logarithm of the density. He uses a Bayesian approach to density estimation in which a stochastic process structure is placed on the log density. This represents an example of penalizing for roughness in the logarithm of the density.

2.4 Evaluation of the Idea of Penalization

De Montricher, Tapia and Thompson (1975) have shown that Good-Gaskin approach does not always yield the true solution. They showed that in case of the roughness functional,

$$\Phi(f) = \int (f'^2/f), \tag{2.11}$$

their approach yields the unique and true solution.

To avoid having to deal with a nonnegative constraint on f, Good and Gaskin suggested an alternative way of expressing the problem. This is done by formulating the problem in terms of the square root of the density function $\gamma = \sqrt{f}$. Thus, the roughness penalty Φ is a functional of the root-density function. De Montricher, Tapia and Thompson (1975) proved that working with $\gamma = f^{\frac{1}{2}}$, is not always equivalent to working with f itself. Also Wegman [1983] concluded that, the price of this nonnegativity trick is to lose the polynomial spline form of the solution. The solution will be an exponential spline instead, also with knots at the sample points.

Silverman B (1982) showed that the limiting case as the parameter α-in relation (2.10) - tends to zero is represented by the functional w. This functional, being dependent only on the logarithm of the density, guarantees us that any density estimates obtained will be positive. This is a major advantage of this method over some other density estimation methods, by which we get some negative estimates.

2.2.5 The Discretized Maximum Penalized Likelihood Estimator

2.2.5.1 The Basic Idea of Discretization

It is not computationally feasible to calculate the spline density estimators described above by theorem (2.2.1). It has been suggested that one should deal with the nonnegativity constraint directly. This avoids the unsatisfactory trick of working with the square root of the density estimator.

A discrete maximum penalized likelihood estimate (D.M.P.L.E) has been proposed as an approximation to the spline maximum penalized likelihood estimate (M.P.L.E) given by theorem 2.2.1 for the Sobolev space H_0^1[a,b].

The idea of (D.M.P.L.E) was to replace the infinite-dimensional problem by a finite-dimensional one. This arises when we restrict attention to (i) piecewise constant simple functions, or (ii) piecewise linear functions defined using a uniform mesh or partition of the interval [a,b].

2.2.5.2 A Discrete Approximation:

Start by recalling definition 2.2.5 for the Sobolev space, and let Ω = [a,b] be a finite interval. Assume a number m, which is moderately sized number (typically m = 40), and cover [a,b] with a regularly spaced mesh of points. The equally spaced mesh will be defined by the nodes $a = t_0 < t_1 < \ldots < t_m = b$, where the mesh interval η is defined by

$$\eta = t_j - t_{j-1} = (b\text{-}a)/m. \tag{2.12}$$

Further, assume that $p(\cdot)$ be a continuous piecewise linear function defined over the mesh $\{t_j\}_{j=0}^m$ and vanishing outside the interval Ω = [a,b]. Let P_m be the space of all such

functions p on the interval [a,b] which are linear on each interval $I_j = [t_j, t_{j+1}]$. Assume that the estimator is completely determined by its value at the mesh nodes, i.e.,

$$p_j = p(t_j), \qquad j = 0,...,m. \tag{2.13}$$

Suppose that

$$\left.\begin{aligned} & p(a) = p(t_0) = p_0 = 0 \\ \text{and} \quad & p(b) = p(t_m) = p_m = 0, \end{aligned}\right] \tag{2.14}$$

then, the linear spline p will belong to $H_0^1(a,b)$, where $H_0^1(a,b)$ is a special case of the Sobolev space $H_0^k(a,b)$. It is a straightforward matter to show that

$$\left.\begin{aligned} & \int_a^b p(t)dt = \eta \sum_{j=1}^{m-1} \\ \text{and} \quad & p(t) \Leftrightarrow p_j \geq , \quad j = 0,...,m. \end{aligned}\right] \tag{2.15}$$

Having mentioned that $p_j = p(t_j)$ - equation (3.2) - a roughness penalty $\Phi(p)$ can be defined as follows.

$$\Phi(p) = \eta \sum_{j=1}^{m} \left\{ \frac{p_j - p_{j-1}}{\eta} \right\} \quad , \tag{2.16}$$

which is a discrete approximation to the integral penalty $\int p'^2$. In other words, the usual infinite dimensional problem, which depends upon maximizing the criterion functional

$$\tilde{L}(f) = \sum_{i=1}^{n} \log f(x_i) - \alpha \int_{-\infty}^{\infty} f'(t)dt, \tag{2.17}$$

has been approximated, specially the differential operator, by a finite differences, represented by (2.16), over values at the mesh nodes.

Definition 2.2.5.1 :

For $x_1, x_2,...,x_n \in [a,b]$ consider the following constrained optimization problem, represented by the following

$$\max L(p_1,...,p_{m-1}) = \sum_{i=1}^{n} \log p(x_i) - \alpha \eta \sum_{j=1}^{m} \left\{ \frac{p_j - p_{j-1}}{\eta} \right\} \tag{2.18}$$

subject to

$$\left.\begin{aligned} &\eta \sum_{j=1}^{m-1} p_j = 1 \\ \text{and} \quad &p_j \geq 0\,, \quad j = 1,\ldots, m-1 \end{aligned}\right] \tag{2.19}$$

The solution to the above problem is called the discrete maximum penalized likelihood estimate (D.M.P.L.E).

2.2.5.3 An Evaluation of the Idea of Discetization

2.2.5.3.1 Remarks on the Discretized Formulation

We make two remarks on the previous formulation, being represented by relation (2.18) and (2.19). Firstly, the penalized likelihood $L(p_1,\ldots,p_{m-1})$ is written indirectly in terms of the $(m-1)$ parameters $p_1,\ldots,p_{m-1}$ and $p(X_i)$ - appearing in (2.18) - is just a linear combination of the two nearest $p(t_j)$. Secondly, with the set of constraints, shown by (2.18), it is ensured that the resulting (D.M.P.L.E) will be (i) nonnegative on the real line, and (ii) integrating to one.

Tapia, Scott and Thompson (1980) have made an important suggestion, by which an improvement of the (D.M.P.L.E) has taken place. They gave a theoretical justification for the intuitive property that $\hat{p}$ will be a good approximation to the exact maximum penalized likelihood estimator with penalty $\Phi(f) = \int f'^2$. In their practical work [1980], they used a penalty representing higher derivatives, namely, the second derivative $\int f''^2$, which is being approximated by the second differences.

2.2.5.3.2 A Limiting Property of the (D.M.P.L.E)

A general conclusion is detected from the above representation, given by relations 2.18 and 2.19, of the (D.M.P.L.E). This suggests that, the (D.M.P.L.E) is an approximation to the spline maximum penalized likelihood estimate (M.P.L.E) given by theorem 2.2.5 for the

Sobolev space $H_0^1[a,b]$. This conclusion is resulted from replacing the infinite-dimensional problem by the finite-dimensional one, which arises by restricting our attention to piecewise linear functions p(·) defined using a regular mesh or a partition of the interval [a,b].

A related point arises here, concerns how good this (D.M.P.L.E) will be in approximating the spline (M.P.L.E). The answer is represented by the fact that the (D.M.P.L.E) approaches the spline (M.P.L.E) as the mesh size η approaches zero. This will be stated by the following theorem.

Theorem 2.2.5.1:

Suppose $\Omega = [a,b]$ is a finite interval, $x_1,...,x_n$ is a fixed sample and that the data outside Ω is ignored. Let η be the size of the mesh - given by (2.12) - used to obtain the discrete maximum penalized likelihood estimate (D.M.P.L.E) guaranteed by theorem 2.2.5 given above. Then the simple function (D.M.P.L.E) converges to $H_0^1[a,b]$ spline maximum penalized likelihooh estimate (M.P.L.E) guaranteed by theorem 2.1 in the sup norm as $\eta \to 0$.

2.2.6 A Bayesian Approach

2.2.6.1 Bayesian Interpretation of the (M.P.L) Approach

The maximum penalized likelihood method has been interpreted using a Bayesian argument. This was supported by an exploratory data analysis procedure for density estimation and bump-hunting introduced by Good and Gaskins (1980). This is done by maximizing a certain score function w. The funciton w is defined as

$$w = w(f) = L - \Phi(f) = L - \alpha \int [\gamma''(x)]^2 dx, \qquad (2.20)$$

where L is the likelihood function and $\Phi(f)$ is the roughness penalty. The function $\Phi(f)$ depends upon the density function f, apart from the proportionality paramter α which depends on the observations.

The coefficient α , appearing in equation (2.20), determines the magnitude of the roughness penalty, and hence is analogous to the smoothing parameter in the context of density estimation methods. According to the Bayesian interpretation of the maximum penalized likelihood method (M.P.L), the coefficient α will be called the "hyperparameter", being a parameter in a prior.

Suppose that the data are categorized in a histogram of J bins, the likelihood L, appearing in (2.20) takes the form

$$L = \sum_{i=1}^{J} n_i \log_e \int_{B_i} f(x)dx, \tag{2.21}$$

where n_i is the sample frequency in the i^{th} bin B_i. Now, recall equation (2.4) and regard $e^{-\Phi(f)}$ as proportional to an "improper" prior density in function space. This space consists of either functions f when estimating a density function (as in (2.20)) or discrete probabilities when estimating the probabilities for categorized data (as in (2.21)). Thus, the maximum penalized likelihooh (M.P.L) method maximizes the pasterior density in the function space.

2.2.6.2 Controlling the Degree of Smoothing : A Bayesian Rationale

The main conclusion of considering the prior of the density f proportion to $\exp(-\alpha\Phi(f))$, is that we will arrive at the fact that the smoothing parameter α is a parameter of the prior - being called the hyperparameter. Thus, the penalized likelihood, in this case, represents the logarithm of the posterior distribution. The maximum likelihood estimate (M.P.L.E) is equivalent to the mode of the posterior distribution over the space of all smooth curves f's.

An important consequence is that we can control the degree of smoothing, using the hyperparameter α, which is analogous to the smoothing parameter in the density estimation context. By varying the hyperparameter α, we get estimates ranging from an infinitely rough

(as α tends to zero) to the infinity smooth maximum likelihood normal fit to the data when α tends to infinity.

The above argument gives us a satisfying rationale for a particular choice of the hyperparameter α (or the roughness functional). Such decision, previously, used to be made either in an ad hoc way or for reasons of mathematical conveniences.

There is an important recommendation, ought to be mentioned, in the context of using the Bayesian interpretation of the maximum penalized likelihood approach. That is, the hyperparameter α be necessary estimated from the data. This had been done by Scott (1976) who developed an iterative data-based approach for estimating the smoothing parameter α. His technique only requires the prior knowledge that the unknown density has a square integrable second derivative.

2.2.7 General Features of the Penalized Likelihood Approach

From a philosophical point of view, the penalized approach, represented by equation (2.20) for example, makes clear the notion that there are two conflicting goals in density estimation. The first, is to maximize fidelity to the data as measured by the log likelihood $\sum \log f(X_i)$, while the second is to avoid estimates which exhibit too much roughness as measured by $\Phi(f)$. The choice of the smoothing parameter α controls the balance (trade-off) between smoothness and goodness-of-fit. Also, the choice of the roughness penalty $\Phi(f)$ determines what kind of behaviour, in the density estimate, is considered to be undesirable in excess.

The wide applicability of the penalized likelihood approach to a variety of density estimation problems, represents an interesting feature of this approach. As an example, the method of nonparametric regression via the penalized log likelihood functional, which is often called "spline smoothing" (see, for example, De Montricher, Tapia and Thompson (1975) for the

relationship to spline methods).

Another attractive feature of the penalized approach is its readiness to be interpreted using Bayesian argument. Having mentioned that - see subsection (2.6), it seems useful to introduce a comparison between the (M.P.L) and the Bayesian approach. This will be done in the next subsections.

2.3 KERNEL DENSITY ESTIMATION APPROACH

2.3.1 Basic Idea and Definitions

Given a random sample $x_1,...,x_n$ from a continuous but unknown density f, Rosenblatt (1956) proposed an estimate of the form

$$\hat{f}_n(x) = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{h(n)} w\left(\frac{x-x_j}{h(n)}\right), \qquad (3.1)$$

where w(y) is a weight function which equals $\frac{1}{2}$ if $|y| < 1$, and is equal to zero otherwise.

Although Rosenblatt suggested generalizing (3.1) to estimates using different bases than step function, the detailed explication of kernel estimators is due to Parzen (1962).

Definition 2.3.1:

Parzen [1962] defines the kernel estimator for f(x) as

$$\hat{f}_n(x) = \int_{-\infty}^{\infty} \frac{1}{h(n)} K\left(\frac{x-y}{h(n)}\right) dF_n(y) = \frac{1}{nh(n)} \sum_{j=1}^{n} K\left(\frac{x-x_j}{h(n)}\right) \qquad (3.2)$$

where the kernel function $K(\cdot)$ satisfies the following conditions:

$$\left.\begin{array}{ll} \text{(i)} & \sup_{-\infty<y<\infty} |K(y)| < \infty \\ \text{(ii)} & \int_{-\infty}^{\infty} |K(y)| d\mu(y) < \infty \\ \text{(iii)} & \lim_{y\to\infty} | yK(y) | = 0 \\ \text{and} & \\ \text{(iv)} & \int_{-\infty}^{\infty} |K(y)| d\mu(y) < \infty \end{array}\right] \qquad (3.3)$$

Rosenblatt asserted that h(n), appearing in equation (3.2), satisfies the following conditions

$$\left.\begin{array}{ll} & \lim_{n\to\infty} nh(n) = \infty \\ \text{and} & \lim_{n\to\infty} h(n) = 0 \end{array}\right] \qquad (3.4)$$

In general, the idea of these estimates (3.1) is that (3.2) is a distribution with mass $\frac{1}{n}$ placed at each of the observation. Thus the expressions (3.1) and (3.2) smear this probability out continuously, according to the choice of the kernel function $K(\cdot)$. In other wards, the kernel estimator is constructed by placing a kernel function $K(x; x_j, h(n))$, where h(n) is the smoothing parameter, over each observation in the data $x_1,...,x_n$, being assumed to be a random sample from the distribution is question.

An important conclusion, conerning the smoothness of the kernel estimate, could be realized from the above definitions. That is, the smoothness of the estimated density depends upon two factors, namely, the smoothness of the kernel function $K(\cdot)$ and the data responsiveness of the window size h(n). Thus, we study each factor, in detail in the next subsections.

2.3.2 An Optimal Kernel Function

2.3.2.1 A Bayesian-based approach

Whittle (1958) approached the problem of finding an optimal kernel for estimating the density

at a point, using a prior information about the density.

The idea, in general, is based upon relating the density estimation problem to the theory of stochastic process. Specifically, he assumes that different values of the density are related by a prior covariance structure, and then obtains integral equations for his posterior estimates.

Whittle considers the kernel

$$K_n(x,y) = w_x(y), \tag{3.5}$$

and assumes that the sample size n is Poisson mean M. An estimate of the unnormalized density $\phi = Mf$ will be as follows

$$\hat{\phi}_n(x) = n \int_{-\infty}^{\infty} w_x(y)\, dF_n(y). \tag{3.6}$$

Assumptions and the method:

Whittle assumes a prior distribution for $\phi(y)$, which has the first moment

$$E_p\{\phi(x)\} = \mu(x), \tag{3.7}$$

where E_p, represents the expectation with respect to the prior distribution. Also, he assumes that for his prior distribution, the second moment takes the form

$$E_p\{\phi(x)\phi(y)\} = \mu(x,y). \tag{3.8}$$

Based upon these assumptions, he arrives at the weighting function w_x (appearing in (3.6) by minimizing Δ^2, where

$$\Delta^2 = E_p E_s[\hat{\phi}_n(x) - \phi(x)]^2, \tag{3.9}$$

where E_S is the expectation over sampling variations. The resulting optimal weighting function (or kernel) is the solution of the integral equation

$$\mu(x)w_X(y) + \int \mu(y,z)w_X(z)dz = \mu(y,x). \tag{3.10}$$

An important advantage of this method is that the prior distribution, need not be known, only the first two moments are required.

An interesting feature of this method shows that Parzen's [1962] kernel estimator (given by (3.2)) is a special case of Whittle's estimator, when there is no prior information available. In this case, he considers a normalized covariance function that is "second-order stationary". This assumption allows him to estimate the entire density with one kernel, i.e., to view his estimator as a Parzen kernel estimator. Also, this stationarity assumption gives Whittle's approach a "time series" flavour.

2.3.3.2 An Optimizatgion Techniques for Optimally Chosen Kernels

Watson and Leadbetter [1963], consider an estimator

$$\hat{f}_n(x) = \int_{-\infty}^{\infty} K_n(x,y)d\,F_n(y) = \frac{1}{n}\sum_{j=1}^{n} K_n(x,x_j), \tag{3.11}$$

where F_n is the empirical distribution function.

Definition 2.3.2:

The integrated mean squared error (I.M.S.E) is defined as

$$\text{I.M.S.E} = \int E\{\hat{f}_n(x) - f(x)\}^2 dx$$

$$= E\int \{\hat{f}_n(x) - f(x)\}^2 dx \tag{3.12}$$

Watson and Leadbetter considered (3.12), as a measure of goodness. They found δ_n, satisfying the equation

$$\delta_n(x-y) = K_n(x-y), \tag{3.13}$$

which minimizes the integrated mean squared error (3.12).

In the previous formulation, the problem of choosing the optimal kernel $K_n(x,y)$ has been casted in an optimization framework, guaranteeing that the resulting estimate be nonnegative.

A main disadvantage associated with this technique is that δ_n, defined by (3.13), depends upon the explicit form of the probability density function to be estimated.

A similar approach, had been pursued by Rosenblatt [1971], who suggested that the optimal kernel $K(\cdot)$ can be estimated by solving the following optimization problem:

$$\text{minimize } \int K^2(x)dx \tag{3.14}$$

subject to:

$$\left.\begin{aligned} &\int K(x)dx = 1 \\ &K(-x) = K(x) \geq 0 \\ &\int x^2 K(x)dx = 1. \end{aligned}\right] \tag{3.15}$$

The above formulation, leads to an estimate of the kernel $K^*(\cdot)$ which is nonnegative and has a finite support. Philosophically, kernels with finite support [Wegman (1972)] seem more attractive than those with infinite support, on the grounds that the resulting density has zero mass in the tails. As far as, the numerical calculations are concerned, kernels with finite support has definite computational advantages.

2.3.3 Determination of the correct degree of smoothing

2.3.3.1 The Smoothing Problem

The smoothing problem does exist when estimating a density function, and manifests itself in the ability of the resulting estimator to explain or to fit the observed data.

The problem is similar to the bias-variance tradeoff, which is well-known in spectral analysis of time series. That is, for a very large smoothing parameter h, we have too smooth estimates representing a small variance at the price of a large bias. On the other hand, by a very small value of the smoothing parameter h we may detect the fine structure (i.e., reducing the bias) observed from the data, but at the expense of high variance.

The point where, the bias and variance of the estimate are both acceptable has largely been a subjective decision best resolved in an interactive mode with the computer.

In conclusion, it is more convincing to use the observations themselves to determine an appropriate degree of smoothing, and this general approach is known as data-based smoothing. Scott [1981] asserts that, the data-based algorithm is that can be embodied in a computer subroutine whose input is the data and where the output is the value of the smoothing parameter h, that is approximately equal to the theoretically optimal, but unknown, value of the smoothing parameter.

2.3.3.2 Approaches For Estimating the Optimal h(n)

We have mentioned the need for data-defined procedures of determining the correct degree of smoothing. Some of these approaches depend upon the optimization of certain criteria for the performance of the density estimate. On the other hand, other approaches have the feature of being iterative or quasi-optimal. A common feature is that they are, to a great extent data-

based approaches.

2.3.3.2.1 A Modified Maximum Likelihood Criterion

Duin,R. [1976], studied the choice of the smoothing parameters for Parzen [1962] estimators of probability density function, being given by equation (3.2).

Firstly, he had considered the following problem with a maximum likelihood criterion for choosing h

$$\max_{h>0} L(h) = \prod_{k=1}^{n} \hat{f}(x_k) \tag{3.16}$$

From the definition of the kernel estimator (3.2), it may be seen that $h = 0$ maximizes L(h), corresponding to an estimate with a Dirac function at each of the sample points and value of zero elsewhere.

Duin and Habbema et al (1976), consider problem (3.16) with a slightly modified maximum likelihood criterion, as follows

$$\max_{h>0} \hat{L}(h) = \prod_{k=1}^{n} \hat{f}_k(x_k) \tag{3.17}$$

where

$$\hat{f}_k(x_k) = \frac{1}{nh} \sum_{j \neq k}^{n} K\left(\frac{x_k - x_j}{h}\right). \tag{3.18}$$

They found the optimal smoothing parameter h^* as a solution to problem (3.18).

We notice, in the formulation represented by relations (3.17) and (3.18), that they omit the contribution of the sample itself in the estimation of the density at that point. This is because each term in the product (3.17) becomes infinite if h becomes zero.

2.3.3.2.2 Data-based smoothing by Cross-Validation

The idea of this method is to define a certain smoothing criterion (referred to as a likelihood-like expression) which measures the ability of the estimator to explain the observed data. The optimal smoothing parameter h^*, is then chosen to maximize this measure of explanation.

Now recall the kernel estimator, defined by Parzen [1962] as

$$\hat{f}_{h,n}(t) = \frac{1}{nh} \sum_{j \neq k}^{n} K\left(\frac{t-x_j}{h}\right). \tag{3.19}$$

and denote by $\hat{f}^{j}_{h,n-1}$ the estimator computed after deleting the j^{th} observation, i.e.,

$$\hat{f}_{h,n-1}(t) = \frac{1}{(nh)} \sum_{j \neq k}^{n} K\left(\frac{t-x_j}{h}\right). \tag{3.20}$$

Now, $\hat{f}^{j}_{h,n-1}$ is not dependent on x_j, and $\hat{f}^{j}_{h,n-1}$ may be taken as a measure of the appropriateness of h as a value of the smoothing parameter. As the supper-script j ranges through the full sample, we obtain n such measures of explanations, by which we define the likelihood-like expression

$$L_h = \prod_{j=1}^{n} \hat{f}^{j}_{h,n-1}(x_i), \tag{3.21}$$

as the smoothing criterion.

Definition 2.3.3.1 Cross-Validated Smoothing Parameters

The optimal smoothing parameter h^* is that value of h, which maximizes the smoothing criterion L_h, given by equation (3.21), and is called the cross-validated smoothing parameter.

Definition 2.3.3.2 Cross-validated Kernel Estimator

The cross-validated kernel estimator $\hat{f}_{h^*,n}$ is that estimator which results from the substitution

by the optimal h^* in the kernel estimator 3.19.

2.3.3.2.3 Evaluation of the Cross-Validation Technique

The main advantage of the cross-validation method is that, it can be considered as a natural develpment (or extension) of the idea of using the likelihood principle to judge the adequacy of fit of a statistical model.

The method has applications to other estimation problems. A wide variety of applications, to ridge regression (Wahba (1979)), spline smoothing [Wahba (1973), (1975) and Boneva (1970)] and density estimation, have demonstrated a good performance of estimators smoothed by cross-validation techniques.

The cross-validation technique, as presented in subsection 3.3.2.2, has a fruitful feature of being able to incorporate some Bayesian-based ideas of smoothing a curve. As far as, constructing a stochastic model for curve smoothing, Wahba [1970] presented some theoretical results in this context. By using such results, the selection of a smoothing criterion L_h - given by equation (3.21)- has been related to the specification of a prior probability measure over a function space.

The main difficulty with this method, is that there is no guarantee of having a consistent cross-validated kernel estimators, when the kernel has a compact support. Schuster and Gregory [1981] have given some conditions under which compact kernel density estimators are not consistent.

Chow, Geman and Wu [1983] suggested, without proof, that kernels with heavy tails will undersmooth densities with light tails (such as densities with compact suport). Hall [1984] has come to a similar conclusion stating that : cross-validation tends to undersmooth, when restricted to finite intervals on which the density is smooth and bounded away from zero.

2.4 THEORETICAL COMPARISONS

2.4.1 The Kernel Approach versus the (M.P.L) Approach

The kernel density estimate has been defined as

$$\hat{f}_n(t) = \frac{1}{nh} \sum_{j=1}^{n} K\Big(\frac{t-x_j}{h}\Big), \qquad (4.1)$$

where $K(\cdot)$ is the kernel function which integrates to one. The estimate $\hat{f}_p$ which maximizes - over all values of f - the penalized likelihood funciton

$$L_\alpha(f) = \sum_{j=1}^{n} \log f(X_j) - \alpha\, \Phi(f) \qquad (4.2)$$

is called the (M.P.L.) estimate. In equation (4.2), $\Phi(f)$ refers to the roughness penalty functional, and α is the smoothing parameter.

Similarities between the two approaches, are summed up in the following remarks:

(i) The parameter α in equation (4.2), represents the smoothing parameter and corresponds to the kernel scaling-parameter h shown by equation (4.1).

(ii) The parameter α controls the balance between smoothness (as represented by the penalty functional $\Phi(f)$) and the goodness-of-fit (as measured by $\sum \log f(X)$). Thus, it is analogous, in its effect, to the kernel smoothing parameter h.

(iii) In equation (4.1) the averaging operation is in fact a smoothing operation. Thus it is similar to giving the parameter α - in equation (4.2)

- values which are greater than zero.

Since the data is used to determine the estimate (4.2) with relatively minor arbitrary choices, the resulting (M.P.L.E) is derived in a more data-oriented (nonparametric) sense. On the other hand, the choice of the kernel function $K(\cdot)$, in relation 4.1 is not determined by the data, but rather by the user.

Secondly, the kernel estimate $\hat{f}_n(t)$, given in (4.1), depends linearly on the kernel function $K(\cdot)$. Thus, the appearence and properties of $\hat{f}_n$ are dictated more by the arbitrary choice of $K(\cdot)$.

Finally. if we take equation (4.2) to be represented by

$$L_\alpha(f) = \sum_{j=1}^{n} \log f(X_j) - \alpha \int_{-\infty}^{\infty} f'(t)dt, \qquad (4.3)$$

we realize the difficulty, in the infinite-dimensional problem, of maximizing the penalized likelihood 4.3. For this reason, Scott, Tapia and Thompson (1980) introduced a discrete approximation to the integral $\int_{-\infty}^{\infty} f'$ using the first differences. In the kernel approach, there is no need for such approximations, because there is no restrictions to estimating on a finite domain as there is with the (M.P.L) method.

2.4.2 The Bayesian Approach versus the (M.P.L.) Approach

Bayesian advocates argued that, the roughness penalties methods, do not incorporate prior estimates for the density. These methods are, in fact, roughly equivalent to a Bayesian approach, where the prior estimate is a uniform density. Consquently, the tails of these posterior estimates will tend to be relatively thicker than under using more reasonably prior estimates.

Apart from the Bayesian approach, the majority of nonparametric density estimation methods have a smoothing parameter which must be chosen by the experimenter. Regarding the penalty function methods, this parameter has to be estimated, preferably using data-based technique (see Wahba (1975) and (1977)).

Another possible drawback of using the (M.P.L) approach, will be associated with using a roughness penalty term $\Phi(f)$ basaed only upon the first derivative of f. In this case, the

resulting estimates will be having jags at the data points.

Because of the previous undesirable features of the (M.P.L) approach, Leonard, T. (1978) proposed an interesting approach for desnity estimation. His approach, could be considered, from a roughness penalty viewpoint, as a device to interpret penalty methods within a prior-informative Bayesian framework.

2.5 CONCLUDING RECOMMENDATIONS

Practically all density estimation methods have the property that the limiting estimate as the amount of smoothing (controlled by α in (4.2) or h in (4.1)) decreases is a sum of spikes at the observations.

From the previous comparisons, we introduce a linkage between the Bayesian analysis and the above two approaches of density estimation. This suggests, choosing the smallest value of α (or h in the case of the kernel estimate) which reveals fine structure without too much oscillary behaviour consistent with "prior knowledge". This remark represents a compromise between the Bayesian approach and the nonparametric approaches of density estimation.

There is a disappointed fact associated with the kernel approach of density estimation. That is, for very small samples, the direct use for computation of formula 4.1, for the kernel estimate, is highly inefficient. Thus, it is more convincing to think of the kernel estimate as a convolution of the data with the kernel. We recommend, here, the use of an integral transform (Laplace transform, for example) to perform this convolution. This remark constitutes the basic idea upon which we introduce the Laplace-based technique of mixing density estimation. This will be done in the next chapter.

Finally, the introduction of the Bayesian interpretation of some nonparamtric density estimation methods, is helpful in expressing the need to consider some Bayesian-based approaches for desnity estimation. This is our justification for proposing a theoretical Bayesian-based kernel estimator for the mixing density in the mixed exponentials problem. This will be given in chapter V.

CHAPTER THREE

THE LAPLACE-BASED TECHNIQUE FOR ESTIMATING THE MIXING DENSITY FUNCTION IN THE MIXED EXPONENTIALS CASE

INTRODUCTION

In this chapter we propose a simple technique for using the Laplace transformation in solving the problem of estimating the mixing density function for the mixture of exponentials problem.

We start with some necessary definitions and notation. Then a set of assumptions will be proposed, upon which the Laplace-based estimation technique has been based. This is done by : (i) introducing an approximation formula for the Laplace integral in the exponential case, and (ii) inverting it to estimate the mixing density function.
An estimation procedure using these steps will be described and applied to some simulated data.

