Aspects of the Statistical Analysis of Data from Mixture Distributions

Athanase Polymenis

A Dissertation Submitted to the

University of Glasgow

for the degree of

Doctor of Philosophy

Department of Statistics

August 1997

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

Abstract

The purpose of this thesis is to give insight into major problems arising in the theory of mixture distributions, and, more importantly, to improve and extend some of the results that are given in the literature. We especially focus on the problem of estimating the number of components that underlie the probability distribution of a data sample. This is among the most difficult problems encountered in this area.

In principle, there are two main approaches to the problem; the theoretical approach studies the asymptotic distribution of the likelihood ratio test under the null hypothesis for testing for k_1 versus k_2 components, where $k_1 < k_2$, and the algorithmic approach uses simulations in order to overcome some of the theoretical difficulties. In this work, we use both methodologies, and we give illustrations of the methods' performances in some practical examples. We emphasise, now, the approaches that we adopt for dealing with this problem.

In a Monte-Carlo context, we propose a technique that uses an information theory criterion inside a parametric bootstrap procedure. The performance of this technique is then assessed, and comparison is made to a method using a similar type of bootstrap procedure, but where the decision criterion is based on likelihood ratio inference, and to Windham and Cutler's (1992) information theory based method. Another combined approach is also suggested.

Using a stochastic algorithmic methodology, Celeux (1987) proposes a test for the number of components, and claims that it follows a Hotelling's distribution. We argue about the reasons why this does not hold, and we derive the asymptotic distribution of this test statistic. This theoretical investigation leads us to study some problems that go beyond the scope of the mixture framework, since they are related to the theory of autoregressive processes. Some simulation results are also provided in some simple situations.

In the case where the mixing proportions are known, there is a result in the literature (Goffinet et al, 1992) that provides the asymptotic distribution of the likelihood ratio test under the null hypothesis. However, this result is not useful in some cases, corresponding to some values of the proportions. Using, then, theoretical arguments supported by simulation results, we provide this distribution in those cases.

Thus, in summary, there are three main directions in this thesis: the information based approach that mainly uses computational tools arising from recent developments in the theory of the EM algorithm; the stochastic approach that uses mathematical tools from the theory of stochastic processes; and the study of the likelihood ratio test for known proportions, whose general techniques arise from the theory of asymptotic statistics.

To the memory of my father, George Theodore,
and to my mother, Angeliki

Acknowledgements

- I would first like to express my deepest heartfelt gratitude to my supervisor, Professor D. Michael Titterington, to whom I am profoundly indebted for his unstinting help and encouragement, for providing me with excellent references, and for his kindness and patience with me throughout the whole duration of this thesis. Furthermore, I thank him again for giving me the opportunity to participate in the international workshop on statistical mixture modelling, held in France (1995), and for encouraging me to present a contribution to a discussion to the Royal Statistical Society.

- I am also indebted to the current Head of Department, Mr. Peter Breeze, as well as the former Head, Professor Ian Ford, for providing me with a very comfortable working environment.

- My gratitude goes also to the members of staff and the research students of the Department of Statistics of Glasgow University, who helped me in very many ways and contributed to making my life more comfortable. More particularly, I would like to thank: Professor Adrian Bowman for his hospitality and for his advice when I first came to Glasgow; Dr. Ben Torsney for his assistance with my subject and for his encouragement; Dr. Alexander Ivanov for constructive discussions on my subject; Dr. Alan Dunmur and Dr. Ilya Molchanov for their helpful advice concerning the typing of my text; Dr. Jim Kay for his constant encouragement and moral support; Miss Mary Nisbet and Mrs. Myra Smith for their kindness.

- I would specially like to thank my friend, Shahram Zare, for his help with the typing of my text.

- I would also like to thank Professor Bruno Goffinet, from the I.N.R.A. (France), who was very cooperative in sending me very quickly some vital information for my thesis.

- I feel also grateful to my friends, Angela Diblasi, Necla Gunduz, and Saumen Mandal, among many others, for their invaluable moral support.

- My final acknowledgement is reserved for my mother, Angeliki, to whom I owe a huge debt of gratitude for all her understanding and support without which the present thesis would not have been achieved.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Definition of the problem and generalities

In order to define mixture distributions and, subsequently, the theoretical problems pertaining to them, we first give a brief description of the general missing data context, which is as follows.

Let us consider two sample spaces, Υ and Ξ, and two sets of data, Y and X, which are respective realisations from the above sample spaces, and suppose that, instead of observing "complete data" Y, "incomplete data" X are observed. Then, we can assume that there is a mapping T from Υ onto Ξ, defined by $Y \rightarrow T(Y) = X$. If q is a parameter such that the probability density of Y is $f(y) = f(y|q)$, and that of X is $f(x) = f(x|q)$, then these two

densities are connected by the equation

$$f(x|q) = \int_{T^{-1}(X)} f(y|q)\mu(dY) \tag{1.1}$$

where $\mu(dY)$ denotes a dominating measure.

We now direct attention specifically at the mixture problem itself. For a sample size N, if we set $Y = (Y_1, ..., Y_N)$, $X = (X_1, ..., X_N)$ and $Z = (Z_1, ..., Z_N)$, the components of the complete data Y can be written in a mixture model as $Y_i = (X_i, Z_i)$ with $i = 1, ..., N$, where each $Z_i = (Z_{ik})$, for $k = 1, ..., K$, is an indicator vector such that Z_{ik} takes the value 1 if X_i belongs to the the k^{th} category and the value 0 otherwise. Then, $\Upsilon = \Xi \times \mathbf{Z}$, where $\mathbf{Z}$ is the sample space for Z. It is clear, then, that mixture models can be regarded as a particular case of the missing data model.

Now, we can write

$$f(x_i, z_i|q) = f(x_i|z_i, q)f(z_i|q),$$

so that equation (1.1) becomes

$$f(x_i|q) = \int_Z f(x_i|z_i, q)f(z_i, q)dz_i. \tag{1.2}$$

Now, Z is a finite set with K elements; if 1 is the k^{th} component of Z_i, $f(z_i|q)$ will be denoted by p_k and will be the probability that an observed variable X_i arises from a probability distribution F_k with density $f(X_i|\theta_k)$. So, q_k, which is the k^{th} component of q, will be written as (p_k, θ_k) where the p_k's are the mixing weights and the θ_k's the component parameters.

