PARAMETER ESTIMATION BASED ON EMPIRICAL TRANSFORMS

A THESIS SUBMITTED TO

THE UNIVERSITY OF KENT AT CANTERBURY

IN THE SUBJECT OF STATISTICS

FOR THE DEGREE

OF DOCTOR OF PHILOSOPHY.

By

Panagiotis Besbeas

June 1999

Abstract

In this thesis, we provide a unified treatment of the topic of parameter estimation using integral transforms, such as the characteristic function and moment generating function. This topic encompasses a wealth of methods, which typically vary from each other in relation to the type of weight function and choice of integral transform that is being employed.

We show that the integrated squared error method dominates alternative transform methods, particularly in terms of robustness. We present a convenient and flexible approach to dealing with the difficulty here surrounding the necessary weight function, and illustrate the success of this approach on the mixture of two normal distributions. Furthermore, we show that the integrated squared error method also outperforms the maximum likelihood method for this distribution, particularly with samples with outliers or a small number of observations.

Acknowledgements

I am indebted to Professor Byron J. T. Morgan and Dr. Qiwei Yao for their skillful supervision, and the Institute of Mathematics and Statistics for its co-operative climate and facilities.

I am grateful to the EPSRC for their financial support and to my family to whom I owe a great deal and more.

I would also like to thank Alastair Duncombe for helpful discussions concerning the presentation of this thesis.

Contents

Abstract ii

Acknowledgements iii

1 Estimation based on transforms 1

1.1 The problem 1

1.2 Motivation for transform-based inference 2

1.3 The empirical transform 4

1.4 Minimum distance methods 5

1.5 The empirical characteristic function 7

1.5.1 Graphical investigation 7

1.5.2 Theoretical investigation 9

1.6 The integrated squared error method 11

1.6.1 Regularity conditions 12

1.6.2 Properties of the integrated squared error estimator 13

1.7 Application to the normal distribution 16

1.8 Selecting a value for λ 19

1.8.1 The automatic approach 20

1.8.2 The precision approach 20

1.8.3 The robustness approach 21

1.9 A note on standardising the data 22

1.10 Location and scale invariance of $\hat{\boldsymbol{\theta}}$ 24

1.11 The family of stable laws 25

1.12 The Cauchy distribution 27
1.13 Estimation in the Cauchy distribution 28
1.13.1 Maximum likelihood inference 29
1.13.2 Integrated squared error inference 29
1.14 Integrated distance estimation based on moment generating functions 37

2 Density representation **40**
2.1 Scope of this chapter 40
2.2 Density representation of the ISE function 41
2.3 Kernel density estimation 43
2.4 Properties of the integrated squared error estimator 45
2.5 The mean integrated squared error criterion 48
2.6 The asymptotic $MISE$ approximation 49
2.7 Optimum $MISE$ weight function 50
2.8 Theoretical difficulties 52
2.9 Application to the negative exponential distribution 53
2.10 Optimum weight function theory 56
2.10.1 Practical issues 56
2.10.2 The basic idea 58
2.10.3 A further indication 61
2.11 Choice of value for the parameter λ 63
2.11.1 Theoretical selection of λ 63
2.11.2 Practical selection of λ 64
2.12 Application to the negative exponential mixture 68
2.13 Conclusions 73

3 Mixtures of normal distributions **74**
3.1 Introduction 74
3.2 Finite mixture distributions 75
3.3 Mathematical aspects of mixtures 76
3.3.1 Identifiability 77

3.3.2 Information 77
3.4 Mixtures of two normal distributions 78
3.5 Estimation in the mixture of two normal distributions 80
3.5.1 The method of moments 81
3.5.2 Graphical methods 83
3.5.3 The method of maximum likelihood 83
3.5.4 The Bayesian approach 86
3.6 Robust estimation in the mixture of two normal distributions . . . 87
3.6.1 Minimum distance estimation based on distribution functions 87
3.6.2 Minimum distance estimation based on integral transforms 88
3.7 The method of integrated squared error 90
3.7.1 The integrated squared error estimator 90
3.7.2 The robustness properties of the estimator 94
3.7.3 The asymptotic properties of the estimator 100
3.7.4 Appropriate values for λ 102
3.7.5 The density representation of the ISE function 102
3.7.6 Links with density estimation 104
3.7.7 The $MISE$ and $AMISE$ criteria 106
3.7.8 The smoothed cross-validation selector 109
3.8 Estimation in the mixture of k normal distributions 110
3.8.1 The method of integrated squared error 112
3.9 Concluding remarks 114

4 Sampling experiments **115**
4.1 Introduction 115
4.2 Comparison tools 116
4.3 Asymptotic theory 117
4.4 Simulation details 119
4.5 Random variate generation 121
4.6 Starting values 122

4.7 Optimisation using simulated annealing 123
4.8 Hybrid algorithm 125
4.9 Computational details 127
4.10 Comparison of two selectors for λ 128
4.11 Simulation results 131
4.11.1 Experiment 1 ($p = 0.5, \mu_1 = -2, \sigma_1 = 1, \mu_2 = 2, \sigma_2 = 1$) . . 131
4.11.2 Experiment 2 ($p = 0.5, \mu_1 = -1, \sigma_1 = \sqrt{3}, \mu_2 = 1, \sigma_2 = 1$) . 135
4.11.3 Experiment 3 ($p = 0.5, \mu_1 = 0, \sigma_1 = \sqrt{3}, \mu_2 = 3, \sigma_2 = 1$) . . 140
4.11.4 Experiments 4–7 and evaluation 143
4.12 Simulation results when the true parameter values were used to start the recursions 152

5 Least-squares transform estimation **155**
5.1 Introduction 155
5.2 Motivation for step weight functions 156
5.3 The moment generating function estimator 158
5.3.1 Regularity conditions 160
5.3.2 Properties of the moment generating function estimator . . 160
5.3.3 Selecting a value for $\boldsymbol{t}$ 162
5.4 Application to the normal distribution 163
5.5 Limiting forms of the moment generating function estimator . . . 169
5.6 Insights into the moment generating function method 171
5.7 The preferred moment generating function method 172
5.7.1 Properties of the preferred moment generating function estimator 174
5.8 The modified moment generating function method 177
5.8.1 Properties of the modified moment generating function estimator 178
5.9 The q-L method 181
5.9.1 Regularity conditions 183

5.9.2 Properties of the q-L estimator 184
5.10 Estimation in the Cauchy distribution 186
5.10.1 Estimation using a single point 186
5.10.2 Estimation using $q > 1$ points 192
5.11 Concluding remarks 197

6 Conclusions and future work **199**
6.1 Conclusions 199
6.2 Future work 200
6.2.1 Application to kernel density estimation 200
6.2.2 Multivariate random variables 201
6.2.3 Indexed random variables 205

References **213**

Chapter 1

Estimation based on transforms

1.1 The problem

One of the oldest problems in statistical literature is that of making inferences about a parent population on the basis of information provided by a random sample. The conclusion drawn depends not only on the data, i.e., on what is being observed, but also on background knowledge of the situation. The latter is formalised in the assumptions under which the analysis is undertaken. We shall distinguish between two principal lines of approach:

1. the non-parametric approach;

2. the parametric approach.

In the non-parametric approach, the observations are analysed on their own terms, essentially without extraneous assumptions. The principal aim of the analysis is to describe the data in ways that reveal its underlying structure.

In the parametric approach, however, the observations are presumed to be the values taken on by random variables which follow a distribution function, $F(x)$, belonging to some known family $\mathcal{F}$. Frequently, the distribution is indexed by parameters, $\boldsymbol{\theta}$, taking values in a space, Θ, so that

$$\mathcal{F} = \{F(x;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}. \tag{1.1}$$

The objective of the analysis is then to specify a plausible value for $\boldsymbol{\theta}$ (point estimation), or at least to determine a subspace of Θ of which we can assert that it does, or does not, contain $\boldsymbol{\theta}$ (hypothesis testing).

Although both approaches are equally useful, the parametric approach is more often considered. The aim in writing this chapter is to introduce a parametric method of model-fitting that is based on integral transforms of the distribution function. The work herein focuses on the case of independent and identically distributed (i.i.d.) random variables. The chapter begins with the motivation for transform-based inference, with the necessary theoretical background being covered in Sections 1.3–1.5. The integrated squared error method is then introduced, and subsequently applied to the estimation of the parameters of the normal distribution. With the integrated squared error method there is an accompanying parameter, which is open to choice. Section 1.8 provides several approaches to selecting this parameter, with two arising practical issues considered in the following two sections. We introduce the family of stable laws in Section 1.11, before we pay particular attention on a member of this family, namely the Cauchy distribution. Finally, we consider the real-variable formulation of the integrated squared error method.

1.2 Motivation for transform-based inference

The estimation problem described in the previous section imposes a family of distribution functions, $\mathcal{F}$, and requires determination of an element, $F(x;\hat{\boldsymbol{\theta}})$ in $\mathcal{F}$, which is in some sense optimal. To make this requirement precise, it is necessary to specify a definition of optimality.

There is typically no unique definition of optimality (see, for example, *Lehmann, 1983, p. 2*). An intuitively appealing concept is that of *uniformly minimum variance unbiased* (UMVU) estimation, but this tends to be very restrictive in practice. A theory of much wider applicability is obtained by adopting a large-sample approach. The resulting definition is referred to as asymptotic optimality and, in

view of its relation with maximum likelihood, determination of $F(x;\hat{\boldsymbol{\theta}})$ by this method is preferred in theory as well as in practice.

There are, however, some difficulties with this traditional approach to parametric estimation. First, the optimality of the maximum likelihood method depends rather heavily on the precise nature of the assumed distribution family $\mathcal{F}$. Since the imposition of $\mathcal{F}$ often rests on rather weak assumptions, it becomes important to consider the performance of the maximum likelihood method under departures from the conjectured model (robustness). In general, the lack of robustness of the maximum likelihood method is easily demonstrated.

Secondly, there exist several probability distributions which are only conveniently represented by integral transforms, such as the characteristic function or moment generating function. The absence of closed form expressions for the relevant densities complicates the implementation of the maximum likelihood method. A representative example is the stable laws, the class of limit distributions of sums of independent identically distributed random variables, which are often used to model such noisy processes as common-stock returns.

Thirdly, there is a variety of probability distributions in which integral transforms of the distribution function are of simpler form than the distribution function itself. These distributions often arise as convolutions of independent random variables, and are inconvenient to fit by maximum likelihood. A typical example is the lagged-normal distribution, the convolution of normal and gamma random variables, which is often used to model dilution curves.

One way of overcoming all of these difficulties is to use a method based on integral transforms. This is the approach we have adopted. The resulting method can be employed for parameter estimation in any distribution family. However, it may be argued that using transforms for parameter estimation does not generally result in a simple method.

1.3 The empirical transform

Let $X_1, X_2, \ldots, X_n$ be a set of independent, identically distributed random variables with distribution function

$$F(x; \boldsymbol{\theta}) = \Pr(X_1 \leq x),$$

and suppose that $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is a $p \times 1$ vector of unknown parameters. The empirical distribution function of $X_1, X_2, \ldots, X_n$ is defined by

$$F_n(x) = n^{-1} \sum_{i=1}^{n} I_x(X_i),$$

where $I_x(X)$ is the function

$$I_x(X) = \begin{cases} 1 & \text{if } X \leq x, \\ 0 & \text{otherwise.} \end{cases}$$

Many standard methods of statistical inference rely on the empirical distribution function, and the integral transform method is no exception.

The method is based on a possibly complex-valued function $g(t, x)$ such that

$$G(t; \boldsymbol{\theta}) = \int_{-\infty}^{\infty} g(t, x) \, \mathrm{d}F(x; \boldsymbol{\theta}) \tag{1.2}$$

exists and is finite for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $t \in T \subseteq \mathbb{R}$. An empirical version of this transform may be defined as

$$G_n(t) = \int_{-\infty}^{\infty} g(t, x) \, \mathrm{d}F_n(x) = n^{-1} \sum_{j=1}^{n} g(t, X_j).$$

It turns out that the transform functions $g(t, x)$ that are potentially of interest are numerous. *Feuerverger and McDunnough (1984)* indicate that the following choices are typical:

1. Take $g(t,x) = \exp(tx)$. Then $G(t;\boldsymbol{\theta})$ and $G_n(t)$ coincide with the moment generating function (mgf) $M(t;\boldsymbol{\theta}) = \int \exp(tx)\,\mathrm{d}F(x;\boldsymbol{\theta})$, when it exists, and the empirical mgf $M_n(t) = n^{-1}\sum_{j=1}^{n}\exp(tX_j)$ respectively.

2. Take $g(t,x) = \exp(itx)$. Now $G(t;\boldsymbol{\theta})$ and $G_n(t)$ coincide with the characteristic function (cf) $\phi(t;\boldsymbol{\theta}) = \int \exp(itx)\,\mathrm{d}F(x;\boldsymbol{\theta})$ and the empirical cf $\phi_n(t) = n^{-1}\sum_{j=1}^{n}\exp(itX_j)$ respectively.

Focusing on the empirical transform $G_n(t)$, it follows from the Strong Law of Large Numbers that, for any fixed $t \in T$,

$$G_n(t) \to G(t;\boldsymbol{\theta})$$

almost surely as $n \to \infty$. Furthermore, by the Central Limit Theorem, the stochastic process $\{n^{1/2}[G_n(t) - G(t;\boldsymbol{\theta})] : t \in T\}$ will, asymptotically, converge to a Gaussian process. These properties suggest a variety of inferential procedures for study.

1.4 Minimum distance methods

Minimum distance estimation was first subjected to comprehensive study in a series of papers culminating in *Wolfowitz (1957)*, and has since been considered as a method for deriving robust estimators. The basic philosophy of minimum distance estimation is to match the empirical distribution function $F_n(x)$ to an element, $F(x;\hat{\boldsymbol{\theta}})$, of the family $\{F(x;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ as closely as possible.

This methodology suggests a broadening of the class of estimators. If we define the random functions

$$\Delta[G_n(\,.\,), G(\,.\,;\boldsymbol{\theta})], \tag{1.3}$$

which are distance functions measuring the discrepancy between the transforms $G_n(t)$ and $G(t;\boldsymbol{\theta})$ then, for a suitably chosen distance function Δ, a *minimum*

distance estimator, $\hat{\boldsymbol{\theta}}$, for $\boldsymbol{\theta}$ is a value of $\boldsymbol{\theta}$ which satisfies

$$\Delta[G_n(\,.\,), G(\,.\,;\hat{\boldsymbol{\theta}})] = \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \Delta[G_n(\,.\,), G(\,.\,;\boldsymbol{\theta})]. \tag{1.4}$$

It is clear at the outset that some distance functions will produce estimates with better properties than others. In particular, a distance function measuring a "supremum-type" discrepancy, such as the Kolmogorov discrepancy

$$\Delta[G_n(\,.\,), G(\,.\,;\boldsymbol{\theta})] = \sup_{t\in T} |G_n(t) - G(t;\boldsymbol{\theta})|,$$

would be an unwise choice, for then the corresponding asymptotic theory is typically not normal (see, for example, *Parr and Schucany, 1982*). A more propitious choice for Δ would measure an "integral-type" discrepancy. The function leading to the greatest degree of mathematical tractability is

$$\Delta[G_n(\,.\,), G(\,.\,;\boldsymbol{\theta})] = \int_T |G_n(t) - G(t;\boldsymbol{\theta})|^2 \,\mathrm{d}W(t), \tag{1.5}$$

which we shall use throughout this work.

The function $W(t)$ is referred to as the weight function and is open to choice. The weight functions that have been adopted in the literature can be divided into two basic types:

1. monotonic non-decreasing step functions;

2. monotonic non-decreasing continuous functions.

The use of the former type of weight function appears to have been initiated by *Press (1972)*, while the latter was introduced by *Paulson, Holcomb and Leitch (1975)*.

The performances of the resulting estimators in terms of (i) efficiency, (ii) robustness, and (iii) computational feasibility will be examined in this thesis. We shall start with a method which is known as the *integrated squared error* method. This is a particularly important member of the general class of integrated distance

methods, so we shall pay special attention to it. In this method, the weight function is a continuous function and the choice of transform is the characteristic function. First then, we shall review some properties of the empirical characteristic function.

1.5 The empirical characteristic function

Let $X_1, X_2, \ldots, X_n$ be a set of independent, identically distributed random variables with distribution function $F(x; \boldsymbol{\theta})$, and suppose that

$$\phi(t; \boldsymbol{\theta}) = \int_{-\infty}^{\infty} e^{itx} \, \mathrm{d}F(x; \boldsymbol{\theta}) = U(t; \boldsymbol{\theta}) + i\, V(t; \boldsymbol{\theta}) \tag{1.6}$$

is the characteristic function corresponding to $F(x; \boldsymbol{\theta})$. As noted in Section 1.3, the characteristic function can be estimated by the empirical characteristic function

$$\phi_n(t) = n^{-1} \sum_{j=1}^{n} e^{itX_j} = U_n(t) + i\, V_n(t). \tag{1.7}$$

The empirical characteristic function seems to appear for the first time in *Cramér (1946, p. 342)* where it plays only an auxiliary role. It appears for the second time in *Parzen (1962)* in the context of non-parametric density estimation. In its own right, the empirical characteristic function entered into statistical literature in *Press (1972)*. He realised that the estimation of the parameters of stable distributions can be based on $\phi_n(t)$. Since then, the empirical characteristic function has been applied to a very wide range of problems, as cited by *Epps (1993)*.

1.5.1 Graphical investigation

A plot of the empirical characteristic function, as a function of t, can be helpful for indicating how adequately one might expect it to estimate the theoretical characteristic function. For example, if $X_1, X_2, \ldots, X_n$ are independent, identically

distributed random variables with characteristic function

$$\phi(t;\boldsymbol{\theta}) = \exp(i\,\theta_1 t - |\theta_2 t|^{\theta_3}), \tag{1.8}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)^\top$ belongs to the parameter space

$$\Theta = \{|\theta_1| < \infty,\ \theta_2 > 0,\ 0 < \theta_3 \leq 2\},$$

then it may be useful to replace the left hand side of (1.8) by $\phi_n(t)$ and examine the resulting plots.

Figure 1.1 shows four realisations of $U_n(t)$ based on simulated samples from a population whose characteristic function is given by (1.8) with $\theta_1 = 5$, $\theta_2 = 2$, together with

1. $\theta_3 = 1$ and $n = 20$;
2. $\theta_3 = 1$ and $n = 50$;
3. $\theta_3 = 2$ and $n = 20$;
4. $\theta_3 = 2$ and $n = 50$.

The values $\theta_3 = 1$ and $\theta_3 = 2$, which give rise to the Cauchy and normal distributions, respectively, were chosen to make the necessary simulation procedures as simple as possible. Nevertheless, the results should hold for all $\boldsymbol{\theta}$. To attach insight to these graphs, the corresponding $U(t;\boldsymbol{\theta})$ is also included.

As anticipated, there is a good overall agreement between $U_n(t)$ and $U(t;\boldsymbol{\theta})$. It is clear that increasing n increases the agreement between the two transforms, and this is as expected. On the other hand, less expectedly, $U_n(t)$ is more accurate for small than for moderate t. Another important question which is raised but cannot be answered from the figure, is how the variability present in $U_n(t)$ varies with $\boldsymbol{\theta}$.

A similar pattern emerged for $V_n(t)$ so we have omitted the corresponding plots. It appears, therefore, that the empirical characteristic function may have

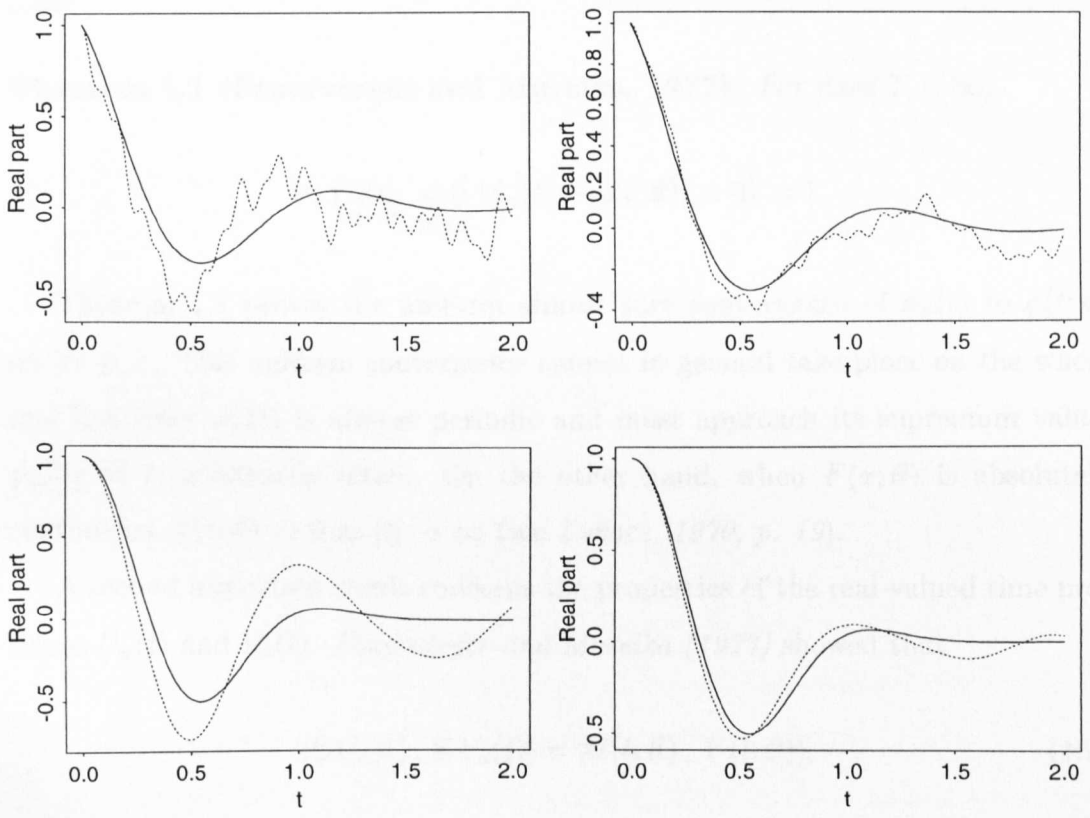

Figure 1.1: The real part of the characteristic function (1.8) (solid line) overlaid with the real part of the empirical characteristic function (dotted line) in four different contexts. The contexts are (reading from left-right, top-bottom): (1) $\theta_3 = 1, n = 20$; (2) $\theta_3 = 1, n = 50$; (3) $\theta_3 = 2, n = 20$; and (4) $\theta_3 = 2, n = 50$. In each case, $\theta_1 = 5$ and $\theta_2 = 2$.

good properties as an estimator of the characteristic function (1.8), and is likely to retain these properties in different settings. We investigate this issue in detail in the next section.

1.5.2 Theoretical investigation

The asymptotic properties of the empirical characteristic function have been thoroughly investigated by *Feuerverger and Mureika (1977)* and *Csörgő (1981)*.

For any fixed t, $\phi_n(t)$ is an average of bounded independent identically distributed random variables having mean $\phi(t; \boldsymbol{\theta})$. It follows from the Strong Law of Large Numbers that $\phi_n(t)$ converges almost surely to $\phi(t; \boldsymbol{\theta})$. Furthermore, we

have:

Theorem 1.1 (Feuerverger and Mureika, 1977). *For fixed $T < \infty$,*

$$\Pr\Big(\lim_{n\to\infty} \sup_{|t|\le T} |\phi_n(t) - \phi(t;\boldsymbol{\theta})| = 0\Big) = 1.$$

Theorem 1.1 proves the uniform almost sure convergence of $\phi_n(t)$ to $\phi(t;\boldsymbol{\theta})$ on $|t| \le T$. This uniform convergence cannot in general take place on the whole real line since $\phi_n(t)$ is almost periodic and must approach its supremum value, $\phi_n(0) = 1$, arbitrarily often. On the other hand, when $F(x;\boldsymbol{\theta})$ is absolutely continuous, $\phi(t;\boldsymbol{\theta}) \to 0$ as $|t| \to \infty$ (see *Lukacs, 1970, p. 19*).

A second important result concerns the properties of the real-valued time processes $U_n(t)$ and $V_n(t)$. *Feuerverger and Mureika (1977)* showed that

$$[\mathrm{E}\,U_n(t)\,,\,\mathrm{E}\,V_n(t)] = [U(t;\boldsymbol{\theta})\,,\,V(t;\boldsymbol{\theta})], \tag{1.9}$$

while

$$\left.\begin{aligned}
2n\,\mathrm{Cov}[U_n(t_1), U_n(t_2)] &= U(t_1+t_2;\boldsymbol{\theta}) + U(t_1-t_2;\boldsymbol{\theta}) - 2U(t_1;\boldsymbol{\theta})U(t_2;\boldsymbol{\theta}) \\
2n\,\mathrm{Cov}[U_n(t_1), V_n(t_2)] &= V(t_1+t_2;\boldsymbol{\theta}) - V(t_1-t_2;\boldsymbol{\theta}) - 2U(t_1;\boldsymbol{\theta})V(t_2;\boldsymbol{\theta}) \\
2n\,\mathrm{Cov}[V_n(t_1), V_n(t_2)] &= U(t_1-t_2;\boldsymbol{\theta}) - U(t_1+t_2;\boldsymbol{\theta}) - 2V(t_1;\boldsymbol{\theta})V(t_2;\boldsymbol{\theta}).
\end{aligned}\right\} \tag{1.10}$$

The behaviour of $U_n(t)$ and $V_n(t)$ has been investigated by *Koutrouvelis (1980)*. He showed, from equations (1.9)–(1.10), that as $|t| \to \infty$

$$[\mathrm{E}\,U_n(t)\,,\,\mathrm{E}\,V_n(t)] \to (0\,,\,0) \quad \text{and} \quad [\mathrm{Var}\,U_n(t)\,,\,\mathrm{Var}\,V_n(t)] \to (\tfrac{1}{2n}\,,\,\tfrac{1}{2n}),$$

while as $|t| \to 0$

$$[\mathrm{E}\,U_n(t)\,,\,\mathrm{E}\,V_n(t)] \to (1\,,\,0) \quad \text{and} \quad [\mathrm{Var}\,U_n(t)\,,\,\mathrm{Var}\,V_n(t)] \to (0\,,\,0).$$

These relations indicate that the tails of $U_n(t)$ and $V_n(t)$ are completely "noise"

around zero means, while $[U_n(t)\,,\,V_n(t)]$ estimates $[U(t;\boldsymbol{\theta})\,,\,V(t;\boldsymbol{\theta})]$ with increasing accuracy as $t \to 0$.

The convergence properties of the processes $U_n(t)$ and $V_n(t)$ have been studied through the empirical characteristic process

$$Y_n(t;\boldsymbol{\theta}) = n^{1/2}[\phi_n(t) - \phi(t;\boldsymbol{\theta})]. \tag{1.11}$$

Feuerverger and Mureika (1977) showed that

1. $\mathrm{E}[Y_n(t;\boldsymbol{\theta})] = 0$,
2. $\mathrm{Cov}[Y_n(t_1;\boldsymbol{\theta}), Y_n(t_2;\boldsymbol{\theta})] = \phi(t_1 + t_2;\boldsymbol{\theta}) - \phi(t_1;\boldsymbol{\theta})\,\phi(t_2;\boldsymbol{\theta})$,

and proved that $Y_n(t)$ converges (both weakly and in distribution) to a complex-valued Gaussian process having zero mean and the same covariance structure as $Y_n(t)$. The weak convergence of $Y_n(t)$ is also treated by *Csörgő (1981)*.

The results of this section demonstrate that the empirical characteristic function has good properties as an estimator of the theoretical characteristic function, and the remainder of this chapter is concerned with how we might exploit these.

1.6 The integrated squared error method

Let $W(t)$ be a monotonically non-decreasing continuous function, and consider the estimator $\hat{\boldsymbol{\theta}}$ which minimises

$$\Delta[\phi_n(\,.\,), \phi(\,.\,;\boldsymbol{\theta})] = \int_{-\infty}^{\infty} |\phi_n(t) - \phi(t;\boldsymbol{\theta})|^2 \,\mathrm{d}W(t) \tag{1.12}$$

with respect to $\boldsymbol{\theta}$. It appears that *Press (1972)* was the first to propose this estimator to estimate the parameters in a stable distribution. It was first studied, again, as applied to stable distributions, by *Paulson, Holcomb and Leitch (1975)*, with the specific weight function $W(t) = \int_{-\infty}^{t} e^{-y^2}\mathrm{d}y$.

In general, the weight function will be open to choice. The simplest choice is, of course, $W(t) = t$, but this may only be used if $F(x;\boldsymbol{\theta})$ is discrete. In fact,

the role of the weight function can be related to the geometrical behaviour of the empirical characteristic function. We have seen that the characteristic function of an absolutely continuous distribution approaches zero as $|t| \to \infty$, whereas the empirical characteristic function is periodic. One purpose of $W(t)$ is, therefore, to assure convergence of the integral in (1.12). Another is to give the discrepancies $|\phi_n(t) - \phi(t;\boldsymbol{\theta})|^2$ high influence at values of t where $\phi_n(t)$ has high precision. Regardless of purpose, we shall make the weight function depend on a parameter. There are definite advantages to such a course of action, which will be discussed in a later section. For our purposes then, $\hat{\boldsymbol{\theta}}$ will be defined as follows:

Definition 1.1. The integrated squared error (ISE) estimator, $\hat{\boldsymbol{\theta}}$, for $\boldsymbol{\theta}$ is the value of $\boldsymbol{\theta}$ which minimises

$$I(\boldsymbol{\theta};\lambda) = \int_{-\infty}^{\infty} |\phi_n(t) - \phi(t;\boldsymbol{\theta})|^2 \, \mathrm{d}W(t;\lambda), \tag{1.13}$$

where $W(t;\lambda)$ is a suitably selected weight function depending on a positive parameter λ.

The asymptotic properties of this estimator were investigated by *Thornton and Paulson (1977)*. The same was done independently, and under much less restrictive conditions by *Heathcote (1977)*. He called $\hat{\boldsymbol{\theta}}$ the integrated squared error estimator, and in the present note we shall rely upon Heathcote's regularity conditions for the consistency and asymptotic normality of $\hat{\boldsymbol{\theta}}$.

1.6.1 Regularity conditions

Consider the family of characteristic functions

$$\mathcal{F} = \{\phi(t;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p\}, \quad p \geq 1,$$

where $\phi(t;\boldsymbol{\theta}) = U(t;\boldsymbol{\theta}) + iV(t;\boldsymbol{\theta})$. The following conditions constitute what *Heathcote (1977)* calls the regular case.

1. For $i = 1, 2, \ldots, p$, $(\partial/\partial\theta_i)U(t;\boldsymbol{\theta})$ and $(\partial/\partial\theta_i)V(t;\boldsymbol{\theta})$ exist for all $t \in \mathbb{R}$, and are uniformly bounded by functions that are integrable with respect to $W(t;\lambda)$.

2. For $i, j = 1, 2, \ldots, p$, $(\partial^2/\partial\theta_i\partial\theta_j)U(t;\boldsymbol{\theta})$ and $(\partial^2/\partial\theta_i\partial\theta_j)V(t;\boldsymbol{\theta})$ exist for all $t \in \mathbb{R}$, and are uniformly bounded by functions that are integrable with respect to $W(t;\lambda)$.

3. For $i, j = 1, 2, \ldots, p$, $U(t;\boldsymbol{\theta})$, $V(t;\boldsymbol{\theta})$, $(\partial/\partial\theta_i)U(t;\boldsymbol{\theta})$, $(\partial/\partial\theta_i)V(t;\boldsymbol{\theta})$, $(\partial^2/\partial\theta_i\partial\theta_j)U(t;\boldsymbol{\theta})$, and $(\partial^2/\partial\theta_i\partial\theta_j)V(t;\boldsymbol{\theta})$ are all jointly continuous in $t \in \mathbb{R}$ and $\boldsymbol{\theta} \in \Theta_0$, where Θ_0 is the closure of some neighbourhood of the true parameter value.

It should be pointed out that these conditions are not very stringent on $\phi(t;\boldsymbol{\theta})$. Furthermore, the first two can be regarded as (mild) restrictions on the choice of the weight function $W(t;\lambda)$.

1.6.2 Properties of the integrated squared error estimator

Let $\boldsymbol{\theta}_0 \in \Theta$ be the true parameter value we seek to estimate. We have:

Theorem 1.2 (Heathcote, 1977). *Under the three above regularity conditions, the integrated squared error estimator, $\hat{\boldsymbol{\theta}}$, for $\boldsymbol{\theta}_0 \in \Theta \subseteq \mathbb{R}^p$ is strongly consistent and*

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \Sigma(\boldsymbol{\theta}_0)],$$

where

$$\Sigma(\boldsymbol{\theta}) = K^{-1}(\boldsymbol{\theta})\,\Omega(\boldsymbol{\theta})\,K^{-1}(\boldsymbol{\theta}). \tag{1.14}$$

In this expression, $K(\boldsymbol{\theta})$ is the $p \times p$ symmetric matrix whose (i,j)th element is

$$\kappa_{ij}(\boldsymbol{\theta}) = \int_{-\infty}^{\infty} \left[\frac{\partial U(t;\boldsymbol{\theta})}{\partial\theta_i}\,\frac{\partial U(t;\boldsymbol{\theta})}{\partial\theta_j} + \frac{\partial V(t;\boldsymbol{\theta})}{\partial\theta_i}\,\frac{\partial V(t;\boldsymbol{\theta})}{\partial\theta_j}\right] \mathrm{d}W(t;\lambda),$$

and $\Omega(\boldsymbol{\theta})$ *is the* $p \times p$ *covariance matrix of the random variables*

$$\tau_i(X_1;\boldsymbol{\theta}) = \int_{-\infty}^{\infty} \{[\cos(tX_1) - U(t;\boldsymbol{\theta})]\frac{\partial U(t;\boldsymbol{\theta})}{\partial \theta_i} + [\sin(tX_1) - V(t;\boldsymbol{\theta})]\frac{\partial V(t;\boldsymbol{\theta})}{\partial \theta_i}\}\, \mathrm{d}W(t;\lambda), \quad i = 1, 2, \ldots, p.$$

In Theorem 1.2, $\xrightarrow{d}$ denotes convergence in distribution and $\mathcal{N}(\mathbf{0}, \Sigma)$ is the p-variate normal distribution with mean vector zero and covariance matrix Σ.

While the precision of an estimator is a very important aspect to consider, it is not the only one. We may also consider how contamination influences a given estimator. For example, how is the effect on the estimator related to the magnitude of the contaminants? What is the worst possible effect that a single contaminant can have? Is this effect bounded or not? Aspects such as these require a powerful array of tools based on the *influence function.* The approach is due to *Hampel (1971)*; see also *Hampel (1974).*

The viewpoint of the influence function is very easily described. Given a random sample $X_1, X_2, \ldots, X_n$ from a distribution $F(x;\boldsymbol{\theta})$, we seek to measure the effect of adding an observation ξ to the sample on the given estimator for $\boldsymbol{\theta}$. In the context of integrated squared error estimation, the influence function was first subjected to study by *Paulson and Nicklin (1983)* and *Bryant and Paulson (1983).* More recently, *Campbell (1993)* contributed the following theorem:

Theorem 1.3 (Campbell, 1993). *Under regularity conditions (1), (2) and (3), the integrated squared error estimator,* $\hat{\boldsymbol{\theta}}$*, for* $\boldsymbol{\theta}_0 \in \Theta \subseteq \mathbb{R}^p$ *has joint influence function*

$$IF(\xi;\hat{\boldsymbol{\theta}}) = K^{-1}(\boldsymbol{\theta}_0)\, \boldsymbol{\tau}(\xi;\boldsymbol{\theta}_0), \tag{1.15}$$

where $K(\boldsymbol{\theta})$ *is the* $p \times p$ *symmetric matrix whose* (i,j)*th element is*

$$\kappa_{ij}(\boldsymbol{\theta}) = \int_{-\infty}^{\infty} [\frac{\partial U(t;\boldsymbol{\theta})}{\partial \theta_i}\frac{\partial U(t;\boldsymbol{\theta})}{\partial \theta_j} + \frac{\partial V(t;\boldsymbol{\theta})}{\partial \theta_i}\frac{\partial V(t;\boldsymbol{\theta})}{\partial \theta_j}]\, \mathrm{d}W(t;\lambda),$$

and $\boldsymbol{\tau}(\xi;\boldsymbol{\theta})$ *is the* $p \times 1$ *vector whose ith element is*

$$\tau_i(\xi;\boldsymbol{\theta}) = \int_{-\infty}^{\infty} \{IF[\xi;U_n(t)]\frac{\partial U(t;\boldsymbol{\theta})}{\partial\theta_i} + IF[\xi;V_n(t)]\frac{\partial V(t;\boldsymbol{\theta})}{\partial\theta_i}\}\,\mathrm{d}W(t;\lambda),$$

where

$$IF[\xi;U_n(t)] = \cos(t\xi) - U(t;\boldsymbol{\theta})$$
$$IF[\xi;V_n(t)] = \sin(t\xi) - V(t;\boldsymbol{\theta}).$$

The influence function is, of course, only one measure of robustness. However, we need only emphasise the influence function here because other measures of robustness such as gross-error sensitivity, local-shift sensitivity, rejection point, and breakdown point (see, for example, *Barnett and Lewis, 1994, p. 72*) can be obtained once influence functions have been provided. In addition, an important property of the influence function (see, for example, *Huber, 1981, p. 14*) is that if we regard the argument ξ as a random quantity distributed according to the underlying model, then its expectation with respect to ξ,

$$\int IF(\xi;\hat{\boldsymbol{\theta}})\,\mathrm{d}F(\xi;\boldsymbol{\theta}),$$

is zero, while its mean squared error,

$$\int IF(\xi;\hat{\boldsymbol{\theta}})\,IF(\xi;\hat{\boldsymbol{\theta}})^{\mathsf{T}}\,\mathrm{d}F(\xi;\boldsymbol{\theta}), \tag{1.16}$$

is equal to the asymptotic variance of $\hat{\boldsymbol{\theta}}$. Thus we have a direct connection between Theorems 1.2 and 1.3.

Since the statistical properties of the integrated squared error estimator depend on the weight function $W(t;\lambda)$, it is in principle possible to define an optimal $W(t;\lambda)$. However, the complicated dependence of either efficiency or robustness on $W(t;\lambda)$ suggests that the optimal $W(t;\lambda)$, in either sense, will be difficult to obtain. Consequently, present practice regarding the choice of $W(t;\lambda)$ is a combination of common-sense and convenience. As we shall see in Chapter 2, a

frequency-domain analysis of (1.13) can yield a great amount of insight into the choice of the weight function. Meantime, we shall illustrate how the integrated squared error method may be used in practice.

1.7 Application to the normal distribution

Let $X_1, X_2, \ldots, X_n$ be independent, identically distributed random variables from a normal distribution with mean μ and variance σ^2, and suppose that $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$ is unknown. Since the characteristic function of X_1 is

$$\phi(t; \boldsymbol{\theta}) = \exp(it\mu - \sigma^2 t^2/2),$$

the integrated squared error estimator, $\hat{\boldsymbol{\theta}}$, for $\boldsymbol{\theta}$ is the value of $\boldsymbol{\theta}$ which minimises

$$I(\boldsymbol{\theta}; \lambda) = \int_{-\infty}^{\infty} |n^{-1} \sum_{j=1}^{n} e^{itX_j} - e^{it\mu - \sigma^2 t^2/2}|^2 \, \mathrm{d}W(t; \lambda) \tag{1.17}$$

for some non-decreasing weight function $W(t; \lambda)$. In practice, one might not contemplate using integral transform-based inference to estimate $\boldsymbol{\theta}$, since maximum likelihood can be readily implemented. Nevertheless, the integrated squared error estimator was considered by *Thornton and Paulson (1977)*, with the specific weight function $W(t) = \int_{-\infty}^{t} e^{-y^2} \mathrm{d}y$. This weight function can be regarded as a special case of $W(t; \lambda) = \int_{-\infty}^{t} e^{-\lambda^2 y^2} \mathrm{d}y$, which results in

$$I(\boldsymbol{\theta}; \lambda) = \frac{\pi^{1/2}}{\lambda n^2} \sum_{j=1}^{n} \sum_{k=1}^{n} \exp[-\frac{1}{4} \frac{(X_j - X_k)^2}{\lambda^2}] - \frac{2}{n} (\frac{\pi}{\lambda^2 + \frac{1}{2}\sigma^2})^{1/2} \sum_{j=1}^{n} \exp[-\frac{1}{4} \frac{(X_j - \mu)^2}{\lambda^2 + \frac{1}{2}\sigma^2}] + (\frac{\pi}{\lambda^2 + \sigma^2})^{1/2}. \tag{1.18}$$

It is reasonable to expect that the integrated squared error estimator will be less efficient but more robust than the maximum likelihood estimator. We find, from Theorem 1.2, that $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is asymptotically normally distributed with

mean vector zero and covariance matrix $\Sigma(\boldsymbol{\theta})$, where the elements of $\Sigma(\boldsymbol{\theta})$ are

$$\sigma_{11}(\boldsymbol{\theta}) = \frac{\sigma^2(\lambda^2+\sigma^2)^3}{(\lambda^4+2\lambda^2\sigma^2+\frac{3}{4}\sigma^4)^{3/2}}$$

$$\sigma_{12}(\boldsymbol{\theta}) = 0$$

$$\sigma_{22}(\boldsymbol{\theta}) = \frac{16}{9}\frac{(\lambda^2+\sigma^2)^5(\lambda^4+2\lambda^2\sigma^2+\frac{3}{2}\sigma^4)}{(\lambda^4+2\lambda^2\sigma^2+\frac{3}{4}\sigma^4)^{5/2}} - \frac{16}{9}(\lambda^2+\sigma^2)^2.$$

The Fisher information matrix for a single observation is given by

$$\mathcal{I}(\boldsymbol{\theta}) = \begin{pmatrix} 1/\sigma^2 & 0 \\ & 1/2\sigma^4 \end{pmatrix} \tag{1.19}$$

so that, for example, the asymptotic efficiency of $\hat{\mu}$ is

$$eff(\hat{\mu}) = (\lambda^4+2\lambda^2\sigma^2+\tfrac{3}{4}\sigma^4)^{3/2}/(\lambda^2+\sigma^2)^3. \tag{1.20}$$

Figure 1.2 shows that the efficiency of $\hat{\mu}$ varies from 0.65 to 1.00 as λ and σ vary from 0 to 3 and 0 to 4 respectively. Clearly, the integrated squared error estimator can perform adequately in these terms.

The robustness properties of $\hat{\boldsymbol{\theta}}$ may be examined through the behaviour of its influence function. If we define

$$g(x) = \exp[-\frac{1}{4}\frac{(x-\mu)^2}{(\lambda^2+\frac{1}{2}\sigma^2)}], \quad x \in \mathbb{R},$$

then, from Theorem 1.3, we obtain

$$IF(\xi;\hat{\boldsymbol{\theta}}) = \begin{pmatrix} [\frac{(\lambda^2+\sigma^2)}{(\lambda^2+\sigma^2/2)}]^{3/2}(\xi-\mu)g(\xi) \\ \frac{4}{3}(\lambda^2+\sigma^2) + \frac{2}{3}[\frac{(\lambda^2+\sigma^2)}{(\lambda^2+\sigma^2/2)}]^{5/2}[(\xi-\mu)^2-(2\lambda^2+\sigma^2)]g(\xi) \end{pmatrix}. \tag{1.21}$$

It is interesting to evaluate this influence function at the standard normal distribution, as the standard normal is a familiar point of reference. Figure 1.3 gives the individual influence functions for $\hat{\boldsymbol{\theta}}$ with $\lambda = 1, 2, 3$. As observed in

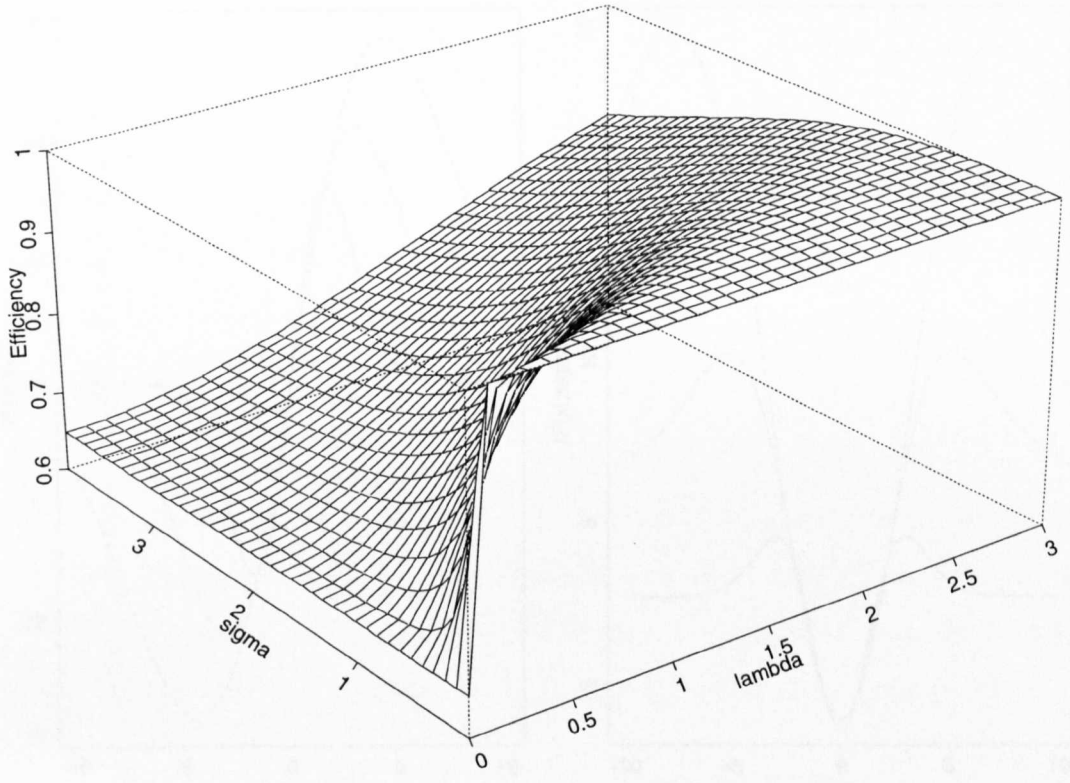

Figure 1.2: Perspective view of the asymptotic efficiency of $\hat{\mu}$. Note that $eff(\hat{\mu})$ is independent of μ.

the figure, the individual influence functions are bounded in ξ, and decline as $|\xi| \to \infty$. This implies that outlying observations have little effect on $\hat{\boldsymbol{\theta}}$. Such estimators are said to perform robustly in the presence of outliers.

In contrast, the joint influence function for the maximum likelihood estimator, $\tilde{\boldsymbol{\theta}}$, may be shown to be

$$IF(\xi; \tilde{\boldsymbol{\theta}}) = \begin{pmatrix} \xi - \mu \\ (\xi - \mu)^2 - \sigma^2 \end{pmatrix}.$$

The individual influence functions for $\tilde{\boldsymbol{\theta}}$ are unbounded in ξ, so that a single additional observation can completely change the value of the parameter estimates.

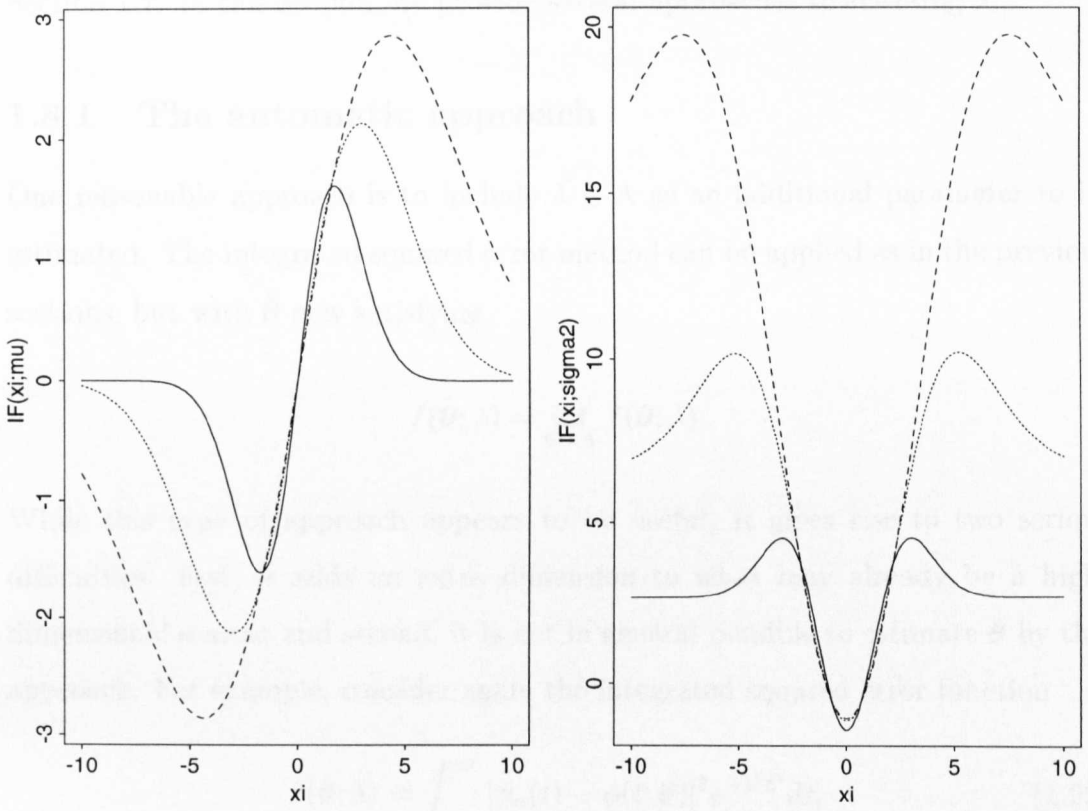

Figure 1.3: Individual influence functions for the integrated squared error estimator $\hat{\boldsymbol{\theta}} = (\hat{\mu}, \hat{\sigma}^2)^\top$ with $\lambda = 1$ (solid line), $\lambda = 2$ (dotted line), and $\lambda = 3$ (dashed line). In each case the influence functions are evaluated at the standard normal distribution.

Thus, the maximum likelihood estimator does not perform robustly in the presence of outliers.

In summary, the integrated squared error method can provide a flexible estimator for the parameters of a normal distribution. This is because the resulting estimator can attain arbitrarily high efficiency on the one hand, but can also be made increasingly robust on the other. However, the statistical properties of this estimator depend on λ, and we now investigate its choice.

1.8 Selecting a value for λ

Practical implementation of the integrated squared error estimator requires a value for the parameter λ. This choice is very important, as was shown graphically in

Section 1.7. In this section, we provide several approaches to selecting λ.

1.8.1 The automatic approach

One reasonable approach is to include $\lambda \in \Lambda$ as an additional parameter to be estimated. The integrated squared error method can be applied as in the previous sections, but with $\hat{\boldsymbol{\theta}}$ now satisfying

$$I(\hat{\boldsymbol{\theta}}; \hat{\lambda}) = \inf_{\boldsymbol{\Theta} \times \Lambda} I(\boldsymbol{\theta}; \lambda).$$

While this type of approach appears to be useful, it gives rise to two serious difficulties: first, it adds an extra dimension to what may already be a high-dimensional search; and second, it is not in general possible to estimate $\boldsymbol{\theta}$ by this approach. For example, consider again the integrated squared error function

$$I(\boldsymbol{\theta}; \lambda) = \int_{-\infty}^{\infty} |\phi_n(t) - \phi(t; \boldsymbol{\theta})|^2 e^{-\lambda^2 t^2} \mathrm{d}t, \tag{1.22}$$

which was applied in Section 1.7 to the normal distribution. Since

$$|\phi_n(t) - \phi(t; \boldsymbol{\theta})| \leq 2$$

by the triangle inequality, it follows that

$$I(\boldsymbol{\theta}; \lambda) \leq \int_{-\infty}^{\infty} 4e^{-\lambda^2 t^2} \mathrm{d}t = \frac{4\pi^{1/2}}{\lambda}$$

tends to its infimum (zero) as λ tends to infinity. Due to possible non-uniqueness of the value $\hat{\boldsymbol{\theta}}$ achieving this infimum, external criteria for selecting λ must be considered.

1.8.2 The precision approach

The problems entailed in the automatic approach may be avoided if λ is selected by considering the mean squared error of the estimator. In particular, if $\hat{\boldsymbol{\theta}}$ is to

estimate $\boldsymbol{\theta}$ using a sample of size n, one possibility would be to choose λ so as to minimise the determinant of the mean squared error

$$\mathrm{E}[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^{\top}]. \tag{1.23}$$

More often the calculation of (1.23), other than by simulation, is difficult or intractable. A criterion of much wider applicability is obtained by adopting a large-sample approach. In this case, (1.23) would be replaced by its asymptotic version $\Sigma(\boldsymbol{\theta})$, as stated in (1.14).

This approach may be criticised in two ways: first, it is based on asymptotic variances whereas finite-sample variances are the appropriate ones to use; and second, the optimum λ will depend on the unknown parameter values. Nevertheless, the precision approach has been exercised, for example, by *Bryant and Paulson (1983)* and *Paulson and Nicklin (1983)*.

1.8.3 The robustness approach

Alternatively, we can select λ by considering the robustness properties of the estimator. In this case, the value of λ leading to the best possible performance in relation to some measure of robustness would be selected. There are several measures of robustness to consider; one of those mentioned in Section 1.6 was gross-error sensitivity. This is defined for an estimator $\hat{\theta}$ as the supremum of the absolute value of its influence function,

$$\gamma(\hat{\theta}) = \sup_{\xi} |IF(\xi; \hat{\theta})|. \tag{1.24}$$

In words, the gross error sensitivity measures the worst possible effect a fixed amount of contamination can have on the estimator. If it is used as measure of robustness in the robustness approach, then one would select λ so as the effect of contamination is made as small as possible.

This approach may be criticised in the following ways: first, the measure of

robustness is open to choice; second, the optimum λ will depend on the unknown parameter values; and third, it is possible to obtain different optimal values of λ for different components of $\hat{\boldsymbol{\theta}}$.

In summary, we can regard the integrated squared error method as both an efficient as well as a robust method. As an efficient method, it can be made increasingly efficient by selecting λ by the precision approach. As a robust method, it becomes increasingly robust as λ is selected by the robustness approach. This feature is not enjoyed by other estimation methods such as, for example, maximum likelihood or method of moments.

1.9 A note on standardising the data

Let X be a random variable taking on values in a sample space $\mathcal{X}$ according to a distribution $F(x;\boldsymbol{\theta})$, which is known to belong to a family $\mathcal{F}$. In this section, we shall consider one-to-one transformations g of the sample space onto itself so that, for each $\boldsymbol{\theta}$, the distribution of $Z = g(X)$, denoted $F(z;\boldsymbol{\theta}')$, is again a member of $\mathcal{F}$.

The study of this type of transformations has mainly been motivated by numerical considerations. For example, some numerical difficulties can be eliminated by *standardising* the data prior to estimation. This is effected by transforming the realisations $x_1, x_2, \ldots, x_n$ of X to

$$z_j = (x_j - \mu)/\sigma, \quad j = 1, 2, \ldots, n, \tag{1.25}$$

where $\mu \in \mathbb{R}$ and $\sigma > 0$ are appropriate constants. The standardised values $z_1, z_2, \ldots, z_n$ may be regarded as realisations of the random variable Z with distribution $F(z;\boldsymbol{\theta}')$.

In the context of integrated squared error estimation, the process of standardisation was first employed by *Paulson, Holcomb and Leitch (1975)*, but later

Paulson and Delehanty (1984) argued that it is not necessary. This view is consistent with our results from a wide variety of simulation trials. In addition, we have been successful in creating instances where the integrated squared error function was flatter when it was based on the standardised data than when based on the original data. One such instance is depicted in Figure 1.4.

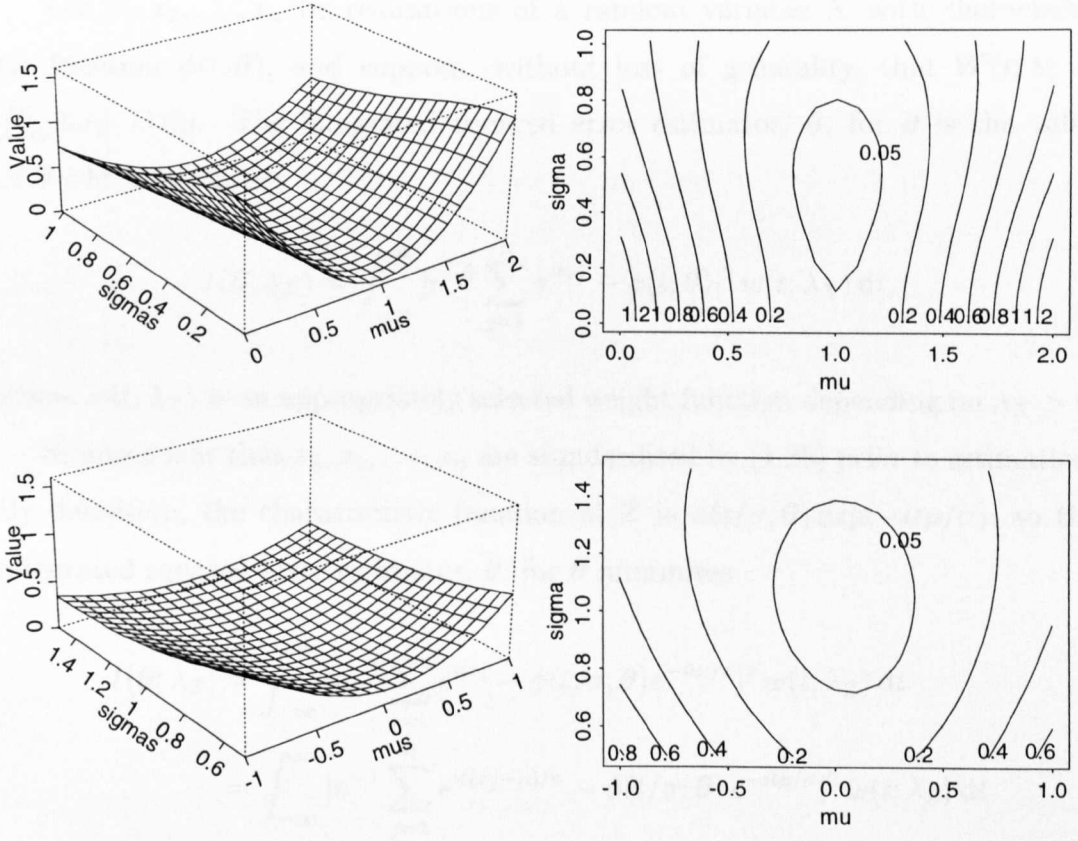

Figure 1.4: Perspective plot and contour-level plot (top) of the integrated squared error function (1.18) based on a sample of size $n = 50$ from a normal distribution with mean $\mu = 1$ and standard deviation $\sigma = 0.5$. Also included are the perspective plot and contour-level plot (bottom) of (1.18) based on the same but standardised sample. In either case, we have used $\lambda = 3/4$.

Perhaps of more importance, however, was the result that integrated squared error estimation was not invariant with respect to standardisations. This issue is examined in the next section.

1.10 Location and scale invariance of $\hat{\boldsymbol{\theta}}$

As indicated in the previous section, integrated squared error estimation has the disadvantage of not being invariant with respect to standardisations. In this section, we shall provide a mild restriction on the weight function $W(t;\lambda)$, which will then ensure that this invariance property is satisfied.

Let $x_1, x_2, \ldots, x_n$ be realisations of a random variable X with characteristic function $\phi(t;\boldsymbol{\theta})$, and suppose, without loss of generality, that $W(t;\lambda) = \int_{-\infty}^{t} w(y;\lambda)\,\mathrm{d}y$. The integrated squared error estimator, $\hat{\boldsymbol{\theta}}$, for $\boldsymbol{\theta}$ is the value of $\boldsymbol{\theta}$ which minimises

$$I(\boldsymbol{\theta};\lambda_X) = \int_{-\infty}^{\infty} |n^{-1}\sum_{j=1}^{n} e^{itx_j} - \phi(t;\boldsymbol{\theta})|^2\, w(t;\lambda_X)\,\mathrm{d}t,$$

where $w(t;\lambda_X)$ is an appropriately selected weight function depending on $\lambda_X > 0$.

Suppose now that $x_1, x_2, \ldots, x_n$ are standardised by (1.25) prior to estimation. By definition, the characteristic function of Z is $\phi(t/\sigma;\boldsymbol{\theta})\exp(-it\mu/\sigma)$, so the integrated squared error estimator, $\hat{\boldsymbol{\theta}}$, for $\boldsymbol{\theta}$ minimises

$$\begin{aligned}
I(\boldsymbol{\theta};\lambda_Z) &= \int_{-\infty}^{\infty} |n^{-1}\sum_{j=1}^{n} e^{itz_j} - \phi(t/\sigma;\boldsymbol{\theta})e^{-it\mu/\sigma}|^2\, w(t;\lambda_Z)\,\mathrm{d}t \\
&= \int_{-\infty}^{\infty} |n^{-1}\sum_{j=1}^{n} e^{it(x_j-\mu)/\sigma} - \phi(t/\sigma;\boldsymbol{\theta})\, e^{-it\mu/\sigma}|^2\, w(t;\lambda_Z)\,\mathrm{d}t \\
&= \int_{-\infty}^{\infty} |n^{-1}\sum_{j=1}^{n} e^{i(t/\sigma)x_j} - \phi(t/\sigma;\boldsymbol{\theta})|^2\, w(t;\lambda_Z)\,\mathrm{d}t \\
&= \sigma I(\boldsymbol{\theta};\lambda_X)
\end{aligned}$$

if and only if

$$w(\sigma t;\lambda_Z) = w(t;\lambda_X). \tag{1.26}$$

When (1.26) holds, integrated squared error estimation will be invariant with

respect to standardisations, since the problems of estimating $\boldsymbol{\theta}$ on the basis of $x_1, x_2, \ldots, x_n$ and on the basis of $z_1, z_2, \ldots, z_n$ are formally identical. The class of weight functions satisfying (1.26) is particularly rich. In general, any weight function of the form $w(t;\lambda) = w_1(\lambda^\alpha t^\beta)$, for some function $w_1(x)$ and $\alpha, \beta \in \mathbb{R}$, can be shown to satisfy (1.26) by taking

$$\lambda_Z = \sigma^{-\beta/\alpha}\lambda_X. \tag{1.27}$$

In view of (1.27), the parameter λ ensures that integrated squared error estimation is invariant with respect to standardisations. This provides a novel interpretation of this parameter and illustrates that its value should depend on the actual sample to hand.

We have reached a point where the theory that has been presented can be illustrated on more complicated distributions than the normal distribution. We shall take this opportunity to focus on a particular element of the family of stable laws, namely the Cauchy law. In the next section we introduce this important family of distributions and subsequently concentrate on the Cauchy law.