Simulations are presented that illustrate the proposed estimation technique and relate the behaviour of the estimated mixing density to changes of its parameters.
The idea of using the maximum likelihood method is applied to the image density function for estimating the original (or the mixing in our case) density function. An extensive set of figures will show the effect on the shape of the resulting estimated mixing density of varying its parameters.

Some suggestions on the choice of an appropriate degree of smoothing of the estimated mixing density, will be made.

Finally, we make some remarks on the differences between our estimation technique and the kernel approach of density estimation.

3.2 MAIN REQUIREMENTS FOR THE ESTIMATION PROBLEM

3.2.1 Definitions and Notations

We start with some basic definitions, a theorem, and notation which will be helpful in proving the main results, being proposed in the next section of this chapter.

Definition 3.2.1: A function $h(\lambda)$ is called sectionally (or piecewise) continuous in an interval $a \leq \lambda \leq b$ if the interval can be subdivided into a finite number of intervals in each of which the function is continuous and has finite right and left hand limit.

Definition 3.2.2: If there exist a real constants $M > 0$ and α, such that for all λ, we have that

$$|h(\lambda)| < Me^{\alpha\lambda} \tag{2.1}$$

we say that $h(\lambda)$ is a function of exponential order α as $\lambda \to \infty$, or, briefly, is of exponential order.

Theorem 3.2.1:

Suppose that $h(\lambda)$ is a piecewise continuous function, is of exponential order and its Laplace transform is

$$\phi(s) = \int_0^\infty e^{-s\lambda} h(\lambda) d\lambda \tag{2.2}$$

Thus if $\phi^{(n)}(s)$ is the n^{th} derivative of $\phi(s)$ with respect to s, then we have that

$$\mathcal{L}^{-1}\{\phi^{(n)}(s)\} = (-1)^n \lambda^n h(\lambda), \qquad (2.3)$$

where $\mathcal{L}^{-1}$ denotes the inverse Laplace operator.

In equation (2.2), the function $h(\lambda)$ is usually referred to as the original function, and the function $\phi(s)$ as the image function.

Definition 3.2.3: The set of all original functions $h(\lambda)$ will be called the original space and the set of all image functions $\phi(s)$ will be referred to as the image space.

3.2.2 The Uniqueness Feature

In the Laplace terminology, the concept of a mapping (or a transformation) means that the Laplace integral (2.2) expresses a relation which states that every original function $h(\lambda)$ is related to an image function $\phi(s)$. Therefore, one of the main topics, associated with the mapping representation (2.2) is the question of recovering (or estimating) the original function $h(\lambda)$ from the image function $\phi(s)$.

In order to find a solution to our estimation problem we have to pay much attention to the conditions under which the uniqueness of the estimated original function $h(\lambda)$ (mixing in our case) could be guaranteed. Continuity is sufficient in most cases to determine the original function $h(\lambda)$ uniquely from the image function $\phi(s)$.

3.3 THE MAIN RESULTS

3.3.1 Assumptions and Formulation

Assumption 1:

Assume that we define the density function f(x,r) as

$$f(x,r) = \int_0^\infty \lambda e^{-\lambda x} \pi(\lambda,r) d\lambda \qquad (3.1)$$

The above function (3.1), being defined in the image space, is the Laplace transform of the function $\lambda\pi(\lambda,r)$.

Assumption 2:

Suppose that the density function is assumed to take the mixed Gamma form

$$\pi(\lambda,r) = \sum_{k=1}^{m} \alpha_k \frac{\beta_k^r \lambda^{r-1} e^{-\lambda\beta_k}}{\Gamma(r)} , \qquad (3.2)$$

where

$$\sum_{k=1}^{m} \alpha_k = 1, \qquad \alpha_k \geq 0 \qquad (3.3)$$

Under the above assumptions, the mixture density will be taking the following form

$$f(x,r) = \int_0^\infty \lambda e^{-\lambda x} \sum_{k=1}^{m} \alpha_k \frac{\beta_k (\lambda\beta_k)^{r-1} e^{-\lambda\beta_k}}{\Gamma^1(r)}$$

$$= \sum_{k=1}^{m} \alpha_k \frac{\beta_k^r}{(x+\beta_k)^{r+1}} \cdot \frac{1}{\Gamma(r)} \int_0^\infty e^{-\tau} \tau^r \, d\tau$$

$$= \sum_{k=1}^{m} \alpha_k \frac{r\beta_k^r}{(x+\beta_k)^{r+1}} \qquad (3.4)$$

$$= \sum_{k=1}^{m} \alpha_k \, q_k(x;\, r,\, \beta) \qquad (3.5)$$

In equations (3.4) and (3.5) the function $q_k(x;\, r,\, \beta)$ is a Pareto density function.

We notice that the Pareto density function, has a mean equals to $\frac{\beta}{r-1}$, while the expression for $E(X^k)$ is equal to

$$E(X^k) = \frac{k!\ \beta^k}{(r-1)(r-2)..(r-k)} \tag{3.6}$$

where $r > k$.

A special case of the second assumption, being represented by equation (3.2), results when r equals one. Thus, the density function (3.2), takes the form

$$p(\lambda,1) = \sum_{k=1}^{m} \alpha_k\ \beta_k\ e^{-\lambda\beta_k}, \tag{3.7}$$

which represents a mixture of exponentials.

Another way of getting the mixed exponentials (3.7) is found by keeping x fixed in relation (3.4), and letting both r and β_k tend to infinity in such a way that the ratio $\frac{r}{\beta_k}$ equals a constant, say, t_k. This is represented by

$$\frac{r\beta_k^r}{(x+\beta_k)^{r+1}} = \frac{r}{\beta} \cdot \frac{\beta_k^{r+1}}{(x+\beta_k)^{r+1}}$$

$$= \frac{r}{\beta_k} \cdot \frac{1}{\left(1+\frac{x}{\beta_k}\right)^{r+1}}$$

$$= t_k \cdot \frac{1}{\left(1+\frac{t_k x}{r}\right)^{r+1}} \tag{3.8}$$

By taking the limit as r tends of infinity of relation (3.8), we find that the density function $q_k(x;\ r,\beta)$, appearing in equation (3.5), tends to the density $t_k\ e^{-t_k x}$. In other words, we

have that

$$\lim_{r,\,\beta_k} q_k(x;\, r,\, \beta) = t_k\, e^{-t_k x} \tag{3.9}$$

We discuss the implementation of the above result, as follows:

The limiting relation, could be represented by

$$\phi(t) = \sum_{k=1}^{m} \alpha_k\, \lambda_k\, e^{-\lambda_k t}$$

$$(\lambda), \qquad = \int_0^{\infty} \lambda\, e^{-\lambda t}\, d\mu_m(\lambda) \tag{3.10}$$

where, in terms of probability measure, the measure $\mu_m(\lambda)$ has jumps of α_k. Supposing that this measure is entirely concentrated at λ, i.e.,

$$\mu_m = \delta(\lambda) = \lambda e^{-\lambda t}, \tag{3.11}$$

where $\delta(\lambda)$ is the Delta function. Knowing that $\delta(\lambda) = 0$ at $\lambda = \infty$ and $\delta(\lambda) = 1$ at $\lambda = 0$, so that at these limits of integration - appearing in (3.10) - the value of $\delta(\lambda)$ will be zero or one. In other words, equation (3.10) represents an integration with respect to a measure μ_m taking either zero or one, so $\phi(t) = \lambda e^{-\lambda t}$. The final conclusion from the above argument is that the limiting case of our formulation (3.4) gives us the discrete case of Lindsay (1983) as a special case.

3.3.2 An Estimation Procedure

A natural extension of the assumptions, being mentioned in subsection 3.3.1, is to suggest a procedure for estimating the mixing density in our mixture problem.

The basic structure of this estimation procedure consists of the following theoretical steps.

Step 1:

Assume the existence of an asymptotic expansion for the following Laplace transform, which has been proposed in our assumptions.

$$f(x,r) = \int_0^\infty \lambda e^{-\lambda x}\, \pi(\lambda,r) d\lambda. \tag{3.12}$$

Notice that the uniqueness of the inversion of formula 3.12 is being guaranteed by the continuity of the mixing density.

Step 2:

We assume that the asymptotic expansion of the image function f(x,r) could be represented as

$$f(x,r) = \sum_{j=0}^{\infty} \alpha_j\, (x+\beta_j)^{-(r+1)}\, r\beta_j^r\,, \tag{3.13}$$

where α_j, β_j are coefficients to be determined, and the parameter $r > 0$.

Step 3:

Taking a partial sum of the formal series 3.13, it becomes as

$$f(x,r) = \sum_{j=1}^{m} \alpha_j\, (x+\beta_j)^{-(r+1)}\, r\beta_j^r\,, \tag{3.14}$$

This partial sum 3.14 can be considered as an asymptotic approximation of the image function f(x,r).

Step 4:

By employing the inversion operator of Lapalce transform, we recover the original (mixing) density function

$$\pi(\lambda,r) = \sum_{j=1}^{m} \alpha_j \frac{\beta_k (\beta_k \lambda)^{r-1} e^{-\beta_k \lambda}}{\Gamma(r)} \qquad (3.15)$$

Notice that for the case where $r = 1$, we have the following mixing density

$$\pi(\lambda,r) = \sum_{j=1}^{m} \alpha_j \beta_k e^{-\lambda\beta_k}, \qquad (3.16)$$

which is the mixed exponentials density.

As a justification for deriving formula 3.15 of the original mixing density from the image 3.14, we argue that, becasue of having a finite number of terms in relation 3.14 - in the image space -, the inversion operation could be performed term by term to recover the original mixing density by equation 3.15.

3.4 SIMULATION STUDY

3.4.1 Introduction

In this section, we start by recalling the estimation procedure being proposed in the previous section to deal with the problem of estimating the mixing density $\pi(\lambda,r)$ which is given by equation 3.15. Numerical representation for this estimation procedure will employ the maximum likelihood method for estimating the parameters of the mixing density 3.15, which has been recovered by using the procedure.

A new idea will be suggested through our numerical study of the estimated mixing density. This idea will be introduced by trying to explore a possible connection between one of the parameters - being the truncation point m - of the estimated mixing density and the well-known notion of the "smoothing parameter" as one of the key factors in the probability

density estimation context.

A simulated examples, comparing the effect of varing the parameter (which is analogous to the smoothing parameter) will be given. This will be associated with a graphical representation of the results, exhibiting how sharp these variations will be in smoothing the resulting estimated density under different sample sizes. In other words, comparisons will be performed by introducing an extensive set of curves expressing the behaviour of the estimated mixing density, being influenced by variations in sample size n as well as changes in the parameter m. This latter parameter is considered to be analogous to the smoothing parameter.

3.4.2 A Numerical Representation Scheme

We will be recalling the previous idea, being reviewed in subsection 3.2.2 concerning the significance of the Laplace formulation of the mixture setting. This idea assures that dealing with the image function f(x,r) is much easier than with the original density function $\pi(\lambda,r)$, being shown in equation (3.15),

For handling our problem of estimating the mixing density, the following suggested scheme will be as a numerical representation of the estimation procedure, being proposed in subsection 3.3.2. The numerical scheme will aim to solve the proposed integral equation 3.12 for the sake of recovering the mixing density function, which has been given by equation 3.15. In fact, we mean by recovering the density function 3.15, the estimation of its parameters.

By invoking the asymptotic approximation 3.14, being mentioned in the third step of the procedure, and by employing the maximum likelihood criterion, the previous estimation procedure (introduced in (3.3.2)) can be represented by the following scheme :

First:

Start by taking a sample $x_1,...,x_n$, consists of n independent observations from the density function f(x,r) given by 3.14 as

$$f(\underset{\sim}{x},r) = \sum_{k=1}^{m} \alpha_k (x+\beta_k)^{-(r+1)} r\beta_k. \qquad (4.1)$$

We notice that the partial sum in equation (4.1) is determined by the truncation point m.

Second:

Find the likelihood function of these n independent observations. This is represented as follows:

$$L(\underset{\sim}{x},r) = \prod_{j=1}^{n} f(x_j,r)$$

$$= \prod_{j=1}^{n} \left\{ \sum_{k=1}^{m} \alpha_k (x_j+\beta_k)^{-(r+1)} r\beta_k^r \right\}, \qquad (4.2)$$

where, for a given real value of r, the constants α_k and β_k(k=1,...,m) are the parameters to be determined by the maximum likelihood estimation method.

Third:

Maximize the logarithm of the likelihood function 4.2 to estimate the parameters α_k and β_k(k = 1,...,m), i.e.,

$$\max \log L(\underset{\sim}{x},r)$$

$$\equiv \max_{\alpha_k,\beta_k} \sum_{j=1}^{n} \log \left\{ \sum_{k=1}^{m} \alpha_k (x_j+\beta_k)^{-(r+1)} r \beta_k^r \right\} \qquad (4.3)$$

Fourth:

Substitute the estimated values of the parameters $\hat{\alpha}_k$ and $\hat{\beta}_k$(k=1,...,m) in the formula (3.15), given in the fourth step of the previous estimation procedure. This gives us the following estimate of the mixing density function

$$\hat{\pi}(\lambda,r) = \sum_{k=1}^{m} \hat{\alpha}_k \frac{\hat{\beta}_k^r \lambda^{r-1} e^{-\hat{\beta}_k \lambda}}{\Gamma(r)} \qquad (4.4)$$

3.4.3 An Illustrative Examples:

We start by the proposed likelihood function, which has been given by relation (4.3). A maximization of this formulation results two sets of estimates for $\hat{\alpha}_k$ and $\hat{\beta}_k$(k=1,...,m). This values are substituted into equation (4.4) to get the estimate $\hat{\pi}(\lambda,r)$ for the mixing density function.

Different sets of observations (having different sizes) are generated from the same mixture of exponentials which consists of two components with parameters ($\lambda_1 = 3$, $\lambda_2 = 8$), and the corresponding mixing probabilities are ($\pi_1 = .7$, $\pi_2 = .3$), respectively. These sets of observations will represent the data (X_j, $j = 1,...,n$), and will be used in formula (4.3), shown above.

Throughout the estimation process (of the parameters α_k and β_k, $k = 1,...,m$) and for reasons of comparison the value of the parameter r, appearing in equations (4.3) and (4.4), has been unified and taken to be equal to two (i.e., r =2) in all cases of the numerical results.

3.4.3.1 Levels of Analysis

In our numerical example, we have carried out the analysis by using different sample sizes, being classified as

(i) small sample size (as n = 15)

(ii) moderate sample size (as n = 40 or n = 50),

and (iii) large sample size (as n equals 100 or 120).

The value of the parameter m denotes the number of terms (or the truncation point) to be summed up by the internal summation of formula (4.3), and also, represents the number of the parameters $\alpha_k$and β_k(k = 1,...,m) which have to be estimated.

The criterion for the estimation of the above-mentioned parameters is the maximization of the likelihood function given by equation (4.3). These parameters have been estimated for cases where m takes the values 2,5,10 and 15, using different sample sizes for each of the previous values of m.

Having mentioned that we will be changing both the parameter m and the sample size n, it is convenient, for exploring the behaviour of our estimated density, to perform the analysis on two levels. These are

Level 1:

Studying the behaviour of the estimated density, being recovered using a specific value of m, under various values of the sample size n.

Level 2:

Analyzing the sensitivity of the estimated density, being estimated using a specific sample size n, to various values of the parameter m.

3.4.3.2 The Effect of the Sample Size n:

We start by a certain small value of the parameter m, then we increase the sample size and measure the sensitivity of the results of these changes, to the sample size by giving a graphical representation of the estimated density. Then, under a greater value of m, we repeat this process of increasing the sample size and displaying graphically the effect on the estimated density.

We suggest a criterion for judging the resulting estimated density. This is based upon the relative capability of the density in recapturing the true values of the parameters we have started with.

In all the figures, being given in this chapter, a rescaling operation has been carried out to place the different curves onto the same picture. The aim of this operation is to make all the curves having the same height. In figure (3.1), we find that under a small value of m(say m=2), the smaller the value of n the more peaked the estimated density will be.

As the sample size n increases we get a better curve for the estimated density. Our judgement is based upon the criterion that the peak for the estimated curve (when n = 100)

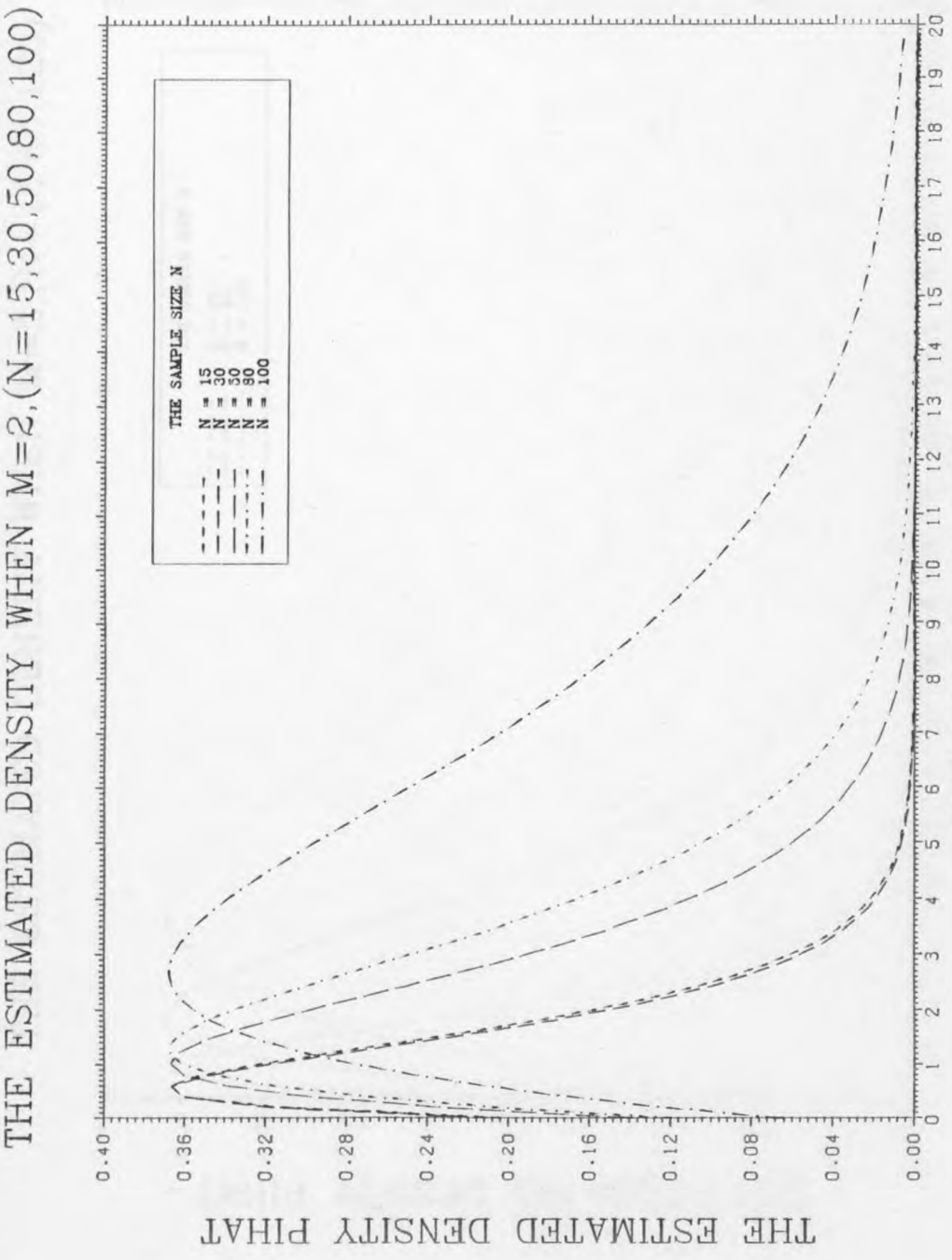
THE ESTIMATED DENSITY WHEN M=2,(N=15,30,50,80,100)
THE SAMPLE SIZE N
N = 15
N = 30
N = 50
N = 80
N = 100
THE ESTIMATED DENSITY PIHAT
0.40
0.36
0.32
0.28
0.24
0.20
0.16
0.12
0.08
0.04
0.00
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
THE VALUES OF LAMBDAHAT

Fig. 3.1

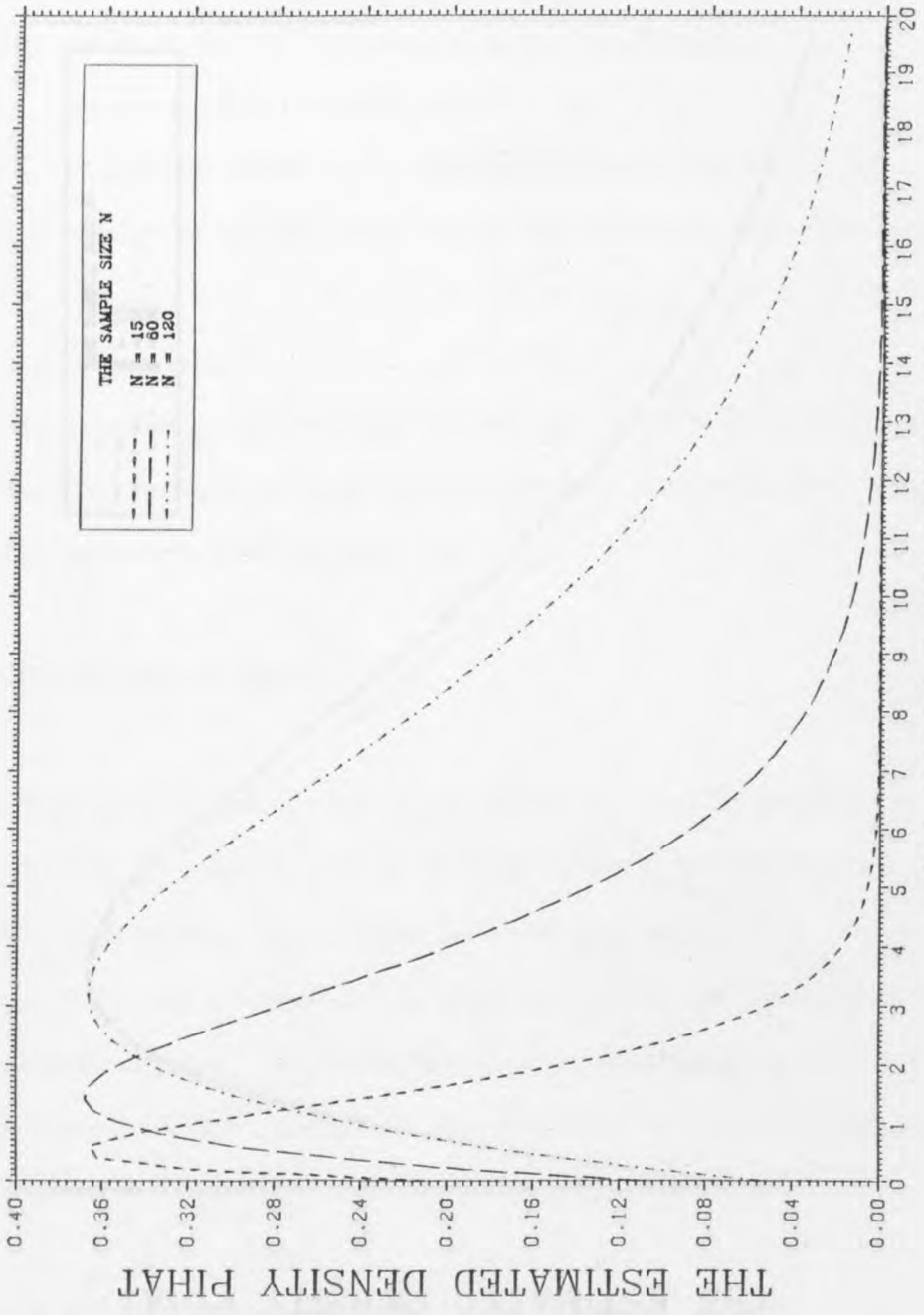
THE ESTIMATED DENSITY WHEN M=5,(N=15,60,120)
THE SAMPLE SIZE N
N = 15
N = 60
N = 120
THE ESTIMATED DENSITY PIHAT
0.40
0.36
0.32
0.28
0.24
0.20
0.16
0.12
0.08
0.04
0.00
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
THE VALUES OF LAMBDAHAT

Fig. 3.2

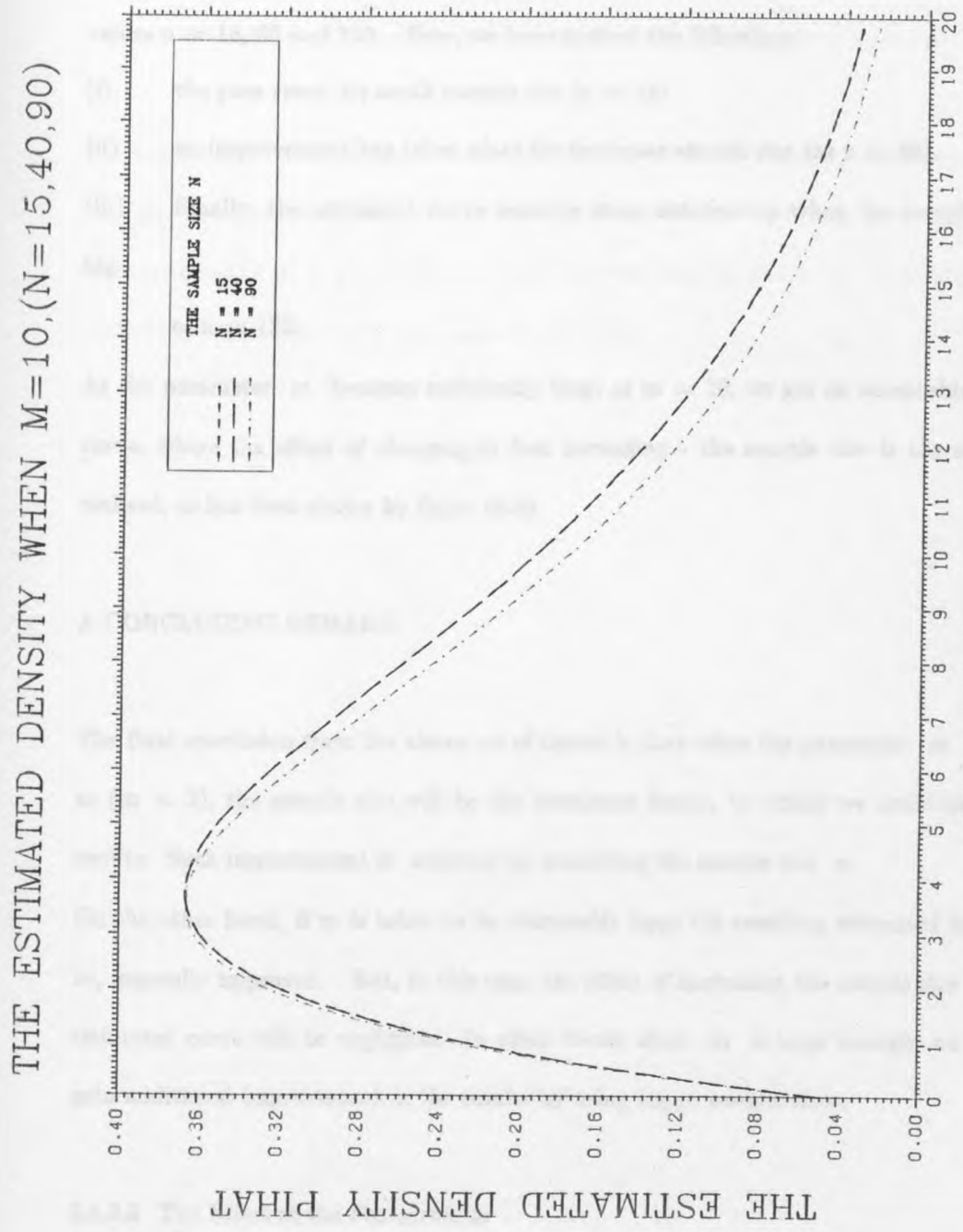
THE ESTIMATED DENSITY WHEN M=10,(N=15,40,90)
THE SAMPLE SIZE N
N = 15
N = 40
N = 90
THE ESTIMATED DENSITY PIHAT
0.40
0.36
0.32
0.28
0.24
0.20
0.16
0.12
0.08
0.04
0.00
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
THE VALUES OF LAMBDAHAT

Fig. 3.3

happens at a value of $\hat{\lambda}$ equal, approximately, one of the true values of the parameter (being $\lambda_1 = 3$).

Notice also, the poor results for the estimated density at moderate values (n = 50).

A greater value of m = 5 (see figure (3.2)) has been used in case the sample size takes the values n = 15, 60 and 120. Here, we have noticed the following:

(i) the poor result for small sample size (n = 15)

(ii) an improvement has taken place for moderate sample size (as n = 60).

(iii) Finally, the estimated curve becomes more satisfactory when the sample size is as big

as n = 120.

As the parameter m becomes sufficiently large as m = 10, we get an acceptable estimated curve, where the effect of changing-in fact increasing - the sample size is too small to be realized, as has been shown by figure (3.3)

A CONCLUDING REMARK

The final conclusion from the above set of curves is that when the parameter m is as small as (m = 2), the sample size will be the dominant factor, by which we could improve the results. Such improvement is achieved by increasing the sample size n.