Equation (1.2) then becomes

$$f(x_i|q) = \sum_{k=1}^{K} p_k f(x_i|\theta_k). \tag{1.3}$$

For $k = 1, ..., K$, p_k is a probability, and therefore $0 < p_k < 1$ and $\sum_{k=1}^{K} p_k = 1$.

Let us now suppose that the observed data arise from a probability distribution whose density $f(x)$ has the form of equation (1.3); then, the main problem that we shall concentrate on, in the present work, will be to estimate the number of mixture components K. There has been a substantial amount of publications concerning this area. Extensive overviews of the techniques used to deal with the problem can be found in Titterington et al (1985), McLachlan and Basford (1988), and towards the end of the monograph by Everitt and Hand (1981). A more recent compendium of techniques can be found in Titterington (1997a).

Before going more deeply into the mathematical theory that is used to deal with these models, we first digress to mention briefly some of their applications.

One of the most prominent applications concerns medical diagnosis; in this case, the observed data X consist of a series of clinical tests performed on some patient, and the missing data Z indicate the disease category; it is then crucial to identify the number of underlying disease categories. Another field of interest is related to image analysis; in this case, the observed data X are the colours of a blurred image made up of a number of pixels, and the missing

data Z indicate the true colours of each pixel (Titterington, 1990). In general, then, the missing data are two-dimensional and are assumed to follow a Markov Random Field; this type of modelling occurs in remote sensing as well. A last area of interest that we describe is related to fisheries studies; for example, the observed data X might describe the length of the fish, and the missing data Z denote the underlying age category.

In another well known field of application, speech recognition, the element of interest is that the missing data follow a Markov chain.
There are many other applications concerning mixture models. The brief description that we give here is not meant to cover the very wide range of applications, but rather to give a flavour of the field, illustrate the connection between the general missing data problem and the mixture models, and emphasise how important it is to know the number of underlying categories in an unclassified set of data. A detailed account of examples of applications can be found in Titterington et al (1985).

We now return to the mathematical considerations of these problems. We shall assume in the sequel that the mixtures that we study here are identifiable, that is, distinct parameter values determine distinct members of the family. In a more formal way, a class of mixture distributions is said to be identifiable if and only if the fact that any two members of that family $\sum_{k=1}^{K} p_k F_k$ and $\sum_{k=1}^{K'} p'_k F'_k$ are equal implies that $K = K'$ and there is a permutation of the indices $(1, ..., K)$ such that $p_k = p'_k$ and $F_k = F'_k$.

An important theorem is as follows: *a necessary and sufficient condition for a class of distributions to be identifiable is that the component distributions*

be a linearly independent set over the field of real numbers, $\mathbb{R}$.

This result comes from Yakowitz and Spragins (1968). Identifiability problems have also been discussed in Teicher (1963). We shall use the identifiability notion later in this work.

1.2 The maximum-likelihood approach and the EM algorithm

In order to solve the problem of the estimation of the number of components, a most attractive way is to use the usual maximum-likelihood technique; that is, if we consider first the likelihood function of the sample X, $l(q|X) = \Pi_{i=1}^{N} f(x_i|q)$, the log-likelihood function can be written as

$$L(q) = L(q|X) = \sum_{i=1}^{N} \log f(x_i|q),$$

and the method consists of taking as estimator a value $\hat{q}$ of q that solves the likelihood equations

$$\frac{\partial L(q)}{\partial q} = 0;$$

$\hat{q}$ is a maximum likelihood estimate (MLE) of the parameter q. In general, this technique works quite well in the case of single-component distributions, and the asymptotic theory based on it is well developed. However, the situation is not so simple when dealing with mixtures and we shall describe the main difficulties in the sequel.

From the literature, the main asymptotic theorem states that, under some assumptions, that we present below as the standard regularity conditions 1, 2, 3 and 4, and N sufficiently large, there is a unique strongly consistent solution q_N of the likelihood equations, and this solution locally maximises the log-likelihood function. Furthermore, if we denote the true parameter by q_0, $\sqrt{N}(q_N - q_0)$ is asymptotically normally distributed with mean $\mathbf{0}$, and variance-covariance matrix $[E(\frac{\partial \log f(x)}{\partial q}\frac{\partial \log f(x)}{\partial q^T})]^{-1}$, which is the inverse of the Fisher information matrix. We now give these conditions:

1) q_0 is interior to the parameter space Ω, where $\Omega \subseteq \mathbb{R}^p$.

2) For $i, j, k = 1, ..., p$, the partial derivatives $\partial f/\partial q_i$, $\partial^2 f/\partial q_i \partial q_j$ and $\partial^3 f/\partial q_i \partial q_j \partial q_k$ exist and satisfy

$$|\frac{\partial f(x)}{\partial q_i}| \leq M_i(x),$$

$$|\frac{\partial^2 f(x)}{\partial q_i \partial q_j}| \leq M_{ij}(x),$$

$$|\frac{\partial^3 \log f(x)}{\partial q_i \partial q_j \partial q_k}| \leq M_{ijk}(x).$$

3) The functions $M_i(x)$ and $M_{ij}(x)$ are integrable and the function $M_{ijk}(x)$ satisfies

$$\int M_{ijk}(x) f(x|q_0) d\mu(x) < \infty.$$

4) The Fisher information matrix is well defined and positive definite at q_0. Furthermore the following relationships are verified:

$$E[\frac{\partial \log f(x)}{\partial q}] = \mathbf{0},$$

$$E[\frac{\partial \log f(x)}{\partial q}\frac{\partial log f(x)}{\partial q^T}] = -E[\frac{\partial^2 \log f(x)}{\partial q \partial q^T}].$$

These equations are a natural consequence of the previous conditions. These assumptions are very well known and can be found in most of the texts on statistical inference (for example, Zacks (1971), Cox and Hinkley (1974), Redner and Walker (1984)).