1.11 The family of stable laws

The family of stable laws has considerable importance in probability theory, though statistical applications appear to be rather limited. Nevertheless, *Paulson, Holcomb and Leitch (1975)* and *Leitch and Paulson (1975)* have applied stable laws to stock market data; see also *Koutrouvelis (1980).* In order for a random variable X to be stably distributed, it is necessary and sufficient that its characteristic function $\phi(t;\boldsymbol{\theta})$ be representable in the form (see, for example, *Lukaks, 1970, p. 136*):

$$\phi(t;\boldsymbol{\theta}) = \exp\{i\delta t - |ct|^\alpha[1 + i\beta\frac{t}{|t|}\omega(t,\alpha)]\}, \tag{1.28}$$

where $\boldsymbol{\theta} = (\alpha, \beta, c, \delta)^\top$, $t/|t| \equiv 0$ at $t = 0$, and

$$\omega(t, \alpha) = \begin{cases} \tan(\frac{\pi\alpha}{2}) & \text{if } \alpha \neq 1, \\ \frac{2}{\pi}\log(|t|) & \text{if } \alpha = 1. \end{cases}$$

The parameter space $\boldsymbol{\Theta}$ is given by

$$\boldsymbol{\Theta} = \{0 < \alpha \leq 2,\ |\beta| \leq 1,\ c > 0,\ |\delta| < \infty\},$$

where α is the characteristic exponent of the distribution, β is the skewness parameter, c (or sometimes $\gamma = c^\alpha$) is the scale parameter, and δ is the location parameter.

Stable laws are important largely because they are the only possible limiting laws for sums of independent and identically distributed random variables. All but one of the laws have infinite variance and this has negative implications for standard statistical methods. However, this has enabled stable laws to be considered as possible models for the distribution of noisy processes which typically arise in business and economics. A mathematical discussion of the properties of the stable laws may be found in *Gnedenko and Kolmogorov (1954, Chapter 7)* and *Lukacs (1970, Chapter 5).*

Let $F(x; \boldsymbol{\theta})$ and $f(x; \boldsymbol{\theta})$ be the distribution and density function corresponding to (1.28) respectively. An integral expression for $F(x; \boldsymbol{\theta})$ may be found in, for example, *Leitch and Paulson (1975).* A series representation for $f(x; \boldsymbol{\theta})$ may be found in *Johnson, Kotz and Balakrishnan (1994, p. 57).* In general, simple expressions for the density do not exist except in the cases

1. $\alpha = 1/2$, $\beta = 1$ (Levy law);

2. $\alpha = 1$, $\beta = 0$ (Cauchy law);

3. $\alpha = 2$ (normal law).

It is precisely for this reason that inference problems in the stable laws are more

naturally approached through the characteristic function. *Paulson, Holcomb and Leitch (1975)* provide a computationally attractive method for estimating $\boldsymbol{\theta}$ which is essentially based on the integrated squared error function. At the time of writing, we are not aware of any published application of the integrated squared error method specific to the Cauchy law. In addition, the Cauchy law will allow the theoretical performance of integrated squared error estimation relative to maximum likelihood estimation to be evaluated. This is lacking from *Paulson, Holcomb and Leitch (1975)*. As such, we shall now consider the Cauchy law in greater detail.

1.12 The Cauchy distribution

Let us consider the family of all stable laws with parameters $\alpha = 1$ and $\beta = 0$. This sub-family is commonly referred to as the Cauchy family of distributions, and has been studied in the mathematical world for over three centuries. An excellent account of the Cauchy distribution has been prepared by *Johnson, Kotz and Balakrishnan (1994, Chapter 16)*.

The characteristic function of the Cauchy distribution is given by

$$\phi(t;\boldsymbol{\theta}) = \exp(i\delta t - c|t|), \tag{1.29}$$

where $\boldsymbol{\theta} = (c, \delta)^\top$ belongs to the parameter space

$$\boldsymbol{\Theta} = \{c > 0,\ |\delta| < \infty\}.$$

The Cauchy distribution can be alternatively represented by its probability density function which, as previously indicated, can be expressed in simple terms. In particular, the probability density function corresponding to (1.29) is

$$f(x;\boldsymbol{\theta}) = \frac{c}{\pi[c^2 + (x-\delta)^2]}, \quad x \in \mathbb{R}. \tag{1.30}$$

The most notable difference between the Cauchy and normal distributions is

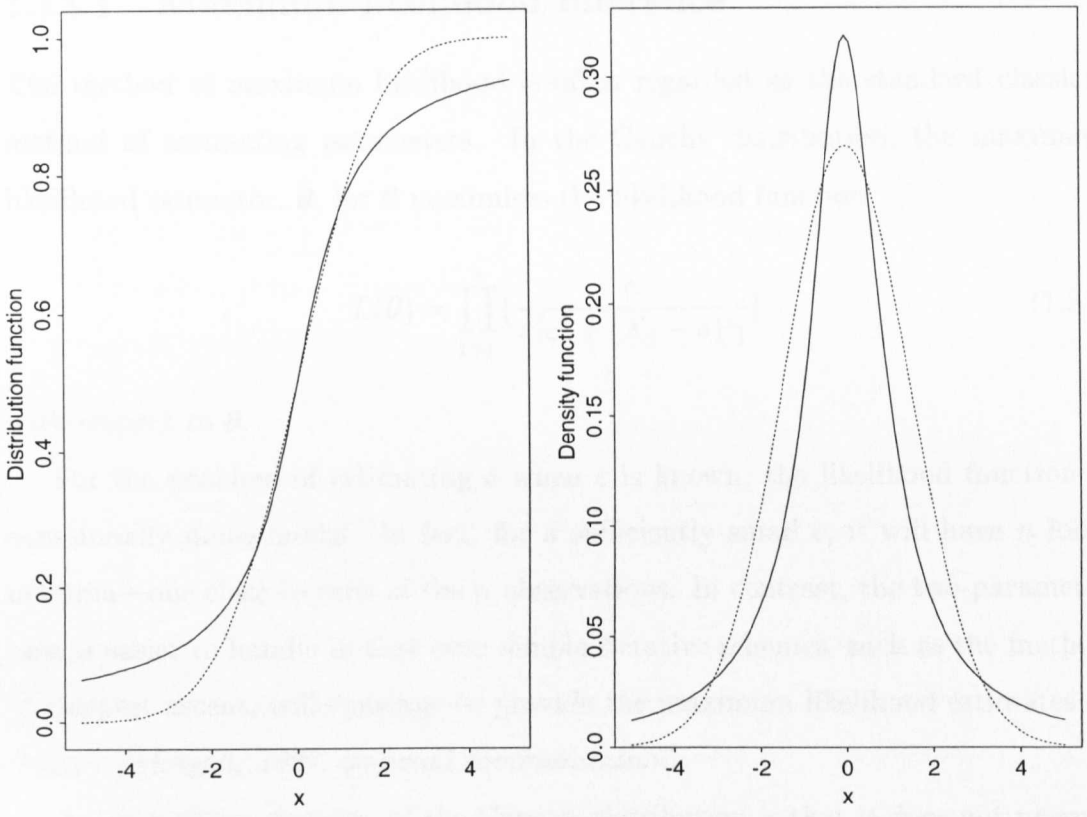

Figure 1.5: The Cauchy distribution and density function (solid lines) overlaid with the normal distribution and density function (dotted lines). The Cauchy distribution is in standard form, obtained by putting $\delta = 0, c = 1$; the normal distribution has mean zero and standard deviation $(0.67445)^{-1}$. The two distributions have the same median ($x = 0$) and upper and lower quartiles ($x = \pm 1$).

in the longer and flatter tails of the former. This difference is illustrated in Figure 1.5, and implies in the Cauchy distribution a greater frequency of both smaller and larger observations than would be expected under conditions of normality.

1.13 Estimation in the Cauchy distribution

Let $X_1, X_2, \ldots, X_n$ be independent, identically distributed random variables from a Cauchy distribution, and suppose that $\boldsymbol{\theta} = (c, \delta)^\top$ is unknown. The estimation of $\boldsymbol{\theta}$ has been investigated by numerous methods. *Johnson, Kotz and Balakrishnan (1994, Chapter 16)* discuss various methods including those based on order statistics, maximum likelihood, and Bayes theorem.

1.13.1 Maximum likelihood inference

The method of maximum likelihood is often regarded as the standard classical method of estimating parameters. In the Cauchy distribution, the maximum likelihood estimator, $\tilde{\boldsymbol{\theta}}$, for $\boldsymbol{\theta}$ maximises the likelihood function

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \{ \frac{c}{\pi[c^2 + (X_j - \delta)^2]} \} \tag{1.31}$$

with respect to $\boldsymbol{\theta}$.

For the problem of estimating δ when c is known, the likelihood function is occasionally multi-modal. In fact, for a sufficiently small c, it will have n local maxima—one close to each of the n observations. In contrast, the two-parameter case is easier to handle in that even simple iterative schemes, such as the method of steepest ascent, will converge to provide the maximum likelihood estimates of δ and c (*Morgan, 1997, personal communication*).

An interesting property of the Cauchy distribution is that it does not possess a finite mean or variance. This implies that observations of very large magnitude can be expected. As such, we seek to examine the robustness of the maximum likelihood estimator. The customary way of investigating the robustness of an estimator is through the behaviour of its influence function. The joint influence function for $\tilde{\boldsymbol{\theta}}$ is given by (see, for example, Campbell, 1992)

$$IF(\xi; \tilde{\boldsymbol{\theta}}) = \begin{pmatrix} 2c - 4c^3[c^2 + (\xi - \delta)^2]^{-1} \\ 4c^2(\xi - \delta)[c^2 + (\xi - \delta)^2]^{-1} \end{pmatrix}$$

so that the maximum likelihood estimator is robust against outliers.

1.13.2 Integrated squared error inference

It is straightforward to use the integrated squared error method to obtain parameter estimates for $\boldsymbol{\theta}$. In particular, the integrated squared error estimator, $\hat{\boldsymbol{\theta}}$, for

$\boldsymbol{\theta}$ is the value of $\boldsymbol{\theta}$ which minimises

$$I(\boldsymbol{\theta};\lambda) = \int_{-\infty}^{\infty} |n^{-1}\sum_{j=1}^{n} e^{itX_j} - e^{i\delta t - c|t|}|^2 \, \mathrm{d}W(t;\lambda) \tag{1.32}$$

for some appropriately selected weight function $W(t;\lambda)$. The function leading to the greatest degree of mathematical tractability is $W(t;\lambda) = \int_{-\infty}^{t} e^{-\lambda|y|}\mathrm{d}y$. This weight function will be adopted for this work. The integrated squared error function (1.32) becomes

$$I(\boldsymbol{\theta};\lambda) = \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}\frac{2}{\lambda^2+(X_j-X_k)^2} - \frac{1}{n}\sum_{j=1}^{n}\frac{4(\lambda+c)}{(\lambda+c)^2+(X_j-\delta)^2} + \frac{2}{\lambda+2c}, \tag{1.33}$$

and is illustrated in Figure 1.6 for a particular sample. Also depicted, in Figure 1.7, is the contour-level plot of this integrated squared error function.

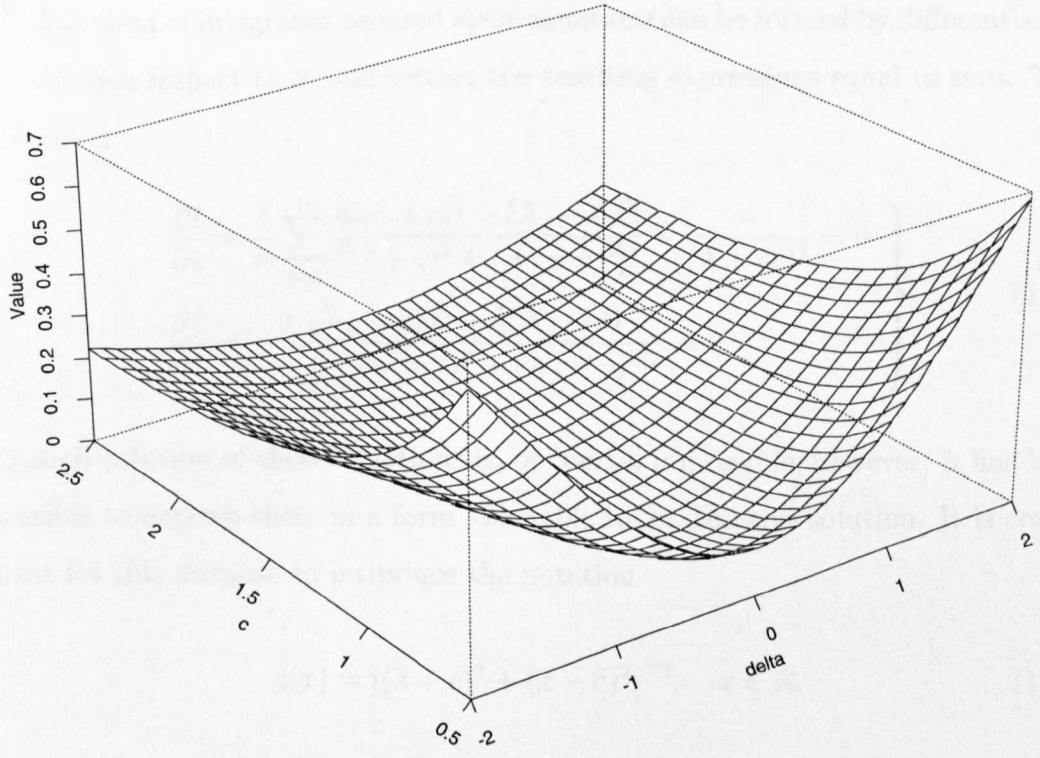

Figure 1.6: Perspective view of the integrated squared error function (1.33) based on a sample of size $n = 50$ from a Cauchy distribution with $\delta = 0, c = 1$. We have used $\lambda = 1$.

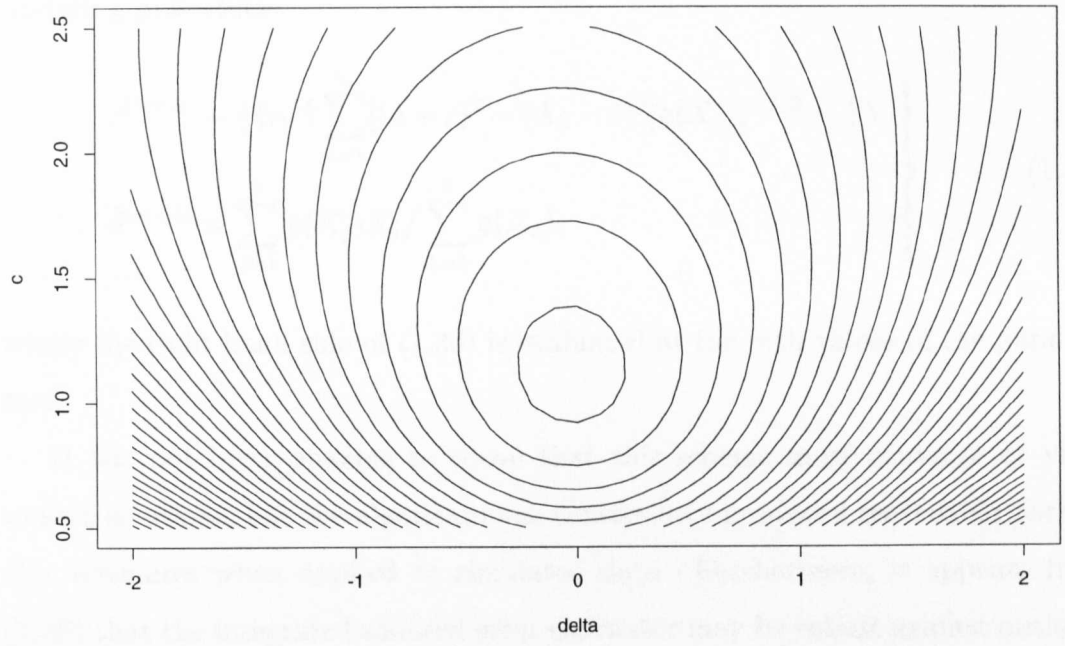

Figure 1.7: Contour-level plot of the integrated squared error function shown in Figure 1.6.

A system of integrated squared error equations can be formed by differentiating (1.33) with respect to $\boldsymbol{\theta}$, and setting the resulting expressions equal to zero. This results in

$$\left.\begin{aligned} \frac{\partial I}{\partial c} &= \frac{1}{n}\sum_{j=1}^{n}\frac{4[(\lambda+c)^2-(X_j-\delta)^2]}{[(\lambda+c)^2+(X_j-\delta)^2]^2} - \frac{4}{(\lambda+2c)^2} = 0 \\ \frac{\partial I}{\partial \delta} &= -\frac{1}{n}\sum_{j=1}^{n}\frac{8(\lambda+c)(X_j-\delta)}{[(\lambda+c)^2+(X_j-\delta)^2]^2} = 0. \end{aligned}\right\} \tag{1.34}$$

Explicit solution of these equations is, of course, impossible; however, it has been possible to express them in a form amenable to an iterative solution. It is convenient for this purpose to introduce the notation

$$g(x) = [(\lambda+c)^2+(x-\delta)^2]^{-2}, \quad x \in \mathbb{R}. \tag{1.35}$$

Then, if $\hat{\delta}^{(m)}$ and $\hat{c}^{(m)}$ represent the parameter estimates at the mth iteration of the iterative algorithm, the equations (1.34) may be manipulated to suggest the

updating procedure

$$\left.\begin{aligned} \hat{c}^{(m+1)} &= \tfrac{1}{2}\{n^{-1}\sum_{j=1}^{n}[(\lambda+c)^2-(X_j-\delta)^2]g(X_j)\}^{-1/2}-\tfrac{1}{2}\lambda \\ \hat{\delta}^{(m+1)} &= \sum_{j=1}^{n} g(X_j)X_j/\sum_{k=1}^{n} g(X_k), \end{aligned}\right\} \tag{1.36}$$

where the right hand side of (1.36) is evaluated at the mth values of the parameters.

It has not been possible to prove that this scheme must converge or yield unique solutions, but the algorithm has consistently produced reasonable parameter estimates when applied to simulated data. Furthermore, it appears from (1.36) that the integrated squared error estimator may be robust against outliers. Consider for example the relationship

$$\hat{\delta} = \sum_{j=1}^{n} a(X_j)X_j,$$

which follows from (1.36) if the iteration converges, with $\hat{\delta}^{(m)} \to \hat{\delta}$ as $m \to \infty$. In this expression, $a(X_j) = g(X_j)/\sum_k g(X_k)$ and it it is clear from (1.35) that the weights $a(X_j)$ are smallest for those X_j which are most removed from $\hat{\delta}$. The estimator $\hat{\delta}$ is thus robust against outliers. This can be formally examined, for both $\hat{\delta}$ and $\hat{c}$, by Theorem 1.3 and some straightforward but tedious computations. We find that the joint influence function for $\hat{\boldsymbol{\theta}}$ is given by

$$IF(\xi;\hat{\boldsymbol{\theta}}) = \begin{pmatrix} \tfrac{1}{2}(\lambda+2c)-\tfrac{1}{2}(\lambda+2c)^3[(\lambda+c)^2-(\xi-\delta)^2]g(\xi) \\ (\lambda+c)(\lambda+2c)^3(\xi-\delta)g(\xi) \end{pmatrix} \tag{1.37}$$

so that the integrated squared error estimator is indeed robust against outlying observations.

The asymptotic distribution of $\hat{\boldsymbol{\theta}}$ is readily obtained from Theorem 1.2. We find that $n^{1/2}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})$ is asymptotically normally distributed with mean vector zero

and covariance matrix

$$\Sigma(\boldsymbol{\theta}) = \frac{c\,(\lambda + 2c)^2\,(5\lambda^2 + 14\lambda c + 10c^2)}{16\,(\lambda + c)^3} I_2, \tag{1.38}$$

where I_2 is the 2×2 identity matrix. The Fisher information matrix for a single observation is

$$\mathcal{I}(\boldsymbol{\theta}) = \frac{1}{2c^2}\, I_2$$

so that, for example, the asymptotic efficiency of $\hat{\delta}$ is given by

$$\mathit{eff}(\hat{\delta}) = \frac{32\,c\,(\lambda + c)^3}{(\lambda + 2c)^2\,(5\lambda^2 + 14\lambda c + 10c^2)}. \tag{1.39}$$

Note that the efficiency of $\hat{\delta}$ is independent of δ, and that it also happens to coincide with the efficiency of $\hat{c}$. Figures 1.8 and 1.9 show that this efficiency is very high for most values of λ.

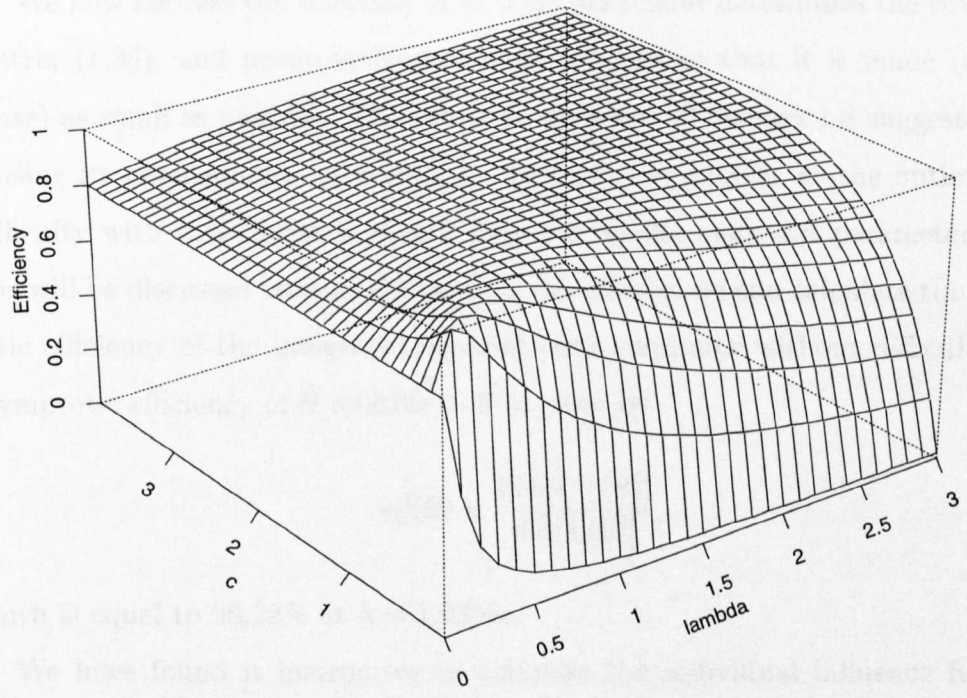

Figure 1.8: Perspective view of the asymptotic efficiency of $\hat{\delta}$.

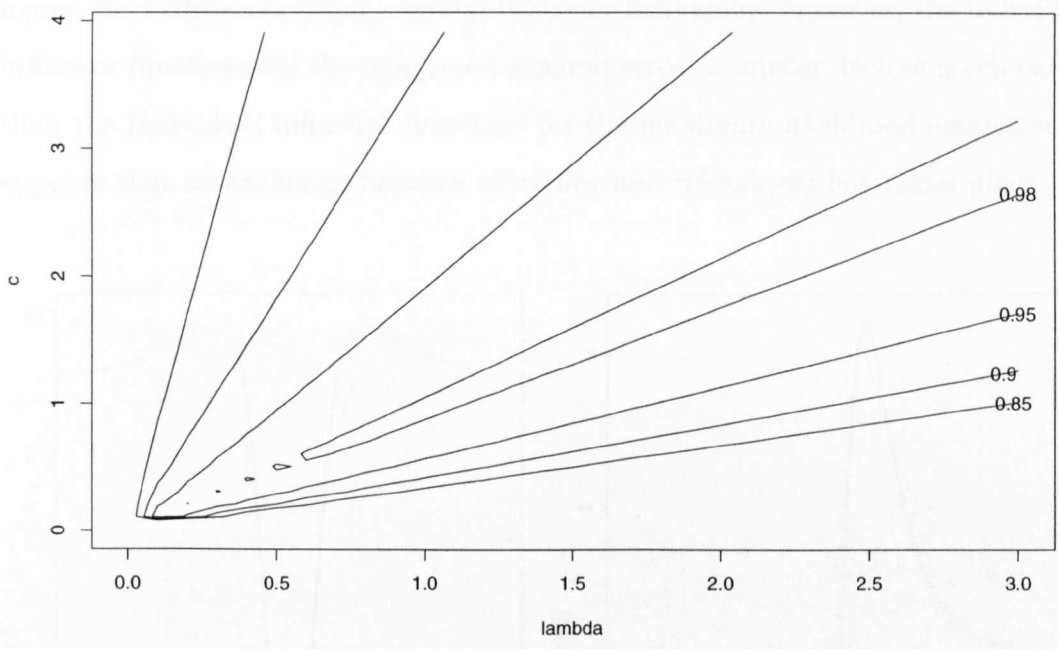

Figure 1.9: Contour-level plot of the efficiency of $\hat{\delta}$ as shown in Figure 1.8. The plot illustrates that $\hat{\delta}$ can attain efficiencies of over 98%.

We now address the selection of λ. This parameter determines the covariance matrix (1.38), and presumably should be selected so that it is made (in some sense) as small as possible. The precision approach of Section 1.8 suggested minimising its determinant. Accordingly, we find $\lambda = 1.0315c$ at the optimum. A difficulty with this choice is that it depends on the unknown parameter c, and this will be discussed later in this section. Meantime, we can calculate the asymptotic efficiency of the integrated squared error estimator without difficulty. The asymptotic efficiency of $\hat{\boldsymbol{\theta}}$ relative to $\tilde{\boldsymbol{\theta}}$ is given by

$$\mathit{eff}(\hat{\boldsymbol{\theta}}) = \frac{\det[\mathcal{I}^{-1}(\boldsymbol{\theta})]}{\det[\Sigma(\boldsymbol{\theta})]},$$

which is equal to 96.22% at $\lambda = 1.0315c$.

We have found it instructive to compare the individual influence functions for the integrated squared error estimator with $\lambda = 1.0315c$ and the individual influence functions for the maximum likelihood estimator. These are depicted in Figure 1.10 for the Cauchy distribution with $\delta = 0$ and $c = 1$. As observed in the

figure, the estimators display similar influence behaviour. However, the individual influence functions for the integrated squared error estimator decline much earlier than the individual influence functions for the maximum likelihood estimator. It appears that an exchange between efficiency and robustness has taken place.

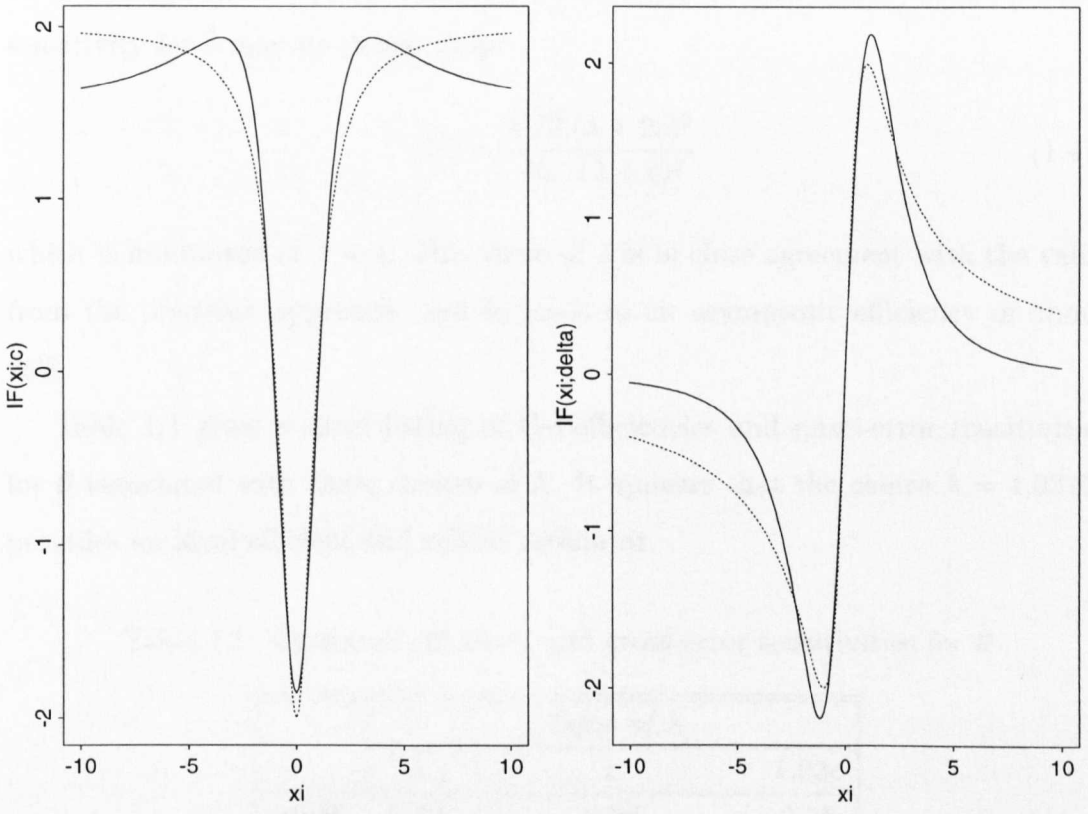

Figure 1.10: Individual influence functions for the integrated squared error estimator with $\lambda = 1.0315$ (solid lines) overlaid with individual influence functions for the maximum likelihood estimator (dotted lines). The influence functions are evaluated at the Cauchy distribution with $\delta = 0, c = 1$.

Alternatively, since the parameter λ also determines the joint influence function (1.37), it could be selected by optimising some measure of robustness of the estimators. The robustness approach of Section 1.8 suggests the gross-error sensitivity (1.24), which we shall now consider. The gross-error sensitivity for $\hat{c}$ may

be shown to be

$$\gamma(\hat{c}) = \frac{1}{2}(\lambda + 2c) + \frac{1}{16}\frac{(\lambda + 2c)^3}{(\lambda + c)^2}, \tag{1.40}$$

which is minimised over the positive real axis as $\lambda \to 0$. This value for λ leads to an asymptotic efficiency of about 64%. On the other hand, the gross-error sensitivity for $\hat{\delta}$ may be shown to be

$$\gamma(\hat{\delta}) = \frac{3\sqrt{3}}{16}\frac{(\lambda + 2c)^3}{(\lambda + c)^2}, \tag{1.41}$$

which is minimised at $\lambda = c$. This value of λ is in close agreement with the value from the precision approach, and so leads to an asymptotic efficiency of about 96%.

Table 1.1 gives a short listing of the efficiencies and gross-error sensitivities for $\hat{\boldsymbol{\theta}}$ associated with these choices of λ. It appears that the choice $\lambda = 1.0315c$ provides an ideal efficient and robust estimator.

Table 1.1: Optimum efficiency and gross-error sensitivities for $\hat{\boldsymbol{\theta}}$

	Value of λ		
	0	c	$1.03c$
$eff(\hat{\boldsymbol{\theta}})$	0.64	0.96	0.96
$\gamma(\hat{c})$	$1.50c$	$1.92c$	$1.94c$
$\gamma(\hat{\delta})$	$2.60c$	$2.19c$	$2.19c$

Finally, we must discuss the problem which arises from the optimal choice of λ depending on the true parameter value c. Some proposals have been made in the context of generalised moment estimation by *Ball and Milne (1996)*. In the context of integrated squared error estimation, this concept can be employed as follows. Denote the integrated squared error estimator by $\hat{\boldsymbol{\theta}}(\lambda)$ to emphasise its dependence on λ. Then, in a similar fashion to *Ball and Milne (1996)*, choose $\lambda^{(0)}$ arbitrarily and let $\hat{\boldsymbol{\theta}}^{(0)} = \hat{\boldsymbol{\theta}}(\lambda^{(0)})$; next, for $m = 1, 2, \ldots$ let $\lambda^{(m)} = \lambda(\hat{\boldsymbol{\theta}}^{(m-1)})$ and

$\hat{\boldsymbol{\theta}}^{(m)} = \hat{\boldsymbol{\theta}}(\lambda^{(m)})$; finally, if this iteration converges, with $\lambda^{(m)} \to \lambda^A$ as $m \to \infty$, then $\boldsymbol{\theta}$ may be estimated by the *adaptive estimator* $\hat{\boldsymbol{\theta}}(\lambda^A)$.

In the present context, the adaptive estimator was found to be very effective in a wide variety of computer simulations.

In summary, we can outline the results of this section as follows. The integrated squared error estimator for the Cauchy parameters is developed. The asymptotic variance and influence function of the estimator play a double role in our analysis: first, they provide a basis for the theoretical evaluation of the estimator; and second, they prove useful in selecting λ. A difficulty which remains is that considering different aspects of the estimators leads to different values of λ. This problem will be dealt with in Chapter 2.

1.14 Integrated distance estimation based on moment generating functions

The range of problems to which the integrated squared error method is appropriate appears to be very wide. This is because the characteristic function is in one-to-one correspondence with the distribution function while behaving simply under shifts and scale changes. In addition, it allows an easy characterisation of independence and of symmetry. This statement is not to deny, however, that there exist other functionals with similar properties. The moment generating function is one obvious alternative which leads to the special case

$$J(\boldsymbol{\theta}) = \int_T [M_n(t) - M(t;\boldsymbol{\theta})]^2 \, \mathrm{d}W(t) \tag{1.42}$$

of (1.5). The integrated distance estimator based on (1.42) can be regarded as a generalised moment estimator, and was first proposed by *Leslie (1970)*.

In analogy with the integrated squared error method, the weight function $W(t)$ is open to choice. Nevertheless, the consistency and asymptotic normality of this estimator have been investigated by *Quandt and Ramsey (1978)* when $W(t)$ is a

step function, and *Leslie and Khalique (1980)* when $W(t)$ is a continuous function. In this section, we shall restrict attention to continuous weight functions in order to compare the resulting method with the integrated squared error method.

Integrated distance estimation based on moment generating functions appears to be easier to deal with algebraically because:

1. it does not contain complex numbers;
2. empirical moment generating functions are much smoother than empirical characteristic functions (see *Kumar, Nicklin and Paulson, 1979*).

While this is a plausible argument on the surface, there can be serious difficulties associated with the use of the moment generating function. For example, consider the distance

$$J(\boldsymbol{\theta};\lambda) = \int_{-\infty}^{\infty} [n^{-1}\sum_{j=1}^{n} e^{tX_j} - e^{\mu t+\sigma^2 t^2/2}]^2 \, \mathrm{d}W(t;\lambda)$$

as an alternative to distance (1.17). Setting $W(t;\lambda) = \int_{-\infty}^{t} e^{-\lambda^2 y^2} \, \mathrm{d}y$ results in

$$J(\boldsymbol{\theta};\lambda) = \frac{\pi^{1/2}}{\lambda n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}\exp[\frac{1}{4}\frac{(X_j+X_k)^2}{\lambda^2}] - \frac{2}{n}(\frac{\pi}{\lambda^2-\frac{1}{2}\sigma^2})^{1/2}\sum_{j=1}^{n}\exp[\frac{1}{4}\frac{(X_j+\mu)^2}{\lambda^2-\frac{1}{2}\sigma^2}]$$
$$+ (\frac{\pi}{\lambda^2-\sigma^2})^{1/2}\exp(\frac{\mu^2}{\lambda^2-\sigma^2})$$

for $\lambda > \sigma$. In this case, the integrated squared error method is favoured because:

1. numerical problems with large exponential terms do not occur;
2. numerical results show greater stability;
3. the admissible range of values for λ is not curtailed.

All these features result from the uniform boundedness of the characteristic function, and do not, of course, depend on the underlying distribution being normal. In addition, the integrated squared error method is superior to the method based on moment generating functions because:

4. it can be used when no moments exist;

5. the practical range of choice of $W(\,.\,)$ will be wider (see *Clarke and Heathcote, 1978*).

In conclusion, the form of the integrated squared error function might be more complicated, but it is analytically sounder. In light of this result, we do not recommend the use of integrated distance methods based on moment generating functions. Accordingly, for the most part, we shall concentrate on the integrated squared error method.

Chapter 2

Density representation of the ISE function

2.1 Scope of this chapter

In view of the one-to-one correspondence between probability density functions and characteristic functions, their Fourier transforms, models of distributions can be represented equivalently by either function. In practice, the density function is the usual representation, because it is the more intuitive concept and because the method of maximum likelihood relies on it. On the other hand, the characteristic function is the canonical representation of some useful distributions whose density functions cannot be expressed in closed form. The difficulty of applying maximum likelihood to these models led to the advent of methods based on characteristic functions. The integrated squared error method of Chapter 1 is a paradigm of this approach, which has received significant attention in the literature. In some situations, such as the stable laws, it may be argued that it is the best method available. However, it is safe to say that the integrated squared error method is not yet widely used by applied statisticians.

A factor that has possibly limited the use of the integrated squared error method is the difficulty behind the choice of the weight function $W(t;\lambda)$. In the general case, it is not clear how one might select a suitable weight function. This

is partly a consequence of the complicated dependence of the integrated squared error estimators on $W(t;\lambda)$.

This chapter uses an alternative representation of the integrated squared error function to make three innovative contributions: first, it provides a viable way of selecting the weight function; second, it demonstrates that the choice of the weight function is not particularly important but the choice of its scaling is; and third, it describes how this scaling can be selected in practice.

2.2 Density representation of the ISE function

Let $\{F(x;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ be a family of distribution functions indexed by the parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)^\top$, and suppose that $X_1, X_2, \ldots, X_n$ is a random sample from some population whose distribution function is a member of this family with parameter vector $\boldsymbol{\theta}_0$. In Chapter 1 we considered parameter estimation based on the empirical characteristic function $\phi_n(t) = n^{-1}\sum_{j=1}^{n} e^{itX_j}$, which involved minimising the integrated squared error function

$$I(\boldsymbol{\theta};\lambda) = \int_{-\infty}^{\infty} |\phi_n(t) - \phi(t;\boldsymbol{\theta})|^2 \,\mathrm{d}W(t;\lambda)$$

with respect to $\boldsymbol{\theta}$. The integrated squared error function may be regarded as a measure of the deviation between $\phi_n(t)$ and the characteristic function $\phi(t;\boldsymbol{\theta}) = \int_{-\infty}^{\infty} e^{itx}\,\mathrm{d}F(x;\boldsymbol{\theta})$, which corresponds to $F(x;\boldsymbol{\theta})$. The weight function $W(t;\lambda)$ was a monotonic non-decreasing continuous function, which we shall now assume to be given by

$$W(t;\lambda) = \int_{-\infty}^{t} |w(y;\lambda)|^2 \,\mathrm{d}y$$

for some function $w(y;\lambda)$. There is no loss of generality in so doing, since the choice of the weight function is arbitrary. As a result, the integrated squared error function becomes

$$I(\boldsymbol{\theta};\lambda) = \int_{-\infty}^{\infty} |\varphi_n(t;\lambda) - \varphi(t;\boldsymbol{\theta},\lambda)|^2 \,\mathrm{d}t, \tag{2.1}$$

where

$$\varphi_n(t;\lambda) = \phi_n(t)w(t;\lambda) \tag{2.2}$$

$$\varphi(t;\boldsymbol{\theta},\lambda) = \phi(t;\boldsymbol{\theta})w(t;\lambda). \tag{2.3}$$

If we could determine that $\varphi_n(t;\lambda)$ and $\varphi(t;\boldsymbol{\theta},\lambda)$ are characteristic functions, then Parseval's theorem (see, for example, *Priestley, 1981, p. 201*) allows an expression of $I(\boldsymbol{\theta};\lambda)$ in terms of densities. It turns out that if $w(t;\lambda)$ is a characteristic function, then so are (2.2) and (2.3). In this case, we may re-write (2.1) as

$$I(\boldsymbol{\theta};\lambda) = 2\pi \int_{-\infty}^{\infty} [f_{\varphi_n}(x;\lambda) - f_{\varphi}(x;\boldsymbol{\theta},\lambda)]^2 \,\mathrm{d}x, \tag{2.4}$$

where $f_{\varphi_n}(x;\lambda)$ and $f_{\varphi}(x;\boldsymbol{\theta},\lambda)$ are the densities corresponding to the characteristic functions $\varphi_n(t;\lambda)$ and $\varphi(t;\boldsymbol{\theta},\lambda)$ respectively. These densities may be expressed in terms of the convolutions

$$f_{\varphi_n}(x;\lambda) = f_{\phi_n}(x) * f_w(x;\lambda) \tag{2.5}$$

$$f_{\varphi}(x;\boldsymbol{\theta},\lambda) = f(x;\boldsymbol{\theta}) * f_w(x;\lambda), \tag{2.6}$$

where $f_{\phi_n}(x)$ and $f_w(x;\lambda)$ are the probability mass and density functions, respectively, corresponding to the characteristic functions $\phi_n(t)$ and $w(t;\lambda)$.

The density representation of the integrated squared error function, as stated in (2.4), has been noted by a number of authors, including *Heathcote (1977)*, *Paulson and Nicklin (1983)*, and *Bryant and Paulson (1983)*. The practical utilities of density and characteristic function representations will depend on the underlying model distribution. However, (2.4) demonstrates that the integrated squared error method conforms to the more common formulation of minimum distance methods, in which the distance is based on density functions. This aspect is important to our work for two reasons:

1. it readily shows that the integrated squared error method possesses good

robustness properties (see, for example, *Parr and Schucany, 1980*);

2. it leads to the interpretation of the weight function $w(t;\lambda)$ in terms of a smoothing or randomising density $f_w(x;\lambda)$ (see *Heathcote, 1977*).

Heathcote (1977) observed that the smoothing by $f_w(x;\lambda)$ operation results in certain desirable properties, and suggested that a sufficiently smooth and tractable weight function should be selected. He also recognised that the optimum (in some sense) weight function will generally depend on $F(x;\boldsymbol{\theta})$, but otherwise the choice of $W(t;\lambda)$ remains unsupported by the literature. As a consequence, the important mathematical properties of the normal distribution have resulted in the choice

$$w(t;\lambda) = \exp(-\tfrac{1}{2}\lambda^2 t^2), \tag{2.7}$$

or, equivalently,

$$f_w(x;\lambda) = (2\pi\lambda^2)^{-1/2}\exp[-\tfrac{1}{2}(x/\lambda)^2], \tag{2.8}$$

being extensively used in applications.

However, it is not difficult to think of applications where other choices of $w(t;\lambda)$ may be used more effectively than (2.7). One such example is the Cauchy distribution discussed in Chapter 1 (Section 1.12), where the weight function $w(t;\lambda) = \exp(-\lambda|t|/2)$ was used. Consequently, there is considerable scope for investigating the problem of obtaining the optimum (in some sense) weight function. In the following sections we shall exploit the density representation of the integrated squared error function in order to obtain a satisfactory solution to this problem.

2.3 Kernel density estimation

As demonstrated in the previous section, the integrated squared error function can admit a density representation provided that the weight function is chosen

as a characteristic function. In fact, it is through this density representation that we encounter a link between the methods of integrated squared error and kernel density estimation. We can see this explicitly by writing (2.5) in the form

$$f_{\varphi_n}(x;\lambda) = \int f_w(x-y;\lambda)\,\mathrm{d}F_{\phi_n}(y), \tag{2.9}$$

where $F_{\phi_n}(y)$ is the distribution function of $f_{\phi_n}(y)$. Since $F_{\phi_n}(y)$ is the empirical distribution function of $X_1, X_2, \ldots, X_n$, (2.9) becomes

$$f_{\varphi_n}(x;\lambda) = n^{-1}\sum_{j=1}^{n} f_w(x-X_j;\lambda). \tag{2.10}$$

By writing (2.5) in the form (2.10) we are immediately made aware that $f_{\varphi_n}(x;\lambda)$ is a kernel density estimator employing the density $f_w(x;\lambda)$ which is called the *kernel*, and a positive number λ which is called the *bandwidth*.

Kernel density estimators have been around since the seminal papers of *Rosenblatt (1956)* and *Parzen (1962)*. To date articles written about kernel density estimators number in the thousands. This chapter makes no endeavour to survey the field of kernel density estimation; such a review can be found in, for example, *Wand and Jones (1995)*. Instead, our goal is to present the aspects of kernel density estimation which we see as being relevant to integrated squared error estimation. In particular, we intend to use developments from kernel density estimation since:

1. the weight function employed in the integrated squared error estimation of $\boldsymbol{\theta}_0$ corresponds to the kernel employed in the kernel density estimation of $f(x;\boldsymbol{\theta}_0)$;

2. the scaling of the weight function in the above integrated squared error problem corresponds to the bandwidth employed in the above kernel density estimation problem.

On this basis, we conjecture that the problem of selecting the weight function

in integrated squared error estimation is in some respects similar to the problem of selecting the kernel in kernel density estimation; the methods employed here are inspired by the methods used in the treatment of the latter problem. Likewise, the methods used to select the scaling of the weight function in integrated squared error estimation are inspired by the techniques used in selecting the bandwidth in kernel density estimation.

This discussion suggests that the density representation of the integrated squared error function will be of substantial theoretical utility. We have, therefore, found it useful to reiterate the statistical properties of the estimators from this perspective. As we shall see, there are advantages to such a course of action.

2.4 Properties of the integrated squared error estimator

As usual, let $\{F(x;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ be a family of distribution functions indexed by $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)^\top$, and suppose that $X_1, X_2, \ldots, X_n$ is a random sample from some population whose distribution function is a member of this family with parameter vector $\boldsymbol{\theta}_0$. Furthermore, suppose that $w(t;\lambda)$ is a characteristic function with corresponding density $f_w(x;\lambda)$. In this case, we have:

Theorem 2.1. *Under the regularity conditions of Section 1.6.1, the integrated squared error estimator, $\hat{\boldsymbol{\theta}}$, for $\boldsymbol{\theta}_0$ has joint influence function*

$$IF(\xi;\hat{\boldsymbol{\theta}}) = K^{-1}(\boldsymbol{\theta}_0)\,\boldsymbol{\tau}(\xi;\boldsymbol{\theta}_0),$$

where $K(\boldsymbol{\theta})$ is the $p \times p$ symmetric matrix whose (i,j)th element is

$$\kappa_{ij}(\boldsymbol{\theta}) = \int_{-\infty}^{\infty} \frac{\partial f(x;\boldsymbol{\theta}) * f_w(x;\lambda)}{\partial \theta_i} \frac{\partial f(x;\boldsymbol{\theta}) * f_w(x;\lambda)}{\partial \theta_j}\,\mathrm{d}x,$$

and $\boldsymbol{\tau}(\xi;\boldsymbol{\theta})$ *is the* $p \times 1$ *vector whose* i*th element is*

$$\tau_i(\xi;\boldsymbol{\theta}) = \int_{-\infty}^{\infty} \frac{\partial f(x;\boldsymbol{\theta}) * f_w(x;\lambda)}{\partial \theta_i} \left[f_w(x-\xi;\lambda) - f(x;\boldsymbol{\theta}) * f_w(x;\lambda) \right] \mathrm{d}x.$$

To prove Theorem 2.1 we need the following results which may be proved by standard arguments (see, for example, *Campbell, 1993*):

Result 2.4.1. Given a distribution function $F(x)$, we may define a functional $T = T(F)$ by

$$T(F) = \int t(x)\, \mathrm{d}F(x).$$

This functional may then be estimated by $T_n = T(F_n)$, where $F_n(x)$ is the empirical distribution function based on a random sample $X_1, X_2, \ldots, X_n$ from $F(x)$. In this case,

1. $IF(\xi; T_n) = t(\xi) - T$.

2. If $h_n = h(T_n)$, where h is a differentiable function, then
   $IF(\xi; h_n) = (\partial h/\partial T)\, IF(\xi; T_n)$.

3. If $T = (T_1, T_2, \ldots, T_p)^\top$ in the second result above, then
   $IF(\xi; h_n) = \sum_{i=1}^{p} (\partial h/\partial T_i)\, IF(\xi; T_{in})$.

4. If $h_n = \int T(F_n; s)\mathrm{d}s$, then
   $IF(\xi; h_n) = \int IF(\xi; T_n(s))\mathrm{d}s$.

Proof (Theorem 2.1). The integrated squared estimator $\hat{\boldsymbol{\theta}}$ is found in practice as the solution of

$$\frac{\partial I(\boldsymbol{\theta};\lambda)}{\partial \theta_i} = 0, \quad i = 1, 2, \ldots, p,$$

which, from expression (2.4), may be written as

$$\int_{-\infty}^{\infty} \frac{\partial f_\varphi(x;\boldsymbol{\theta},\lambda)}{\partial \theta_i} [f_{\varphi_n}(x;\lambda) - f_\varphi(x;\boldsymbol{\theta},\lambda)]\, \mathrm{d}x = 0, \quad i = 1, 2, \ldots, p.$$

Applying first (4) and then (3) of Result 2.4.1 to the ith ($i = 1, 2, \ldots, p$) equation

of this system, we find the ith influence equation

$$\int_{-\infty}^{\infty} \frac{\partial f_\varphi(x;\boldsymbol{\theta},\lambda)}{\partial \theta_i}\{IF[\xi; f_{\varphi_n}(x;\lambda)] - \sum_{j=1}^{p} \frac{\partial f_\varphi(x;\boldsymbol{\theta},\lambda)}{\partial \theta_j} IF(\xi;\hat{\theta}_j)\}\,\mathrm{d}x = 0.$$

Re-arranging we obtain

$$\int_{-\infty}^{\infty} \frac{\partial f_\varphi(x;\boldsymbol{\theta},\lambda)}{\partial \theta_i} IF[\xi; f_{\varphi_n}(x;\lambda)]\,\mathrm{d}x = \int_{-\infty}^{\infty} \frac{\partial f_\varphi(x;\boldsymbol{\theta},\lambda)}{\partial \theta_i} \sum_{j=1}^{p} \frac{\partial f_\varphi(x;\boldsymbol{\theta},\lambda)}{\partial \theta_j} IF(\xi;\hat{\theta}_j)\,\mathrm{d}x,$$

or, equivalently,

$$\tau_i(\xi;\boldsymbol{\theta}) = \sum_{j=1}^{p} \kappa_{ij}(\boldsymbol{\theta}) IF(\xi;\hat{\theta}_j),$$

since $IF[\xi; f_{\varphi_n}(x;\lambda)] = f_w(x-\xi;\lambda) - f_\varphi(x;\boldsymbol{\theta},\lambda)$ by (1) of Result 2.4.1 and $f_\varphi(x;\boldsymbol{\theta},\lambda)$ is given by (2.6). Bringing these p influence equations together, we may form the single matrix equation

$$\boldsymbol{\tau}(\xi;\boldsymbol{\theta}) = K(\boldsymbol{\theta}) IF(\xi;\hat{\boldsymbol{\theta}}),$$

or, equivalently, $IF(\xi;\hat{\boldsymbol{\theta}}) = K^{-1}(\boldsymbol{\theta})\,\boldsymbol{\tau}(\xi;\boldsymbol{\theta})$. □

As noted in Chapter 1, influence functions do not only convey information regarding robustness, they also provide a convenient method for calculating asymptotic variances. On this basis, the density representation of Theorem 1.2 follows.

Theorem 2.2. *Under the regularity conditions of Section 1.6.1, the integrated squared error estimator, $\hat{\boldsymbol{\theta}}$, for $\boldsymbol{\theta}_0$ is strongly consistent and*

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}[\mathbf{0}\,,\, \Sigma(\boldsymbol{\theta}_0)],$$

where

$$\Sigma(\boldsymbol{\theta}) = K^{-1}(\boldsymbol{\theta})\,\Omega(\boldsymbol{\theta})\,K^{-1}(\boldsymbol{\theta}).$$

In this expression, $K(\boldsymbol{\theta})$ is the $p \times p$ symmetric matrix whose (i,j)th element is

$$\kappa_{ij}(\boldsymbol{\theta}) = \int_{-\infty}^{\infty} \frac{\partial f(x;\boldsymbol{\theta}) * f_w(x;\lambda)}{\partial \theta_i} \frac{\partial f(x;\boldsymbol{\theta}) * f_w(x;\lambda)}{\partial \theta_j} \, \mathrm{d}x,$$

and $\Omega(\boldsymbol{\theta})$ is the covariance matrix of the p random variables

$$\tau_i(X_1;\boldsymbol{\theta}) = \int_{-\infty}^{\infty} \frac{\partial f(x;\boldsymbol{\theta}) * f_w(x;\lambda)}{\partial \theta_i} [f_w(x - X_1;\lambda) - f(x;\boldsymbol{\theta}) * f_w(x;\lambda)] \, \mathrm{d}x.$$

Proof. The consistency and asymptotic normality of $\hat{\boldsymbol{\theta}}$ follow from Theorem 1.2, while the form of the asymptotic covariance matrix follows immediately from Theorem 2.1 and expression (1.16). □

2.5 The mean integrated squared error criterion

As suggested in Section 2.3, we intend to bring in developments from kernel density estimation to integrated squared error estimation. In the former context, estimators are compared with reference to various error criteria. This is perhaps the most critical stage in the entire undertaking. In this section we present one such error criterion, which we shall use extensively later.

Let $X_1, X_2, \ldots, X_n$ be independent, identically distributed random variables with density $f(x;\boldsymbol{\theta}_0)$, and suppose that

$$f_n(x;h) = (nh)^{-1} \sum_{j=1}^{n} K[(x - X_j)/h] \tag{2.11}$$

is a kernel estimator of this density, with kernel $K(x)$ and bandwidth h. A slightly more convenient formula for the kernel estimator can be obtained by introducing the rescaling notation $K(x;h) = h^{-1}K(x/h)$. This allows us to write

$$f_n(x;h) = n^{-1} \sum_{j=1}^{n} K(x - X_j;h), \tag{2.12}$$

which is the notation used in (2.10). The analysis of the performance of the

kernel density estimator requires the specification of appropriate error criteria for measuring the error when estimating the density at a single point as well as the error when estimating the density over the whole real line.

When estimating the density at a fixed point x, it is common to measure the performance of the kernel estimator by the size of the *mean squared error* (MSE)

$$MSE[f_n(x;h)] = \mathrm{E}[f_n(x;h) - f(x;\boldsymbol{\theta}_0)]^2.$$

The MSE is rarely referred to in the context of density estimation because it is usually desirable to estimate the density over the entire real line. In this case it is more appropriate to consider an error criterion that globally measures the distance between the density and its kernel estimator. One such criterion is the *mean integrated squared error* ($MISE$) defined by

$$MISE[f_n(\,.\,;h)] = \mathrm{E}\int [f_n(x;h) - f(x;\boldsymbol{\theta}_0)]^2\,\mathrm{d}x. \tag{2.13}$$

The mean integrated squared error may be recognised as the global counterpart of the mean squared error, since $MISE(f_n) = \int MSE(f_n)$. The decision to work with the mean integrated squared error is largely because of its mathematical simplicity. There are also good reasons for working with other criteria, such as the weighted mean integrated squared error (see, for example, *Fryer, 1976*), and the mean integrated absolute error (see, for example, *Devroye and Györfi, 1985, p. 1*), but the analysis of these quantities is substantially more complicated.

2.6 The asymptotic $MISE$ approximation

A problem which arises with the $MISE$ is that it involves several integrals which, in general, would need to be evaluated by Monte Carlo methods. One way of overcoming this problem involves the derivation of its large sample approximation. This approximation admits a very simple expression and allows a deeper appreciation of the role of the bandwidth. It can also be used to obtain the rate

of convergence of the kernel estimator.

The assumptions for the asymptotic approximation of the $MISE$ are (see, for example, *Wand and Jones, 1995, pp. 19–20*):

1. the density $f(x;\boldsymbol{\theta})$ is such that its second derivative $f''(x;\boldsymbol{\theta})$ is continuous, square integrable and ultimately monotone;

2. the kernel $K(x)$ is a bounded probability density function having finite fourth moment and symmetry about the origin;

3. the bandwidth $h = h_n$ is a non-random sequence of positive numbers which approaches zero at a rate slower than n^{-1}.

In this case, *Rosenblatt (1956)* defined the *asymptotic mean integrated squared error* ($AMISE$)

$$AMISE[f_n(\,.\,;h)] = (nh)^{-1}\int K(x)^2\mathrm{d}x + \tfrac{1}{4}h^4\mu_2(K)^2\int f''(x;\boldsymbol{\theta}_0)^2\mathrm{d}x, \qquad (2.14)$$

where $\mu_2(K) = \int x^2K(x)\mathrm{d}x$.

The $AMISE$ is a much simpler quantity to comprehend than the $MISE$. *Silverman (1986, p. 40)* noted that it gives the $MISE$ as the sum of the asymptotic integrated variance and the asymptotic integrated squared bias of the kernel density estimator. The asymptotic integrated variance is proportional to h^{-1}, so for this quantity to decrease one needs to take h to be large. However, taking h large means an increase in the asymptotic integrated squared bias, since this quantity is proportional to h^4. This is known as the variance-bias trade-off and is a mathematical quantification for the role of the bandwidth. Another advantage of the $AMISE$ over the $MISE$ will become apparent as we proceed.

2.7 Optimum $MISE$ weight function

We have already emphasised the reasons for selecting the weight function as a characteristic function. This section is concerned with how we might select a

suitable weight function, which would enable us to exploit the integrated squared error method. As indicated, this problem is in some respects similar to the problem of selecting a suitable kernel in kernel density estimation.

The latter problem has been extensively studied by a number of authors, including *Watson and Leadbetter (1963)*. These authors showed how to derive the kernel that is optimal in the $MISE$ sense for a given density $f(x)$ and sample size n. In particular, they considered the complex variable version of the $MISE$,

$$MISE[f_n(\,.\,)] = \mathrm{E}\int [f_n(x) - f(x)]^2 \mathrm{d}x \tag{2.15}$$

$$= \frac{1}{2\pi}\,\mathrm{E}\int |\phi_{f_n}(t) - \phi_f(t)|^2 \mathrm{d}t, \tag{2.16}$$

and showed that it is minimised with respect to the kernel $K(x;h)$ by taking

$$\phi_K(t) = \frac{n\,|\phi_f(t)|^2}{1 + (n-1)|\phi_f(t)|^2}, \tag{2.17}$$

where the notation $\phi_g(t)$ denotes the characteristic function of $g(x)$.

In the context of kernel density estimation, there are two serious problems associated with the use of (2.17). These are:

1. the integral involved in the Fourier inversion of (2.17) is not always easy to evaluate (see *Abdous, 1993*);

2. the optimum kernel depends on the underlying distribution, which is in theory unknown.

Consequently, (2.17) has made only a limited impact in this field.

However, in the context of integrated squared error estimation, (2.17) need not be inverted. Furthermore, there is *a priori* knowledge of the underlying distribution and so the difficulties above are not a problem. Consequently, we can rewrite the information conveyed by (2.17) in terms of the weight function itself.

Proposition 2.1. *Let $I(\boldsymbol{\theta})$ be the integrated squared error function for a distribution with characteristic function $\phi(t;\boldsymbol{\theta})$, and suppose that $X_1, X_2, \ldots, X_n$ is a*

random sample from this distribution with parameter vector $\boldsymbol{\theta}_0$. *Then the optimum (in MISE sense) weight function is such that*

$$w_{MISE}(t) = \frac{n\,|\phi(t;\boldsymbol{\theta}_0)|^2}{1+(n-1)|\phi(t;\boldsymbol{\theta}_0)|^2}, \tag{2.18}$$

in which case, the minimum MISE value is given by

$$\inf MISE[f_{\varphi_n}(\,.\,)] = \frac{1}{2\pi}\int \frac{|\phi(t;\boldsymbol{\theta}_0)|^2\,[1-|\phi(t;\boldsymbol{\theta}_0)|^2]}{1+(n-1)|\phi(t;\boldsymbol{\theta}_0)|^2}\,\mathrm{d}t. \tag{2.19}$$

2.8 Theoretical difficulties

The approach of Section 2.7 uses the $MISE$ as a criterion to derive the optimum weight function for the integrated squared error method. In general, the success of this approach will depend on:

1. the availability of a good omnibus numerical optimisation package, since closed form solutions will be relatively rare;

2. knowledge of the true parameter value $\boldsymbol{\theta}_0$, which is an impossible situation.

Since point (1) above is nowadays not a major problem, this section is devoted to how we might overcome the second difficulty.

One possibility is to regard $w_{MISE}(t)$ as a function of t and $\boldsymbol{\theta}$, denoted $w_{MISE}(t,\boldsymbol{\theta})$, and minimise the integral

$$J(\boldsymbol{\theta}) = \int_{-\infty}^{\infty} |\phi_n(t) - \phi(t;\boldsymbol{\theta})|^2\,|w_{MISE}(t,\boldsymbol{\theta})|^2\,\mathrm{d}t \tag{2.20}$$

with respect to $\boldsymbol{\theta}$. In this case, however, we will no longer optimise the integrated squared error function but some other more complicated criterion.

Alternatively, we may designate the solutions for $\boldsymbol{\theta}$ from the equations

$$\int_{-\infty}^{\infty} \frac{\partial\,|\phi_n(t)-\phi(t;\boldsymbol{\theta})|^2}{\partial\theta_j}|w_{MISE}(t,\boldsymbol{\theta})|^2\,\mathrm{d}t = 0, \quad j = 1,2,\ldots,p,$$

namely the equations which result from differentiation of (2.20) with respect to $\boldsymbol{\theta}$ while holding the weight function out of the differentiation process. The motivation for this approach revolves partly on the notion of χ^2 minimum procedures. In χ^2 minimum estimation, one subdivides $\mathbb{R}$ into cells $R_1, R_2, \ldots, R_m$ and seeks a choice of parameters which minimises

$$\chi^2(\boldsymbol{\theta}) = \sum_{j=1}^{m} \frac{[N_j - \mathrm{E}_j(\boldsymbol{\theta})]^2}{\mathrm{E}_j(\boldsymbol{\theta})} \tag{2.21}$$

or, more conveniently, which solves the system of equations

$$\sum_{j=1}^{m} \frac{N_j - \mathrm{E}_j(\boldsymbol{\theta})}{\mathrm{E}_j(\boldsymbol{\theta})} \frac{\partial \mathrm{E}_j(\boldsymbol{\theta})}{\partial \theta_k} = 0, \quad k = 1, 2, \ldots, p,$$

obtained from (2.21) by holding the denominator out of the differentiation process (see *Cramér, 1946, pp. 424-428*). In these expressions, N_j and $\mathrm{E}_j(\boldsymbol{\theta})$ are, respectively, the observed and expected number of observations in R_j $(j = 1, 2, \ldots, m)$.

Minimising different distance functions will lead to estimators with generally different properties. We shall not conduct a detailed comparative study here since, as we shall see, point (2) above will not be a problem once additional developments from kernel density estimation have been adopted. First, we have found it useful to apply Proposition 2.1 to an example. We have chosen the negative exponential distribution which, while analytically simple, is believed to illustrate the concepts clearly. In addition the negative exponential will motivate further research. As such, this is a very important example.

2.9 Application to the negative exponential distribution

Let $X_1, X_2, \ldots, X_n$ be independent, identically distributed random variables from a negative exponential distribution, with probability density function

$$f(x; \theta) = \theta \exp(-\theta x), \quad x > 0, \ \theta > 0, \tag{2.22}$$

and suppose that θ_0 is the true but unknown parameter value. The characteristic function corresponding to (2.22) is

$$\phi(t;\theta) = \theta/(\theta - it), \quad \theta > 0, \tag{2.23}$$

and so the integrated squared error estimator, $\hat{\theta}$, for θ_0 minimises

$$I(\theta) = \int_{-\infty}^{\infty} |n^{-1} \sum_{j=1}^{n} e^{itX_j} - \phi(t;\theta)|^2 \, |w(t)|^2 \, \mathrm{d}t \tag{2.24}$$

with respect to θ. As usual, the weight function $w(t)$ is open to choice. In the absence of an obvious choice, it is commonplace to select the normal characteristic function. This was shown to be equivalent to selecting the normal kernel in the kernel estimation of the underlying density. However, *Fryer (1977)* refers to the inadequacy of the normal kernel for the kernel estimation of the negative exponential density. On these grounds, we may concern ourselves with finding an alternative weight function.