On the other hand, if m is taken to be reasonably large the resulting estimated density will be, generally improved. But, in this case, the effect of increasing the sample size n on the estimated curve will be negligible. In other words when m is large enough, we could not gain additional improvement in the results by using bigger sample sizes.

3.4.3.3 The Effect of the Parameter m

By using a small size (as n = 15), we notice that (see figure (3.4)), the samller the value of

m(m = 2 or m = 5), the more likely we get an undesirable spiky estimated density. These results (curves) will improve (even at such small n) when the parameter m takes values m = 10 and m = 15.

At a moderate value of the sample size (n = 50), we have a similar desirable effect on the estimated density for the increased values of m = 2, 5 and 10. This is shown in figure (3.5).

Taking samples of size as big as n = 100, we notice that the result is, in general, acceptable. In this case, the improvement in the estimated curve, resulting from increasing the parameter m(m = 2, 4 and 10), is slower than the cases in which m is increasing but the sample is smaller than n = 100. See figure (3.6).

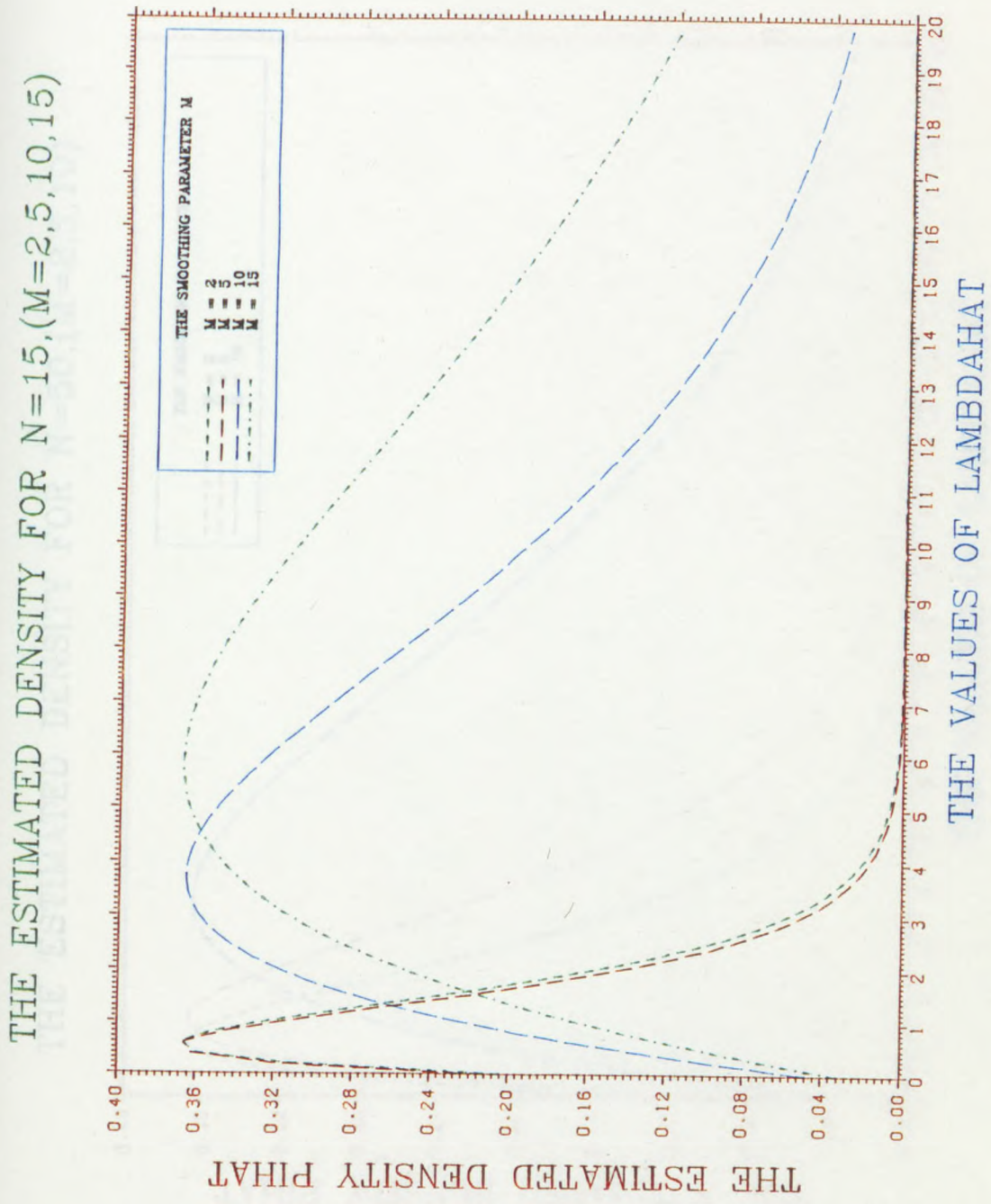
THE ESTIMATED DENSITY FOR N=15,(M=2,5,10,15)
THE SMOOTHING PARAMETER M
M = 2
M = 5
M = 10
M = 15
THE ESTIMATED DENSITY PHAT
0.40
0.36
0.32
0.28
0.24
0.20
0.16
0.12
0.08
0.04
0.00
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
THE VALUES OF LAMBDAHAT

Fig. 3.4

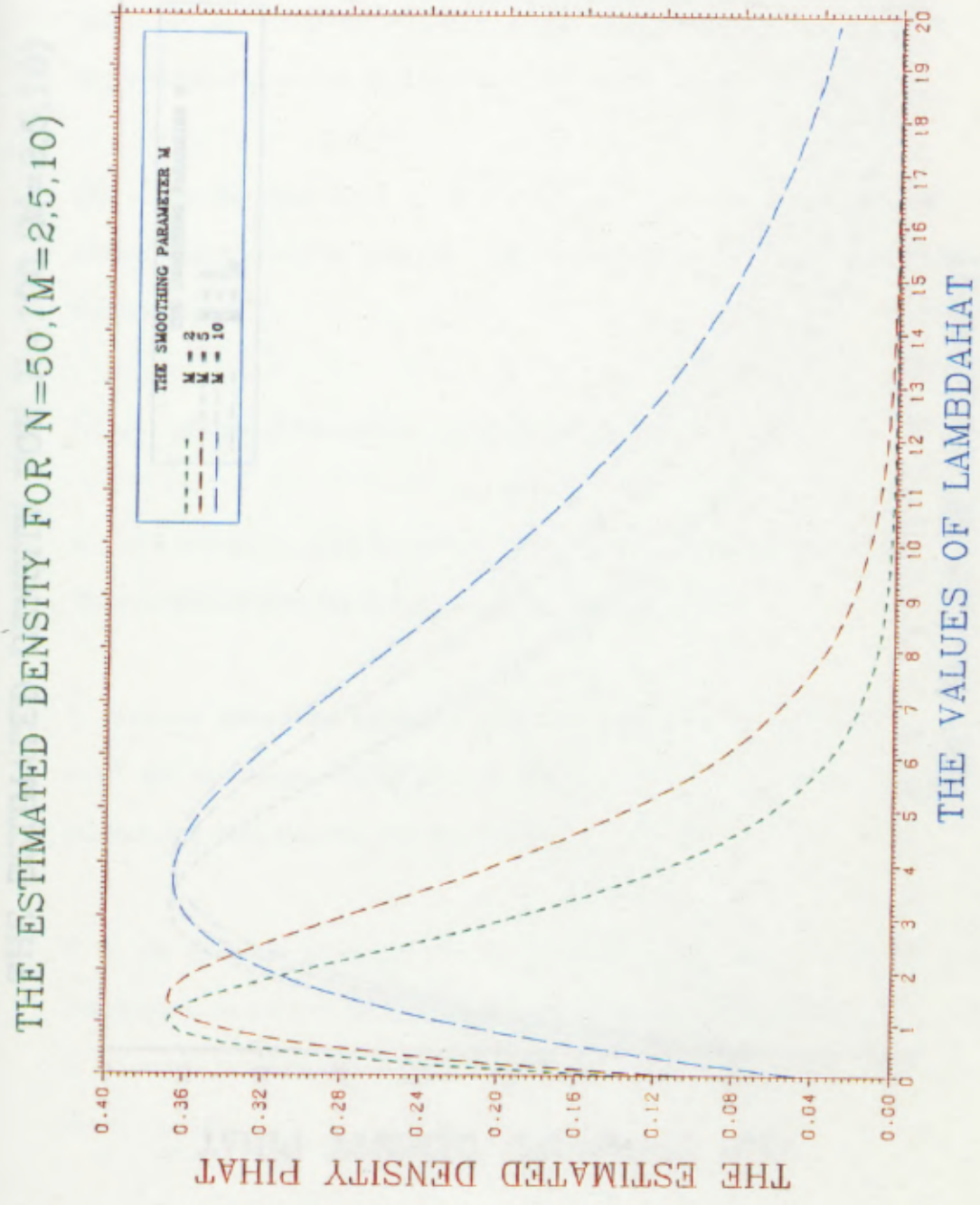
THE ESTIMATED DENSITY FOR N=50,(M=2,5,10)
THE SMOOTHING PARAMETER M
M = 2
M = 5
M = 10
THE ESTIMATED DENSITY PIHAT
0.40
0.36
0.32
0.28
0.24
0.20
0.16
0.12
0.08
0.04
0.00
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
THE VALUES OF LAMBDAHAT

Fig. 3.5

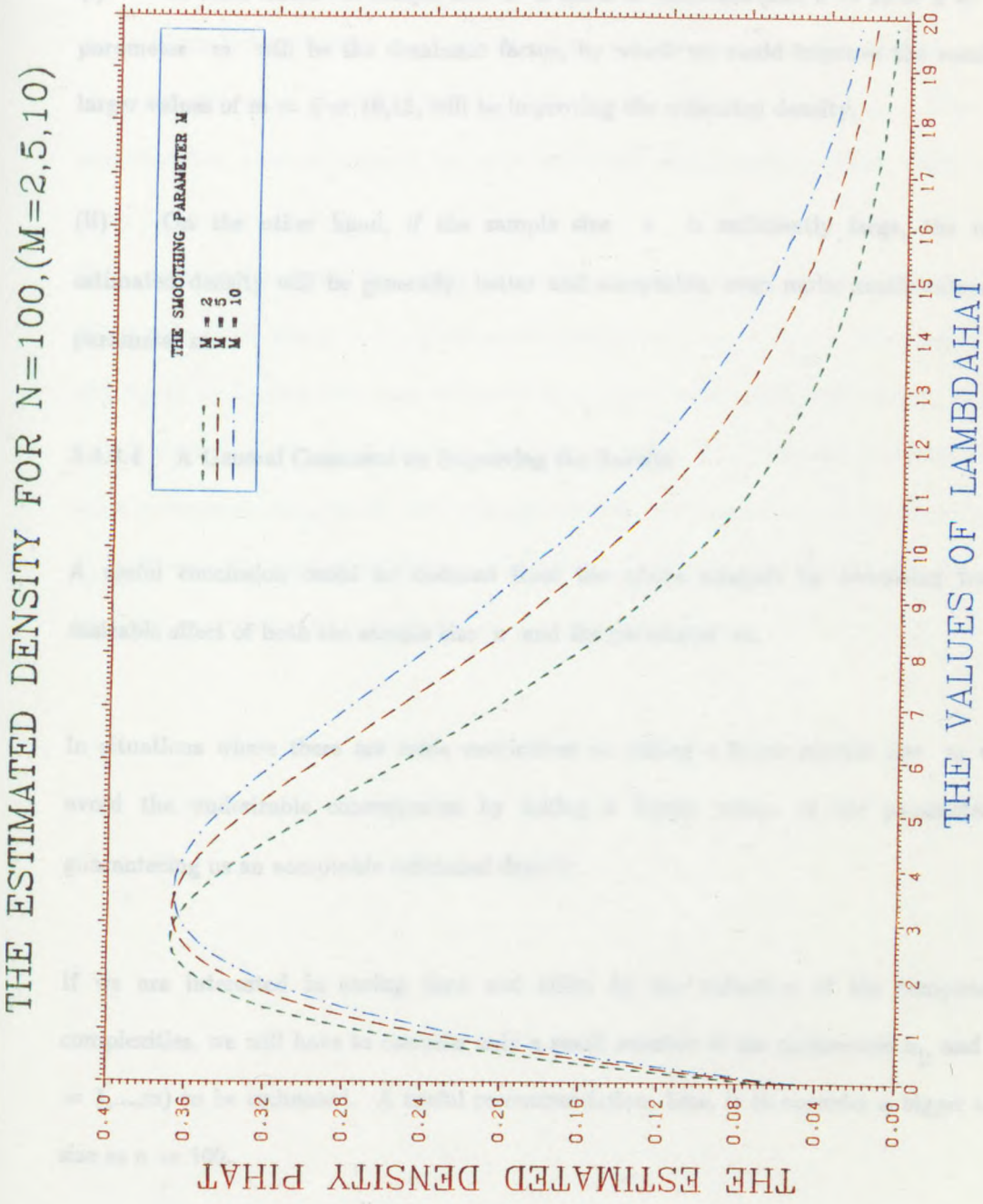
THE ESTIMATED DENSITY FOR N=100,(M=2,5,10)
THE SMOOTHING PARAMETER M
M = 2
M = 5
M = 10
THE ESTIMATED DENSITY PIHAT
0.40
0.36
0.32
0.28
0.24
0.20
0.16
0.12
0.08
0.04
0.00
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
THE VALUES OF LAMBDAHAT

Fig. 3.6

A CONCLUDING REMARK

The final conclusion from the previous set of figures is summed up as follows:

(i) In cases where the sample size n is small or moderate (i.e., n = 15 or n = 50), the parameter m will be the dominant factor, by which we could improve the results. A larger values of m = 5 or 10,15, will be improving the estimated density.

(ii) On the other hand, if the sample size n is sufficiently large, the resulting estimated density will be generally, better and acceptable, even under small values of the parameter m.

3.4.3.4 A General Comment on Improving the Results

A useful conclusion could be deduced from the above analysis by benefiting from the desirable effect of both the sample size n and the parameter m.

In situations where there are some restrictions on taking a larger sample size n, we can avoid the undesirable consequences by taking a bigger values of the parameter m, guaranteeing us an acceptable estimated density.

If we are interested in saving time and effort by the reduction of the computational complexities, we will have to consider only a small number of the parameters α_k and β_k (k = 1,...,m) to be estimated. A useful recommendation, here, is to consider a bigger sample size as n = 100.

In conclusion, our proposed Laplace-based technique of estimating the mixing density is offering us two tuning factors, namely the sample size n and the truncation parameter m.

The latter is found to be analogous to the smoothing parameter in the density estimation context. By these tuning parameters, we can adapt the resulting shape of the estimated mixing density to cope with or even avoid the drawbacks of some imposed limitations and thus widening the chance of getting an acceptable curve for this estimated density.

As an example for choosing an acceptable shape of the estimated mixing density, by picking up a suitable value of m, we give the following set of figure (3.7-3.10). These figures represent four possible views of the same estimated mixing density $\hat{\pi}(\lambda,r)$, which has been given by equation (4.4).

Here, we have found that the choice of a moderate value of $m = 10$ (i.e., taking ten terms of the right-hand side of (4.1)), gives us this acceptable shape of $\hat{\pi}(\lambda,r)$. The parameters α_k and β_k, ($k = 1,...,10$) have been estimated by maximizing the likelihood function (4.3).

In this example a sample of size $n = 40$ (see, also figure (3.3)) has been generated from the same mixture of exponentials which consists of two components with parameters ($\lambda_1 = 3$, $\lambda_2 = 8$) with the associated mixing probabilities ($\pi_1 = .7$, $\pi_2 = .3$). The set of data will represent x_j, $j = 1,...,40$. The value of r is assumed to have a value of $r = 2$.

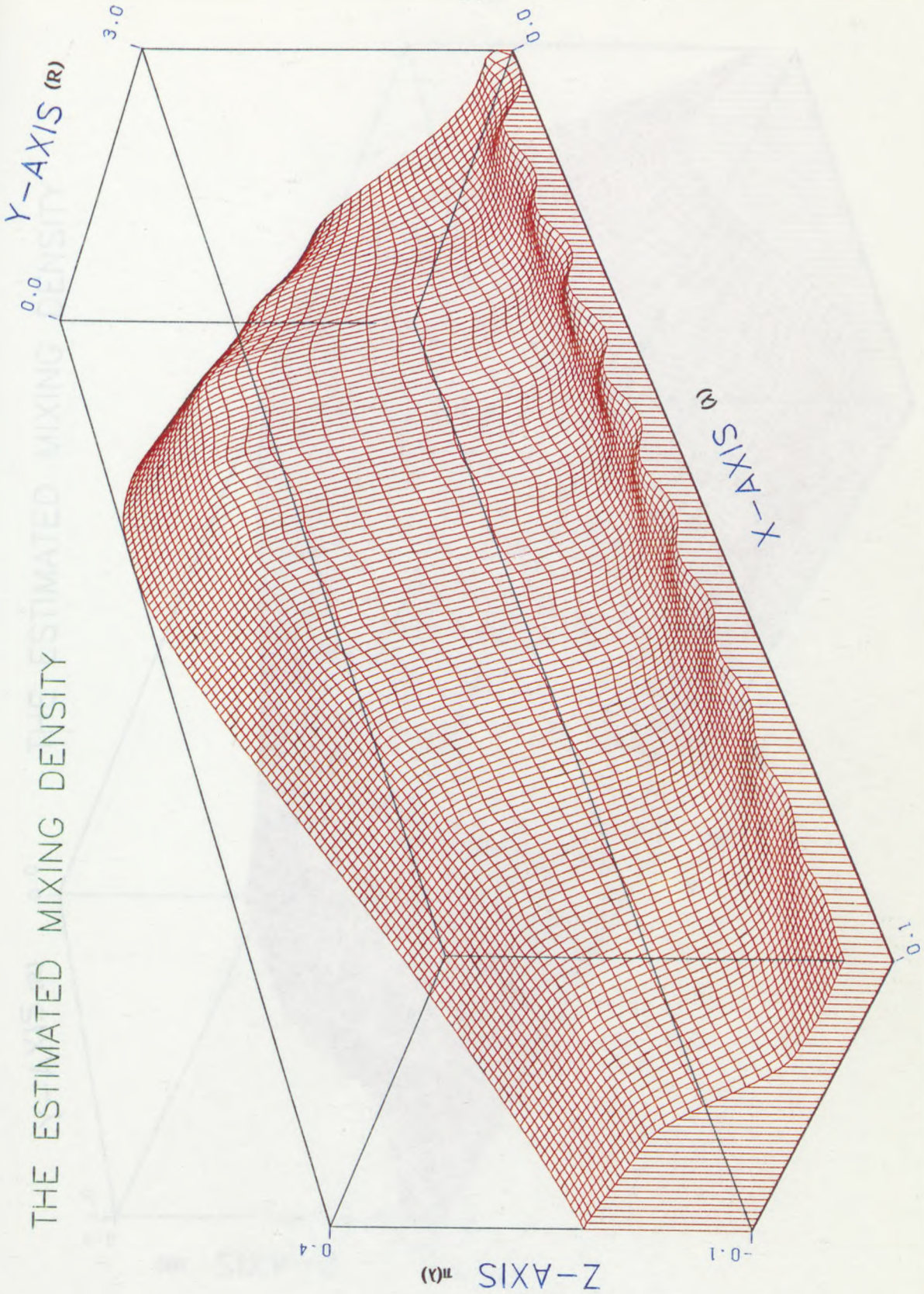
THE ESTIMATED MIXING DENSITY
Y-AXIS (R)
3.0
0.0
0.0
X-AXIS (I)
1.0
-0.1
Z-AXIS π(λ)
0.4
-0.1

Fig. 3.7

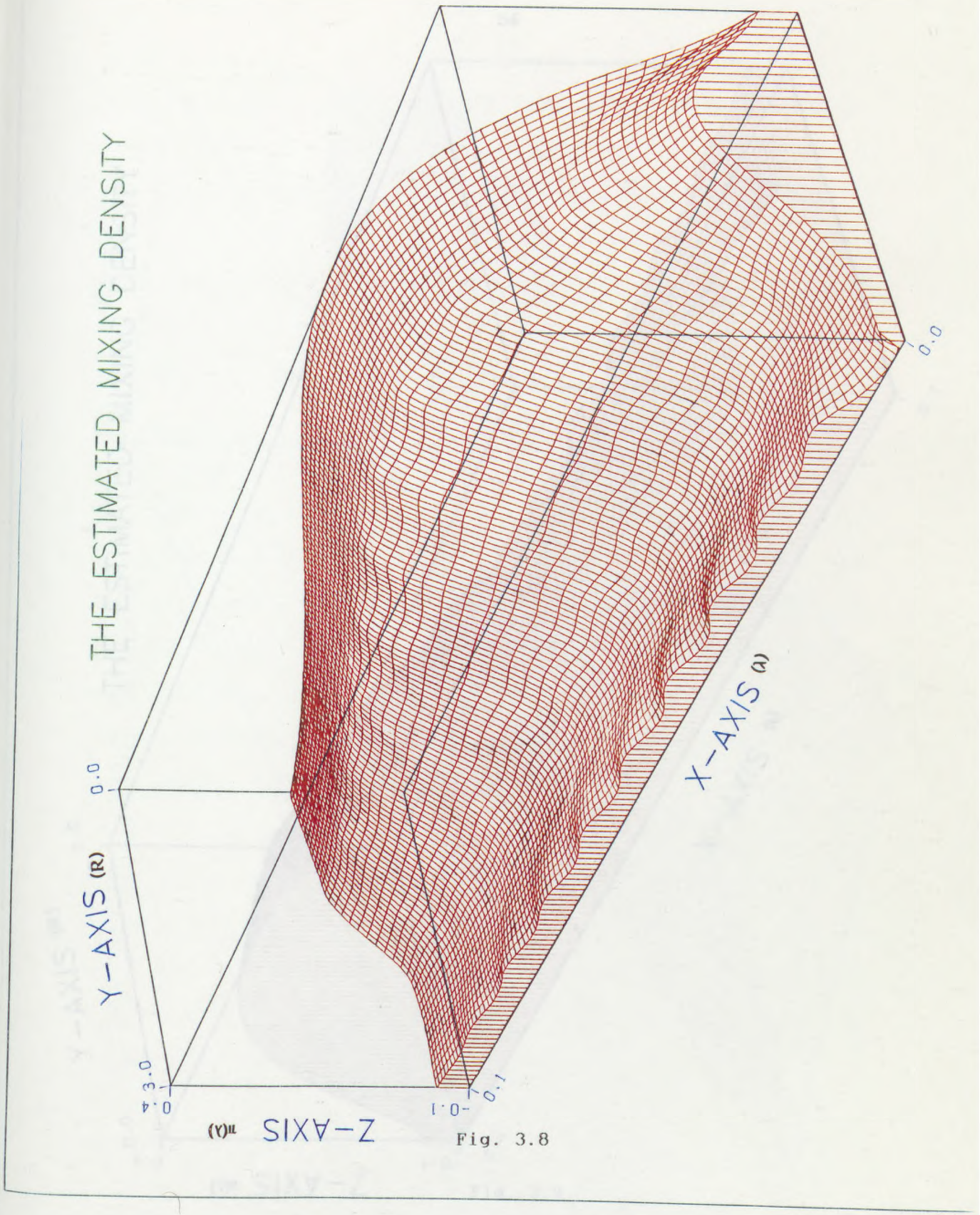
THE ESTIMATED MIXING DENSITY
Y-AXIS (R)
0.0
3.0
0.4
Z-AXIS (Y)
0.1
-0.1
X-AXIS (λ)
0.0

Fig. 3.8

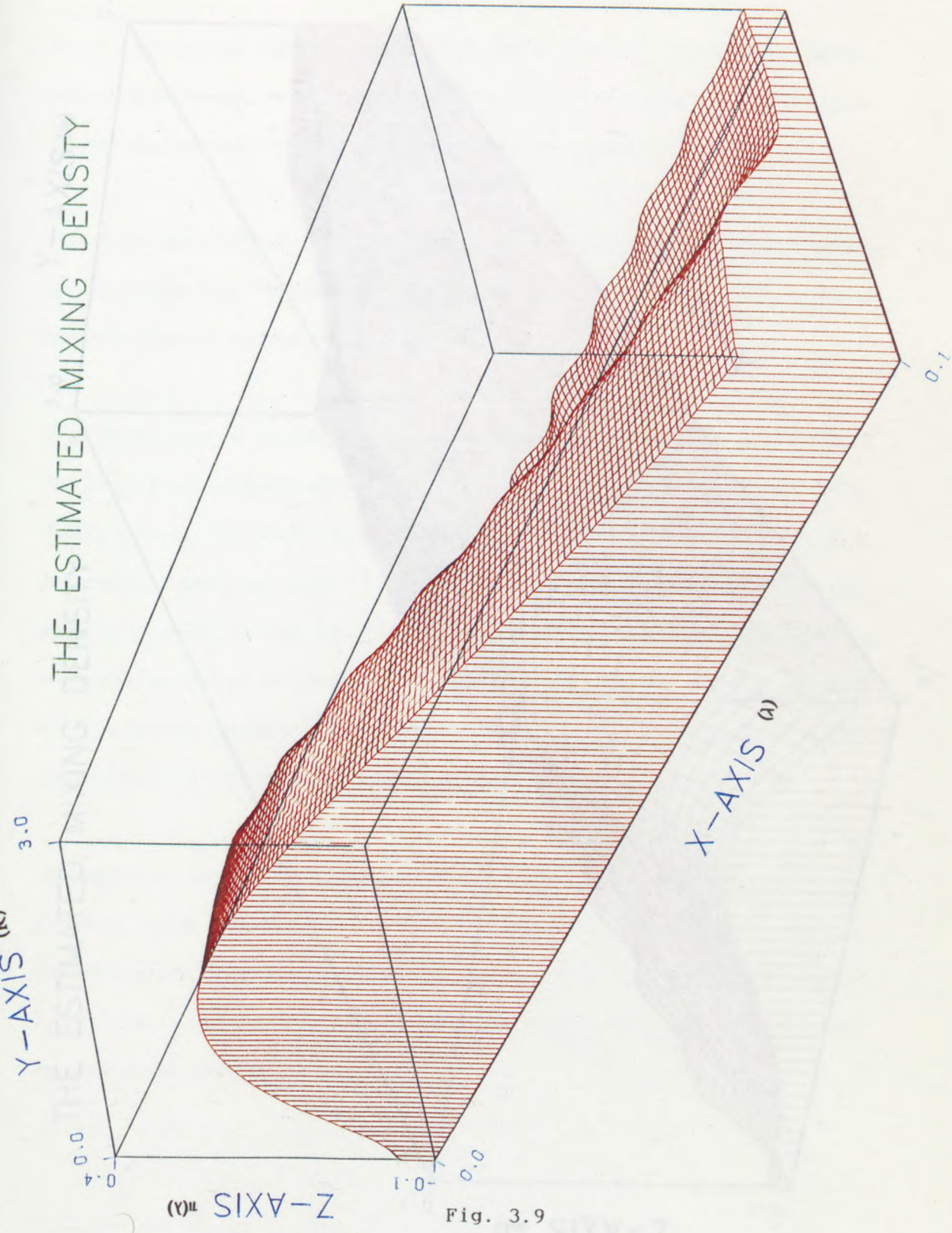
THE ESTIMATED MIXING DENSITY
Y-AXIS (R)
3.0
0.0
Z-AXIS (Y)
0.4
0.0
-0.1
X-AXIS (λ)
0.0
0.1

Fig. 3.9

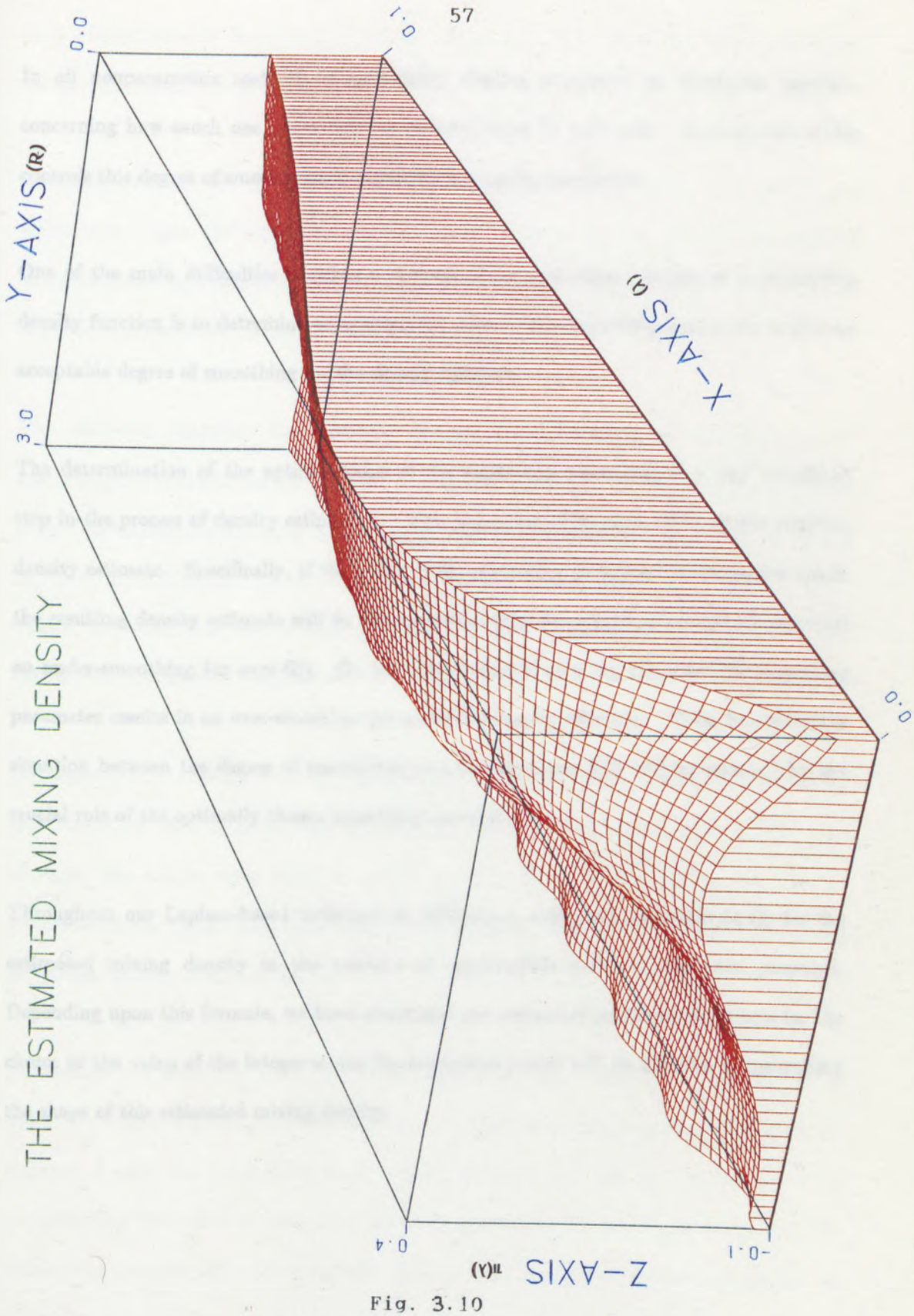
THE ESTIMATED MIXING DENSITY
Y-AXIS (R)
0.0
3.0
X-AXIS (X)
1.0
0.0
Z-AXIS (Y)
0.4
-0.1

Fig. 3.10

3.5 THE DEGREE OF SMOOTHNESS

3.5.1 Introduction

In all nonparametric methods of probability density estimation an inevitable question, concerning how much one does have to smooth, must be answered. A parameter which controls this degree of smoothness is called the smoothing parameter.