Now, the problem that one encounters when applying this method to estimating parameters of mixture densities is that, usually, the likelihood equations are non-linear, and analytic computation of $\hat{q}$ is, thus, typically impossible. Therefore, one has to resort to iterative procedures; we shall concentrate on one of the most popular, the Expectation-Maximisation (EM) algorithm. We now describe how this algorithm works, first in the missing data case, and secondly in the more particular mixture model case.

The EM algorithm starts with an initial point q^0; if we denote by q^m an estimate of the parameter q at the m^{th} iteration, then iteration $m+1$ is:

E-STEP: Determine $Q(q|q^m) = E[\log f(y|q)|X, q^m] = \int \log(f(y|q)) f(z|x, q = q^m) dz$
M-STEP: Find q^{m+1} *which maximises* $Q(q|q^m)$.
(E stands for expectation and M stands for maximisation.)

The idea that lies behind this algorithm is the following. Ideally, one wants to maximise the function $\log f(y|q)$ using a maximum likelihood technique, but since the random variable Y is partially unobserved, its expectation given the data X and the parameter q^m, $Q(q|q^m) = E[\log f(y|q)|X, q^m]$ is computed; then q^{m+1} is the parameter value which maximises $Q(q|q^m)$.

Let us now consider a finite mixture problem, where the observed data have the probability density

$$f(x|q) = \sum_{k=1}^{K} p_k f(x|\theta_k).$$

The EM algorithm then becomes (Celeux and Diebolt (1988), or Aitkin and Rubin (1985))

E-step: for $k = 1, ..., K$ *and* $i = 1, ..., N$, *compute* $t_k^m(x_i)$ *where*

$$t_k^m(x_i) = \frac{p_k^m f(x_i|\theta_k^m)}{\sum_{j=1}^{K} p_j^m f(x_i|\theta_j^m)},$$

which is the posterior probability that X_i *has been drawn from the* k^{th} *component.*

M-step: Compute

$$p_k^{m+1} = \frac{1}{N} \sum_{i=1}^{N} t_k^m(x_i),$$

and solve the equations

$$\sum_{i=1}^{N} t_k^m(x_i)\left(\frac{\partial \log f(x_i|\theta_k)}{\partial \theta_k}\right) = 0$$

for $k = 1, ..., K$, *where* N *is the sample size and* K *the number of components.*

The EM algorithm has been extensively studied by Dempster et al (1977); one essential result that they derived was that the log-likelihood is increased at each iteration, and this feature makes it very attractive for applications. The log-likelihood limit is a stationary point of the log-likelihood, and it can also be a local maximum, but it is pointed out in Wu (1983) that this may

be difficult to verify; since the choice of a starting point for the algorithm influences this convergence, Wu (1983) suggests trying different starting points. In the mixtures context, since, in general, mixture log-likelihoods are multimodal, the algorithm can converge to a point that is not global maximum. One suggestion for overcoming this problem, was by Thode et al (1987), where a set of starting points is proposed in the case of two-component univariate mixtures.

It is worthwhile to mention briefly a result in the general theoretical context by Wu (1983), namely that the EM sequence does not in general converge to one point but to a compact, connected component of either the set of stationary points or the set of local maxima in the interior of the parameter space; we do not know if these points are local maxima, but it is pointed out in Wu (1983) that this feature is not as important as the behaviour of the log-likelihood.

Another problem that arises with EM is that, sometimes, convergence is very slow. For instance, the degree of separation between the components can be a major factor in that context. We shall discuss the speed of convergence in more detail in Chapter 2.

A useful local asymptotic result, which is the version, for EM, of the main asymptotic theorem stated previously when dealing with mixtures from the exponential family (that is, where each component density is a member of an exponential family with density $f(x|\theta_k) = b(x)\exp[(q(\theta_k))^T T_k(x))]/a(\theta_k)$, where $a(\theta_k)$ is a normalising constant), is given in the following theorem (Redner and Walker, 1984).

THEOREM

Under the assumption that the true mixing proportions are strictly positive, and under condition 4 stated above, the strongly consistent solution q_N of the likelihood equations is well defined for N sufficiently large, with probability 1. Furthermore, there exists a certain norm $||.||$ on the parameter space Ω, in which the EM sequence q^m converges linearly to q_N whenever q^0 is sufficiently near q_N; that is, there exists a constant r, with $0 \leq r < 1$, for which

$$||q^{m+1} - q_N|| \leq r||q^m - q_N||,$$

whenever q^0 is sufficiently near q_N.

This result is then twofold: it suggests first that, if the starting points are close to the MLE, then, for a reasonable sample size, the algorithm will always converge towards the MLE, thereby avoiding the stationary points that we mentioned previously. On the other hand, it states that EM converges linearly, and defines its rate of convergence r.

In order to apply EM, one has to know the number of components of the mixture in question, because of the assumption that the true proportions are strictly positive. This is one of the major drawbacks of this algorithm, especially in cases which are of paramount interest in this work, that is, where the correct number of components is unknown and has to be estimated. In fact, in those cases, there has been a considerable amount of work in the literature, but still much remains to be solved. In the next section, we define the test that the asymptotic theory proposes in that regard, we emphasise the major problems, and we give a flavour of various attempts to tackle them.

1.3 Major problems that arise when testing for the number of components

We have seen in the previous section the problems that arise when we want to estimate the parameters in a mixture distribution. We shall now see that a much worse problem is posed when we try to use the asymptotic theory for assessing the number of components. Indeed, the most natural test that we use for testing for the number of components in a mixture, and which is based on parameter estimation via the likelihood equations, is the Likelihood Ratio Test (LRT). The LRT tests between the hypotheses H_0: $q \in \Omega_0 \subset \mathbb{R}^{p_0}$ and H_1: $q \in \Omega_1 \subset \mathbb{R}^p$, where Ω_0 is the parameter space corresponding to the null hypothesis H_0, Ω_1 the parameter space corresponding to the alternative H_1, and Ω_0 is a subset of Ω_1. It can be defined in the following way:

$$T(X) = \frac{\sup_{q\in\Omega_1} l(q|X)}{\sup_{q\in\Omega_0} l(q|X)}.$$

Then H_0 is rejected if $2\log T(X)$ is larger than a constant c.
The main asymptotic result concerning $T(X)$ is the following (Wilks, 1963)

THEOREM
Under H_0, and under the standard regularity conditions,

$$2\log T(X) \to \chi^2(p - p_0)$$

in distribution as $N \to \infty$.