The optimum (in $MISE$ sense) weight function may be obtained from Proposition 2.1 and some straightforward computations; we find

$$w_{MISE}(t) = n\theta_0^2/(n\theta_0^2 + t^2). \tag{2.25}$$

This weight function corresponds to the special case of

$$w(t;\lambda) = (1 + \lambda^2 t^2)^{-1} \tag{2.26}$$

in which $\lambda = (n\theta_0^2)^{-1/2}$.

Comparison of (2.26) with the normal weight function (2.7) is not a clear-cut problem since the form of $w(t;\lambda)$ is coupled with the parameter λ. However, we can compare corresponding densities since parameters often have clear geometrical interpretations in this context. The integral involved in the Fourier inversion of

(2.26) may be evaluated analytically to give

$$f_w(x;\lambda) = (2\lambda)^{-1}\exp(-|x|/\lambda), \quad x \in \mathbb{R},\ \lambda > 0. \tag{2.27}$$

This is the double exponential or Laplace density (see, for example, *Johnson, Kotz and Balakrishnan, 1995, Chapter 24*) with mean zero and variance $2\lambda^2$. Figure 2.1 compares the double exponential distribution with the normal distribution with the same mean and variance.

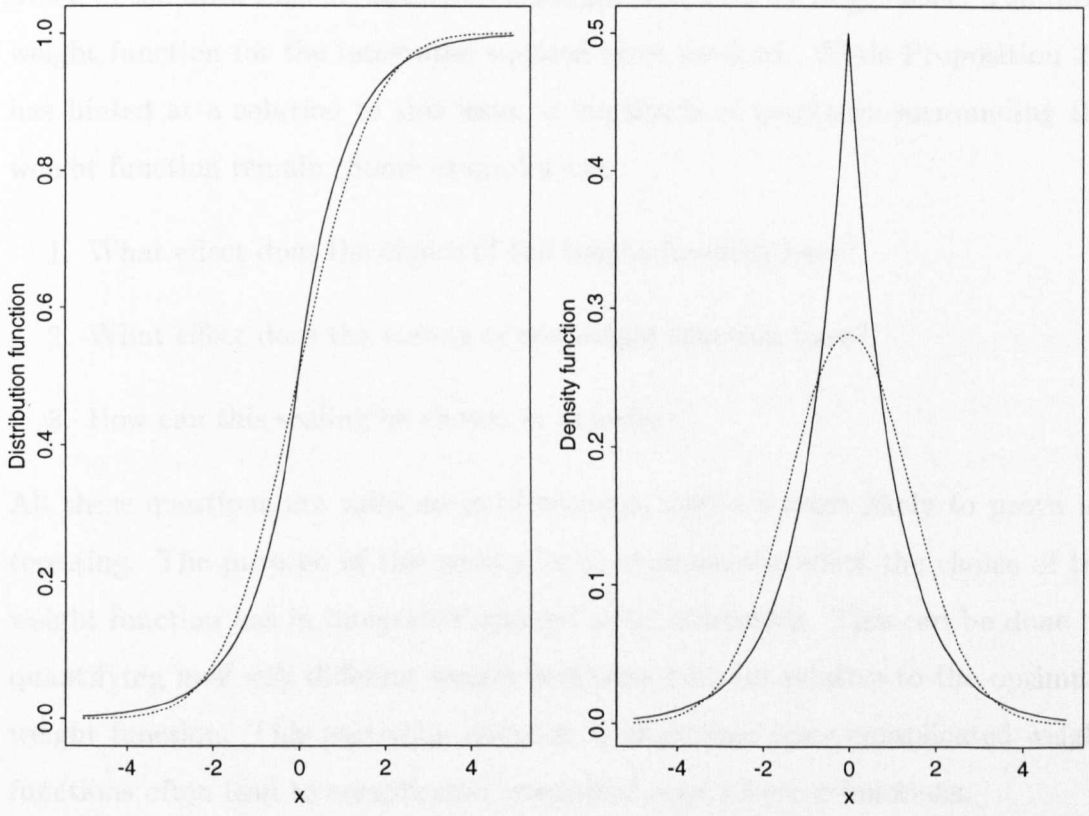

Figure 2.1: The double exponential distribution and density function (solid line) overlaid with the normal distribution and density function (dotted line). The two distributions have the same mean, namely zero, and variance, namely 2.

The most notable difference between the double exponential and normal density functions depicted in the figure is in the dramatic peak of the former. However, this difference is the product of the two distributions having equal variances and could be made comparatively smaller if the variances were allowed to vary. Of

course, this would increase the discrepancy in the tails, but it may be that, at a certain point, a better overall agreement would be obtained. At that point, the agreement between the corresponding characteristic functions would be analogous and so the performances of (2.26) and (2.7) in integrated squared error estimation would be more comparable. This issue will be examined in the following section.

2.10 Optimum weight function theory

Much of the work thus far has been concerned with how we might select a suitable weight function for the integrated squared error method. While Proposition 2.1 has hinted at a solution to this issue, a multitude of questions surrounding the weight function remain. Some examples are:

1. What effect does the choice of the weight function have?
2. What effect does the scaling of the weight function have?
3. How can this scaling be chosen in practice?

All these questions are valid areas of research, and are most likely to prove interesting. The purpose of this section is to examine the effect the choice of the weight function has in integrated squared error estimation. This can be done by quantifying how well different weight functions perform relative to the optimum weight function. This particular question is important since complicated weight functions often lead to complicated integrated squared error functions.

To answer this question we will first derive a useful representation of the $MISE$ that allows a more direct application of this criterion in the context of integrated squared error estimation.

2.10.1 Practical issues

As indicated above, we intend to use the $MISE$ to judge between weight functions in the same way as the $MISE$ is used to judge between kernel density

estimators. One possible problem with this approach is that the $MISE$ depends on the underlying density function which, in our context, either may not exist or may be prohibitively complex in form. In this section we make a suggestion for overcoming this problem.

The motivation necessary for this comes from expanding (2.13) to obtain

$$MISE[f_n(\,.\,;h)] = \mathrm{E}\int f_n(x;h)^2\,\mathrm{d}x - 2\,\mathrm{E}\int f_n(x;h)f(x;\boldsymbol{\theta}_0)\,\mathrm{d}x + \int f(x;\boldsymbol{\theta}_0)^2\,\mathrm{d}x.$$

We may then apply Parseval's theorem, giving

$$MISE[f_n(\,.\,;h)] = \frac{1}{2\pi}[\mathrm{E}\int |\phi_{f_n}(t;h)|^2\,\mathrm{d}t - 2\,\mathrm{E}\int \phi_{f_n}(t;h)\,\overline{\phi(t;\boldsymbol{\theta}_0)}\,\mathrm{d}t + \int |\phi(t;\boldsymbol{\theta}_0)|^2\,\mathrm{d}t], \tag{2.28}$$

where $\overline{\phi(t;\boldsymbol{\theta})}$ is the complex conjugate of $\phi(t;\boldsymbol{\theta})$. Suppose now that the kernel in $f_n(x;h)$ is symmetric about the origin. This is a reasonable and desirable assumption to make in practice, as shown by *Rosenblatt (1956)*. Under this assumption, the characteristic function of the kernel is real-valued. Consequently, changing the order of integration in (2.28) and using (see, for example, *Koutrouvelis, 1980*)

$$\mathrm{E}[|\phi_n(t)|^2] = |\phi(t;\boldsymbol{\theta}_0)|^2 + n^{-1}[1 - |\phi(t;\boldsymbol{\theta}_0)|^2],$$

results in

$$MISE[f_n(\,.\,;h)] = \frac{1}{2\pi}\{\int [(1-n^{-1})\phi_K(t;h)^2 - 2\,\phi_K(t;h) + 1]|\phi(t;\boldsymbol{\theta}_0)|^2\,\mathrm{d}t + \int n^{-1}\phi_K(t;h)^2\,\mathrm{d}t\}. \tag{2.29}$$

The $MISE$ as stated in (2.29) does not explicitly depend on $f(x;\boldsymbol{\theta})$, and so the difficulty above has been overcome. In addition, (2.29) is often much simpler

to apply than (2.13) when dealing with integrated squared error estimation. In this context, we can rewrite the information conveyed by (2.29) in terms of the weight function itself.

Proposition 2.2. *Let $I(\boldsymbol{\theta};\lambda)$ be the integrated squared error function for a distribution with characteristic function $\phi(t;\boldsymbol{\theta})$, and suppose that $w(t;\lambda)$ is a real-valued characteristic function. Then the $MISE$ of the kernel density estimator, $f_{\varphi_n}(x;\lambda)$, in the density representation of $I(\boldsymbol{\theta};\lambda)$ is such that*

$$MISE[f_{\varphi_n}(\,.\,;\lambda)] = \frac{1}{2\pi}\{\int [(1-n^{-1})w(t;\lambda)^2 - 2\,w(t;\lambda) + 1]|\phi(t;\boldsymbol{\theta}_0)|^2\,\mathrm{d}t + \int n^{-1}w(t;\lambda)^2\,\mathrm{d}t\}, \tag{2.30}$$

where n and $\boldsymbol{\theta}_0$ are the sample size and true parameter value respectively.

2.10.2 The basic idea

We can investigate the effect of the choice of the weight function in the following intuitive way. Given a weight function $w(t;\lambda)$, we can measure how well it performs relative to the optimum (in $MISE$ sense) weight function by comparing corresponding $MISE$ values. In particular, our judgement will be based on the difference between the two values.

A simple illustration of this idea is provided by the negative exponential distribution of Section 2.9. The optimum (in $MISE$ sense) weight function was given by (2.25). The minimum $MISE$ value, corresponding to (2.25), may be shown to be

$$\inf MISE[f_{\varphi_n}(\,.\,)] = \frac{(n^{1/2}-1)\theta_0}{2(n-1)}, \tag{2.31}$$

where n and θ_0 are the sample size and true parameter value respectively.

Suppose now that the normal weight function (2.7) was employed instead. The

$MISE$ corresponding to the normal weight function may be shown to be

$$\begin{aligned} MISE[f_{\varphi_n}(.;\lambda)] = (2\sqrt{\pi}n\lambda)^{-1} + \tfrac{1}{2}(1-n^{-1})\,\theta_0\,\exp(\lambda^2\theta_0^2)\,[1-\mathrm{erf}(\lambda\theta_0)] \\ -\,\theta_0\,\exp(\lambda^2\theta_0^2/2)\,[1-\mathrm{erf}(\lambda\theta_0/\sqrt{2})] + \tfrac{1}{2}\theta_0, \end{aligned} \tag{2.32}$$

where $\mathrm{erf}(x)$ denotes the error function

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2)\,\mathrm{d}t.$$

Figure 2.2 depicts (2.31) and (2.32) plotted against λ for $n = 25$ and $\theta_0 = 1$. As observed in the figure, the normal weight function can lead to a range of $MISE$ performance. However, in the region of λ where good performance is obtained, namely $\lambda \approx 0.2$, there is a very good agreement between (2.32) and (2.31).

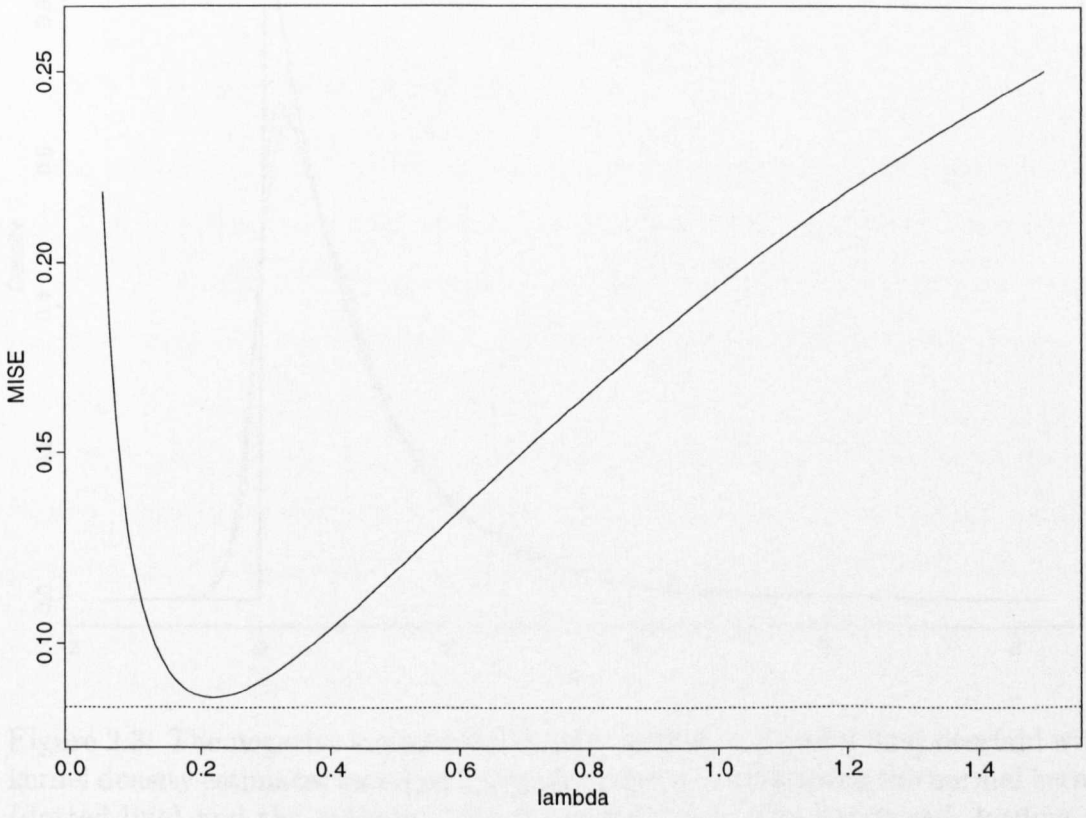

Figure 2.2: The mean integrated squared error associated with the normal weight function (solid line) overlaid with the minimum $MISE$ value (dotted line) for the negative exponential density with $\theta_0 = 1$, $n = 25$.

We conjecture that in this region of λ the performances of the integrated squared error estimators based on (2.25) and (2.7) will be comparable. A simple way to examine this is by comparing the performances of the kernel density estimators in the corresponding integrated squared error functions. This is permissible for two reasons. First, the sample information is processed in the estimation procedure only through the kernel density estimator (2.10). Secondly, if for two different weight functions the left hand sides of (2.10) are comparable, then so are the weight functions and hence the resulting parameter estimates. Figure 2.3 shows the corresponding kernel estimates of $f(x) = \exp(-x)$, $x > 0$ based on a sample of size $n = 1000$. The true density is also shown for comparison.

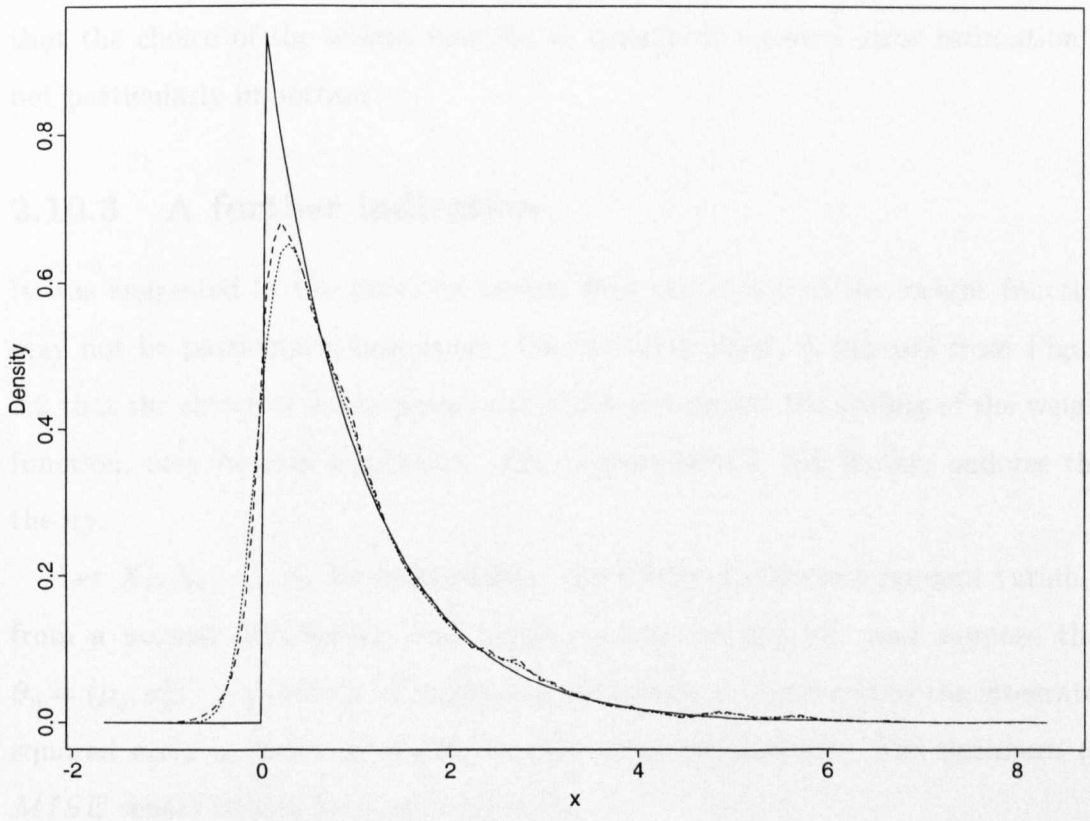

Figure 2.3: The negative exponential density with $\theta_0 = 1$ (solid line) overlaid with kernel density estimates based on a sample of size $n = 1000$ using the normal kernel (dotted line) and the optimum kernel (dashed line). The bandwidth leading to the smallest possible value of (2.32) was used for the normal kernel.

There are two important issues that are apparent from the figure. The first

issue is that neither estimator performs adequately near the origin. This is due to kernel estimator having to find a compromise between estimating the two distinct values of the density on either side of zero. As noted by *Silverman (1986, pp. 29–32)* and *Wand and Jones (1995, pp. 46–49)*, many modifications have been proposed and studied to improve kernel estimation of densities with bounded domains. However, these modifications are not relevant to integrated squared error estimation and hence are not discussed in this thesis. The second issue, which has importance in integrated squared error estimation, is that of the closeness of the two density estimates. This result is consistent with the general view (see, for example, *Wand and Jones, 1995, pp. 28–31*) that the choice of the kernel in kernel density estimation is not particularly important. On this basis, we conjecture that the choice of the weight function in integrated squared error estimation is not particularly important.

2.10.3 A further indication

It was suggested in the previous section that the choice of the weight function may not be particularly important. On the other hand, it appears from Figure 2.2 that the choice of λ, the parameter which determines the scaling of the weight function, may be very important. The present section will further endorse this theory.

Let $X_1, X_2, \ldots, X_n$ be independent, identically distributed random variables from a normal distribution with mean μ_0 and variance σ_0^2, and suppose that $\boldsymbol{\theta}_0 = (\mu_0, \sigma_0^2)^\top$ is unknown. If parameter estimation is to proceed by the integrated squared error method, a weight function must be selected. The optimum (in $MISE$ sense) weight function is given by

$$w_{MISE}(t) = \frac{n\, e^{-\sigma_0^2 t^2}}{1 + (n-1)e^{-\sigma_0^2 t^2}}. \tag{2.33}$$

In view of (2.33), the optimum weight function may be criticised in two ways: first, it depends on the unknown parameter σ_0^2; and second, it gives rise to an integrated

squared error function that can not be evaluated explicitly. Consequently, we may want to investigate alternative choices of weight function. Some possibilities are:

1. the normal characteristic function (2.7);
2. the double exponential characteristic function (2.26);
3. the uniform characteristic function $w(t;\lambda) = \sin(\lambda t)/(\lambda t)$;
4. the triangular characteristic function $w(t;\lambda) = 2[1-\cos(\lambda t)]/(\lambda t)^2$.

In Chapter 1 (Section 1.7) we have seen the application of choice (1) above. We can now examine how this choice compares to the above alternatives. A simple way to do this is by comparing their $MISE$ performance. Table 2.1 presents the minimum $MISE$ value,

$$\inf MISE[f_{\varphi_n}(\,.\,)] = \frac{1}{2\pi}\int \frac{e^{-\sigma_0^2 t^2}(1-e^{-\sigma_0^2 t^2})}{1+(n-1)e^{-\sigma_0^2 t^2}}\,\mathrm{d}t, \tag{2.34}$$

and the smallest possible $MISE$ value for estimating the normal density using the kernels corresponding to the above characteristic functions. The minimising values for λ are also included. The results are based on a sample of size $n = 25$ from the normal distribution with mean $\mu_0 \in \mathbb{R}$ and variances $\sigma_0^2 = 1, 4$.

Table 2.1: Minimum $MISE$ and smallest $MISE$ values for several kernels

Kernel	$\sigma_0^2 = 1$		$\sigma_0^2 = 4$	
	$MISE$	λ	$MISE$	λ
Optimum $MISE$	0.0117	-	0.0059	-
Normal	0.0137	0.6094	0.0069	1.2188
Double exponential	0.0164	0.5081	0.0082	1.0162
Uniform	0.0149	1.0131	0.0075	2.0261
Triangular	0.0134	1.4488	0.0067	2.8969

As anticipated, the smallest possible $MISE$ values are relatively similar, but on the other hand, the minimising values for λ are very different. This result is

consistent with the general view (see, for example, *Wand and Jones, 1995, Chapter 3*) that the choice of the bandwidth in kernel density estimation is particularly important. On this basis, we conjecture that the scaling of the weight function in integrated squared error estimation is also important.

2.11 Choice of value for the parameter λ

It appears from Section 2.10 that the increased flexibility due to the inclusion of the parameter λ allows for considerable freedom of choice in the selection of the weight function. However, this increased flexibility has its costs and leads to new questions. Of these, the most important is how to select a suitable value for this parameter, an issue which is studied below.

2.11.1 Theoretical selection of λ

The approach of Section 2.7 in deriving the optimum weight function was based on minimising the $MISE$. Since the $MISE$ also depends on λ, then λ could be selected so that the $MISE$ is made as small as possible. However, the resulting selector, denoted λ_{MISE}, may be criticised in three ways:

1. the $MISE$ may not be of closed form;
2. an explicit solution for λ_{MISE} will be relatively rare;
3. the value for λ_{MISE} will depend on the unknown parameters $\boldsymbol{\theta}_0$.

The idea of the $AMISE$ approach is to remove the first two of these difficulties. This approach is based on the formula for the asymptotic $MISE$, which is very easy to compute. Further, if h_{AMISE} is the quantity minimising the $AMISE$ with respect to h, then (see, for example, *Wand and Jones, 1995, p. 22*)

$$h_{AMISE} = [\frac{\int K(x)^2 \mathrm{d}x}{n\,\mu_2(K)^2 \int f''(x;\boldsymbol{\theta}_0)^2 \mathrm{d}x}]^{1/5}. \tag{2.35}$$

The analogue of (2.35) for integrated squared error estimation is

$$\lambda_{AMISE} = [\frac{\int f_w^\star(x)^2 \mathrm{d}x}{n\,\mu_2(f_w^\star)^2 \int f''(x;\boldsymbol{\theta}_0)^2 \mathrm{d}x}]^{1/5},$$

where $f_w^\star(x) = \lambda\, f_w(\lambda x;\lambda)$ permits the "decoupling" of $f_w(x;\lambda)$ and λ. Thus, λ_{AMISE} has a closed form expression and so the first two criticisms above have been overcome. The third criticism remains, however, and without knowledge of $\boldsymbol{\theta}_0$, the optimum λ cannot be determined. Some of the suggestions proposed in practice are discussed in the following section. The topic under study then coincides with that of bandwidth selection.

2.11.2 Practical selection of λ

There is currently a vast literature on bandwidth selection and any attempt to present this topic is necessarily of limited scope. This section is no exception. In our selection, we were motivated by the need to have a simple selector that requires very little calculation and a sophisticated selector that aims to give reasonable results each time. Our exposition of this topic is far from exhaustive and is based on *Wand and Jones (1995, Chapter 3).*

One strategy in kernel density estimation is to choose the bandwidth subjectively by eye. This would involve looking at several density estimates and selecting the estimate that is the most suitable in some sense. Although this approach has its advantages, there is much to be said for choosing the bandwidth automatically. This way, its value will not be allowed to depend on x, and ideally it will not depend on the parameters that have to be estimated.

Normal scale bandwidth selector

First and foremost among all bandwidth selectors is the normal scale selector h_{NS}. The normal scale selector evaluates h_{AMISE} at the normal distribution. This yields

the simple formula

$$h_{NS} = [\frac{8\pi^{1/2} \int K(x)\mathrm{d}x}{3n\,\mu_2(K)^2}]^{1/5}\hat{\sigma}, \tag{2.36}$$

where $\hat{\sigma}$ is a robust estimate of the standard deviation (see, for example, *Silverman, 1986, p. 47*). Furthermore, if a normal kernel is being used, then the bandwidth obtained from (2.36) would be

$$h_{NS} = (\frac{4}{3n})^{1/5}\hat{\sigma}.$$

Clearly, the assumption of an underlying normal distribution is potentially dangerous but it may be that, for unimodal distributions at least, h_{NS} gives a useful choice of smoothing parameter, which requires very little calculation. However, for departures from normality such as multi-modality, normal scale selectors will tend to over-smooth and mask important features in the data. As such, the normal scale selector can not be recommended for general use.

Within the past few years, considerable strides have been made in the development of more sophisticated selectors. Most sophisticated selectors fall into one of two categories: the first category houses all the selectors which estimate either the $MISE$ or the $AMISE$ from the data and then locate its minimum; the second category groups the selectors which minimise either criterion theoretically and then estimate this minimising value directly.

These two categories contain a variety of data-driven selectors and this variety can be bewildering. Some insight into the relative merits of competing selectors can be obtained through asymptotic arguments. However, the main tool for the comparison of bandwidth selectors is simulation. In particular, simulation suggests that selectors from the second category are often subject to less variability. The approach we present below takes this issue into account and results in a class of selectors which have been shown by *Sheather and Jones (1991)* to excel in practice.

Plug-in bandwidth selection

Since the earliest days of density estimation, iterative procedures have been proposed in which an estimate of the unknown $\int f''(x;\boldsymbol{\theta}_0)\mathrm{d}x$ is "plugged-in" the formula for the optimum $AMISE$ bandwidth

$$h_{AMISE} = [\frac{\int K(x)^2\mathrm{d}x}{n\,\mu_2(K)^2\int f''(x;\boldsymbol{\theta}_0)^2\mathrm{d}x}]^{1/5}.$$

Scott, Tapia and Thompson (1977) provided an early example of this approach. Since then, a number of improvements have been made. In particular, *Sheather and Jones (1991)*, extending work of *Park and Marron (1990)*, described a selection procedure with excellent properties. This is based on a clever method for estimation of $\int f''(x;\boldsymbol{\theta}_0)\mathrm{d}x$, using a kernel-driven functional. Unfortunately, another selection problem now arises, namely, how do we choose the bandwidth of the auxiliary kernel estimator? This choice can be made using (2.35), although this would lead to yet another selection problem. The usual strategy for overcoming this problem is to use, at some stage, a quick a simple estimate, such as the normal scale selector (2.36). This means that we have a family of plug-in selectors that depend on the number of stages before a quick a simple estimate is used. At times, we shall refer to any of these selectors as h_{PI}.

The plug-in selector h_{PI} is relatively easy to implement and has been recommended by a number of comparative studies, including that of *Jones, Marron and Sheather (1996)*. However, it should be remembered that h_{PI} is based on $AMISE$, which is only a large sample approximation to $MISE$. In many circumstances the minimiser of $AMISE$ is a good approximation to the minimiser of $MISE$, but sometimes it is not, as indicated by *Marron and Wand (1992)*. The following bandwidth selector takes this issue into account.

Smoothed cross-validation bandwidth selection

The smoothed cross-validation approach is similar to the plug-in approach in that it estimates unknown quantities in the criterion which is being optimised. The

difference is that smoothed cross-validation is based on the $MISE$ rather than the $AMISE$. This has the intuitively appealing feature of having less dependence on asymptotic approximations. On the other hand, the smoothed cross-validation selector is not as easy to implement and somewhat more difficult to analyse.

The essential idea is to estimate the unknown $f(x;\boldsymbol{\theta}_0)$ in (2.13) by a kernel estimator using a kernel $L(x)$ and a bandwidth g. The resulting objective function may be shown to be (see, for example, *Wand and Jones, 1995, p. 76*)

$$SCV(h) = (nh)^{-1}\int K(x)^2\,\mathrm{d}x + n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n} l(X_i - X_j; g, h), \tag{2.37}$$

where

$$l(x; g, h) = (K_h * K_h * L_g * L_g - 2K_h * L_g * L_g + L_g * L_g)(x).$$

In this expression $K_h(x) = h^{-1}K(x/h)$, $L_g(x) = g^{-1}L(x/g)$, and the asterisk represents the operation of convolution between the indicated densities.

Unfortunately, the bandwidth minimising (2.37) is not fully automatic, since it depends on the auxiliary kernel $L(x)$ and its bandwidth g. The choice of $L(x)$ is not critical and may be based on grounds of computational convenience. However, the choice of g is, and considerable theory has been devoted to it. This involves precisely the same considerations as were necessary for the selection of the auxiliary bandwidth in h_{PI}. This means that we also have a family of smoothed cross-validation selectors. At times, we shall refer to any of these selectors as h_{SCV}.

Concluding remarks

Any attempt at drawing conclusions or giving decisive recommendations for the choice of bandwidth selector seems destined for failure. Clearly one's choice must depend to a large extent on the application at hand. Taking Wand and Jones's comments as a basis however, my recommendation is to use a version of h_{SCV},

unless of course the smoothed cross-validation objective function is inconvenient to apply. If such a case arises, then a version of h_{PI} would constitute a good alternative choice.

2.12 Application to the negative exponential mixture

The integrated squared error method has thus far been applied to a number of rather straightforward examples. These examples were specifically selected in order to motivate and illustrate the theory of integrated squared error estimation without any unnecessary complications. On the other hand, the simplicity of these examples is self-defeating, since maximum likelihood can be readily implemented. However, there exist more complicated examples for which maximum likelihood cannot be easily implemented. One such example is the finite mixture of normal distributions.

The formal introduction to the family of mixture distributions will be reserved for Chapter 3, where the mixture of two normal distributions will be extensively studied. In this section, attention is concentrated on the mixture of two negative exponential distributions. Such mixtures arise in industrial applications, notably in the analysis of failure time data, and have important mathematical properties.

The probability density function of a mixture of two negative exponential distributions is given by

$$f(x;\boldsymbol{\theta}) = p\,\theta_1 \exp(-\theta_1 x) + (1-p)\,\theta_2 \exp(-\theta_2 x), \quad x > 0, \tag{2.38}$$

where $\boldsymbol{\theta} = (p, \theta_1, \theta_2)^\top$ belongs to the parameter space

$$\boldsymbol{\Theta} = \{0 \leq p \leq 1,\ \theta_1 > 0,\ \theta_2 > 0\}.$$

The characteristic function corresponding to (2.38) is

$$\phi(t;\boldsymbol{\theta}) = p\frac{\theta_1}{\theta_1 - it} + (1-p)\frac{\theta_2}{\theta_2 - it}, \tag{2.39}$$

which, based on a random sample $X_1, X_2, \ldots, X_n$ from (2.38), may be estimated by the empirical characteristic function $\phi_n(t)$.

With a parametric formulation of the negative exponential mixture, a matter of initial concern is the estimation of the parameters. If estimation is to proceed by the integrated squared error method, then a weight function must be selected. The optimum (in $MISE$ sense) weight function is a function of

$$|\phi(t;\boldsymbol{\theta})|^2 = \frac{[p\theta_1 + (1-p)\theta_2]^2 t^2 + \theta_1^2\theta_2^2}{(\theta_1^2 + t^2)(\theta_2^2 + t^2)} \tag{2.40}$$

evaluated at the true parameter value $\boldsymbol{\theta}_0$. Since (2.40) has a complicated form, it follows that the optimum weight function may be impractical. A more convenient choice is the normal weight function (2.7) but there is also something to be said for choosing the double exponential weight function (2.26). This was the optimum (in $MISE$ sense) weight function for the single exponential density, which may be regarded as the special case of the mixture exponential density with either $\theta_1 = \theta_2$ or $p = 1$. This weight function will be adopted for this work.

Before we continue with the resulting integrated squared error function, we may want to examine how the double exponential weight function compares to the optimum weight function. A simple way to do this is by comparing the $MISE$'s of the corresponding kernel density estimators. The minimum $MISE$ value, corresponding to the optimum weight function, is given by

$$\inf MISE[f_{\varphi_n}(\,.\,)] = (2\pi)^{-1}\int \frac{|\phi(t;\boldsymbol{\theta}_0)|^2\,[1 - |\phi(t;\boldsymbol{\theta}_0)|^2]}{1 + (n-1)|\phi(t;\boldsymbol{\theta}_0)|^2}\,\mathrm{d}t. \tag{2.41}$$

The integral which appears in the right hand side of (2.41) can be explicitly integrated but the result is too complicated to give here. The $MISE$ corresponding

to the double exponential weight function is, from Proposition (2.2), given by

$$MISE[f_{\varphi_n}(\,.\,;\lambda)] = (4n\lambda)^{-1} + \boldsymbol{p}^\top[(1-n^{-1})M_2 - 2M_1 + M_0]\boldsymbol{p}, \tag{2.42}$$

where $\boldsymbol{p} = (p, 1-p)^\top$, and M_k $(k = 0, 1, 2)$ are the 2×2 matrices having (i, j) elements equal to

$$\begin{aligned} m_{0;ij} &= \frac{\theta_i\theta_j}{\theta_i + \theta_j} \\ m_{1;ij} &= \frac{\theta_i\theta_j[(\theta_i + \theta_j)\lambda + 2]}{2(\theta_i + \theta_j)(\theta_i\lambda + 1)(\theta_j\lambda + 1)} \\ m_{2;ij} &= \frac{\theta_i\theta_j[(\theta_i\lambda + 2)(\theta_j\lambda + 1)^2 + (\theta_j\lambda + 2)(\theta_i\lambda + 1)^2]}{4(\theta_i + \theta_j)(\theta_i\lambda + 1)^2(\theta_j\lambda + 1)^2}. \end{aligned}$$

The right hand side of (2.42) is evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

Figure 2.4 depicts (2.41) and (2.42) plotted against λ for $p = 0.5$ and: (1) $\theta_1/\theta_2 = 0.01, n = 25$; (2) $\theta_1/\theta_2 = 1, n = 25$; (3) $\theta_1/\theta_2 = 0.01, n = 50$; and (4) $\theta_1/\theta_2 = 1, n = 50$.

As anticipated, the double exponential weight function (2.26) can lead to a range of $MISE$ performance. However, in the region of λ where good performance is obtained, the optimum weight function gains very little. The performance of the double exponential weight function improves as $\theta_1/\theta_2 \to 1$, and this is reasonable. The same can be said as $p \to 1$, but this is not shown here. This illustrates that the double exponential weight function can be used profitably in the integrated squared error estimation of the parameters relating to the mixture of two negative exponential distributions.

We shall therefore proceed to estimate the parameters from consideration of

$$I(\boldsymbol{\theta};\lambda) = \int_{-\infty}^{\infty} |\phi_n(t) - \phi(t;\boldsymbol{\theta})|^2 \,(1+\lambda^2 t^2)^{-2}\,\mathrm{d}t, \tag{2.43}$$

or, equivalently,

$$I(\boldsymbol{\theta};\lambda) = 2\pi \int_{-\infty}^{\infty} [n^{-1}\sum_{j=1}^{n} f_w(x - X_j;\lambda) - f(x;\boldsymbol{\theta}) * f_w(x;\lambda)]^2\,\mathrm{d}x. \tag{2.44}$$

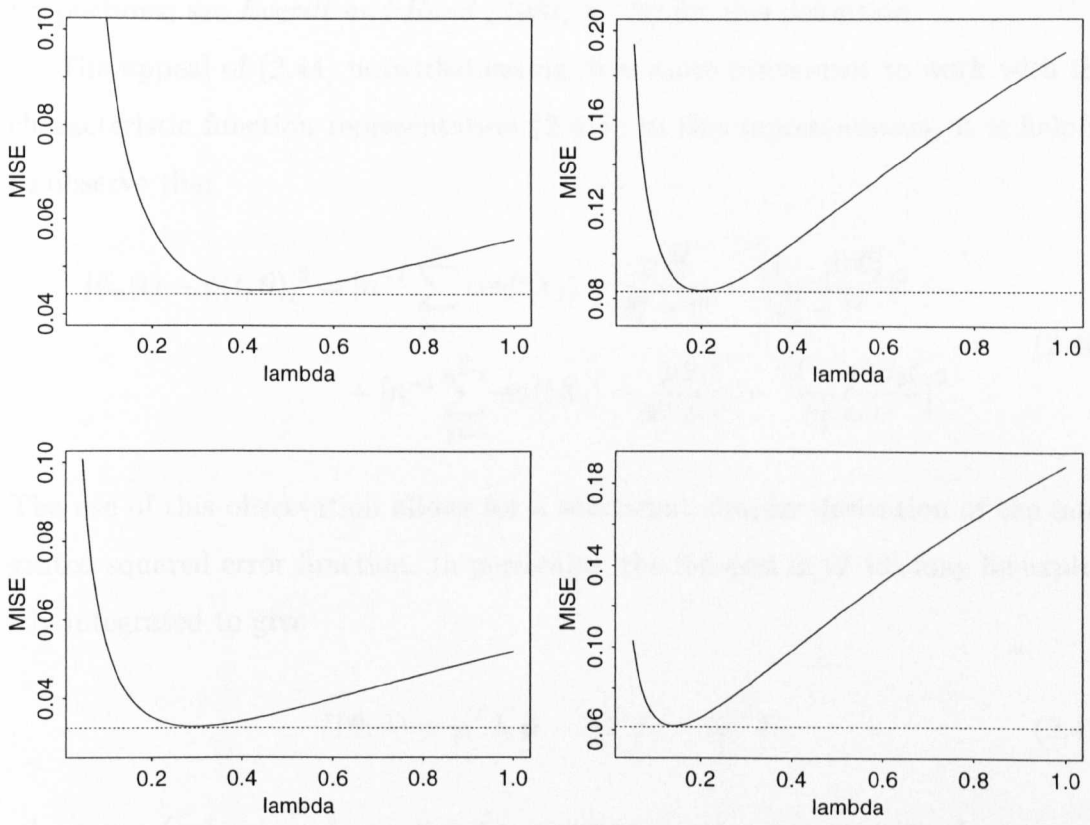

Figure 2.4: The $MISE$ associated with the double exponential weight function (solid line) overlaid with the $MISE$ associated with the optimum weight function (dotted line) for the mixture of two exponential distributions in four different contexts. The contexts are (reading from left-right, top-bottom): (1) $\theta_1/\theta_2 = 0.01, n = 25$; (2) $\theta_1/\theta_2 = 1, n = 25$; (3) $\theta_1/\theta_2 = 0.01, n = 50$; and (4) $\theta_1/\theta_2 = 1, n = 50$. In each case, the mixing proportion was $p = 0.5$.

In the latter representation, the density $f_w(x;\lambda)$ is given by (2.27), while the convolution $f(x;\boldsymbol{\theta}) * f_w(x;\lambda)$ may be shown to be given by

$$f(x;\boldsymbol{\theta}) * f_w(x;\lambda) = p\, g(x;\theta_1,\lambda) + (1-p)\, g(x;\theta_2,\lambda), \quad x \in \mathbb{R}, \tag{2.45}$$

where

$$g(x;\theta,\lambda) = \frac{\theta + \theta\,\mathrm{sgn}(x)}{2(1-\lambda^2\theta^2)} \exp(-\theta\,|x|) - \frac{\theta^2\lambda + \theta\,\mathrm{sgn}(x)}{2(1-\lambda^2\theta^2)} \exp(-\frac{1}{\lambda}|x|).$$

Thus, the density (2.45) may be regarded as a mixture of double exponential distributions, if we slightly generalise the definition of mixtures to permit negative

proportions; see *Everitt and Hand (1981, p. 79)* for this definition.

The appeal of (2.44) notwithstanding, it is more convenient to work with the characteristic function representation (2.43). In this representation, it is helpful to observe that

$$\begin{aligned}|\phi_n(t) - \phi(t;\boldsymbol{\theta})|^2 = [n^{-1}\sum_{j=1}^{n}\cos(tX_j) - \frac{p\,\theta_1^2}{\theta_1^2+t^2} - \frac{(1-p)\,\theta_2^2}{\theta_2^2+t^2}]^2 \\ + [n^{-1}\sum_{j=1}^{n}\sin(tX_j) - \frac{p\,\theta_1 t}{\theta_1^2+t^2} - \frac{(1-p)\,\theta_2 t}{\theta_2^2+t^2}]^2.\end{aligned} \tag{2.46}$$

The use of this observation allows for a somewhat simpler derivation of the integrated squared error function. In particular, the integral in (2.43) may be explicitly integrated to give

$$I(\boldsymbol{\theta};\lambda) \propto \boldsymbol{p}^\top I_0\,\boldsymbol{p} - 2\boldsymbol{p}^\top \boldsymbol{I}_1 - 2\boldsymbol{p}^\top \boldsymbol{I}_2, \tag{2.47}$$

where $\boldsymbol{p} = (p, 1-p)^\top$, I_0 is a 2×2 symmetric matrix whose (i,j)th element is

$$I_{0;ij} = \frac{\theta_i\theta_j[(\theta_i\lambda+2)(\theta_j\lambda+1)^2 + (\theta_j\lambda+2)(\theta_i\lambda+1)^2]}{2(\theta_i+\theta_j)(\theta_i\lambda+1)^2(\theta_j\lambda+1)^2},$$

$\boldsymbol{I}_1$ is a 2×1 vector whose ith element is

$$I_{1;i} = \frac{1}{n}\sum_{j=1}^{n}\{\frac{\theta_i}{(1-\theta_i^2\lambda^2)^2}\exp(-\theta_i X_j) - [\frac{\theta_i^2\lambda}{(1-\theta_i^2\lambda^2)^2} + \frac{\theta_i^2(\lambda+X_j)}{2(1-\theta_i^2\lambda^2)}]\exp(-\frac{1}{\lambda}X_j)\},$$

and $\boldsymbol{I}_2$ is a 2×1 vector whose ith element is

$$I_{2;i} = \frac{1}{n}\sum_{j=1}^{n}\{\frac{\theta_i}{(1-\theta_i^2\lambda^2)^2}\exp(-\theta_i X_j) - [\frac{\theta_i}{(1-\theta_i^2\lambda^2)^2} + \frac{\theta_i X_j}{2\lambda(1-\theta_i^2\lambda^2)}]\exp(-\frac{1}{\lambda}X_j)\}.$$

A system of integrated squared error equations can be formed by differentiating (2.47) with respect to p, θ_1, θ_2, and setting the resulting expressions equal to zero. Explicit solution of these equations is, of course, impossible. It has not been possible to prove that the integrated squared error method must yield unique

parameter estimates, but it has consistently produced reasonable parameter estimates when applied to simulated data.

The practical implementation of the integrated squared error method has required the choice of the parameter λ. This choice is similar to the choice of bandwidth in the kernel density estimation of the negative exponential mixture density. On this basis, the normal scale selector, λ_{NS}, was deemed inappropriate. A more propitious choice would be provided by the smoothed cross-validation selector, λ_{SCV}. However, the smoothed cross-validation selector was not easy to implement. For ease of calculation, the plug-in selector, λ_{PI}, was selected.

2.13 Conclusions

This chapter has used the density representation of the integrated squared error function to address the choice of the weight function in the integrated squared error method. In particular, through this density representation we encountered a link between the methods of integrated squared error and kernel density estimation. This has enabled us to:

1. provide the optimum (in $MISE$ sense) weight function;

2. demonstrate that the choice of the weight function is not important, but the choice of its scaling is;

3. describe how the scaling of the weight function can be selected in practice.

These developments have a wide-ranging applicability and have been demonstrated by a number of examples, including the mixture of two negative exponential distributions.

Chapter 3

Mixtures of normal distributions

3.1 Introduction

Mixtures of distributions have received widespread attention in the statistical literature. This is partly because of interest in their mathematical properties, but mainly because of the considerable number of areas in which they are encountered. Examples date back to the end of the last century with typical areas of application ranging from the study of failure time distributions for electronic valves (see, for example, *Davis, 1952*) to the study of length distributions for fish (see, for example, *Hosmer, 1973*). Applications to other examples in fishery studies, as well as in genetics, medicine, chemistry, psychology and other fields, can be found in *Everitt and Hand (1981)* and *Titterington, Smith and Makov (1985).*

The problem of estimating the parameters in mixture distributions has also been the subject of a large, diverse body of literature. This problem has generally proved not to be straightforward, for two main reasons. First, explicit parameter estimators generally do not exist, so that numerical methods are required. Secondly, practical difficulties which arise in certain aspects of the analysis reveal some "non-standard" theoretical problems. As a result, parameter estimation in mixture distributions requires more than just a straightforward application of conventional methods.

This chapter is concerned with mixtures of normal distributions. In practice,

these are the most widely used mixture distributions. The chapter begins with the parametric formulation of finite mixture distributions, with the related mathematical aspects of identifiability and information being discussed in Section 3.3. The mixture of two normal distributions is then introduced and a substantial account of the currently used methods for the estimation of its parameters follows in the next two sections. Section 3.7 provides a comprehensive account of the theoretical and computational issues in integrated squared error estimation of this mixture. Finally, some consideration is also given to the mixture of k ($k > 2$) normal distributions.

3.2 Finite mixture distributions

Let $\{F(x;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ be a family of distribution functions indexed by the parameter vector $\boldsymbol{\theta}$, and suppose that $X_1, X_2, \ldots, X_n$ is a random sample from some population whose distribution function is a member of this family. Often it is known or suspected that the random sample has arisen from a population Π which is a mixture of a finite number, k, of populations $\Pi_1, \Pi_2, \ldots, \Pi_k$ in some proportions $p_1, p_2, \ldots, p_k$, respectively, where

$$p_i \geq 0 \quad \text{and} \quad \sum_{i=1}^{k} p_i = 1.$$

The probability density function corresponding to Π can therefore be represented in the finite mixture form

$$f(x;\boldsymbol{\theta}) = \sum_{i=1}^{k} p_i g_i(x;\boldsymbol{\theta}_i), \tag{3.1}$$

where $g_i(x;\boldsymbol{\theta}_i)$ is the probability density function corresponding to Π_i, and $\boldsymbol{\theta}_i$ denotes the parameters in the adopted parametric form of this density. The vector

$$\boldsymbol{\theta} = (p_1, p_2, \ldots, p_{k-1}, \boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top, \ldots, \boldsymbol{\theta}_k^\top)^\top$$

of all unknown parameters belongs to some parameter space Θ. Note that, since the proportions p_i $(i = 1, 2, \ldots, k)$ sum to one, one of them is redundant, so we have modified $\boldsymbol{\theta}$ accordingly. We shall refer to the p_i as *mixing proportions*, and the densities $g_i(x; \boldsymbol{\theta}_i)$ as *component densities* of the mixture.

It is straightforward to verify that (3.1) does, indeed, define a probability density function. There is no requirement that the component densities should all belong to the same parametric family, but in most applications this will be the case. The finite mixture density function will then have the form

$$f(x; \boldsymbol{\theta}) = \sum_{i=1}^{k} p_i g(x; \boldsymbol{\theta}_i), \tag{3.2}$$

where $g(x; \boldsymbol{\theta}_i)$ is the probability density function of the parametric family.

In what follows, all of our detailed discussion is directed towards this model. Thus, we shall assume that $X_1, X_2, \ldots, X_n$ are independent, identically distributed random variables with probability density function (3.2). In addition, we shall assume that there is no *a priori* knowledge of the population of origin of each random variable. This should be appropriate in many situations in practice.

3.3 Mathematical aspects of mixtures

With a parametric formulation of finite mixture models, we digress to two topics which have to do with the general well-posedness of estimation problems rather than with any particular method of estimation. The first topic, identifiability, addresses the theoretical question of whether it is possible to uniquely estimate a parameter from a sample, however large. The second topic, information, relates to the practical matter of how good one can reasonably hope for an estimate to be. A thorough survey of these topics is far beyond the scope of this thesis; we try however to cover those aspects of them which have a specific bearing on the sequel.

3.3.1 Identifiability

In general, the family of density functions $\{f(x;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ is said to be *identifiable* if distinct values of $\boldsymbol{\theta}$ determine distinct members of the family. This can be interpreted as follows in the case where $f(x;\boldsymbol{\theta})$ defines a family of finite mixture densities according to (3.2). A family of finite mixtures is said to be identifiable for $\boldsymbol{\theta} \in \Theta$ if for any two members $f(x;\boldsymbol{\theta}) = \sum_{i=1}^{k} p_i g(x;\boldsymbol{\theta}_i)$ and $f(x;\boldsymbol{\theta}') = \sum_{i=1}^{k'} p'_i g(x;\boldsymbol{\theta}'_i)$, then

$$f(x;\boldsymbol{\theta}) \equiv f(x;\boldsymbol{\theta}')$$

if and only if $k = k'$ and there is a permutation π of $(1, 2, \ldots, k)$ such that $p_i = p'_{\pi(i)}$ and, if $p_i \neq 0$, $\boldsymbol{\theta}_i = \boldsymbol{\theta}'_{\pi(i)}$ for $i = 1, 2, \ldots, k$. Here $\equiv$ implies equality of the densities for almost all x relative to the underlying measure on $\mathbb{R}$ appropriate for $f(x;\boldsymbol{\theta})$.

Titterington, Smith and Makov (1985, pp. 35–42) have given a coherent account of the concept of identifiability for finite mixtures, including theorems which establish the identifiability for mixtures of normal, gamma, and other continuous distributions. A second more recent reference is *Sapatinas (1995)*.

3.3.2 Information

In general, the Fisher information matrix for an observation X_1 with density $f(x;\boldsymbol{\theta})$ is given by

$$\mathcal{I}(\boldsymbol{\theta}) = \mathrm{E}\{[\frac{\partial}{\partial\boldsymbol{\theta}} \log f(X_1;\boldsymbol{\theta})][\frac{\partial}{\partial\boldsymbol{\theta}} \log f(X_1;\boldsymbol{\theta})]^\top\},$$

provided that this expression exists. (In writing $\partial/\partial\boldsymbol{\theta}$, we suppose that $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)^\top$, and we take $\partial/\partial\boldsymbol{\theta} = (\partial/\partial\theta_1, \partial/\partial\theta_2, \ldots, \partial/\partial\theta_p)^\top$.) The Fisher information matrix has general significance concerning the distribution of unbiased and asymptotically unbiased estimators. For the present purposes, the importance of the Fisher information matrix lies in its role in determining the asymptotic distribution of maximum likelihood estimators.

A number of authors have considered the Fisher information matrix for finite mixture densities in a variety of contexts. *Hill (1963)* gave a general power series expansion of the Fisher information about the mixing proportion in a univariate mixture of two normal or negative exponential distributions. *Behboodian (1972)* presented a method for the numerical calculation of the information matrix for a mixture of two univariate normal distributions with arbitrary variances, while *Tan and Chang (1972)* considered the equal variance case in deriving the asymptotic relative efficiency of the moment estimator. For two univariate normal populations with no restrictions on the variances, *Hosmer and Dick (1977)* studied the information matrix, where in addition to the unclassified observations, there were also some observations of known origin.

3.4 Mixtures of two normal distributions

In practice, the most widely used finite mixture distributions are those involving normal components. Of these, the mixture involving two components is most commonly used.

In its most general form, the mixture of two normal distributions is defined by the probability density function

$$f(x;\boldsymbol{\theta}) = p\, g(x;\boldsymbol{\theta}_1) + (1-p)\, g(x;\boldsymbol{\theta}_2), \quad x \in \mathbb{R}, \tag{3.3}$$

where $g(x;\boldsymbol{\theta}_j) = (2\pi\sigma_j^2)^{-1/2}\exp[-(x-\mu_j)^2/(2\sigma_j^2)]$ is the density of the normal distribution with mean μ_j and standard deviation σ_j. In this case, $\boldsymbol{\theta}_1 = (\mu_1, \sigma_1)^\top$, $\boldsymbol{\theta}_2 = (\mu_2, \sigma_2)^\top$ and $\boldsymbol{\theta} = (p, \mu_1, \sigma_1, \mu_2, \sigma_2)^\top$. The parameter space associated with the mixture of two normal distributions is given by

$$\Theta = \{0 \le p \le 1,\ |\mu_1| < \infty,\ \sigma_1 > 0,\ |\mu_2| < \infty,\ \sigma_2 > 0\}.$$

The characteristic function corresponding to (3.3) is given by

$$\phi(t;\boldsymbol{\theta}) = p\,\psi(t;\boldsymbol{\theta}_1) + (1-p)\,\psi(t;\boldsymbol{\theta}_2), \tag{3.4}$$

where $\psi(t;\boldsymbol{\theta}_j) = \exp(it\mu_j - \sigma_j^2 t^2/2)$ is the characteristic function of the normal distribution with mean μ_j and standard deviation σ_j.

The mixture of two normal distributions can provide a flexible framework for approximating distributions which are not well-modelled by any standard parametric family. This flexibility is illustrated in Figure 3.1. The introduction of more components would, of course, be expected to yield a better approximation and, in fact, this feature is exploited in kernel density estimation.

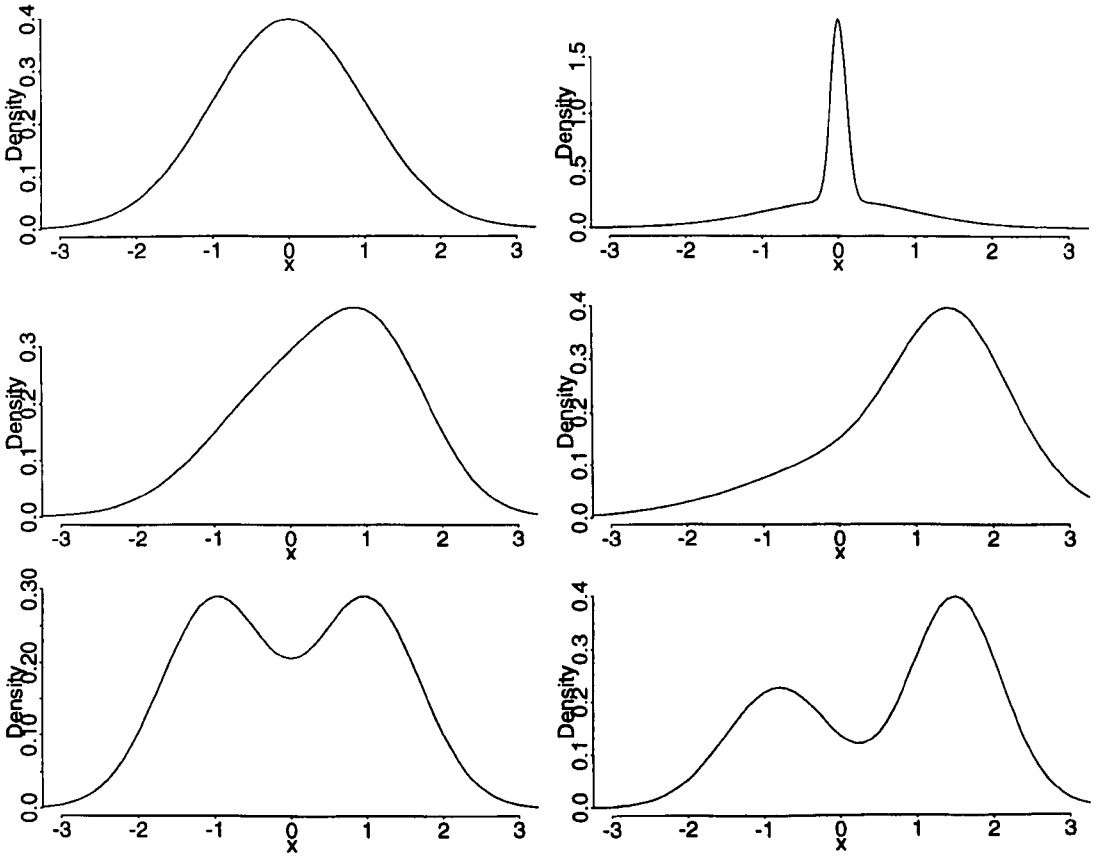

Figure 3.1: A number of shapes the normal mixture density (3.3) can take. The shapes are (reading from left-right, top-bottom): (1) Gaussian; (2) kurtotic; (3) unimodal; (4) skewed; (5) bimodal; and (6) asymmetric.

The six densities of the figure have been selected because they represent a

variety of possible forms of (3.3). The most striking feature of the density corresponding to a mixture of two normal distributions is often that of bimodality. The figure, however, illustrates that a mixture of two normals, differing in means, can still be unimodal. The values for the parameters of these densities are given in Table 3.1. For ease in plotting, these have been chosen so that the densities diminish outside $x \in [-3, 3]$.

Table 3.1: Parameters for the six example density functions of Figure 3.1

Density type	$\boldsymbol{\theta}$				
	p	μ_1	σ_1	μ_2	σ_2
Gaussian	1	0	1	-	-
Kurtotic	0.6	0	1	0	0.1
Unimodal	0.6	0	1	1.2	0.7
Skewed	0.5	0.6	1.5	1.5	0.7
Bimodal	0.5	−1	0.7	1	0.7
Asymmetric	0.4	−0.8	0.7	1.5	0.6

3.5 Estimation in the mixture of two normal distributions

With a parametric formulation of the mixture of two normal distributions, a matter of initial concern is the estimation of the parameters. This is one of the oldest problems in the statistical literature, dating back to *Pearson (1894)*. Since then, a remarkable variety of estimation methods have been applied.

The following is an outline summary of some of the methods which have been considered. This summary is not intended to be exhaustive, but it is hoped that it will provide some perspective in which to view the remainder of this chapter. Additional details on the methods reviewed below as well as on other less well-known methods can be found in the monographs by *Everitt and Hand (1981)*, *Titterington, Smith and Makov (1985)* and *McLachlan and Basford (1988)*.

3.5.1 The method of moments

Estimation using the method of moments is often regarded as the starting point of the analysis of mixtures. *Pearson (1894)* used this method in one of the earliest studies of the mixture of two normal distributions.

The method of moments involves equating the first five sample moments given by

$$m_r = n^{-1} \sum_{i=1}^{n} (X_i - \overline{X})^r, \quad r = 1, 2, \ldots, 5$$

to their theoretical counterparts

$$\mu_r = \int (x - \mu)^r \mathrm{d}F(x; \boldsymbol{\theta}), \quad r = 1, 2, \ldots, 5,$$

where $\overline{X} = \sum_{i=1}^{n} X_i$ and $\mu = \mathrm{E}(X_1)$. By some considerable algebraic manipulation, these equations may be reduced to the famous "nonic", originally derived by *Pearson (1894)*. This takes the form

$$\sum_{i=0}^{9} a_i y^i = 0, \tag{3.5}$$

where

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

and where $k_4 = m_4 - 3m_2^2$ and $k_5 = m_5 - 10 m_2 m_3$ are the fourth and fifth sample cumulants respectively. If a solution to the system of moment equations exists, then it may be obtained as follows.

Let y be a real negative root of (3.5); then calculate

$$\rho = \frac{-8m_3y^3 + 3k_5y^2 + 6m_3k_4y + 2m_3^3}{y(2y^3 + 3k_4y + 4m_3^2)},$$

and solve the quadratic equation $\delta^2 - \rho\delta + y = 0$, giving roots

$$\delta_1 = [\rho - (\rho^2 - 4y)^{1/2}]/2$$
$$\delta_2 = [\rho + (\rho^2 - 4y)^{1/2}]/2.$$

The moment estimates of the five mixture parameters may now be computed from

$$\left.\begin{aligned} \hat{p} &= \delta_2/(\delta_2 - \delta_1) \\ \hat{\mu}_j &= \delta_j + \overline{X} \\ \hat{\sigma}_j &= [\frac{1}{3}\delta_j(2\rho - m_3/y) + m_2 - \delta_j^2]^{1/2} \end{aligned}\right\} \tag{3.6}$$

for $j = 1, 2$.

The most attractive feature of the method of moments is its relative ease of application. Indeed, the method of moments was usually the method of choice until the arrival of high-speed computing. However, this method may be criticised in three ways:

1. it is based on the sample cumulants k_4 and k_5 which are not unbiased estimators for the corresponding population cumulants (see, for example, *Bryant and Paulson, 1982*);

2. it is based on high order sample moments which have poor sampling properties;

3. it may give rise to non-unique parameter estimates, or fail to give any at all.

These problems have led investigators to consider alternative estimation methods.

3.5.2 Graphical methods

As speculated by *Fowlkes (1979)*, it was probably because of the problems of the moment estimators and the absence of modern computing technology that attention was turned to various graphical methods. The majority of these methods were attempts to obtain crude parameter estimates, but there was a great need for some kind of solution, however crude.

There are two main types of plot for univariate data, depending on whether the density function or distribution function is being depicted. In particular, of course, the former plots include the histogram and the latter the normal quantile-quantile or Q-Q plot. The best of these graphical methods can be found in *Titterington, Smith and Makov (1985, pp. 52–71)*, but nonetheless, they may still be criticised in two ways:

1. they require relatively large sample sizes;

2. they are generally unlikely to lead to accurate parameter estimates.

Consequently investigators have not highly valued graphical methods.

3.5.3 The method of maximum likelihood

With the arrival of high-speed computing, attention was turned to the method of maximum likelihood. The customary approach to determining a maximum likelihood estimate is first to derive a system of equations (called the *likelihood equations*) which are satisfied by the maximum likelihood estimate, and then to obtain the maximum likelihood estimate by solving these equations. The likelihood equations are found by differentiating the logarithm of the likelihood function with respect to the parameters, and setting the derivatives equal to zero.

In the present mixture context, the logarithm of the likelihood function is given by

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(X_i; \boldsymbol{\theta}),$$

where $f(x;\boldsymbol{\theta})$ is the density function (3.3). The likelihood equations are complicated in form (see, for example, *Everitt and Hand, 1981, p. 36*) and beyond hope of solution by analytic means. Consequently, one must resort to seeking an approximate solution via some iterative procedure.

There are, of course, many general iterative procedures for numerically approximating maximum likelihood estimates. Our main interest here, however, is in a special iterative method which can be interpreted as an application of the EM algorithm of *Dempster, Laird and Rubin (1977)*. This algorithm proceeds in two steps, E for expectation and M for maximisation, and has certain desirable theoretical properties by its very definition (see, for example, *Meng and vanDyk, 1997*). The application of the EM algorithm to the present mixture context can be described as follows.

Let

$$\boldsymbol{\theta}^{(0)} = (p^{(0)}, \mu_1^{(0)}, \sigma_1^{(0)}, \mu_2^{(0)}, \sigma_2^{(0)})^\top$$

be a set of initial parameter estimates, and suppose that $\boldsymbol{\theta}^{(m)}$ are the parameter estimates at the mth iteration. At this iteration, define the weights

$$w_{i1}(\boldsymbol{\theta}^{(m)}) = \frac{p^{(m)} g(X_i; \boldsymbol{\theta}_1^{(m)})}{f(X_i; \boldsymbol{\theta}^{(m)})}, \quad i = 1, 2, \ldots, n$$

$$w_{i2}(\boldsymbol{\theta}^{(m)}) = \frac{(1 - p^{(m)}) g(X_i; \boldsymbol{\theta}_2^{(m)})}{f(X_i; \boldsymbol{\theta}^{(m)})}, \quad i = 1, 2, \ldots, n,$$

and compute the sums

$$n_j^{(m)} = \sum_{i=1}^{n} w_{ij}(\boldsymbol{\theta}^{(m)}), \quad j = 1, 2.$$

Then, the parameter estimates at iteration $m+1$ may be obtained by

$$\left.\begin{aligned} p^{(m+1)} &= \frac{n_1^{(m)}}{n} \\ \mu_j^{(m+1)} &= \frac{1}{n_j^{(m)}} \sum_{i=1}^{n} w_{ij}(\boldsymbol{\theta}^{(m)}) X_i \\ \sigma_j^{(m+1)} &= [\frac{1}{n_j^{(m)}} \sum_{i=1}^{n} w_{ij}(\boldsymbol{\theta}^{(m)})(X_i - \mu_j^{(m+1)})^2]^{1/2} \end{aligned}\right\} \tag{3.7}$$

for $j = 1, 2$. These iterations are continued until some suitable convergence criterion has been satisfied.