One of the main difficulties of using a nonparametric estimation method of a probability density function is to determine an appropriate value of this smoothing parameter to give an acceptable degree of smoothing for the density estimate.

The determination of the optimal value of the smoothing parameter is a very significant step in the process of density estimation. This is becasue of its great effect on the resulting density estimate. Specifically, if the value of the smoothing parameter is chosen too small, the resulting density estimate will be unacceptably spiky (or rough), a case which reprsents an under-smoothing (or over-fit). On the other hand, a very big value for the smoothing parameter results in an over-smoothed (or under-fit) density estimate. Thus, balancing the situation between the degree of smoothness and the goodness-of-fit will be achieved by the crucial role of the optimally chosen smoothing parameter.

Throughout our Laplace-based technique of estimation, a formula (equation (4.4)) for the estimated mixing density in the mixture of exponentials problem, has been proposed. Depending upon this formula, we have structured our numerical results to assess how far the choice of the value of the integer m (or the truncation point) will be effective in controlling the shape of this estimated mixing density.

3.5.2 An analogy of the Smoothing Parameter

It is well-known that the choice of the value of the smoothing parameter in the probability density estimation context is the controlling factor in the final resulting density estimate.

A similar situation will be faced in handling our Laplace-based technique of mixing density estimation. Here, the crucial issue is found to be the determination of a suitable value of the parameter m, (assuming that it is not equal to the sample size n) which gives us an acceptable shape of the mixing density estimate.

The similarity manifests itself through the fact that we have two conflicting goals, concerning the choice of the value of the integer m. That is, if m is chosen too large, this entails an undesirable difficulty in estimating too many parameters (represented by $\hat{\alpha}_k$ and $\hat{\beta}_k$, k = 1,...,m) relative to the available sample size n. On the other hand, if m is chosen too small, important detectable features of the estimated mixing density may not appear in our estimate, being represented by formula 4.4.

Our approach for arriving at an appropriate value of m, is to reach an acceptable situation (see subsection 4.4.3.4) about how sensitive the resulting estimated density would be to changing the sample sizes under a specific value of m. In other words, we use the variations in the sample sizes as the decisive factor in choosing the appropriate value of the parameter m (see subsection 3.4.3.3).

Throughout our numerical example, it has been noticed that changing the sample size n will be having an undesirable sharp effect on the estimated mixing density, which has been constructed using too small value of m relative to the sample size n. In adition to the computational difficulties of estimating too many parameters ($\hat{\alpha}_k$ and $\hat{\beta}_k$, k = 1,...,m) - by maximizing formula 4.3 - the sensitivity of the estimated mixing density (represented by

relation 4.4)) to changes in the sample size n will be unacceptably low.

A balancing situation has been detected, by the analysis of the simulated example, being introduced in the previous subsection 3.4.3. This suggests that, the choice of a moderate value of the parameter m, by which we promote the relative capability of the resulting density estimate (equation (4.4)) in recapturing the true values of the parameters of the mixture we have started with. This value of m should, also, be reasonably sensitive to the changes in the sample size n.

3.6 SOME COMPARATIVE REMARKS

3.6.1 Introduction

Given a kernel function K, which is a probability density function symmetric about zero, a positive smoothing parameter h and a sample $x_1,...,x_n$, the kernel estimate of the density f at each fixed point t is

$$\hat{f}(t) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{t-x_i}{h}\right) \tag{6.1}$$

The kernel estimate (6.1) is nonnegative and integrates to one.

Recall the assumed formulation, which has been mentioned in subsection 3.1, specifically relation 3.4. Using the additional set of assumptions, which is

$$\left.\begin{array}{lll} \text{(i)} & m = n & \\ \text{(ii)} & \alpha_i = \frac{1}{n} & \forall\ i = 1,...,n \\ \text{and} \quad \text{(iii)} & \beta_i = x_i(r-1) & \end{array}\right] \tag{6.2}$$

we get the following estimator represented by

$$\hat{f}(t) = \frac{1}{n} \sum_{i=1}^{n} \frac{r[x_i(r-1)]^r}{[t+x_i(r-1)]^{r+1}} \tag{6.3}$$

This estimator represents a kernel-type estimator, which will be discussed, in details, in the next chapter.

An evaluation of our Laplace-based technique of estimation in the mixture of exponentials setting, will be mentioned. This will be done by considering the formula 6.3 as respresenting the estimator of the density 3.1 in the image space. Then, a comparison will

be performed between this estimator and the usual kernel density estimator, being represented by equation (6.1).

3.6.2 A Theoretical Comparison

For the previously defined estimators, we notice that they have the same nature. In equation (6.1), Rosenblatt (1956) places a kernel function $K(\cdot)$ at each observation x_i, $i = 1,...,n$. Also, in the same sense, our estimator 6.3 places a Gamma $(\lambda; r, x_i(r-1))$ with its maximum $\hat{\lambda} = \frac{(r-1)}{\beta_i}$ at each $\frac{1}{x_i}$, $i = 1,...,n$.

In regard to having an estimate, which is itself a density, the two approaches achieve this goal. That is, if the kernel function K(.) (in equation (6.1)) is simply chosen as a density, then the estimate $\hat{f}$ will, also, be.

By reviewing our assumptions, specially equation (3.2), we realize that this feature does exist.

Concerning the number of observations upon which the density estimate will be based, the Laplace-based approach marks an advantage over the kernel one. In the Laplace-based technique, specially the following formula for the estimator of the mixing desnity function

$$\hat{\pi}(\lambda;r) = \sum_{k=1}^{m} \hat{\alpha}_k \frac{\lambda^{r-1}\hat{\beta}_k^r e^{-\lambda\hat{\beta}_k}}{\Gamma(r)}, \qquad (6.4)$$

it is recommended to avoid taking too many parameters $(\hat{\alpha}_k, \hat{\beta}_k, k=1,...,m)$ relative to the available sample size n. This indicates that our technique rarely invloves more than twenty terms (i.e., the value of the truncation point m in equation 6.4). But, the kernel estimate, being represented by equation (6.1), involves as many terms as there are observations.

Finally, the issue of correctly choosing the parameter m, appearing in equation (6.4), is a significant practical problem in our technique. Because if m is chosen too large, this requires estimating too many parameters α_k and β_k from the relation

$$\max \log L(\underset{\sim}{x};r)$$

$$\equiv \max_{\alpha_k,\beta_k} \sum_{j=1}^{n} \log\Big\{ \sum_{k=1}^{m} \alpha_k (x_j+\beta_k)^{-(r+1)}{}_r\beta_k^r \Big\}. \qquad (6.5)$$

This means, in fact, facing computational difficulties. On the other hand, if m is chosen too small, an important detectable features of the true mixing density may not appear by the estimated density 6.4.

This situation is exactly similar to the problem of optimally choosing the smoothing parameter h in the kernel approach, being represented by equation (6.1).

CHAPTER FOUR

A KERNEL METHOD OF ESTIMATION

4.1 INTRODUCTION

This chapter starts with some basic definitions of the kernel density estimators. In this context, two main approaches, introduced by Rosenblatt (1956) and Parzen (1962), will be defined, associated with some of their main features.

A kernel-type estimation method for the mixing density say, $\pi(\lambda)$ in the mixture of exponentials setting has been proposed. This is done by introducing a theoretical form for the estimator $\hat{\pi}(\lambda)$, which is essentially based upon an empirical Bayes formulation. This estimator has been constructed by assuming the kernel density function to be equal to the conditional distribution of λ given a single observation x. Under this assumption an unbiased estimator of $\pi(\lambda)$ will be derived. In this context an artificial example will be given in which we implement the estimator.

Some limiting features of the proposed theoretical kernel estimator have been studied. In this context, a lemma will be introduced, by which we investigate how well the assumed kernel function (on which the theoretical estimator has been based) behaves for a general density f(x) as r tends to infinity.

Having considered (r^{-1}) as the smoothing parameter in our theoretical kernel estimator $\hat{\pi}(\lambda)$, then the limiting behaviour is shown to be analogous to the behaviour of a density estimator (kernel, M.P.L., etc.) when the smoothing parameter approaches zero.

Also, an investigation of the bias of the proposed estimator will be carried out. We assess

how far the moments of the derived estimator $\hat{\pi}(\lambda)$ mimic the moments of the underlying density function $\pi(\lambda)$.

Finally a graphical representation of the kernel-type estimator will be given for two values of r ,namely, r = 1 and r = 2. the first case gives us the mixed exponential density as a special case of our estimator $\hat{\pi}(\lambda)$. This has been done using different sets of real data.

4.2 GENERAL CONSIDERATIONS

4.2.1 An Abstract Framework

For the sake of clarification, it is appropriate that the general abstract framework, upon which the estimation problem will be based, be given.

Let the random variable X be defined on the probability space ($\mathfrak{S}$, $\mathcal{A}$, P). For reasons of presentation, $\mathfrak{S}$ may be taken to be the real line and $\mathcal{A}$ to be the family of Borel sets. It is assumed that $f = dP/d\mu$ is the desired density function of the random variable X with respect to Lebesgue measure μ.

Let $x_1, x_2, \ldots, x_n$ be a random sample of n independent realizations of the random variable X. The estimate , say, $\hat{f}_n(x)$ is then some specified function of these sample values at the point x.

4.2.2 Some Basic Definitions

An earliest definition of the kernel estimate had been introduced by Rosenblatt (1956), who considered using a central difference of the sample distribution function as an estimate of the density.

Definition 4.2.2.1: Given a random sample $x_1,\ldots,x_n$ from a continuous but unknown

density f, Rosenblatt (1956) proposed an estimate of the form

$$\hat{f}_n(x) = \frac{1}{2h(n)}\Big\{F_n(x+h(n)) - F_n(x - h(n))\Big\}, \tag{2.1}$$

where $F_n(x)$ is the empirical distribution function defined by

$$F_n(x) = \frac{1}{n}\Big\{\text{number of sample points} \leq x\Big\} \tag{2.2}$$

and where $h(n) \to 0$ as $n \to \infty$.

Another representation for the Rosenblatt estimate 2.1, has been given through defining a weight function "w" by

$$w(y) = \begin{cases} \frac{1}{2} & \text{if } | y | < 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.3}$$

Thus the estimate (2.1) is expressed - in this sense - as

$$\hat{f}_n(x) = \frac{1}{n}\sum_{j=1}^{n} \frac{1}{h(n)}\, w\left(\frac{x-x_j}{h(n)}\right), \tag{2.4}$$

where x_1 ,..., x_n are the data points.

An important remark concerning Rosenblatt approach is that, it is simply a histogram which, for estimating the desnity of x, say, has been shifted so that x lies at the centre of a mesh interval. For evaluating the density at another point, say y, the mesh is shifted again so that y is at the centre of a mesh interval.

Although Rosenblatt suggested generalizing (2.4) to estimates using different bases than step functions, the detailed explication of kernel estimators is due to Parzen (1962).

Definition 4.2.2.2: Parzen (1962) defines the kernel estimator for f(x) as

$$\hat{f}_n(x) = \int_{-\infty}^{\infty} \frac{1}{h(n)} K\left(\frac{x-y}{h(n)}\right) dF_n(y) = \frac{1}{nh(n)} \sum_{j=1}^{n} K\left(\frac{x-x_j}{h(n)}\right), \qquad (2.5)$$

where the kernel $K(\cdot)$, is defined to be a Borel measurable function satisfying the conditions:

(i) $\sup_{-\infty<y<\infty} |K(y)| < \infty$

(ii) $\int_{-\infty}^{\infty} |K(y)|\, d\mu(y) < \infty \qquad (2.6)$

(iii) $\lim_{y\to\infty} |yK(y)| = 0$

and

(iv) $\int_{-\infty}^{\infty} K(y) d\mu(y) = 1.$

The idea of the estimate $\hat{f}_n(x)$ is that it has been constructed by placing a kernel function $K(x; x_j, h(n))$, where $h(n)$ is the smoothing parameter, over each observation in the data $X_1,...X_n$, being assumed to be a random sample from the distribution in question.

Concerning the above-mentioned definitions 4.2.1 and 4.2.2 we give the following *remarks*:

(i) The condition that the kernel $K(x; x_j, h(n))$ is a density function guarantees that the estimate $\hat{f}_n(x)$ will also be.

(ii) The kernel function $K(x; y, h(n))$, for example, is a density function with location parameter y and scale parameter $h(n)$.

(iii) The estimate $\hat{f}_n(x)$ has equal weights of $(\frac{1}{n})$ on each of the n kernels centered at the data points.

(iv) The parameter h(n) is a scale parameter which reflects the spread or support of $K(x; x_j, h(n))$.

(v) The choice of the kernel function is very important in determining the properties of the estimate $\hat{f}_n(x)$. Nevertheless, it is generally accepted that the choice of the kernel function is less crucial than determining the value of the smoothing parameter h(n).

An important conclusion, from the last remark on the choice of the smoothing parameter (s), has been commonly accepted. That, it is natural and more convincing to try to use the observations themselves to determine an appropriate degree of smoothing. This generally accepted approach is known as "*data-based* smoothing". The idea of this approach is to define a certain "smoothing" criterion (referred to as a likelihood-like expression) which measures the ability of the estimator to explain the observed data. The optimal smoothing parameter $h^*(n)$ is then chosen to maximize this measure of explanation. The method is usually referred to as data-based smoothing by cross-validation.

4.3 A GENERAL FORMULATION OF THE THEORETICAL KERNEL ESTIMATOR

4.3.1 Empirical Bayes Framework

The methods of kernel estimation, mentioned above in section II of this chapter, are all for a simple density. We seek, here, a kernel estimation method for the mixing density $\pi(\lambda)$. Any such method will be having the assumed form:

$$\hat{\pi}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \pi(\lambda, X_i), \tag{3.1}$$

where the function $\pi(\lambda,\cdot)$ is the kernel function and $X_1,...,X_n$ are the observations in a random sample of size n.

If $\pi(\lambda)$ is known, then a theoretical form for $\hat{\pi}(\lambda)$ can be found using an empirical Bayes formulation as follows:

Suppose that $f(x/\lambda)$ is taken to be the sampling distribution of the random variable X, then the mixing density $\pi(\lambda)$ may be interpreted as a prior density for λ. Thus, the joint density function of X and λ is defined as

$$f(x,\lambda) = f(x/\lambda)\, \pi(\lambda) \tag{3.2}$$

Also, suppose that $\pi(\lambda/x)$ is the conditional distribution of λ given a single observation x, then the posterior distribution for λ, will be

$$\pi(\lambda/x) = \frac{f(x,\lambda)}{f(x)} \tag{3.3}$$

We consider, now, setting the kernel, appearing in (3.1), equal to the conditional distribution of λ given a single observation x. In other wards, assume that

$$\pi(\lambda,x) = \pi(\lambda/x) \tag{3.4}$$

Given an n observations $X_1,...,X_n$ of the random variable X, then the expectation of $\hat{\pi}(\lambda)$ with respect to these observations will be

$$E\{\hat{\pi}(\lambda)\} = E_X[\frac{1}{n}\sum_{i=1}^{n}\pi(\lambda, x_i)]$$

$$= \frac{1}{n}\sum_{i=1}^{n}E_X[\pi(\lambda, x_i)]$$

$$= \frac{1}{n}\sum_{i=1}^{n}E_X[\pi(\lambda/ x_i)]$$

$$= E_X[\pi(\lambda/ x_i)] \tag{3.5}$$

But, we have that

$$E_X[\pi(\lambda/ x_i)] = \int_{-\infty}^{\infty}\pi(\lambda/x)f(x)dx, \tag{3.6}$$

then we have, fom (3.5) that

$$E_X[\pi(\lambda/ x_i)] = \int_{-\infty}^{\infty}f(x,\lambda)dx$$

$$= \pi(\lambda). \tag{3.7}$$

Thus, $\hat{\pi}(\lambda)$ is an unbiased estimator for $\pi(\lambda)$.

In evaluating assumption (3.4), we argue that : despite the fact that $\pi(\lambda)$ is unknown, the above result is promising because it may be helpful in suggesting a suitable form for the kernel.

4.3.2 The Exponential Case

Suppose that

$$f(x/\lambda) = \lambda e^{-\lambda x}, \tag{3.8}$$

and

$$\pi(\lambda) = \frac{\beta(\lambda\beta)^{r-1} e^{-\beta\lambda}}{\Gamma(r)} \tag{3.9}$$

Thus, we have that

$$f(x) = \frac{r\beta^r}{(x+\beta)^{r+1}}. \tag{3.10}$$

knowing that

$$\pi(\lambda/x) = \frac{f(x/\lambda)\pi(\lambda)}{f(x)}, \tag{3.11}$$

then we have that

$$\pi(\lambda/x) = \left(\frac{(x+\beta)^{r+1}}{r\beta^r}\right) \cdot \frac{\beta(\lambda\beta)^{r-1} e^{-\beta\lambda}}{\Gamma(r)} \cdot \lambda e^{-\lambda x}$$

$$= \frac{(x+\beta)^{r+1}}{\Gamma(r+1)} \lambda^r e^{-(x+\beta)\lambda} \tag{3.12}$$

Equation (3.12) represents the Gamma density for λ, i.e., it is Gamma (r+1, x+β). Thus $\hat{\pi}(\lambda,x)$ takes the following form

$$\hat{\pi}(\lambda,x) = \frac{1}{n} \sum_{i=1}^{n} \frac{(x_i+\beta)^{r+1} \lambda^r e^{-(x_i+\beta)\lambda}}{\Gamma(r+1)} \tag{3.13}$$

From the above example, it is seen that the shape parameter r will be considered as a smoothing parameter in same way as the usual scale parameter h in kernel density estimator. One obvious drawback in formula (3.13) is its dependence on β.

Consider the case where $\beta = 0$, then from equation (3.13) we have that

$$\hat{\pi}(\lambda, x_i) = \frac{(x_i)^{r+1} \lambda^r e^{-\lambda x_i}}{\Gamma(r+1)} . \tag{3.14}$$

Recalling that

$$f(x) = \frac{r\beta^r}{(x+\beta)^{r+1}} , \tag{3.15}$$

we can recapture $\pi(\lambda)$ by replacing β by $(x_i+\beta)$ and r by $(r+1)$ in the same way we obtain the kernel appearing in equation (3.14).

By taking equation (3.15) we can obtain the maximum likelihood estimator of β based on a single observation x, as follows

$$\text{Log } f(x) = \text{Log } r + r \text{ Log } \beta - (r+1) \text{ Log } (x+\beta)$$

$$\frac{\partial \text{Log } f(x)}{\partial \beta} = \frac{r}{\beta} - \frac{(r+1)}{(x+\beta)} = 0 \tag{3.16}$$

Equation (3.16) gives us, the maximum likelihood estimator as

$$\hat{\beta}_{MLE} = xr. \tag{3.17}$$

Thus, another way of interpreting the method in the case where $\beta = 0$ and $r = 1$ is to substitute $x_i = \hat{B}_{MLE}$ in equation (3.14), replace r by $(r+1)$ and then recapture $\pi(\lambda)$ in the same way. This suggests a slightly different form for the kernel, represented by

$$\hat{\pi}(\lambda,x) = \frac{(xr)^{r+1} \lambda^r e^{-x\lambda r}}{\Gamma(r+1)} . \tag{3.18}$$

We notice that, the case where r = 1 is established, also, when the Gamma mixing density $\pi(\lambda)$ reduces to an exponential. This case has been demonstrated , by giving a graphical representation of our estimator for two comparative values of r , namely , r = 1 and r = 2. This is applied to different sets of real data shown in Table 4.1 below (see Smith(1986)). The graphs are given at the end of this chapter. Another interpretation is to put $\beta = x_i(r-1)$ in equation (3.13). Note that, the mean of f(x), being given in equation (3.15) is equal to $\frac{\beta}{r-1}$.

TABLE 4.1

60	100	199	141	118	173	156	230	155
51	90	105	143	273	218	173	169	397
83	59	147	98	192	162	125	178	1063
140	80	113	122	238	288	852	271	738
109	128	98	110	105	394	559	129	140
106	117	118	132	398	585	442	568	364
119	177	182	194	108	295	168	115	218
76	98	131	155	182	262	286	280	461
68	158	156	104	130	127	261	305	174
67	107	78	83	170	151	227	326	326
111	125	84	125	181	181	285	1101	504
57	118	103	165	119	209	253	285	374
69	99	89	146	152	141	166	734	321
75	186	124	100	199	186	133	177	169
122	66	71	318	89	309	309	493	426
128	132	65	136	211	192	247	218	248
95	97	220	200	324	117	112	342	350
87	87	109	201	164	203	202	431	348
82	69	93	251	133	198	365	143	265
132	109	171	111	121	255	702	381	293

Each column of Table 4.1 represents a set of data . These are the lifetimes (in units of 1000 cycles) of steel specimens under various stress amplitudes.

4.4 SOME LIMITING PROPERTIES of $\hat{\pi}(\lambda)$

4.4.1 Introduction

In this section we pay attention to the estimator $\hat{\pi}(\lambda,x)$, which has been given by equation (3.18) and which is shown to be unbiased as $\beta \to 0$ and $r = 1$. This estimator will be proven to have an interesting properties as $r \to \infty$.

The maximum of $\pi(\lambda,x)$ is obtained at

$$\lambda^* = \frac{1}{x}. \tag{4.1}$$

Moreover, as r tends to infinity the mean of the density (3.18) which is Gamma (r+1, xr) density, is $\frac{r+1}{xr}$. The variance, also, which equals to $\frac{(r+1)}{(xr)^2}$ will be tending to zero as r tends to infinity.

Thus, the kernel $\pi(\lambda,x)$ tends to a Dirac-delta function spiking at the point $(\frac{1}{x})$.

Thus, for the full kernel given by equation (3.1) as

$$\hat{\pi}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \hat{\pi}(\lambda, x_i) \tag{4.2}$$

tends to a series of spikes at the points $\frac{1}{x_i}$ (i=1,...,n).

It is worth investigating how the kernel behaves for a general density f(x) as r tends to infinity.

We give now the following Lemma, associated with its proof.

Lemma 4.1

For the density $\hat{\pi}(\lambda,x)$ represented by

$$\hat{\pi}(\lambda,x) = \frac{(xr)^{r+1}\, \lambda^r\, e^{-xr\lambda}}{\Gamma(r+1)} \quad , \tag{4.3}$$

the integral

$$\int_0^\infty \frac{1}{x^{k+1}}\,\hat{\pi}(\lambda,x)dx \tag{4.4}$$

tends to λ^{k-1} as r tends to infinity.

Proof

Starting by the integral (4.4) , we have that

$$\int_0^\infty \frac{1}{x^{k+1}}\,\hat{\pi}(\lambda, x)dx = \int_0^\infty \frac{(xr)^{r+1}\,\lambda^r\, e^{-xr\lambda}}{x^{k+1}\,\Gamma(r+1)}\,dx$$

$$= \frac{(r)^{r+1}\,\lambda^r}{\Gamma(r+1)}\int_0^\infty x^{r-k}\, e^{-xr\lambda}\,dx \tag{4.5}$$

By making the substitution

$$xr\lambda = y \qquad , \quad dy = r\lambda dx$$

then relation (4.5) is written as

$$\int_0^\infty \frac{1}{x^{k+1}}\,\hat{\pi}(\lambda, x)dx = \frac{(r)^{r+1}\,\lambda^r}{\Gamma(r+1)} \int_0^\infty \left[\frac{y}{r\lambda}\right]^{r-k} e^{-y}\,\frac{dy}{r\lambda}$$

$$= \frac{(r)^{r+1}\,\lambda^r}{(r\lambda)^{r-k+1}\,\Gamma(r+1)} \int_0^\infty e^{-y}\, y^{r-k}\,dy$$

$$= \frac{(r)^{r+1}\,\lambda^r}{(r\lambda)^{r-k+1}} \cdot \frac{\Gamma(r-k+1)}{\Gamma(r+1)} \tag{4.6}$$

From Stirling's approximation formula

$$\Gamma(n+1) = n! = (2\pi n)^{\frac{1}{2}} (n)^n e^{-n},$$

we get from (4.6) that

$$\int_0^\infty \frac{1}{x^{k+1}} \hat{\pi}(\lambda, x)dx \simeq \lambda^{k-1}(r)^k \left(\frac{2\pi(r-k)}{2\pi r} \right)^{\frac{1}{2}} \cdot \frac{(r-k)^{r-k}}{(r)^r} \cdot \frac{e^{-(r-k)}}{e^{-r}}. \qquad (4.7)$$

From equation (4.7), we have that $\left(\frac{r-k}{r} \right)^{\frac{1}{2}}$ tends to one as r tends to infinity. Also, we have from (4.7) that

$$\frac{(r)^k (r-k)^{r-k}}{(r)^r} = \left(\frac{r-k}{r} \right)^{r-k}$$

$$= \left(\frac{r}{r-k} \right)^k \left(\frac{r-k}{r} \right)^r$$

$$= \left(\frac{1}{1-\frac{k}{r}} \right)^k \left(1 - \frac{k}{r} \right)^r. \qquad (4.8)$$

By substituting from (4.8) into (4.7) and taking the limit of (4.7) as r tends to infinity, we find that the whole expression tends to λ^{k-1}, completing the proof. □

We will be using the result of the lemma to evaluate the bias in our estimator of $\pi(\lambda)$.

4.4.2 The Tail Limiting Behaviour

We will start by the following example, and then state two concluding remarks about this limiting behaviour. Let us define a density on the interval [0,1] as

$$\begin{aligned} \pi(\lambda) &= 2(1-\lambda) && \lambda \in [0,1] \\ &= 0 && \text{otherwise.} \end{aligned} \qquad (4.9)$$

Thus f(x) takes the form

$$f(x) = \int_0^1 \lambda e^{-\lambda x}\, \pi(\lambda) d\lambda$$

$$= 2 \int_0^1 \lambda(1-\lambda) e^{-\lambda x} d\lambda \qquad (4.10)$$

By putting $u = \lambda$ and $dv = e^{-\lambda x} d\lambda$ in relation (4.10) and integrating it by parts, we get that

$$f(x) = \frac{2}{x^2}\left(e^{-x} + 1\right) + \frac{4}{x^3}\left(e^{-x} - 1\right). \qquad (4.11)$$

Taking the limit of relation (4.11) as x tends to infinity we have the following relation

$$f(x) \sim \frac{2}{x^2} - \frac{4}{x^3}. \qquad (4.12)$$

Notice that relation (4.12) is derived in the sense of the following definition:

Definition 4.4.2.1 If a_n and b_n are two sequences, and if

$$\lim_{n \to \infty} \frac{a_n}{b_n} = 1,$$

then a_n is asymptotically equivalent to b_n. This relation will be denoted by $a_n \sim b_n$.