The main problem that arises in the mixture case is that the conditions

underlying the above theorem do not hold. The reason for this is that H_0 corresponds to a boundary for H_1, and this breaks condition 1. On the other hand, as pointed out in Aitkin and Rubin (1985), under H_0 the log-likelihood is not of full rank: therefore, problems arising with mixture distributions are non-regular maximum likelihood problems (Cheng and Traylor, 1995). The problem is that, even if the mixtures are identifiable, the parameters are not. An account of these problems is also given by Ghosh and Sen (1985), who propose in a theoretical context a method based on the supremum of a normal process; these techniques are further investigated by Berdaï and Garel (1996) and by Garel (1996). For testing between a single distribution and a mixture of two, in the case that the components are known, Titterington et al (1985) prove that $T(X)$ is asymptotically distributed as $0.5\chi^2(0) + 0.5\chi^2(1)$, whereas, if the regularity conditions were valid, one should obtain a $\chi^2(1)$. This result has been extended to testing K versus $K+1$ components, by Chen and Cheng (1992). This type of asymptotics have also been derived in the context of some special mixtures by Böhning et al (1994), who use some techniques from Lindsay (1983).

We give in the sequel a brief description of some of the alternatives proposed in the literature for tackling these irregularities. This description is by no means extensive, since it is not meant to give a full overview of the techniques that have been used by very many authors, but, instead, to highlight the various difficulties that one has to grapple with in dealing with these types of problem. The techniques that are mainly used in the present work will be presented in the next section.

We first present an approach by Aitkin and Rubin (1985). In an attempt

to move the parameters away from the boundary under the null hypothesis, they proposed to place a prior distribution $h(p)$ on the mixing proportions, and considered maximising the log-likelihood function

$$l'(\theta|X) = \int l(q|X)h(p)dp.$$

In order to do that, they derived a more complicated version of EM. Then, supposing that the null hypothesis is defined by a common θ, they proposed the likelihood ratio statistic

$$T'(X) = \frac{l'(\hat{\theta}|X)}{\max_\theta \Pi_{i=1}^N f(X_i|\theta)}.$$

However, as shown by Quinn et al (1987), even after this modification, there is a break in the standard regularity conditions, so that $2\log T'(X)$ does not follow a χ^2 distribution. This problem is also very relevant to the difficulties that we are faced with in Chapter 5.

Among other suggestions, we mention those that use simulations in order to detect the number of degrees of freedom of the χ^2 distribution for $T(X)$. In an example where a single normal is tested against a mixture of two, Thode et al (1988) showed that the usual asymptotic theory holds only for very large samples, and that, if samples of moderate size are used, strict application of the standard theory will lead to overestimating the significance levels. In the same spirit, Wolfe (1971) suggested approximating $2\log T(X)$ by a χ^2 with degrees of freedom equal to twice the difference in dimensionality between the component parameters under H_1 and under H_0. Aitkin et al (1981) showed that this approximation was not correct.

Another type of technique arising from cluster analysis is Akaike's Information Criterion AIC (Bozdogan and Sclove, 1984). The problem is then the lack of theoretical justification since the conditions underlying these types of criterion are the same as those for the LRT (Titterington et al, 1985). In a simulation exercise, Windham and Cutler (1992) showed that, for a poor separation of the mixture components, AIC always overestimates the correct number.

Finally, a class of very important approaches which has been used more recently, with the advent of computers, are the Monte-Carlo based methods (Hope, 1968). Aitkin et al (1981) showed how bootstrap replications can be used to provide a test of size α, and applied this technique to reject Wolfe's suggestion. This type of approach has been brought further by McLachlan (1987), for the test of a single normal distribution versus a mixture of two normals. The validity and theoretical properties of the bootstrap likelihood are discussed by Feng and McCulloch (1996).

It follows from this discussion that the determination of the number of components in a mixture distribution constitutes a very difficult problem, which has only been partly solved. The present work deals with a variety of techniques that have been used in the statistical literature, and will consist of improving existing methodology and extending some theorems. Simulation results will be presented for all envisaged problems.

1.4 Description of chapters

Chapter 2: We present here an information ratio technique proposed by Windham and Cutler (1992), and discuss some validity problems concerning it. In order to improve this methodology, we propose a modification of their method which we believe is better supported by the theory. These methods are implemented in a Monte-Carlo computational scheme. More specifically, our modified method computes the same values as do Windham and Cutler, but uses these values inside a different decision-making criterion. Furthermore, we propose to combine our procedure with the LRT for dealing with data from one-component distributions.

Chapter 3: We present here a stochastic version of the EM algorithm, namely the SEM algorithm (Celeux and Diebolt, 1988). Formal theory serves to back up an asymptotic theorem, and we compare it with the theory that lies behind EM. Following Celeux (1987), we give a test statistic for the number of components that is based on random variables following an autoregressive process. However, there are serious theoretical problems when we consider it to be a standard Hotelling's test, because the random variables that underlie it are not mutually independent.

Chapter 4: We derive here a result for the asymptotic distribution of the statistic presented in the previous chapter. Then, using simulations, we provide some numerical results in order to assess the performance of this statistic. We also construct a new statistic based on the residuals of the autoregressive process in hand, which seems to yield similar kind of results as the statistic under consideration.

Chapter 5: In this part we consider a theorem by Goffinet et al (1992), who derive the limiting distribution of the LRT statistic for testing between a single component and a mixture of two, in the case where the mixing proportions are known. The limiting distribution depends discontinuously on the proportion parameter, and huge sample sizes are needed, near the discontinuity, in order to obtain the asymptotic results predicted by the theorem. We give some heuristic insight to that intricate area, and we compute characteristics of the LRT under the null hypothesis. In doing that, we use a technique for the detection of the zeroes of the LRT which is more formal than the ad-hoc approach of Goffinet et al (1992). Thus, we find the distribution of the LRT statistic in those cases, for reasonable sample sizes.