However, in defining a maximum likelihood estimate in this way, we have failed to consider two technical difficulties associated with likelihood estimation for mixtures of normal distributions. These are:

1. the likelihood function is not bounded above in Θ;

2. the likelihood equations will generally have multiple roots.

Point (1) above was perhaps first noted by *Kiefer and Wolfowitz (1956)*, who observed that if one of the component means coincides with a sample observation and the corresponding variance tends to zero, then the likelihood function increases without bound. However, as remarked in *McLachlan and Bashford (1988, p. 38)*, this is not a problem since the essential aim of likelihood estimation is to find a sequence of roots of the likelihood equations which is consistent, and hence efficient if the usual regularity conditions hold. *Redner and Walker (1984)* verified that for a mixture of univariate normal distributions there is a sequence of roots with the desired asymptotic properties, but there is, of course, the problem of identifying it. This means that in many cases it will be difficult to justify choosing one set of parameter estimates above another. Furthermore, the extent to which these results hold in small samples is a contentious issue. These problems have encouraged the development of a Bayesian approach to the problem.

3.5.4 The Bayesian approach

The general Bayesian approach to estimating an unknown parameter vector $\boldsymbol{\theta} \in \Theta$ is very easily described. If $f(\boldsymbol{X}; \boldsymbol{\theta})$ denotes the probability density function of the sample $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$, Bayes theorem provides the mechanism whereby beliefs about $\boldsymbol{\theta}$ *prior* to observing $\boldsymbol{X}$, expressed as a density $p(\boldsymbol{\theta})$, are updated into beliefs about $\boldsymbol{\theta}$ *posterior* to observing $\boldsymbol{X}$, denoted by $p(\boldsymbol{\theta} \mid \boldsymbol{X})$ and given by

$$p(\boldsymbol{\theta} \mid \boldsymbol{X}) = \frac{p(\boldsymbol{\theta}) f(\boldsymbol{X}; \boldsymbol{\theta})}{\int_{\Theta} p(\boldsymbol{\theta}) f(\boldsymbol{X}; \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}}.$$

In general, all aspects of the posterior distribution are valid quantities for inference. However, in most cases quantities of interest require the evaluation of integrals of the form

$$\int_{\Theta} h(\boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{X}) \, \mathrm{d}\boldsymbol{\theta},$$

where $h(\boldsymbol{\theta})$ is a function of the parameters.

In theory the Bayesian approach appears to have many advantages over the method of maximum likelihood. In practice it presents new problems of its own. Some of these problems are:

1. choosing a suitable prior;

2. carrying out efficient numerical integration in several dimensions.

A recent discussion of these problems, together with details of some proposed strategies, is given in *Diebolt and Robert (1994)*; see also *Escobar and West (1995)*. We shall not enter into detail here, since, once a numerical integration strategy such as Markov chain Monte Carlo integration is invoked, a finite mixture model is simply a special case of a Bayesian inference problem.

3.6 Robust estimation in the mixture of two normal distributions

We shall now consider for the mixture of two normal distributions some robust estimation methods whereby observations assessed as atypical of a component or the mixture itself are automatically given a reduced weight in the computation of the parameter estimates.

The problem of robust estimation in a general estimation context has been a major concern in recent statistical literature. Minimum distance estimation has been shown by *Parr and Schucany (1980)* to provide a reasonable mode of attack for this problem. To reiterate the working definition given in Chapter 1 (Section 1.4), the basic philosophy of minimum distance estimation is to estimate the parameters by minimising some specified distance measure over the parameter space.

There is clearly a wide range of estimators that can derived by this method, depending on the choice of the distance measure. The choice leading to the greatest degree of mathematical tractability is

$$\Delta[K(\,.\,), L(\,.\,)] = \int_{-\infty}^{\infty} |K(x) - L(x)|^2 \, \mathrm{d}W(x), \tag{3.8}$$

where $K(x)$, $L(x)$ and $W(x)$ are suitably selected functions. Concerning the choice of the functions $K(x)$ and $L(x)$, the published literature contains references to two major categories, namely distributions functions and integral transforms.

3.6.1 Minimum distance estimation based on distribution functions

Minimising distance measures between distribution functions is the most common application of minimum distance estimators. In the present mixture context, the

vector of unknown parameters, $\boldsymbol{\theta}$, is estimated by minimising

$$\Delta[F_n(\,.\,), F(\,.\,;\boldsymbol{\theta})],$$

the distance between the empirical distribution function $F_n(x)$ and the mixture distribution function $F(x;\boldsymbol{\theta})$. For a particular form of the weight function $W(x)$, *Woodward, Parr, Schucany and Lindsey (1984)* estimated the mixing proportion, with the unknown remaining parameters being viewed as incidental parameters. The distance used by these authors is closely related to that of *Choi and Bulgren (1968)*, in which a non-decreasing step weight function was used. Thus, *Choi and Bulgren (1968)* approximated the integral in (3.8) by a sum.

The usefulness of this type of approximation was addressed by *Everitt and Hand (1981, p. 20)*. They noted that these approximations would cease to be approximate if the data were grouped. This practice leads, of course, to a considerable loss of information so minimising an approximate distance is, perhaps, not the best approach to adopt in all cases. Consequently, there is considerable scope for investigating alternative distance measures.

Minimising a distance between the empirical and theoretical distribution functions is an intuitive concept. However, in view of the one-to-one correspondence between the distribution function and integral transforms thereof, minimum distance estimation may also be based on the latter. This approach is taken on by the second category of minimum distance methods.

3.6.2 Minimum distance estimation based on integral transforms

Perhaps the most critical aspect of the construction of a minimum distance method based on integral transforms is the choice of transform. There is considerable freedom of choice, but transforms such as the characteristic function and moment generating function are most common in practice. Early applications favoured the algebraic simplicity of the moment generating function. For example, *Quandt and*

Ramsey (1978) proposed estimating the parameters of the mixture by minimising

$$S(\boldsymbol{\theta}) = \sum_{i=1}^{q}[n^{-1}\sum_{j=1}^{n} e^{t_i X_j} - pe^{\mu_1 t_i + \sigma_1^2 t_i^2/2} - (1-p)e^{\mu_2 t_i + \sigma_2^2 t_i^2/2}]^2, \qquad (3.9)$$

the sum of squared deviations between the empirical and theoretical moment generating functions evaluated at points t_j $(j = 1, 2, \ldots, q)$. Obviously crucial in this approach is the choice of these points. *Quandt and Ramsey (1978)* suggested that very large or very small values of t should be avoided and that q should be not less that five. Their choice in the examples they gave was $q = 5$ with t values -0.2, -0.1, 0.1, 0.2, 0.3.

As in *Choi and Bulgren (1968)*, *Quandt and Ramsey (1978)* used (3.8) with a step weight function. This type of weight function gives rise to a very flexible estimation method, which will be discussed in detail in Chapter 5. However, in the present mixture context there are numerical complexities. We repeated the simulation experiments performed by *Quandt and Ramsey (1978)* and found that the resulting parameter estimates were extremely sensitive to:

1. the values of the points t_j $(j = 1, 2, \ldots, 5)$;

2. the iteration starting point used in the minimisation of (3.9).

These results are consistent with the simulations of *Kumar, Nicklin and Paulson (1979)* and *Everitt and Hand (1981, pp. 53–56).*

Nevertheless, the choice of t_j $(j = 1, 2, \ldots, q)$ was taken up by *Schmidt (1982).* He showed that a more efficient estimator than that of *Quandt and Ramsey (1978)* could be obtained by minimising a generalised, instead of a simple, sum of squares. He also pointed out that by increasing q, one can, not surprisingly, eliminate the role played by the particular choice of the t_j $(j = 1, 2, \ldots, q)$. However, as q is increased indefinitely, it was suggested in Chapter 1 (Section 1.14) that minimum distance estimation based on moment generating functions was inferior to that based on characteristic functions. The remainder of this chapter, therefore, will be directed towards minimum distance methods based on characteristic functions.

In this case, the distance in question becomes none other than the integrated squared error function.

3.7 The method of integrated squared error

The integrated squared error estimation of the parameters in the two-component normal mixture appears to have been neglected in the literature. *Binder (1978)* and *Clarke and Heathcote (1978)* merely referred to it as an alternative to the method of *Quandt and Ramsey (1978). Kumar, Nicklin and Paulson (1979)* provided a minimum distance method based on characteristic functions, but the particular choice of objective function needs justification, especially with respect to the double role played by p. Perhaps the most systematic application of the integrated squared error method to the present mixture context may be found in *Bryant and Paulson (1983).* However, as the authors admit, the problem considered therein is overly simplistic in that the component distributions are taken to be completely specified and attention focuses only on the mixing proportion.

3.7.1 The integrated squared error estimator

In a general estimation context, the integrated squared error estimator, $\hat{\boldsymbol{\theta}}$, for $\boldsymbol{\theta}$ is the value of $\boldsymbol{\theta}$ which minimises

$$I(\boldsymbol{\theta}) = \int_{-\infty}^{\infty} |\phi_n(t) - \phi(t;\boldsymbol{\theta})|^2 |w(t)|^2 \mathrm{d}t,$$

where $w(t)$ is a suitably selected weight function. In the present mixture context, the optimum (in $MISE$ sense) weight function is, from Proposition 2.1, given by

$$w_{MISE}(t) = \frac{n\,|\phi(t;\boldsymbol{\theta}_0)|^2}{1 + (n-1)|\phi(t;\boldsymbol{\theta}_0)|^2}, \tag{3.10}$$

where n is the sample size, $\boldsymbol{\theta}_0$ is the true parameter value, and

$$|\phi(t;\boldsymbol{\theta})|^2 = p^2 e^{-\sigma_1^2 t^2} + 2p(1-p)e^{-(\sigma_1^2+\sigma_2^2)t^2/2}\cos[(\mu_1-\mu_2)t] + (1-p)^2 e^{-\sigma_2^2 t^2}. \tag{3.11}$$

Since (3.11) has a complicated form, it follows that the optimum weight function (3.10) may be impractical. An alternative choice is provided by the normal characteristic function

$$w(t;\lambda) = \exp(-\tfrac{1}{2}\lambda^2 t^2). \tag{3.12}$$

This weight function was considered by *Bryant and Paulson (1983)* in the estimation of the mixing proportion.

It is instructive to examine the $MISE$ performance of (3.12) relative to that of (3.10). The minimum $MISE$ value, corresponding to (3.10), is, from Proposition 2.1, given by

$$\inf MISE[f_{\varphi_n}(\,.\,)] = \frac{1}{2\pi}\int \frac{|\phi(t;\boldsymbol{\theta}_0)|^2\,[1-|\phi(t;\boldsymbol{\theta}_0)|^2]}{1+(n-1)|\phi(t;\boldsymbol{\theta}_0)|^2}\,\mathrm{d}t. \tag{3.13}$$

The $MISE$ corresponding to (3.12) is given by (see *Marron and Wand, 1992*)

$$MISE[f_{\varphi_n}(\,.\,;\lambda)] = (2\pi^{1/2}n\lambda)^{-1} + \boldsymbol{p}^\top[(1-n^{-1})M_2 - 2M_1 + M_0]\boldsymbol{p}, \tag{3.14}$$

where $\boldsymbol{p} = (p, 1-p)^\top$, and M_ℓ ($\ell = 0, 1, 2$) are the 2×2 matrices having (i,j) elements equal to

$$m_{\ell;ij} = [2\pi(\sigma_i^2 + \sigma_j^2 + \ell\lambda^2)]^{-1/2}\exp[-\tfrac{1}{2}(\mu_i-\mu_j)^2/(\sigma_i^2+\sigma_j^2+\ell\lambda^2)].$$

The right hand side of (3.14) should, of course, be evaluated at the true parameter values $\boldsymbol{\theta}_0$.

Figure 3.2 depicts (3.13) and (3.14) plotted against λ for the six normal mixture densities of Figure 3.1 when $n = 50$.

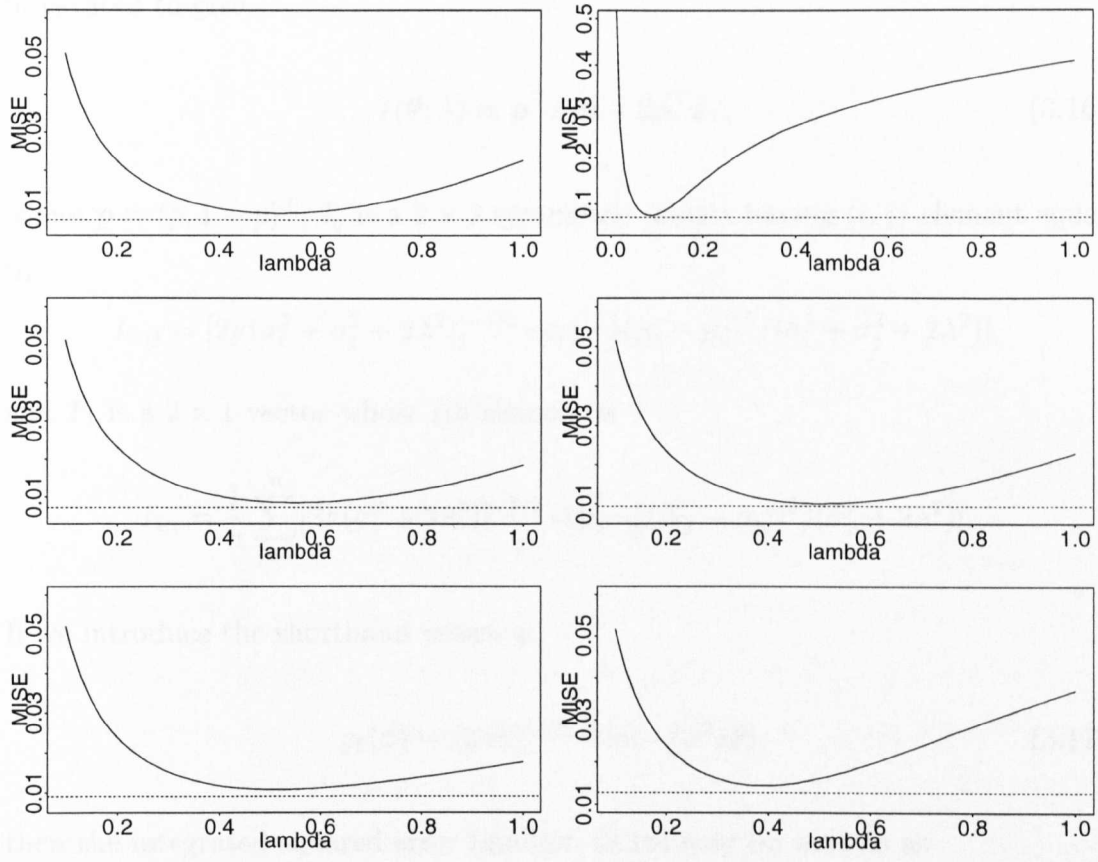

Figure 3.2: The $MISE$ (3.14) (solid line) overlaid with the minimum $MISE$ value (3.13) (dotted line) for six normal mixture densities when $n = 50$. The densities are the (reading from left-right, top-bottom): (1) Gaussian; (2) kurtotic; (3) unimodal; (4) skewed; (5) bimodal; and (6) asymmetric densities of Figure 3.1.

As anticipated, the weight function (3.12) leads to a range of $MISE$ performance. However, in the region of λ where good performance is obtained, the optimum weight function gains very little. This result is relatively independent of the parameter values and is consistent with the general view of Chapter 2. On this basis, we proceed to estimate the normal mixture parameters by minimising

$$I(\boldsymbol{\theta}; \lambda) = \int_{-\infty}^{\infty} |\phi_n(t) - \phi(t; \boldsymbol{\theta})|^2 \exp(-\lambda^2 t^2)\, \mathrm{d}t \tag{3.15}$$

with respect to $\boldsymbol{\theta}$.

The integral which appears in the left hand side of (3.15) may be explicitly

integrated to give

$$I(\boldsymbol{\theta};\lambda) \propto \boldsymbol{p}^\top I_0 \boldsymbol{p} - 2\boldsymbol{p}^\top \boldsymbol{I}_1, \tag{3.16}$$

where $\boldsymbol{p} = (p, 1-p)^\top$, I_0 is a 2×2 symmetric matrix having (i,j) element equal to

$$I_{0;ij} = [2\pi(\sigma_i^2 + \sigma_j^2 + 2\lambda^2)]^{-1/2} \exp[-\tfrac{1}{2}(\mu_i - \mu_j)^2/(\sigma_i^2 + \sigma_j^2 + 2\lambda^2)],$$

and $\boldsymbol{I}_1$ is a 2×1 vector whose ith element is

$$I_{1;i} = \frac{1}{n}\sum_{j=1}^{n}[2\pi(\sigma_i^2 + 2\lambda^2)]^{-1/2} \exp[-\tfrac{1}{2}(X_j - \mu_i))^2/(\sigma_i^2 + 2\lambda^2)].$$

If we introduce the shorthand notation

$$g_\theta(x) = (2\pi\theta)^{-1/2} \exp(-\tfrac{1}{2}x^2/\theta), \tag{3.17}$$

then the integrated squared error function (3.16) may be written as

$$\begin{aligned} I(\boldsymbol{\theta};\lambda) \propto\ & p^2 g_{2\sigma_1^2+2\lambda^2}(0) + 2p(1-p)g_{\sigma_1^2+\sigma_2^2+2\lambda^2}(\mu_1 - \mu_2) \\ & + (1-p)^2 g_{2\sigma_2^2+2\lambda^2}(0) - 2p\frac{1}{n}\sum_{j=1}^{n} g_{\sigma_1^2+2\lambda^2}(X_j - \mu_1) \\ & - 2(1-p)\frac{1}{n}\sum_{j=1}^{n} g_{\sigma_2^2+2\lambda^2}(X_j - \mu_2). \end{aligned} \tag{3.18}$$

A system of integrated squared error equations can be formed by differentiating (3.18) with respect to the parameters and setting the resulting expressions equal to zero. This gives

$$\begin{aligned} \frac{\partial I}{\partial p} =\ & 2p g_{2\sigma_1^2+2\lambda^2}(0) + 2(1-2p)g_{\sigma_1^2+\sigma_2^2+2\lambda^2}(\mu_1 - \mu_2) - 2(1-p)g_{2\sigma_2^2+2\lambda^2}(0) \\ & - 2\frac{1}{n}\sum_{j=1}^{n} g_{\sigma_1^2+2\lambda^2}(X_j - \mu_1) + 2\frac{1}{n}\sum_{j=1}^{n} g_{\sigma_2^2+2\lambda^2}(X_j - \mu_2) = 0 \end{aligned}$$

$$\frac{\partial I}{\partial \mu_i} = (-1)^i\, 2p(1-p)\frac{\mu_1-\mu_2}{\sigma_1^2+\sigma_2^2+2\lambda^2} g_{\sigma_1^2+\sigma_2^2+2\lambda^2}(\mu_1-\mu_2)$$
$$- 2p_i \frac{1}{\sigma_i^2+2\lambda^2}\frac{1}{n}\sum_{j=1}^{n}(X_j-\mu_i) g_{\sigma_i^2+2\lambda^2}(X_j-\mu_i) = 0$$
$$\frac{\partial I}{\partial \sigma_i} = -p_i^2 \frac{\sigma_i}{\sigma_i^2+\lambda^2} g_{\sigma_i^2+\lambda^2}(0)$$
$$- 2p(1-p)\frac{\sigma_i}{\sigma_1^2+\sigma_2^2+2\lambda^2}\left[1-\frac{(\mu_1-\mu_2)^2}{\sigma_1^2+\sigma_2^2+2\lambda^2}\right] g_{\sigma_1^2+\sigma_2^2+2\lambda^2}(\mu_1-\mu_2)$$
$$+ 2p(1-p)\frac{\sigma_i}{\sigma_i^2+2\lambda^2}\frac{1}{n}\sum_{j=1}^{n}\left[1-\frac{(X_j-\mu_i)^2}{\sigma_i^2+2\lambda^2}\right] g_{\sigma_i^2+2\lambda^2}(X_j-\mu_i) = 0$$

for $i = 1, 2$, where $p_1 = p$ and $p_2 = 1-p$.

Explicit solution of these equations is, of course, impossible. Consequently, one must resort to seeking an approximate solution via some iterative procedure. There are many general iterative procedures which are suitable for this purpose. We have in mind here the Newton-Raphson method, various quasi-Newton methods, and conjugate gradient methods. Alternatively, we may consider abandoning the integrated squared equations altogether and minimising the integrated squared error function directly. One method which is suitable for this purpose, and which we describe in detail in Chapter 4, is simulated annealing.

3.7.2 The robustness properties of the estimator

As emphasised in the beginning of Section 3.6, we are considering a class of estimation methods which possess good robustness properties. The customary way of examining the robustness of an estimator is through the behaviour of its influence function. In short, the influence function describes the response of the estimator to an additional observation. An additional observation in our framework will be denoted by ξ. Then, since the regularity conditions of Chapter 1 (Section 1.6.1) are satisfied (see, for example, *Heathcote, 1977*), the joint influence function for the integrated squared error estimator, $\hat{\boldsymbol{\theta}}$, is, from Theorem 1.3, given by

$$IF(\xi; \hat{\boldsymbol{\theta}}) = K^{-1}(\boldsymbol{\theta})\, \boldsymbol{\tau}(\xi; \boldsymbol{\theta}). \tag{3.19}$$

In this expression, $K(\boldsymbol{\theta})$ is the 5×5 symmetric matrix with elements

$$\kappa_{11}(\boldsymbol{\theta}) = (\frac{\pi}{\sigma_1^2 + \lambda^2})^{1/2} - 4\pi g_{\sigma_1^2+\sigma_2^2+2\lambda^2}(\mu_1 - \mu_2) + (\frac{\pi}{\sigma_2^2 + \lambda^2})^{1/2},$$

$$\kappa_{12}(\boldsymbol{\theta}) = p\frac{2\pi(\mu_1 - \mu_2)}{\sigma_1^2 + \sigma_2^2 + 2\lambda^2} g_{\sigma_1^2+\sigma_2^2+2\lambda^2}(\mu_1 - \mu_2),$$

$$\kappa_{13}(\boldsymbol{\theta}) = p\frac{2\pi\sigma_1[(\sigma_1^2 + \sigma_2^2 + 2\lambda^2) - (\mu_1 - \mu_2)^2]}{(\sigma_1^2 + \sigma_2^2 + 2\lambda^2)^2} g_{\sigma_1^2+\sigma_2^2+2\lambda^2}(\mu_1 - \mu_2)$$
$$- p\frac{\pi^{1/2}\sigma_1}{2(\sigma_1^2 + \lambda^2)^{3/2}},$$

$$\kappa_{14}(\boldsymbol{\theta}) = (1-p)\frac{2\pi(\mu_1 - \mu_2)}{\sigma_1^2 + \sigma_2^2 + 2\lambda^2} g_{\sigma_1^2+\sigma_2^2+2\lambda^2}(\mu_1 - \mu_2),$$

$$\kappa_{15}(\boldsymbol{\theta}) = (1-p)\frac{2\pi\sigma_2[(\mu_1 - \mu_2)^2 - (\sigma_1^2 + \sigma_2^2 + 2\lambda^2)]}{(\sigma_1^2 + \sigma_2^2 + 2\lambda^2)^2} g_{\sigma_1^2+\sigma_2^2+2\lambda^2}(\mu_1 - \mu_2)$$
$$+ (1-p)\frac{\pi^{1/2}\sigma_2}{2(\sigma_2^2 + \lambda^2)^{3/2}},$$

$$\kappa_{22}(\boldsymbol{\theta}) = p^2\frac{\pi^{1/2}}{2(\sigma_1^2 + \lambda^2)^{3/2}},$$

$$\kappa_{23}(\boldsymbol{\theta}) = 0,$$

$$\kappa_{24}(\boldsymbol{\theta}) = p(1-p)\frac{2\pi[(\sigma_1^2+\sigma_2^2+2\lambda^2)-(\mu_1-\mu_2)^2]}{(\sigma_1^2+\sigma_2^2+2\lambda^2)^2} g_{\sigma_1^2+\sigma_2^2+2\lambda^2}(\mu_1-\mu_2),$$

$$\kappa_{25}(\boldsymbol{\theta}) = p(1-p)\frac{2\pi\sigma_2(\mu_1-\mu_2)[3(\sigma_1^2+\sigma_2^2+2\lambda^2)-(\mu_1-\mu_2)^2]}{(\sigma_1^2+\sigma_2^2+2\lambda^2)^3} g_{\sigma_1^2+\sigma_2^2+2\lambda^2}(\mu_1-\mu_2),$$

$$\kappa_{33}(\boldsymbol{\theta}) = p^2\frac{3\pi^{1/2}\sigma_1^2}{4(\sigma_1^2+\lambda^2)^{5/2}},$$

$$\kappa_{34}(\boldsymbol{\theta}) = p(1-p)\frac{2\pi\sigma_1(\mu_1-\mu_2)[(\mu_1-\mu_2)^2-3(\sigma_1^2+\sigma_2^2+2\lambda^2)]}{(\sigma_1^2+\sigma_2^2+2\lambda^2)^3} g_{\sigma_1^2+\sigma_2^2+2\lambda^2}(\mu_1-\mu_2),$$

$$\kappa_{35}(\boldsymbol{\theta}) = p(1-p)\frac{2\pi\sigma_1\sigma_2}{(\sigma_1^2+\sigma_2^2+2\lambda^2)^4}[12\lambda^2(\sigma_1^2+\sigma_2^2+\lambda^2)+3(\sigma_1^2+\sigma_2^2)^2+(\mu_1-\mu_2)^4$$
$$-6(\mu_1-\mu_2)^2(\sigma_1^2+\sigma_2^2+2\lambda^2)]g_{\sigma_1^2+\sigma_2^2+2\lambda^2}(\mu_1-\mu_2),$$

$$\kappa_{44}(\boldsymbol{\theta}) = (1-p)^2\frac{\pi^{1/2}}{2(\sigma_2^2+\lambda^2)^{3/2}},$$

$$\kappa_{45}(\boldsymbol{\theta}) = 0,$$

$$\kappa_{55}(\boldsymbol{\theta}) = (1-p)^2\frac{3\pi^{1/2}\sigma_2^2}{4(\sigma_2^2+\lambda^2)^{5/2}},$$

and $\boldsymbol{\tau}(\xi;\boldsymbol{\theta})$ is the 5×1 vector with elements

$$\begin{aligned}\tau_1(\xi;\boldsymbol{\theta}) &= 2\pi g_{\sigma_1^2+2\lambda^2}(\xi-\mu_1) - 2\pi g_{\sigma_2^2+2\lambda^2}(\xi-\mu_2) - p\left(\frac{\pi}{\sigma_1^2+\lambda^2}\right)^{1/2} \\ &\quad + (1-p)\left(\frac{\pi}{\sigma_2^2+\lambda^2}\right)^{1/2} - 2(1-2p)\pi g_{\sigma_1^2+\sigma_2^2+2\lambda^2}(\mu_1-\mu_2), \end{aligned} \tag{3.20}$$

$$\begin{aligned}\tau_2(\xi;\boldsymbol{\theta}) &= p(1-p)\frac{2\pi(\mu_1-\mu_2)}{\sigma_1^2+\sigma_2^2+2\lambda^2} g_{\sigma_1^2+\sigma_2^2+2\lambda^2}(\mu_1-\mu_2) \\ &\quad + p\frac{2\pi(\xi-\mu_1)}{\sigma_1^2+2\lambda^2} g_{\sigma_1^2+2\lambda^2}(\xi-\mu_1), \end{aligned} \tag{3.21}$$

$$\begin{aligned}\tau_3(\xi;\boldsymbol{\theta}) &= p(1-p)\frac{2\pi\sigma_1[(\sigma_1^2+\sigma_2^2+2\lambda^2)-(\mu_1-\mu_2)^2]}{(\sigma_1^2+\sigma_2^2+2\lambda^2)^2} g_{\sigma_1^2+\sigma_2^2+2\lambda^2}(\mu_1-\mu_2) \\ &\quad + p\frac{2\pi\sigma_1[(\xi-\mu_1)^2-(\sigma_1^2+2\lambda^2)]}{(\sigma_1^2+2\lambda^2)^2} g_{\sigma_1^2+2\lambda^2}(\xi-\mu_1) \\ &\quad + p^2\frac{\pi^{1/2}\sigma_1}{2(\sigma_1^2+\lambda^2)^{3/2}}, \end{aligned} \tag{3.22}$$

$$\begin{aligned}\tau_4(\xi;\boldsymbol{\theta}) &= p(1-p)\frac{2\pi(\mu_2-\mu_1)}{\sigma_1^2+\sigma_2^2+2\lambda^2} g_{\sigma_1^2+\sigma_2^2+2\lambda^2}(\mu_1-\mu_2) \\ &\quad + (1-p)\frac{2\pi(\xi-\mu_2)}{\sigma_2^2+2\lambda^2} g_{\sigma_2^2+2\lambda^2}(\xi-\mu_2), \end{aligned} \tag{3.23}$$

$$\begin{aligned}\tau_5(\xi;\boldsymbol{\theta}) &= p(1-p)\frac{2\pi\sigma_2[(\sigma_1^2+\sigma_2^2+2\lambda^2)-(\mu_1-\mu_2)^2]}{(\sigma_1^2+\sigma_2^2+2\lambda^2)^2} g_{\sigma_1^2+\sigma_2^2+2\lambda^2}(\mu_1-\mu_2) \\ &\quad + (1-p)\frac{2\pi\sigma_2[(\xi-\mu_2)^2-(\sigma_2^2+2\lambda^2)]}{(\sigma_2^2+2\lambda^2)^2} g_{\sigma_2^2+2\lambda^2}(\xi-\mu_2) \\ &\quad + (1-p)^2\frac{\pi^{1/2}\sigma_2}{2(\sigma_2^2+\lambda^2)^{3/2}}. \end{aligned} \tag{3.24}$$

The difficulty of obtaining a symbolic expression for the joint influence function (3.19) is immediately apparent; deriving (3.19) requires the inversion of the algebraic matrix $K(\boldsymbol{\theta})$. This is, of course, not an elementary calculation, and when the structure of $K(\boldsymbol{\theta})$ is also considered, the calculation becomes impractical. Nevertheless, we can investigate the influence behaviour of $\hat{\boldsymbol{\theta}}$ in the following

intuitive way.

Let $\kappa^{ij}(\boldsymbol{\theta})$ be the (i,j)th element of $K^{-1}(\boldsymbol{\theta})$. Then it is clear from (3.19) that the ith $(i = 1, 2, \ldots, 5)$ individual influence function for the integrated squared error estimator is given by

$$IF(\xi; \hat{\theta}_i) = \sum_{j=1}^{5} \kappa^{ij}(\boldsymbol{\theta}) \tau_j(\xi; \boldsymbol{\theta}), \tag{3.25}$$

where $\tau_j(\xi; \boldsymbol{\theta})$ has, from equations (3.20)–(3.24), the form

$$\tau_j(\xi; \boldsymbol{\theta}) = A_j(\boldsymbol{\theta}) + \sum_{k=1}^{2} e^{-B_k(\boldsymbol{\theta})(\xi - \mu_k)^2} [C_{jk}(\boldsymbol{\theta})(\xi - \mu_k)^2 + D_{jk}(\boldsymbol{\theta})(\xi - \mu_k) + E_{jk}(\boldsymbol{\theta})] \tag{3.26}$$

for suitably selected functions $A_j(\boldsymbol{\theta}), B_k(\boldsymbol{\theta}), C_{jk}(\boldsymbol{\theta}), D_{jk}(\boldsymbol{\theta})$ and $E_{jk}(\boldsymbol{\theta})$. Substituting (3.26) into (3.25) results in

$$IF(\xi; \hat{\theta}_i) = A_i^{\star}(\boldsymbol{\theta}) + \sum_{j=1}^{2} e^{-B_j(\boldsymbol{\theta})(\xi - \mu_j)^2} P_{ij}(\xi; \boldsymbol{\theta}), \tag{3.27}$$

where $A_i^{\star}(\boldsymbol{\theta})$ is a function of $\boldsymbol{\theta}$, and $P_{ij}(\xi; \boldsymbol{\theta})$ is a second-order polynomial in ξ. This implies that

$$\lim_{|\xi| \to \infty} IF(\xi; \hat{\theta}_i) = A_i^{\star}(\boldsymbol{\theta}), \tag{3.28}$$

since $B_j(\boldsymbol{\theta}) = \frac{1}{2}(\sigma_j^2 + 2\lambda^2)^{-1} > 0$. The integrated squared error estimator is thus robust against outliers.

We now illustrate the robustness of the integrated squared error estimator by evaluating (3.19) at a particular parameter set. We have selected the set $\boldsymbol{\theta} = (0.4, -0.8, 0.7, 1.5, 0.6)^{\top}$, which gives rise to the asymmetric density of Figure 3.1. The individual influence functions for $\hat{\boldsymbol{\theta}}$ are depicted graphically in Figures 3.3 and 3.4 below.

Figure 3.3 exhibits the influence function for $\hat{p}$ with $\lambda = \frac{1}{2}, 1, \frac{3}{2}$. As expected,

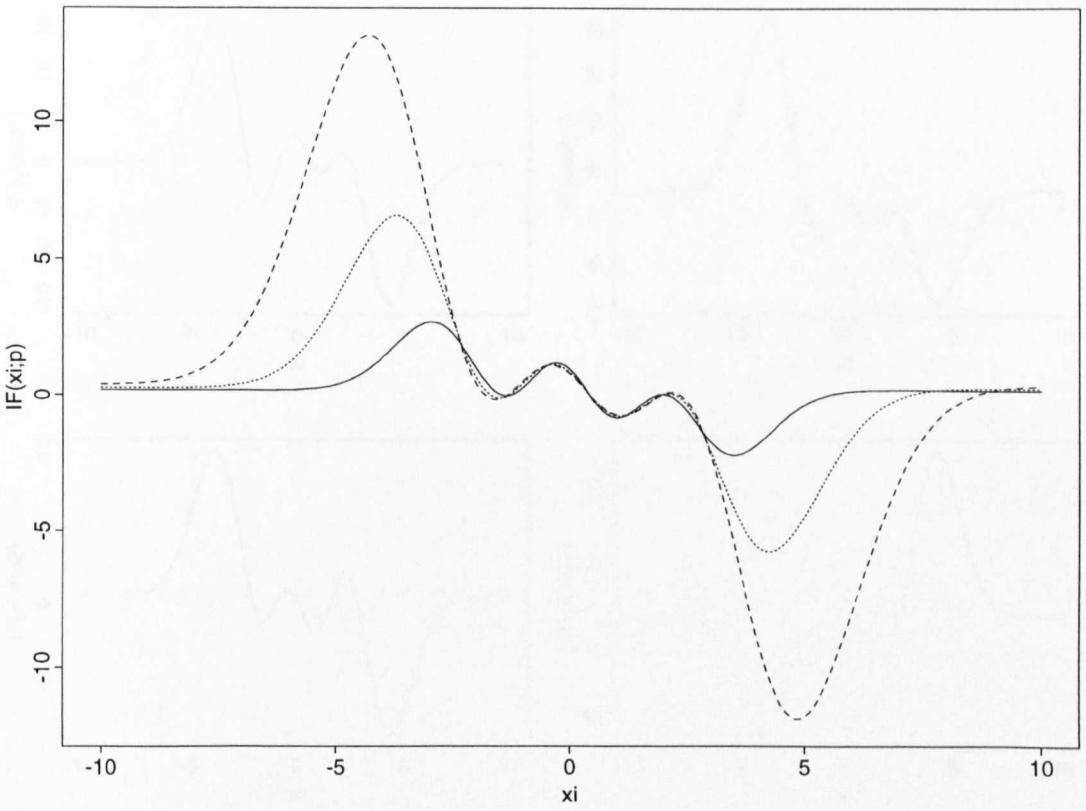

Figure 3.3: The influence function for the integrated squared error estimator $\hat{p}$ with $\lambda = 1/2$ (solid line), $\lambda = 1$ (dotted line), and $\lambda = 3/2$ (dashed line). The influence function is evaluated at the asymmetric distribution of Table 3.1.

the estimator $\hat{p}$ is robust against outliers. However, the estimator becomes increasingly robust as λ decreases. As we shall see, this is due to the connection between integrated squared error and kernel density estimation. Meantime, we can explain this result as follows. If ξ can take values on the whole real line, then the behaviour of the influence function for $\hat{p}$ depends critically on the functions $B_1(\boldsymbol{\theta})$ and $B_2(\boldsymbol{\theta})$. In particular, large values of these functions lead to influence functions which degenerate into their asymptotes (3.28) much earlier than small values. The aforementioned influence behaviour of $\hat{p}$ follows since the functions $B_1(\boldsymbol{\theta})$ and $B_2(\boldsymbol{\theta})$ are inversely proportional to λ.

The argument above holds, of course, for the whole set of influence functions. Consequently, we concentrate on a single value for λ. In Figure 3.4 we plot the influence functions for the remaining integrated squared error estimators with

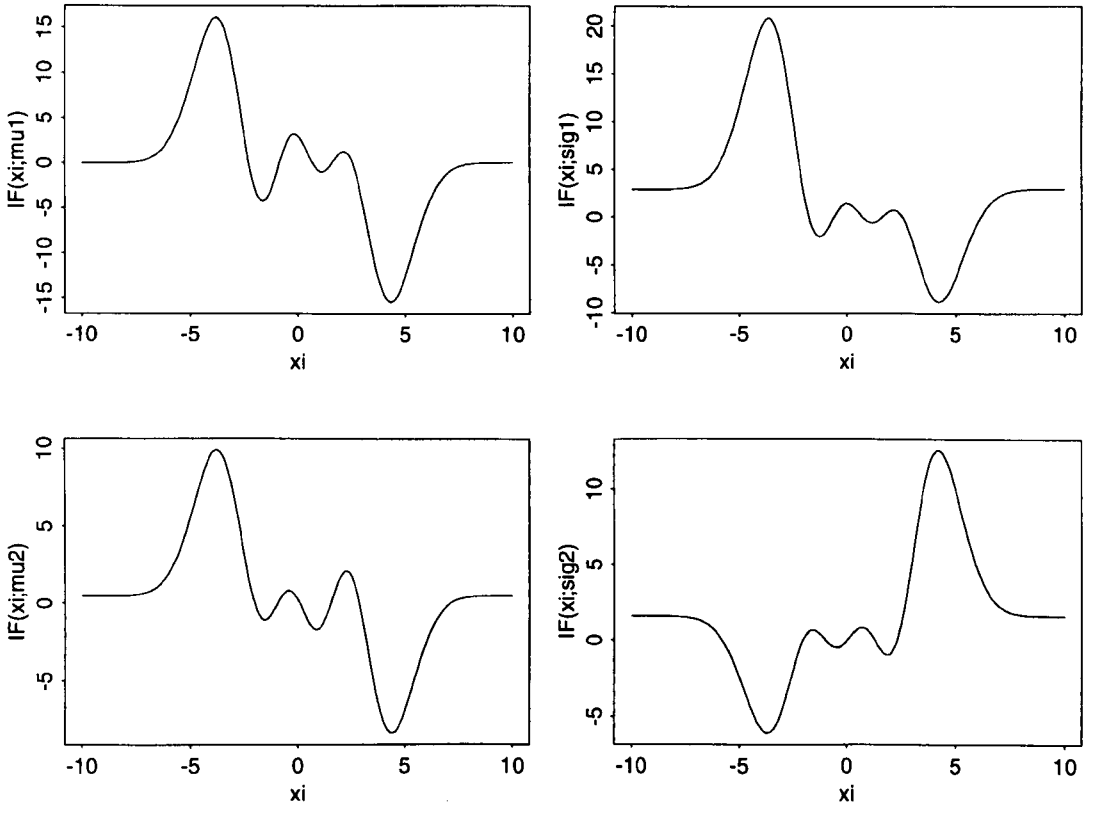

Figure 3.4: The influence functions for (reading from left-right, top-bottom): (1) $\hat{\mu_1}$; (2) $\hat{\sigma_1}$; (3) $\hat{\mu_2}$; and (4) $\hat{\sigma_2}$ with $\lambda = 1$. These influence functions are evaluated at the asymmetric distribution of Table 3.1.

$\lambda = 1$. The conclusions from these figures are entirely consistent with those from Figure 3.3. This illustrates the robustness of the integrated squared error estimator.

3.7.3 The asymptotic properties of the estimator

The asymptotic properties of the integrated squared error estimator in a general estimation context were presented in Chapter 1 (Section 1.6). The present mixture context is simply a special case of this. Thus, since the regularity conditions of Chapter 1 (Section 1.6.1) are satisfied (see, for example, Section 3.7.2), the integrated squared error estimator, $\hat{\boldsymbol{\theta}}$, for $\boldsymbol{\theta}$ is strongly consistent and

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \Sigma(\boldsymbol{\theta})],$$

where

$$\Sigma(\boldsymbol{\theta}) = K^{-1}(\boldsymbol{\theta})\,\Omega(\boldsymbol{\theta})\,K^{-1}(\boldsymbol{\theta}). \tag{3.29}$$

In this expression, $K(\boldsymbol{\theta})$ is the 5×5 symmetric matrix as stated in (3.19), and $\Omega(\boldsymbol{\theta})$ is the covariance matrix of the random variables $\tau_i(X_1; \boldsymbol{\theta})$, $i = 1, 2, \ldots, 5$ as stated in equations (3.20)–(3.24). The derivation of $\Sigma(\boldsymbol{\theta})$ has therefore been reduced to the calculation of

$$\Omega(\boldsymbol{\theta}) = \mathrm{E}[\boldsymbol{\tau}(X_1; \boldsymbol{\theta})\,\boldsymbol{\tau}(X_1; \boldsymbol{\theta})^\top]. \tag{3.30}$$

The expectations in (3.30) are in principle straightforward to evaluate but the results will be very lengthy. It practice, it may be more convenient to approximate these expectations by

$$\Omega_n(\boldsymbol{\theta}) = n^{-1} \sum_{i=1}^{n} \boldsymbol{\tau}(X_i; \boldsymbol{\theta})\,\boldsymbol{\tau}(X_i; \boldsymbol{\theta})^\top. \tag{3.31}$$

The empirical approximation to the asymptotic covariance matrix, denoted $\Sigma_n(\boldsymbol{\theta})$, is obtained by substituting (3.31) for (3.30) into (3.29).

We have found it useful to compare $\Sigma_n(\boldsymbol{\theta})$ and the negative of the Hessian matrix

$$H(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\partial^2 \log f(X_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^\top},$$

which provides an estimate of the information matrix for a random sample, $n\mathcal{I}(\boldsymbol{\theta})$. For example, for a sample of $n = 1000$ from the asymmetric distribution of Table 3.1, we have estimated the asymptotic efficiency of $\hat{\boldsymbol{\theta}}$ using

$$\widehat{eff}(\hat{\boldsymbol{\theta}}) = \frac{\det[-nH^{-1}(\boldsymbol{\theta})]}{\det[\Sigma_n(\boldsymbol{\theta})]}.$$

These efficiencies are provided in Table 3.2 for selected values of λ. As observed in the table, a value of $\lambda = 0.8$ produces efficiencies of about 74%. The efficiency

declines as λ deviates from this value. Clearly the choice of λ is crucial.

Table 3.2: Efficiency of $\hat{\boldsymbol{\theta}}$ for selected values of λ

	Values of λ							
	0	.2	.4	.6	.8	1	1.2	1.4
$\widehat{eff}(\hat{\boldsymbol{\theta}})$	.20	.28	.49	.69	.74	.68	.60	.52

3.7.4 Appropriate values for λ

Practical implementation of the integrated squared error estimator requires the specification of the parameter λ. This parameter is open to choice, and presumably should be chosen so that the determinant of $\Sigma(\boldsymbol{\theta})$ is made as small as possible. The potential for application of this approach is large. However, there are three significant drawbacks to its use here. First, it is based on asymptotic variances whereas finite-sample variances are the appropriate ones to use. Second, the covariance matrix $\Sigma(\boldsymbol{\theta})$ is difficult to compute. Third, it depends on the unknown parameter values.

Alternatively the choice of the value λ may be based on robustness considerations. In the present context, the estimator becomes increasingly robust as λ decreases from infinity. However, λ cannot be decreased indefinitely since then the method will begin to cluster groups of observations and ultimately each observation will be regarded as a separate cluster. This property is due to the connection between integrated squared error estimation and kernel density estimation, an association which we shall now explore.

3.7.5 The density representation of the ISE function

The density representation of the integrated squared error function was the theme of Chapter 2, in which it was demonstrated that for a particular form of the weight

function, the integrated squared error method is equivalently a density estimation method. In particular, if $w(t;\lambda)$ is a characteristic function with corresponding density $f_w(x;\lambda)$, then

$$I(\boldsymbol{\theta};\lambda) = \int_{-\infty}^{\infty} |\phi_n(t) - \phi(t;\boldsymbol{\theta})|^2 \, |w(t;\lambda)|^2 \, \mathrm{d}t \tag{3.32}$$

$$= 2\pi \int_{-\infty}^{\infty} [n^{-1} \sum_{j=1}^{n} f_w(x - X_j;\lambda) - f(x;\boldsymbol{\theta}) * f_w(x;\lambda)]^2 \, \mathrm{d}x, \tag{3.33}$$

where the asterisk denotes the operation of convolution between the indicated densities.

The relative utilities of (3.33) as opposed to (3.32) will depend on the underlying model distribution. For the present mixture model, the density representation (3.33) has the following advantages. First, it shows that $w(t;\lambda) = \exp(-\lambda^2 t^2/2)$, or, equivalently, $f_w(x;\lambda) = (2\pi\lambda^2)^{-1/2} \exp[-x^2/(2\lambda^2)]$, provides an attractive choice, since then the convolution $f(x;\boldsymbol{\theta}) * f_w(x;\lambda)$ is of closed form. In particular, following the notation of (3.17), we find

$$f(x;\boldsymbol{\theta}) * f_w(x;\lambda) = p \, g_{\sigma_1^2+\lambda^2}(x - \mu_1) + (1-p) \, g_{\sigma_2^2+\lambda^2}(x - \mu_2)$$

so that the convolution in (3.33) is again a normal mixture density.

Secondly, it shows that the integrated squared error method is equivalent to a density estimation method. The density $f(x;\boldsymbol{\theta})$ is assumed *a priori*, while the estimate $n^{-1} \sum_{j=1}^{n} f_w(x - X_j;\lambda)$ is a kind of posterior estimate based on the kernel $f_w(x;\lambda)$ and the data. The parameter λ plays the role of the bandwidth in this estimate and, as we shall see, this allows for a somewhat simpler selection of its value.

Thirdly, it constitutes a more practical context in which the robustness of the integrated squared error estimator can be investigated. In particular, the ith ($i = 1, 2, \ldots, 5$) individual influence function for the integrated squared error

estimator is, from Theorem 2.1, given by

$$IF(\xi;\hat{\theta}_i) = \sum_{j=1}^{5} \kappa^{ij}(\boldsymbol{\theta}) \int_{-\infty}^{\infty} \frac{\partial f(x;\boldsymbol{\theta}) * f_w(x;\lambda)}{\partial \boldsymbol{\theta}} [f_w(x-\xi;\lambda) - f(x;\boldsymbol{\theta}) * f_w(x;\lambda)]\,\mathrm{d}x,$$

where $\kappa^{ij}(\boldsymbol{\theta})$ is the (i,j)th element of $K^{-1}(\boldsymbol{\theta})$. This influence function depends on ξ only through

$$h(\xi) = \int_{-\infty}^{\infty} \frac{\partial f(x;\boldsymbol{\theta}) * f_w(x;\lambda)}{\partial \boldsymbol{\theta}} f_w(x-\xi;\lambda)\,\mathrm{d}x,$$

which becomes

$$h(\xi) = \frac{\partial}{\partial \boldsymbol{\theta}} [f(\xi;\boldsymbol{\theta}) * f_w(\xi;\lambda) * f_w(\xi;\lambda)]$$

by changing the order of integration and differentiation. This change of order can be verified here symbolically, for example using Maple, but a more formal justification comes from applying the dominated convergence theorem (see, for example, *Karr, 1993, pp. 108-109*). Since the convolution $f(x;\boldsymbol{\theta}) * f_w(x;\lambda) * f_w(x;\lambda)$ is again a normal mixture density, the individual influence functions for $\hat{\boldsymbol{\theta}}$ are bounded in ξ. This implies that the integrated squared error estimator is robust against outliers.

3.7.6 Links with density estimation

As indicated in Section 3.7.4, the robustness properties of the integrated squared error method are due to its connection with kernel density estimation. We now illustrate and expand on this result. The basic idea of this section is to examine the response of the density estimate in (3.33) to variation in the parameter λ. The approach is that of *Paulson and Nicklin (1983)*.

The twenty-five observations in Table 3.3 are taken from *Everitt and Hand (1981, p. 42)*, and are theoretically from a two-component normal mixture with $\boldsymbol{\theta} = (0.33, -2, 1, 2, 1)^\top$. For our analysis, a choice of the value λ is required. Somewhat arbitrarily, we have chosen $\lambda = 1, \frac{1}{2}, \frac{1}{4}$. Figure 3.5 depicts what the

Table 3.3: Twenty-five observations from a two-component normal mixture with $p = 0.33, \mu_1 = -2, \mu_2 = 2, \sigma_1 = \sigma_2 = 1$

Observation	x_i	*Observation*	x_i
1	0.608	14	2.400
2	−1.590	15	−2.499
3	0.235	16	2.608
4	3.949	17	−3.458
5	−2.249	18	0.257
6	2.704	19	2.569
7	−2.473	20	1.415
8	0.672	21	1.410
9	0.262	22	−2.653
10	1.072	23	1.396
11	−1.773	24	3.286
12	0.537	25	−0.712
13	3.240		

density estimate looks like at these values of λ. Incidentally, for large λ the density estimate is approximately uniform.

At $\lambda = 1$ the density estimate can distinguish between two populations, while at $\lambda = 1/2$ it can distinguish between three populations. As λ further decreases, the density estimate starts distinguishing between observations and ultimately becomes a set of Dirac delta functions located at each observation. Thus, the reason for the increased robustness of the integrated squared error method as λ decreases has become clear.

We now examine the response of the parameter estimates to variation in λ. Table 3.4 tabulates the parameter estimates for selected values of λ.

As λ decreases from four to zero, the parameter estimates fluctuate appreciably. We should point out that it is possible to produce mixture data for which variation in λ can lead to dramatic changes in the parameter estimates. Since the sample information is processed in the integrated squared error method only

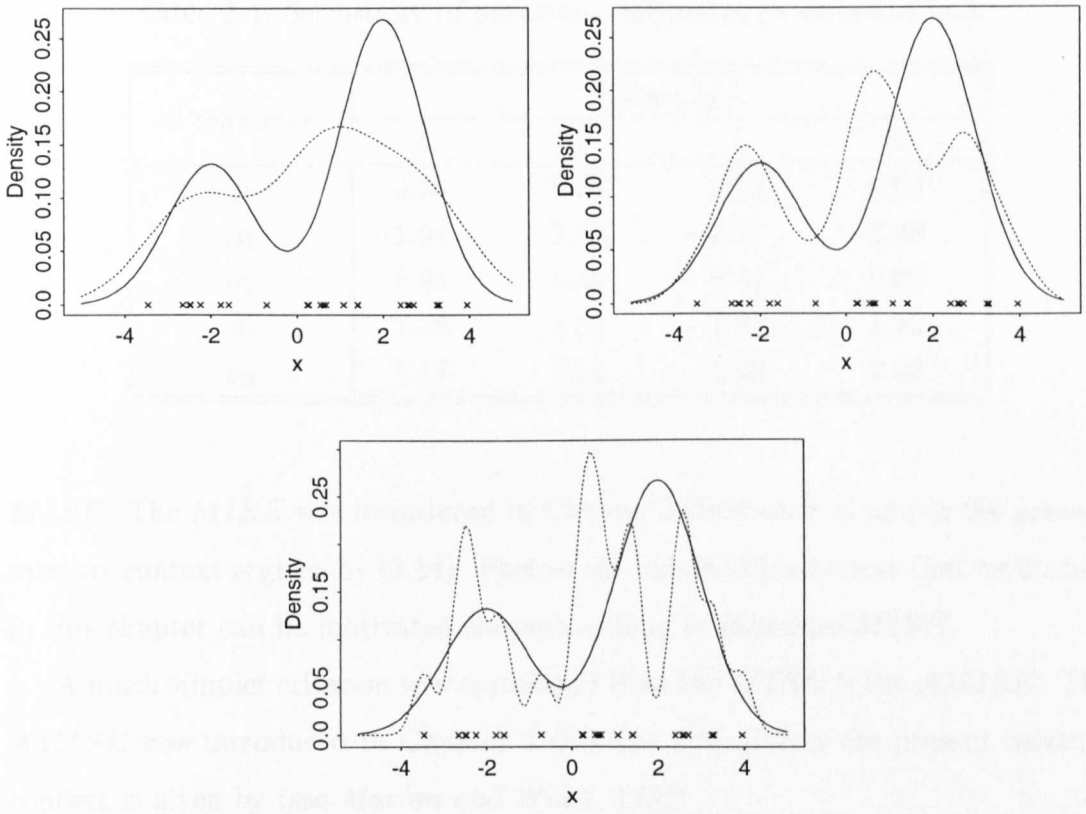

Figure 3.5: The normal mixture density (3.3) with $\boldsymbol{\theta} = (0.33, -2, 1, 2, 1)^\top$ (solid lines) overlaid with kernel density estimates based on the the twenty-five observations of Table 3.3 (dotted lines). The kernel estimates were constructed using the normal kernel and the bandwidths (reading from left-right, top-bottom): (1) $\lambda = 1$; (2) $\lambda = 1/2$; and (3) $\lambda = 1/4$. Also depicted are the observation values (crosses).

through the kernel density estimate, we are recommending that λ be selected in the way given in the next section.

3.7.7 The *MISE* and *AMISE* criteria

The selection of λ in a general estimation context was discussed in detail in Chapter 2. In brief, Chapter 2 drew attention to the connection between integrated squared error and kernel density estimation and provided a natural choice for the parameter λ. This is a very convenient approach which has not been previously adopted.

In kernel density estimation, estimators are compared with reference to the

Table 3.4: Sensitivity of parameter estimates to variation in λ

Parameter	*Values of* λ			
	4	1	$\frac{1}{2}$	0
p	0.36	0.20	0.22	0.09
μ_1	−1.94	−2.59	−2.47	−2.48
σ_1	0.93	0.33	0.41	0.02
μ_2	1.79	1.31	1.35	1.10
σ_2	1.17	1.61	1.60	2.02

MISE. The *MISE* was introduced in Chapter 2 (Section 2.5) and in the present mixture context is given by (3.14). Each of the bandwidth selectors that we discuss in this chapter can be motivated through aiming to minimise *MISE*.

A much simpler criterion to comprehend than the *MISE* is the *AMISE*. The *AMISE* was introduced in Chapter 2 (Section 2.6) and in the present mixture context is given by (see *Marron and Wand, 1992*)

$$AMISE[f_{\varphi_n}(\,.\,;\lambda)] = (2\pi^{1/2}n\lambda)^{-1} + \tfrac{1}{4}\lambda^4 \boldsymbol{p}^\top A \boldsymbol{p}. \tag{3.34}$$

Here $\boldsymbol{p} = (p, 1-p)^\top$, and A is the 2×2 matrix having (i,j) element equal to

$$a_{ij} = g^{(4)}_{\sigma_i^2+\sigma_j^2}(\mu_i - \mu_j), \tag{3.35}$$

where $g^{(r)}_\theta(x) = (\mathrm{d}^r/\mathrm{d}x^r)g_\theta(x)$ is the notation for the rth derivative of $g_\theta(x)$. The bandwidth aiming to minimise the *AMISE* can be easily derived by differentiating (3.34) with respect to λ and setting the derivative equal to zero. This results in the closed form expression

$$\lambda_{AMISE} = \left[\frac{1}{2\pi^{1/2}n\,\boldsymbol{p}^\top A \boldsymbol{p}}\right]^{1/5}.$$

A very important application of the exact *MISE* and *AMISE* expressions is to the problem of quantifying how well the latter approximates the former. This

issue is especially important in practice because there is a definite price to be paid by the $MISE$. This is due to the bandwidth aiming to minimise $MISE$ being only implicitly defined. Figure 3.6 depicts how well $AMISE$ approximates $MISE$ for the asymmetric density of Figure 3.1 when the sample sizes are 50 and 200.

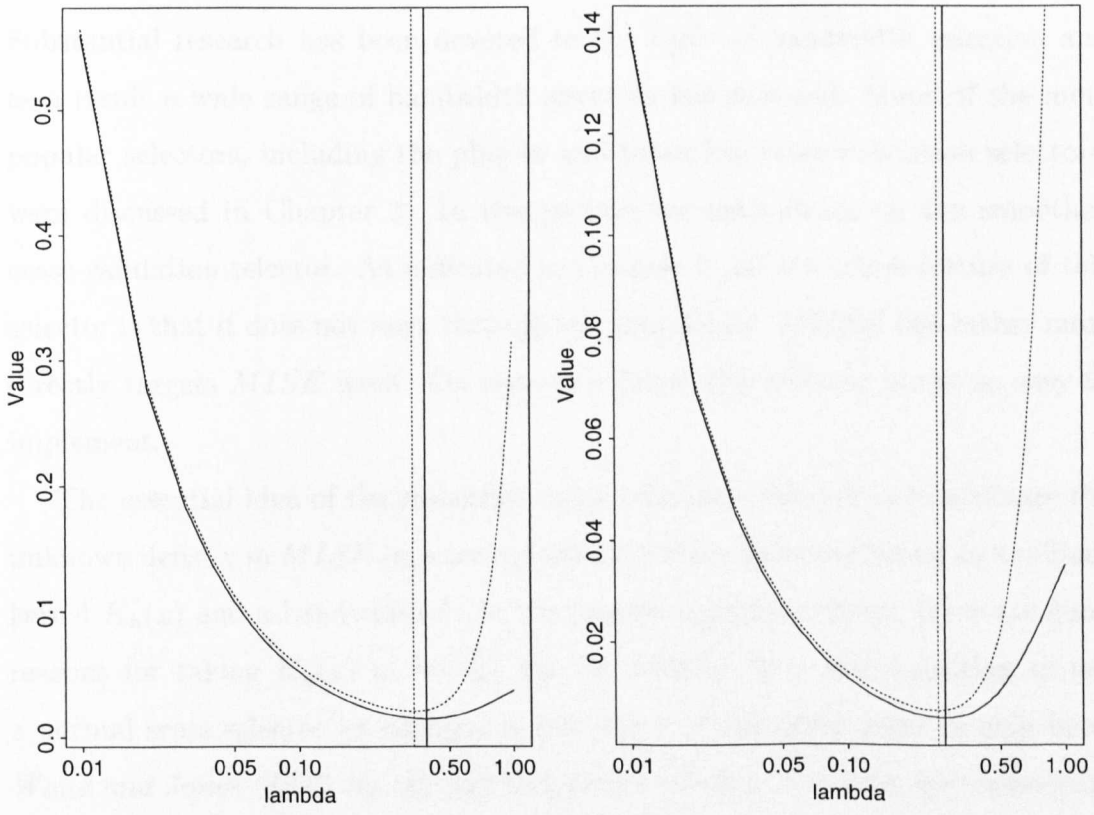

Figure 3.6: The $MISE$ (3.14) (solid lines) and the $AMISE$ (3.34) (dotted lines) for the asymmetric density of Figure 3.1 based on samples of size $n = 50$ (left) and $n = 200$ (right). Also plotted are their respective minimisers (vertical lines).

As one would expect, the $AMISE$ approximation to the $MISE$ improves as the sample size increases. On the other hand, the approximation of these curves worsens considerably as λ grows. This pattern was found to be typical for all the densities of Figure 3.1. In fact, this pattern seems typical for most densities since the bias approximation in the $AMISE$ is based on the assumption that $\lambda \to 0$.

In summary, the $AMISE$ approximation to $MISE$ is quite good for small

λ, but can be very poor for large λ. One consequence of this behaviour is that the bandwidth aiming to minimise $AMISE$ may not always provide a decent approximation to the bandwidth aiming to minimise $MISE$. For essentially this reason, we consider the selector described in the following section.

3.7.8 The smoothed cross-validation selector

Substantial research has been devoted to the topic of bandwidth selection and as a result a wide range of bandwidth selectors has emerged. Some of the more popular selectors, including the plug-in and smoothed cross-validation selectors, were discussed in Chapter 2. In this section we concentrate on the smoothed cross-validation selector. As indicated in Chapter 2, an attractive feature of this selector is that it does not work through the asymptotic $AMISE$ but rather more directly targets $MISE$ itself. On the other hand, this selector is not so easy to implement.

The essential idea of the smoothed cross-validation selector is to estimate the unknown density in $MISE$ by a second kernel density estimator using an auxiliary kernel $K_h(x)$ and a bandwidth h. In the present mixture context, there are good reasons for taking $K_h(x)$ to be the normal density. It is also tempting to use a normal scale selector to estimate h but this is is not quite good enough here. *Wand and Jones (1995, pp. 82–84)* find it best to allow h to have dependence on λ of the form

$$h = Cn^p\lambda^m$$

for parameters C, p and m. An optimal choice of these parameters is discussed in *Wand and Jones (1995, p. 79)*. The smoothed cross-validation selector that evolves from these choices is described below.

Let the auxiliary kernel $K_h(x)$ be the normal density $g_h(x)$, and define

$$\hat{\psi}_r(h) = n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n} g_h^{(r)}(X_i - X_j),$$

where $g_h^{(r)}(x)$ denotes the rth derivative of $g_h(x)$ as in (3.35).

Step 1 Compute kernel estimates $\hat{\psi}_6(h_1)$ and $\hat{\psi}_{10}(h_2)$, where

$h_1 = 2^{1/2}[2/(7n)]^{1/9}\hat{\sigma}$,

$h_2 = 2^{1/2}[2/(11n)]^{1/13}\hat{\sigma}$,

and $\hat{\sigma}$ is a robust estimate of the standard deviation.

Step 2 Compute kernel estimates $\hat{\psi}_4(h_3)$ and $\hat{\psi}_8(h_4)$, where

$h_3 = \{-6/[(2\pi)^{1/2}\hat{\psi}_6(h_1)n]\}^{1/7}$,

and

$h_4 = \{-210/[(2\pi)^{1/2}\hat{\psi}_{10}(h_2)n]\}^{1/11}$.

Step 3 Choose λ to minimise

$$SCV(\lambda) = (2\pi^{1/2}n\lambda)^{-1} + n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}(g_{2\lambda^2+2h^2} - 2g_{\lambda^2+2h^2} + g_{2h^2})(X_i - X_j),$$

where

$h = \hat{C}n^{-23/45}\lambda^{-2}$,

and

$\hat{C} = [441/(64\pi)]^{1/18}(4\pi)^{-1/5}\hat{\psi}_4(h_3)^{-2/5}\hat{\psi}_8(h_4)^{-1/9}$.