It is common that for large x, the density f(x) will have an expansion of the form

$$f(x) \sim \frac{a_{k+1}}{(x)^{k+1}} + \frac{a_{k+2}}{(x)^{k+2}}. \qquad (4.13)$$

Now, assuming that f(x) can be expressed by the expansion (4.13), then the above

lemma 4.1 allows us to express the expectation of $\hat{\pi}(\lambda,x)$ as

$$E_X\left\{\hat{\pi}(\lambda,x)\right\} = \int_0^\infty \lambda e^{-\lambda x} f(x)dx$$

$$\sim a_{k+1}\,\lambda^{k-1} + a_{k+2}\,\lambda^k + .. \qquad (4.14)$$

In the above example if $k = 1$, the expectation (4.14) can be approximated by

$$E\{\hat{\pi}(\lambda,x)\} \sim 2(1 - 2\lambda). \qquad (4.15)$$

Notice that this is no longer a density.

This example shows two things: First, the example demonstrates that the bias can be removed by replacing (2λ) by λ, which represents a rescaling of the λ axis. Second, the example shows us a way of extending the applicability of the lemma to any density f(x), which takes the form

$$f(x) \sim \frac{a_2}{x^2} + \frac{a_3}{x^3} + \ldots, \qquad (4.16)$$

as x tends to infinity. That is, the bias up to linear terms in λ, in the estimation of $\pi(\lambda)$, can be eliminated by the rescaling.

An important special case results if we take $r = 1$, so that the mixing density is exponential, and the density f(x) takes the form

$$f(x) = \frac{\beta}{(x+\beta)^2}$$

$$= \frac{1}{x^2\beta}\left(\frac{x\beta}{x+\beta}\right)^2.$$

By applying the lemma, we get the following relation for the density

$$f(x) \sim \frac{1}{x^2\beta}\left(1 - \frac{x}{\beta} + \frac{x^2}{\beta^2} - \ldots\right). \qquad (\beta > 0)\,. \tag{4.17}$$

4.4.3 The Limiting Behaviour of the Mode

We can investiagte the mode of the estimator $\hat{\pi}(\lambda)$ in the limit as r rends to infinity. In the case of a single point mass for $\pi(\lambda)$ at λ_0 (i.e., $\pi(\lambda)$ is a Dirac-delta function at the point λ_0), we have

$$f(x) = \lambda_0\, e^{-\lambda_0 x}. \tag{4.18}$$

Thus, the expectation of $\hat{\pi}(\lambda,x)$ could be written as

$$E_x[\hat{\pi}(\lambda,x)] = \int_0^\infty \lambda_0\, e^{-\lambda_0 x}\, \hat{\pi}(x,\lambda)dx\,. \tag{4.19}$$

By taking $\hat{\pi}(\lambda,x)$ as proposed by equation (3.18), we have that

$$E_x[\hat{\pi}(\lambda,x)] = \frac{\Gamma(r+1)\lambda_0(r)^{r+1}\lambda^{r-1}}{(r\lambda+\lambda_0)^{r+1}\,\Gamma(r+1)}$$

$$= \lambda_0\left(1 + \frac{\lambda_0}{r\lambda}\right)^{-(r+1)}. \tag{4.20}$$

Taking the limit as r tends to infinity of the right-hand side of equation (4.20), we have that

$$E_x[\hat{\pi}(\lambda,x)] \sim \frac{\lambda_0}{\lambda^2}\, e^{-\frac{\lambda_0}{\lambda}}\,. \tag{4.21}$$

The above approximate result, represented by relation 4.21 achieves a maximum at

$$\lambda^* = \frac{\lambda_0}{2} \ . \tag{4.22}$$

Thus, we see that the mode of the expected density (for one observation) is achieved at $\frac{1}{2} \times$ the correct value.

The final conclusion is that, for obtaining an unbiased estimator of the density $\pi(\lambda,x)$, a scale factor of two is appropriate in the limit. This holds for the case of mode as with the linear tail.

4.4.4 A General Remark

We have introduced the definition of the full kernel-type estimator as

$$\hat{\pi}(\lambda) = \sum_{i=1}^{n} \hat{\pi}(\lambda; x_i), \tag{4.23}$$

which is a linear function of the estimator $\hat{\pi}(\lambda, x_i)$. Thus, all the results, being stated above for the density $\hat{\pi}(\lambda, x_i)$, hold in moving to the kernel-type estimator, represented by equation (4.23).

The kernel density estimate $\hat{f}_n$ is defined as

$$\hat{f}_n(x) = \frac{1}{nh_n} \sum_{j=1}^{n} K\left(\frac{x-x_j}{h_n}\right), \tag{4.24}$$

where $K(\cdot)$ is the kernel function which integrates to one. This estimate depends linearly on the kernel function $K(\cdot)$, hence the appearence and properties of $\hat{f}_n$ will be dictated by the choice of the kernel $K(\cdot)$.

This general remark establishes a connection between our kernel-type estimator and the

usual kernel-density estimator (4.24). That is, discussing the features of the kernel estimator "$\hat{f}_n$" through studying the features of its kernel function $K(\cdot)$ is similar to studying the properties of our estimator $\hat{\pi}(\lambda)$ via results about the estimator $\hat{\pi}(\lambda, x)$.

4.5 A MOMENT PROPERTY

4.5.1 A Motivation for a Study of Moments

One of the mathematical conveniencs of using mixtures is that the moments are inherited in the following way. If, we define

$$\mu_k(\lambda) = \int_{-\infty}^{\infty} x^k f(x/\lambda)dx,$$

then

$$\mu_k = \int_{-\infty}^{\infty} x^k f(x)dx,$$

$$= \int_{-\infty}^{\infty} x^k \Big\{ \int_{0}^{\infty} f(x/\lambda)\pi(\lambda)d\lambda \Big\} dx. \qquad (5.1)$$

Thus, the k^{th} moment, can be written, from (5.1), as

$$\mu_k = \int_{0}^{\infty} \Big\{ \int_{-\infty}^{\infty} x^k f(x/\lambda)dx \Big\} \pi(\lambda)d\lambda$$

$$= \int_{0}^{\infty} \mu_k(\lambda)\pi(\lambda)d\lambda \qquad (5.2)$$

Having mentioned the limiting behaviour of the estimators $\hat{\pi}(\lambda)$, we shall continue to investigate the moments of these estimators (or densities). The aim of such study is to judge how far the moments of the estimator $\hat{\pi}(\lambda)$ mimic the moments of the underlying $\pi(\lambda)$. Thus, if we define

$$\tau_k = \int_{-\infty}^{\infty} \lambda^k \pi(\lambda)d\lambda, \qquad (5.3)$$

then for the theoretical kernel of section (3.2), we have that

$$\hat{\tau}_k = \int_{0}^{\infty} \lambda^k \hat{\pi}(\lambda,x)d\lambda$$

$$= \int_0^\infty \lambda^k \, \hat{\pi}(\lambda/x) d\lambda. \tag{5.4}$$

Then, the expectation of the formula (5.4) is

$$E_x\{\hat{\tau}_k\} = \int_{-\infty}^{\infty} \int_0^\infty \lambda^k \, f(x/\lambda) \, \pi(\lambda) d\lambda \, dx$$

$$= \int_0^\infty \lambda^k \left\{ \int_{-\infty}^{\infty} f(x/\lambda) \, dx \right\} \pi(\lambda) d\lambda \tag{5.5}$$

knowing that the integral, given between brackets, in equation (5.5) is equal to one. Thus, we have for the expectation of $\hat{\tau}_k$, that

$$E_x\{\hat{\tau}_k\} = \int_0^\infty \lambda^k \, \pi(\lambda) d\lambda = \tau_k, \tag{5.6}$$

in which the right-hand side of (5.6) is a result of the definition in equation (5.3).

In conclusion, equations (5.3) and (5.6) tell us that the same properties are inherited, again, by the estimator $\hat{\pi}(\lambda)$.

4.5.2 An Inverse-Mean Type of Bias

The mean of the density represented by equation (3.18) is

$$\hat{\tau}_1 = \frac{(r+1)}{xr}. \tag{5.7}$$

Thus

$$E_X(\hat{\tau}_1^{-1}) = \frac{r}{(r+1)} E(X)$$

$$= \frac{r}{(r+1)} \frac{\beta}{r-1}, \; r > 1, \tag{5.8}$$

where $f(x) = (r\beta^r)/(x+\beta)^{r+1}$. So that

$$E_X(\hat{\tau}_1^{-1}) = \frac{r}{(r+1)} \cdot \frac{\beta}{r-1} = \frac{r^2}{r^2-1} \tau_k^{-1}, \tag{5.9}$$

where $\tau_k = \frac{\beta}{r}$ is the mean of the original $\pi(\lambda)$ distribution given by equation (3.9). Notice that when $r = 1$, then $E_X(\hat{\tau}_1^{-1})$ which is given by (5.9) will be infinite, providing us with a further justification for using a value of "r" greater than one.

Now, from equations (5.6) and (5.7) the mean of the estimator $\hat{\pi}(\lambda)$ will be

$$\hat{\tau}_1 = \frac{r+1}{r} \cdot \frac{1}{n} \sum_{i=1}^{n} \frac{1}{x_i} \geq \frac{r+1}{r} \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right)^{-1}, \tag{5.10}$$

thus, we have

$$E\left(\hat{\tau}_1^{-1}\right) \leq \frac{r}{r+1} \; E\left(\frac{1}{n} \sum_{i=1}^{n} x_i \right) = \frac{r}{r+1} \; E(X)$$

$$= \frac{r}{r+1} \frac{\beta}{r-1} = \frac{r^2}{r^2-1} \tau_1^{-1} \; . \tag{5.11}$$

By considering equation (5.11), and letting r tend to infinity we have, in a sense, an "inverse-mean biasedness".

The conclusion, being mentioned by the last statement, confirms the scaling downwards which was clear for the cases of the mode and the tail of the estimator $\hat{\pi}(\lambda)$.

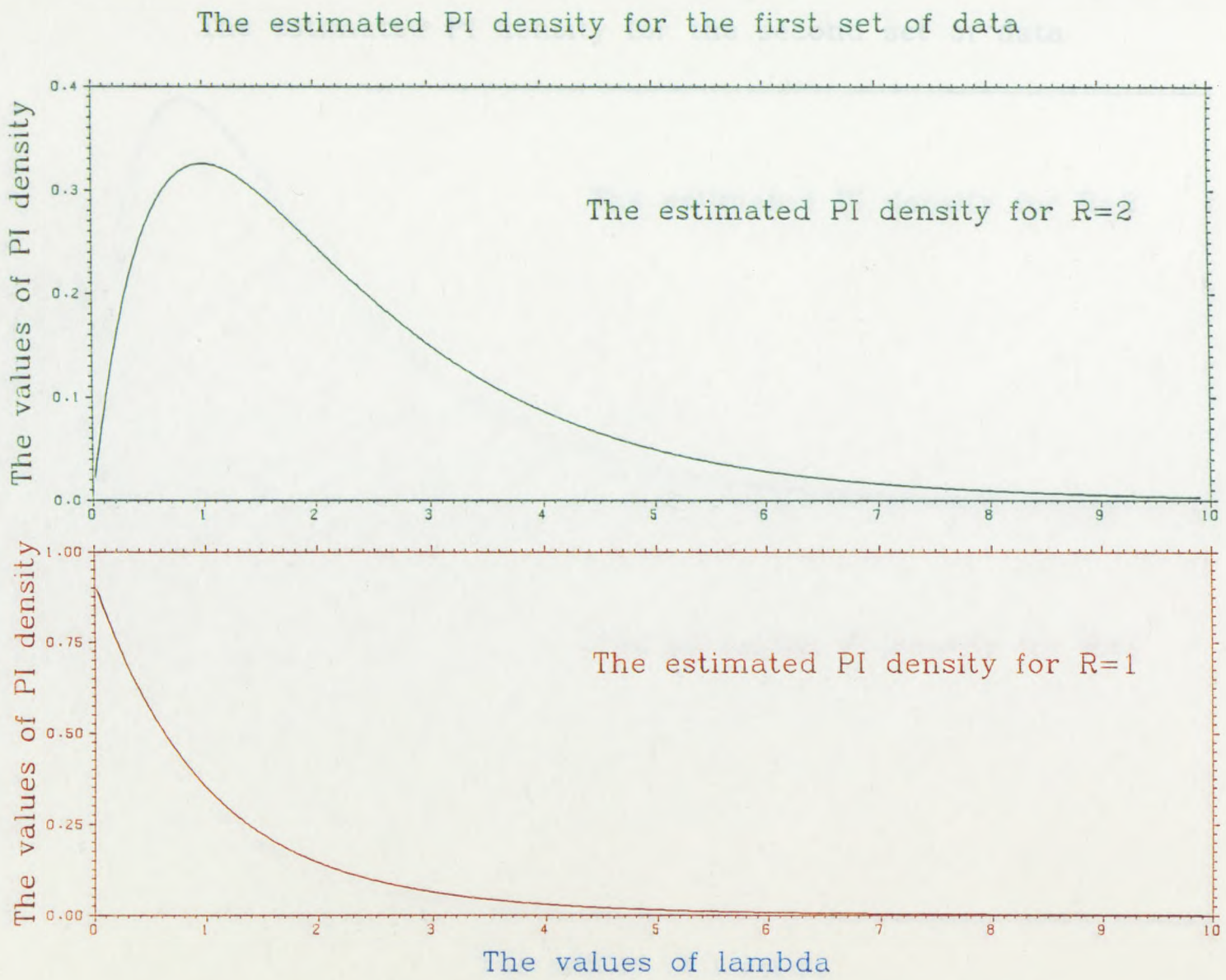
The estimated PI density for the first set of data
The values of PI density
The estimated PI density for R=2
The values of PI density
The estimated PI density for R=1
The values of lambda

Fig. 4.1

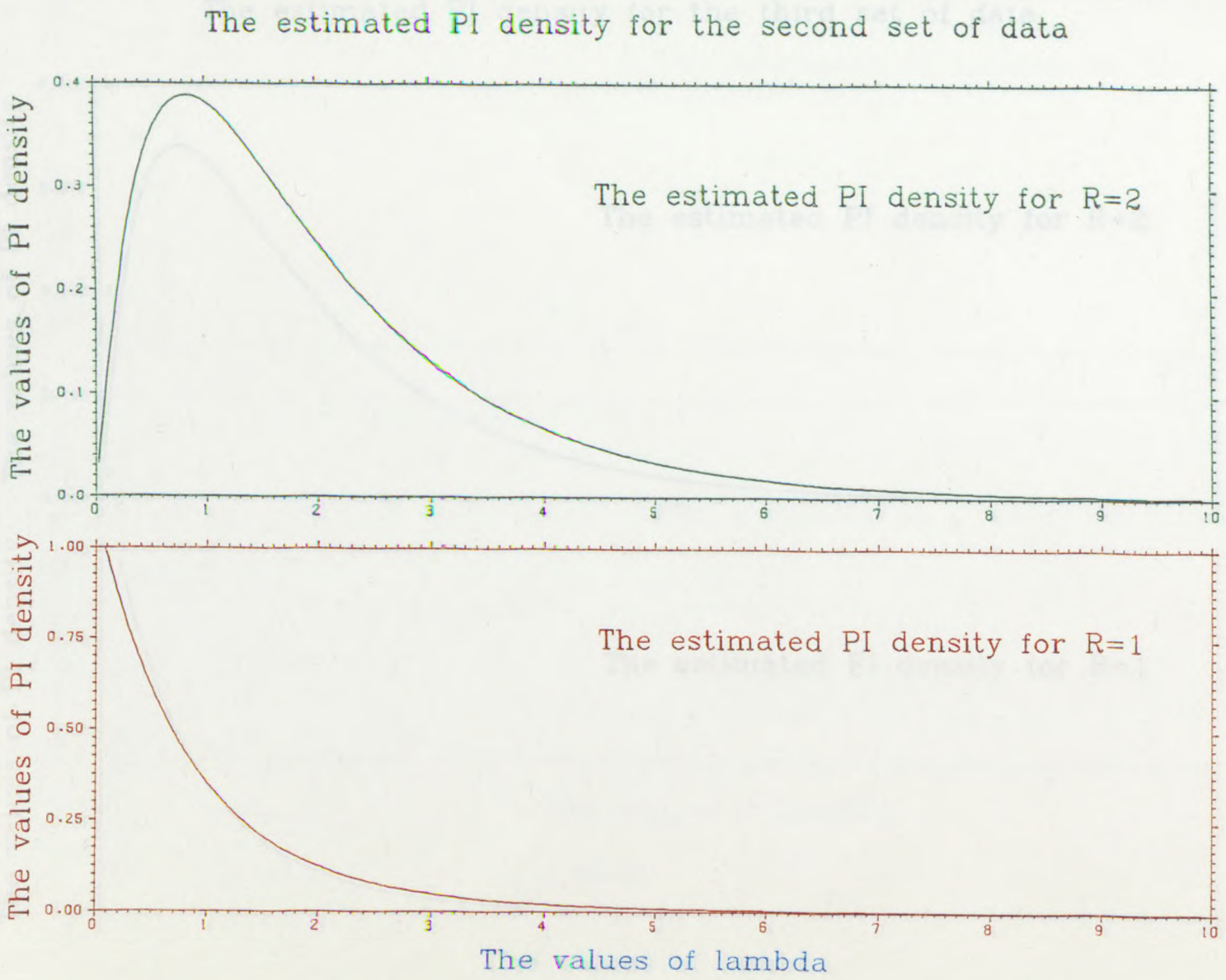
The estimated PI density for the second set of data
The values of PI density
The estimated PI density for R=2
0.4
0.3
0.2
0.1
0.0
0 1 2 3 4 5 6 7 8 9 10
The values of PI density
The estimated PI density for R=1
1.00
0.75
0.50
0.25
0.00
0 1 2 3 4 5 6 7 8 9 10
The values of lambda

Fig 4.2

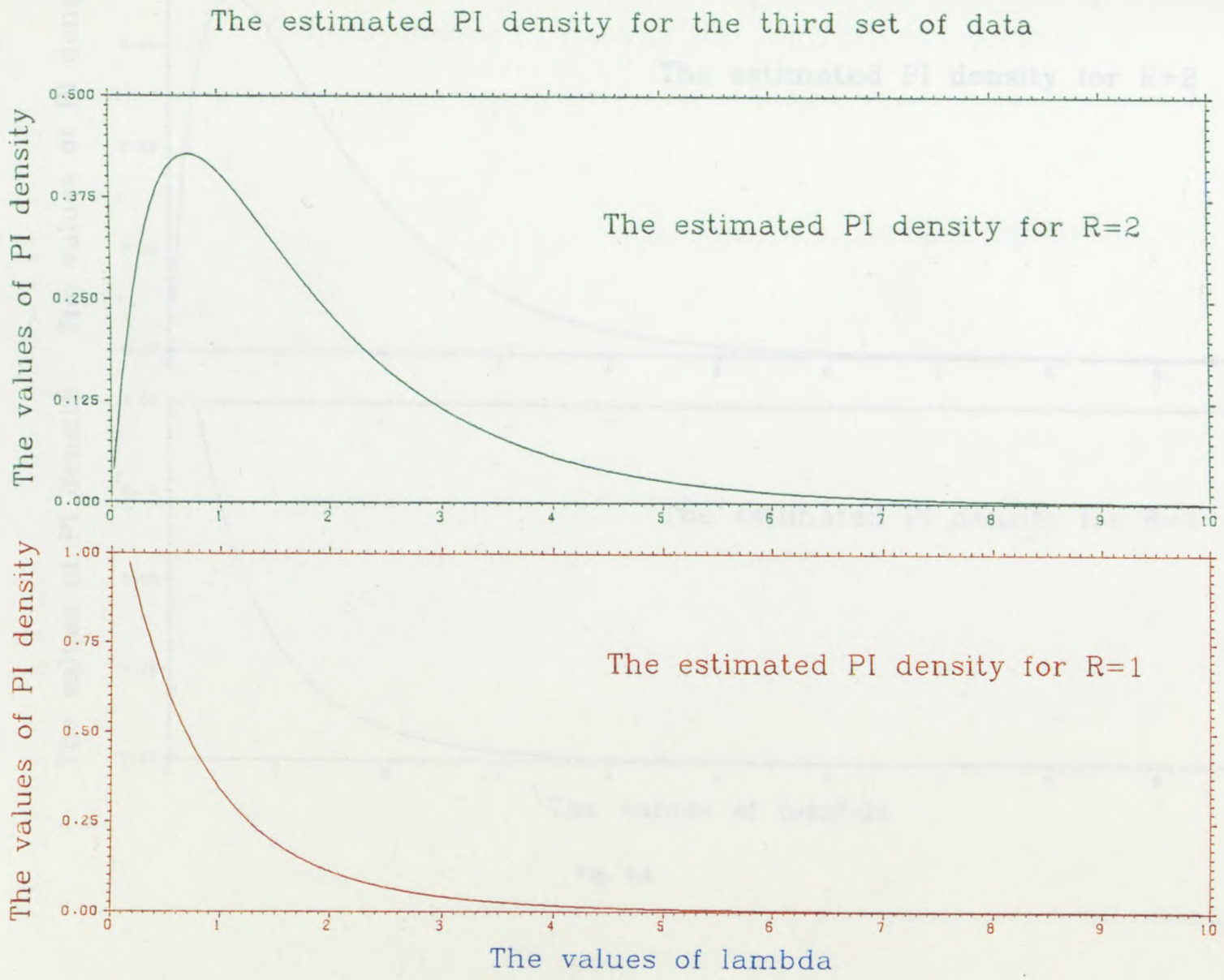
The estimated PI density for the third set of data
The values of PI density
0.500
0.375
0.250
0.125
0.000
The estimated PI density for R=2
0
1
2
3
4
5
6
7
8
9
10
The values of PI density
1.00
0.75
0.50
0.25
0.00
The estimated PI density for R=1
The values of lambda

Fig. 4.3

The estimated PI density for the fourth set of data

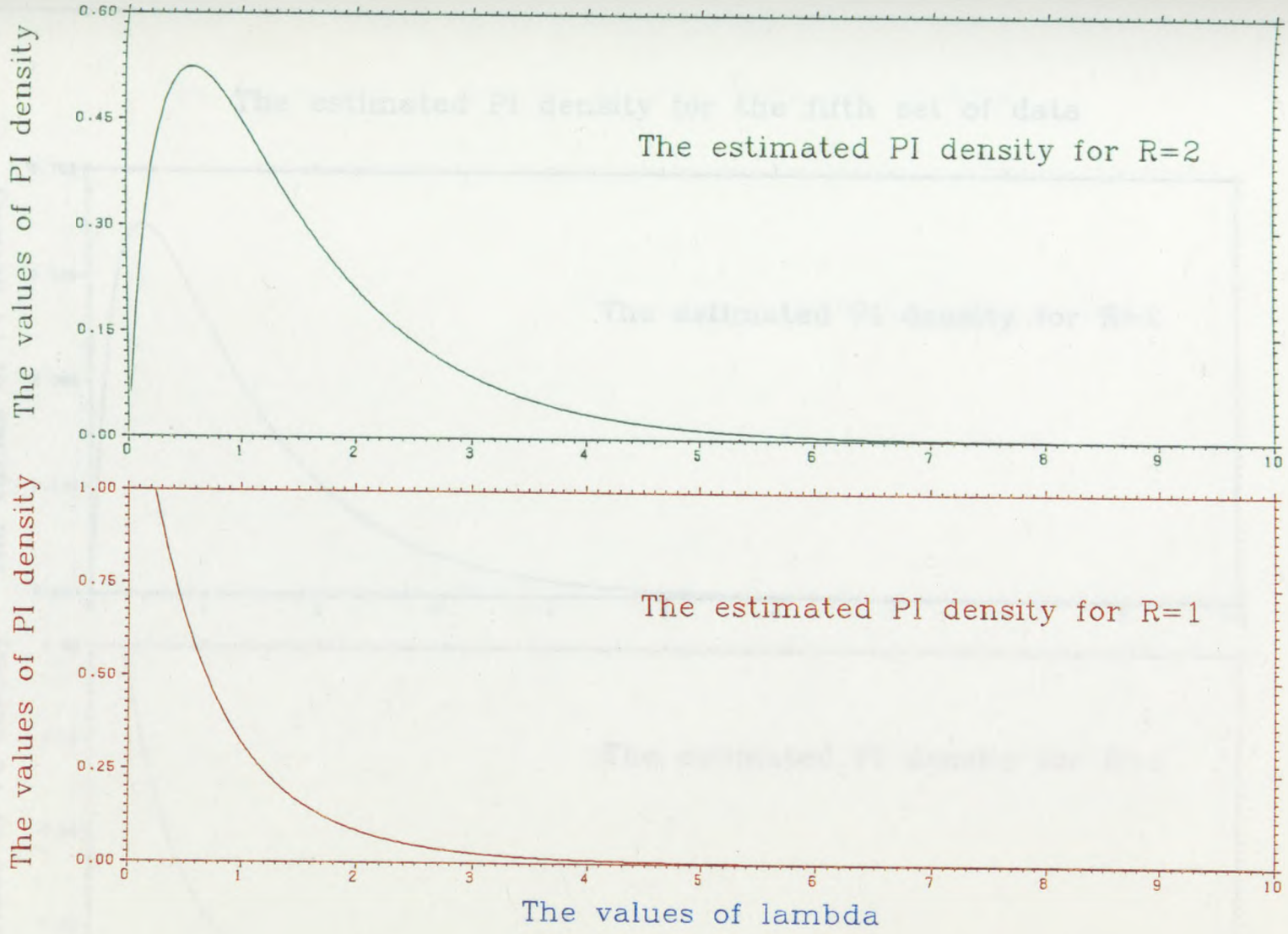

Fig. 4.4

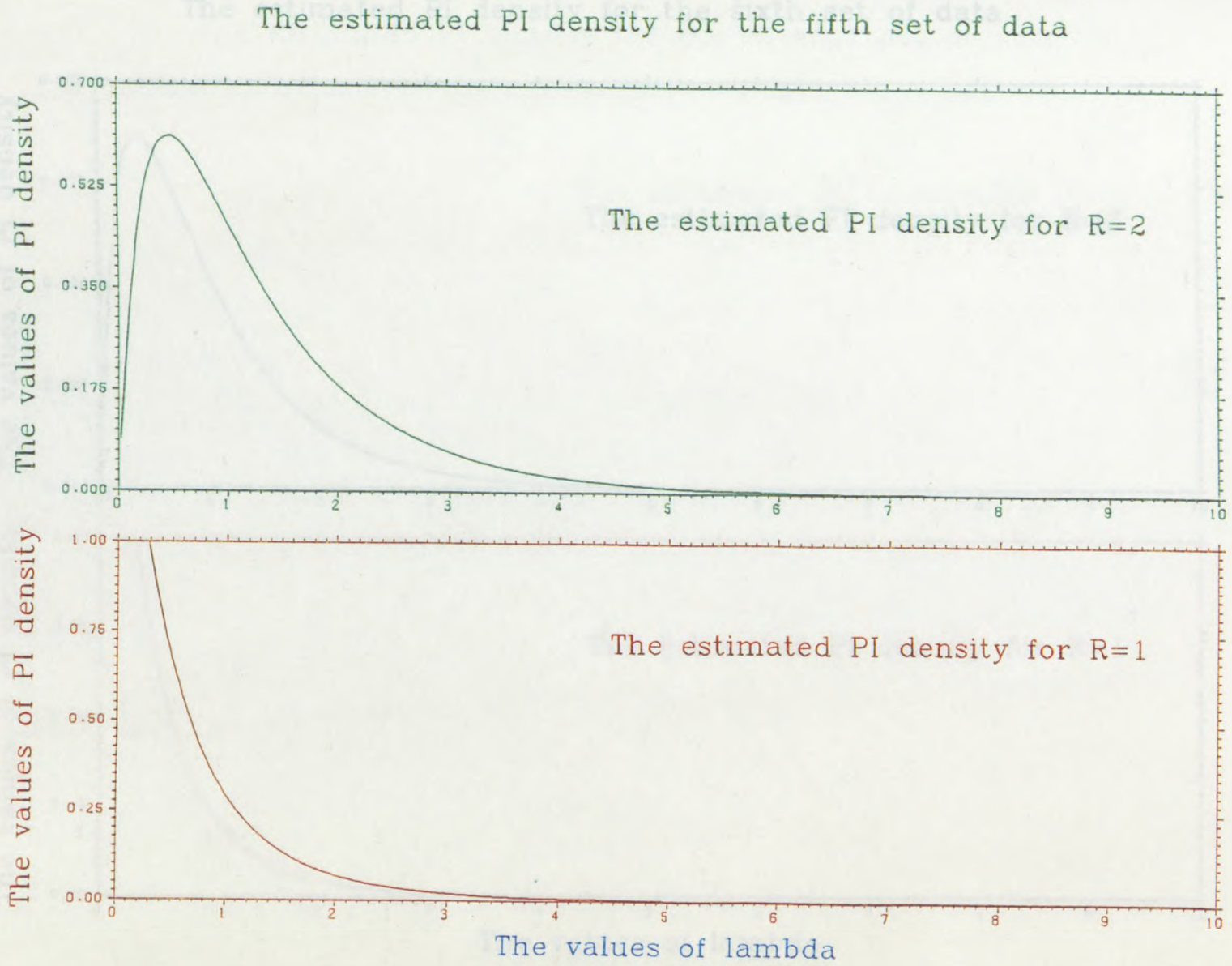
The estimated PI density for the fifth set of data
The values of PI density
0.700
0.525
0.350
0.175
0.000
The estimated PI density for R=2
0 1 2 3 4 5 6 7 8 9 10
The values of PI density
1.00
0.75
0.50
0.25
0.00
The estimated PI density for R=1
0 1 2 3 4 5 6 7 8 9 10
The values of lambda