Chapter 6: This part is mainly twofold: on the one hand, it describes the conclusions of our work, and, on the other hand, gives some directions for further investigation in the area of the estimation of the number of components in a mixture.

Chapter 2

On the Determination of the Number of Components using Information Ratio Techniques

2.1 Information ratios and the EM rate of convergence

2.1.1 The EM algorithm and its connection to information ratios

We have defined "complete" and "incomplete" data in the introductory part of the present work, as preliminaries to the presentation of the mixture problem. In this chapter, we analyse further these ideas, as they will be used to derive some criteria for the identification of the number of components.

For that purpose, let us now define I_X to be the observed-data observed information matrix evaluated at the MLE, I_Y the conditional expectation of the complete-data observed information matrix evaluated at the MLE, and $I_{Y|X}$ the information matrix for the conditional density $f(z|x,q)$ evaluated at the MLE (equation 2.1 below will show that $I_{Y|X}$ characterises the loss of information in observing X instead of Y).

If we denote by $\hat{q}$ the maximum likelihood estimator of the parameters involved in our problem, then

$$I_Y = [E(-\frac{\partial^2 \log f(y|q)}{\partial q \partial q^T}|X,q)]_{q=\hat{q}} = [\int -\frac{\partial^2 \log f(y|q)}{\partial q \partial q^T} f(z|x,q)dz]_{q=\hat{q}}.$$

$$I_X = [-\frac{\partial^2 \log f(x|q)}{\partial q \partial q^T}]_{q=\hat{q}}.$$

$$I_{Y|X} = [E(-\frac{\partial^2 \log f(z|x,q)}{\partial q \partial q^T}|X,q)]_{q=\hat{q}} = [\int -\frac{\partial^2 \log f(z|x,q)}{\partial q \partial q^T} f(z|x,q)dz]_{q=\hat{q}}.$$

Then, these three matrices are related by the following equation, called also "missing information principle", by Orchard and Woodbury (1972).

$$I_X = I_Y - I_{Y|X}. \tag{2.1}$$

This result is easy to obtain; indeed we can write

$$f(y|q) = f(x|q)f(z|x,q),$$

so that

$$L(q) = \log f(x|q) = \log f(y|q) - \log f(z|x,q);$$

taking then the second derivatives of this expression with respect to q, averaging over $f(z|x,q)$, multiplying it by (-1), and evaluating it at $q = \hat{q}$ yields the result. This result has also been given in Sundberg (1974).

In order to obtain $\hat{q}$ we use the Expectation-Maximisation (EM) algorithm, that we described in Chapter 1. This algorithm defines then a mapping M: $q \mapsto M(q)$, from the parameter space Ω to itself.

Let DM be the Jacobian matrix of M. Then a useful formula connecting DM with the information matrices is the following (Dempster et al, 1977):

$$DM(\hat{q}) = I_{Y|X} I_Y^{-1}. \tag{2.2}$$

Since this result is essential for the theory presented in this chapter, we give an idea of the proof. This is as follows. Using a Taylor expansion, about the MLE $\hat{q}$, of the function $\frac{\partial}{\partial q'} E[\log f(y|q')|X,q]$, and substituting $q = q^{(m)}$ and $q' = q^{(m+1)}$, we obtain in the limit (i.e as $m \to \infty$)

$$DM(\hat{q}) I_Y + [\frac{\partial}{\partial q}(\frac{\partial}{\partial q'} Q(q'|q))_{q'=\hat{q}}]_{q=\hat{q}} = [0],$$

where [0] denotes the matrix with elements all equal to 0, and $Q(q'|q) = E[\log f(y|q')|X,q]$. On the other hand, from the expression $\log f(y|q') = L(q') + \log f(z|x,q')$, we obtain

$$E[\log f(y|q')|X,q] = L(q') + E[\log f(z|x,q')|X,q]$$

so that, taking first derivatives with respect to q, and evaluating this expression at $q = \hat{q}$, we obtain

$$[\frac{\partial}{\partial q} Q(q'|q)]_{q=\hat{q}} = [\frac{\partial}{\partial q} E(\log(f(z|x,q')|X,q)]_{q=\hat{q}}$$

and therefore

$$DM(\hat{q})I_Y + [\frac{\partial}{\partial q'}[\frac{\partial}{\partial q} E(\log(f(z|x,q')|X,q)]_{q=\hat{q}}]_{q'=\hat{q}} = [0].$$

The conclusion comes using Lemma 2 of Dempster et al (1977) (the second term of the left-hand side of the above equality equals $-I_{Y|X}$).

$DM(\hat{q})$ is also called matrix of fractions of missing information, since $I_{Y|X}$ measures the loss of information due to missing data (by the Missing Information Principle), and I_Y measures the information for the complete data. Combining equations 2.1 and 2.2, we obtain

$$DM(\hat{q}) = Id - I_X I_Y^{-1}, \tag{2.3}$$

where Id denotes the identity matrix.

2.1.2 Some theoretical results concerning the minimum information ratio (MIR)

Let us suppose that we consider the general case of missing data, whose distribution belongs to the exponential family. Then a root $\hat{q}$ is called point of attraction for the iterative process $q^{m+1} = M(q^m)$ if, for a starting point q^0 for

that process, which is in the neighbourhood of $\hat{q}$, the above process converges to $\hat{q}$, and $\hat{q}$ satisfies the equation $\hat{q} = M(\hat{q})$ (Ostrowski, 1960); then the main criterion, as defined in Ostrowski (1960) and Sundberg (1976), is the following.

For a root $\hat{q}$ to be a point of attraction, it is necessary that the absolute value of the largest eigenvalue of DM at $q = \hat{q}$, that we call r, is less than or equal to 1 and sufficient that it is strictly less than 1; this value is the factor (rate) of convergence.

These results are of general interest, since they can be applied to any iterative algorithm for the general missing data case.