This selector will be used extensively in Chapter 4.

3.8 Estimation in the mixture of k normal distributions

Partly as a consequence of its mathematical properties and partly as a consequence of its practical importance, the mixture distribution which has received the most attention is the mixture of two normal distributions. However, there are circumstances in which a mixture of k ($k > 2$) normal distributions is more appropriate. This section is concerned with the problem of estimating the parameters of this mixture.

The probability density function of a mixture of k normal distributions is given

by

$$f(x;\boldsymbol{\theta}) = \sum_{j=1}^{k} p_j g(x;\boldsymbol{\theta}_j), \quad x \in \mathbb{R}, \tag{3.36}$$

where $g(x;\boldsymbol{\theta}_j) = (2\pi\sigma_j^2)^{-1/2}\exp[-(x-\mu_j)^2/(2\sigma_j^2)]$ is the density of the normal distribution with mean μ_j and standard deviation σ_j. In this notation, $\boldsymbol{\theta}_j = (\mu_j, \sigma_j)^\top$ and

$$\boldsymbol{\theta} = (p_1, p_2, \ldots, p_{k-1}, \boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top, \ldots, \boldsymbol{\theta}_k^\top)^\top$$

is the vector of all unknown parameters.

The characteristic function corresponding to (3.36) is

$$\phi(t;\boldsymbol{\theta}) = \sum_{j=1}^{k} p_j \psi(t;\boldsymbol{\theta}_j),$$

where $\psi(t;\boldsymbol{\theta}_j) = \exp(it\mu_j - \sigma_j^2 t^2/2)$ is the characteristic function of the normal distribution with mean μ_j and standard deviation σ_j.

The estimation methods proposed for the mixture of two normal distributions can be extended, with appropriate modifications, to cover the mixture of k normal distributions. However, many new difficulties are encountered in even the simplest $k = 3$ case. For example, the method of moments requires moments up to order eight and the resulting system of equations is likely to be very difficult to solve with any accuracy, quite apart from the poor sampling properties of such high-order sample moments. The computation of the maximum likelihood estimators will, obviously, be even more demanding than the $k = 2$ case. Nevertheless, it is clearly desirable to have some kind of solution to the problem, however complicated. The emphasis of this section is on the integrated squared error method. Incidentally, *Richardson and Green (1997)* have developed a Bayesian method for estimating the parameters when the number of components is considered unknown.

3.8.1 The method of integrated squared error

The application of the integrated squared error method to the mixture of k normal distributions requires the selection of a weight function. The optimum (in $MISE$ sense) weight function is given by (3.10), where

$$|\phi(t;\boldsymbol{\theta})|^2 = \sum_{i=1}^{k}\sum_{j=1}^{k} p_i p_j \exp[-\tfrac{1}{2}(\sigma_i^2+\sigma_j^2)t^2]\cos[(\mu_i-\mu_j)t]. \tag{3.37}$$

It follows from the degree of complexity of (3.37) that the optimum $MISE$ weight function may be impractical. The weight function leading to the greatest degree of mathematical tractability is

$$w(t;\lambda) = \exp(-\tfrac{1}{2}\lambda^2 t^2), \tag{3.38}$$

or, equivalently,

$$f_w(x;\lambda) = (2\pi\lambda^2)^{-1/2}\exp[-\tfrac{1}{2}(x/\lambda)^2].$$

This weight function has resulted in a novel method of parameter estimation which is presented below.

The integrated squared error function for the mixture of k normal distributions based on (3.38) may be explicitly integrated to give

$$I(\boldsymbol{\theta};\lambda) \propto \boldsymbol{p}^\top I_0\,\boldsymbol{p} - 2\boldsymbol{p}^\top \boldsymbol{I}_1, \tag{3.39}$$

where $\boldsymbol{p} = (p_1, p_2, \ldots, p_k)^\top$, I_0 is a $k \times k$ symmetric matrix with (i,j) element

$$I_{0;ij} = [2\pi(\sigma_i^2+\sigma_j^2+2\lambda^2)]^{-1/2}\exp[-\tfrac{1}{2}(\mu_i-\mu_j)^2/(\sigma_i^2+\sigma_j^2+2\lambda^2)],$$

and $\boldsymbol{I}_1$ is a $k \times 1$ vector whose ith element is

$$I_{1;i} = \frac{1}{n}\sum_{j=1}^{n}[2\pi(\sigma_i^2+2\lambda^2)]^{-1/2}\exp[-\tfrac{1}{2}(X_j-\mu_i))^2/(\sigma_i^2+2\lambda^2)].$$

A system of integrated squared error equations can be formed by differentiating (3.39) with respect to the parameters and setting the resulting expressions equal to zero. These equations are, of course, beyond hope of solution by analytic means. Consequently, one must resort to seeking an approximate solution via some iterative procedure. Nevertheless, we can investigate the robustness of the estimators without much difficulty. The first approach we used to find the joint influence function for the integrated squared error estimator was based on Theorem 1.3. However, this was not very promising due to the complexity of the results. It is precisely for this reason that the density representation of Theorem 1.3, stated as Theorem 2.1, was developed.

In the present mixture context, the density representation of the integrated squared error function is given by

$$I(\boldsymbol{\theta};\lambda) = 2\pi \int_{-\infty}^{\infty} [n^{-1}\sum_{j=1}^{n} f_w(x - X_j;\lambda) - f(x;\boldsymbol{\theta}) * f_w(x;\lambda)]^2 \,\mathrm{d}x, \qquad (3.40)$$

where the asterisk denotes the operation of convolution between the indicated densities. It follows from Theorem 2.1 that the ith ($i = 1, 2, \ldots, 3k-1$) individual influence function for the integrated squared error estimator, $\hat{\boldsymbol{\theta}}$, has the form

$$IF(\xi;\hat{\theta}_i) = \sum_{j=1}^{3k-1} \kappa^{ij}(\boldsymbol{\theta}) \int_{-\infty}^{\infty} \frac{\partial f(x;\boldsymbol{\theta}) * f_w(x;\lambda)}{\partial \boldsymbol{\theta}} [f_w(x-\xi;\lambda) - f(x;\boldsymbol{\theta}) * f_w(x;\lambda)] \,\mathrm{d}x,$$

where $\kappa^{ij}(\boldsymbol{\theta})$ denotes the (i,j) element of $K^{-1}(\boldsymbol{\theta})$. The robustness of the integrated squared error estimator follows by employing similar arguments to the $k = 2$ case.

The integrated squared error method cannot be applied profitably without a good choice for λ. In parallel to the mixture of two normal distributions, one can select λ by minimising either $MISE$ or $AMISE$. Taking the results of Section 3.7.7 as a basis, we recommend minimising the former. The theoretical expression for $MISE$ is given by (see *Marron and Wand, 1992*)

$$MISE[f_{\varphi_n}(\,.\,;\lambda)] = (2\pi^{1/2} n\lambda)^{-1} + \boldsymbol{p}^{\top}[(1-n^{-1})M_2 - 2M_1 + M_0]\boldsymbol{p}, \qquad (3.41)$$

where $\boldsymbol{p} = (p_1, p_2, \ldots, p_k)^\top$, and M_ℓ $(\ell = 0, 1, 2)$ is the $k \times k$ matrix having (i, j) element equal to

$$m_{\ell;ij} = [2\pi(\sigma_i^2 + \sigma_j^2 + \ell\lambda^2)]^{-1/2} \exp[-\tfrac{1}{2}(\mu_i - \mu_j)^2/(\sigma_i^2 + \sigma_j^2 + \ell\lambda^2)].$$

The practical minimisation of (3.41) may be achieved through a version of the smoothed cross-validation selector.

3.9 Concluding remarks

The estimation of the parameters in mixture distributions is one of the oldest problems in the statistical literature. This chapter has systematically considered the mixture of two normal distributions. In particular, it has summarised some of the currently used estimation methods and has also provided a comprehensive account of the theoretical and computational issues in integrated squared error estimation. With this latter approach, a choice must be made for the scaling of the weight function. This choice was addressed using the developments of Chapter 2. Finally, these results were extended to the mixture of k $(k > 2)$ normal distributions.

Chapter 4

Sampling experiments

4.1 Introduction

The problem of estimating the parameters of a mixture of normal distributions was considered in Chapter 3, with emphasis on the mixture involving two components. This particular mixture has received substantial attention in the statistical literature, resulting in the wide variety of estimation methods seen in the chapter. This variety of applicable methods can be bewildering. To judge rapidly between them, it is necessary to bear in mind accuracy, robustness, and ease of calculation. The relative importance of these factors varies with circumstance, but they should always be taken into account. Thus, for example, graphical inference may be dismissed on account of accuracy, whilst the perceived computational complexity of the Bayesian approach may still prove too daunting for many people. However, it is not immediately apparent how to choose from among the methods of:

1. moments;

2. maximum likelihood;

3. integrated squared error.

The purpose of this chapter is to examine the performance of these methods for estimating the parameters of a mixture of two normal distributions. The

chapter begins by introducing the main tools for assessing the performance of an estimator, with details of these in the mixture of two normal distributions being presented in the following two sections. The problem of simulating from this mixture is considered in Section 4.5, and the importance of starting values for maximum likelihood and integrated squared error estimation is then discussed. Two stochastic optimisation methods are introduced in the following two sections, and the computational details of the simulation study that follows are presented in Section 4.9. A comparison of two selectors for the scaling of the weight function in integrated squared error estimation is carried out next, and the results of the simulation study are presented in Section 4.11. Finally, we present the results of the simulation study when the true parameter values were used to start the recursions.

4.2 Comparison tools

In investigating the behaviour of a parameter estimator, an analysis of its finite sampling distribution is often undertaken in statistics. This analysis involves deriving expressions for the expectation and variability of the estimator, quantities which can also be used to decide among estimators.

More often the derivation of finite sampling distributions is difficult or intractable, particularly for estimators with complicated structure. It is precisely for this reason that statisticians have devised two important tools to facilitate the study of these estimators. These are *asymptotic theory* and *simulation.*

The strength of asymptotic theory is that it provides an effective means of examining the behaviour of an estimator, through general results such as the laws of large numbers and the central limit theorem. The weakness of asymptotic theory is that it only describes the behaviour of the estimator in large samples. This behaviour does not necessarily coincide with what is happening in smaller samples.

Simulation helps in overcoming some of the problems entailed in asymptotic

theory. In short, simulation enables one to examine the behaviour of an estimator for any sample size. Despite this important strength, the weakness of simulation should also be recognised. This is that the lessons are limited to only the set of examples that can be studied. These limits are of practical importance, because very substantial effort is required to carry out even a moderate scale simulation study.

Clearly, asymptotic theory and simulation can be used to complement each other and so consequently this is the approach we have adopted below. The strength of this approach is that one can gain much more from this than when the same effort is devoted to either simulation or to complicated theoretical work.

4.3 Asymptotic theory

As noted in the previous section, a typical asymptotic analysis is based on repeated sampling ideas, and aims to establish the consistency and asymptotic normality of the estimator in question. These properties are a consequence of applying, respectively, the laws of large numbers and the central limit theorem to an appropriately derived sequence of random variables.

A common characteristic of the moment, maximum likelihood and integrated squared error estimators is their well-developed asymptotic theory. In particular, the consistency and asymptotic normality of each estimator can be easily derived in a wide variety of distributions. However, in the case of a mixture of two normal distributions the asymptotic theory is not so straightforward. A brief review of past work related to this mixture is presented below.

1. We begin with the maximum likelihood (ML) estimator since it naturally forms a basis for our comparisons. The derivation of the asymptotic distribution of the ML estimator reduces to the calculation of the Fisher information matrix $\mathcal{I}(\boldsymbol{\theta})$. Explicit evaluation of $\mathcal{I}(\boldsymbol{\theta})$ is not possible here and several authors, including *Hill (1963)*, *Behboodian (1972)*, and *Dick and Bowden (1973)*, have considered approximate information matrices instead.

A common finding of these approximations is that in some sections of the parameter space very large samples may be needed for accurate parameter estimation.

2. In parallel to the ML estimator, the derivation of the asymptotic distribution of the integrated squared error (ISE) estimator reduces to the calculation of the asymptotic covariance matrix $\Sigma(\boldsymbol{\theta})$. We indicated how this matrix can be calculated in Chapter 3 (Section 3.7.3), but an effective comparison of $\Sigma(\boldsymbol{\theta})$ and $\mathcal{I}(\boldsymbol{\theta})$ may only be carried out numerically. It is clear at the outset that the efficiency of the ISE estimator relative to the ML estimator will be less than unity, but *Bryant and Paulson (1983)* report optimistic efficiencies for the ISE estimator of the mixing proportion.

3. Finally, calculation of the covariance matrix for the moment estimator tends to be restricted to Taylor expansion-based approximations (see, for example, *Robertson and Fryer, 1970*). However, *Tan and Chang (1972)* and *Fryer and Robertson (1972)*, amongst others, have shown that the method of moments (MM) is generally inferior to the method of ML for this problem.

In theory, the asymptotic results of the previous paragraphs suggest that the ML estimator should always be used. In practice, these results need to be viewed with some caution because they do not necessarily apply to smaller samples. This issue is highlighted in simulations by *Dick and Bowden (1973)*, where the sample variances of the ML estimates often exceeded the estimated asymptotic variances by several factors. Furthermore, asymptotic results disregard certain aspects, such as ease of calculation and feasibility, which play an important role in the selection of an estimator in practice.

The main tool for assessing the practical performance of an estimator is simulation. In the remainder of this chapter, we shall use simulation to examine the performance of the above methods.

4.4 Simulation details

When a simulation study is being designed, a compromise between two contradictory issues must be reached. On the one hand it is reassuring to examine a wide variety of examples, but on the other it is practical to keep the number of examples small.

This compromise is very important in the selection of examples that can be considered. In this study we appreciate the need to examine examples which provide varying degrees of difficulty in the estimation of the mixture parameters. The degree of difficulty depends on either the separation between the component densities or the mixing proportion. It initially appears that a very large number of experiments might be necessary. Fortunately, we can significantly reduce the number of experiments for two reasons. First, large degrees of separation are not likely to be of much comparative value as all methods can be expected to perform equally well. Secondly, parameter sets in which $p \geq 0.5$ need only be examined. This is due to a phenomenon called "label switching" by *Redner and Walker (1984)*, which points out that the mixture densities evaluated at $\boldsymbol{\theta} = (p, \mu_1, \sigma_1, \mu_2, \sigma_2)^\top$ and $\boldsymbol{\theta} = (1-p, \mu_2, \sigma_2, \mu_1, \sigma_1)^\top$ are necessarily identical. Consequently, attention shall be restricted to the experiments summarised in Table 4.1.

Table 4.1: Summary Characteristics of Experiments

Experiment	*Population Parameters*					*Sample Size*	*Number of Samples*
	p	μ_1	σ_1^2	μ_2	σ_2^2		
1	0.5	−2	1	2	1	50	100
2	0.5	−1	3	1	1	50	100
3	0.5	0	9	3	1	50	100
4	0.8	−2	1	2	1	50	100
5	0.8	−1	3	1	1	50	100
6	0.8	−1	9	1	1	50	100
7	0.8	0	1	0	16	50	100

The density functions corresponding to these experiments are shown visually in Figure 4.1. These seven densities have been carefully selected because they cover a wide spectrum of the possible mixture problems that may occur in practice. It is, therefore, reasonable to expect that the lessons learnt from this simulation study will be of wider applicability than to just the parameter sets studied here.

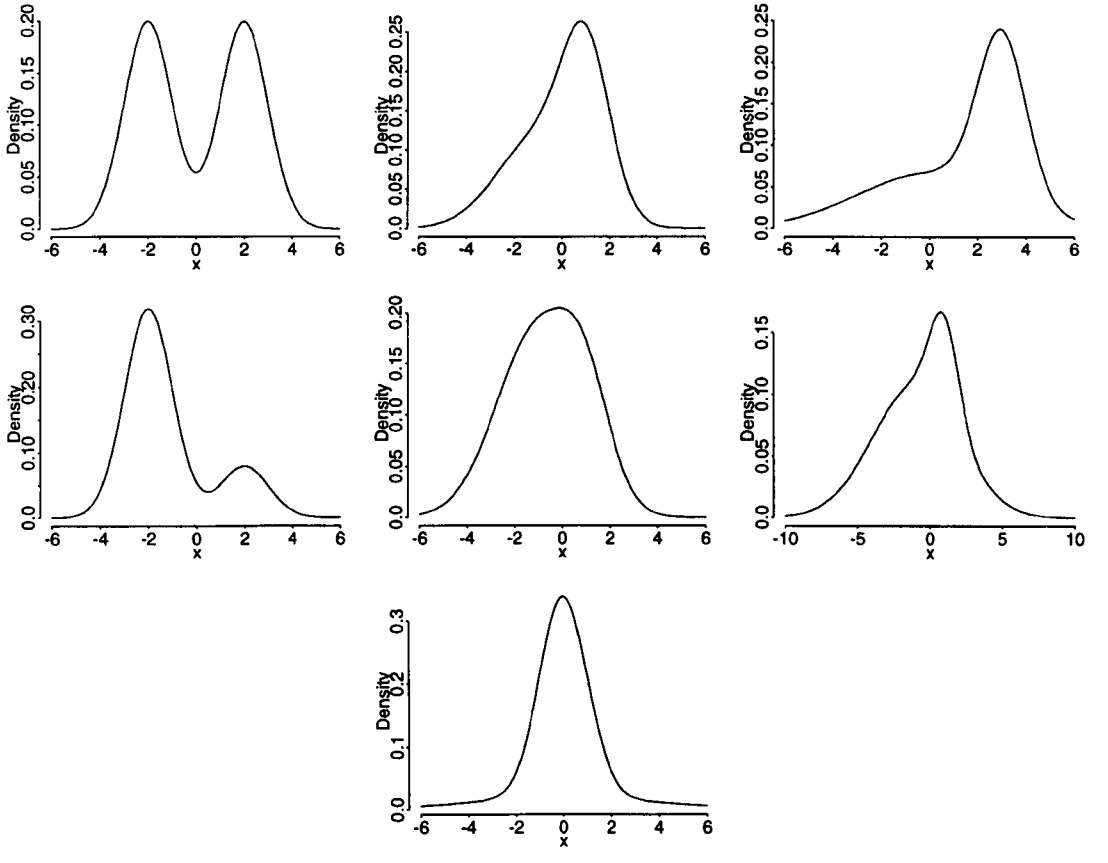

Figure 4.1: Probability density functions corresponding to (reading from left-right, top-bottom): (1) experiment 1; (2) experiment 2; (3) experiment 3; (4) experiment 4; (5) experiment 5; (6) experiment 6; and (7) experiment 7.

In contrast, the sample size does impose a restriction on the simulation conclusions. It is worth mentioning that the primary purpose of this study was to compare the small-sample characteristics of the three methods. Although this may seem a paradox, small-sample characteristics are often examined in the literature for sample sizes in excess of 100 observations. However, it may be argued that such sample sizes are larger than those that typically arise in practice. On

the other hand, it was felt that in very small samples all methods might produce bad results and no difference would be noticed. Judiciously a compromise was reached by taking $n = 50$ throughout.

4.5 Random variate generation

For each experiment in Table 4.1, one hundred random samples of size $n = 50$ from the corresponding mixture distribution were needed. Nowadays, computer programmes that simulate from a variety of distributions have been written and are available in sources such as the IMSL and NAG libraries. However, it is unusual to encounter library subroutines that simulate from mixture distributions, even those with normal components. The following probabilistic meaning of the mixture distribution function might therefore prove useful for this purpose.

Suppose X is a random variable with distribution function $F(x)$, and let Π_j $(j = 1, 2)$ be a population with distribution function $F_j(x)$. Furthermore, suppose that E_j is the event that X comes from population Π_j with $\Pr(E_1) = p$ and $\Pr(E_2) = 1-p$. Now, assuming that Π_1 and Π_2 are mutually exclusive populations, we have

$$\Pr(X \leq x) = \Pr(E_1)\ \Pr(X \leq x \mid E_1) + \Pr(E_2)\ \Pr(X \leq x \mid E_2),$$

or, equivalently,

$$F(x) = pF_1(x) + (1-p)F_2(x)$$

for any real number x.

In general, this means that one can simulate from $F(x)$ by simulating from $F_1(x)$ with probability p, and from $F_2(x)$ with probability $1-p$. In this respect, the normal mixture samples necessary for the experiments were generated as follows. For each sample, fifty observations from the standard normal distribution were generated (see, for example, *Morgan, 1984, pp. 78-81*). Each standard normal variate was then appropriately transformed to either component 1 or component

2, depending on whether an independent variate, distributed uniformly over $(0, 1)$, was less than or greater than p. This approach to sampling from a mixture distribution can be generalised to accommodate any number of components, and does not, of course, depend on the component distributions being normal.

Furthermore, this approach allows one to keep track of the origin of each observation. For each sample we can thus determine the mixing proportion and estimate each component distribution separately. This would be equivalent to the situation in practical applications where additional information exists to classify the data. *Hosmer (1973)* showed that considerable gains in efficiency are possible in these situations but occurrences of classified data are rare in practice.

4.6 Starting values

For either the ML or the ISE estimator to be used in practice, one must provide starting values for their optimisation procedures. Clearly, the selection of starting values is not important when the function to be optimised does not possess local optima. *Fowlkes (1979)* observed that the selection of starting values is crucial in ML estimation, while *Hosmer (1973)* and *Woodward, Parr, Schucany and Lindsey (1984)* indicated that starting values are not that critical. The importance of good starting values is not known in ISE estimation.

In order to examine the susceptibility of ML and ISE estimation to starting values, we performed a preliminary simulation study. In particular, for a series of samples of size $n = 50$, the ML and ISE optimisation procedures were initiated from several starting points. The final parameter estimates were then obtained and compared. We found that the optimisation procedures produced substantially different estimates for several of the starting points for some samples, although many starting points produced essentially the same estimates. This feature was worse in ML estimation but demonstrated that in some samples good starting points are important for both methods.

Selecting good starting values is not easy and there is no guarantee that a

certain value will lead to the global optimum. A global optimum can only be ensured by an extensive search of the parameter space, which can be very time-consuming. Alternatively, a stochastic search method could be used.

Stochastic search methods have been designed to deal with the problem of function optimisation in the presence of local optima. There are several methods to choose from and some of the methods are very sophisticated indeed. We based our selection on accounts of well-developed theory and ease of application. A method that satisfied both accounts was *simulated annealing.* In the following section we outline the main steps of a simulated annealing algorithm for the optimisation of continuous functions.

4.7 Optimisation using simulated annealing

Much work has been published on the theoretical aspects of simulated annealing. This section provides a brief overview of this theory as well as an introduction to the practical aspects of function optimisation using this approach. Additional references are *Kirkpatrick, Gelatt and Vecchi (1983)* and *Ingrassia (1992)*; see also *Brooks and Morgan (1995).* The name of the algorithm comes from the analogy with the physical procedure of annealing, which consists of melting and then slowly cooling a physical substance in search of the ground state.

Let $H(\boldsymbol{x})$ be a real-valued function defined on a compact subset $\boldsymbol{D}$ of $\mathbb{R}^d$, and without loss of generality suppose the context is minimisation. The simulated annealing algorithm is an inhomogeneous Markov process on $\boldsymbol{D}$ depending on a positive parameter T such that, for each value of T, the corresponding homogeneous Markov process has the Gibbs distribution as its unique equilibrium distribution:

$$\pi_T(\boldsymbol{x}) = \frac{\exp[-H(\boldsymbol{x})/T]}{\int_{\boldsymbol{D}} \exp[-H(\boldsymbol{y})/T]\mathrm{d}\boldsymbol{y}}, \quad \boldsymbol{x} \in \boldsymbol{D}.$$

From the physical analogy, the function $H(\boldsymbol{x})$ is called *energy* and the parameter T is the *temperature.*

As T tends to 0 from above, it can be shown that $\pi_T(\boldsymbol{x})$ converges to a probability measure which is concentrated on the set of points of global minima for $H(\boldsymbol{x})$. Undoubtedly the most important phase of this process concerns the cooling, especially at low temperatures. This process is dependent on time, and hence we shall write T_t instead of T.

One of the main drawbacks of the algorithm concerns its slow speed of convergence. Although an optimal cooling schedule has been found, it is not normally applied in practical problems since the convergence of such an algorithm is too slow. Instead, many non-optimal cooling schedules have been proposed as a compromise between the speed of convergence and the probability that the algorithm does not get stuck in a metastable state. We recall that a *cooling schedule* is the following set of parameters:

1. an initial value $\boldsymbol{x}_0$;
2. an initial temperature T_0;
3. a criterion for the choice of the point for the subsequent iteration;
4. a criterion for changing the current value of the temperature in between two subsequent Markov chains;
5. the length N of each Markov chain;
6. a stopping criterion for the algorithm.

The simulated annealing algorithm repeats the following two steps N times for each value of the temperature T_t.

Step 1 Generate a point $\boldsymbol{y} \in \boldsymbol{D}$ for the subsequent iteration.

Step 2 If $H(\boldsymbol{y}) \leq H(\boldsymbol{x})$ then we accept $\boldsymbol{y}$ as the new state of the Markov chain at temperature T_t, otherwise we take $\boldsymbol{y}$ as the new state with probability $\exp\{-[H(\boldsymbol{y}) - H(\boldsymbol{x})]/T_t\}$ and $\boldsymbol{x}$ as the new state with probability $1 - \exp\{-[H(\boldsymbol{y}) - H(\boldsymbol{x})]/T_t\}$.

Then, if the stopping criterion is not satisfied, we decrease the temperature, according to a fixed rule, and repeat Steps 1 and 2 again N times. The difference with respect to any deterministic search method comes from the acceptance, with a fixed probability law, of a point in which the function $H(\boldsymbol{x})$ has a value greater than the previous one. In this way we can avoid getting trapped in a local minimum.

The selection of the point $\boldsymbol{y} = (y_1, y_2, \ldots, y_d)^\top$ for the subsequent iteration is quite important. *Brooks and Morgan (1995)* suggested a method, in which $\boldsymbol{y}$ is chosen by first selecting one of the y_i $(i = 1, 2, \ldots, d)$ variables at random, and then randomly selecting a new value for that variable within the bounds set for it by the problem at hand. Thus $\boldsymbol{y}$ takes the same variable values as $\boldsymbol{x}$ except for one.

Finally, the point in which the function $H(\boldsymbol{x})$ has reached the smallest value during the realisation of the algorithm is given as the simulated annealing solution.

4.8 Hybrid algorithm

A characteristic of the annealing algorithm is that it always converges to within a neighbourhood of the global minimum and, furthermore, the size of this neighbourhood can be reduced by altering the parameters of the algorithm. In most cases we would like to find the global minimum to several decimal places. It is clear that the annealing algorithm could produce results with such accuracy, but that the execution time would be prohibitive. On the other hand, a deterministic search method can produce solutions to any accuracy but can have considerable difficulty in finding the correct solution. This suggests that a *hybrid algorithm* would be worth consideration. This algorithm consists of two distinct components. The first component, an annealing algorithm, is used to produce a starting point for the second component, a deterministic search method.

Brooks and Morgan (1995) suggested an alternative approach, in which the annealing component is stopped prematurely, after N_t temperature reductions,

and each of the points accepted at the final temperature, together with the best point overall, are taken as starting points. The second component is then initiated from each starting point, and the best solution generated from these points is given as the hybrid solution.

This algorithm has been found by *Brooks and Morgan (1995)* to be both fast and reliable for a range of functions and will be adopted for this work. The first component of this hybrid algorithm may be regarded as a very sophisticated way of selecting starting values. The quantity and quality of these starting values is controlled by the cooling schedule of the annealing component. The cooling schedule here considered can be described as follows:

1. *Initial solution* $\boldsymbol{\theta}_0$—this is randomly selected in $\boldsymbol{\Theta}$.

2. *Initial value of the temperature* T_0—we have set $T_0 = 10$.

3. *Length of each Markov chain* N—we have adopted $N = 200$.

4. *Rule for changing the temperature between subsequent Markov chains*—we have considered the simple exponential law $T_{n+1} = \rho T_n$ with $\rho = 0.9$.

5. *Temperature reductions* N_t—we have adopted $N_t = 100$.

For the second component of the hybrid algorithm we designated the expectation-maximisation (EM) algorithm for the method of ML, and the Nelder-Mead Simplex search method for the method of ISE. The EM algorithm was described in Chapter 3 (Section 3.5.3). The Nelder-Mead Simplex search method is described in, for example, *Everitt (1987, pp. 16-20)*. In short, the method works by building a simplex, that is, a polytope of $d + 1$ points where d is the dimension of the problem, and expanding and contracting the simplex according to the Nelder-Mead algorithm to search for a local minimum.

The computational details of these hybrid algorithms, together with those of the MM are presented in the following section.

4.9 Computational details

As indicated in Chapter 3 (Section 3.5.1), the computation of the MM estimates requires a negative root of the ninth degree polynomial equation (3.5). Modern iterative procedures can effectively locate the roots of any polynomial equation, and so the MM estimates follow without much difficulty. Unfortunately, for some combinations of sample data:

1. the nonic may have no negative roots;
2. the nonic may have more than one negative roots;
3. the MM estimates may not be admissible, that is, one of the conditions $\hat{p} < 0$, $\hat{p} > 1$, $\hat{\sigma}_1 \leq 0$ or $\hat{\sigma}_2 \leq 0$ may exist.

In this work, the lack of uniqueness implied by the second difficulty was resolved by choosing the set of estimates in closest agreement between the fitted sixth-order moment and its sample counterpart. In the occurrence of the other two difficulties, however, the computation of the MM estimates was classified as a failure.

On the other hand, the ML estimates of the mixture parameters were computed as follows. Starting at a set of initial estimates

$$\boldsymbol{\theta}^{(0)} = (p^{(0)}, \mu_1^{(0)}, \sigma_1^{(0)}, \mu_2^{(0)}, \sigma_2^{(0)})^\top,$$

the iterative calculations of the EM algorithm were repeated, until either of the following conditions occurred:

1. the EM algorithm converged;
2. the estimates $\boldsymbol{\theta}^{(m)}$ $(m = 1, 2, \ldots)$ were inadmissible, that is, the standard deviation of one of the components went to zero.

The EM algorithm was considered to have converged when the proportionate value of the step size dropped below 10^{-4}. If the EM algorithm failed to converge in the

specified number of iterations, the estimates obtained in the last iteration were taken to be the parameter estimates. The computation of the ML estimates would be counted as a failure only if the second condition occurred for each starting value produced by the annealing component of the hybrid algorithm.

Finally, the ISE estimates of the mixture parameters were computed as follows. Starting at a set of initial estimates, the minimisation of the ISE function (3.18) was carried out by the Nelder-Mead Simplex search method until:

1. the Simplex method converged;

2. the specified number of iterations was reached.

In either case, the estimates obtained in the last iteration were taken to be the parameter estimates. The computation of the ISE estimates would be counted as a failure only if inadmissible estimates were obtained from each starting point produced by the annealing component of the hybrid algorithm.

In addition to initial parameter estimates, however, the method of ISE requires a value for λ. This parameter determines the scaling of the weight function in the ISE function, which was shown in Chapter 2 to be quite important. Also shown in the chapter was how this parameter could be selected in practice, and the smoothed cross-validation selector was suggested for this purpose. The aim of the following section is to examine the performance of the smoothed cross-validation selector.

4.10 Comparison of two selectors for λ

The smoothed cross-validation selector for λ was introduced in Chapter 2 (Section 2.11.2), and was discussed in detail in Chapter 3 (Section 3.7.8) within the context of a mixture of two normal distributions. We shall now examine and contrast the performance of this selector with that of an iterative selector.

The iterative selector proposed below is similar to the smoothed cross-validation selector in that it minimises the mean integrated squared error ($MISE$). The

difference is that the iterative selector substitutes estimates for the unknown parameters in the $MISE$ rather than substitute a kernel estimator for the unknown density. This approach is viable in the context of ISE estimation because there is *a priori* knowledge of the parametric family describing the underlying population.

When the parametric family is the mixture of two normal distributions, the $MISE$ is given by (3.14). In this case, the iterative selector may be described as follows: first estimate $\boldsymbol{\theta}_0$ using (3.15) with a normal scale selector for λ; next, substitute these parameter estimates into (3.14) and minimise with respect to λ; and finally, update parameter estimates using (3.15) with the new λ and repeat the cycle until convergence. If this iteration converges, then λ can be selected by the iterative selector.

The comparison of the smoothed cross-validation and the iterative selectors will be based on the simulated samples for the experiments of Table 4.1. Although the performance of the resulting ISE estimator is the main concern, the results of this section will be stated in terms of the selector distribution. This approach is analogous to that of *Park and Marron (1990)*, where the performance of the kernel density estimator is stated in terms of the bandwidth distribution.

Figures 4.2 and 4.3 contain the results of the simulation study. For each experiment, these figures depict the kernel estimates of the density of the smoothed cross-validation and iterative selectors. The bandwidth used for these kernel estimates was the normal scale selector, which seems reasonable in view of the limiting normal distributions generally available. To attach additional insight to these estimates, the value for λ minimising the $MISE$ evaluated at the true parameter values is also depicted.

The performances of the smoothed cross-validation and the iterative selectors were satisfactory in that they produced reasonable values for λ each time. However, the performance of the iterative selector was in a sense disappointing. This selector makes use of the known distribution of X_1 and was therefore expected to produce better results than those produced by the smoothed cross-validation selector. In fact, this turned out not to be the case and the iterative selector tended to

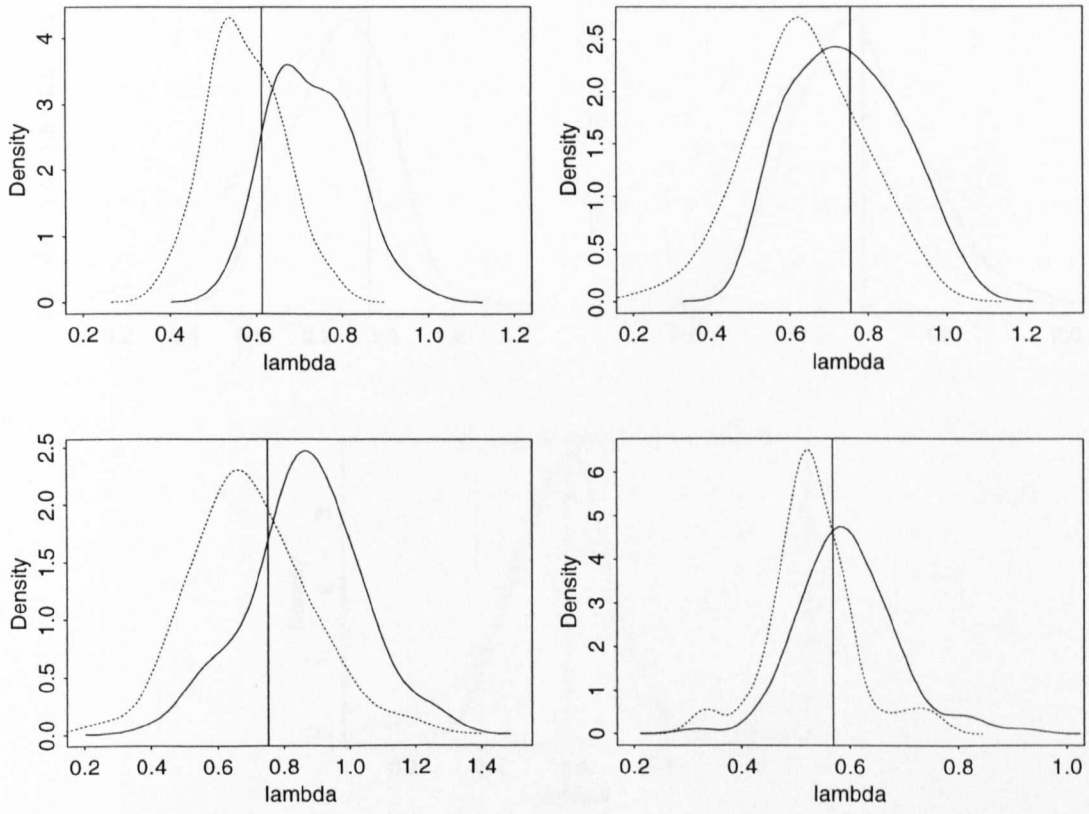

Figure 4.2: Kernel estimates of the density of the smoothed cross-validation selector (solid lines) and the iterative selector (dotted lines) based on the samples for (reading from left-right, top-bottom): (1) experiment 1; (2) experiment 2; (3) experiment 3; and (4) experiment 4. The vertical lines show the values for λ which minimise the $MISE$ evaluated at the true parameter values.

undersmooth. On the other hand, the smoothed cross-validation selector generally oversmoothed, but this at least has the merit of discouraging overinterpretation of features which may be due to sampling variation. In addition the smoothed cross-validation selector was considerably easier to apply since this selector did not depend upon starting values.

We believe that these results (together with the practical advice of *Wand and Jones (1995, p. 86)* that selectors based on the $AMISE$ tend to underperform selectors based on the $MISE$) warrant the use of the smoothed cross-validation selector for the selection of λ in ISE estimation. Accordingly, the smoothed cross-validation selector will be adopted for this work.

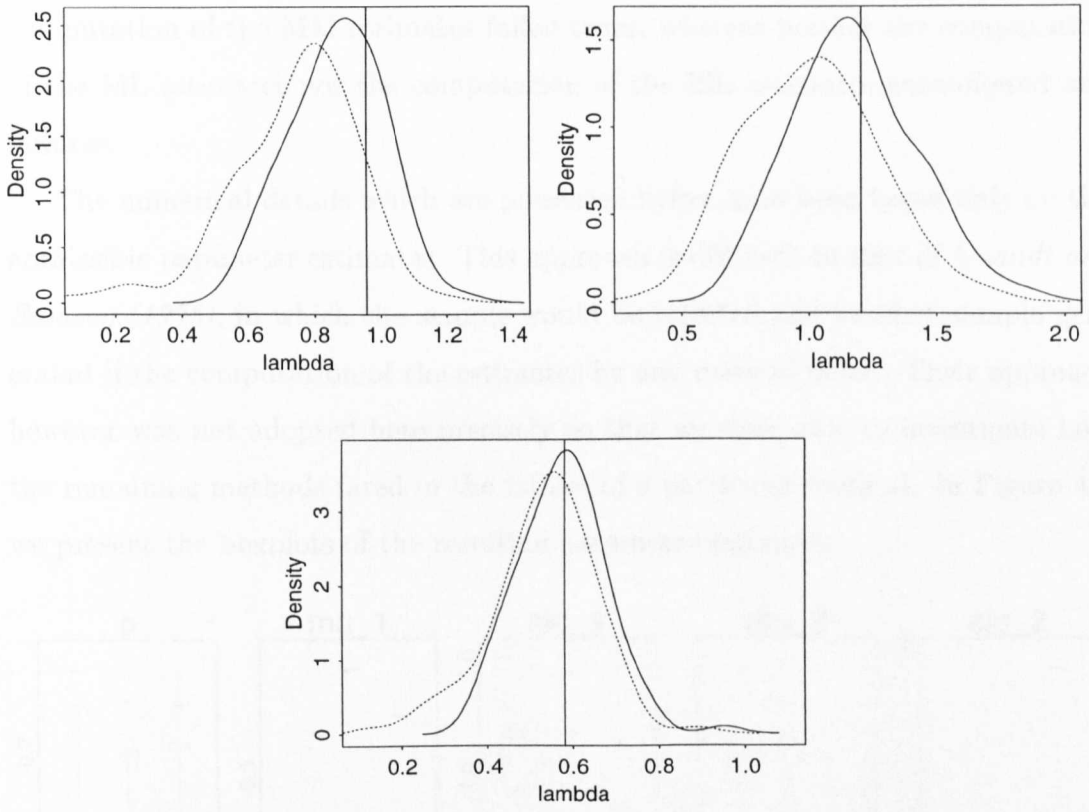

Figure 4.3: Kernel estimates of the density of the smoothed cross-validation selector (solid lines) and the iterative selector (dotted lines) based on the samples for (reading from left-right, top-bottom): (1) experiment 5; (2) experiment 6; and (3) experiments 7. The vertical lines show the values for λ which minimise the $MISE$ evaluated at the true parameter values.

4.11 Simulation results

The results of the simulation study on the methods of MM, ML and ISE will now be presented, with the remainder of the section devoted to the discussion of these results.

4.11.1 Experiment 1 ($p = 0.5, \mu_1 = -2, \sigma_1 = 1, \mu_2 = 2, \sigma_2 = 1$)

The results based on the samples of experiment 1 are reported first. As indicated in Table 4.1, one hundred samples from the corresponding distribution were generated. The estimation methods discussed in Section 4.9 were implemented in the

aforementioned way to produce parameter estimates for each sample. Thus, the computation of the MM estimates failed twice, whereas neither the computation of the ML estimates nor the computation of the ISE estimates encountered any failures.

The numerical details which are presented below have been based only on the admissible parameter estimates. This approach is different to that of *Quandt and Ramsey (1978)*, in which the sample would be rejected and another sample generated if the computation of the estimates by any method failed. Their approach however was not adopted here precisely so that we were able to investigate how the remaining methods fared in the failure of a particular method. In Figure 4.4 we present the boxplots of the resulting parameter estimates.

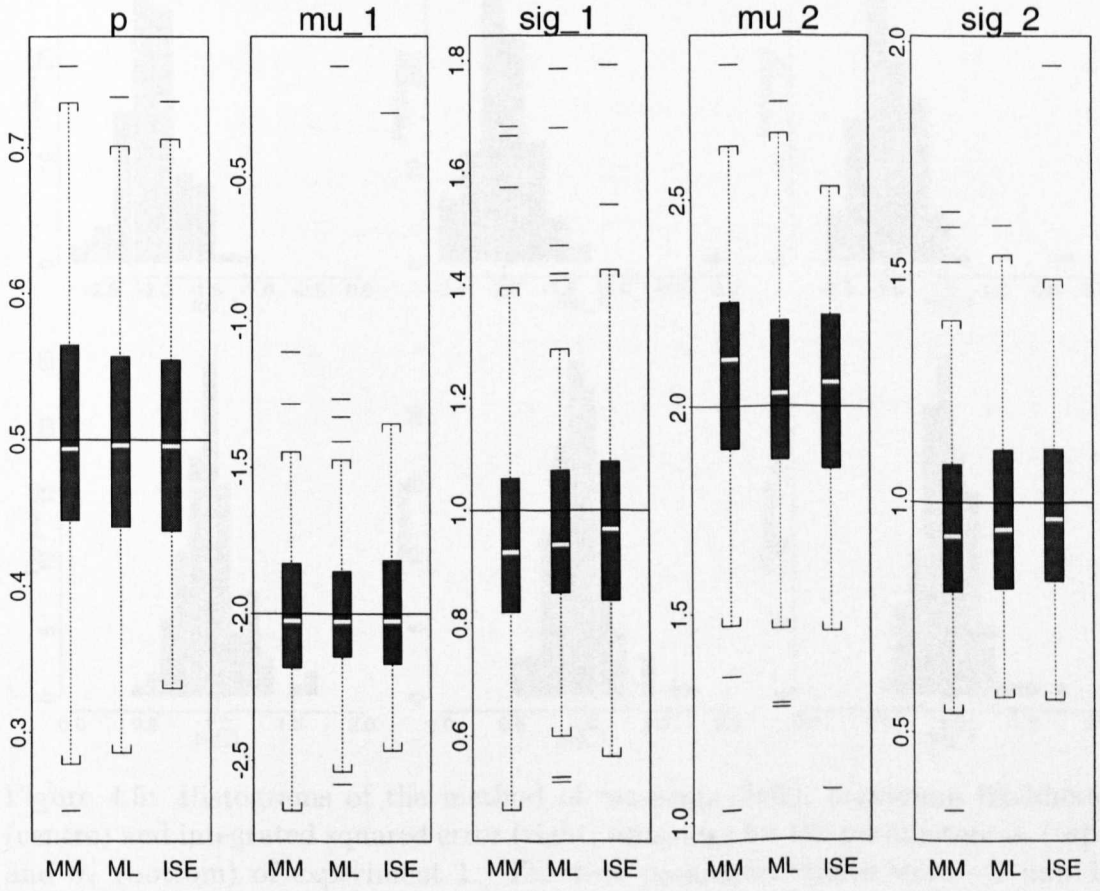

Figure 4.4: Boxplots of the method of moments (MM), maximum likelihood (ML) and integrated squared error (ISE) estimates for the parameters p, μ_1, σ_1, μ_2, and σ_2 of experiment 1. The true parameter values were 0.5, -2, 1, 2, and 1, respectively, and are shown in the boxplots as horizontal lines.

As observed in the figure, the performances of the three estimators were comparable. For example, the means and variances of these parameter estimates were essentially identical. As a result, there is little to choose amongst the three methods, hence the MM may be preferred.

We have found it instructive to compare the distribution of the MM estimates with the distributions of the ML and ISE estimates. The general shape of these distributions can be inferred from the boxplots, but the commonly used diagnostic technique is the histogram. Figure 4.5 depicts histograms of the estimates for the parameters μ_1 and σ_1 obtained from each method.

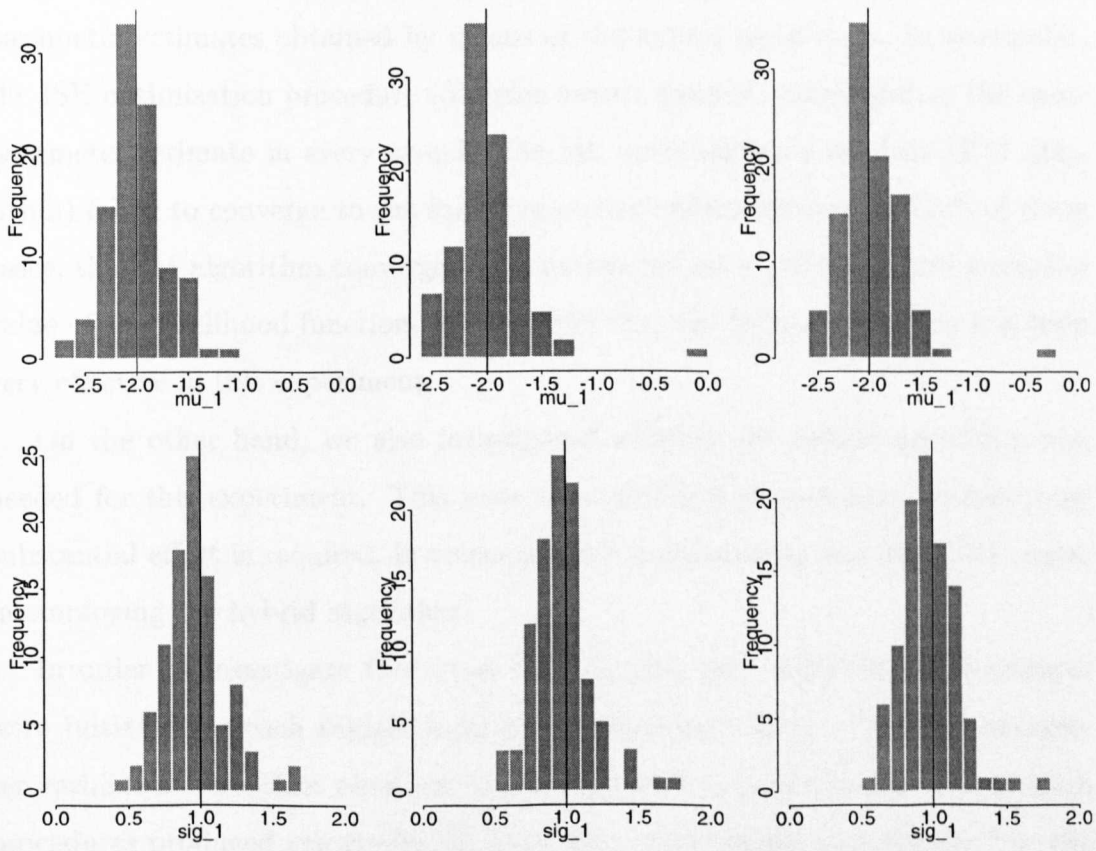

Figure 4.5: Histograms of the method of moments (left), maximum likelihood (centre) and integrated squared error (right) estimates for the parameters μ_1 (top) and σ_1 (bottom) of experiment 1. The true parameter values were -2 and 1, respectively, and are shown in the histograms as vertical lines.

The shapes of the histograms are as expected. They are evidently of approximately "normal" shape and noticeably centred around the true parameter values.

The histograms of the remaining parameters are totally analogous to those of Figure 4.5 and have been omitted. Such results provide an indication that for this sample size and parameter set, the usual asymptotic approximations can be quite accurate. Thus, in theory the ML method looks preferable.

Another issue we investigated was the impact the hybrid algorithm has had on the ML and ISE parameter estimates. In order to assess this impact, the ML and ISE optimisation procedures were initiated only once for each sample, with the true parameters values as the starting values.

The optimisation procedures converged in a normal way for all the samples. The convergence from the true parameter values was almost always to the same parameter estimates obtained by means of the hybrid algorithms. In particular, the ISE optimisation procedure (Simplex search method) converged to the same parameter estimate in every sample; the ML optimisation procedure (EM algorithm) failed to converge to the same parameter estimate twice. In both of these cases, the EM algorithm converged to a parameter estimate which had a smaller value of the likelihood function. This implies that the hybrid algorithm has been very effective in this experiment.

On the other hand, we also investigated whether the hybrid algorithm was needed for this experiment. This issue is of practical importance, because very substantial effort is required, in terms of both programming and also CPU time, in employing the hybrid algorithm.

In order to investigate this issue, the ML and ISE optimisation procedures were initiated for each sample from several starting values. The final parameter estimates were then obtained and compared. In general, the optimisation procedures produced essentially the same estimates, which were identical to the estimates previously obtained. This implies that good starting values are not that important for this experiment, hence the longer CPU time required by the hybrid algorithms cannot be justified.

In terms of practical performance, the main findings from experiment 1 may be summarised as follows:

1. the methods of moments, maximum likelihood and integrated squared error provide comparable parameter estimates;

2. the usual asymptotic approximations can be quite accurate;

3. the choice of starting values does not appear to be crucial in maximum likelihood and integrated squared error estimation.

4.11.2 Experiment 2 ($p = 0.5, \mu_1 = -1, \sigma_1 = \sqrt{3}, \mu_2 = 1, \sigma_2 = 1$)

At this point, we may wonder what it is about the mixture distribution of experiment 1 that leads to accurate parameter estimates, and to what extent this useful result will hold more widely. Previous work by *Hosmer (1973)* has given some evidence that parameter estimation may be unreliable for sample sizes $n \leq 300$ and values of $\boldsymbol{\theta}$ such that $|\mu_1 - \mu_2| \leq 3\min(\sigma_1, \sigma_2)$. The need for either large samples or well-separated components has been noted by other authors, including *Day (1969)*, *Fryer and Robertson (1972)*, and *Hosmer and Dick (1977)*.

In order to investigate the extent to which poorly separated component densities affect the performance of each method, mixture distributions involving varying degrees of separation need to be studied. We start by halving the degree of separation (compared to Experiment 1) while keeping the mixing proportion unchanged. Summary numerical details from fitting a mixture distribution which fits this description can be found in Figure 4.6.

In contrast to experiment 1, a range of performance was now observed. Starting from the frequency of failures, the MM experienced severe problems in this experiment. In particular, the computation of the MM estimates failed a total of 28 times. On the contrary, neither the computation of the ML estimates nor the computation of the ISE estimates encountered any failures. In this respect, the hybrid algorithm has worked very well.

Turning to the values of the obtained parameter estimates, the performance of the MM does not improve. This method performs poorly because the corresponding estimates have the highest bias overall. For example, in Figure 4.6 observe

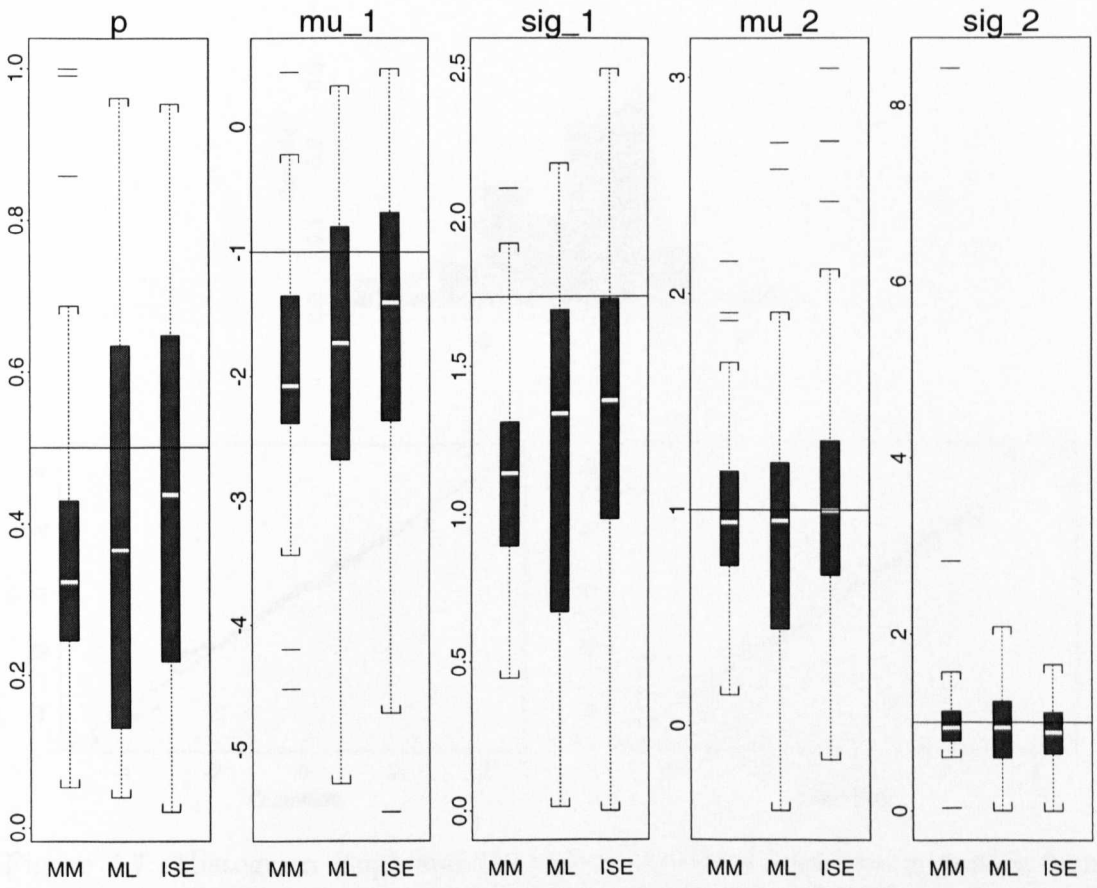

Figure 4.6: Boxplots of the method of moments (MM), maximum likelihood (ML) and integrated squared error (ISE) estimates for the parameters p, μ_1, σ_1, μ_2, and σ_2 of experiment 2. The true parameter values were 0.5, -1, 1.73, 1, and 1, respectively, and are shown in the boxplots as horizontal lines.

that the median of the MM estimates is generally further from the true parameter value than the median of either the ML or the ISE estimates. Such results tend to indicate that the MM is, in general, inadequate for estimating the parameters of the present mixture distribution.

Concentrating on the performances of the ML and ISE estimators, of particular importance were the samples in which the computation of the MM estimates failed. In general, we found the performances of the estimators to be analogous, producing parameter estimates which provided good fits to the sample distributions. For example, Figure 4.7 shows a histogram of one such sample with the fitted mixture densities superimposed. Also shown in the figure are Q-Q plots confirming the adequacy of the fitted densities.

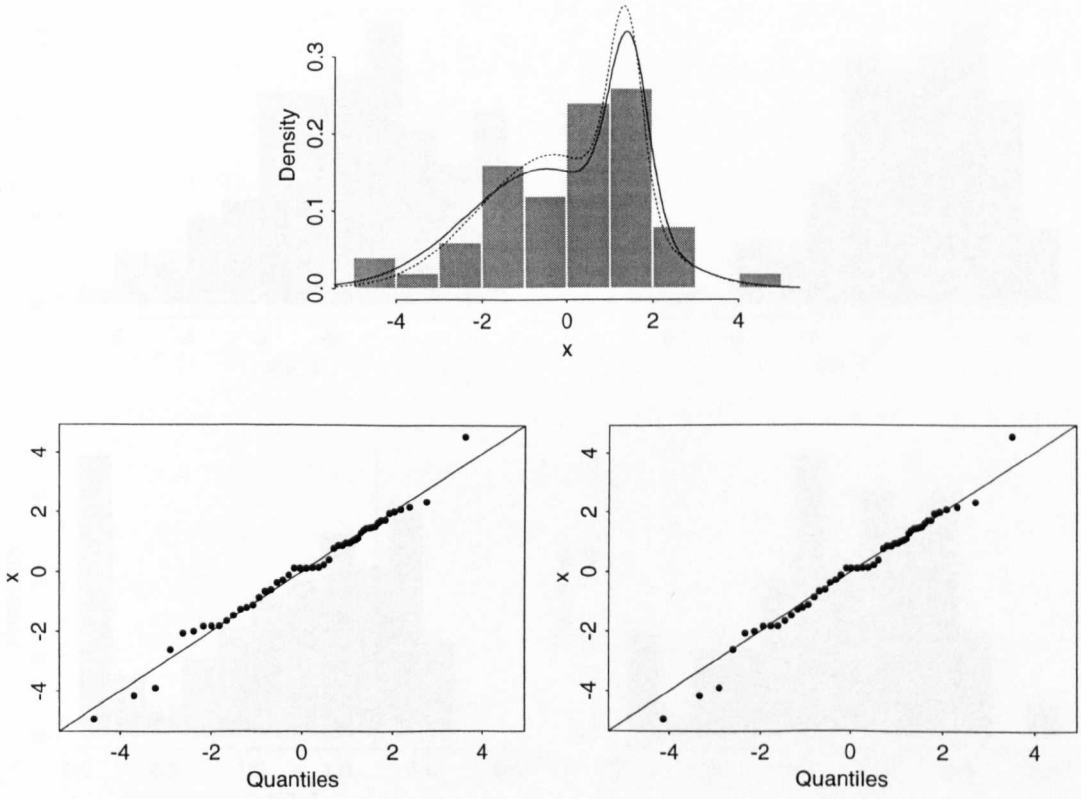

Figure 4.7: Histogram (top) and *Q-Q* plots (bottom) based on a sample from experiment 2 in which the computation of the MM estimates failed. The histogram is overlaid with density functions fitted by maximum likelihood (solid line) and integrated squared error (dotted line). The *Q-Q* plot on the left is for the density fitted by maximum likelihood, and the *Q-Q* plot on the right is for the density fitted by integrated squared error.

As observed in the figure, the fitted densities are in fair agreement with each other. However, the differences between the fitted densities are considerably smaller than the differences between the corresponding parameter estimates. In fact, for the sample size and parameter set studied here, the variances of the parameter estimators are so large that the parameter estimates are probably of limited practical value, as also suggested by *Leytham (1984)*. An indication of the severity of this problem is provided in Figure 4.8. This figure presents the histograms of the ML and ISE estimates for the parameters μ_1 and σ_1.

In addition to the high variability of the parameter estimates, there are two features of the histograms that give reason for concern. The first feature regards

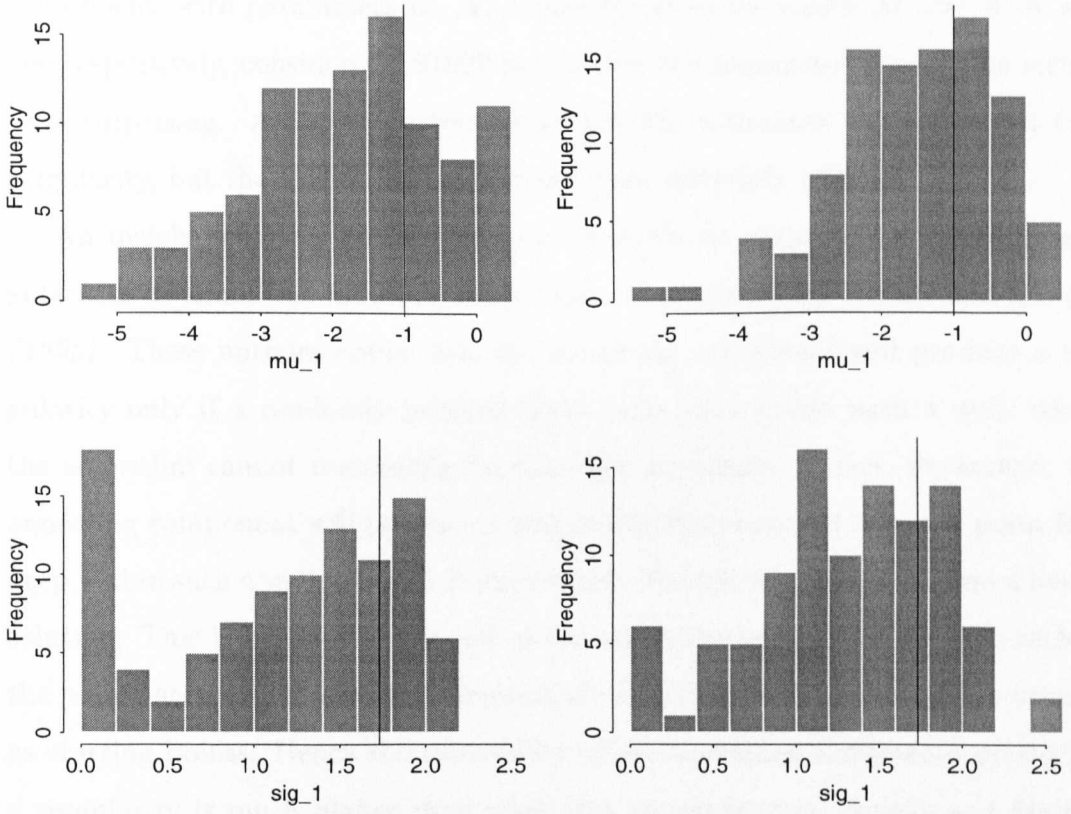

Figure 4.8: Histograms of the maximum likelihood (left) and integrated squared error (right) estimates for the parameters μ_1 (top) and σ_1 (bottom). The true parameter values were -1 and 1.73, respectively, and are shown in the histograms as vertical lines.

their "lack of normality". This indicates a weak relationship between finite sample and asymptotic properties. The weak relationship may be due to several reasons, such as (1) the precision of the asymptotic approximation, (2) the smallness of the sample size, (3) the presence of outliers amongst the estimates, or (4) the acceptance of estimates that do not truly correspond to the global optimum.

The second feature, which is perhaps the more important of the two, regards the number of parameter estimates in which one of the estimated standard deviations was near zero (singularity). Such estimates have a long history in normal mixture problems and have been associated with the presence of outliers, or the use of unfavourable starting values (see, for example, *Titterington, Smith and Makov, 1985, p. 94*). In order to eliminate the latter possibility, we employed a second hybrid algorithm involving a very extensive annealing component. This annealing

component, with parameters T_0, N, ρ, and N_t set to the values 20, 500, 0.98, and 500 respectively, considered 250 000 points over the parameter space. The results were surprising. A higher percentage of the ML estimates had converged to a singularity, but the ISE estimates had not been adversely affected.