Fig. 4.5

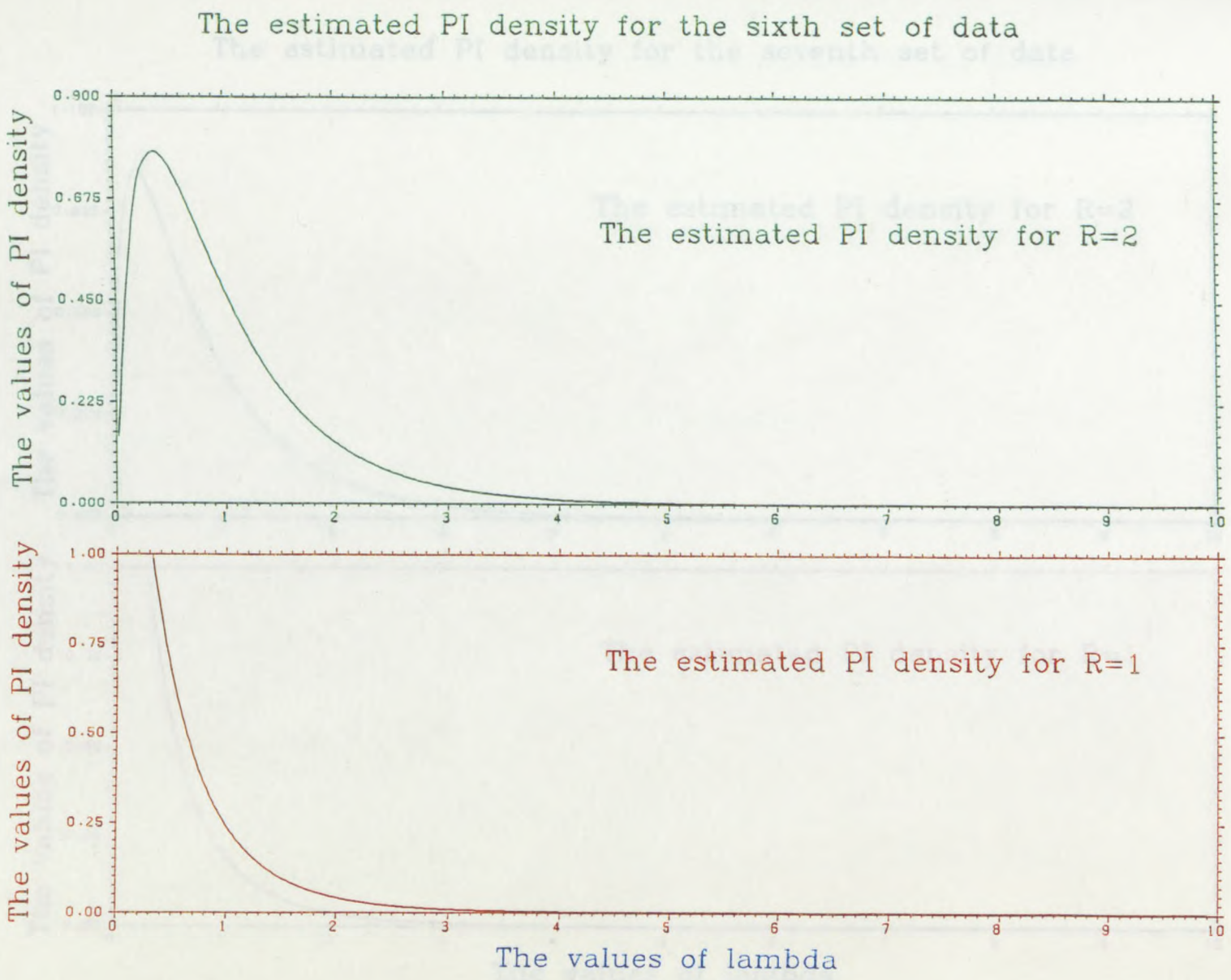
The estimated PI density for the sixth set of data
The values of PI density
0.900
0.675
0.450
0.225
0.000
0 1 2 3 4 5 6 7 8 9 10
The estimated PI density for R=2
The values of PI density
1.00
0.75
0.50
0.25
0.00
0 1 2 3 4 5 6 7 8 9 10
The estimated PI density for R=1
The values of lambda

Fig. 4.6

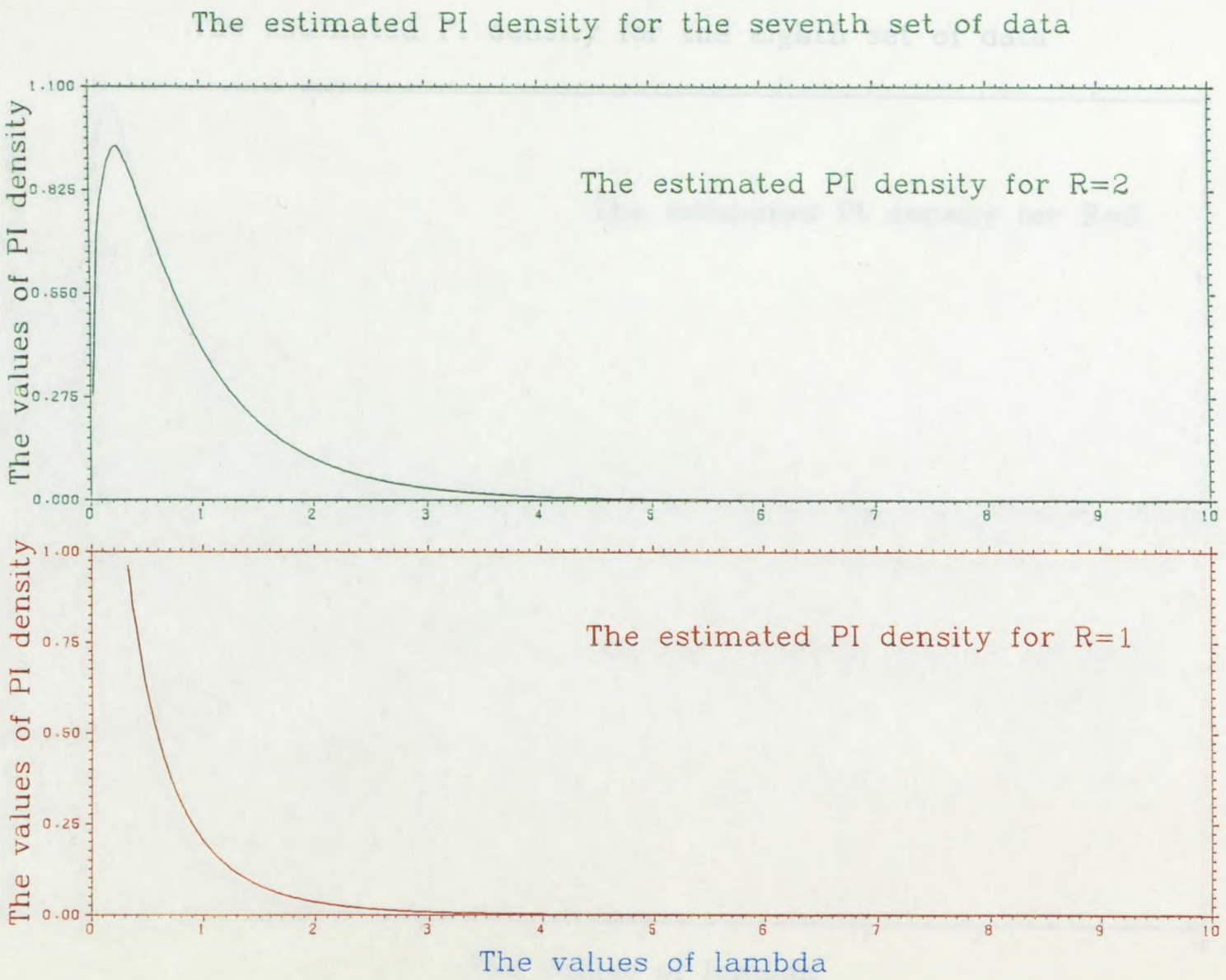
The estimated PI density for the seventh set of data
The estimated PI density for R=2
The estimated PI density for R=1
The values of PI density
The values of PI density
The values of lambda
1.100
0.825
0.550
0.275
0.000
1.00
0.75
0.50
0.25
0.00
0
1
2
3
4
5
6
7
8
9
10

Fig. 4.7

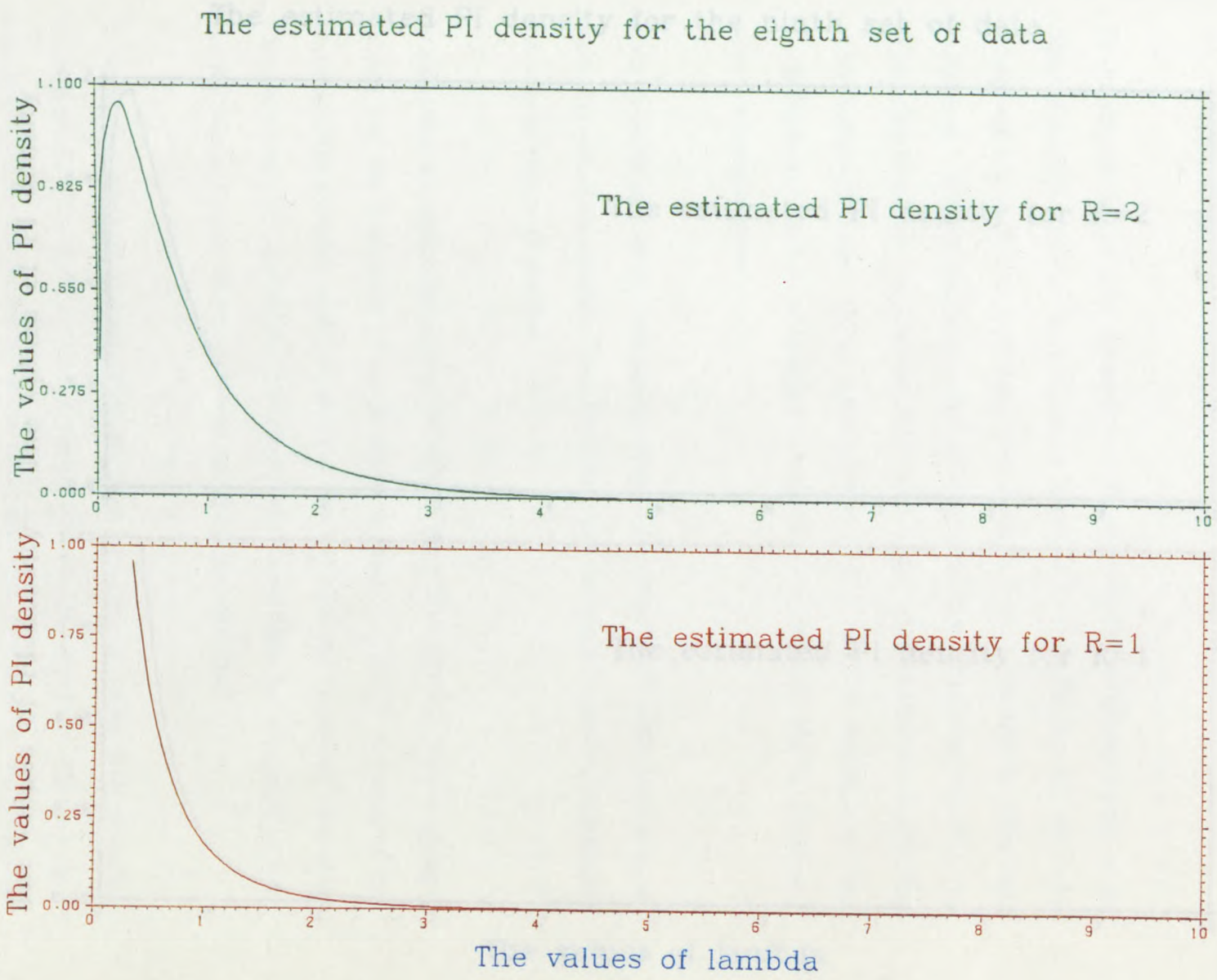
The estimated PI density for the eighth set of data
The values of PI density
1.100
0.825
0.550
0.275
0.000
0 1 2 3 4 5 6 7 8 9 10
The estimated PI density for R=2
The values of PI density
1.00
0.75
0.50
0.25
0.00
0 1 2 3 4 5 6 7 8 9 10
The estimated PI density for R=1
The values of lambda

Fig. 4.8

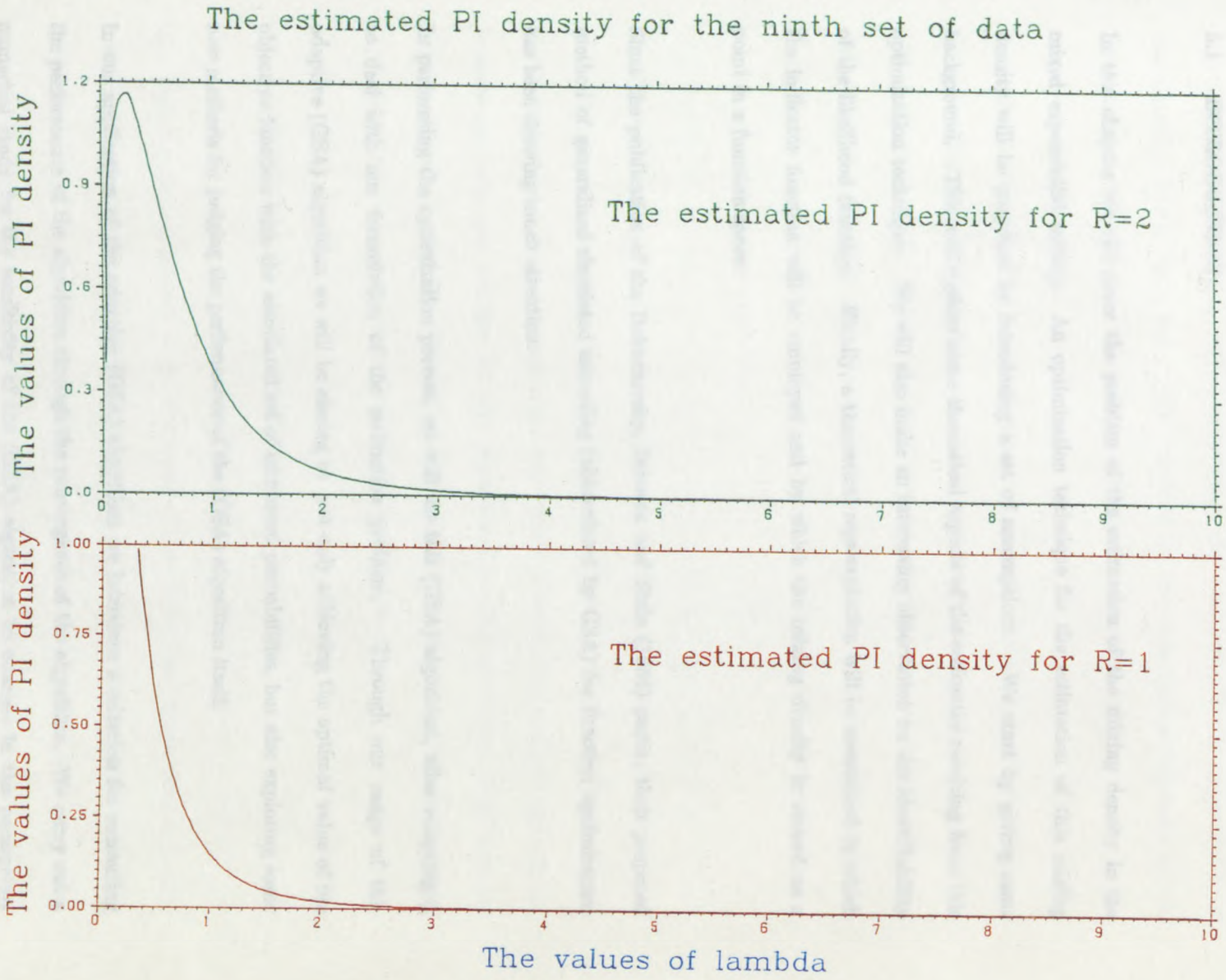
The estimated PI density for the ninth set of data
The values of PI density
1.2
0.9
0.6
0.3
0.0
0
1
2
3
4
5
6
7
8
9
10
The estimated PI density for R=2
The values of PI density
1.00
0.75
0.50
0.25
0.00
0
1
2
3
4
5
6
7
8
9
10
The estimated PI density for R=1
The values of lambda

Fig. 4.9

CHAPTER FIVE

AN OPTIMIZATION TECHNIQUE OF ESTIMATION

5.1 INTRODUCTION

In this chapter we will cover the problem of the estimation of the mixing density in the mixed exponential setting. An optimization technique for the estimation of this mixing density will be proposed by introducing a set of assumptions. We start by giving some background. This will explore some theoretical aspects of the estimator resulting from this optimization technique. We will also make an interesting observation on the identifiability of the likelihood function. Finally, a theoretical representation will be mentioned in which the indicator function will be employed and by which the mixing density is viewed as a point in a function space.

Since the publication of the Bohachevsky, Johnson and Stein (1986) paper, their proposed method of generalized simulated annealing (abbreviated by GSA) for function optimization has been drawing much attention.

In performing the optimization process, we will use this (GSA) algorithm, after adapting it to deal with our formulation of the estimation problem. Through our usage of this adaptive (GSA) algorithm we will be aiming at not only achieving the optimal value of the objective function with the associated set of estimated probabilities, but also exploring some new methods for judging the performance of the (GSA) algorithm itself.

In our application of the adaptive (GSA) algorithm we introduce a criterion for measuring the performance of the algorithm through the convergence of the algorithm. We carry out a numerical study for the sensitivity of the (GSA) algorithm to changes in the parameters

affecting our estimation problem. The numerical examples will involve exploring possible ways of improving the algorithm performance measured by the success in achieving the desired characteristics of the algorithm namely convergence and optimality.

The numerical study is of importance because it may be helpful in: (i) discovering some new features of the adaptive (GSA) algorithm, (ii) enhancing better understanding of the algorithm performance, and (iii) making us more able to conclude some general characteristics of the problems to which the algorithm is applicable.

A justification for adopting the (GSA) algorithm in performing our optimization technique for estimating the mixing probabilities will be based upon two important features of the (GSA) algorithm itself. Firstly, the (GSA) algorithm is mainly useful for finding a global extremum of a function that has many local extrema such as the present problem. Secondly, the algorithm is useful for functions which may not be smooth, because it does not require any calculations of derivatives.

Finally, this chapter ends with a comparative study between our optimization technique and two other density estimation methods, namely, the kernel method and the maximum likelihood method. In this context some connections between these methods (specifically, between our technique and the two above-mentioned methods of density estimation) are drawn. Finally, a suggested rationale has been given by which our optimization technique is related to other nonparametric methods of density estimation.

5.2 DEFINITIONS AND NOTATIONS

This section puts the problem of the estimation of the mixing distribution from a random sample of observations in an abstract representation, and refers to the issue of identifiability as one of the main requirements for the estimation problem to be a meaningful one. Also, an interesting observation is made concerning the uniqueness of the solution of our problem.

Let $(x,\mathcal{F})$ and (Θ,β) be two measurable spaces such that β contains all Borel subsets of Θ. Let $\Phi = \{P_\theta, \theta \in \Theta\}$ be a family of probability measures on $(x,\mathcal{F})$ such that the mapping $\theta \to P_\theta(A)$ is $\beta-$measurable for each $A \in \mathcal{F}$. Suppose that if $\theta_1 \neq \theta_2$, then $P_{\theta_1} \neq P_{\theta_2}$.

Definition 5.2.1: If G and H are two probability measures defined on the probability spaces (Θ,β) and $(x,\mathcal{F})$ respectively, such that

$$H(A) = \int_\Theta P_\theta(A)\, dG(\theta) \quad , \qquad A \in \mathcal{F} \tag{2.1}$$

then the probability measure H is a mixture of $\Phi = \{P_\theta, \theta \in \Theta\}$, and the probability measure G is mixing distribution.

The problem of estimating the mixing distribution, which has been expressed by equation (2.1), can be considered meaningful only if there is a one-to-one correspondence between the mixing distribution G and the resulting mixture H.

Definition 5.2.2: Identifiability: Let Λ be the class of all mixing distributions on (Θ,β) and ξ be the corresponding class of mixtures. Let Q be a mapping $Q : \Lambda \to \xi$ defined by

$$Q(G) \equiv H. \tag{2.2}$$

Here the class Λ of all mixing distributions is said to be "identifiable" if Q is a one-to-one mapping.

5.3 NONPARAMETRIC TECHNIQUE OF ESTIMATION

5.3.1 The basic idea

The optimization technique here for a nonparametric estimation of the mxing density, is based upon generalizing some concepts which have been used in the histogram and kernel density estimation methods. In order to demonstrate this idea we start by the following definition.

Definition 5.3.1: Parzen Kernel Estimator: Given a random sample $x_1,...x_n$ from a continuous but unknown density f. Parzen (1962) defines the kernel density estimator as

$$\hat{f}_n(x) = \int_{-\infty}^{\infty} \frac{1}{h_n} K\left(\frac{x-y}{h_n}\right) dF_n(y) \qquad (3.1)$$

$$= \frac{1}{nh_n} \sum_{j=1}^{n} K\left(\frac{x-x_j}{h_n}\right).$$

Denoting $K_h(y) = \frac{1}{h} K\left(\frac{y}{h}\right)$ as the scaled kernal we could rewrite equation (3.1) for a given random sample $x_1,...x_n$ as

$$\hat{f}(x) = \frac{1}{n} \sum_{j=1}^{n} K_h(x-x_j) \qquad (3.2)$$

This estimate has equal weights of $\frac{1}{n}$ on each of the n kernels centered at the data points $x_j (j=1,...,n)$.

In fact, the kernel estimator- shown by equation (3.2) – is simply a histogram where every point (in case of estimating the density at that point) is the center of a smapling interval, so by using definition 5.3.1 we have the advantage of freeing the histogram from a particular choice of bin position. This argument could be considered as a justification for some of our main assumptions, specifically those which have been based upon a kernel structure.

The optimization technique, we propose here, is based upon exploiting the connection

between the kernel and the histogram methods of density estimation and upon making a certain generalization of the idea of equally weightening the n kernels being presented by the above-mentioned equation (3.2).

5. 3.2: The Main Assumptions:

The optimization technique for estimating the mixing probabilities is based upon the following set of assumptions:

Assumption 1

Having said that the kernel estimator (3.2) has equal weights of $\frac{1}{n}$ on each of the n kernels centered at the data points, we make a more general assumption of unequal weights. We denote these weights by π_j , j=1,...,n.

Assumption 2:

The kernel function $K_h(x_j)$, j=1,...,k, which appears in equation (3.2) is replaced by the histogram-like mapping I_j, j=1,...,k defined by

$$I_j = \int_{c_{j-1}}^{c_j} f(x;\lambda)d\lambda \quad , \forall \text{ integer } j \tag{3.3}$$

where $(c_j - c_{j-1})$ is the width of the j^{th} cell (c_{j-1}, c_j).

In fact I_j is a mapping $M : \mathcal{F} \rightarrow \mathbb{H}$, where $\mathcal{F}$ is the space of functions f and $\mathbb{H}$ is the space of reals (or the space of integrals I_j which have been shown by (3.3)). Thus, each I_j, j=1,...,k is in fact a functional.

Assumption 3:

Assume that we have k equally spaced cells (c_{j-1}, c_j), j = 1,...,k, and that we have given a sample $x_1,...,x_n$ of size n from f, then we can evaluate the integral, represented by equation (3.3), over each cell (see the appendix for such evaluation). Define the likelihood

function of these n observations as

$$L(\underset{\sim}{\pi};\underset{\sim}{x}) = \prod_{i=1}^{n}\left\{\sum_{j=1}^{k}\pi_j I_j(x_i)\right\} \tag{3.4}$$

The above reprsentation (3.4) is, in fact, a polynomial of the n^{th} degree in π's.

Assumption 4:

Estimate the unequal weights π's, by maximizing the likelihood function $L(\underset{\sim}{\pi};\underset{\sim}{x})$ given in the previous assumption by equation (3.4). The resulting set of estimated weights $\hat{\pi}_j$, $j=1,...,k$ will be representing the mixing probabilities in our case of the mixture of exponentials.

Assumption 5:

From the previous set of assumptions we present the estimation problem in an optimization representation as

$$\max L(\underset{\sim}{\pi};\underset{\sim}{x}) = \max_{\underset{\sim}{\pi}} \log \prod_{i=1}^{n}\left\{\sum_{j=1}^{k}\pi_j I_j(x_j)\right\}$$

$$= \max_{\underset{\sim}{\pi}}\left\{\sum_{i=1}^{n}\log\left[\sum_{j=1}^{k}\pi_j I_j(x_i)\right]\right\} \tag{3.5}$$

subject to the linear constraints

$$\sum_{j=1}^{k}\pi_j(c_j - c_{j-1}) = 1$$

$$\pi_j \geq 0 \quad \text{and} \quad c_0 = 1$$

$$j = 1,...,k \tag{3.6}$$

5.3.3 Some Theoretical Aspects of the Optimization Technique:

5.3.3.1 Introduction .

We discuss here some theoretical consequences of the above assumptions.

Let

$$I_j(x) = \int_{c_{j-1}}^{c_j} f(x/\lambda)d\lambda = \int_{c_{j-1}}^{c_j} \lambda e^{-\lambda x}\, d\lambda \qquad (3.7)$$

By evaluating the integrals I_j (j=1,...,k) by parts - see the appendix - we obtain

$$I_k(x) = -\frac{1}{x^2}\left[e^{-c_k x}(c_k x + 1) - e^{-c_{k-1} x}(c_{k-1} x + 1)\right], \qquad (3.8)$$

where $x > 0$.

Note that the above formula (3.8), after multiplying it by x^2, becomes an exponential polynomial.

By the third assumption, having assumed that we have k intervals associated with k heights π_j(j=1,...,k), the joint density function (for a particualr x) will be

$$f(x) = \sum_{j=1}^{k} \pi_j I_j(x) \qquad (3.9)$$

Take a sample of n independent observations on the random variable X with the density function (3.9). The likelihood function $L(\underset{\sim}{\pi}; \underset{\sim}{x})$ will be represented by equation (3.4).

Now, suppose that we have from equation (3.9),

$$f(x_i) = y_i = \sum_{j=1}^{k} \pi_j I_j(x_i) \qquad (3.10)$$

Suppose that there exists an estimator for $\pi_j(j=1,...,k)$ denoted by $\hat{\underset{\sim}{\pi}} = (\hat{\pi}_1,...,\hat{\pi}_k)$, which represents the solution of the optimization setting (3.5), we get the following relation from (3.10),

$$\hat{y}_i = \sum_{j=1}^{k} \hat{\pi}_j \, I_j(x_i) \tag{3.11}$$

We can interpret formula (3.11) as representing a mapping from the measure space, say, $\prod$ to R^n, in which the point $(\hat{y}_1,...,\hat{y}_n)$ in the image space determines the measure $\hat{\underset{\sim}{\pi}} = (\hat{\pi}_1,...,\hat{\pi}_k)$ in this measure space $\prod$, such that

$$\psi(\hat{\underset{\sim}{\pi}}) = (\hat{y}_1,...,\hat{y}_n), \tag{3.12}$$

where ψ is the logarithm of the likelihood (3.4).

5.3.3.2 A Characterization of the Estimator $\hat{\pi}$:

From the fifth assumption, specially the optimization formulation (3.5), we realize that the solution of it can be charaterized by two things:

(i) The width of the intervals of the original histogram $(c_j - c_{j-1})$, $(j=1,...,k)$, knowing that this number is assumed to be k.

(ii) the amount of probability (height $\hat{\pi}_j$) at each support interval of the histogram.

Concerning the width of the interval, we will be studying its effect, proposing linking this with the idea of controlling the smoothness of the density estimator by varying the smoothing parameter. This link will be demonstrated by a numerical example in the next subsections.

A partial uniqueness result can be established using a theorem of Polya and Szego (1925): a non-identically vanishing exponential polynomial taking the form

$$\sum_{j=1}^{k} p_j(y)\, e^{k_j y} \tag{3.13}$$

where the p_j is an ordinary polynomial of degree n_j, $(j=1,\ldots,k)$ has at most $\sum_{j=1}^{k} (n_j+1)-1$ zeros. We shall show that if $n \geq 2k$, then no two different vectors $(\hat{\pi}_1,\ldots,\hat{\pi}_k)$ can determine the same values

$$\hat{f}(x_i) = \hat{y}_i = \sum_{j=1}^{k} \hat{\pi}_j I_j(x_i), \qquad i = 1,\ldots,n. \tag{3.14}$$

The proof runs as follows:

Since

$$I_j(x) = -\frac{1}{x^2}\left[e^{-c_j x}(c_j x + 1) - e^{-c_{j-1} x}(c_{j-1} x + 1)\right], \tag{3.15}$$

we have that

$$x^2 \sum_{j=1}^{k} \pi_j I_j(x) \tag{3.16}$$

is a polynomial of the Polya-Szego form with each $n_j = 1$. Now, consider $x^2 \sum_{j=1}^{k} \hat{\pi}'_j I_j(x_i)$. If this yields the same $\hat{y}_i$ then

$$x^2 \sum_{j=1}^{k} (\hat{\pi}_j - \hat{\pi}'_j) I_j(x) = 0 \tag{3.17}$$

has zeros at $x_1,\ldots,x_n$. Thus it is either identically zero or $n \leq 2k-1$. In other words, if $n \geq 2k$ the $\hat{\pi}_j$ uniquely determine the $\hat{y}_j$.