More particularly, in the case of the EM algorithm, the same ideas are also expressed in the papers by Dempster et al (1977), and by Meng and Rubin (1994), again for the missing data situation: indeed, denoting by $\hat{q}$ the maximum likelihood estimator (MLE) of the parameters, it is stated that the eigenvalues of $DM(\hat{q})$ all lie in $[0, 1[$, if I_X is positive definite, and this is a sufficient condition for $\hat{q}$ to be a local maximum likelihood estimate. Then, in the case that the eigenvalues of $DM(\hat{q})$ are all less than 1, the largest such eigenvalue gives the rate of convergence of the algorithm.

Now, since I_X is the Fisher information matrix for the unclassified (incomplete) sample, and I_Y the Fisher information matrix for the classified (complete) sample, it results that the information ratio matrix $I_X I_Y^{-1}$ measures the proportion of information about q from the unclassified sample. Hence, Windham and Cutler (1992) use the term *minimum information ratio*, or MIR, for the smallest eigenvalue of $I_X I_Y^{-1}$. From 2.3, we see that the MIR

can also be considered as 1 minus the largest eigenvalue of $DM(\hat{q})$.

Hence the MIR can be written as 1 minus the rate of convergence of EM, which can also be defined as the speed of convergence of EM (Meng 1994), and takes values between 0 and 1. Moreover, a key remark in Dempster et al (1977) is that, if the information loss due to incompleteness is small, then the algorithm converges rapidly. This idea appears also in Redner and Walker (1984), for mixtures of densities from exponential families, as a consequence of their main theorem concerning EM, which has been stated in the introductory part of this work. The conclusion they come to is that, if the mixture components are well separated, then EM converges rapidly, whereas, in cases where the components are poorly separated, the convergence of EM is slow. In view of these results, one should expect that large values of MIR suggest a good clustering of the data, whereas small values suggest a poor clustering. On the other hand, if we had observed the complete sample, we would have obtained in theory that $I_X = I_Y$, and thus $DM(\hat{q}) = [0]$, so that the largest eigenvalue of $DM(\hat{q})$ would have been 0, implying that MIR=1; this argument suggests that big values of MIR point towards the right number of components.

On the basis of this analysis, the method suggested in Windham and Cutler's paper is to fit to the data mixture models with different numbers of components, the smallest number being 2, and choose the model that provides the largest MIR; this is called the *estimation step*.

The implementation of the estimation step of MIR in Windham and Cutler's paper is as follows.

a) Choose k_1 *and* k_2 *with* $2 \leq k_1 \leq k_2$.

b) For each k, with $k_1 \leq k \leq k_2$, obtain the MIR(k), assuming the mixture has k components.
c) Estimate the number of components to be $\hat{k}$, the value of k for which the MIR(k) is largest.

Remark: Windham and Cutler choose $k_1 = 2$ and $k_2 = 5$, in applying their method. We will use the same values in the present work, for the sake of comparison.

In the sequel, we will call this method *basic MIR*, in order to differentiate it from the other approaches that we introduce and which also include computation of the MIR values.

In addition to the estimation step, Windham and Cutler propose, as a second step in their methodology, to define a "confidence measure" of the reliability of the estimation step, called $\hat{p}$; to do that, the following *validation step* is applied.

a) Obtain m bootstrap samples from the original data.
b) Repeat the estimation procedure for each one of these samples, and derive the bootstrapped $\hat{\hat{k}}_1,...,\hat{\hat{k}}_m$.
c) Calculate the probability $\hat{p}$ that the maximum MIR occurred at the estimate of the number of components $\hat{k}$, given by the estimation step, using the formula $\hat{p}$=(number of times $\hat{\hat{k}}_j = \hat{k}$)/m, for $j = 1, ..., m$.

However, as is stressed at the conclusion of their paper, there is an open problem concerning the theoretical validity of this measure.

This whole method, including estimation step and validation step, is called the *minimum information ratio estimation and validation* ($MIREV$) procedure.

2.1.3 Applying the information ratio to accelerate EM

Louis (1982) derives equations (2.1) and (2.3) in a slightly different context. Equation (2.1) allows us to compute the observed information using the EM algorithm. Using his methodology in a practical example where the data arise from a mixture of two univariate normal distributions, Louis calculates complete and observed Fisher information matrices, in the EM, along with the EM convergence rate. This rate can also be used in some applications, and we briefly report a few, in order to emphasise the importance of the area. The first example comes from Louis (1982), where, from equation (2.3), he uses $DM(\hat{q})$ to speed up the EM algorithm, in the following procedure:

$$q_{acc} = q^m + (1 - DM(\hat{q}))^{-1}(q^{m+1} - q^m),$$

where q^m is the parameter estimate at iteration m of EM, q_{acc} the accelerated estimate, and $DM(\hat{q})$ is an estimate of $DM(q^m)$; this procedure is a special case of *Aitken's acceleration method.* A second slightly different example comes from Böhning et al (1994), where they use the same type of procedure in order to find a "good" stopping rule for the EM algorithm, as a by-product. Some other examples of procedures that accelerate EM are the methods proposed in Peters and Walker (1978); applied to mixture models, these methods can be written in the following way.

If q^m is the parameter estimate at iteration m of EM, then the accelerated estimate takes the form

$$q_{acc} = (1 - \epsilon)q^m + \epsilon M(q^m).$$

It has been shown in Peters and Walker (1978) that, with probability 1, as the sample size tends to infinity, the above iterative procedure converges locally to the strongly consistent maximum-likelihood estimate whenever the parameter ϵ is strictly between 0 and 2. For well separated mixtures, the optimal ϵ is close to 1, and rapid local convergence is expected asymptotically, whereas for poorly separated mixtures the optimal ϵ is close to 2, and slow convergence has to be expected. In drawing these conclusions, the role of $DM(\hat{q})$ is crucial: indeed, as for Redner and Walker's theorem, it is the largest eigenvalue of $E[DM(q_0)]$, where q_0 is the true parameter value to be estimated, that will be used to determine the rate of convergence of the algorithm.