An insight into this performance can perhaps be obtained by regarding singularities as small but infinitely deep wells, as suggested by *Brooks and Morgan (1995)*. These authors noted that the annealing component will produce a singularity only if a randomly selected point falls deep within such a well, where the algorithm cannot reasonably be expected to escape. In fact, we amend, the annealing component will produce a singularity if a randomly selected point falls deep within such a well, where it is improbable that the algorithm will find a better solution. This is because at the end of the annealing component we take each of the points accepted at the final temperature, together with the best point overall, as starting points. Hence the probability of the annealing component producing a singularity is much higher than what was anticipated by *Brooks and Morgan (1995)*, and increases considerably with the number of new points considered. We believe this explains the performance of the second ML hybrid algorithm.

In contrast, this problem does not arise for the ISE hybrid algorithm since the ISE function does not contain singularities. To illustrate this, consider Figure 4.9, which displays the ISE function, as a function of σ_1, for a particular sample from experiment 2.

As observed in the figure, there is a well-defined minimum near $\sigma_1 = 0.8$. However, the main message of the figure is that as $\sigma_1 \to 0$, the value of the ISE function is not affected. This explains the performance of the second ISE hybrid algorithm.

In terms of practical performance, the main findings from experiment 2 may be summarised as follows:

1. the method of moments is, in general, inferior to the methods of maximum likelihood and integrated squared error;

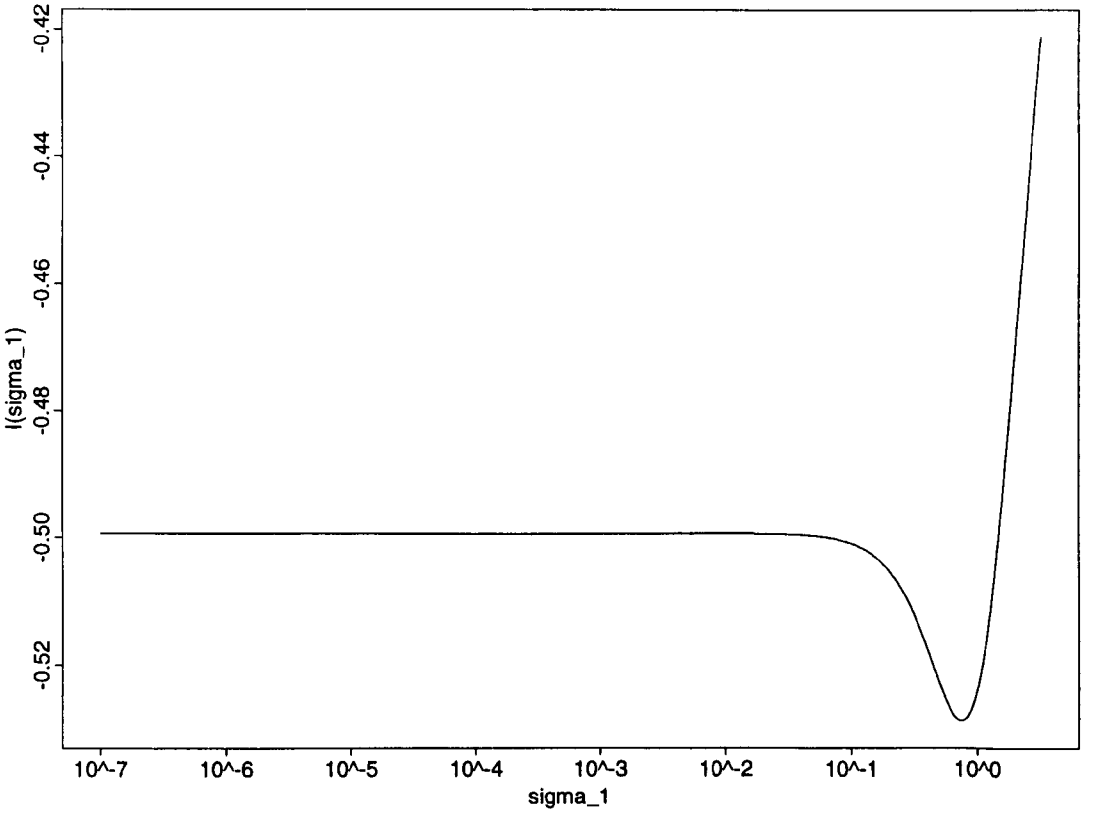

Figure 4.9: The integrated squared error function (3.18) evaluated at $p = 0.5$, $\mu_1 = -1$, $\mu_2 = 1$, $\sigma_2 = 1$, and $\lambda = 0.73$ for a sample from experiment 2.

2. the usual asymptotic approximations can be quite inaccurate;

3. good starting values are substantially easier to find for integrated squared error than for maximum likelihood estimation.

4.11.3 Experiment 3 ($p = 0.5, \mu_1 = 0, \sigma_1 = \sqrt{3}, \mu_2 = 3, \sigma_2 = 1$)

The results of experiments 1 and 2 are consistent with the intuitive idea that decreasing the separation between the component densities produces a deterioration in the results for all methods. Of course, some methods were less affected than others and in real terms the ISE method performed best and the method of moments worst. However, it would be presumptuous to evaluate each method without additional empirical evidence. The numerical details which are presented below have been obtained from experiment 3.

As usual, the MM exhibited a very high failure rate. More specifically, the computation of the MM estimates failed in 21 cases. On the contrary, neither the ML nor the ISE hybrid algorithms produced any failures. This is a typical characteristic of the hybrid algorithm.

In addition to the high failure rate, the MM estimates, when they existed, were generally inferior to either the ML or the ISE estimates. This is illustrated in Figure 4.10 below, where the MM estimates are shown to have the highest bias overall. As commented in the previous section, such results indicate that the MM is also inadequate for estimating the parameters of this mixture distribution.

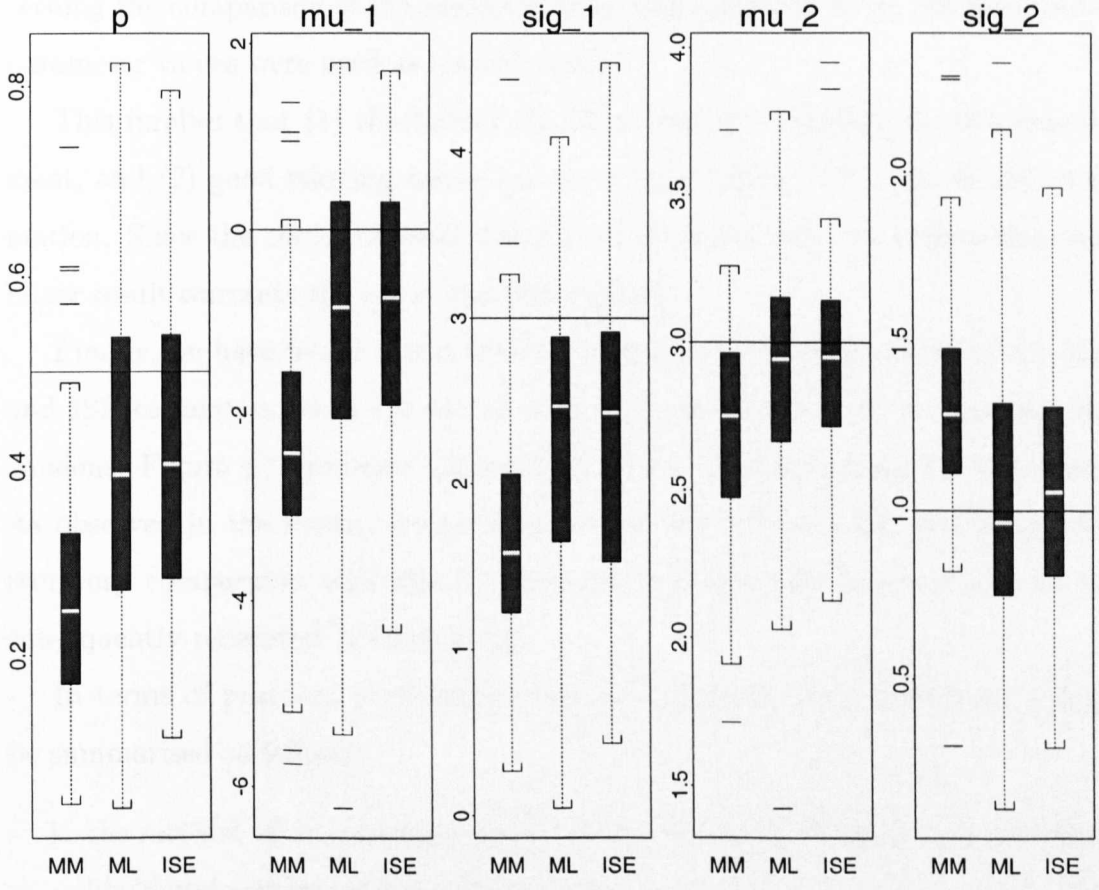

Figure 4.10: Boxplots of the method of moments (MM), maximum likelihood (ML) and integrated squared error (ISE) estimates for the parameters p, μ_1, σ_1, μ_2, and σ_2 of experiment 3. The true parameter values were 0.5, 0, 3, 3, and 1, respectively, and are shown in the boxplots as horizontal lines.

On the other hand, the performances of the ML and ISE estimators were comparable. In fact, based on the sample variances of these estimators it is difficult

to choose one estimator rather than another. However, if ease of computation is an issue, then the ISE estimator may be preferred since it is less dependent upon good starting values. In particular, when the EM algorithm was initiated from the true parameter values, the resulting ML estimates differed to the estimates obtained by means of the hybrid algorithm in 17 cases, in all of which it was the EM algorithm estimates that had a smaller value of the likelihood function. The corresponding number of cases for the ISE estimates was just 9; in these cases, the Simplex search method converged to a parameter estimate which had a greater value of the integrated squared error function. Equivalent conclusion concerning the comparison of the methods were obtained when other less favourable parameter values were used as starting values.

This implies that (1) the hybrid algorithm was also effective for this experiment, and (2) good starting values are more important in ML than in ISE estimation. Since the choice of good starting values is not easy, we believe that the latter result warrants the use of the ISE method.

Finally, we have found it instructive to compare the performances of the ML and ISE estimators when the true parameter values were used to start the recursions. Figure 4.11 presents the boxplots of the resulting parameter estimates. As observed in the figure, the performances of the ML and ISE estimators are still very comparable, and this is consistent with the simulation results to be subsequently presented in Section 4.12.

In terms of practical performance the main findings from experiment 3 may be summarised as follows:

1. the method of moments generally underperforms the methods of maximum likelihood and integrated squared error;

2. the methods of maximum likelihood and integrated squared error provide comparable parameter estimates;

3. the choice of starting values is more important in maximum likelihood than in integrated squared error estimation.

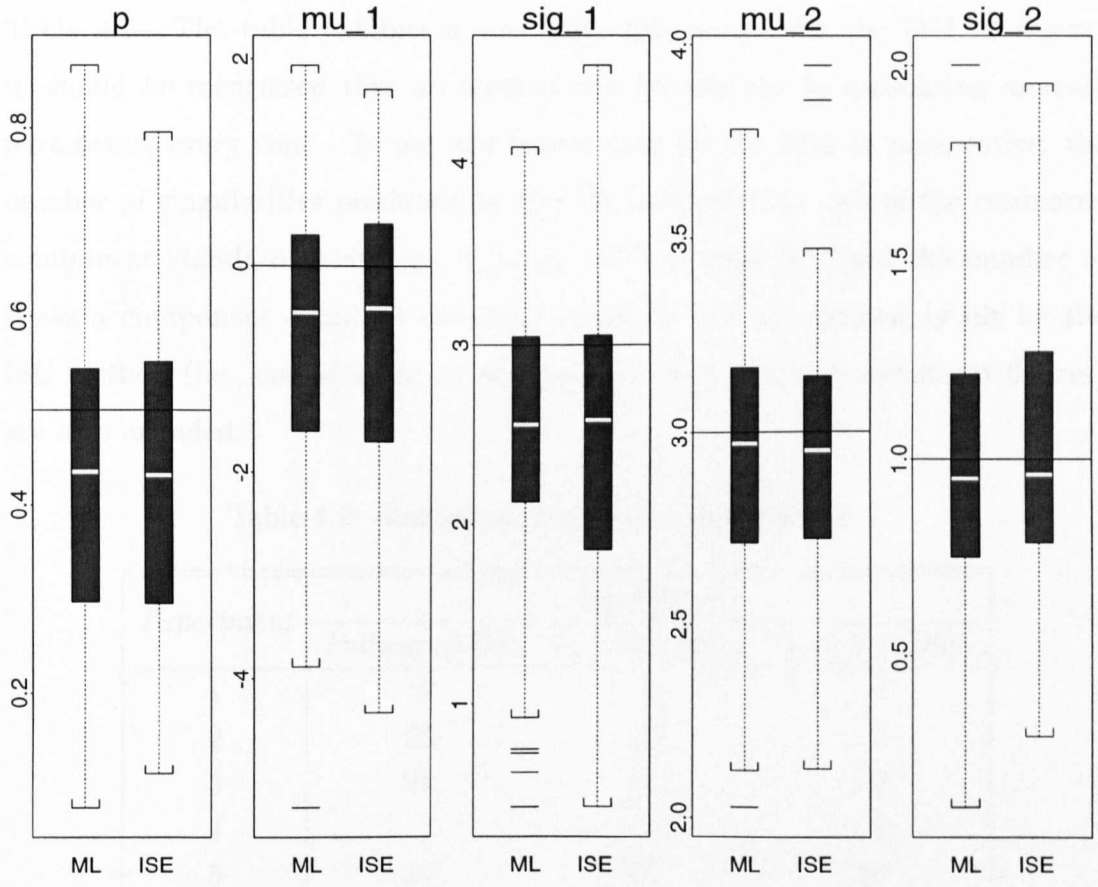

Figure 4.11: Boxplots of the maximum likelihood (ML) and integrated squared error (ISE) estimates for the parameters p, μ_1, σ_1, μ_2, and σ_2 of experiment 3 derived by using the true parameter values. The true parameter values were 0.5, 0, 1.73, 3, and 1, respectively, and are shown in the boxplots as horizontal lines.

4.11.4 Experiments 4–7 and evaluation

As indicated in Section 4.4, the difficulty in the estimation of the mixture parameters depends on the separation between the components and the mixing proportion. Experiments 1–3 examined the performances of the methods under varying degrees of separation between the two components. Experiments 4–7 then examine the performances of the methods under changes in both separation and mixing proportion.

In general, the patterns found in experiments 1–3 were also present in these experiments. For this reason we present only a limited summary of our results. The frequency of failures in the computation of the MM estimates is given in

Table 4.2. The table exhibits a very high failure rate for the MM. However, it should be recognised that no method can be effective in estimating mixture parameters every time. To put the failure rate for the MM in perspective, the number of singularities produced by the ML method (i.e., one of the estimated component standard deviations $\hat{\sigma}_1$ or $\hat{\sigma}_2$ was less than 0.1) and the number of times a component standard deviation endpoint was approximately hit by the ISE method (i.e., one of $\hat{\sigma}_1$ or $\hat{\sigma}_2$ was equal to zero to 4 or 5 significant figures) are also included.

Table 4.2: Simulation details of experiments 1–7

Experiment	*Number of*		
	Failures (MM)	$\hat{\sigma}_j < 0.1$ (ML)	$\hat{\sigma}_j \approx 0.0$ (ISE)
1	2	0	0
2	28	22	8
3	21	1	0
4	7	1	0
5	31	26	20
6	36	14	14
7	11	6	8

These details for the methods of ML and ISE do not correspond to failures in the computation of the parameter estimates, and hence are not totally comparable to the MM failures. However they can be viewed as difficulties encountered by these methods, and so Table 4.2 can be used as a basis to compare the three methods. Based on these results, it is clear that the method of ISE had the best overall performance and the MM the worst.

Additional information about the performance of each method is provided in Figures 4.12–4.15. These figures present the boxplots of the MM, ML and ISE parameter estimates obtained in experiments 4–7 respectively.

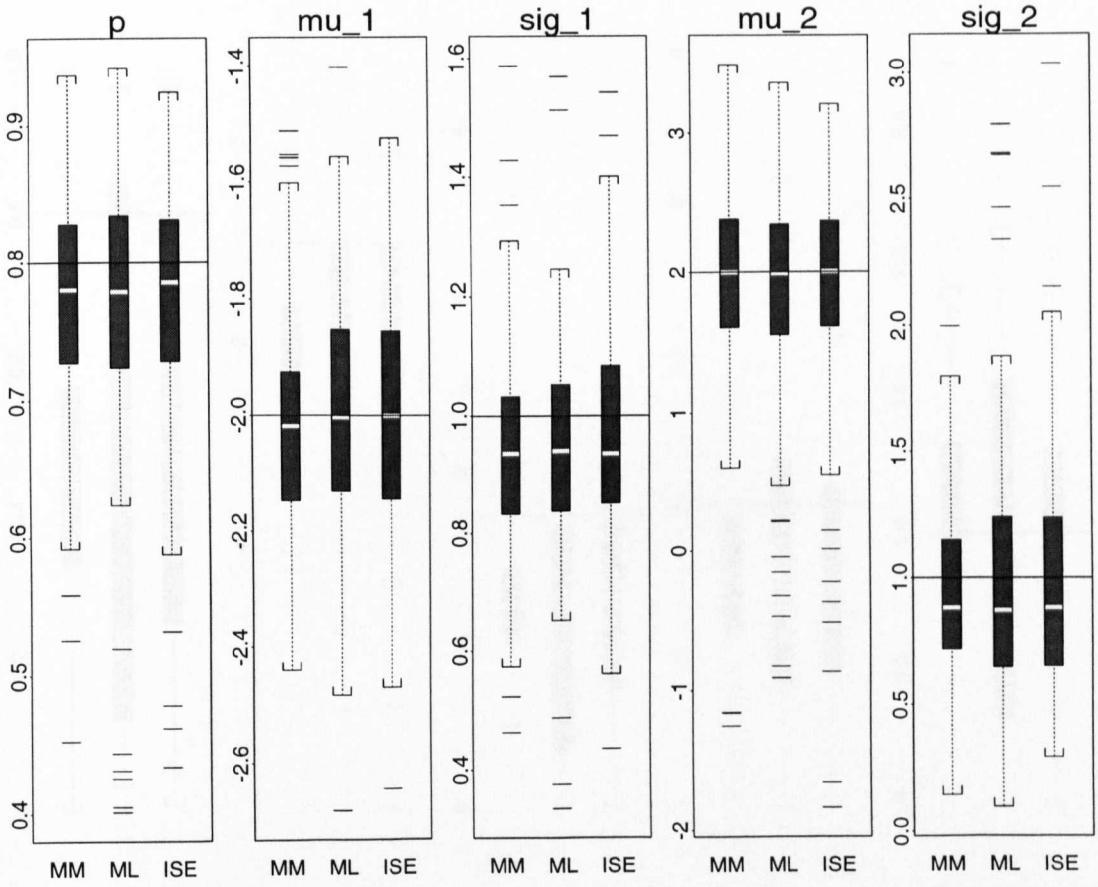

Figure 4.12: Boxplots of the method of moments (MM), maximum likelihood (ML) and integrated squared error (ISE) estimates for the parameters p, μ_1, σ_1, μ_2, and σ_2 of experiment 4. The true parameter values were 0.8, -2, 1, 2, and 1, respectively, and are shown in the boxplots as horizontal lines.

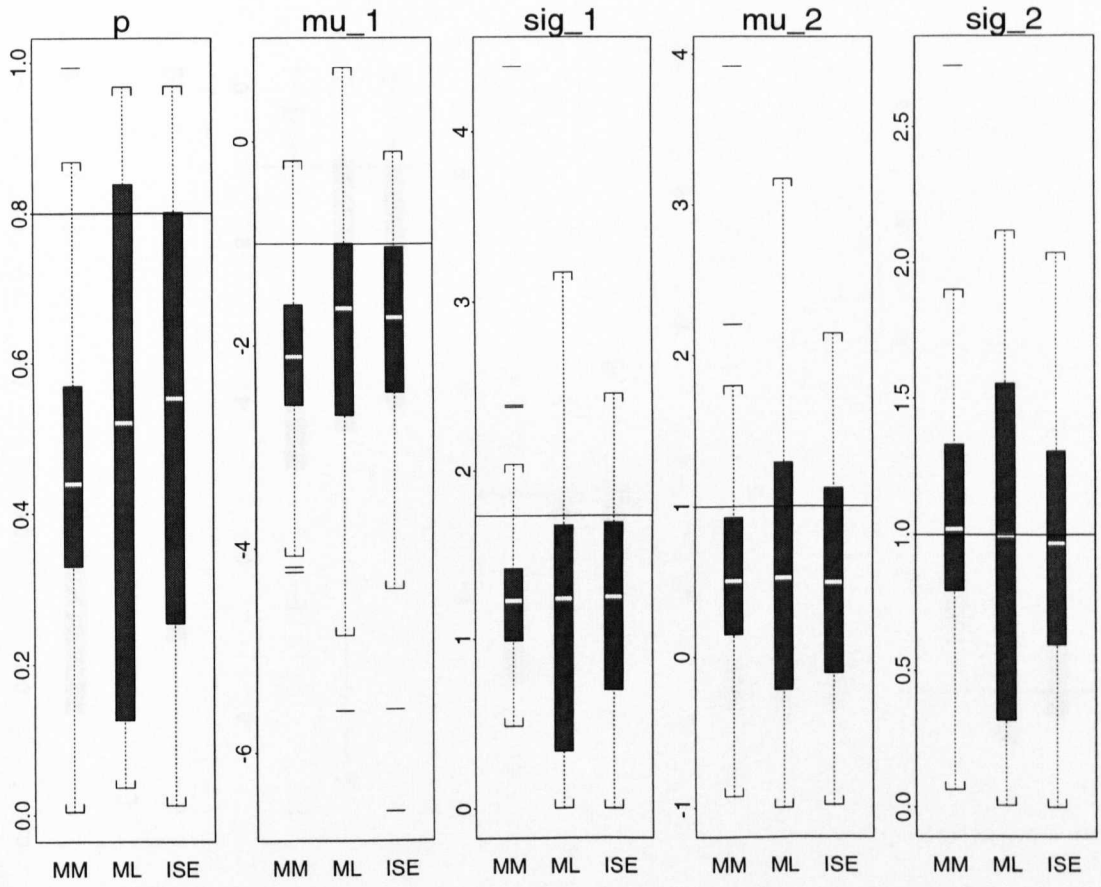

Figure 4.13: Boxplots of the method of moments (MM), maximum likelihood (ML) and integrated squared error (ISE) estimates for the parameters p, μ_1, σ_1, μ_2, and σ_2 of experiment 5. The true parameter values were 0.8, -1, 1.73, 1, and 1, respectively, and are shown in the boxplots as horizontal lines.

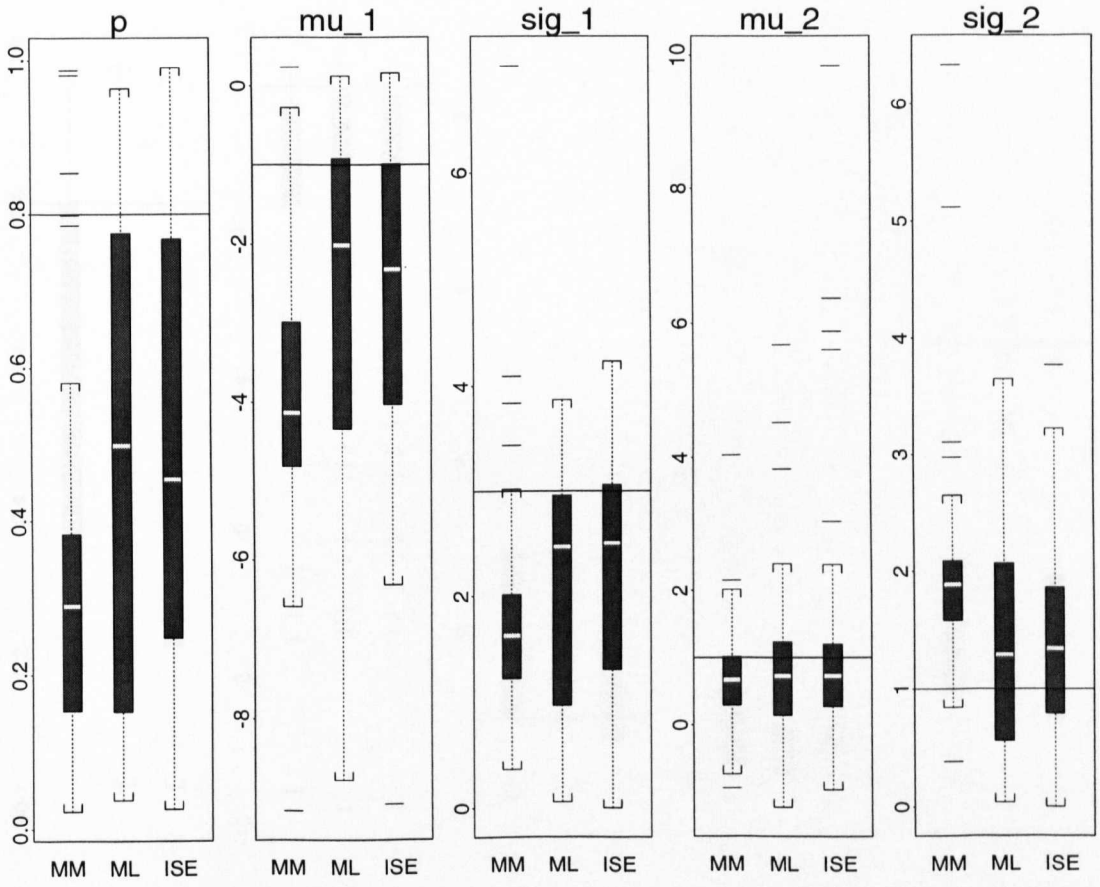

Figure 4.14: Boxplots of the method of moments (MM), maximum likelihood (ML) and integrated squared error (ISE) estimates for the parameters p, μ_1, σ_1, μ_2, and σ_2 of experiment 6. The true parameter values were 0.8, -1, 3, 1, and 1, respectively, and are shown in the boxplots as horizontal lines.

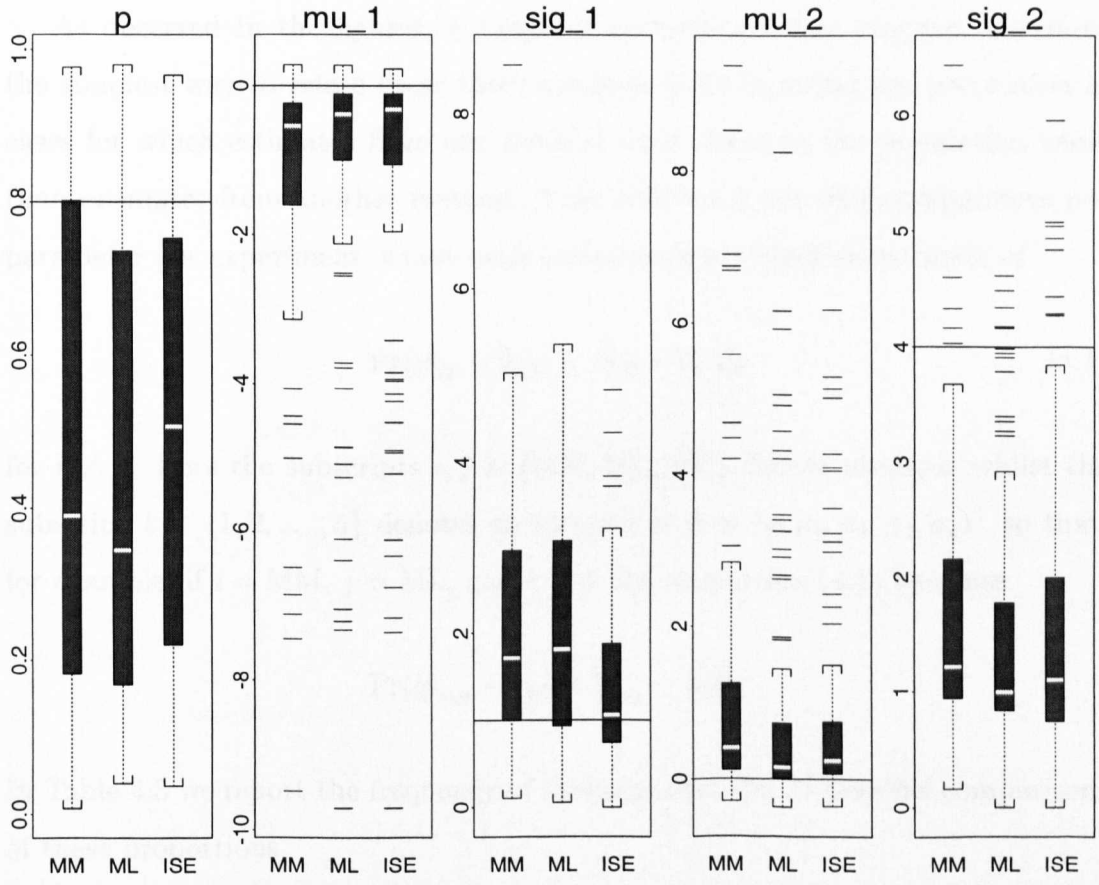

Figure 4.15: Boxplots of the method of moments (MM), maximum likelihood (ML) and integrated squared error (ISE) estimates for the parameters p, μ_1, σ_1, μ_2, and σ_2 of experiment 7. The true parameter values were 0.8, 0, 1, 0, and 4, respectively, and are shown in the boxplots as horizontal lines.

As observed in the figures, a range of performance has emerged. Perhaps the simplest way to relate these three methods is by counting the proportion of cases for which estimates from one method were closer to the population value than estimates from another method. This involved 3 pairwise comparisons per parameter per experiment, where each comparison provided an estimate of

$$\Pr(|\hat{\theta}_{ik} - \theta_{k;0}| < |\hat{\theta}_{jk} - \theta_{k;0}|) \tag{4.1}$$

for $i \neq j$. Here the subscripts $i, j \in \{\text{MM}, \text{ML}, \text{ISE}\}$ denote methods whilst the subscript $k \in \{1, 2, \ldots, 5\}$ denotes an element of $\boldsymbol{\theta} = (p, \mu_1, \sigma_1, \mu_2, \sigma_2)^\top$ so that, for example, if $i = \text{MM}$, $j = \text{ML}$, and $k = 1$ the proportion (4.1) becomes

$$\Pr(|\hat{p}_{\text{MM}} - p_0| < |\hat{p}_{\text{ML}} - p_0|).$$

In Table 4.3 we report the frequency of rankings over the 35 possible comparisons of these proportions.

Table 4.3: Frequency of rankings for experiments 1–7

Method	*Rank*		
	1	2	3
MM	8	3	24
ML	10	16	9
ISE	17	16	2

The comparisons of Table 4.3 have been based on the samples from experiments 1–7 in which the computation of the MM estimates was successful. Although this has lead to an overly optimistic assessment of the MM, it exhibited the worst overall performance. This is consistent with what has been reported in the literature (see, for example, *McLachlan and Basford, 1988, p. 4*) about the performance of the MM.

Of the ML and ISE methods, the ML method was better in only 10 of the

35 possible comparisons. To provide insight into what affects their relative performance, several individual cases were examined in detail. In most cases, the ML and ISE estimates gave rise to densities which were of nearly equal quality, as observed in Figure 4.7. A difference in performance unfolded when either estimate involved a value of less than 0.1 for one of the standard deviations. A feature of all the estimates falling in this category was that the value of the mixing proportion was either very small (i.e., less than 10%) or very large (i.e., greater than 90%). A second feature exclusive to the ISE estimates was that almost all of these very small standard deviations were actually equal to zero to 4 or 5 significant figures. In other words, the ISE method visually fitted a "single-component" mixture rather than a two-component mixture. This feature was not common amongst the ML estimates for the following reason.

The ML estimator often interpreted an extreme observation as being the only sample value from one of the populations with the remaining observations belonging to the other. The estimator would therefore fit a two-component mixture with one component centred on the extreme observation and the other weighted in favour of the remaining observations. This illustrates the lack of robustness of the ML estimator.

Unlike the ML method, the ISE method places direct emphasis on the robustness of the resulting estimator. This implies that observations assessed as atypical are automatically given reduced weight in the computation of the parameter estimates. Thus, mixture distributions involving components centred on extreme observations are rarely fitted. In addition to robustness, the ISE method is not troubled by singularities. This was shown graphically in Figure 4.9, where we observed that letting $\sigma_1 \to 0$ did not affect the ISE function. This implies that a single-component mixture would be fitted only if the parameter set minimising the ISE function involved an approximately zero standard deviation. Visual insight into these effects can be obtained from Figure 4.16, whereas an indication of the frequency of singularities and single-component mixtures in experiments 1–7 was given in Table 4.2.

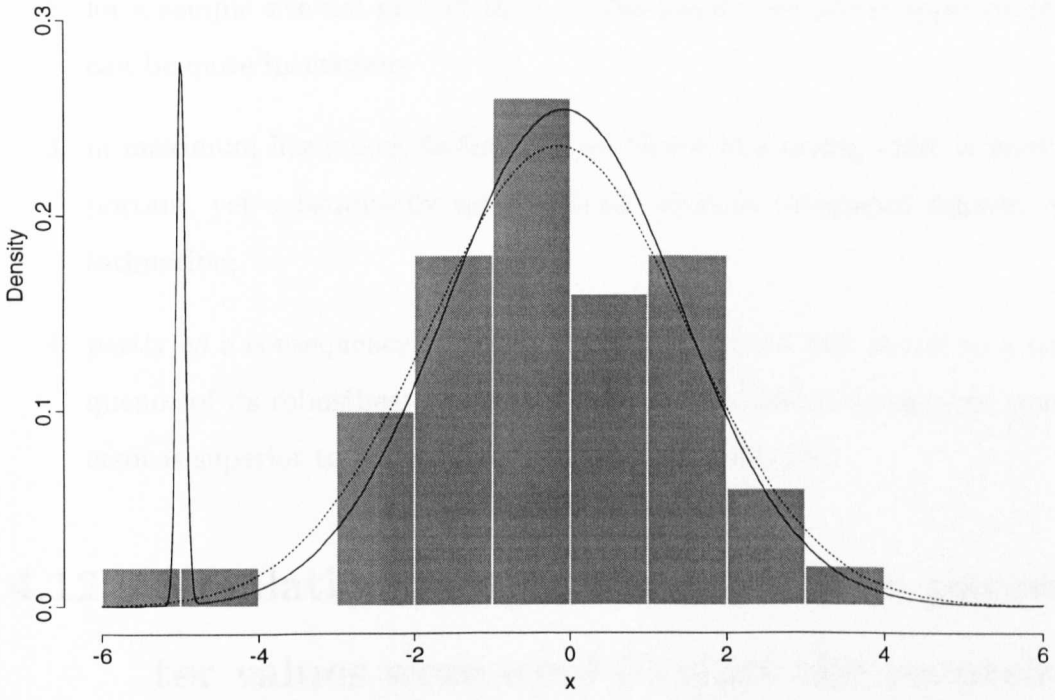

Figure 4.16: Histogram of a sample from experiment 2 overlaid with two density functions fitted by maximum likelihood (solid line) and integrated squared error (dotted line).

Figure 4.16 suggests that the ML estimates are likely to be more variable than their ISE counterparts. The results of the simulation study verify that this is indeed the case, particularly when the estimation of the parameters was difficult. These cases account for the Table 4.3 superiority of the ISE method.

A problem common to both ML and ISE methods is that for the sample sizes and parameter sets considered here the variances of the parameter estimators were very large. This feature may well be due, in general, to the presence of a few outliers among the estimates and was also observed by *Leytham (1984)*. In fact, it is precisely for this reason that comparisons in terms of the proportions (4.1) and not mean squared errors were carried out in Table 4.3.

In terms of practical performance, the main findings from this simulation study may be summarised as follows:

1. the method of moments is, in general, inferior to the methods of maximum likelihood and integrated squared error;

2. for a sample size not greater than 50, the usual asymptotic approximations can be quite inaccurate;

3. in maximum likelihood, finding a good choice of starting value is more important, yet substantially more difficult, than in integrated squared error estimation;

4. partly as a consequence of points (2) and (3) above and partly as a consequence of its robustness against outliers, the method of integrated squared error is superior to the method of maximum likelihood.

4.12 Simulation results when the true parameter values were used to start the recursions

In the preceding section, the ML and ISE estimates were obtained by means of the hybrid algorithm, since the true parameter vales were assumed unknown. While this is most appropriate in practice, it is also appropriate to investigate the impact the hybrid algorithm has had on the relative performance of the ML and ISE methods. This is a two-fold investigation consisting of the impact of

1. an extensive hybrid algorithm;

2. a limited hybrid algorithm.

Concerning the impact of (1) above, we do not feel motivated to investigate this further since, as observed in experiment 2, an extensive hybrid algorithm generally deteriorates the performance of the ML estimator. However the impact of (2) would be worth investigation. This is because in many practical applications limited hybrid algorithms may be the ones of choice for reasons of economy, in money, time, and effort. In this respect, we decided to initiate the ML and ISE optimisation procedures only once, with the true parameter values as the starting values. We recognise that this is likely to lead to an overly optimistic assessment of the performance of either estimator, but this does not matter here since we are

only interested in their relative performance. The frequency of failures and some additional details for the ML and ISE methods are presented in Table 4.4 below. For comparative purposes, the frequency of failures for the MM is also included.

Table 4.4: Failure details of experiments 1–7 when true parameter values were used to start the recursions

Experiment	*Number of*				
	MM	ML		ISE	
	Failures	*Failures*	$\hat{\sigma}_j < 0.1$	*Failures*	$\hat{\sigma}_j \approx 0.0$
1	2	0	0	0	0
2	28	0	4	0	7
3	21	0	0	0	0
4	7	0	0	0	0
5	31	4	3	0	16
6	36	1	4	0	15
7	11	2	4	0	3

As observed in the table, the computation of the ML estimates occasionally resulted in a failure. In addition to failures, the ML estimator was prone to singularities despite the fact that the true parameter values were used to start the recursions. In contrast, the computation of the ISE estimates did not result in any failures. However the ISE estimator did fit several single-component mixtures. On the other hand, in these cases the data may be such that a mixture of two normals is not a plausible model and the ISE estimator is pointing this out. Nevertheless, even if these cases are regarded as failures, the failure rate of either the ML or the ISE method was appreciably lower than that of the MM.

From the point of view of the obtained parameter estimates, the numerical details of this simulation study are summarised in Table 4.5. This table shows the frequency of rankings of the mean squared errors of the MM, ML and ISE estimates for the parameters of the mixture distributions in the seven experiments.

In parallel to Table 4.3, the comparisons of Table 4.5 have been based on the samples in which the computation of all three estimates was successful. In fact,

Table 4.5: Ranking details of experiments 1–7 when true parameter values were used to start the recursions

Method	*Rank*		
	1	2	3
MM	7	3	25
ML	15	14	6
ISE	13	18	4

these were mainly the samples in which the computation of the MM estimates was successful. Nevertheless, the performance of the MM estimator, as judged by these comparisons, was substantially inferior to that of either the ML or ISE estimator. These conclusions are consistent with what has already been observed regarding the inadequacy of the MM to fit mixture data.

Concentrating on the performance of the ML and ISE estimators, the relationship between them was reversed compared to the relationship observed in Section 4.11. In this case, the ML estimator exhibited a better overall mean squared error performance than did the ISE estimator. Similar conclusions were reached by examining the overall proportion by which the ML method produced estimates closer to the true parameter values than the ISE method. However, the difference between the two methods was seldom large and the superiority of the ML method was not completely uniform. Furthermore, as indicated in Section 4.11, the ISE method is more robust against outliers and less dependent upon good starting values than the ML method. In addition, the ISE method is considerably easier to apply in that when good starting values are needed, they are substantially easier to find. In practice, therefore, we recommend fitting a mixture of two normal distributions by the ISE method, particularly with samples with outliers or a small number of observations.

Chapter 5

Least-squares transform estimation

5.1 Introduction

A general introduction to minimum distance estimation involving integral transforms was given in Chapter 1, with emphasis on the integrated squared error method. This method was found to be an attractive alternative to maximum likelihood possessing appealing statistical properties. In particular, the integrated squared error estimator was shown to be consistent, asymptotically normal, and robust against outlying observations. However, the asymptotic efficiency of the estimator has been shown by *Heathcote (1977)* to be generally less than unity and, furthermore, the weight function $W(t;\lambda)$ was open to choice. In the general case, the efficiency of the integrated squared error estimator will be a function of $W(t;\lambda)$, and presumably $W(t;\lambda)$ should be chosen so as to maximise this. Unfortunately, the complicated dependence of the efficiency on $W(t;\lambda)$ suggests that this ideal solution is not practicable. This has led to the mean integrated squared error developments of Chapter 2.

Aside from efficiency and mean integrated squared error considerations, computational complexity may be the most important factor in the choice of weight function. The type of weight function leading to the greatest degree of numerical

simplicity is a monotonic non-decreasing step function. This chapter will investigate some properties of minimum distance estimators which are based on integral transforms and step weight functions. The chapter begins with the motivation for step weight functions, with the moment generating function method being introduced and applied in Sections 5.3–5.4. Two practical issues in the general application of this method are considered in the following two sections, whilst in Sections 5.7–5.9 we introduce three extensions of the moment generating function method (the preferred moment generating function method, the modified moment generating function method, and the q-L method). Finally, we illustrate the performance of the q-L method for estimating the parameters of a Cauchy distribution.

5.2 Motivation for step weight functions

In the context of minimum distance estimation involving integral transforms, it was noted in Chapter 1 that the most commonly used weight functions divide into two basic types:

1. monotonic non-decreasing continuous functions;

2. monotonic non-decreasing step functions.

A comprehensive account of the theoretical and practical issues in estimation utilising the first type of weight function was presented in Chapters 1–4. In general, continuous weight functions seem to produce sensible results.

However, there are three reasons one might want to select a step rather than a continuous weight function. First, the numerical computation of the distance function (1.5) is not generally straightforward when continuous weight functions are used. Of course, one view of current statistical modelling is that intractable distance functions, or likelihoods can be submitted to numerical analytical black-boxes which, with the aid of sophisticated computers, will readily provide parameter estimates and measures of error. However, it is the experience of *Morgan*

(1998, personal communication) that such black-boxes do not always work, and this may lead to promising directions for research being abandoned.

Secondly, there could arise applications in which it is more efficient to use a step rather than a continuous weight function. One such example is the normal distribution discussed in Chapter 1 (Section 1.7), where the weight function $W(t;\lambda) = \int_{-\infty}^{t} e^{-\lambda^2 y^2} \mathrm{d}y$ was used. In particular, we found that the efficiency of the resulting estimator improved with increases in λ. As λ increases, this weight function approaches the shape of a step function, as illustrated in Figure 5.1 below.

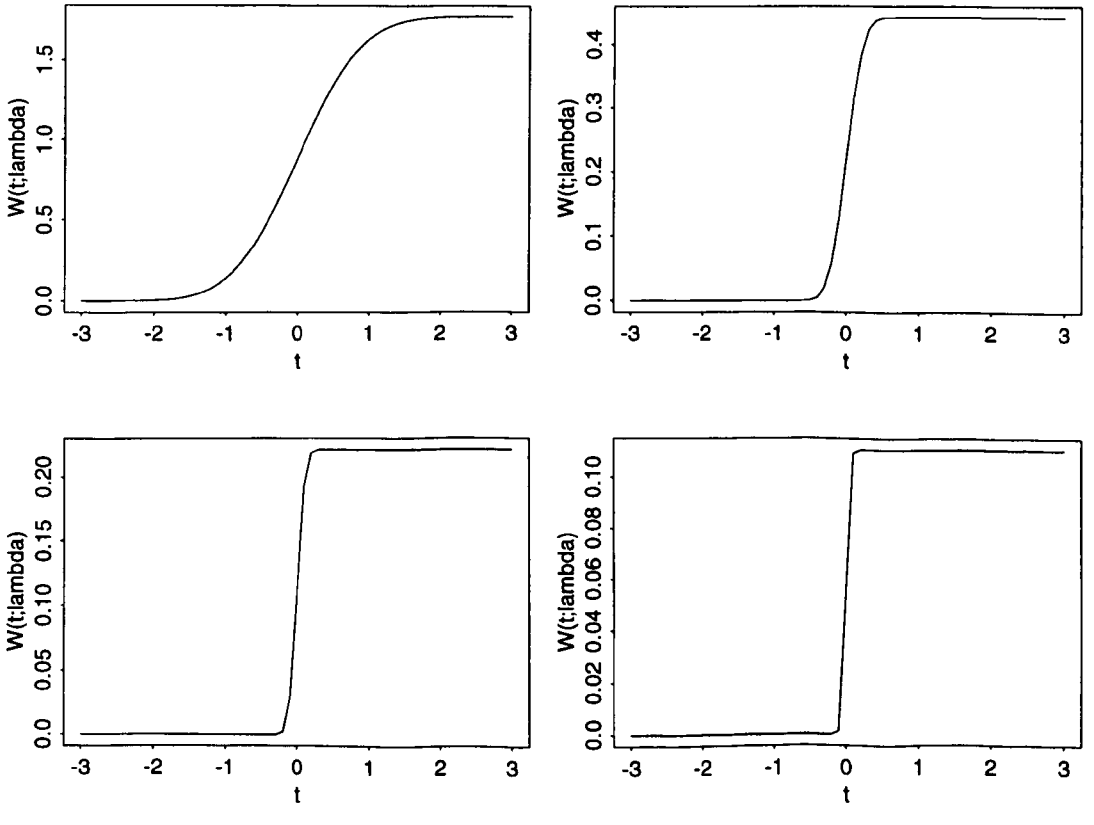

Figure 5.1: The weight function $W(t;\lambda) = \int_{-\infty}^{t} e^{-\lambda^2 y^2} \mathrm{d}y$ with (reading from left-right, top-bottom): (1) $\lambda = 1$; (2) $\lambda = 4$; (3) $\lambda = 8$; and (4) $\lambda = 16$.

Thirdly, the attraction of being able to obtain explicit estimators might lead one in some applications to prefer step to continuous weight functions. One such example is the stable laws, in which a step weight function enabled *Press (1972)* to obtain explicit parameter estimators.

For these reasons it is important to consider minimum distance estimation utilising step weight functions. This is the approach we have adopted in this chapter.

5.3 The moment generating function estimator

Let $\{F(X;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ again be a family of distribution functions indexed by the parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)^\top$, and suppose that $X_1, X_2, \ldots, X_n$ is a random sample from some population whose distribution function is a member of this family with parameter vector $\boldsymbol{\theta}_0$. In Chapter 1 we considered parameter estimation based on the empirical transform $G_n(t) = n^{-1}\sum_{j=1}^{n} g(t, X_j)$, which involved minimising the integral

$$\Delta[G_n(\,.\,), G(\,.\,;\boldsymbol{\theta})] = \int_T |G_n(t) - G(t;\boldsymbol{\theta})|^2 \,\mathrm{d}W(t) \tag{5.1}$$

with respect to $\boldsymbol{\theta}$. This integral may be regarded as a measure of the deviation between $G_n(t)$ and the transform $G(t;\boldsymbol{\theta}) = \int_{-\infty}^{\infty} g(t,x)\,\mathrm{d}F(x;\boldsymbol{\theta})$, which corresponds to $F(x;\boldsymbol{\theta})$. Previously, the weight function $W(t)$ was a monotonic non-decreasing continuous function. However, for the rest of this section we shall assume it is a step function with equal increments at the points t_j $(j = 1, 2, \ldots, p)$ in T. As a result, the integral in (5.1) becomes

$$\Delta[G_n(\,.\,), G(\,.\,;\boldsymbol{\theta})] \propto \sum_{j=1}^{p} |G_n(t_j) - G(t_j;\boldsymbol{\theta})|^2.$$

The range of possible functions $g(t,x)$ is much as it was in Chapter 1 (Section 1.3). Since in this method it is customary to employ real-valued functions, a typical choice for $g(t,x)$ is $g(t,x) = \exp(tx)$. Then $G(t;\boldsymbol{\theta})$ and $G_n(t)$ are the moment generating function (mgf) $M(t;\boldsymbol{\theta}) = \int \exp(tx)\,\mathrm{d}F(x;\boldsymbol{\theta})$, when it exists, and the empirical mgf $M_n(t) = n^{-1}\sum_{j=1}^{n}\exp(tX_j)$ respectively.

The resulting estimator (called by *Quandt and Ramsey (1978)* the moment

generating function estimator) is defined as follows:

Definition 5.1. The moment generating function (MGF) estimator, $\hat{\boldsymbol{\theta}}$, for $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$, is the value of $\boldsymbol{\theta}$ which minimises

$$S(\boldsymbol{\theta}; \boldsymbol{t}) = \sum_{j=1}^{p} [M_n(t_j) - M(t_j; \boldsymbol{\theta})]^2 \tag{5.2}$$

for a suitable value of $\boldsymbol{t} = (t_1, t_2, \ldots, t_p)^\top$ in T^p.

Accordingly, parameter estimation by way of (5.2) may be effected through consideration of the normal equations

$$\sum_{j=1}^{p} \frac{\partial M(t_j; \boldsymbol{\theta})}{\partial \theta_k} [M_n(t_j) - M(t_j; \boldsymbol{\theta})] = 0, \quad k = 1, 2, \ldots, p, \tag{5.3}$$

or, equivalently, by solving the system of equations

$$M_n(t_j) = M(t_j; \boldsymbol{\theta}), \quad j = 1, 2, \ldots, p \tag{5.4}$$

for $\boldsymbol{\theta}$, if such a solution exists. Subject to the existence and uniqueness of this solution, (5.4) demonstrates that the moment generating function estimator essentially equates the empirical and theoretical moment generating functions at the points t_j $(j = 1, 2, \ldots, p)$. This has the intuitively appealing feature of a relatively easy computation problem to overcome. On the other hand, the resulting estimator suffers from arbitrariness in the choice of these points.

This discussion indicates some of the criteria involved in the choice of $\boldsymbol{t}$. Most importantly, the points t_j $(j = 1, 2, \ldots, p)$ must satisfy the requirement that the system (5.3) should be non-singular. In addition, values of t which cause $M(t; \boldsymbol{\theta})$ to become computationally intractable need to be avoided. Further consideration of how to select $\boldsymbol{t}$ must be preceded by a discussion of the properties of the resulting estimator. These properties are subject to certain regularity conditions, and these, along with others which will be required later in this chapter are listed below.

5.3.1 Regularity conditions

Let $\{F(x;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p\}$ be a family of distribution functions, and suppose that the moment generating function $M(t;\boldsymbol{\theta})$, which corresponds to $F(x;\boldsymbol{\theta})$, exists for all $t \in T$. If, for $\boldsymbol{t} = (t_1, t_2, \ldots, t_q)^\top$ in T^q $(q \geq p)$, we define the $q \times p$ matrix $K(\boldsymbol{\theta})$ with (i,j) element

$$\kappa_{ij}(\boldsymbol{\theta}) = \frac{\partial M(t_i;\boldsymbol{\theta})}{\partial \theta_j},$$

and the $q \times q$ matrix $\Omega(\boldsymbol{\theta})$ with (i,j) element

$$\omega_{ij}(\boldsymbol{\theta}) = M(t_i + t_j;\boldsymbol{\theta}) - M(t_i;\boldsymbol{\theta})M(t_j;\boldsymbol{\theta})$$

then the properties of the moment generating function-based estimators require the following conditions:

1. Θ is an open rectangle;
2. $M(t;\boldsymbol{\theta})$ is continuously differentiable (in $\boldsymbol{\theta}$) at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$;
3. if $q = p$ then the matrix $K(\boldsymbol{\theta})$ is invertible at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$;
4. the matrix $K(\boldsymbol{\theta})^\top K(\boldsymbol{\theta})$ is invertible at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$;
5. the matrix $\Omega(\boldsymbol{\theta})$ is invertible at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$;
6. the matrix $K(\boldsymbol{\theta})^\top \Omega(\boldsymbol{\theta}) K(\boldsymbol{\theta})$ is invertible at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

We remark that conditions 3–6 will hold very generally provided that the points t_j $(j = 1, 2, \ldots, q)$ in T are non-zero and distinct.

5.3.2 Properties of the moment generating function estimator

The principal asymptotic properties of an estimator are consistency and asymptotic normality, with a computable covariance matrix. For the moment generating function estimator, these properties have been established by *Quandt and Ramsey*

(1978). In particular, in the context of a mixture of two normal distributions (so that conditions 1–3 are satisfied), they presented:

Theorem 5.1 (Quandt and Ramsey, 1978). *Given a set of non-zero and distinct points t_j $(j = 1, 2, \ldots, p)$ in T, the moment generating function estimator, $\hat{\boldsymbol{\theta}}$, for $\boldsymbol{\theta}_0 \in \Theta \subseteq \mathbb{R}^p$ is strongly consistent and*

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \Sigma(\boldsymbol{\theta}_0)],$$

where

$$\Sigma(\boldsymbol{\theta}) = K^{-1}(\boldsymbol{\theta})\, \Omega(\boldsymbol{\theta})\, K^{-1}(\boldsymbol{\theta})^{\top}. \tag{5.5}$$

In this expression, $K(\boldsymbol{\theta})$ is the $p \times p$ matrix whose (i, j)th element is

$$\kappa_{ij}(\boldsymbol{\theta}) = \frac{\partial M(t_i; \boldsymbol{\theta})}{\partial \theta_j}$$

and $\Omega(\boldsymbol{\theta})$ is the $p \times p$ symmetric matrix whose (i, j)th element is

$$\omega_{ij}(\boldsymbol{\theta}) = M(t_i + t_j; \boldsymbol{\theta}) - M(t_i; \boldsymbol{\theta}) M(t_j; \boldsymbol{\theta}).$$

Theorem 5.1 demonstrates that the asymptotic covariance matrix of the moment generating function estimator is, in principle, straightforward to calculate. This gives it an advantage over the integrated squared error estimator, whose asymptotic covariance matrix is often intractable.

The robustness properties of an estimator are generally described through the behaviour of its influence function, as noted in Chapter 1. This function examines the response of the estimator to an additional observation ξ. For the moment generating function estimator, and under conditions 1–3, *Campbell (1993)* presented the following theorem:

Theorem 5.2 (Campbell, 1993). *Given a set of non-zero and distinct points t_j $(j = 1, 2, \ldots, p)$ in T, the moment generating function estimator, $\hat{\boldsymbol{\theta}}$, for $\boldsymbol{\theta}_0 \in$*

$\Theta \subseteq \mathbb{R}^p$ *has joint influence function*

$$IF(\xi; \hat{\boldsymbol{\theta}}) = K^{-1}(\boldsymbol{\theta}_0)\, \boldsymbol{\tau}(\xi; \boldsymbol{\theta}_0), \tag{5.6}$$

where $K(\boldsymbol{\theta})$ *is the* $p \times p$ *matrix whose* (i,j)*th element is*

$$\kappa_{ij}(\boldsymbol{\theta}) = \frac{\partial M(t_i; \boldsymbol{\theta})}{\partial \theta_j}$$

and $\boldsymbol{\tau}(\xi; \boldsymbol{\theta})$ *is the* $p \times 1$ *vector whose* i*th element is*

$$\tau_i(\xi; \boldsymbol{\theta}) = \exp(t_i \xi) - M(t_i; \boldsymbol{\theta}).$$

It is clear from Theorem (5.2) that the ith ($i = 1, 2, \ldots, p$) individual influence function for the moment generating function estimator may be expressed as

$$IF(\xi; \hat{\theta}_i) = \sum_{j=1}^{p} \kappa^{ij}(\boldsymbol{\theta}_0)[\exp(t_j \xi) - M(t_j; \boldsymbol{\theta}_0)], \tag{5.7}$$

where $\kappa^{ij}(\boldsymbol{\theta})$ is the (i,j)th element of $K^{-1}(\boldsymbol{\theta})$. Since (5.7) is not bounded in ξ, the moment generating function estimator cannot be robust over the entire real line. This is in contrast with the integrated squared error estimator.

5.3.3 Selecting a value for t

The moment generating function method requires a value for $\boldsymbol{t}$ before it can be implemented. There seems to be no clear-cut way for selecting $\boldsymbol{t}$, and as a result several practical approaches have been proposed. Some of the possibilities include:

1. using cross-validation (see *Laurence and Morgan, 1987a, b*);

2. maximising the likelihood function subject to the constraints imposed by (5.4) (see *Laurence, Morgan and Tweedie, 1987; Laurence and Morgan, 1987a, b*);

3. minimising a measure of the asymptotic covariance matrix $\Sigma(\boldsymbol{\theta}_0)$ (see *Read, 1981; Schmidt, 1982; Ball and Milne, 1996*);

4. searching only along, or close to, the "diagonal" line $t_1 = t_2 = \cdots = t_p$ (see *Tweedie, Zhy and Choy, 1995; Yao and Morgan, 1999*).

The four approaches above should, in theory, provide a reasonable value for $\boldsymbol{t}$ in many situations. In practice, they may be criticised as follows. The first approach may involve extensive numerical computations. In the second, the likelihood function may be difficult to produce, while the third possibility may not be suitable for small samples. Finally, in the last approach it is not clear which criterion is being optimised. In fact, based on extensive empirical evidence it has been related by *Yao and Morgan (1999)* to the approach of minimising the determinant of $\Sigma(\boldsymbol{\theta})$.

The moment generating function method can be employed for parameter estimation in any distribution family having a closed form moment generating function, as illustrated in the following section. However, the greatest utility of the method may be with families more complicated than simple location-scale models (see, for example, *Quandt and Ramsey, 1978*).

5.4 Application to the normal distribution

Let $X_1, X_2, \ldots, X_n$ be independent, identically distributed random variables from a normal distribution with mean μ and variance σ^2, and suppose that $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$ is unknown. The moment generating function of X_1 is

$$M(t; \boldsymbol{\theta}) = \exp(t\mu + \sigma^2 t^2/2), \quad t \in \mathbb{R}$$

and, for given $\boldsymbol{t} = (t_1, t_2)^\top$ in $\mathbb{R}^2$, we may use the solution of

$$M_n(t_j) = \exp(t_j\mu + \sigma^2 t_j^2/2), \quad j = 1, 2 \tag{5.8}$$

to estimate $\boldsymbol{\theta}$. It is straightforward to verify that if $t_2 \neq t_1$ and $t_j \neq 0$, then these equations result in the explicit estimators

$$\left.\begin{aligned} \hat{\mu} &= \frac{t_2^2 \log[M_n(t_1)] - t_1^2 \log[M_n(t_2)]}{t_1 t_2 (t_2 - t_1)} \\ \hat{\sigma}^2 &= 2\frac{t_1 \log[M_n(t_2)] - t_2 \log[M_n(t_1)]}{t_1 t_2 (t_2 - t_1)}. \end{aligned}\right\} \tag{5.9}$$

The limits of these estimators as $t_2 \to t_1$ or $t_j \to 0$ $(j = 1, 2)$ can be evaluated by l'Hôpital's rule so that, for example, as $t_2 \to t_1$ these estimators converge to

$$\left.\begin{aligned} \hat{\mu} &= \frac{2 \log[M_n(t_1)]}{t_1} - \frac{M_n^{(1)}(t_1)}{M_n(t_1)} \\ \hat{\sigma}^2 &= 2\{\frac{M_n^{(1)}(t_1)}{t_1 M_n(t_1)} - \frac{\log[M_n(t_1)]}{t_1^2}\}, \end{aligned}\right\} \tag{5.10}$$

where $M_n^{(1)}(t)$ denotes the derivative of $M_n(t)$.

Figure 5.2 depicts the moment generating function estimators of μ and σ^2 for a random sample of size $n = 50$ from a standard normal distribution. As observed in the figure, the estimators depend upon t_1 and t_2, and this problem will be considered later in this section.

The asymptotic distribution of the moment generating function estimator, $\hat{\boldsymbol{\theta}}$, can be obtained from Theorem 5.1. We find that $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is asymptotically normally distributed with mean vector zero and covariance matrix $\Sigma(\boldsymbol{\theta})$, where the elements of $\Sigma(\boldsymbol{\theta})$ are given by $(t_2 \neq t_1,\ t_j \neq 0)$

$$\sigma_{11}(\boldsymbol{\theta}) = \frac{t_2^2}{t_1^2(t_2 - t_1)^2}(e^{\sigma^2 t_1^2} - 1) - \frac{2}{(t_2 - t_1)^2}(e^{\sigma^2 t_1 t_2} - 1) + \frac{t_1^2}{t_2^2(t_2 - t_1)^2}(e^{\sigma^2 t_2^2} - 1)$$

$$\sigma_{12}(\boldsymbol{\theta}) = \frac{-2t_2}{t_1^2(t_2 - t_1)^2}(e^{\sigma^2 t_1^2} - 1) + \frac{2(t_1 + t_2)}{t_1 t_2 (t_2 - t1)^2}(e^{\sigma^2 t_1 t_2} - 1) - \frac{2t_1}{t_2^2(t_2 - t_1)^2}(e^{\sigma^2 t_2^2} - 1)$$

$$\sigma_{22}(\boldsymbol{\theta}) = \frac{4}{t_1^2(t_2 - t_1)^2}(e^{\sigma^2 t_1^2} - 1) - \frac{8}{t_1 t_2 (t_2 - t_1)^2}(e^{\sigma^2 t_1 t_2} - 1) + \frac{4}{t_2^2(t_2 - t_1)^2}(e^{\sigma^2 t_2^2} - 1).$$

In parallel to the estimators, the limits of these elements as $t_2 \to t_1$ or $t_j \to 0$ can be evaluated by l'Hôpital's rule. We find, for example, that as $t_2 \to t_1$ these

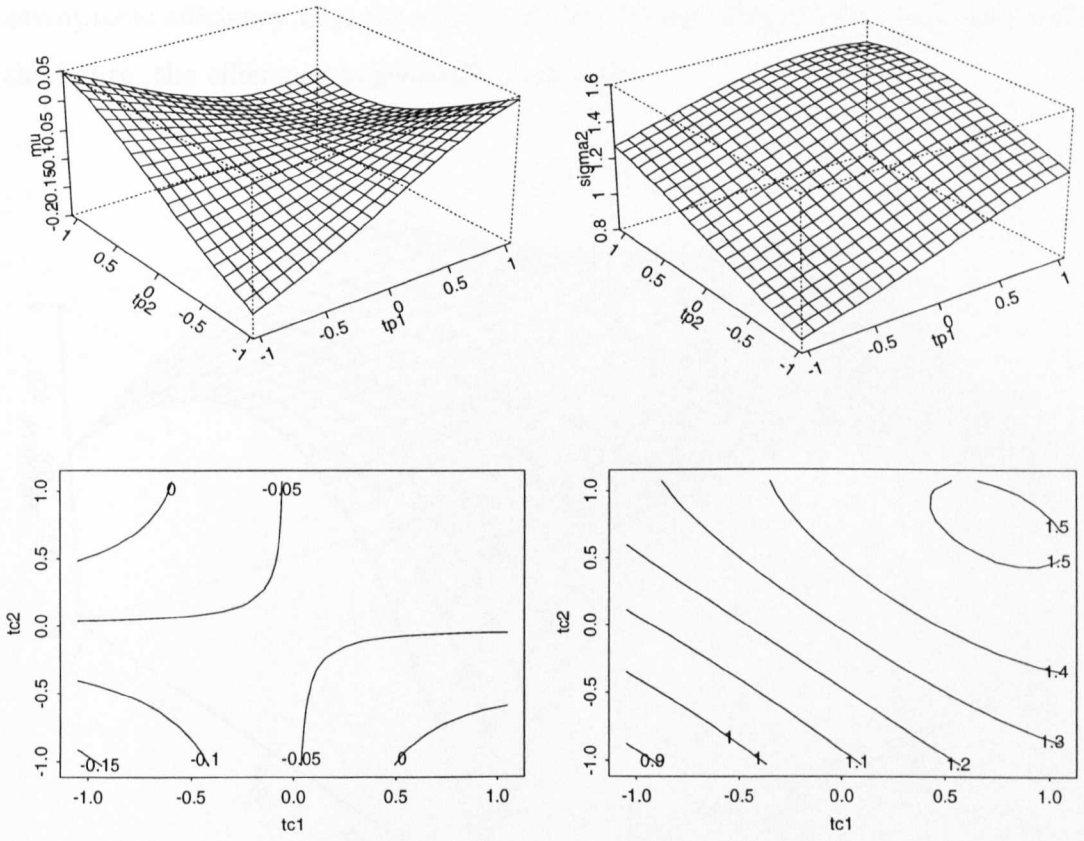

Figure 5.2: Perspective view (top) and contour level-plot (bottom) of the moment generating function estimator for μ (left) and σ^2 (right) based on a random sample of size $n = 50$ from a standard normal distribution.

elements become

$$\sigma_{11}(\boldsymbol{\theta}) = 4t_1^{-2}(e^{\sigma^2 t_1^2} - 1) - 3\sigma^2 e^{\sigma^2 t_1^2} + \sigma^4 t_1^2 e^{\sigma^2 t_1^2}$$
$$\sigma_{12}(\boldsymbol{\theta}) = -4t_1^{-3}(e^{\sigma^2 t_1^2} - 1) + 4\sigma^2 t_1^{-1} e^{\sigma^2 t_1^2} + 2\sigma^4 t_1 e^{\sigma^2 t_1^2}$$
$$\sigma_{22}(\boldsymbol{\theta}) = 4t_1^{-4}(e^{\sigma^2 t_1^2} - 1) - 4\sigma^2 t_1^{-2} e^{\sigma^2 t_1^2} + 4\sigma^4 e^{\sigma^2 t_1^2}.$$

The Fisher information matrix for a single observation is given by (1.19) so that, for example, the asymptotic efficiency of $\hat{\mu}$ is

$$\mathit{eff}(\hat{\mu}) = \sigma^2/\sigma_{11}(\boldsymbol{\theta}).$$

Note that this expression does not depend on μ. Figure 5.3 then depicts the

asymptotic efficiency of $\hat{\mu}$ for a normal distribution with $\sigma^2 = 1$. As observed in the figure, the efficiency is generally very high.