This is essentially an identifiability result and should be the starting point for more detailed uniqueness questions.

5.3.4 An Algorithm for Optimization:

5.3.4.1 Introduction

This section covers generalized simulated annealing GSA for function optimization as introduced recently by Bohachevsky, Johnson and Stein (1986), and uses it for the optimization problem given in the fifth assumption.

We will attack the problem of the estimation of the mixing density, throughout our usage of the GSA method by:

(i) discussing the applicability of the algorithm to the mixture of exponentials case, and

(ii) investigating the possible adaptations required for such application.

The stages of our application are

First: We start by generating a set of data from a mixture consists of, for example, two exponential distributions, say, F_1 and F_2.

Second: Set up initial values by solving the optimization problem being represented by (3.5) and (3.6). The result of this step is the following set of initial values:

(1) Consider the optimal set of the estimated probabilities (or heights $\hat{\underset{\sim}{\pi}}^*$) among the total estimated sets is one thousand sets). Take the optimal value of the objective function L^* at this optimal set $\hat{\underset{\sim}{\pi}}^*$. Denote this optimal value L^* by ϕ_{optimal} or just ϕ_m.

(2) Exclude this optimal case, and consider another (any other) set of estimated

probabilities $\hat{\underline{\pi}}$. This set of initial probabilities (or heights) will be referred to as $\hat{\underline{\pi}}_0$.

(3) Take the corresponding value of the objective function $L(\hat{\underline{\pi}}_0)$ and denote this initial value of L at $\underline{\pi}_0$ by ϕ_0.

We give the following definition:

Definition 5.3.4.1 Assumethat the initial set of heights $\hat{\underline{\pi}}$ lie in the set $\Omega \in R^n$ which is assumed to be the unit interval [0,1]. This set will be referred to as the domain of definition of the estimated probabilities.

5.3.4.2 Steps of the GSA Algorithms:

In the framework of the previous "estimated " sets of initial values, being determined in the previous subsection, our problem of estimation has been "adapted" to the application of the (GSA) algorithm. The steps of (GSA) in our case will be:

The First Step:

Start by setting a level of accuracy, denoted by "η", which is chosen arbitrary as small as possible. Compute the difference $|\phi_0 - \phi_m|$ and compare it with "η". The result will be either

(1) $|(\phi_0 - \phi_m)| \leq \eta$ we stop. This is not likely to happen in early stages of iterations because of choosing "η" too small.

(2) or, $|(\phi_0 - \phi_m)| > \eta$ we go to the second step.

The Second Step:

We will be performing this step in two stages:

(1) Compute the step "direction", denoted by "SDIR", by generating a

uniform random variable [0,1].

(2) Assume arbitrary value for the step size, denoted by "SSIZE" or just "SS".

The Third Step:

Compute the new point after making a step of size SS in the direction SDIR and denote it by $\underset{\sim}{\pi}^*$, where

$$\underset{\sim}{\pi}^* = \underset{\sim}{\pi}_0 + (\text{SS})(\text{SDIR}) \tag{3.18}$$

Notice that $\underset{\sim}{\pi}^*$ is referred to as "XNEW" in our computer program.

The Fourth Step:

Check if the value $\underset{\sim}{\pi}^*$, given by (3.18), will be lying in the domain of definition of the estimated probabilities given by definition (3.4.1). In other words, we check if the elements of $\underset{\sim}{\pi}^*$ will be summing to unity. We have one of the following two cases:

(1) $\underset{\sim}{\pi}^*$ is not in the set $\Omega = [0,1]$, we go back to the second step to compute another random direction.

(2) If $\underset{\sim}{\pi}^*$ is an element of Ω, we compute:

(a) the value of the objective function $\phi(\underset{\sim}{\pi}^*)$ or "ϕ_1", and

(b) the difference $|\phi_1 - \phi_0|$, denoted by "$\Delta\phi$".

The Fifth Step:

Compare the values of ϕ_1 and ϕ_0, the outcome will be one of

(1) If $\underline{\phi_1 \geq \phi_0}$, we accept the step, being determined by (3.18) and proceed by

(a) taking the most recent value $\underset{\sim}{\pi}^*$ to be our initial $\underset{\sim}{\pi}_0$.

(b) taking ϕ_1 to be our initial ϕ_0, and

(c) computing the difference $|\phi_0 - \phi_m|$ which will be

(i) either $|\phi_0 - \phi_m| \geq \eta$, we take the new values to start another random direction by repeating step (2).

(2) If $\underline{\phi_1 < \phi_0}$, i.e., if we are diverging from the maximum we are aiming to achieve , then

(a) Consider the greater value ϕ_0, to define the conditional probability of acceptance p as

$$p = \exp(-B\ \phi_0^G\ \Delta\phi) \tag{3.19}$$

where "B" is arbitrary constant, and "G" is a negative arbitrary constant.

(b) Generate a uniform [0,1] random number denoted by V.

(c) Compare p with the value of V , we have one of the two cases

(i) If $p \geq V$, we accept the step, and so

- set $\underline{\pi}^* \equiv \underline{\pi}_0$ and $\phi_1 \equiv \phi_0$,
- repeat the loop starting from finding another direction

(ii) if $p < V$, reject the step, and go back to step (2) to try another step moving from another random direction.

5.4.1 A SIMULATION STUDY

4.1 Introduction and Motivation

In the framework of our structure of the problem of estimating the mixing probabilities in a mixture of exponentials, we carry out simulation study to assess the general impact of changing the algorithm's parameters on two features of the solution. These features of the resulting solution are:

(i) its sensitivity to changes in these parameters.

(ii) its behaviour from the convergence point of view.

The two aspects will be achieved by performing a sensitivity analysis in which we propose a criterion for the convergence of the results. Thus, by this analysis we will be able to explore some new characteristics of our problem to which the adaptive GSA algorithm is applied.

The set of data is simulated, using a sample of size n = 30, from a mixture of exponential distributions with parameters ($\lambda_1 = 1$, $\lambda_2 = 3$, $\lambda_3 = 6$, $\lambda_4 = 9$), associated with the mixing probabilities ($\pi_1 = .2$, $\pi_2 = .3$, $\pi_3 = .2$, $\pi_4 = .3$). These data will be substituted into our main program for the adaptive GSA algorithm. The initial values for π's (denoted by $\underset{\sim}{\pi}_0$) will be shown in the first two columns of each table. We consider four equally spaced intervals (i.e., the value of k in equation (3.4) equals four). These intervals are (0,2),...,(6,8), where we notice that the width of the intervals is equal to two, and $c_0 = 0$, which is required by equation (3.5). The level of accuracy is denoted by (ETA), and defined in the steps of the adaptive GSA algorithm. Similar definitions of the parameters of the algorithm (like, for example, B, G and p) have been given in the above-mentioned subsection (5.3.4.2).

The simulation study will be performed by applying the previously mentioned steps of the adaptive algorithm, considering two different cases for the number of iterations. A comparison will be made between the results obtained in these cases.

The simulation study will be helpful in clarifying

(i) the meaning and the effect of the parameters, which are specifying the adaptive algorithm and affecting its performance.

(ii) the relations between these parameters. Thus, such information will enhance better understanding for the numerical results and for the algorithm's performance in general.

5.4.2 A Criterion for Convergence:

By recalling the notion of the conditional probability of acceptance p, being defined by equation (3.19), we extend that notion, here, by proposing the following definition

Definition 5.4.2.1 : The unconditional probability of acceptance is the limiting case of the conditional probability of acceptance when the exponent in the right-hand side of equation (3.19) is equal to zero. In other words, in situation when $p = 1$ we will accept the step directly (or unconditionally).

From the previous definition(5.4.2.1) an important relation has been realized, which will be confirmed by our simulation results. That is, the lower the value of the exponent, the more likely the step to be accepted.

Also, using the above definition, we could classify a step to be either beneficial or detrimental. If p equals one, the step is a beneficial one (i.e., will be accepted unconditionally), while if p is less than one, the step is a detrimental one. The later type of steps, will be accepted according to an auxiliary experiment, being shown by case (i) in (c) of the fifth step.

We introduce a second convergence indicator, by calculating the ratio (R) of the number of the conditionally accepted (conditioned upon $p \geq V$) steps - denoted by CT(4) in our computer program - to the total number of detrimental steps - denoted by CT(3).

From the conditional probability of acceptance p and the calculated ratio R we suggest a criterion for measuring the convergence of the adaptive GSA algorithm. Thus, we give the

following definition:

Definition 5.4.2.2 The convergence of the adaptive GSA algorithm could be expressed by the closeness of p to the ratio R. In other words, the closer the ratio R to the value p, the better the results will be from the point of view of having an improved convergence feature.

By the above criterion, we will be able to assess the impact of varying the parameters, specifying the algorithm, on its performance, measured from a convergence point of view. This assessment will be made in our simulation study.

5.4.3 A Sensitivity Analysis

A sensitivity analysis will be carried out, considering the previously mentioned definitions and criteria, and using the set of simulated data together with the other assumptions which have been referred to in subsection (5.4.1). This analysis will be done within the framework of our main problem of estimating the mixing probabilities in a mixture of exponentials.

The sensitivity analysis will produce extensive numerical results. These results show the effectiveness of varying the parameters of the algorithm on its performance from two points of view:

(i) achieving the optimal (in our case the maximum) value of the objective function

$$L = L(\underset{\sim}{\pi}; \underset{\sim}{x}) = \sum_{i=1}^{n} \log \left[\sum_{j=1}^{k} \pi_j I_j(x_i) \right] \qquad (4.1)$$

associated with the optimal set of estimated probabilities, which is denoted by $\underset{\sim}{\pi}^*$. (in our case, it is π_j^*, j=1,...,4 - see subsection (5.4.1))

(ii) reaching a satisfactory rate of convergence, where the definiiton of convergence has

been given in subsection (5.4.2).

The basic numerical results, calculated by performing such sensitivity analysis, can be summed up in the following:

5.4.3.1 The Effect of the Step Size (SS)

The numerical results, show that the step size (SS) is the predominate factor affecting the performance of the algorithm, as far as achieving the optimal value of the objective function, shown by equation (4.1), is concerned.

Table 4.1 shows the results of varying the step size SS on both the objective function (stated by equation (4.1)) and the convergence of the algorithm as given by definition (5.4.2.2).

TABLE 4.1

INITIAL VALUES

		THE ESTIMATED PROBABILITIES ($\hat{\pi}$)				
π_0		SS=.30	SS=.20	SS=.15	SS=.10	SS=.03
π_1	.170	.255	.226	.212	.198	.214
π_2	.318	.342	.334	.330	.326	.305
π_3	.189	.193	.192	.191	.190	.163
π_4	.323	.210	.248	.266	.286	.320
VALUE OF L	12.049	20.313	20.446	20.486	20.492	20.492
VALUE OF (B)		1	1	1	1	1
VALUE OF (G)		−1	−1	−1	−1	−1
LEVEL OF ACCURACY "ETA"		1E-4	1E-4	1E-4	1E-4	1E-4
CONVERGENCE	p	.88610	.87360	.95950	.75419	.91309
INDICATORS	R	1.0	1.0	.95000	1.0	1.0

The above results, given by table 4.1, have been derived in the framework of subsection 5.4.1, specifically for a true mixing distribution which places discrete probabilities ($\pi_1 = .2$, $\pi_2 = .3$, $\pi_3 = .2$ and $\pi_4 = .3$), at ($\lambda_1 = 1$, $\lambda_2 = 3$, $\lambda_3 = 6$, $\lambda_4 = 9$).

Concluding Remarks

(i) By keeping the values of the step size (SS) as small as .10 or .03 we always achieve the optimal value of the objective function (L). This objective function has been shown by equation (4.1).

(ii) A review of the values of the conditional probability of acceptance p under various step sizes SS, shows us that: the larger the step size the smaller the probability p of accepting this step.

(iii) As far as the convergence of the solution is concerned, we notice that the best result, according to the criterion being introduced by definition 5.4.2.2, is reached when the step size (SS) is equal to .15. The poor convergence results of the optimal L^* cases (SS = .10 and .03) will be analyzed further in the light of changing the other parameters of the algorithm such as G and B. This will be made to explore ways of improving such results.

5.4.3.2 The Effect of the Parameter G

It has been mentioned in relation (3.19) that G is an arbitrary negative constant. We give the following table 4.2 which considers the same setting given in the introductory subsection 5.4.1, but assumes decreasing the value of the parameter G. The results will be as follows:

TABLE 4.2

INITIAL VALUES		THE ESTIMATED PROBABILITIES ($\hat{\pi}$)			
π_0		G = −2	G = −3	G = − 4	G = −5
π_1	.170	.198	.198	.198	.198
π_2	.318	.326	.326	.326	.326
π_3	.189	.190	.190	.190	.190
π_4	.323	.285	.285	.286	.285
VALUE OF L	12.049	20.492	20.492	20.492	20.492
LEVEL OF ACCURACY "ETA"		1E−6	1E−6	1E−6	1E−6
STEP SIZE (SS)		.10	.10	.10	.10
CONVERGENCE p		.97042	.99681	.99966	.99996
INDICATORS R		1.0	1.0	1.0	1.0

Concluding Remarks

(i) Under a suitably chosen small step size SS = .10, the resulting solution is always optimal from the point of view of achieving the maximum value of the objective function L.

(ii) By reviewing the convergence results of table 4.2, given above, we notice that the smaller the value of the parameter G the better these results (measured by the closeness of p and R) will be.

5.4.3.3 The Effect of the Parameter B

The numerical results, show an undesirable effects of increasing the value of B on the convergence results. This is realized from the observation that the larger the value of B the more divergent the convergence indicators (p and R) will be.

Table 4.3 shows this efect.

TABLE 4.3

INITIAL VALUES		THE ESTIMATED PROBABILITIES ($\hat{\pi}$)				
π_0		β=1	β=2	β=3	β=4	β=5
π_1	.170	.214	.214	.214	.214	.214
π_2	.318	.303	.303	.303	.303	.303
π_3	.189	.163	.163	.163	.163	.163
π_4	.323	.320	.320	.320	.320	.320
VALUE OF L	12.049	20.492	20.492	20.492	20.492	20.492
VALUE OF (G)		−5	−5	−5	−5	−5
STEP SIZE (SS)		.03	.03	.03	.03	.03
CONVERGENCE p		.99999	.99999	.99998	.99997	.99996
INDICATORS R		1.0	1.0	1.0	1.0	1.0

As with the previous cases, the data used in the above table 4.3, is simulated - see subsection 5.4.1 - using a true discrete mixing probabilities which are $\pi_1 = .2$, $\pi_2 = .3$, $\pi_3 = .2$ and $\pi_4 = .3$, at ($\lambda_1 = 1$, $\lambda_2 = 3$, $\lambda_3 = 6$, $\lambda_4 = 9$).

Concluding Remarks

(i) The optimal value of the objective function L^* (20.492) is achieved by choosing a small step size SS = .03 in all cases of B. This value of L^* is associated with the set of estimated mixing probabilities $\hat{\pi}_1 = .214$, $\hat{\pi}_2 = .303$, $\hat{\pi}_3 = .163$ and $\hat{\pi}_4 = .320$. This is as far as the optimality of the results of table 4.3 is concerned.

(ii) From the convergence point of view, table 4.3 shows us that increasing the value of B has diverted the value of p (from .99999 to .99996) from the value of R.

The above convergence remark, concerning the negative effect of increasing B, coincides with the well-known result in statistical mechanics which states that : the probability of a transit from a state of energy E_1 to another state of energy E_2 (assuming $E_2 > E_1$) is equal to $[\exp\left(\frac{-\Delta E}{KT}\right)]$, where k is a Boltzman's constant, and T is the temperature. Here, the relation is the lower the temperature the smaller the probability of a transition.
Recall equation (3.19) which is

$$p = \exp(-B\, \phi_0^G\, \Delta\phi) \qquad (4.2)$$

and compare it with the previous notion we notice that B in equation (4.2) corresponds to $\frac{1}{KT}$ and the larger the value of B, the smaller the probability of accepting the step.

The sensitivity analysis applied with respect to the parameter G may be helpful in seeking a possible way of overcoming the negative effect (as shown above in (ii)) of increasing B on the convergence results. This will be demonstrated in the following table 4.4, where we have recorded the convergence indicators for increasing values of B coupled with decreasing values of G.

TABLE 4.4

VALUES OF (B)	VALUES OF THE PARAMETER (G)			
	G=−1		G=−1	
	CONVERGENCE INDICATOR		CONVERGENCE INDICATOR	
	p	R	p	R
B=1	.75419	1.0	.99993	1.0
B=2	.56881	1.0	.99989	1.0
B=3	.42899	1.0	.99986	1.0

Concluding Remarks

The above table 4.4 shows that decreasing the parameter G from −1 to −5 has a profound influence on the convergence results. It is not only improving the convergence results (measured by the closeness of p and R) at each individual level of B, but it is also slowing down the rate of deterioration in convergence associated with higher values of B.

5.4.3.4 The Effect of the Level of Accuracy (ETA)

We have repeated running the above-mentioned cases of table 4.1 using higher levels of accuracy, such as ETA = 1E−5 and ETA = 1E−6. We have found that there is no effect for choosing higher levels of accuracy either on the optimal value of the objective function (represented by equation (4.1)) or on the convergence results.

A sensitivity analysis may be employed, here, for improving the convergence results (measured by definition 5.4.2.2), irrespective of the level of accuracy ETA, by either decreasing G or decreasing B or both, as shown by table 4.5

TABLE 4.5

LEVEL OF ACCURACY(ETA)		ETA=1E−3		ETA=1E−4		ETA=1E−5	
		VALUES OF B		VALUES OF B		VALUES OF B	
		B = 3	B = 2	B = 3	B = 2	B = 3	B = 2
G=−1	p	.42899	.56881	.42899	.56881	.42899	.56881
	R	1.0	1.0	1.0	1.0	1.0	1.0
G=−2	p	.91387	.94172	.91387	.94172	.91387	.94172
	R	1.0	1.0	1.0	1.0	1.0	1.0
G=−3	p	.99046	.99363	.99046	.99363	.99046	.99363
	R	1.0	1.0	1.0	1.0	1.0	1.0
G=−4	p	.99898	.99932	.99898	.99932	.99898	.99932
	R	1.0	1.0	1.0	1.0	1.0	1.0

A Concluding Remarks

From the numerical results, shown by table 4.5 we can conclude the following:

(i) Decreasing the parameter G as well as decreasing B will have a positive effect on the convergence results, represented by the closeness of p to R. This effect is still valid even under lower levels of accuracy ETA.

(ii) The rate of improvement in convergence associated with decreased values of G is more faster than with decreased values of B.

5.4.3.5 The Effect of the Number of Iterations

All the previously mentioned results, discussed in the above four subsections, have been calculated for a total number of iterations (denoted by J) equals 10.000.

We have made an attempt to reduce the number of iterations to only J = 1000, and have given the results in the following table 4.6. This table demonstrated the effect of varying the parameter G - see its effect in subsection 5.4.3.2 - with the number of iterations J. The results have been given in the sequel.

The results of the following table are produced using a set of data simulated from a mixture of exponentials with the true mixing probabilities $\pi_1 = .2$, $\pi_2 = .3$, $\pi_3 = .2$ and $\pi_4 = .3$. The objective function, shown by equation (4.1) achieves its optimal value $L^* = 20.492$. This optimal value L^* is associated with the estimated mixing probabilities, which are $\hat{\pi}_1 = .214$, $\hat{\pi}_2 = .303$, $\hat{\pi}_3 = .163$ and $\hat{\pi}_4 = .320$. (see subsection 4.1 for details).

A Concluding Recommendation

From the previous table 4.6, we see that at a certain value of G the convergence results, being measured by the closeness of p and R are better for smaller number of iterations. Having concluded that, and because of the need to minimize the cost of computer time by using smaller number of iterations, then we recommend using the smaller number of iterations J = 1000 instead of J = 10000.

TABLE 4.6

		G = -1		G = -2		G = -3		G = -4	
NUMBER OF ITERATIONS		J=10000	J=1000	J=10000	J=1000	J=10000	J=1000	J=10000	J=1000
CONVERGENCE	P	.42899	.96279	.91387	.99699	.99046	.99976	.99898	.99998
INDICATORS	R	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
STEP SIZE "SS"		.10	.10	.10	.10	.10	.10	.10	.10
LEVEL OF ACCURACY "ETA"		1E-3	1E-3	1E-3	1E-3	1E-3	1E-3	1E-3	1E-3
VALUE OF "B"		3	3	3	3	3	3	3	3

5.5 A COMPARATIVE STUDY OF SOME RELATED METHODS OF ESTIMATION

5.5.1 Introduction

Having mentioned - see section three of this chapter - that our technique for estimating the mixing probabilities is based upon some concepts of a two nonparametric methods of density estimation, namely, the kernel method and the maximum likelihood method, then we introduce a brief comparative study between these methods and our optimization method.

The comparative study will be aiming at (i) exploring the relations between these methods and ours, and (ii) discussing the connections between some of their main characteristics. Such comparisons will be helpful in enhancing better understanding of the features of our optimization technique of estimation.

Having based our optimization technique on an idea borrowed from the kernel method of density estimation, we will start by making a comparison between the two methods. Throughout this comparison an analogy between these methods has been deduced and presented by a numerical example. This example exhibits some sort of common feature between our technique of estimation and the kernel method of density estimation.

Because of using the likelihood principle as a corner stone for constructing our optimization technique of estimation (notice the third assumption equation (3.4)), we will establish some comparisons between our technique and the nonparametric maximum likelihood method of density estimation. Also, an analogy between the two methods has been realized, and its impact on the results has been discussed.

5.5.2 The Optimization Method Versus the Kernel Method

We start by reviewing Parzen (1962) kernel estimator, being given in section 3 equation

(3.1), as

$$\hat{f}_n(x) = \int_{-\infty}^{\infty} \frac{1}{h_n} K\Big(\frac{x-y}{h_n} \Big) dF_n(y) = \frac{1}{nh_n} \sum_{j=1}^{n} K\Big(\frac{x-x_j}{h_n} \Big), \qquad (5.1)$$

where $K(\cdot)$ is the kernel function.

Knowing that the empircial distribution function "F_n", appearing in 5.1, is a discrete distribution placing a mass ($\frac{1}{n}$) at each of the n observations, thus the idea of the estimator is that, formula (5.1) smears this probability out continuously, according to the "choice" of the kernel $K(\cdot)$.

5.5.2.1 A Comparison between the Optimization and Kernel Methods

From a practical point of view, it is more convincing and realistic to place unequal weights on the integrals I_j,$(j = 1,...,k)$(as has been assumed by the optimization technique equation (3.4)) than placing equal weights $\frac{1}{n}$ on each of the n kernels shown by equation (5.1).

In the optimization technique, and because of using the likelihood criterion to estimate these unequal weights (π_j, $j = 1,...,k$) (as shown by the fourth assumption), we could achieve the following advantages over the kernel method:

(1) reflect the structure of the data

(2) avoid the effect of the arbitrary choice of the kernels $K(\cdot)$ by the experimenter.

An important remark, to be made here is that the kernel estimate $\hat{f}_n$, as appears in (5.1), depends linearly on the kernel function $K(\cdot)$, hence the appearence and properties of $\hat{f}_n$ are dictated more by the arbitrary choice of the kernel $K(\cdot)$. Such choice is not determined by the data, but rather by the user himself. In our optimization method, since the data is used to determine the estimated probabilities with relatively minor arbitrary choices, then

the resulting estimates are determined in more data-oriented (or nonparametric) sense.

5.5.2.2 The Connection between the Optimization and Kernel Methods

The connection between our optimization technique and the kernel method is based upon the notion that the cell width $(c_j - c_{j-1})$, $(j = 1,...,k)$, is being considered as the smoothing parameter (or tuning parameter) in our method. This will be shown to be analogous to the smoothing parameter h_n, shown by equation (5.1), for the kernel estimate. Thus, widening the cell width in our optimization technique will improve the results, similar to the case of increasing the smoothing parameter h in the kernel method.

In our optimization technique, the improvement of the results is represented by achieving an optimal (maximum) value of the objective function (shown by equation (3.4)). But, in the kernel method such improvement is represented by getting a smoother density estimate.

The following table 5.1 sums up the results of an example, where we use a mixture of four exponentials (i.e., k is equal to four in equation (3.4)) with paramerters ($\lambda_1 = 1$, $\lambda_2 = 3$, $\lambda_3 = 6$, $\lambda_4 = 4$) associated with the mixing probabilities ($\pi_1 = .2$, $\pi_2 = .3$, $\pi_3 = .1$, $\pi_4 = .4$) respectively.

TABLE 5.1

VALUES OF $c_j(c_0=0)$					THE CELL WIDTH $(c_J - C_{J-1})$	THE ESTIMATED PROBABILITIES				THE MAXIMUM OF THE FUNCTION
c_0	c_1	c_2	c_3	c_4	J=1,...,4	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	$\hat{\pi}_4$	$L(\hat{\underset{\sim}{\pi}}; \underset{\sim}{x})$
0	3	6	9	12	3	.174	.316	.186	.324	31.253
0	4	8	12	16	4	.178	.314	.187	.321	37.529
0	5	10	15	20	5	.175	.314	.187	.324	41.663
0	6	12	18	24	6	.173	.318	.186	.323	44.592
0	7	14	21	28	7	.176	.314	.187	.323	46.772

For reasons of comparison, table 5.1 fixes the number of intervals $k = 4$, varies the width of these intervals, and presents the results of five of these cases.

The above table 5.1 illustrates a way for improving the results (notice the last column), obtained from the optimization technique. Such improvement is achieved by increasing the cell width $(c_j - c_{j-1})$, $j = 1,...,k$, being shown in the second column. This is analogous to the way of improving the kernel density estimate (i.e. getting a reasonably smoother estimate) by increasing the smoothing parameter h in formula (5.1).

5.5.3 The Optimization Method versus the (M.P.L) Method

5.5.3.1 Definitions and Notations

Given a random sample $x_1,...,x_n$ from a density function f defined on the set $\Omega = (a,b)$,

we define

Definition 5.5.3.1 Let $H(a,b) = H(\Omega)$ be a manifold in $L_1(\Omega)$. A manifold is a set of reasonably similiar functions.

Definition 5.5.3.2 Consider a manifold $H(\Omega) \subset L_1(\Omega)$, and the following constrained optimization problem

$$\begin{aligned} &\text{maximize } L(f) \\ \text{subject to } f \in H(\Omega),\ &\int f(x)dx = 1 \qquad (5.2) \\ &\text{and } f(x) \geq 0 \ \forall \quad x \in \Omega \end{aligned}$$

The integration in 5.2 is with respect to Lebesgue measure. Any solution to problem 5.2 is defined to be a maximum likelihood estimate based on the sample $x_1,...x_n$.

The main drawback of problem 5.2 is that the likelihood considered as a functional is unbounded, i.e.,

$$\text{Max} \prod_{j=1}^{n} f(x_j) = +\infty, \qquad (5.3)$$

where the maximum is taken over all probability densities on the real line. In other words, a linear combination of Dirac delta function at the sample points satisfies the constraints and results in a value of $+\infty$ for the objective likelihood functional.

For the sake of avoiding this Dirac catastroph and guaranteeing the existence of problem 5.2, Good and Gaskins (1971) introduced the maximum penalized likelihood method (M.P.L) of density estimation.

Definition 5.5.3.3

Assume $x_1,\ldots,x_n$ be independent observations from a distribution function F with density function f. Define the penalized likelihood function as

$$L_\alpha(f) = \sum_{j=1}^{n} \log f(x_j) - \alpha\, \Phi(f), \tag{5.4}$$

where the positive number α is the smoothing parameter, and $\Phi(f)$ is the roughness penalty functional

Definition 5.5.3.4 An estimate $\hat{f}$ will be called a maximum penalized likelihood density estimate if

$$\begin{aligned} &\hat{f} \text{ maximizes } L_\alpha(f) \\ &\underline{\text{subject to}} \int f(x)dx = 1, \quad f(x) \geq 0 \quad \forall x \\ &\text{and } \Phi(f) < \infty, \end{aligned} \tag{5.5}$$

In other words, $\hat{f}$ maximizes $L_\alpha(f)$ over the class of all f satisfying 5.5.

5.5.3.2 A Comparison between the Optimization and (M.P.L) Methods

Some remarks, which are necessary for constructing a comparison between the optimization and the (M.P.L) methods of estimation will be discussed briefly, with emphasis on detecting an analogy between one of the main features of them.