2.1.4 A measure for the global rate of convergence

From a practical point of view, one uses the relationship MIR=1-rate of convergence of EM, to compute numerical values of MIR, since, by definition (Meng and Rubin (1991), Windham and Cutler (1992), Meng (1994)), if q^m is the parameter estimate given at the m^{th} iteration of the EM algorithm, then the formal definition of r is

$$r = \lim_{m \to \infty} \frac{||q^{m+1} - \hat{q}||}{||q^m - \hat{q}||},$$

provided this limit exists. Now, for practical reasons, the formula used to compute this ratio is

$$r = \lim_{m \to \infty} \frac{||q^{m+1} - q^m||}{||q^m - q^{m-1}||},$$

and the norm $||.||$ used by Windham and Cutler has not been clearly specified; they just stated that it is "any convenient norm on Euclidian space". Thus, here the problem is how to choose the 'global' norm, $||.||$, that we are going to use in order to compute the global rates of convergence. It is obvious then that we should somehow use the component-wise rates to compute the global rate; for that purpose, we give the formal definition of the component-wise rate, as follows.

The j^{th} component-wise rate of convergence is defined as

$$r_j = \lim_{m \to \infty} \frac{|q_j^{m+1} - \hat{q}_j|}{|q_j^m - \hat{q}_j|},$$

provided this limit exists. $|.|$ denotes the classical absolute value.

These definitions are valid for the general missing data problem. Certainly, in our case the global rate will be found for the more special mixture model case. Again, for practical reasons, the formula used to compute the component-wise rates in the mixture problem that we are dealing with, is the following:

$$r_j = \lim_{m \to \infty} \frac{|q_j^{m+1} - q_j^m|}{|q_j^m - q_j^{m-1}|},$$

where q_j^m denotes the value of the j^{th} parameter at iteration m. All the above definitions concerning global and component-wise rates of convergence are stated in Meng (1994).

In order to give a satisfactory answer to this problem, we start by mentioning first the following remark from Dempster et al (1977): "the fraction of information may vary across different components of q, suggesting that certain components of q may approach $\hat{q}$ rapidly using EM, while other components may require many iterations". An example, from Little and Rubin (1987), where this situation may occur is where the sample arises from a univariate contaminated model of the form $f(x|\mu,\sigma^2) = (1-\pi)N(x : \mu,\sigma^2) + \pi N(x : \mu,\sigma^2/\lambda)$, where $0 < \pi < 1$, $\lambda > 0$, and both π and λ are known. The problem is to compute the maximum likelihood estimator $\hat{q}$ of $q = (\mu,\sigma^2)$. In order to do that, the EM algorithm is implemented in Little and Rubin (1987). The same example is considered by Meng and Rubin (1994), and the component-wise rates of convergence of EM are computed; the matrix of fractions of missing information $DM(\hat{q})$ is found to be asymptotically diagonal, where in general the diagonal elements representing these rates are different. However, it is also stated in Meng and Rubin (1994) that this is most unlikely to happen: "in most practical situations, all components converge at the global rate, which equals the largest eigenvalue of the matrix of fractions of missing information" -unless the matrix $DM(\hat{q})$ has the special form of Little and Rubin's example. In general, $DM(\hat{q})$ is not diagonal, and an example of that is found in Louis (1982), and has been mentioned previously: to be specific, a mixture of two normal distributions is considered, where the parameters are the means and the mixing weights (the variances are considered equal and

known); using EM, Louis computes $DM(\hat{q})$, which is found not to be diagonal.

In the same context, it is worthwhile mentioning briefly a more recent technique introduced by Meng and Rubin (1991), and called the *Supplemented EM algorithm* (to differentiate from the Stochastic EM algorithm that we use in other chapters) where the authors take these ideas a step further. Indeed, in order to compute numerically the asymptotic variance matrix of the parameter in question, one has to know the values of the elements of the fractions of the missing information matrix, $DM(\hat{q})$; this algorithm computes these elements iteratively using the EM algorithm.

In this chapter, the examples that we will deal with are completely different from Little and Rubin's example, so that, from what was said before, the component-wise rates of convergence are the same. On the other hand, as a consequence of a proposition in Meng and Rubin (1994), it is stated that "the global rate of convergence should be equal to the component-wise rate of convergence of the slowest component(s), since the whole algorithm converges if and only if all components converge". Now, in our case, since the parameter that we need to estimate is a vector, and since all components converge at the same rate, the global rate is calculated as the average of the component-wise rates; that is

$$r = \frac{1}{s+1} \sum_{j=1}^{s+1} r_j$$

where s is the dimension of the parameter space.

2.1.5 Performance of the basic MIR procedure

The MIR seems to be fallible; indeed, Windham and Cutler mention a case where the MIR validates a two-component model whereas the data arise from a four-component model (they compute exact values of MIR, using numerical integration). In the afore-mentioned case, the validation step gives the four-component model as the second best choice, since 60 per cent of the bootstrap samples indicated two components and 32 per cent indicated four components. In order to cope with these problems, we will try and improve this methodology.

The numerical experiments that we use in the sequel, in order to illustrate the estimation step in their method, as well as the subsequent modifications that are derived, are based on the same example as that used by Windham and Cutler. In the example, samples of size 100 are drawn from equally weighted mixtures of 3 bivariate circular normal distributions. The means are 4 units apart, forming an equilateral triangle. The parameter q consists of the means and the mixing weights; the standard deviation σ associated with the circular component densities is assumed known, and experiments are carried out for 4 different values of σ, namely for σ=1.5, 1.33, 1.0 and 0.67. For each case, 100 replications are carried out. The tables that we obtain, using the basic MIR and its various modifications, give the number of times a particular value of k was chosen.

Remark: the identification of the number of components will depend heavily on the value of σ, as we will see in what follows. Indeed, the more σ is reduced, the more the components are well separated so that, in decreasing the value of σ, one increases the number of times that the right number of

components is detected.

Figures 2.1, 2.2 and 2.3 produced at the end of this chapter, represent data corresponding respectively to the cases $\sigma = 1.5$, 0.67, 1. The difference between first and second cases is striking; indeed, in the first case it is impossible to make out any separation into classes, whereas three clusters clearly appear in the second case. In between those two extreme degrees of separation, we have the intermediate degrees $\sigma = 1.33$ and $\sigma = 1$; even in the latter case, where the spread is small, it is not very obvious to identify, from Figure 2.3, a three-component clustering.