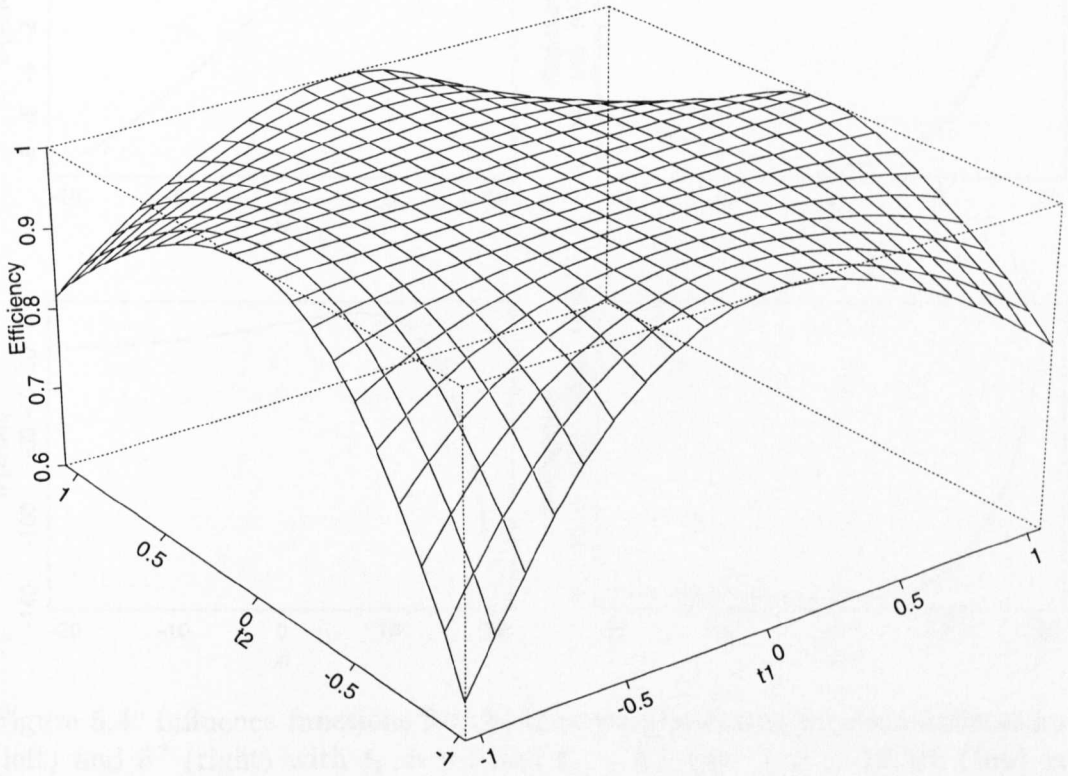

Figure 5.3: Perspective view of the asymptotic efficiency of $\hat{\mu}$ when $\sigma^2 = 1$.

The joint influence function for the moment generating function estimator is, from Theorem 5.2, given by ($t_2 \neq t_1$, $t_1 \neq 0$, $t_2 \neq 0$)

$$IF(\xi; \hat{\boldsymbol{\theta}}) = (t_2 - t_1)^{-1} \begin{pmatrix} \dfrac{t_2}{t_1 M(t_1; \boldsymbol{\theta})}[e^{t_1 \xi} - M(t_1; \boldsymbol{\theta})] - \dfrac{t_1}{t_2 M(t_2; \boldsymbol{\theta})}[e^{t_2 \xi} - M(t_2; \boldsymbol{\theta})] \\ \dfrac{2}{t_2 M(t_2; \boldsymbol{\theta})}[e^{t_2 \xi} - M(t_2; \boldsymbol{\theta})] - \dfrac{2}{t_1 M(t_1; \boldsymbol{\theta})}[e^{t_1 \xi} - M(t_1; \boldsymbol{\theta})] \end{pmatrix}.$$

It is instructive to evaluate this influence function at the standard normal distribution, as the standard normal has been in this work a point of reference. Figure 5.4 provides the individual influence functions for $\hat{\boldsymbol{\theta}}$ with $t_1 = 0.1$, $t_2 = 0.2$, and an extended range of ξ values. The centre portion is most informative in comparison with Figure 1.3. In particular, it illustrates that the moment generating function

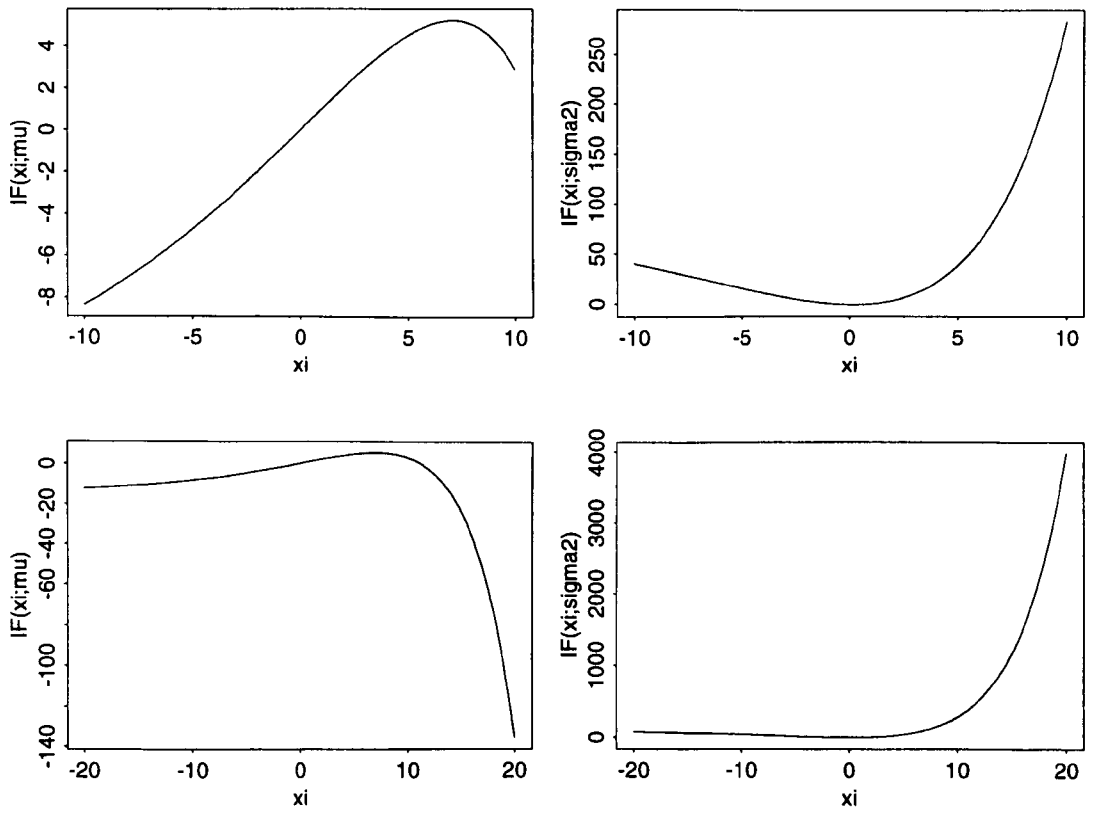

Figure 5.4: Influence functions for the moment generating function estimators $\hat{\mu}$ (left) and $\hat{\sigma}^2$ (right) with $t_1 = 0.1$ and $t_2 = 0.2$ over $\xi \in [-10, 10]$ (top) and $\xi \in [-20, 20]$ (bottom). The influence functions are evaluated at the standard normal distribution.

estimator is less robust than the integrated squared error estimator. The outer portions illustrate that the influence functions are not bounded in ξ.

We finally consider the selection of suitable values for t_1 and t_2. One possibility is to use the second approach of Section 5.3.3. In this approach, the likelihood function

$$L(\boldsymbol{\theta}) = (2\pi\sigma^2)^{-n/2} \exp[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2]$$

is considered as a function of $\hat{\boldsymbol{\theta}}$, given by (5.9), and maximised with respect to $\boldsymbol{t} = (t_1, t_2)^\top$. When $L(\hat{\boldsymbol{\theta}})$ was maximised with respect to $\boldsymbol{t}$, we found that the optimum value often resulted from $\boldsymbol{t} \approx (0, 0)^\top$. We investigated this result further by means of a simulation experiment. In particular, one hundred samples of sizes 25, 50 and 100 from a standard normal distribution were generated and, for each

sample, the value $\hat{\boldsymbol{t}} = (\hat{t}_1, \hat{t}_2)^\top$ which maximised $L(\hat{\boldsymbol{\theta}})$ was obtained. The results of the experiment are summarised in Table 5.1 below. We see from this experiment that $\boldsymbol{t} \approx (0,0)^\top$ in far more than a single instance, and we shall comment on this later in Section 5.6.

Table 5.1: An illustration of the constrained maximum likelihood approach applied to the standard normal distribution

Mean	*Sample size*		
	$n = 25$	$n = 50$	$n = 100$
$\hat{t}_1$	3×10^{-4}	-6×10^{-5}	3×10^{-4}
$\lvert\hat{t}_1 - \hat{t}_2\rvert$	1×10^{-3}	6×10^{-4}	9×10^{-4}

Alternatively, we may use the third approach of Section 5.3.3. Here we select $\boldsymbol{t}$ by minimising some measure, typically the determinant, of the asymptotic covariance matrix $\Sigma(\boldsymbol{\theta})$. In the present context, the determinant of $\Sigma(\boldsymbol{\theta})$ is given by

$$\det[\Sigma(\boldsymbol{\theta})] = 4\frac{e^{\sigma^2(t_1^2+t_2^2)} - e^{\sigma^2 t_1^2} - e^{\sigma^2 t_2^2} - e^{2\sigma^2 t_1 t_2} + 2e^{\sigma^2 t_1 t_2}}{t_1^2\, t_2^2\, (t_1 - t_2)^2}$$

and is minimised with respect to $\boldsymbol{t}$ at $\boldsymbol{t} = (0,0)^\top$. The result that $t_1 = t_2$ at the optimum supports the theory that "diagonal optimisation", as defined in point (4) of Section 5.3.3, occurs as a minimum-variance criterion. This diagonal optimisation result has been observed more generally and for other models (see, for example, *Schmidt, 1982*), and implies that we can reduce a p-dimensional search of an optimum $\boldsymbol{t}$ to a one-dimensional problem. However, it also renders the system of normal equations (5.3) singular, and we consider this further in the next section.

5.5 Limiting forms of the moment generating function estimator

The moment generating function estimator essentially equates empirical and theoretical moment generating functions for a set t_j $(j = 1, 2, \ldots, p)$ so as to give a system of equations to solve for the parameters. Clearly, the set must satisfy the requirement that the system is non-singular. We shall now demonstrate, for the case $p = 2$, how the moment generating function estimator may be derived when this requirement is not satisfied. This can occur in the three ways:

1. $t_2 \to t_1$;
2. $t_1 \to 0$;
3. $t_1 \to 0$, $t_2 \to 0$.

These three cases will now be considered in turn below.

Let $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ be the parameter vector, and consider initially the situation in which $t_2 \to t_1$. If the moment generating function estimator with $t_1 \neq t_2$ has an explicit form, then we may use l'Hôpital's rule to evaluate the limit of the estimator as $t_2 \to t_1$. This approach was illustrated in Section 5.4. Alternatively, or if an explicit solution does not exist, we may evaluate this limit in the following way. Let $t_2 = t_1 - \varepsilon$ so that the estimation equations are, from (5.4),

$$M_n(t_1) = M(t_1; \boldsymbol{\theta}) \tag{5.11}$$

$$M_n(t_1 - \varepsilon) = M(t_1 - \varepsilon; \boldsymbol{\theta}). \tag{5.12}$$

For any value of $t \in T$, we can write down the Taylor series expansions

$$M_n(t_1 - \varepsilon) = M_n(t_1) - \varepsilon M_n^{(1)}(t_1) + \tfrac{1}{2}\varepsilon^2 M_n^{(2)}(t_1 - \varepsilon_1), \quad 0 < \varepsilon_1 < \varepsilon \tag{5.13}$$

$$M(t_1 - \varepsilon; \boldsymbol{\theta}) = M(t_1; \boldsymbol{\theta}) - \varepsilon M^{(1)}(t_1; \boldsymbol{\theta}) + \tfrac{1}{2}\varepsilon^2 M^{(2)}(t_1 - \varepsilon_2; \boldsymbol{\theta}), \quad 0 < \varepsilon_2 < \varepsilon, \tag{5.14}$$

where the superscript denotes the order of differentiation with respect to t. Subtracting (5.14) from (5.13) gives

$$M_n(t_1 - \varepsilon) - M(t_1 - \varepsilon; \boldsymbol{\theta}) = M_n(t_1) - M(t_1; \boldsymbol{\theta}) - \varepsilon[M_n^{(1)}(t_1) - M^{(1)}(t_1; \boldsymbol{\theta})] + \tfrac{1}{2}\varepsilon^2[M_n^{(2)}(t_1 - \varepsilon_1) - M^{(2)}(t_1 - \varepsilon_2; \boldsymbol{\theta})]. \tag{5.15}$$

Substituting (5.11) and (5.12) in (5.15), and dividing through by ε gives

$$M_n^{(1)}(t_1) - M^{(1)}(t_1; \boldsymbol{\theta}) = \tfrac{1}{2}\varepsilon[M_n^{(2)}(t_1 - \varepsilon_1) - M^{(2)}(t_1 - \varepsilon_2; \boldsymbol{\theta})],$$

which, in the limit, as $\varepsilon \to 0$, becomes

$$M_n^{(1)}(t_1) = M^{(1)}(t_1; \boldsymbol{\theta}),$$

provided the second derivatives are bounded; this will certainly be the case if the distribution has finite variance.

Next we consider the situation in which $t_1 \to 0$. Expanding $M_n(0 - \varepsilon) = M(0 - \varepsilon; \boldsymbol{\theta})$ and using $M_n(0) = M(0; \boldsymbol{\theta})$, we find that, as $t_1 \to 0$, (5.11) tends to the first moment equation

$$M_n^{(1)}(0) = M^{(1)}(0; \boldsymbol{\theta}).$$

Finally we examine the case where in addition $t_2 \to 0$. Expanding $M_n(0-\delta) = M(0 - \delta; \boldsymbol{\theta})$ as far as the third derivatives gives

$$M_n(0 - \delta) - M(0 - \delta; \boldsymbol{\theta}) = M_n(0) - M(0; \boldsymbol{\theta}) - \delta[M_n^{(1)}(0) - M^{(1)}(0; \boldsymbol{\theta})] + \tfrac{1}{2}\delta^2[M_n^{(2)}(0) - M^{(2)}(0; \boldsymbol{\theta})] - \tfrac{1}{6}\delta^3[M_n^{(3)}(0) - M^{(3)}(0; \boldsymbol{\theta})],$$

which implies the second moment equation

$$M_n^{(2)}(0) = M^{(2)}(0; \boldsymbol{\theta}),$$

provided the third moment exists.

To summarise, we have shown that as $t_2 \to t_1$ the estimation equations become

$$M_n^{(k)}(t) = M^{(k)}(t; \boldsymbol{\theta}), \quad k = 0, 1,$$

as $t_1 \to 0$ they become

$$M_n^{(1)}(0) = M^{(1)}(0; \boldsymbol{\theta})$$
$$M_n(t_2) = M(t_2; \boldsymbol{\theta}),$$

and as both t_1 and t_2 tend to zero, they simplify to

$$M_n^{(k)}(0) = M^{(k)}(0; \boldsymbol{\theta}), \quad k = 1, 2.$$

Although we have outlined above only the two-parameter case, these results generalise directly to the p-parameter case.

5.6 Insights into the moment generating function method

As seen in the previous section, the moment generating function method reduces to the method of moments when $\boldsymbol{t} = (0, 0, \ldots, 0)^\top$. For arbitrary $\boldsymbol{t}$, *Kiefer (1978)* compared these two methods using

$$M(t; \boldsymbol{\theta}) = \sum_{j=1}^{\infty} \frac{\mu_j t^j}{j!},$$

where μ_k is the kth moment about the origin. In particular, he observed that the moment generating function method uses information in all the moments to estimate $\boldsymbol{\theta}$, while the method of moments uses information in only the first p moments. Furthermore, the moment generating function method weights the

moments so that low-order moments can have greater weights than high-order moments. The moment method weights all moments equally. Since we normally expect a given sample to determine low-order moments more accurately than high-order moments, the moment generating function method might be presumed to be more efficient than the method of moments.

On the other hand, there exist simple cases in which the methods of moments and maximum likelihood coincide. In one-parameter distributions *Tallis and Light (1968)* have established a sufficient condition for the maximum likelihood estimator to result as a moment estimator. In particular, they presented:

Lemma 5.1 (Tallis and Light, 1968). *If an unbiased estimator, $T(x)$, of some strictly monotone, differentiable function of θ, $\tau(\theta)$, satisfying the Cramér-Rao lower bound exists for θ in some interval, then there exists a moment estimator of θ, $\tilde{\theta}$, such that $\tilde{\theta} = \hat{\theta}$, where $\hat{\theta}$ is the maximum likelihood estimator of θ.*

In fact, *Tallis and Light (1968)* pointed out that a necessary and sufficient condition for the Cramér-Rao lower bound to be attained by an unbiased estimator $T(x)$ of some function $\tau(\theta)$ is that $f(x;\theta)$ be of exponential form.

In multi-parameter distributions, the method of moments is generally inferior in almost all respects to the method of maximum likelihood. Such results tend to indicate that the choice $\boldsymbol{t} = (0, 0, \ldots, 0)^\top$ in the moment generating function method should generally be avoided. (An exception occurs with the two-parameter normal distribution; here moment and maximum likelihood estimators coincide and this provides the theoretical underpinning for the results of Section 5.4.) This is consistent with the view of *Ball and Milne (1996)*.

5.7 The preferred moment generating function method

The moment generating function method is the simplest parameter estimation method based on transforms. This simplicity gives it an advantage over, for

example, the integrated squared error method, but it presents new problems of its own. Of these the most important is the choice for $\boldsymbol{t}$. In the remainder of this chapter we shall describe some extensions of the moment generating function method, which aim to increase its flexibility and asymptotic efficiency. On the other hand these extensions will forfeit the simplicity of their predecessor and the possibility of deriving explicit parameter estimators.

One extension, which was also suggested by *Quandt and Ramsey (1978)*, is the preferred moment generating function method. This method is similar to the moment generating function method in that it uses essentially the same criterion to measure the deviation between the empirical and theoretical moment generating functions. The only difference is that the preferred moment generating function method uses more values for t than parameters. The motivation for this comes from the fact that by increasing the number of values for t one can, ultimately, eliminate the role played by the particular choice of these values. The following definition follows from the work of *Quandt and Ramsey (1978)*:

Definition 5.2. The preferred moment generating function (PMGF) estimator, $\hat{\boldsymbol{\theta}}$, for $\boldsymbol{\theta}_0 \in \Theta \subseteq \mathbb{R}^p$, is any value of $\boldsymbol{\theta}$ which minimises

$$S(\boldsymbol{\theta}; \boldsymbol{t}) = \sum_{j=1}^{q} [M_n(t_j) - M(t_j; \boldsymbol{\theta})]^2 \tag{5.16}$$

for a suitable value of $\boldsymbol{t} = (t_1, t_2, \ldots, t_q)^\top$ in T^q, where q is an integer greater than p.

In theory, the preferred moment generating function method has certain advantages over the moment generating function method. In practice, the complication of selecting a value for $\boldsymbol{t}$ remains. This has now become a two-fold problem consisting of:

1. the problem of selecting q;

2. the problem of selecting, for a given q, values for the t_j $(j = 1, 2, \ldots, q)$.

Trade-off between computational complexity and sensitivity to the particular choice of the t_j $(j = 1, 2, \ldots, q)$ is accomplished in the selection of q. For given q, the practical approaches of Section 5.3.3 could be used to select a value for $\boldsymbol{t}$. The following section provides the inferential foundation for the methodology proposed.

5.7.1 Properties of the preferred moment generating function estimator

The asymptotic properties of the preferred moment generating function estimator were investigated by *Quandt and Ramsey (1978)*. In the context of a mixture of two normal distributions (so that conditions 1, 2 and 4 are satisfied), they established:

Theorem 5.3 (Quandt and Ramsey, 1978). *Given a set of non-zero and distinct points t_j $(j = 1, 2, \ldots, q)$ in T, there exists a preferred moment generating function estimator, $\hat{\boldsymbol{\theta}}$, for $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$ which is strongly consistent and for which*

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \Sigma(\boldsymbol{\theta}_0)],$$

where

$$\Sigma(\boldsymbol{\theta}) = [K(\boldsymbol{\theta})^\top K(\boldsymbol{\theta})]^{-1} K(\boldsymbol{\theta})^\top \Omega(\boldsymbol{\theta}) K(\boldsymbol{\theta}) [K(\boldsymbol{\theta})^\top K(\boldsymbol{\theta})]^{-1}. \tag{5.17}$$

Here $K(\boldsymbol{\theta})$ is the $q \times p$ matrix whose (i, j)th element is

$$\kappa_{ij}(\boldsymbol{\theta}) = \frac{\partial M(t_i; \boldsymbol{\theta})}{\partial \theta_j}$$

and $\Omega(\boldsymbol{\theta})$ is the $q \times q$ symmetric matrix whose (i, j)th element is

$$\omega_{ij}(\boldsymbol{\theta}) = M(t_i + t_j; \boldsymbol{\theta}) - M(t_i; \boldsymbol{\theta}) M(t_j; \boldsymbol{\theta}).$$

It is clear from (5.17) that, in general, the form of $\Sigma(\boldsymbol{\theta})$ will be very complicated. However, it is interesting to see what happens when $q = p$, that is, when the number of values for t is equal to the number of parameters being estimated. In this case the preferred moment generating function estimator coincides with the moment generating function estimator. Correspondingly, when $q = p$, (5.17) simplifies to (5.5).

The robustness properties of the preferred moment generating function estimator do not appear to have been investigated. We thus present and prove the following theorem, which is subject to conditions 1, 2 and 4.

Theorem 5.4. *Given a set of non-zero and distinct points t_j $(j = 1, 2, \ldots, q)$ in T, the (consistent) preferred moment generating function estimator, $\hat{\boldsymbol{\theta}}$, for $\boldsymbol{\theta}_0 \in \Theta \subseteq \mathbb{R}^p$ has joint influence function*

$$IF(\xi; \hat{\boldsymbol{\theta}}) = [K(\boldsymbol{\theta}_0)^\top K(\boldsymbol{\theta}_0)]^{-1} K(\boldsymbol{\theta}_0)^\top \boldsymbol{\tau}(\xi; \boldsymbol{\theta}_0), \tag{5.18}$$

where $K(\boldsymbol{\theta})$ is the $q \times p$ matrix whose (i, j)th element is

$$\kappa_{ij}(\boldsymbol{\theta}) = \frac{\partial M(t_i; \boldsymbol{\theta})}{\partial \theta_j}$$

and $\boldsymbol{\tau}(\xi; \boldsymbol{\theta})$ is the $q \times 1$ vector whose ith element is

$$\tau_i(\xi; \boldsymbol{\theta}) = \exp(t_i \xi) - M(t_i; \boldsymbol{\theta}).$$

Proof. The preferred moment generating function estimator is found in practice as the solution of

$$\frac{\partial S(\boldsymbol{\theta}; \boldsymbol{t})}{\partial \theta_k} = 0, \quad k = 1, 2, \ldots, p,$$

which, from expression (5.16), may be written as

$$\sum_{i=1}^{q} \frac{\partial M(t_i; \boldsymbol{\theta})}{\partial \theta_k} [M_n(t_i) - M(t_i; \boldsymbol{\theta})] = 0, \quad k = 1, 2, \ldots, p.$$

For the kth ($k = 1, 2, \ldots, p$) equation of this system, we may write down the influence equation

$$\sum_{i=1}^{q} \frac{\partial M(t_i;\boldsymbol{\theta})}{\partial \theta_k} \{IF[\xi; M_n(t_i)] - \sum_{j=1}^{p} \frac{\partial M(t_i;\boldsymbol{\theta})}{\partial \theta_j} IF(\xi;\hat{\theta}_j)\} = 0.$$

Re-arranging we obtain

$$\sum_{i=1}^{q} \frac{\partial M(t_i;\boldsymbol{\theta})}{\partial \theta_k} IF[\xi; M_n(t_i)] = \sum_{i=1}^{q}\sum_{j=1}^{p} \frac{\partial M(t_i;\boldsymbol{\theta})}{\partial \theta_k} \frac{\partial M(t_i;\boldsymbol{\theta})}{\partial \theta_j} IF(\xi;\hat{\theta}_j),$$

or, equivalently,

$$\sum_{i=1}^{q} \kappa_{ik}(\boldsymbol{\theta})\tau_i(\xi;\boldsymbol{\theta}) = \sum_{i=1}^{q}\sum_{j=1}^{p} \kappa_{ik}(\boldsymbol{\theta})\kappa_{ij}(\boldsymbol{\theta}) IF(\xi;\hat{\theta}_j).$$

Bringing these p influence equations together, we may form the single matrix equation

$$K(\boldsymbol{\theta})^\top \boldsymbol{\tau}(\xi;\boldsymbol{\theta}) = [K(\boldsymbol{\theta})^\top K(\boldsymbol{\theta})] IF(\xi;\hat{\boldsymbol{\theta}}),$$

thus proving the theorem. □

We may verify this result by exploiting the relationship between influence functions and asymptotic variances, as stated in (1.16). In addition, it is straightforward to show that if $q = p$ then (5.18) simplifies to (5.6), as it should. On the other hand, if $q > p$ then the ith ($i = 1, 2, \ldots, p$) individual influence function for the preferred moment generating function estimator may be expressed as

$$IF(\xi;\hat{\theta}_i) = \sum_{j=1}^{q} \kappa^{\star}_{ij}(\boldsymbol{\theta}_0)[\exp(t_j\xi) - M(t_j;\boldsymbol{\theta}_0)],$$

where $\kappa^{\star}_{ij}(\boldsymbol{\theta})$ is the (i,j)th element of $[K(\boldsymbol{\theta})^\top K(\boldsymbol{\theta})]^{-1} K(\boldsymbol{\theta})^\top$. This implies that the preferred moment generating function estimator is, like the moment generating function estimator, not robust over the entire real line.

5.8 The modified moment generating function method

The preferred moment generating function method essentially sets up a large number of equations which, if they were consistent, could be reduced to remove dependencies and, in principle, inverted. Unfortunately sampling fluctuations mean that they will not be consistent so that, rather than trying to invert, we minimise an error criterion as in (5.16). *Schmidt (1982)* pursued the observation that the empirical moment generating function values for different t are not independent, and have unequal variances. Thus what should be minimised is not a simple sum of squares, but rather a generalised sum of squares. This approach is adopted by the modified moment generating function method.

Explicitly, if we define

$$\boldsymbol{M}(\boldsymbol{t};\boldsymbol{\theta}) = [M(t_1;\boldsymbol{\theta}), M(t_2;\boldsymbol{\theta}), \ldots, M(t_q;\boldsymbol{\theta})]^\top$$
$$\boldsymbol{M}_n(\boldsymbol{t}) = [M_n(t_1), M_n(t_2), \ldots, M_n(t_q)]^\top$$

then the modified moment generating function estimator is defined as follows:

Definition 5.3. The modified moment generating function (MMGF) estimator, $\hat{\boldsymbol{\theta}}$, for $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$, is any value of $\boldsymbol{\theta}$ which minimises

$$S(\boldsymbol{\theta};\boldsymbol{t}) = [\boldsymbol{M}_n(\boldsymbol{t}) - \boldsymbol{M}(\boldsymbol{t};\boldsymbol{\theta})]^\top \Omega^{-1}(\boldsymbol{\theta}_0)[\boldsymbol{M}_n(\boldsymbol{t}) - \boldsymbol{M}(\boldsymbol{t};\boldsymbol{\theta})] \tag{5.19}$$

for a suitable value of $\boldsymbol{t} = (t_1, t_2, \ldots, t_q)^\top$ in T^q, where q is an integer greater than p and $\Omega(\boldsymbol{\theta})$ is the $q \times q$ symmetric matrix whose (i,j)th element is

$$\omega_{ij}(\boldsymbol{\theta}) = M(t_i + t_j;\boldsymbol{\theta}) - M(t_i;\boldsymbol{\theta})M(t_j;\boldsymbol{\theta}).$$

The modified moment generating function method has been employed for parameter estimation in a variety of distribution families, including the mixture of

normal distributions (see *Schmidt, 1982*) and the three parameter gamma distribution (see *Koutrouvelis and Canavos, 1997*). In general, this method may be criticised in three ways:

1. it is numerically complicated, particularly in view of the matrix inversion involved.

2. it depends on the unknown parameter value $\boldsymbol{\theta}_0$ through the matrix $\Omega(\boldsymbol{\theta}_0)$;

3. it entails the selection of $\boldsymbol{t}$, which involves precisely the same considerations as were necessary for the selection of $\boldsymbol{t}$ in the preferred moment generating function method.

In practice, the second difficulty could be overcome by regarding $\Omega(\boldsymbol{\theta}_0)$ as a function of the minimising variable $\boldsymbol{\theta}$ although such a procedure would require additional computational expense. In spite of these problems, the modified moment generating function method has distinct advantages over the previous methods. Insights into these can be obtained from the properties of the resulting estimator.

5.8.1 Properties of the modified moment generating function estimator

The asymptotic theory needed for the modified moment generating function method is given by *Schmidt (1982)*. In the context of a mixture of two normal distributions (so that conditions 1, 2, 5 and 6 are satisfied), he established:

Theorem 5.5 (Schmidt, 1982). *Given a set of non-zero and distinct points* t_j $(j = 1, 2, \ldots, q)$ *in* T, *there exists a modified moment generating function estimator,* $\hat{\boldsymbol{\theta}}$, *for* $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$ *which is strongly consistent and for which*

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \Sigma(\boldsymbol{\theta}_0)],$$

where

$$\Sigma(\boldsymbol{\theta}) = [K(\boldsymbol{\theta})^\top \Omega^{-1}(\boldsymbol{\theta}) K(\boldsymbol{\theta})]^{-1}. \tag{5.20}$$

In this expression, $K(\boldsymbol{\theta})$ *is the* $q \times p$ *matrix whose* (i,j)*th element is*

$$\kappa_{ij}(\boldsymbol{\theta}) = \frac{\partial M(t_i;\boldsymbol{\theta})}{\partial \theta_j}$$

and $\Omega(\boldsymbol{\theta})$ *is the* $q \times q$ *symmetric matrix whose* (i,j)*th element is*

$$\omega_{ij}(\boldsymbol{\theta}) = M(t_i + t_j;\boldsymbol{\theta}) - M(t_i;\boldsymbol{\theta})M(t_j;\boldsymbol{\theta}).$$

With this result, *Schmidt (1982)* showed that the modified moment generating function estimator outperforms the preferred moment generating function estimator in terms of asymptotic efficiency. He also showed that the asymptotic efficiency of the modified moment generating function estimator will generally be increased by adding one more point to a given set t_j $(j = 1, 2, \ldots, q)$. Accordingly, *Schmidt (1982)* conjectured that as $q \to \infty$ the modified moment generating function estimator is asymptotically efficient relative to the maximum likelihood estimator, and this was later established by *Feuerverger and McDunnough (1984).*

As with the preferred moment generating function estimator, the robustness properties of the modified moment generating function estimator do not appear to have been investigated. We thus present and prove the following theorem, which is subject to conditions 1, 2, 5 and 6.

Theorem 5.6. *Given a set of non-zero and distinct points* t_j $(j = 1, 2, \ldots, q)$ *in* T*, the (consistent) modified moment generating function estimator,* $\hat{\boldsymbol{\theta}}$*, for* $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$ *has joint influence function*

$$IF(\xi;\hat{\boldsymbol{\theta}}) = [K(\boldsymbol{\theta}_0)^\top \Omega^{-1}(\boldsymbol{\theta}_0) K(\boldsymbol{\theta}_0)]^{-1} K(\boldsymbol{\theta}_0)^\top \Omega^{-1}(\boldsymbol{\theta}_0) \boldsymbol{\tau}(\xi;\boldsymbol{\theta}_0), \tag{5.21}$$

where $K(\boldsymbol{\theta})$ *is the* $q \times p$ *matrix whose* (i,j)*th element is*

$$\kappa_{ij}(\boldsymbol{\theta}) = \frac{\partial M(t_i;\boldsymbol{\theta})}{\partial \theta_j},$$

$\Omega(\boldsymbol{\theta})$ *is the* $q \times q$ *symmetric matrix whose* (i,j)*th element is*

$$\omega_{ij}(\boldsymbol{\theta}) = M(t_i + t_j;\boldsymbol{\theta}) - M(t_i;\boldsymbol{\theta})M(t_j;\boldsymbol{\theta}),$$

and $\boldsymbol{\tau}(\xi;\boldsymbol{\theta})$ *is the* $q \times 1$ *vector whose ith element is*

$$\tau_i(\xi;\boldsymbol{\theta}) = \exp(t_i\xi) - M(t_i;\boldsymbol{\theta}).$$

Proof. The modified moment generating function estimator is found in practice as the solution of

$$\frac{\partial S(\boldsymbol{\theta};\boldsymbol{t})}{\partial \theta_k} = 0, \quad k = 1,2,\ldots,p,$$

or, equivalently,

$$\frac{\partial}{\partial \theta_k}\sum_{i=1}^{q}\sum_{j=1}^{q}[M_n(t_i) - M(t_i;\boldsymbol{\theta})]\omega^{ij}(\boldsymbol{\theta})[M_n(t_j) - M(t_j;\boldsymbol{\theta})] = 0, \quad k = 1,2,\ldots,p,$$

where $w^{ij}(\boldsymbol{\theta})$ is the (i,j)th element of $\Omega^{-1}(\boldsymbol{\theta})$. For the kth ($k = 1,2,\ldots,p$) equation of this system

$$\sum_{i=1}^{q}\sum_{j=1}^{q}\{\frac{\partial M(t_i;\boldsymbol{\theta})}{\partial \theta_k}\omega^{ij}(\boldsymbol{\theta})[M_n(t_j) - M(t_j;\boldsymbol{\theta})] + [M_n(t_i) - M(t_i;\boldsymbol{\theta})]\omega^{ij}(\boldsymbol{\theta})\frac{\partial M(t_j;\boldsymbol{\theta})}{\partial \theta_k}$$
$$- [M_n(t_i) - M(t_i;\boldsymbol{\theta})]\frac{\partial \omega^{ij}(\boldsymbol{\theta})}{\partial \theta_k}[M_n(t_j) - M(t_j;\boldsymbol{\theta})]\} = 0,$$

we may write down the influence function

$$\sum_{i=1}^{q}\sum_{j=1}^{q}\{\frac{\partial M(t_i;\boldsymbol{\theta})}{\partial \theta_k}\omega^{ij}(\boldsymbol{\theta})\{IF[\xi;M_n(t_j)] - \sum_{\ell=0}^{p}\frac{\partial M(t_j;\boldsymbol{\theta})}{\partial \theta_\ell}IF(\xi;\hat{\theta}_\ell)\}$$
$$+ \{IF[\xi;M_n(t_i)] - \sum_{\ell=0}^{p}\frac{\partial M(t_i;\boldsymbol{\theta})}{\partial \theta_\ell}IF(\xi;\hat{\theta}_\ell)\}\omega^{ij}(\boldsymbol{\theta})\frac{\partial M(t_j;\boldsymbol{\theta})}{\partial \theta_k}\} = 0.$$

Re-arranging and simplifying results in

$$\sum_{i=1}^{q}\sum_{j=1}^{q}\frac{\partial M(t_i;\boldsymbol{\theta})}{\partial\theta_k}\omega^{ij}(\boldsymbol{\theta})IF[\xi;M_n(t_j)] = \sum_{i=1}^{q}\sum_{j=1}^{q}\sum_{\ell=0}^{p}\frac{\partial M(t_i;\boldsymbol{\theta})}{\partial\theta_k}\omega^{ij}(\boldsymbol{\theta})\frac{\partial M(t_j;\boldsymbol{\theta})}{\partial\theta_\ell}IF(\xi;\hat{\theta}_\ell)$$

or, equivalently,

$$\sum_{i=1}^{q}\sum_{j=1}^{q}\kappa_{ik}(\boldsymbol{\theta})\omega^{ij}(\boldsymbol{\theta})\tau_j(\xi;\boldsymbol{\theta}) = \sum_{i=1}^{q}\sum_{j=1}^{q}\sum_{\ell=0}^{p}\kappa_{ik}(\boldsymbol{\theta})\omega^{ij}(\boldsymbol{\theta})\kappa_{j\ell}(\boldsymbol{\theta})IF(\xi;\hat{\theta}_\ell).$$

Bringing these p equations together, we may form the single matrix equation

$$K(\boldsymbol{\theta})^\top\Omega^{-1}(\boldsymbol{\theta})\boldsymbol{\tau}(\xi;\boldsymbol{\theta}) = [K(\boldsymbol{\theta})^\top\Omega^{-1}(\boldsymbol{\theta})K(\boldsymbol{\theta})]IF(\xi;\hat{\boldsymbol{\theta}}).$$

The theorem now follows. □

Theorem 5.6 demonstrates that, in general, a symbolic expression for (5.21) will be too complicated to be practicable. Furthermore, it shows that the ith ($i = 1, 2, \ldots, p$) individual influence function for the modified moment generating function estimator will be of the form

$$IF(\xi;\hat{\theta}_i) = \sum_{j=1}^{q}\kappa^\star_{ij}(\boldsymbol{\theta}_0)[\exp(t_j\xi) - M(t_j;\boldsymbol{\theta})], \tag{5.22}$$

where $\kappa^\star_{ij}(\boldsymbol{\theta})$ is the (i,j)th element of $[K(\boldsymbol{\theta})^\top\Omega^{-1}(\boldsymbol{\theta})K(\boldsymbol{\theta})]^{-1}K(\boldsymbol{\theta})^\top\Omega^{-1}(\boldsymbol{\theta})$. The modified moment generating function estimator cannot, therefore, be robust over the entire real line. This is a consequence of basing estimation on moment generating functions.

5.9 The q-L method

As indicated in Chapter 1 (Section 1.14), for continuous weight functions, minimum distance estimation based on characteristic functions can be expected to outperform such estimation based on moment generating functions. This was due

to a number of reasons resulting from the uniform boundedness of the characteristic function. With this in mind, it would be of interest to base minimum distance estimation involving step weight functions on characteristic functions. This approach is adopted in the remainder of this chapter.

The switch from moment generating functions to characteristic functions seems straightforward. For example, in the spirit outlined for the moment generating function method, it seems reasonable to estimate the parameters by minimising

$$\sum_{j=1}^{p} |\phi_n(t_j) - \phi(t_j; \boldsymbol{\theta})|^2$$

with respect to $\boldsymbol{\theta}$, where t_j $(j = 1, 2, \ldots, p)$ are appropriate points in $\mathbb{R}$. However this approach has not been favoured in the literature. *Feuerverger and McDunnough (1981a, b)* discussed some alternatives, including the q-L method.

The motivation for this method comes from the asymptotic joint distribution of $[\phi_n(t_1), \phi_n(t_2), \ldots, \phi_n(t_q)]$. Explicitly, if we define

$$\boldsymbol{z}(\boldsymbol{t}; \boldsymbol{\theta}) = [U(t_1; \boldsymbol{\theta}), U(t_2; \boldsymbol{\theta}), \ldots, U(t_q; \boldsymbol{\theta}), V(t_1; \boldsymbol{\theta}), V(t_2; \boldsymbol{\theta}), \ldots, V(t_q; \boldsymbol{\theta})]^\top$$
$$\boldsymbol{z}_n(\boldsymbol{t}) = [U_n(t_1), U_n(t_2), \ldots, U_n(t_q), V_n(t_1), V_n(t_2), \ldots, V_n(t_q)]^\top,$$

where $U(t; \boldsymbol{\theta})$ and $V(t; \boldsymbol{\theta})$ are the real and imaginary parts of $\phi(t; \boldsymbol{\theta})$ respectively, and $U_n(t)$, $V_n(t)$ are the corresponding parts of $\phi_n(t)$, then the q-L estimator is defined as follows:

Definition 5.4. The q-L estimator, $\hat{\boldsymbol{\theta}}$, for $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$, is any value of $\boldsymbol{\theta}$ which minimises

$$L_q(\boldsymbol{\theta}; \boldsymbol{t}) = [\boldsymbol{z}_n(\boldsymbol{t}) - \boldsymbol{z}(\boldsymbol{t}; \boldsymbol{\theta})]^\top \Omega^{-1}(\boldsymbol{\theta}_0) [\boldsymbol{z}_n(\boldsymbol{t}) - \boldsymbol{z}(\boldsymbol{t}; \boldsymbol{\theta})] \tag{5.23}$$

for a suitable value of $\boldsymbol{t} = (t_1, t_2, \ldots, t_q)^\top$ in $\mathbb{R}^q$, where q is an integer not smaller than $\frac{1}{2}p$ and $\Omega(\boldsymbol{\theta})$ is the $2q \times 2q$ symmetric matrix whose (i, j)th element is (specific

reference to $\boldsymbol{\theta}$ is omitted for convenience of printing)

$$\omega_{ij} = \begin{cases} \frac{1}{2}[U(t_i + t_j) + U(t_i - t_j)] - U(t_i)U(t_j) & 1 \leq i, j \leq q, \\ \frac{1}{2}[V(t_i + t_{j-q}) - V(t_i - t_{j-q})] - U(t_i)V(t_{j-q}) & \begin{cases} 1 \leq i \leq q \\ q+1 \leq j \leq 2q, \end{cases} \\ \frac{1}{2}[U(t_{i-q} - t_{j-q}) - U(t_{i-q} + t_{j-q})] - V(t_{i-q})V(t_{j-q}) & q+1 \leq i, j \leq 2q. \end{cases}$$

The term "q-L method" derives from the fact that the procedure is restricted to a set of q points in $\mathbb{R}$, and is based on the likelihood of the asymptotic joint distribution of $[\phi_n(t_1), \phi_n(t_2), \ldots, \phi_n(t_q)]$. The smallest number of points required by the method is $\lceil \frac{1}{2}p \rceil + 1$, where the symbol $\lceil \frac{1}{2}p \rceil$ denotes the greatest integer less than $\frac{1}{2}p$. This is because for each point of t we have two distinct transforms of the form (1.2) available, one with $g(t, x) = \cos(tx)$ and the other with $g(t, x) = \sin(tx)$. We find, therefore, that when p is even and $q = \frac{1}{2}p$, the q-L method essentially solves the system

$$\left.\begin{aligned} U_n(t_j) &= U(t_j; \boldsymbol{\theta}) \\ V_n(t_j) &= V(t_j; \boldsymbol{\theta}) \end{aligned}\right\} \quad j = 1, 2, \ldots, \tfrac{1}{2}p$$

for $\boldsymbol{\theta}$, if such a solution exists. Otherwise, the method proceeds by minimising a generalised sum of squares and would thus seem to be intimately connected with the modified moment generating function method. Indeed, the two methods may be criticised in similar ways. However, the q-L method has certain advantages over the modified moment generating function method, as will become apparent by exploring the properties of the resulting estimator. These are given in Section 5.9.2, but are subject to the regularity conditions below.

5.9.1 Regularity conditions

Feuerverger and McDunnough (1981a) presented four conditions, which together are sufficient requirements for Theorems 5.7 and 5.8. If, for $\boldsymbol{t} = (t_1, t_2, \ldots, t_q)^\top$ in

$\mathbb{R}^q$ $(q \geq \frac{1}{2}p)$, we define the $2q \times p$ matrix $K(\boldsymbol{\theta})$ with (i,j) element

$$\kappa_{ij}(\boldsymbol{\theta}) = \begin{cases} \dfrac{\partial U(t_i;\boldsymbol{\theta})}{\partial \theta_j} & 1 \leq i \leq q, \\ \dfrac{\partial V(t_{i-q};\boldsymbol{\theta})}{\partial \theta_j} & q+1 \leq i \leq 2q, \end{cases}$$

then these four conditions are:

1. $\boldsymbol{\Theta}$ is an open rectangle;
2. $\phi(t;\boldsymbol{\theta})$ is continuously differentiable (in $\boldsymbol{\theta}$) at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$;
3. the matrix $\Omega(\boldsymbol{\theta})$ is invertible at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$;
4. the matrix $K(\boldsymbol{\theta})^\top \Omega(\boldsymbol{\theta}) K(\boldsymbol{\theta})$ is invertible at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

5.9.2 Properties of the q-L estimator

The asymptotic properties of the q-L estimator were investigated by *Feuerverger and McDunnough (1981a)*. They established:

Theorem 5.7 (Feuerverger and McDunnough, 1981a). *Given a set of non-zero and distinct points t_j $(j = 1, 2, \ldots, q)$ in $\mathbb{R}$, there exists a q-L estimator, $\hat{\boldsymbol{\theta}}$, for $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$ which is strongly consistent and for which*

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \Sigma(\boldsymbol{\theta}_0)],$$

where

$$\Sigma(\boldsymbol{\theta}) = [K(\boldsymbol{\theta})^\top \Omega^{-1}(\boldsymbol{\theta}) K(\boldsymbol{\theta})]^{-1}. \tag{5.24}$$

In this expression $K(\boldsymbol{\theta})$ and $\Omega(\boldsymbol{\theta})$ are the $2q \times p$ and $2q \times 2q$ matrices, respectively, as stated in Section 5.9.1 and Definition 5.4.

In addition to consistency and asymptotic normality, *Feuerverger and McDunnough (1981a)* established the asymptotic efficiency of the q-L estimator. In

particular, they showed that by increasing the number of points t_j $(j = 1, 2, \ldots, q)$ in an appropriate way, the resulting estimator can attain arbitrarily high efficiency. This result was also established for the modified moment generating function estimator, and in this respect there is little to choose between them. However, it is in the robustness properties of the two estimators where the real difference occurs. *Feuerverger and McDunnough (1981a)* have derived the influence function for a certain continuous analogue of the q-L estimator. The influence function of the q-L estimator itself is presented below.

Theorem 5.8. *Given a set of non-zero and distinct points t_j $(j = 1, 2, \ldots, q)$ in $\mathbb{R}$, the (consistent) q-L estimator, $\hat{\boldsymbol{\theta}}$, for $\boldsymbol{\theta}_0 \in \Theta \subseteq \mathbb{R}^p$ has joint influence function*

$$IF(\xi; \hat{\boldsymbol{\theta}}) = [K(\boldsymbol{\theta}_0)^\top \Omega^{-1}(\boldsymbol{\theta}_0) K(\boldsymbol{\theta}_0)]^{-1} K(\boldsymbol{\theta}_0)^\top \Omega^{-1}(\boldsymbol{\theta}_0) \boldsymbol{\tau}(\xi; \boldsymbol{\theta}_0), \tag{5.25}$$

where $K(\boldsymbol{\theta})$ and $\Omega(\boldsymbol{\theta})$ are the $2q \times p$ and $2q \times 2q$ matrices, respectively, of Theorem 5.7, and $\boldsymbol{\tau}(\xi; \boldsymbol{\theta})$ is the $2q \times 1$ vector whose ith element is

$$\tau_i(\xi; \boldsymbol{\theta}) = \begin{cases} \cos(t_i \xi) - U(t_i; \boldsymbol{\theta}) & 1 \leq i \leq q, \\ \sin(t_{i-q} \xi) - V(t_{i-q}; \boldsymbol{\theta}) & q+1 \leq i \leq 2q. \end{cases}$$

Proof. The proof of Theorem 5.8 is essentially analogous to the proof of Theorem 5.6 and is thus omitted. □

In parallel to (5.21), a symbolic expression for (5.25) will tend to be too complicated to be practicable. However, it is straightforward to show that the ith $(i = 1, 2, \ldots, p)$ individual influence function for the q-L estimator will have the general form

$$IF(\xi; \hat{\theta}_i) = \sum_{j=1}^{q} \{a_{ij}(\boldsymbol{\theta}_0)[\cos(t_j \xi) - U(t_j; \boldsymbol{\theta}_0)] + b_{ij}(\boldsymbol{\theta}_0)[\sin(t_j \xi) - V(t_j; \boldsymbol{\theta}_0)]\},$$

where the elements $a_{ij}(\boldsymbol{\theta})$ and $b_{ij}(\boldsymbol{\theta})$ are involved but do not depend on ξ. This demonstrates that the q-L influence functions have periodic components in ξ. The

nature of the harmonics depends partially on the closeness of the t_j $(j = 1, 2, \ldots, q)$ to the origin. Nevertheless, it is clear from this that the q-L estimator is robust over the entire line for arbitrary t_j $(j = 1, 2, \ldots, q)$. This is a consequence of basing estimation on characteristic functions.

In summary, the advantages of the q-L method over, for example, the modified moment generating function method may be claimed to be:

1. it may be used when the moment generating function does not exist;
2. it may be used with fewer values for t than parameters;
3. it is not susceptible to numerical problems due to large exponential terms;
4. it results in estimators which possess bounded influence.

These aspects suggest that the q-L method may be preferred in practice. The following section provides an illustration of this method.

5.10 Estimation in the Cauchy distribution

The integrated squared error estimation of the parameters belonging to the Cauchy distribution was discussed in Chapter 1 (Section 1.13). In this section we shall bring out the differences between integrated squared error and q-L estimation for these parameters. In the latter case, the use of only a single point for t represents a situation meriting special study. Later in this section, consideration is also given to estimation using q $(q > 1)$ points.

5.10.1 Estimation using a single point

Let $X_1, X_2, \ldots, X_n$ be independent, identically distributed random variables from a Cauchy distribution with location δ and scale c, and suppose that $\boldsymbol{\theta} = (c, \delta)^\top$ is unknown. The characteristic function of X_1 is, from (1.29), given by

$$\phi(t; \boldsymbol{\theta}) = \exp(-c|t|)[\cos(t\delta) + i\sin(t\delta)]$$

and may be estimated by the empirical characteristic function

$$\phi_n(t) = U_n(t) + iV_n(t).$$

On this basis, the q-L estimator using a single value for t may be obtained by solving the system of equations

$$\left.\begin{aligned} U_n(t) &= \exp(-c|t|)\cos(t\delta) \\ V_n(t) &= \exp(-c|t|)\sin(t\delta) \end{aligned}\right\} \tag{5.26}$$

for $\boldsymbol{\theta}$. It is straightforward to show that if $t \neq 0$, then these equations result in the estimators

$$\hat{c} = \frac{\log[U_n^2(t) + V_n^2(t)]}{-2|t|} \tag{5.27}$$

$$\hat{\delta} = \frac{1}{t}\arctan[\frac{V_n(t)}{U_n(t)}]. \tag{5.28}$$

The explicit form of these estimators gives them an advantage over the integrated squared error estimators. However, there is a difficulty in the computation of $\hat{\delta}$ since, essentially, it is not unique. This is a consequence of the periodicity of the tangent function,

$$\tan(x + k\pi) = \tan(x), \quad k \in \mathbb{Z},$$

which implies that (cf. (5.28))

$$\tan(t\delta) = \frac{V_n(t)}{U_n(t)} \tag{5.29}$$

has an infinite number of solutions for δ. These are

$$\hat{\delta} = \frac{1}{t}\{\text{Arctan}[\frac{V_n(t)}{U_n(t)}] - 2k\pi\}, \quad k \in \mathbb{Z},$$

where Arctan denotes the principal value of the arctangent function. A similar problem has been reported by *Koutrouvelis (1980)* for a regression-type estimator

of the location parameter of a stable law.

In practice, it is of course impossible to select one solution over another without recourse to the original sample. *Morgan (1997, personal communication)* suggested the following approach. Given an initial estimate of δ, we choose from among the solutions of (5.29) the one which is nearest to this estimate. Figure 5.5 shows plots of $\hat{c}$ and the solution of (5.29) nearest to the median for a random sample from a Cauchy distribution with zero location and unit scale.

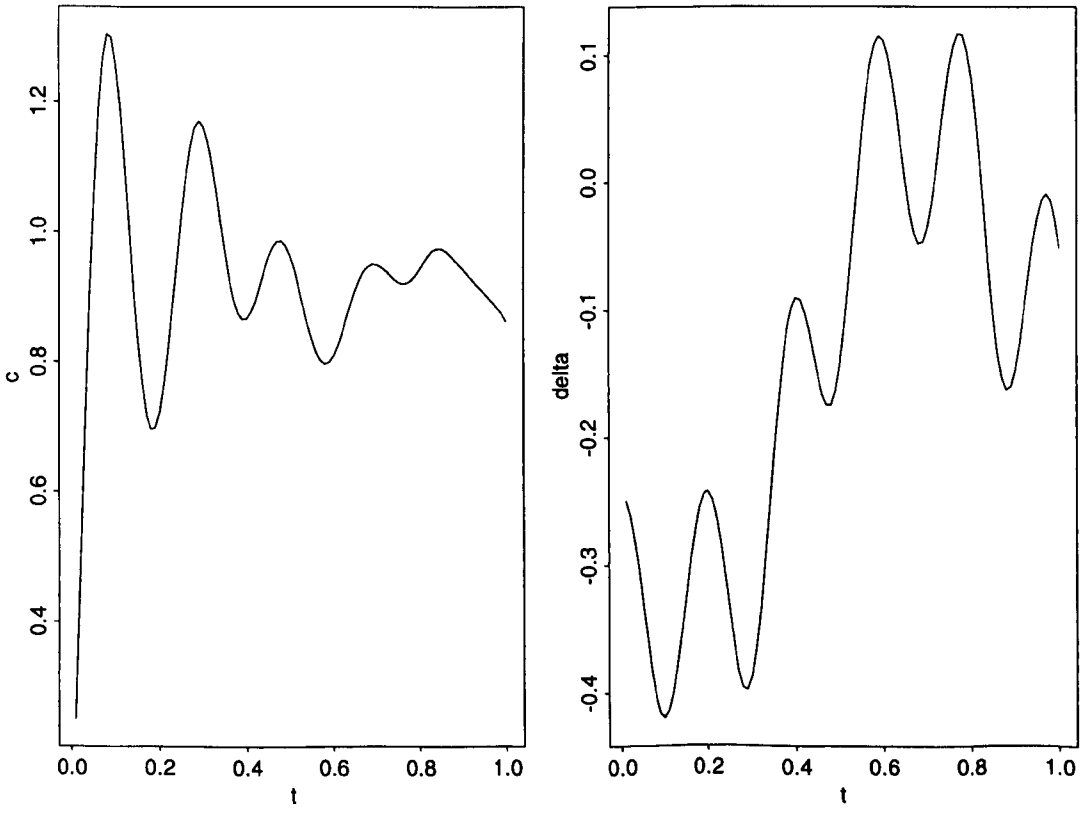

Figure 5.5: The explicit estimators $\hat{c}$ (left) and $\hat{\delta}$ nearest to the median (right) for a random sample of size $n = 50$ from a Cauchy distribution with location $\delta = 0$ and scale $c = 1$. Since the estimators are symmetric about $t = 0$, there is no need to plot negative values of t.

Despite the computational difficulty above, the asymptotic theory of the explicit estimators is quite straightforward. It follows, from Theorem 5.7, that

$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is asymptotically normal with mean vector zero and covariance matrix

$$\Sigma(\boldsymbol{\theta}) = \frac{\exp(2c|t|) - 1}{2t^2} I_2, \tag{5.30}$$

where I_2 denotes the 2×2 identity matrix. The Fisher information matrix for a single observation is $1/(2c^2)I_2$ so that, for example, the asymptotic efficiency of $\hat{\delta}$ is given by

$$\mathit{eff}(\hat{\delta}) = \frac{4c^2t^2}{\exp(2c|t|) - 1}.$$

This also happens to be the asymptotic efficiency of $\hat{c}$, and is depicted in Figures 5.6 and 5.7. As observed in these figures, the efficiency of $\hat{\delta}$ is low and, in fact, never exceeds 64.76%. The corresponding efficiency of the integrated squared error estimator was shown to be 98.09%. Thus, the explicit estimators are less efficient than the integrated squared error estimators.

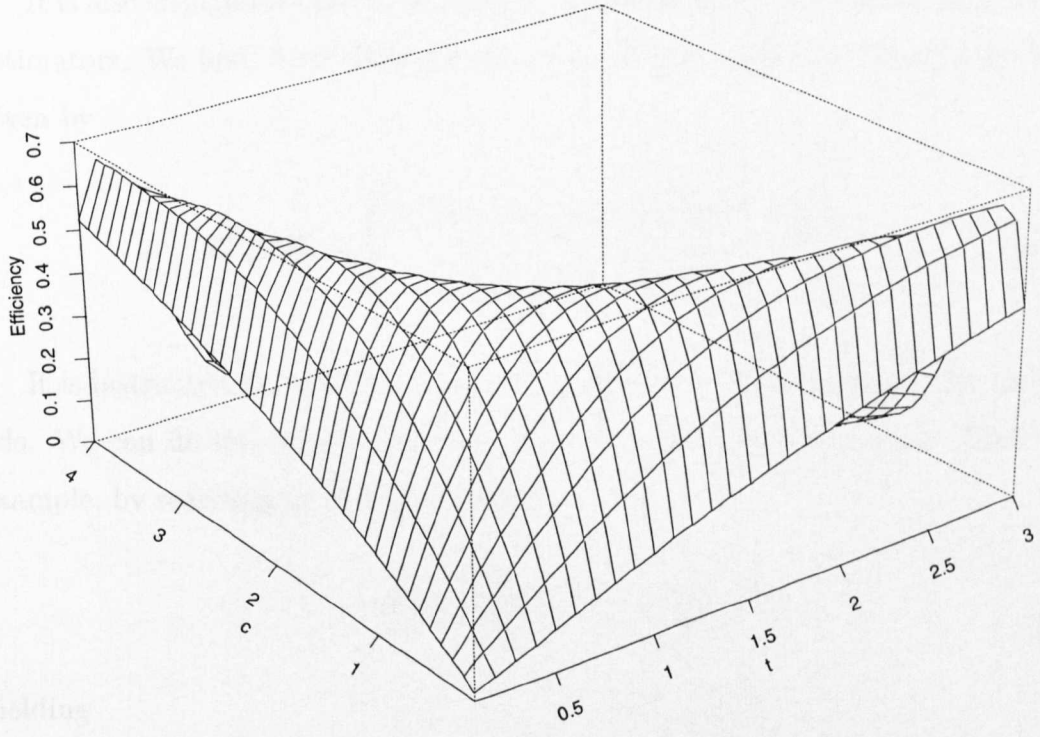

Figure 5.6: Perspective view of the asymptotic efficiency of $\hat{\delta}$. Note that $\mathit{eff}(\hat{\delta})$ is independent of δ.

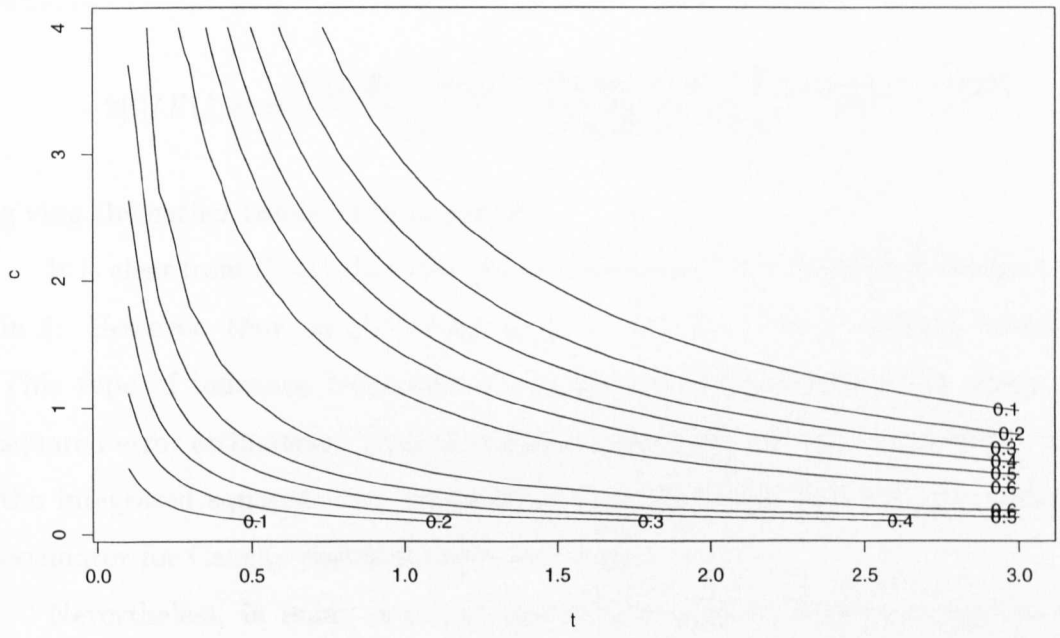

Figure 5.7: Contour-level plot of the efficiency of $\hat{\delta}$ as shown in Figure 5.6. The plot illustrates that $\hat{\delta}$ may be very inefficient.