Having used a histogram-like idea and the maximum likelihood criterion in the optimization technique, then the resulting estimates can be considered as nonparametric ones. Also, unless a specific functional form for the density is assumed in $H(\Omega)$ - see definition 5.5.3.1 -

the (M.L) estimate is a nonparametric one. Therefore, both of the two methods have the same nature of being nonparametric methods of estimation.

Unlike the kernel method, both of the above-mentioned method (namely, the optimization and (M.L)) have the desired property of being data-oriented methods of estimation with a minor degree of arbitrariness in the estimation procedure.

We have mentioned, in the previous subsection, the idea of considering the cell width as an analogy of the smoothing parameter in the density estimation context. Also, from equation (5.4), we notice that as the parameter α gets bigger the resulting density estimate gets smoother. This establishes some sort of connection between the optimization technique and the (M.P.L) method. This connection manifests itself in the fact that increasing the cell width in our method has a positive effect (see table 5.1) as increasing the parameter α in the (M.P.L) method. The improvement in the later case is represented by getting a smoother maximum penalized likelihood estimate.

5.5.4 The limiting Behaviour: An Important Rationale

The limiting behaviour of the resulting estimates will be helpful in realizing the connection between our optimization technique and the (M.P.L) method. That is, by narrowing the cell width (i.e., increasing the number of cells to infinity), we will have a Dirac catastroph (i.e. , too rough estimates). This is exactly the same as when the smoothing parameter α - in equation (5.4) - approaches the value zero in (M.P.L) method. In conclusion there is an analogy between narrowing the cell width (i.e., letting $k \rightarrow \infty$) in our optimization technique and taking α too small in (M.P.L) method.

Finally, we suggest an important rationale which could be relating the previous limiting behaviour of the optimization technique to a general feature of all nonparametric probability

density estimation methods. This states that, since one of the objects of nonparametric methods is to investigate the effect of relaxing parametric assumptions, it seems sensible that the limiting case of a nonparametric density estimate should (or ought to) be a natural parametric estimate.

CHAPTER SIX

THE CONCLUSIONS

6.1 Main Achievements and Contributions

This thesis introduces three main ideas. It proposes two techniques and a kernel method for estimating the mixing density in the mixture of exponentials setting.

The first technique offers us a procedure in which our estimation problem has been represented as an integral equation, specially Laplace transform. Then, a truncated infinite series formula has been suggested, as an approximation for such integral transform. Depending upon this, an analogy has been proposed between the truncation point m (which is assumed to be different from the value of the sample size n) and the notion of the smoothing parameter in the context of the estimation of a probability density function.

In this Laplace-based technique, an estimation procedure has been suggested. This enables us to study, the issue of controlling the shape of the estimated mixing density. It has been found that the determination of a suitable value of the parameter m gives us an acceptable shape of the mixing density estimate. This is because, if m is chosen too large, this causes an undesirable difficulty in estimating too many parameters (represented by $\hat{\alpha}_k$ and $\hat{B}_k$, k = 1,...,m) relative to the available sample size. On the other hand, if m is selected too small, then important detectable features of the estimated mixing density may not appear.

The numerical study of the introduced Laplace-based technique shows its dependence not only upon the parameter m, but also on the sample size n. It has been noticed that an estimate of a mixing density which is based upon too small value of m, will be sharply (or undesirably) affected by varying the sample size n. On the other hand, the sensitivity of the estimated

density to changes in the sample size n will be unacceptably low in case of using too large a value of m. As a result of our simulation study, a balancing situation has been reached, which suggests some sort of compromise between these two factors to get an acceptable shape for the estimated density. That is, the choice of a moderate value of the parameter m (relative to the sample size n), results a mixing density estimate which is reasonably sensitive to the variations in the sample size n.

The second contribution in this thesis is the introduction of a kernel method for estimating the mixing density in our mixture problem. An empirical Bayes framework has been suggested to construct this kernel estimator. This estimator $\hat{\pi}(\lambda)$ is proved to be unbiased under the assumption that the kernel function is equal to the conditional distribution of λ given a single observation x.

A lemma has been introduced, by which some limiting properties of the derived estimator $\hat{\pi}(\lambda)$ have been given. The limiting behaviour of this kernel-type estimator (as $r \to \infty$) is found to be analogous to the behaviour of a density estimator (M.P.L for example) when the smoothing parameter approaches zero. Thus, under the assumption that the values of m and n are equal, the parameter r^{-1} is considered to be the smoothing parameter in our kernel estimator $\hat{\pi}(\lambda)$.

The investigation of the tail limiting behaviour, suggests that a rescaling operation is necessary for the removal of the bias up to a linear terms. An artificial example has been given to illustrate this result. This extends the applicability of the proposed lemma to any other density f(x). Also, by studying the mode, it has been found that a scaling operation (scale factor of two) is appropriate to achieve unbiasedness in the limit.

Some moment properties have been derived. This is done to assess how far the behaviour of the moments of the estimator $\hat{\pi}(\lambda)$ is similar to that of the true underlying density $\pi(\lambda)$. In this case, the limiting behaviour (as $r \to \infty$) has shown us an "inverse-mean" biasedness.

This confirms the conclusions of the scaling downwards which have been derived for the mode and the tail of the estimator $\hat{\pi}(\lambda)$.

Finally, different sets of real data have been used to represent the estimator $\hat{\pi}(\lambda)$ graphically. These graphs show us the mixed exponential density as a special case of our estimator $\hat{\pi}(\lambda)$ when the value of the parameter r equals one.

The third approach introduced in this thesis is an optimization technique for nonparametric mixing density estimation.

Having used the generalized simulated annealing (G.S.A) algorithm, we have been able to : (i) generalize some concepts, borrowed from the histogram and the kernel methods of density estimation, and (ii) discuss the possible adaptations required for the application of the (G.S.A) algorithm to the mixture of exponentials case.

In a simulation study, a sensitivity analysis has been carried out, in which a criterion for the convergence of the adaptive (G.S.A) algorithm has been defined. Thus, we have been able to assess the impact of varying the parameters of the algorithm on its performance, measured from a convergence viewpoint. This sensitivity analysis measures, also, the performance of the algorithm, represented by achieving the maximum (the optimal in our case) value of the objective fucntion.

A connection between our optimization technique and the kernel method of density estimation has been suggested. This is confirmed by giving a numerical example. In this context, it is shown that widening the cell width $(c_j - c_{j-1})$ will improve the results in our optimization technique. This is similar to improving the kernel estimate by increasing the smoothing parameters. Thus, we have a tuning device, represented by the cell-width, by which better results can be obtained.

Having based our optimization technique on the likelihood of the data, then it is more date-oriented (or nonparametric) than the kernel approach, where the arbitrary choice of the kernel function represents its corner-stone.

Finally, in comparing any two methods of density estimation, there is a fundamental issue, represented by the ability of a given method to achieve two goals, namely, the diagnosis and the estimation. By starting with a bimodal histogram, in our optimization technique, we have been interested in the problem of (i) breaking down what appears to be a mixed distribution, and (ii) estimating the mixing density.

6.2 Suggestions for Further Research

Some asymptotic properties of our Laplace-based technique could be derived in conjunction with the optimal choice of our tuning parameters n and m (assuming that they are different).

Firstly, we may study the optimal truncation point m. This optimal choice could be defined in terms of minimizing the integrated mean square error (I.M.S.E) in an asymptotic sense. In this context, we suggest, for reasons of computational convenience, that the (I.M.S.E) could be approximated by the average square error at the observations. This takes the form:

$$\frac{1}{n}\sum_{j=1}^{n}[\hat{f}(x_j) - f(x_j)]^2, \qquad (6.1)$$

where $\hat{f}$ is the density estimator of the true density $f(\cdot)$.

Some ideas of Wahba's paper [1977] may be used in optimally choosing these values of m. For example, the method of cross-validation could be employed as a powerful data-based criterion for this optimal choice.

Secondly, the sample size n could be studied in order to explore the possibility of defining some sort of asymptotic "normality" of the estimated mixing density function. An idea, which has been introduced by Barron, A. [1986] could be helpful in this context. This discusses the convergence to normality of a density function in the sense of relative entropy.

Finally, under the equality assumption of the tuning parameters m and n, we have established a connection between our kernel-type estimator (see chapter IV) and the usual Parzen (1962) kernel estimator. This suggests that r^{-1} is the smoothing parameter in our kernel method. Thus the reviewed methods of chapter II (see for example subsections (2.3.3) and (2.3.3.2)) could be employed for the determination of an "optimal" value of this smoothing parameter.

APPENDIX

(A1) Derivation of the Likelihood Function: The Optimization Technique Case

We will start by recalling our definition, being given in the second assumption, and by which we represent the histogram-like mapping I_j, $j = 1,...,k$, as

$$I_j(x) = \int_{c_{j-1}}^{c_j} f(x; \lambda)d\lambda \quad , \qquad j = 1,...,k \tag{1}$$

where $f(x; \lambda) = \lambda\, e^{-\lambda x}$ is the exponential density function. In the above definition k is the number of equally spaced cells (c_{j-1}, c_j), $j = 1,...,k$. The amount of probability (height) at each support interval of this histogram is denoted by π_j, $j = 1,..., k$.

To evaluate the integral in equation (1), we take the case where $j = 1$, to get for the first interval $(0,c_1)$ the following integral

$$I_1 = \int_0^{c_1} \lambda e^{-\lambda x}\, d\lambda \tag{2}$$

Thus, by making the substitution $u = \lambda$ and $dv = e^{-\lambda x}\, d\lambda$, and integrating by parts, we get from equation (2), the following

$$I_1 = -\frac{\lambda e^{-\lambda x}}{x} \Big|_0^{c_1} + \frac{1}{x}\int_0^{c_1} e^{-\lambda x}\, d\lambda$$

$$I_1 = -\frac{1}{x} c_1 e_x^{-c_1 x} - \frac{1}{x^2}[e^{-c_1 x} - 1]$$

$$I_1 = -\frac{1}{x^2}[e^{-c_1 x}(c_1 x + 1) - 1] \tag{3}$$

For $j = 2$ in equation (1), we get for the second interval (c_1, c_2) the following integral

$$I_2 = \int_{c_1}^{c_2} \lambda e^{-\lambda x}\, d\lambda \qquad (4)$$

Thus, by applying the same rule of integration by parts as before, we get from equation (4) the following

$$I_2 = -\frac{\lambda e^{-\lambda x}}{x}\Big|_{c_1}^{c_2} + \frac{1}{x}\int_{c_1}^{c_2} e^{-\lambda x}\, d\lambda$$

$$I_2 = -\frac{1}{x}[\, c_2 e^{-c_2 x} - c_1 e^{-c_1 x}] - \frac{1}{x^2}[\, e^{-c_2 x} - e^{-c_1 x}]$$

$$I_2 = -\frac{1}{x^2}[\, e^{-c_2 x}(c_2 x + 1) - e^{-c_1 x}(c_1 x + 1)] \qquad (5)$$

From equation (1) and also by taking $j = 3$, we get for the third interval (c_2, c_3), the following integral

$$I_3 = \int_{c_2}^{c_3} \lambda\, e^{-\lambda x}\, d\lambda \qquad (6)$$

Integrating by parts, we get from equation (6) the following

$$I_3 = -\frac{\lambda e^{-\lambda x}}{x}\Big|_{c_2}^{c_3} + \frac{1}{x}\int_{c_2}^{c_3} e^{-\lambda x}\, d\lambda$$

$$I_3 = -\frac{1}{x}[\, c_3 e^{-c_3 x} - c_2 e^{-c_2 x}] - \frac{1}{x^2}[\, e^{-c_3 x} - e^{-c_2 x}]$$

$$I_3 = -\frac{1}{x^2}[\, e^{-c_3 x}(c_3 x + 1) - e^{-c_2 x}(c_2 x + 1)] \qquad (7)$$

By continuing the previous task, we get for the k^{th} interval (c_{k-1}, c_k) the following integral

$$I_k = \int_{c_{k-1}}^{c_k} \lambda\, e^{-\lambda x}\, d\lambda \qquad (8)$$

In line with the previous evaluations of the above integrals, we get, for equation (8) the following result

$$I_k(x) = -\frac{1}{x^2}[\, e^{-c_k x}(c_k x + 1) - e^{-c_{k-1} x}(c_{k-1} x + 1)] \qquad (9)$$

By assuming (as in our assumptions) that we have k of these support intervals associated with k heights π_j,)j = 1,...,k), then the joint density function (for a particular x) will be

$$f(x) = \sum_{j=1}^{k} \pi_j\, I_j(x) \qquad (10)$$

Thus, given "n" independent observations on the random variable X, with the density (10), the likelihood function will be (denoting it by L)

$$L(\underset{\sim}{\pi};\underset{\sim}{x}) = \prod_{i=1}^{n} \Big\{ \sum_{j=1}^{k} \pi_j\, I_j(x_i) \Big\} \qquad (11)$$

Representation (11), given above, is in fact a polynomial of the n^{th} degree in π's. Hence, we have to have a program for maximizing the n^{th} degree polynomial for each n.

REFERENCES

[1] ANDERSON, G. (1969) A comparison of methods for estimating a probability density function. Doctoral dissertation, University of Washington

[2] BARRON, A. (1986) Entropy and central limit theorem. The Annals of Probability. 14, 336-342.

[3] BARTLETT, M.S. (1963) Statistical estimation of density functions. Sankhya Ser. A, 25, 245-254.

[4] BICKEL, P.J. and ROSENBLATT, M. (1973). On some global measures of the deviation of density function estimates. Ann. Statist., 1, 1071-1095.

[5] BLAYDON, C.C. (1967). Approximation of distribution and density functions. Proc. IEEE 55, 231-232.

[6] BLUM, J.R. and SUSARLA, V. (1977). Estimation of a mixing distribution function. Ann. Probability, 5, 200-209.

[7] BOES, D.C. (1966). On the estimation of mixing distributions. Ann. Math. Statist. 37, 177-188.

[8] BOHACEVSKY, J, JOHNSON, I, and STEIN, (1986). Generalized simulated annealing for function optimization.Technometrics, 28 ,209-217.

[9] BONEVA, L.I., KENDALL, D.G. and STEFANOV, I. (1971). Spline transformation : three new diagnostic aids for the statistical data-analyst (with discussion). J. Roy. Statist. Soc. B, 33, 1-70.

[10] BRILLINGER, D. (1969). An asymptotic representation of the sample distribution function. Bull. Amer. Math. Soci., 75, 545-547.

[11] BOWMAN, A.W. (1984). An alternative method of cross-validation for the smoothing of density estimates. Biometrika, 71, 353-360.

[12] BOWMAN, A.W. (1985). A comparative study of some kernel-based nonparametric density estimators. J. Statist. Comput. Simul., 21, 313-327.

[13] BOWMAN, A.W., HALL, P. and TITTERINGTON, D.M. (1984). Cross-validation in nonparametric estimation of probabilities and probability densities. Biometrika, 71, 341-351.

[14] ČENCOV, N.N. (1962). Evaluation of an unknown distribution density from observations. Soviet Math. 3, 1559-1562.

[15] CHOI, I. and BULGREN, J. (1968). An estimation procedure for mixtures of distributions. (with discussion). J. Roy Statist. Soc. B, 33, 326-329.

[16] CHOW, Y.S. GEMAN, S. and WU, L.D. (1983). Consistent cross-valiation density estimation. Ann. Statist., 11, 25-38.

[17] COPAS, J.B. (1978). Discussion of Dr. Leonard's Paper. J. Roy. Statist. Soc. B, 40, 113-146.

[18] CRAVEN, P. and WAHBA, G. (1976). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the generalized cross-validation. Numer. Math., 31, 377-403.

[19] DANKMAR, B. and HOFFEMANN, K. (1982). Numerical techniques for estimating probabilities. J. Statist. Comput. SImul., 14, 283-293.

[20] DE MONTRICHER, G.M. (1973). Nonparametric Bayesian estimation of probability densities by function space techniques. Doctoral dissertation, Rice University, Houston, Texas.

[21] DE MONTRICHER, G.M., TAPIA, R.A. and THOMPSON, J.R. (1975). Nonparametric maximum likelihood estimation of probability densities by penalty function methods. Ann. Statist., 3, 1329-1348.

[22] DUIN, R.P.W. (1976). On the choice of smoothing parameters for Parzen estimators of probability density functions. IEEE Trans. Comput., C-25, 1175-1179.

[23] EPANECHNIKOV, V.A. (1969). Nonparametric estimation of a multivariate probability density. Theor. Prob. Appl. 14, 153-158.

[24] FERGUSON, T.S. (1973). A Bayesian analysis for some nonparametric estimates. Ann. Statist., 1, 209-230.

[25] FRYER, M.J. (1976). Some errors associated with the nonparametric estimation of density functions. I. Inst. Maths. Applics., 18, 371-376.

[26] FRYER, M.J. (1977). A review of some nonparametric methods of density estimation. J. Inst. Maths. Applics., 20, 335-354.

[27] GHORAI, J. and RUBIN, H. (1979). Computational procedure for maximum penalized likelihood estimate. J. Statist. Comput. Simul., 10, 65-78.

[28] GOLDSTEIN, M. (1975). A note on some Bayesian nonparametric estimates. Ann. Statist., 3, 736-740.

[29] GGOD, I.J. (1965). The estimation of probabilities: An essay on Modern Bayesian Methods. Massachusetts Institute of Technology, Cambridge.

[30] GOOD, I.J. (1971). A nonparametric roughness penalty for probability densities. Nature, Lond., 219-239.

[31] GOOD, I.J. and GASKINS, R.A. (1971). Nonparametric roughness penalties for probability densities. Biometrika, 58, 255-277.

[32] GOOD, I.J. and GASKINS, R.A. (1972). Global nonparametric estimation of probability densities. Virginia Journal of Science, 23, 171-193.

[33] GOOD, I.J. and GASKINS, R.A. (1980). Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. J. Amer. Statist. Assoc., 75m 42-73.

[34] HALL, P. (1981). On the nonparametric estimation of mixture proportions. J.R. Statist. Soc. B, 43, 147-156.

[35] HALL, P. (1984). An optimal property of kernel estimators of a probability density. J. Roy. Statist. Soc. B, 46, 134-138.

[36] HANSEN, E., DOYLE, J. and MCNOLTY, F. (1980). Properties of the mixed exponential failure process. Technometrics., 22, 555-565.

[37] HILL, D.L., SAUNDERS, R. and LAUD, P.W. (1980). Maximum likelihood estimation for mixtures. Canad. J. Statist. 8, 87-93.

[38] JEFFREY, H. (1946). An invariant form for the prior probability in estimation problems. Proc. Roy. Soc. A. 186, 453-461.

[39] JEWELL, N.P. (1982). Mixtures of exponential distributions. Ann. Statist. 10, 479-484.

[40] KRONMAL, R. and TARTER, M. (1968). The estimation of probability densities and cumulatives by Fourier series methods, J. Amer. Statist. Assoc. 63, 925-952.

[41] LAIRD, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. J. Amer. Statist. Assoc., 73, 805-811.

[42] LEONARD, T., (1973). A Bayesian method for histograms. Biometrika, 60, 297-308.

[43] LEONARD, T. (1978). Density estimation, stochastic process and prior information (with discussion). J. Roy. Statist. Soc. Ser. B, 40, 113-146.

[44] LINDSAY, B. (1983I). The geometry of mixture likelihoods : A general theory . Ann. Statist . 11 , 86-94.

[45] LOFTSGAARDEN, D.D. and QUESENBERRY, C.P. (1965). A nonparametric estimate of a multivariable density function. Ann. Math. Statist. 38, 1049-1051.

[46] MACDONALD, P.D.M. (1971). Comment on "An estimation procedure for mixtures of distributions" by Choi and Bulgren. J. Roy. Statist. Soc. B, 33, 326-329.

[47] _____(1975). Estimation of finite distribution mixtures. In applied statistics (R.P. Gupta, ed.), 231-245. Amsterdam: North-Holland.

[48] MEEDEN, G. (1972). Bayes estimation of the mixing distribution,the discrete case. Ann. Math . Statist . 43, 1993-1999.

[49] MURRY, G. and TITTERINGTON, D.M. (1978). Estimation problems with data from a mixture. Appl. Statist., 27, 325-334.

[50] MURTHY, V.K. (1965a). Estimation of the probability density. Ann. Math. Statist. 36, 1027-1031.

[51] MURTHY, V.K. (1965b). Estimation of jumps, reliability, and hazard rate. Ann. Math. Statist. 36, 1032-1040.

[52] NADARAYA, E.A. (1965). On nonparametric estimates of density functions and curves. Theory of probability and its applications. 10, 186-190.

[53] NADARAYA, E.A. (1970). Remarks on nonparametric estimates for density functions and regression curves. Theor. Prob. Appl. 15, 134-137.

[54] PARZEN, E. (1962). On the estimation of a probability density function and mode. Ann. Math. Statist., 33, 1065-1076.

[55] _____ (1979). Nonparametric statistical data modelling. J. Amer. Statist. Assoc., 74, 105-131.

[56] PICKANDS, J. (1969). Efficient estimation of a probability density function. Ann. Math. Statist. 40, 854-864.

[57] POINTER, P.C. (1978). Density estimation using orthogonal series - A Bayesian approach. Doctoral dissertation, Oregon State University, Corvallis.

[58] PÓLYA, G. and SZEGO (1925). Aufgaben und Lehrsyze aus der analysis, 2. Springer, Berlin.

[59] QUESENBERRY, C.P. and SCHEULT, A.H. (1971). On unbiased estimation of density function. Ann. Math. Statist., 42, 1434-1438.

[60] RIDER, P.R. (1961a). The method of moments applied to a mixture of two exponential distributions. Ann. Math. Statist., 32, 143-147.

[61] RIESZ, F. and SZ-NAGY, B. (1955). Functional analysis. Unger, New York.

[62] ROBBINS, H. (1948A). Mixtures of distributions. Ann. Math. Statist., 19, 360-369.

[63] ROBBINS, H. (1980). Estimation and prediction for mixtures of the exponential distribution. Proc. Nat. Acad. Science, 77, 2382-2383.

[64] ROSENBLATT, M.M. (1956). Remarks on some nonparametric estimates of a density function. Ann. Math. Statist. 27, 832-837.

[65] _____ (1970). Density estimates and Markov sequences, Nonparametric techniques in statistical inference, ed., M.L. Puri, Cambridge University Press, Cambridge, 199-213.

[66] _____ (1971). Curve estimates. Ann. Math. Statist. 42, 1815-42.

[67] _____ (1979). Global measures of deviation for kernel and nearest neighbour density estimates. In smoothing techniques for curve estimation, 757. Berlin: Springer-Verlag, 181-190.

[68] RUDEMO, M. (1982). Empirical choice of histograms and kernel density estimators. Scand. J. Statist. 9, 65-78.

[69] SCHUSTER, E.F. (1969). Estimation of a probability density and its derivatives. Ann. Math. Statist., 40, 1187-95.

[70] SCHWARTZ, S.C. (1969). On the estimation of a Gaussian convolution probability density. SIAM J. Appl. Math., 17, 447-453.

[71] SCOTT, D.W. (1976). Nonparametric probability density estimation by optimization theoretic techniques. Doctoral dissertation, Rice University, Houston, Texas.

[72] _____ (1979). On optimal and data-based histograms. Biometrika., 66, 605-610.

[73] _____ (1980). Discussion of paper by I.J. Good and R.a. Gaskins. J. Amer. Statist. Assoc. 75, 61-62.

[74] _____ (1985). Averaged shifted histograms : effective nonparametric density estimators in several dimensions. Ann. Statist., 13, 1024-1040.

[75] SCOTT, D.W. and FACTOR, L.E. (1981). Monte Carlo study of three data-based nonparametric density estimators. J. Amer. Statist. Assoc. 76, 9-15.

[76] SCOTT, D.W. and SHEATHER, S.J. (1985). Kernel density estimation with binned data. Comm. Statist.

[77] SCOTT, D.W., TAPIA, R.A. and THOMPSON, J.R. (1976). An algorithm for density estimation. Computer science and statistics : Ninth annual symposium on the interface, Harvard University, Cambridge, Massachusetts.

[78] _____ (1977). Kernel density estimation revisited. Nonlinear analysis, Theory, Methods and Applications, 1, 339-372.

[79] _____ (1980). Nonparametric probability density estimation by discrete maximum penalized-likelihood criteria. Ann. Statist. 8, 820-32.

[80] SHAPIRO, J.S. (1969). Smoothing and approximation of functions. New York: Van Nostrand-Reinhold.

[81] SILVERMAN, B.W. (1978a). Choosing the window width when estimating a density. Biometrika, 65, 1-11.

[82] SILVERMAN, B.W. (1980). Comment on Good and Gaskins paper. J. Amer. Statist. Assoc. 75, 67-68.

[83] _____ (1981b). Using kernel density estimate to investigate multimodality. J. Roy. Statist. Soc. B, 43, 97-99.

[84] _____ (1982b). On the estimation of a probability density function by the maximum penalized likelihood method. Ann. Statist., 10, 795-810.

[85] _____ (1984a). Spline smoothing : the equivalent variable kernel method - Ann. Statist., 12, 898-916.

[86] _____ (1985b). Penalized maximum likelihood estimation. In Kotz, S and Johnson, N.L. (eds.), Wiley Encyclopaedia of Statistrical Sciences, Volume 6. New York:Wiley, 664-667.

[87] SIMAR, L. (1976). Maximum likelihood estimation of a compound Poisson process. Ann. Statist., 6, 1200-1209.

[88] SINGH, R.S. and GASSER, T. (1983). Nonparametric estimates of distribution functions. Commun. Statist. - Theor. Math., 12(18), 2095-2108.

[89] SINGPURWALLA, N.D. and WONG, M.Y. (1983a). Estimation of the failure rate - A study of nonparametric methods, Part I : Non-Bayesian methods, Commun. Statist., Statistical Reviews 559-588.

[90] SMITH, R.L. (1986). A short course in statistical analysis of reliability data. University of Surrey, p48.

[91] TAPIA, R.A. and THOMPSON, J.R. (1978). Nonparametric probability density estimation. Johns Hopkins University Press, Baltimore, Maryland.

[92] TARTAR, M.E. and KRONMAL, R. (1967). A description of new computer methods for estimating the population density. Proc. ACM. 22, 511-519.

[93] _____ (1970). On multivariable density estimates based on orthogonal expansions. Ann. Math. Statist., 41, 718-722.

[94] _____ (1976). An introduction to the implementation and theory of nonparametric density estimation. The American Statistician, 30, 3, 105-112.

[95] TEICHER, H. (1961). Idenrifiability of mixtures. Ann. Math. Statist., 32, 244-248.

[96] _____ (1963). Identifiability of finite mixtures. Ann. Math. Statist., 34, 1265-69.

[97] TERRELL, G.R. and SCOTT, D.W. (1985). Oversmoothed nonparametric density estimates. J. Amer. Statist. Assoc. 80, 209-214.

[98]. TITTERINGTON, D.M. (1983). Kernel-based density estimation using censored, truncated or grouped data. Commun. Statist. - Theor. Math., 12(18), 2151-2167.

[99] _____ (1983). Kernel-based density estimates from incomplete data. J. Roy. Statist. Soc. B, 45-62.

[100] VAN RYZIN, J. (1970). On a histogram method of density estimation. Presented at the institute of Mathematical Statistics Meeting, April 8-10, Dallas.

[101] _____ (1973). A histogram method of density estimation. Commun. Statist., 12, 493-506.

[102] WAHBA, G. (1971). A polynomial algorithm for density estimation. Ann. Math. Statist., 42, 1870-1886.

[103] _____ (1975). Optimal convergence properties of variable knot kernel, and orthogonal series methods for density estimation. Ann. Statist. 3,15-29.

[104] _____ (1975). Smoothing noisy data by spline functions. Numer. Math., 24, 383-394.

[105] _____ (1977). Optimal smoothing of density estimates. In classification and clustering, ed., J. Van Ryzin, New York.

[106] _____ (1977). A survey of some smoothing problems and the method of generalized cross-validation for solving them. P.R. Krishnaiah, ed., Applications of Statistics : North-Holand publishing company - Amsterdam-New York.

[107] WALTER, G. and BLUM, J. (1979). Probability density estimation using delta sequences. Ann. Statist., 7, 328-340.

[108] WATSON, G.S. and LEADBETTER, M.R. (1963). On the estimation of the probability density, I. Ann. Math. Statist., 34, 480-491.

[109] WEHMAN, E.J. (1969). A note on estimating a unimodal density. Ann. Math. Statist., 40, 1661-1667.

[110] _____ (1970). Maximum likelihood estimation of a unimodal density function, I. Ann. Math. Statist., 41, 457-471.

[111] _____ (1970). Maximum likelihood estimation of a unimodal density function, II. Ann. Math. Statist., 41, 2169-2174.

[112] _____ (1972). Nonparametric probability density estimation, I. A survey of available methods. Technometrics, vol. 14, No. 3, 513-546.

[113] _____ (1972). Nonparametric probability density estimation, II. A comparison of density estimaion methods. J. Statist. Comput. Simul., 1, 225-245.

[114] WHITTLE, P. (1958). On the smoothing of probability density functions. J. Roy. Statist. Soc., B, 20, 334-343.

[115] WIDDER, D.V. (1946). The Laplace Transform. Princeton Mathematical Series, (ed.) M. Morse, H.P. Robertson, and A.W. Tucker.

[116] WOODROOFE, M. (1969). On choosing a delta sequence. Ann. Math. Statist., 41, 166-171.