It is also straightforward to investigate the robustness properties of the explicit estimators. We find, from Theorem 5.8, that the joint influence function for $\hat{\boldsymbol{\theta}}$ is given by

$$IF(\xi; \hat{\boldsymbol{\theta}}) = \begin{pmatrix} |t|^{-1}\{1 - \exp(c|t|)\cos[t(\xi - \delta)]\} \\ t^{-1}\exp(c|t|)\sin[t(\xi - \delta)] \end{pmatrix}. \tag{5.31}$$

It is instructive to verify this result by referring to elementary influence methods. We can do this easily here, but this will not always be the case. Thus, for example, by rearranging (5.27) we obtain

$$-2|t|\hat{c} = \log[U_n^2(t) + V_n^2(t)]$$

yielding

$$-2|t|IF(\xi; \hat{c}) = \frac{U(t; \boldsymbol{\theta})IF[\xi; U_n(t)] + 2V(t; \boldsymbol{\theta})IF[\xi; V_n(t)]}{U^2(t; \boldsymbol{\theta}) + V^2(t; \boldsymbol{\theta})}$$

whence

$$-2|t|IF(\xi;\hat{c}) = \frac{U(t;\boldsymbol{\theta})[\cos(t\xi) - U(t;\boldsymbol{\theta})] + 2V(t;\boldsymbol{\theta})[\sin(t\xi) - V(t;\boldsymbol{\theta})]}{U^2(t;\boldsymbol{\theta}) + V^2(t;\boldsymbol{\theta})}$$

giving the earlier result when simplified.

It is clear from (5.31) that the individual influence functions for $\hat{\boldsymbol{\theta}}$ are bounded in ξ. However, they do not decay as $|\xi| \to \infty$, rather they oscillate infinitely. This type of influence behaviour is less attractive than that of the integrated squared error estimators. Thus the explicit estimators are also "less robust" than the integrated squared error estimators. It appears that the cost of an explicit estimator for Cauchy distributions is very high.

Nevertheless, in many practical applications explicit estimators may be the ones of choice for reasons of economy, in money, time and effort. Practical implementation of the present explicit estimators requires the selection of t. This choice is very important as was shown graphically in Figures 5.5, 5.6 and 5.7. One approach to decide on a suitable value for t is to use constrained maximum likelihood, as suggested in Section 5.3.3 and illustrated in Section 5.4. Alternatively, we may decide on a value for t by minimising the determinant of (5.30). This approach results in

$$t = 0.7968/c \tag{5.32}$$

at the optimum, giving the maximum efficiency of 64.76%. A difficulty with this choice is that it depends on the unknown c, and this has been discussed in a general context by *Ball and Milne (1996).*

A general solution to this problem may be described as follows. Denote the explicit estimator for c by $\hat{c}(t)$ to emphasise its dependence on t. Then, in a similar fashion to *Ball and Milne (1996)*, choose $t^{(0)} > 0$ arbitrarily and let $c^{(0)} = \hat{c}(t^{(0)})$; next, for $m = 1, 2, \ldots$ let $t^{(m)} = 0.7968/c^{(m-1)}$ and $c^{(m)} = \hat{c}(t^{(m)})$; finally, if this iteration converges, with $t^{(m)} \to t^A$ as $m \to \infty$, then c can be estimated by the *adaptive estimator* $\hat{c}(t^A)$.

Suppose that the iteration described above converges. Then, from (5.32),

$$\hat{c}(t^A) = 0.7968/t^A \tag{5.33}$$

and substitution in (5.27) shows that t^A satisfies

$$|\phi_n(t^A)|^2 - 0.2032 = 0. \tag{5.34}$$

The function $|\phi_n(t)|^2$ is continuous in t, equals unity at $t = 0$ and tends to zero in an appropriate sense as $t \to \infty$ (see, for example, *Koutrouvelis, 1980*). This implies that (5.34) has at least one root; it will generally have multiple roots, in which case the smallest root should be selected. This is because $\phi_n(t)$ estimates $\phi(t; \boldsymbol{\theta})$ more accurately for small than large values of t, as illustrated in Chapter 1 (Section 1.5). Note that, provided (5.34) has at least one root, (5.33) produces an estimator for c irrespective of whether the iteration above converges.

5.10.2 Estimation using $q > 1$ points

The q-L estimator using a single point involves solving the equation system (5.26) and is thus easily computed. On the other hand, this estimator is not particularly efficient or robust. Of course, the efficiency can improve as more points for t are taken, as shown by *Feuerverger and McDunnough (1981b)* who, when proving the arbitrarily high efficiency of the q-L estimator, used the Cauchy distribution as an example. However, they only used equally spaced points: $\tau, 2\tau, \ldots, q\tau$ $(\tau > 0)$. *Koutrouvelis (1982)* also considered this problem but for general $t_1, t_2, \ldots, t_q$. We shall now present his asymptotic results and investigate the robustness properties of the resulting estimator.

The q-L estimator using $q > 1$ points proceeds by minimising (5.23) with respect to $\boldsymbol{\theta}$. This estimator for δ is not unique, as was the case when $q = 1$. In practice, we essentially choose that estimate which is nearest to the starting value used in the minimisation process. Furthermore, the q-L estimator becomes

increasingly complicated as q increases. This is mainly a consequence of the matrix inversion required for its computation. Nevertheless, *Koutrouvelis (1982)* showed that $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is asymptotically normal with mean vector zero and covariance matrix

$$\Sigma(\boldsymbol{\theta}) = \frac{1}{2}[\sum_{j=1}^{q} \frac{(t_j - t_{j-1})^2}{\exp(2ct_j) - \exp(2ct_{j-1})}]^{-1} I_2, \tag{5.35}$$

where I_2 is the 2×2 identity matrix and $t_0 \equiv 0$. The estimators $\hat{c}$ and $\hat{\delta}$ are thus asymptotically independent and have the same asymptotic variance. They also share the same asymptotic efficiency

$$eff(\hat{\delta}) = \sum_{j=1}^{q} \frac{(2ct_j - 2ct_{j-1})^2}{\exp(2ct_j) - \exp(2ct_{j-1})}. \tag{5.36}$$

By setting $u_j = 2ct_j$ $(j = 1, 2, \ldots, q)$, *Koutrouvelis (1982)* observed that (5.36) coincides with the asymptotic efficiency of an estimator for the scale parameter of an exponential distribution, which is based on q order statistics. Thus the problem of determining the optimum points t_j $(j = 1, 2, \ldots, q)$ for $\hat{\boldsymbol{\theta}}$ is equivalent to the problem of determining the optimum quantiles u_j $(j = 1, 2, \ldots, q)$ for this scale estimator. The latter problem has a known solution (see, for example, *Sarhan, Greenberg and Ogawa, 1963*).

Table 5.2 displays the maximum asymptotic efficiency thus obtained for selected values of q. As a point of comparison, the maximum efficiency of the explicit estimator is also included.

Table 5.2: Maximum efficiency of $\hat{\delta}$ for selected values of q

	Values of q						
	1	2	3	4	5	10	15
$eff(\hat{\delta})$	.6476	.8203	.8910	.9269	.9476	.9832	.9918

As would be expected from the discussion given earlier, the maximum efficiency is a monotonic increasing function in q. It is also apparent from the table that it does not take very many points for t to get a high asymptotic efficiency. As a practical matter, it appears that $q = 5$ would probably be sufficient in most cases. On the other hand, at least $q = 10$ points are required in order to improve on the efficiency of the integrated squared error estimator.

In contrast with the asymptotic efficiency of $\hat{\boldsymbol{\theta}}$, a symbolic expression for its influence function is difficult to produce. For example, even when $q = 2$, we find that

$$IF(\xi; \hat{\boldsymbol{\theta}}) = [\sum_{j=1}^{2} \frac{(t_j - t_{j-1})^2}{\exp(2ct_j) - \exp(2ct_{j-1})}]^{-1} \begin{pmatrix} \rho_1(\xi; \boldsymbol{\theta}) \\ \rho_2(\xi; \boldsymbol{\theta}) \end{pmatrix}, \tag{5.37}$$

where

$$\begin{aligned} \rho_1(\xi; \boldsymbol{\theta}) = & \frac{e^{ct_1}[t_2(e^{2ct_1} - 1) - t_1(e^{2ct_2} - 1)]}{(e^{2ct_1} - 1)(e^{2ct_2} - e^{2ct_1})} \cos[t_1(\xi - \delta)] \\ & - \frac{e^{ct_2}(t_2 - t_1)}{e^{2ct_2} - e^{2ct_1}} \cos[t_2(\xi - \delta)] + \frac{t_1}{e^{2ct_1} - 1} \\ \rho_2(\xi; \boldsymbol{\theta}) = & \frac{e^{ct_1}[t_1(e^{2ct_2} - 1) - t_2(e^{2ct_1} - 1)]}{(e^{2ct_1} - 1)(e^{2ct_2} - e^{2ct_1})} \sin[t_1(\xi - \delta)] \\ & + \frac{e^{ct_2}(t_2 - t_1)}{e^{2ct_2} - e^{2ct_1}} \sin[t_2(\xi - \delta)]. \end{aligned}$$

Since the influence function (5.37) is not trivial, some checks should really be applied. In general, simple checks are not easily constructed but, in this case, it is straightforward to show that (5.37) tends to (5.31) as $t_2 \to t_1$, as it should. A further interesting comparison between (5.37) and (5.31) comes from selecting t_i $(i = 1, 2)$ and t, respectively, so as to maximise efficiency. Figure 5.8 provides plots of the individual influence functions for these estimators evaluated at the Cauchy distribution with $\delta = 0$ and $c = 1$.

As anticipated, these influence functions are bounded in ξ. However, they do not decay as $|\xi| \to \infty$. The q-L influence functions with $q = 1$ are essentially

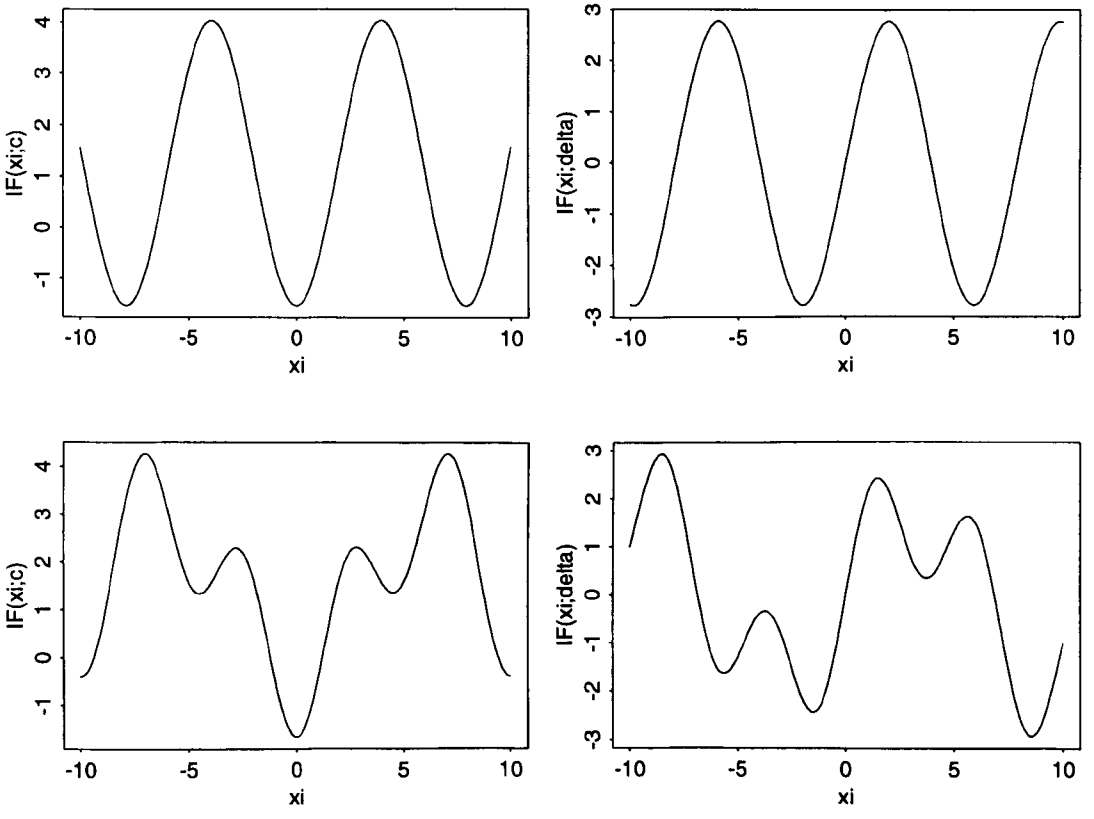

Figure 5.8: Influence functions for the q-L estimators of c (left) and δ (right) with $q = 1$ (top) and $q = 2$ (bottom). The points for t are selected to maximise efficiency at the Cauchy distribution with location $\delta = 0$ and scale $c = 1$.

periodic. The q-L influence functions with $q = 2$ are almost periodic. It is clear that increasing q affects the periodicity of the influence functions. On the other hand, it does not necessarily improve the robustness of the estimator, as noted by *Feuerverger and McDunnough (1981a)*. This is in fact related to the efficiency of the estimator. In particular, according to the results of *Feuerverger and McDunnough (1981a)*, the asymptotic variance of the q-L estimator can be made arbitrarily close to the asymptotic variance of the maximum likelihood estimator by increasing q. Thus, from the relationship between asymptotic variances and influence functions (see expression (1.16)), it seems reasonable to expect that it is also possible to make the q-L influence functions arbitrarily close to those of maximum likelihood. Note, however, that the influence functions of the latter estimator are not necessarily bounded.

This discussion is illustrated in Figure 5.9 below. The figure depicts the individual influence functions for the q-L estimator with $q = 10$. (This estimator is of particular interest here, since it can parallel the asymptotic efficiency of the integrated squared error estimator.) Also included in the figure are the individual influence functions for the maximum likelihood estimator.

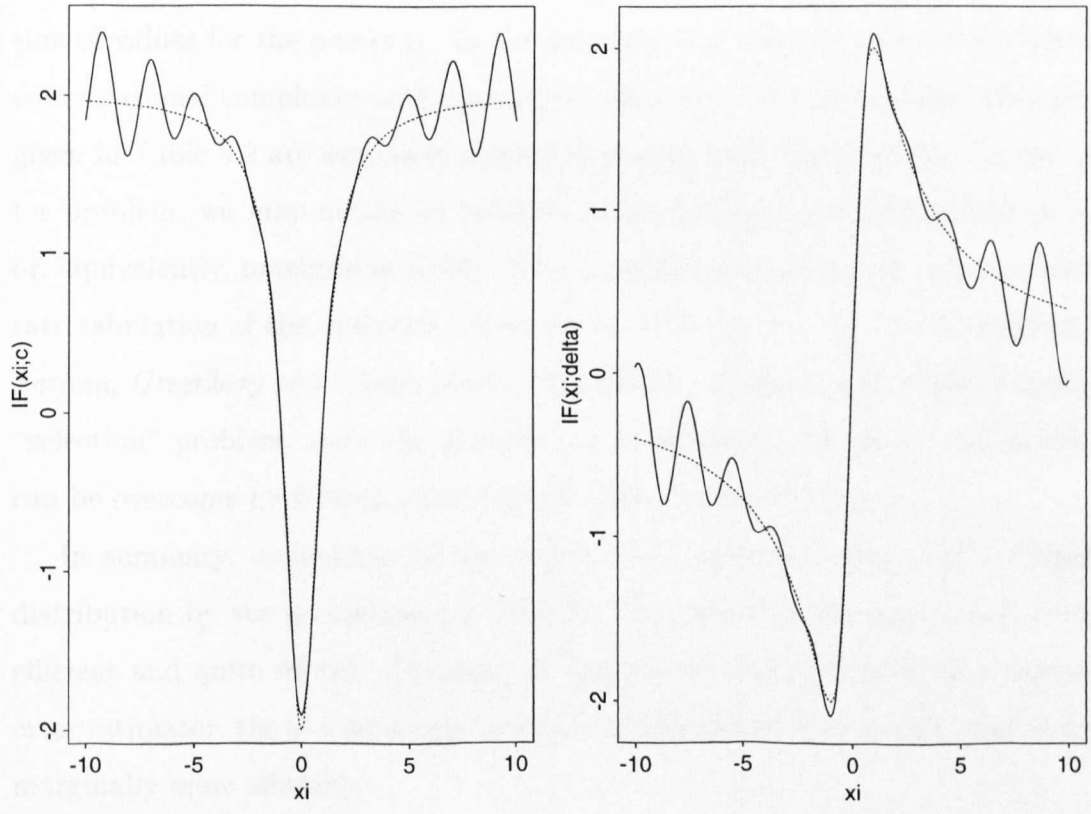

Figure 5.9: Individual influence functions for the q-L estimator with $q = 10$ (solid lines) overlaid with individual influence functions for the maximum likelihood estimator (dotted lines). The points for t are selected to maximise efficiency at the Cauchy distribution with location $\delta = 0$ and scale $c = 1$.

As observed in the figure, the q-L influence functions with $q = 10$ are much closer to those of maximum likelihood than the q-L influence functions with, for example, $q = 1$ or $q = 2$. The figure also illustrates that the q-L estimator with $q = 10$ is more robust than the q-L estimator with, for example, $q = 1$ or $q = 2$. However, this is a consequence of the maximum likelihood estimator being more robust than either of these estimators; this is not a situation that can generally

be expected. In any case, the integrated squared error estimator was found to be more robust than the maximum likelihood estimator (see Figure 1.10) and is, therefore, more robust than any of the q-L estimators.

Nevertheless, practical implementation of the q-L estimator requires the selection of the points t_j $(j = 1, 2, \ldots, q)$. In parallel to the moment generating function-based estimators, this involves (i) the selection of q, and (ii) the selection of values for the points t_j. In the selection of q, there is a trade-off between computational complexity and asymptotic efficiency. The asymptotic efficiencies given in Table 5.2 are especially helpful in dealing with this problem. In the latter problem, we may decide on suitable values for the t_j by minimising (5.35), or, equivalently, maximising (5.36). This is particularly easy here, since an accurate tabulation of the optimum values for $u_j = 2ct_j$ $(j = 1, 2, \ldots, 15)$ is given by *Sarhan, Greenberg and Ogawa (1963).* In practice, of course, this leads to another "selection" problem, since the parameter c is unknown. However, this problem can be overcome by following the principle used in the $q = 1$ case.

In summary, estimation of the location and scale parameters of a Cauchy distribution by the q-L method is feasible. The resulting estimator can be very efficient and quite robust. However, in comparison with the integrated squared error estimator, the q-L estimator is more difficult to use, less robust, and at best marginally more efficient.

5.11 Concluding remarks

This chapter has considered minimum distance transform estimation utilising step weight functions. This involves perhaps the simplest estimation method involving integral transforms, where we equate the empirical and theoretical moment generating functions at as many values for t as there are parameters. However, this method required us to select suitable values for t, was not particularly robust, and may not be especially efficient. Three extensions of the moment generating

function method (the preferred moment generating function method, the modified moment generating function method, and the q-L method) were introduced to help to deal with these difficulties. The latter method was shown to be the better of the three in terms of robustness, but less robust than the integrated squared error method, as illustrated in the Cauchy distribution.

Chapter 6

Conclusions and future work

6.1 Conclusions

The subject matter of this thesis has been parameter estimation using integral transforms, such as characteristic functions and moment generating functions. This type of parameter estimation has an extensive history in statistics, going back nearly forty years. This has resulted in a variety of methods, depending on the type of weight function and choice of integral transform, but little was known about their relative performances. The desirability for such a comparison provided the motivation for this thesis.

On this basis the first method we investigated was the integrated squared error method. An important decision that then needs to be made is the mathematical form for the weight function, but there has been very little practical work on this in the literature. The results of this thesis have shown, however, that it is not so much the choice of the weight function that is paramount but rather its scaling. Accordingly, we proposed several ways of selecting the scaling parameter in practice.

The main application of the integrated squared error method presented was to the mixture of two normal distributions. Theory was developed, which enabled a comparison to be made between this method and the methods of moments

and maximum likelihood. We consequently recommended estimating the parameters of a mixture of two normal distributions using the integrated squared error method, particularly with samples with outliers, or a small number of observations.

As indicated above, we also investigated several alternatives to the integrated squared error method, including the moment generating function method and the q-L method. These alternative means of parameter estimation can lead to considerable computational advantages, but were shown to generally underperform relative to the integrated squared error method. Hence, if parameter estimation is to be based on integral transforms, then the use of the integrated squared error method is recommended *a fortiori*.

6.2 Future work

The ideas that were developed in this thesis can be extended, with appropriate modifications, to cover a variety of settings. This flexibility will be illustrated in this section by means of three applications. Specifically, the first of these is to kernel density estimation and involves, as previously, univariate (independent, identically distributed) random variables. The mixture of multivariate normal distributions provides the focus of the second application, whilst the last application is to quantal assay models, which is a model for random variables indexed by dose.

6.2.1 Application to kernel density estimation

Kernel density estimation has been used extensively throughout this thesis. In essence, the kernel estimate is constructed by centring a kernel at each observation and then summing to obtain the estimate. As indicated in Chapter 2, the choice for the width of the kernel, called the bandwidth, is very important. Although several different methods have been suggested to help in this respect, the choice between them is often not clear-cut. When the kernel is the normal density, the kernel

estimator is essentially a mixture of n normal distributions, with component means coinciding with the observations $X_1, X_2, \ldots, X_n$ and equal mixing proportions of $1/n$ for each component. The difficulty of the bandwidth choice now becomes evident, since the standard deviation cannot be estimated by maximum likelihood (the likelihood function is not bounded above in Θ—see Chapter 3, Section 3.5.3). On the other hand, this mixture representation suggests that integrated squared error estimation could provide an alternative method of selecting the bandwidth.

This method would select the bandwidth by minimising (3.39) with respect to h, where we set $k = n$, $p_i = 1/n$, $\mu_i = X_i$, and $\sigma_i = h$ $(i = 1, 2, \ldots, n)$. A major advantage of the method is that it does not depend on asymptotic approximations but, on the other hand, it depends on the parameter λ. In fact, from the density representation (3.40), this parameter may be regarded as the bandwidth of an auxiliary kernel estimator. Since considerable theory has been devoted to the choice of auxiliary bandwidths (*Wand and Jones, 1995, p. 77*), this consolidates the results of Chapter 2 concerning the choice of weight function in integrated squared error estimation. It also provides a starting point for selecting the parameter λ in the present context. The results of this thesis suggest that this method of bandwidth selection would be worth considering.

6.2.2 Multivariate random variables

Throughout this work the emphasis has been on estimating parameters of univariate distributions. However, there are situations in which a univariate distribution is clearly inappropriate and a multivariate distribution needs to be considered. An important issue to contemplate in these situations is the robustness of the parameter estimators. This is both because it is more difficult to detect outliers and because there is, loosely speaking, more space in which outliers can occur. The results of this thesis suggest that the integrated squared error method will provide attractive robust estimators in the multivariate case, and future work could seek to discover if this is true. We envisage that the mixture of multivariate normal distributions will be a particularly fertile application for this method. This follows

from the work of *Everitt and Hand (1981, p. 44)*, who indicated that the difficulties associated with both the singularities in the likelihood function and with the choice of suitable starting values for the EM algorithm are far more critical in the multivariate than the univariate case. In this section we shall discuss some of these ideas, without getting into the more theoretical issues.

Integrated squared error estimation

An attractive feature of the integrated squared error method of estimating parameters of univariate distributions is that it generalises directly to the multivariate case, as outlined below.

Let $F(\boldsymbol{x};\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$ be a d-variate distribution function with characteristic function

$$\phi(\boldsymbol{t};\boldsymbol{\theta}) = \int_{\mathbb{R}^d} \exp(i\boldsymbol{t}^\top \boldsymbol{x})\, \mathrm{d}F(\boldsymbol{x};\boldsymbol{\theta}), \quad \boldsymbol{t} \in \mathbb{R}^d,$$

and suppose that $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ is a random sample from a population with this distribution function. Then, by analogy with the univariate case, the integrated squared error method of estimating $\boldsymbol{\theta}$ is based on the empirical characteristic function

$$\phi_n(\boldsymbol{t}) = n^{-1} \sum_{j=1}^{n} \exp(i\boldsymbol{t}^\top \boldsymbol{X}_j)$$

and involves minimising the integral

$$I(\boldsymbol{\theta};\Lambda) = \int_{\mathbb{R}^d} |\phi_n(\boldsymbol{t}) - \phi(\boldsymbol{t};\boldsymbol{\theta})|^2 \, \mathrm{d}W(\boldsymbol{t};\Lambda) \tag{6.1}$$

with respect to $\boldsymbol{\theta}$. As previously, $W(\boldsymbol{t};\Lambda)$ is a weight function, but now Λ is a symmetric positive definite matrix of parameters. In general, Λ will have $d(d+1)/2$ independent elements which, even for moderate d, can be a substantial number of parameters to choose. Viewed from this perspective, the arguments for exploiting the link between integrated squared error and kernel density estimation become much stronger in the multivariate case.

The more serious problem that arises with (6.1) is that it involves multivariate

integrals which, in general, would be difficult to evaluate. However, *Paulson and Lawrence (1980)* showed that (6.1) can be evaluated *explicitly* for the multivariate normal distribution. Taking the univariate case as a basis, it seems reasonable to expect that this result will also hold for the mixture of multivariate normal distributions, and this application is considered below.

Application to the mixture of two multivariate normal distributions

Mixtures of multivariate distributions are formed in the same way as for univariate distributions (see Chapter 3, Section 3.2). For example, in the particular case of two d-variate normal component densities to be discussed below, the mixture is defined by the probability density function

$$f(\boldsymbol{x}; p, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, V_1, V_2) = p\, g(\boldsymbol{x}; \boldsymbol{\mu}_1, V_1) + (1-p)\, g(\boldsymbol{x}; \boldsymbol{\mu}_2, V_2), \quad \boldsymbol{x} \in \mathbb{R}^d, \tag{6.2}$$

where

$$g(\boldsymbol{x}; \boldsymbol{\mu}_j, V_j) = (2\pi)^{-d/2} |V_j|^{-1/2} \exp[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_j)^\top V_j^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_j)]$$

is the density function of the d-variate normal distribution with mean vector $\boldsymbol{\mu}_j$ and covariance matrix V_j. The characteristic function corresponding to (6.2) is given by

$$\phi(\boldsymbol{t}; p, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, V_1, V_2) = p\, \psi(\boldsymbol{t}; \boldsymbol{\mu}_1, V_1) + (1-p)\, \psi(\boldsymbol{t}; \boldsymbol{\mu}_2, V_2), \quad \boldsymbol{t} \in \mathbb{R}^d, \tag{6.3}$$

where

$$\psi(\boldsymbol{t}; \boldsymbol{\mu}_j, V_j) = \exp(i\boldsymbol{t}^\top \boldsymbol{\mu}_j - \frac{1}{2}\boldsymbol{t}^\top V_j \boldsymbol{t})$$

is the characteristic function of the d-variate normal distribution with mean vector $\boldsymbol{\mu}_j$ and covariance matrix V_j.

We shall henceforth suppress dependence on the parameters $p, \boldsymbol{\mu}_i, V_i$ $(i = 1, 2)$

for notational convenience. As in the previous work, we shall assume that parameter estimation is based on a sequence $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ of independent, identically distributed random variables. Then the integrated squared error function is given under (6.3) by

$$I(\Lambda) = \int_{\mathbb{R}^d} |\phi_n(\boldsymbol{t}) - \phi(\boldsymbol{t})|^2 \, \mathrm{d}W(\boldsymbol{t}; \Lambda), \tag{6.4}$$

where $W(\boldsymbol{t}; \Lambda)$ is a weight function to be chosen. There are good computational reasons (see below) for selecting the weight function to be of the form

$$W(\boldsymbol{t}; \Lambda) = \int_{Y^d} \exp(-\boldsymbol{y}^\top \Lambda \boldsymbol{y}) \, \mathrm{d}\boldsymbol{y},$$

where $Y^d = (-\infty, t_1] \times (-\infty, t_2] \times \cdots \times (-\infty, t_d]$ and Λ is a positive definite matrix of constants. In this case, the integrated squared error function has, by Parseval's theorem, the equivalent expression

$$I(\Lambda) = (2\pi)^d \int_{\mathbb{R}^d} [f_n(\boldsymbol{x}; \Lambda) - f(\boldsymbol{x}) * g(\boldsymbol{x}; \boldsymbol{0}, \Lambda)]^2 \, \mathrm{d}\boldsymbol{x},$$

where the asterisk represents the operation of convolution and $f_n(\boldsymbol{x}; \Lambda)$ is a d-variate kernel density estimator with kernel $g(\boldsymbol{x}; \boldsymbol{0}, \Lambda)$ and bandwidth matrix Λ.

The work of *Paulson and Lawrence (1980)* on the d-variate normal distribution suggests that this integrated squared error function will admit an explicit form. In addition, *Wand and Jones (1995, pp. 108-109)* discussed the problem of selecting the bandwidth matrix from the data which, as previously indicated, is of considerable practical importance in integrated squared error estimation.

We conjecture from the above that the integrated squared method for estimating the parameters of a mixture of multivariate normal distributions will be feasible. We expect that the method will have its drawbacks, just like any other method. At the very least it could be useful in selecting good starting values for the EM algorithm.

6.2.3 Indexed random variables

Finally in this chapter, we note that the transform methods proposed for identically distributed random variables could also be used for indexed stochastic models. Parameter estimation using integral transforms has been applied to stochastic models, involving indexed random variables, by a number of authors, including *Schuh and Tweedie (1979)*, *Feigin, Tweedie and Belyea (1983)*, *Leedow and Tweedie (1983)*, *Laurence and Morgan (1987b)*, *Tweedie, Zhu and Choy (1995)*, and *Yao and Morgan (1999)*. The original attraction for using transform methods in these models was mainly the possibility of obtaining explicit parameter estimators. Furthermore, these methods were found to produce robust parameter estimators, as illustrated by *Paulson and Nicklin (1983)* and *Campbell (1993)*. In fact, transform methods for indexed random variables have several points of contact with those for identically distributed random variables. However, they are essentially different as they employ a different kind of empirical transform. In this section we shall discuss some of the fundamentals of transform methods in the context of indexed stochastic models.

Transform estimation

A very general class of indexed stochastic models is of the form

$$Y_j = r(t_j; \boldsymbol{\theta}) + \epsilon_j, \quad j = 1, 2, \ldots, n, \tag{6.5}$$

in which the ϵ_j are independent random variables with zero mean and $r(t_j; \boldsymbol{\theta})$ is a deterministic function involving the parameters we wish to estimate. Experimental data used to fit an indexed stochastic model typically consist of measurements on Y_j $(j = 1, 2, \ldots, n)$ taken at a sequence of sampling points $0 < t_1 \leq t_2 \leq \cdots \leq t_n < \infty$. Later in this section we focus on the quantal assay model, which may be expressed in this form.

In order to estimate the parameters of these models, we may adapt the integrated squared error approach of Chapter 1. In particular, if the ϵ_j of (6.5) are

also identically normally distributed with variance σ^2, then we may estimate

$$\exp[isr(t_j;\boldsymbol{\theta}) - \tfrac{1}{2}\sigma^2 s^2]$$

by

$$\exp(isY_j)$$

and the parameters by minimising an integrated distance between these two transforms, as described by *Paulson and Nicklin (1983)* for a variety of linear models for $r(t;\boldsymbol{\theta})$. The work of this thesis suggests that this method is well worth considering. However, this approach has not been favoured in the literature.

On the other hand, the least-squares approach of Chapter 5 has been used, even though it has presented new problems of its own. For example, in the case of identically distributed random variables the quantity to transform was the underlying distribution function. However, in the present context the quantity to transform is often less clear. Furthermore, we must also select the type of transform to employ. Given our previous influence work, the characteristic function is the natural choice. Interestingly, the Laplace transform has been generally employed in this area. Thus, if we decide to transform $r(t;\boldsymbol{\theta})$, then we form

$$L(s;\boldsymbol{\theta}) = \int_0^\infty r(t;\boldsymbol{\theta}) \exp(-st)\, \mathrm{d}t \tag{6.6}$$

and seek to estimate the parameters by equating (6.6) to its empirical version, which can be estimated from the data described above.

This method is analogous to the moment generating function method of Chapter 5 and is independent of the error structure in (6.5). Furthermore, the asymptotic properties of the resulting estimator have been derived by *Yao and Morgan (1999)*. However, the choice of empirical transform is crucial to the performance of this estimator and this is addressed below.

The empirical transform

In contrast to the case of independent identically distributed random variables, the form of the empirical transform for indexed random variables is not unique. One form is based on a Riemann sum approximation to the integral $L(s;\boldsymbol{\theta})$ and is given by

$$L_n(s) = s^{-1}\sum_{j=1}^{n} Y_j[\exp(-sc_{j-1}) - \exp(-sc_j)], \tag{6.7}$$

where $c_0 = 0$, $c_n = \infty$ and c_j are constants satisfying $c_{j-1} \leq t_j \leq c_j$ $(j = 1, 2, \ldots, n-1)$. For example, one might simply take

$$c_j = \frac{1}{2}(t_j + t_{j+1}), \quad j = 1, 2, \ldots, n-1$$

so that the c_j's lie half-way between the design points.

This form of empirical transform was proposed by *Schuh and Tweedie (1979)* and has been employed by a number of authors, including *Yao and Morgan (1999)*. In general, the expectation of (6.7) will not equal to (6.6) and, furthermore, a poor approximation could result if setting $c_0 = 0$ and $c_n = \infty$ can give undue weighting to Y_1 and Y_n, respectively. The first problem arises if t_1 is not close to zero. The simplest way to overcome this problem is by a judicious translation of the sampling points, given by

$$t_j^{\star} = t_j - \tau, \quad j = 1, 2, \ldots, n \tag{6.8}$$

for τ to be determined.

The second problem arises from integrating beyond the range of the data. In order to resolve this problem, an *end-correction* can be made by setting $c_n = t_n$, as suggested by *Leedow and Tweedie (1983)*. However, if we set $c_n = t_n$ in (6.7),

then the resulting empirical transform estimates

$$L(s;\boldsymbol{\theta}) - \int_{t_n}^{\infty} r(t;\boldsymbol{\theta}) \exp(-st)\, \mathrm{d}t$$

so that parameter estimation is inconvenienced by the second term. A similar approach could be used with respect to c_0, as well as c_n, but it frequently occurs that a translation of the sampling points is all that is necessary, as we shall see from the following application.

Application to the quantal assay model

Finney (1971, p. 20) presented the data of Table 6.1, which document the progress of insects sprayed with insecticide. In the jth of five experiments, n_j subjects are examined at log-dose t_j and the number, Y_j, of subjects which have responded is recorded.

Table 6.1: Quantal assay data from *Finney (1971, p. 20)*. Here t is the log, to base 10, of a dose of rotenone in mg/litre, and the subjects are insects, *Macrosiphoniella sanborni*. Response is death, or being seriously affected.

j	t_j	n_j	Y_j
1	0.41	50	6
2	0.58	48	16
3	0.71	46	24
4	0.89	49	42
5	1.01	50	44

If the insects are assumed to respond independently, we may model these data by

$$Y_j \sim \mathrm{Bin}\big(n_j, p(t_j;\boldsymbol{\theta})\big), \quad j = 1, 2, \ldots, 5$$

and

$$p(t;\boldsymbol{\theta}) = \{1 + \exp[-(\alpha + \beta t)]\}^{-1},$$

where $\boldsymbol{\theta} = (\alpha, \beta)^{\top}$ denotes the parameter vector.

Parameter estimation by the method of maximum likelihood results from the maximisation of the likelihood function

$$L(\boldsymbol{\theta}) = \prod_{j=1}^{5} \binom{n_j}{Y_j} p(t_j;\boldsymbol{\theta})^{Y_j}[1-p(t_j;\boldsymbol{\theta})]^{n_j-Y_j} \tag{6.9}$$

with respect to $\boldsymbol{\theta}$. A numerical procedure is needed to maximise (6.9), leading to the estimates $\tilde{\alpha} = -4.839$, $\tilde{\beta} = 7.068$ and a maximised log-likelihood of $\log[L(\tilde{\boldsymbol{\theta}})] = -119.856$.

Laurence, Morgan and Tweedie (1987b) note, however, that the Laplace transform of $[1-p(t;\boldsymbol{\theta})]^{-1} = 1 + e^{\alpha+\beta t}$ has the simple form

$$L(s;\boldsymbol{\theta}) = s^{-1} + (s-\beta)^{-1}\exp(\alpha), \quad s > \beta \tag{6.10}$$

and this is information which is not being utilised. By analogy with the moment generating function method, if we solve the system of equations

$$L_n(s_j) = L(s_j;\boldsymbol{\theta}), \quad j = 1,2,\ s_1 \neq s_2$$

for $\boldsymbol{\theta}$, then we would obtain the explicit estimators

$$\left.\begin{aligned} \hat{\alpha} &= \log\{\frac{(s_2-s_1)[s_1L_n(s_1)-1][s_2L_n(s_2)-1]}{s_1s_2[L_n(s_1)-L_n(s_2)]+s_1-s_2}\} \\ \hat{\beta} &= \frac{(s_1s_2[s_1L_n(s_1)-s_2L_n(s_2)]}{s_1s_2[L_n(s_1)-L_n(s_2)]+s_1-s_2}. \end{aligned}\right\} \tag{6.11}$$

As $s_2 \to s_1$, these estimators converge to

$$\left.\begin{aligned} \hat{\alpha} &= \log\{\frac{-[s_1L_n(s_1)-1]^2}{s_1^2L_n^{(1)}(s_1)+1}\} \\ \hat{\beta} &= \frac{s_1^2[s_1L_n^{(1)}(s_1)+L_n(s_1)]}{s_1^2L_n^{(1)}(s_1)+1} \end{aligned}\right\} \tag{6.12}$$

provided that $L_n^{(1)}(s_1) < -s_1^{-2}$; in fact, this is the solution of system of equations

$$L_n(s_1) = L(s_1; \boldsymbol{\theta})$$
$$L_n^{(1)}(s_1) = L^{(1)}(s; \boldsymbol{\theta}).$$

Figure 6.1 provides graphs of the explicit estimators for the data of Table 6.1. These estimates are based on the empirical transform

$$L_n(s) = s^{-1} \sum_{j=1}^{5} (1 - \frac{Y_j}{n_j})^{-1} [\exp(-sc_{j-1}) - \exp(-sc_j)] \tag{6.13}$$

with $c_0 = 0$, $c_n = \infty$ and $c_j = \frac{1}{2}(t_j + t_{j+1})$, $j = 1, 2, \ldots, 4$.

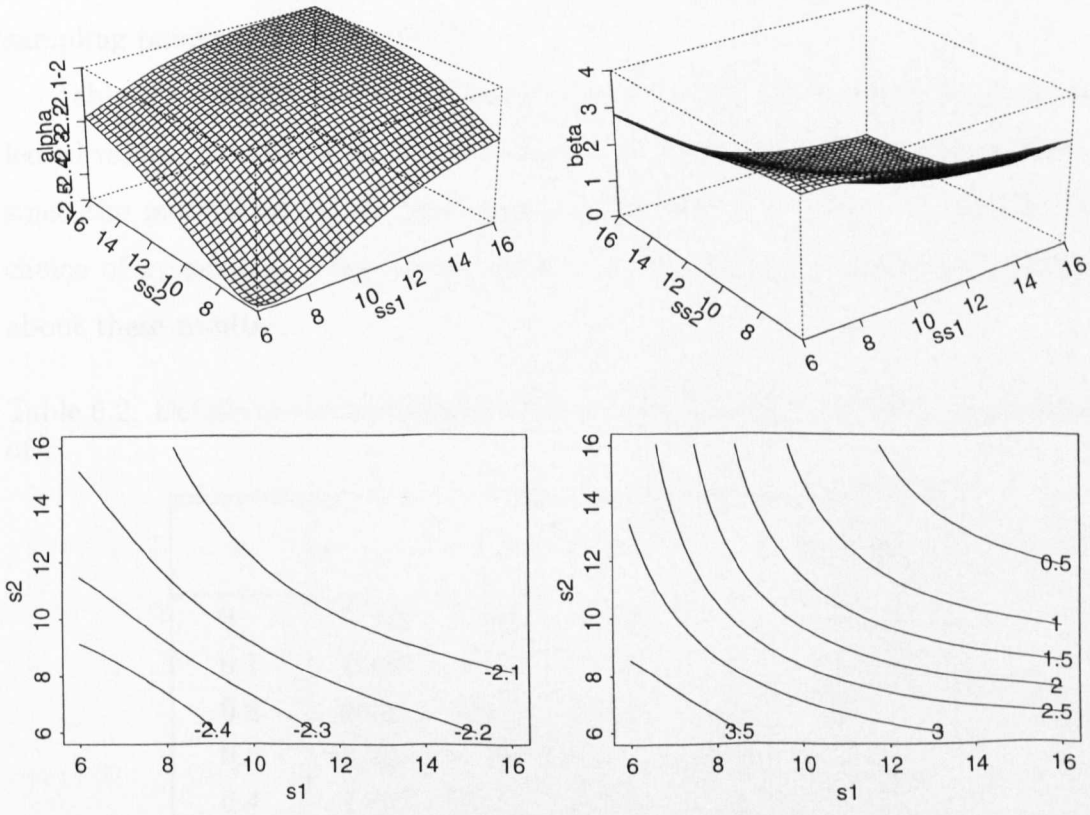

Figure 6.1: Perspective view (top) and contour level-plot (bottom) of the explicit estimator for α (left) and β (right) based on the set of quantal assay data of Table 6.1.

As observed in the figure, the agreement between the maximum likelihood

and explicit estimates is generally very poor. This may be due to a number of reasons, such as (1) the choice of s_1 and s_2, or (2) the poor approximation of (6.10) by (6.13). The choice of s_1 and s_2 may be addressed using one of the methods suggested in Chapter 5 (Section 5.3.3). Since the likelihood function is readily available and the sample sizes are not sufficiently large, we have used the method of constrained maximum likelihood. In this case, the constrained log-likelihood was maximised at $s_1 = 6.389$ and $s_2 = 6.389$, resulting in $\hat{\alpha} = -2.486$, $\hat{\beta} = 3.648$, and $\log[L(\hat{\boldsymbol{\theta}})] = -129.195$. (Recall that the unconstrained maximum log-likelihood was $\log[L(\tilde{\boldsymbol{\theta}})] = -119.856$, with $\tilde{\alpha} = -4.839$ and $\tilde{\beta} = 7.068$.)

The Riemann sum approximation (6.13) can be improved by setting $c_0 = t_1$ and $c_n = t_n$. In the present application the former correction is more beneficial than the latter and, furthermore, can be easily carried out by translating the sampling points according to (6.8).

Table 6.2 presents details of the performance of the explicit estimators for selected values of τ. The criterion of performance which we use is the log-likelihood, since the method of constrained maximum likelihood was used to provide the choice of s_1 and s_2. There seem to be a number of points well worth making about these results.

Table 6.2: Details of the performance of the explicit estimators for selected values of τ

τ	$\boldsymbol{s}$		$\hat{\boldsymbol{\theta}}$		$\log[L(\hat{\boldsymbol{\theta}})]$
	s_1	s_2	$\hat{\alpha}$	$\hat{\beta}$	
0	6.389	6.389	-2.486	3.648	-129.195
0.1	7.453	7.453	-2.886	4.202	-126.249
0.2	9.021	9.021	-3.421	4.961	-123.194
0.3	11.768	11.768	-4.171	6.075	-120.546
0.4	11.837	14.122	-4.839	7.068	-119.856

When $\tau = 0, 0.1, 0.2, 0.3$, the method of constrained maximum likelihood selected $s_1 = s_2$, indicating the possibility of reducing dimensionality by searching

only along this line. On the other hand, constrained maximum likelihood did not result in the maximum likelihood estimate in any of these cases. This is consistent with the intuitive idea that if the Riemann sum approximation is relatively poor for most s, then the constraints imposed by equating empirical and theoretical transforms will exclude the maximum likelihood estimate in most cases. However, as the Riemann sum approximation improves, the constraints have less effect and this can result in parameter estimates close to the maximum likelihood estimate. In fact, as the case $\tau = 0.4$ illustrates, this procedure can result in the maximum likelihood estimate. The unfortunate part is that, in this case, the constrained maximum likelihood was maximised nowhere near the line $s_1 = s_2$.

This last result shows the existence of exceptions to the diagonal optimisation phenomenon of *Yao and Morgan (1999)*. An insight into the existence of exceptions can perhaps be obtained from the fact that constrained maximum likelihood is essentially a reparametrisation for maximum likelihood. In particular, it yields a mapping

$$\begin{array}{ccccc} \mathbb{R}^k & \longrightarrow & \mathbb{R}^p & \longrightarrow & \mathbb{R} \\ \boldsymbol{s} & \longrightarrow & \boldsymbol{\theta} & \longrightarrow & L(\boldsymbol{\theta}) \end{array}$$

and seeks to maximise $L(\boldsymbol{\theta})$ by searching over $\boldsymbol{s}$-space, $\mathbb{R}^k$, rather than over $\boldsymbol{\theta}$-space, $\mathbb{R}^p$. If several derivatives of $L(s;\boldsymbol{\theta})$ and $L_n(s)$ are used, then k can be made considerably less than p, and which then does result in computational advantages. However, inferences cannot be invariant with respect to this reparametrisation unless $k = p$.

In summary, we are convinced that a least-squares transform approach can facilitate parameter estimation in indexed stochastic models in general. However, this approach is not without its problems and needs to be compared with the integrated squared error approach of *Paulson and Nicklin (1983)*. On the basis of the results of this thesis, we conjecture that the integrated squared error estimator will fare much better than its least-squares transform counterpart.

References

[1] Abdous, B. (1993). Note on the minimum mean integrated squared error of kernel estimates of a distribution function and its derivatives. *Communications in Statistics—Theory and Methods* **22**, 603–609.

[2] Azzalini, A. (1996). *Statistical Inference Based on the Likelihood.* London: Chapman and Hall.

[3] Ball, F. and Milne, R. K. (1996). On choosing values of the transform variables in empirical transform based inference. *Technical Report 96-09*, Department of Mathematics, University of Nottingham, United Kingdom.

[4] Barnett, V. and Lewis, T. (1994). *Outliers in Statistical Data.* 3 edn, Chichester: Wiley and Sons.

[5] Behboodian, J. (1970). On a mixture of normal distributions. *Biometrika* **57**, 215–217.

[6] Behboodian, J. (1972). Information matrix for a mixture a two normal distributions. *Journal of Statistical Computation and Simulation* **1**, 295–314.

[7] Binder, D. A. (1978). Comment on "Estimating mixtures of normal distributions and Switching Regressions". *Journal of the American Statistical Association* **73**, 746–747.

[8] Brooks, S. P. and Morgan, B. J. T. (1995). Optimization using simulated annealing. *Technometrics* **44**, 241–257.

[9] Bryant, J. L. and Paulson, A. S. (1979). Some comments on characteristic function-based estimators. *Sankhyā, Series A,* **41**, 109–116.

[10] Bryant, J. L. and Paulson, A. S. (1982). Estimation of the parameters of a modified compound Poisson distribution. *Technical Report 37-82-P2*, Rensselaer Institute, Troy, New York.

[11] Bryant, J. L. and Paulson, A. S. (1983). Estimation of mixing proportions via distance between characteristic functions. *Communications in Statistics—Theory and Methods* **12**, 1009–1029.

[12] Bryant, P. (1978). Comment on "Estimating mixtures of normal distributions and Switching Regressions". *Journal of the American Statistical Association* **73**, 748–749.

[13] Campbell, E. P. (1992). *Robustness of estimation based on empirical transforms*, PhD thesis, Institute of Mathematics and Statistics, University of Kent, United Kingdom.

[14] Campbell, E. P. (1993). Influence for empirical transforms. *Communications in Statistics—Theory and Methods* **22**, 2491–2502.

[15] Choi, K. and Bulgren, W. G. (1968). An estimation procedure for mixtures of distributions. *Journal of the Royal Statistical Society, Series B,* **30**, 444–460.

[16] Clarke, B. R. and Heathcote, C. R. (1978). Comment on "Estimating mixtures of normal distributions and Switching Regressions". *Journal of the American Statistical Association* **73**, 749–750.

[17] Cohen, A. C. (1967). Estimation in mixtures of two normal distributions. *Technometrics* **9**, 15–28.

[18] Cramér, H. (1946). *Mathematical Methods of Statistics.* Princeton: Princeton University Press.

[19] Csörgő, S. (1980). Empirical characteristic functions. Carleton Mathematical Lecture Notes No.26.

[20] Csörgő, S. (1981). Limit behaviour of the empirical characteristic function. *Annals of Probability* **9**, 130–144.

[21] Davis, D. J. (1952). An analysis of some failure data. *Journal of the American Statistical Association* **47**, 113–150.

[22] Day, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika* **56**, 463–474.

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

[24] Devroy, L. and Györfi, L. (1985). *Non-parametric Density Estimation: The L_1 View.* New York: Wiley and Sons.

[25] Dick, N. P. and Bowden, D. C. (1973). Maximum likelihood estimation for mixtures of two normal distributions. *Biometrics* **29**, 781–790.

[26] Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society, Series B,* **56**, 363–375.

[27] Ecker, F. J. (1983). The estimated Laplace transform and function estimation. Unpublished manuscript.

[28] Epps, T. W. (1993). Characteristic functions and their empirical counterparts: geometrical interpretations and applications to statistical inference. *The American Statistician* **47**, 33–38.

[29] Escobar, M. D. and West, M. (1995). Bayesian density-estimation and inference using mixtures. *Journal of the American Statistical Association* **90**, 577–588.

[30] Everitt, B. S. (1987). *Introduction to Optimization Methods and their Application in Statistics.* London: Chapman and Hall.

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

[32] Feigin, P. D. and Heathcote, C. R. (1976). The empirical characteristic function and the Cramér-von Mises statistic. *Sankhyā, Series A,* **38**, 309–325.

[33] Feigin, P. D., Tweedie, R. L. and Belyea, C. (1983). Weighted area techniques for explicit parameter estimation in multi-stage models. *Australian Journal of Statistics* **25**, 1–16.

[34] Feuerverger, A. and McDunnough, P. (1981a). On some Fourier methods for inference. *Journal of the American Statistical Association* **76**, 379–387.

[35] Feuerverger, A. and McDunnough, P. (1981b). On the efficiency of empirical characteristic function procedures. *Journal of the Royal Statistical Society, Series B,* **43**, 20–27.

[36] Feuerverger, A. and McDunnough, P. (1984). On statistical transform methods and their efficiency. *Canadian Journal of Statistics* **12**, 303–317.

[37] Feuerverger, A. and Mureika, R. A. (1977). The empirical characteristic function and its applications. *Annals of Statistics* **5**, 88–97.

[38] Finney, D. J. (1971). *Probit analysis.* 3 edn, Cambridge: Cambridge University Press.

[39] Fowlkes, E. B. (1978). Comment on "Estimating mixtures of normal distributions and Switching Regressions". *Journal of the American Statistical Association* **73**, 747–748.

[40] Fowlkes, E. B. (1979). Some methods for studying the mixture of two normal (lognormal) distributions. *Journal of the American Statistical Association* **74**, 561–575.

[41] Fryer, J. G. and Robertson, C. A. (1972). A comparison of some methods for estimating mixed normal distributions. *Biometrika* **59**, 639–648.

[42] Fryer, M. J. (1976). Some errors associated with the non-parametric estimation of density functions. *Journal of the Institute of Mathematics and its Applications* **18**, 371–380.

[43] Fryer, M. J. (1977). A review of some non-parametric methods of density estimation. *Journal of the Institute of Mathematics and its Applications* **20**, 335–354.

[44] Gnedenko, B. V. and Kolmogorov, A. N. (1954). *Limit Distributions for Sums of Independent Random Variables.* Reading, MA: Addison-Wesley.

[45] Hall, P. (1981). On the non-parametric estimation of mixture proportions. *Journal of the Royal Statistical Society, Series B,* **43**, 147–156.

[46] Hampel, F. R. (1971). A generalized qualitative definition of robustness. *Annals of Mathematical Statistics* **42**, 1887–1896.

[47] Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association* **69**, 383–393.

[48] Heathcote, C. R. (1977). The integrated squared error estimation of parameters. *Biometrika* **64**, 255–264.

[49] Heathcote, C. R. (1978). On parametric density estimators. *Advances in Applied Probability* **10**, 735–740.

[50] Hill, B. M. (1963). Information for estimating the proportions in mixtures of exponential and normal distributions. *Journal of the American Statistical Association* **58**, 918–932.

[51] Hosmer, D. W. (1973). A comparison of iterative maximum likelihood estimates of the parameters of a mixture of two normal distributions under three different types of sample. *Biometrics* **29**, 761–770.

[52] Hosmer, D. W. (1978). Comment on "Estimating mixtures of normal distributions and Switching Regressions". *Journal of the American Statistical Association* **73**, 741–744.

[53] Hosmer, D. W. and Dick, N. P. (1977). Information and mixtures of two normal distributions. *Journal of Statistical Computation and Simulation* **6**, 137–148.

[54] Huber, P. J. (1981). *Robust Statistics.* New York: Wiley and Sons.

[55] Ingrassia, S. (1992). A comparison between the simulated annealing and the EM algorithms in normal mixture decompositions. *Statistics and Computing* **2**, 203–211.

[56] Johnson, N. L. (1978). Comment on "Estimating mixtures of normal distributions and Switching Regressions". *Journal of the American Statistical Association* **73**, 750.

[57] Johnson, N. L., Kotz, S. and Balakrishnan, N. (1994). *Continuous Univariate Distributions.* Vol. 1, 2 edn, New York: Wiley and Sons.

[58] Johnson, N. L., Kotz, S. and Balakrishnan, N. (1995). *Continuous Univariate Distributions.* Vol. 2, 2 edn, New York: Wiley and Sons.

[59] Jones, M. C., Marron, J. S. and Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association* **91**, 401–407.

[60] Karr, A. F. (1993). *Probability.* New York: Springer-Verlag.

[61] Kemp, A. W. and Kemp, C. D. (1987). A rapid and efficient estimation procedure for the negative binomial distribution. *Biometrical Journal* **29**, 865–873.

[62] Kemp, C. D. and Kemp, A. W. (1988). Rapid estimation for discrete distributions. *Technometrics* **37**, 243–255.

[63] Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics* **27**, 887–906.

[64] Kiefer, N. M. (1978). Comment on "Estimating mixtures of normal distributions and Switching Regressions". *Journal of the American Statistical Association* **73**, 744–745.

[65] Kirkpatrick, S., Gelatt, C. D. and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science* **220**, 671–680.

[66] Koutrouvelis, I. A. (1980). Regression-type estimation of the parameters of stable laws. *Journal of the American Statistical Association* **75**, 918–928.

[67] Koutrouvelis, I. A. (1982). Estimation of location and scale in Cauchy distributions using the empirical characteristic function. *Biometrika* **69**, 205–213.

[68] Koutrouvelis, I. A. and Bauer, D. F. (1982). Asymptotic distribution of regression-type estimators of parameters of stable laws. *Communications in Statistics—Theory and Methods* **11**, 2715–2730.

[69] Koutrouvelis, I. A. and Canavos, G. C. (1997). Estimation in the three-parameter Gamma distribution based on the empirical moment generating function. *Journal of Statistical Computation and Simulation* **59**, 47–62.

[70] Kumar, K. D., Nicklin, E. H. and Paulson, A. S. (1979). Comment on "Estimating mixtures of normal distributions and Switching Regressions". *Journal of the American Statistical Association* **74**, 52–55.

[71] Laurence, A. F. and Morgan, B. J. T. (1987a). Quantal assay data with time to response: model-fitting using Laplace transforms. Unpublished manuscript.

[72] Laurence, A. F. and Morgan, B. J. T. (1987b). Selection of the transformation variable in the Laplace transform method of estimation. *Australian Journal of Statistics* **29**, 113–127.

[73] Laurence, A. F., Morgan, B. J. T. and Tweedie, R. L. (1987). Parameter-estimation in non-linear models using Laplace transforms. Unpublished manuscript.

[74] Leedow, M. I. and Tweedie, R. L. (1983). Weighted area techniques for the estimation of the parameters of a growth curve. *Australian Journal of Statistics* **25**, 310–320.

[75] Lehmann, E. L. (1983). *Theory of Point Estimation.* New York: Wiley and Sons.

[76] Leitch, R. A. and Paulson, A. S. (1975). Estimation of Stable Law Parameters: Stock Price Behavior Application. *Journal of the American Statistical Association* **70**, 690–697.

[77] Leslie, R. T. (1970). Parameter estimation via the moment generating function. *Bulletin of the Australian Statistical Society* **1**, 1–6.

[78] Leslie, R. T. and Khalique, A. (1980). Parametric estimation via Laplace transforms. Unpublished manuscript.

[79] Leslie, R. T. and McGilchrist, C. A. (1972). Estimation using the characteristic function of the smoothed probability function. Unpublished manuscript.

[80] Leytham, K. M. (1984). Maximum likelihood estimates for the parameters of mixture distributions. *Water Resources Research* **20**, 896–902.

[81] Lukacs, E. (1970). *Characteristic Functions.* London: Griffin.

[82] Marron, J. S. and Wand, M. P. (1992). Exact mean integrated squared error. *Annals of Statistics* **20**, 712–736.

[83] McLachlan, G. J. and Basford, K. E. (1988). *Mixture Models: Inference and Applications to Clustering.* New York: Marcel Dekker.

[84] Meng, X. L. and vanDyk, D. (1997). The EM algorithm - An old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society, Series B,* **59**, 511–540.

[85] Morgan, B. J. T. (1984). *Elements of Simulation.* London: Chapman and Hall.

[86] Morgan, B. J. T. and Tweedie, R. L. (1981). Stable parameter estimation using Laplace transforms in non linear regression. Unpublished manuscript.

[87] Murray, G. D. and Titterington, D. M. (1978). Estimation problems with data from a mixture. *Annals of Statistics* **27**, 325–334.

[88] Park, B. U. and Marron, J. S. (1990). Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association* **85**, 66–72.

[89] Parr, W. C. and Schucany, W. R. (1980). Minimum distance and robust estimation. *Journal of the American Statistical Association* **75**, 616–624.

[90] Parr, W. C. and Schucany, W. R. (1982). Minimum distance estimation and components of goodness-of-fit statistics. *Journal of the Royal Statistical Society, Series B,* **44**, 178–189.

[91] Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics* **33**, 1065–1076.

[92] Paulson, A. S. and Delehanty, T. A. (1984). Some properties of modified integrated squared error estimators for the stable laws. *Communications in Statistics—Simulation and Computation* **13**, 337–365.

[93] Paulson, A. S. and Delehanty, T. A. (1985). Modified weighted squared error estimation procedures with special emphasis on the stable laws. *Communications in Statistics—Simulation and Computation* **14**, 927–972.

[94] Paulson, A. S., Holcomb, E. W. and Leitch, R. A. (1975). The estimation of the parameters of the stable laws. *Biometrika* **62**, 163–170.

[95] Paulson, A. S. and Lawrence, C. E. (1980). Some modified integrated squared error procedures for multivariate normal data. Unpublished manuscript.

[96] Paulson, A. S. and Nicklin, E. H. (1983). Integrated distance estimators for linear models applied to some published data sets. *Applied Statistics* **32**, 32–50.

[97] Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London, Series A,* **185**, 71–110.

[98] Prakasa Rao, B. L. S. (1987). *Asymptotic Theory of Statistical Inference.* New York: Wiley and Sons.

[99] Press, S. J. (1972). Estimation in univariate and multivariate stable distributions. *Journal of the American Statistical Association* **67**, 842–846.

[100] Priestley, M. B. (1981). *Spectral Analysis and Time Series.* London: Academic Press.

[101] Quandt, R. E. and Ramsey, J. B. (1978). Estimating mixtures of normal distributions and switching regressions. *Journal of the American Statistical Association* **73**, 730–738.

[102] Rao, C. R. (1973). *Linear Statistical Inference and its Applications.* 2 edn, Wiley and Sons, New York.

[103] Read, R. R. (1981). Representation of certain covariance matrices with application to asymptotic efficiency. *Journal of the American Statistical Association* **76**, 148–154.

[104] Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review* **26**, 195–239.

[105] Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, Series B,* **59**, 731–758.

[106] Robertson, C. A. and Fryer, J. G. (1970). The bias and accuracy of moment estimators. *Biometrika* **57**, 57–65.

[107] Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics* **27**, 832–837.

[108] Sapatinas, T. (1995). Identifiability of mixtures of power-series distributions and related characterizations. *Annals of the Institute of Statistical Mathematics* **47**, 447–459.

[109] Sarhan, A. E., Greenberg, B. G. and Ogawa, J. (1963). Simplified estimates for the exponential distribution. *Annals of Mathematical Statistics* **34**, 102–116.

[110] Schmidt, P. (1982). An improved version of the Quandt-Ramsey MGF estimator for mixtures of normal distributions and switching regressions. *Econometrica* **50**, 501–516.

[111] Schuh, H.-J. and Tweedie, R. T. (1979). Parameter estimation using transform estimation in time-evolving models. *Mathematical Biosciences* **45**, 37–67.

[112] Scott, D. W., Tapia, R. A. and Thompson, J. R. (1977). Kernel density estimation revisited. *Nonlinear Analysis Theory Methods and Applications* **1**, 339–372.

[113] Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* New York: Wiley and Sons.

[114] Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B,* **53**, 683–690.

[115] Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* London: Chapman and Hall.

[116] Spurr, B. D. and Koutbeiy, M. A. (1991). A comparison of various methods for estimating the parameters in mixtures of von Mises distributions. *Communications in Statistics—Simulation and Computation* **20**, 725–741.

[117] Stephens, M. (1997). *Bayesian methods for mixtures of normal distributions*, PhD thesis, Department of Statistics, University of Oxford, United Kingdom.

[118] Tallis, G. M. and Light, R. (1968). The use of fractional moments for estimating the parameters of a mixed exponential distribution. *Technometrics* **10**, 161–175.

[119] Tan, W. Y. and Chang, W. C. (1972). Some comparisons of the method of moments and the method of maximum likelihood in estimating parameters of a mixture of two normal densities. *Journal of the American Statistical Association* **67**, 702–708.

[120] Thornton, J. C. and Paulson, A. S. (1977). Asymptotic distribution of characteristic function-based estimators for the stable laws. *Sankhyā, Series A,* **39**, 341–354.

[121] Titterington, D. M. (1983). Minimum distance non-parametric estimation of mixture proportions. *Journal of the Royal Statistical Society, Series B,* **45**, 37–46.

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

[123] Turner, T. R. and Griffiths, D. A. (1992). Estimation based on empirical generating functions. Unpublished manuscript.

[124] Tweedie, R. L., Zhu, Z. Y. and Choy, S. L. (1995). Parameter estimation using Laplace transforms in the M/M/1 queue. Unpublished manuscript.

[125] Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing.* London: Chapman and Hall.

[126] Watson, G. S. and Leadbetter, M. R. (1963). On the estimation of the probability density. *Annals of Mathematical Statistics* **34**, 480–491.

[127] Wise, K. N. (1989). *Statistical aspects of estimating environmental plutonium*, PhD thesis, Department of Statistics, La Trobe University, Australia.

[128] Wolfowitz, J. (1957). The minimum distance method. *Annals of Mathematical Statistics* **28**, 75–88.

[129] Woodward, W. A., Parr, W. C., Schucany, W. R. and Lindsey, H. (1984). A comparison of minimum distance and maximum likelihood estimation of a mixture proportion. *Journal of the American Statistical Association* **79**, 590–598.

[130] Yao, Q. and Morgan, B. J. T. (1999). Empirical transform estimation for indexed stochastic models. *Journal of the Royal Statistical Society, Series B,* **61**, 127–141.