Table 2.1 measures the performance of Windham and Cutler's basic MIR. Note that the results given here are somewhat different from those presented in their paper; this is because we had to reproduce their experiment, since the same datasets as those used in Table 2.1 will be considered for the different modifications that we present.

	$k=2$	$k=3$	$k=4$	$k=5$
$\sigma = 1.50$	55	42	3	0
$\sigma = 1.33$	48	47	5	0
$\sigma = 1.00$	38	62	0	0
$\sigma = 0.67$	25	75	0	0

Table 2.1. Frequencies of Identification of the Number of Components using basic MIR.

Two remarks are stimulated by Table 2.1:

1) As predicted, decreasing σ from 1.5 to 0.67 increases the frequencies of detection of the true underlying model ($k = 3$) from 42 to 75. Moreover, for the poorly separated cases, more than half of the times the method does not

detect the right model, and even for the well separated cases the frequencies of correct model detection are not as good as might be hoped, since for instance, in the case $\sigma = 0.67$, Figure 2.2 shows clearly 3 clusters, but, nevertheless, one time out of four, the method fails to detect the right model.
2) In general, the MIR does not overestimate the number of components. From Table 2.1 one can see that in the well separated cases the method never validates a model with more than 3 components, and in the poorly separated cases overestimation occurs only a very few times; this behaviour of MIR is also reported by Windham and Cutler. The reason for this behaviour, as stated towards the end of their paper, is that, as soon as a mixture with too many components is fitted, the observed Fisher information matrix I_X gets close to singular, so that the MIR gets close to 0.

Finally, it is also worthwhile mentioning that, in their experiment, Windham and Cutler (1992) find this method to be more reliable than the AIC and the Partition Coefficient (Bezdek, 1981).

2.2 On bootstrapping the MIR: the Modified MIR procedure

2.2.1 Main theory and simulation results

We obtain the same result as the one given by remark 2, by putting the question in a more general way: when the model is overfitted, the setup obtained is not identifiable any more. We give an example of this situation.
Let us suppose that the data arise from a mixture with probability density

$$f(x|q_1) = \sum_{j=1}^{K} \pi_j f(x|\phi_j),$$

and we fit to that data the mixture model with probability density

$$g(x|q_2) = \sum_{i=1}^{K+1} p_i f(x|\theta_i).$$

Suppose that the ϕ's and θ's are scalars and are ordered according to increasing indices, that is, $\phi_1 < \phi_2 < ... < \phi_K$ and $\theta_1 \leq \theta_2 \leq ... \leq \theta_{K+1}$.
Clearly, if in model g we set $p_1 = \pi_1,...,p_{K-1} = \pi_{K-1}$, $\theta_1 = \phi_1,...,\theta_{K-1} = \phi_{K-1}$, $\theta_K = \theta_{K+1} = \phi_K$, then density g is equivalent to density f for any p_K, p_{K+1} such that

$$p_K + p_{K+1} = \pi_K.$$

This means that there is an infinite number of representations of the true mixture density in terms of q_1; this establishes the non-identifiability.
This result implies that, when the model is overfitted, the observed Fisher information matrix I_X is singular in theory (Silvey, 1975, pp. 81-82), and therefore the MIR is in theory 0. Based on this argument, our idea is therefore to consider the smallest value of k, for which the corresponding eigenvalue of I_X is theoretically 0 and to select $k-1$ as the true number of components. This choice will correspond to an observed eigenvalue which is close to 0, so that one should observe a sudden drop in the value of MIR when the true number of components is increased by 1; the problem will certainly be, then, how to quantify at what point this happens. In order to account for the drop in the MIR value, we first adopted the following approach.

Suppose that $a \in \mathbb{R}$ is small, and $k_1 \leq k \leq k_2 - 1$. Then, if $\frac{MIR(k+1)}{MIR(k)} < a$, choose k to be the right number, otherwise choose $\arg\max_k(MIR(k))$. In our

case, this becomes

a) If $\frac{MIR(3)}{MIR(2)} < a$, *choose the two-component model and stop.*
b) If $\frac{MIR(4)}{MIR(3)} < a$, *choose the three-component model and stop.*
c) If $\frac{MIR(5)}{MIR(4)} < a$, *choose the four-component model and stop.*
d) Otherwise, go to the estimation step of the basic MIR algorithm.

The problem now is how to choose the value of a. In order to assess the influence of a, we performed the experiment on 100 replicates as before, for four different values of a, namely, 0.1, 0.2, 0.3 and 0.4. Table 2.2 gives the results.

	a	$k=2$	$k=3$	$k=4$	$k=5$
$\sigma = 1.50$	0.1	38	47	15	0
	0.2	26	58	16	0
	0.3	16	61	23	0
	0.4	15	63	22	0
$\sigma = 1.33$	0.1	35	43	22	0
	0.2	18	58	24	0
	0.3	11	66	23	0
	0.4	9	69	22	0
$\sigma = 1.00$	0.1	14	69	17	0
	0.2	8	75	17	0
	0.3	4	82	14	0
	0.4	2	89	9	0
$\sigma = 0.67$	0.1	10	72	18	0
	0.2	5	77	18	0
	0.3	3	86	11	0
	0.4	0	91	9	0

Table 2.2. Frequencies of Identification of the Number of Components chosen for Different Values of a.

Since $\frac{MIR(3)}{MIR(2)}$ is not expected in theory to be small, using this methodology, one should be able to decrease the frequency of selecting a two-component model. Indeed, comparison of Tables 2.1 and 2.2 shows that this is the case for all four values of σ and a. On the other hand, the more a is increased, the more the three-component detection takes on, thus automatically decreasing the two-component validation. Certainly, three-component model detection increases as a increases, so that there exists a minimum value of a, $\min(a)$, such that, for any $a \geq \min(a)$, this methodology performs better than the basic MIR. The problem now is that, because this is an ad-hoc methodology, it is not straightforward to see which value of a to select for each of the σ values. However, the main idea expressed in Table 2.2 will be at the centre of the more formalised approach that we now describe.

2.2.2 Implementing the Modified MIR: methodology and simulation results

The procedure that we will use to identify the number of components is based on two concepts, the first being the idea developed in the previous section, that is, to consider ratios of the form $\frac{MIR(k+1)}{MIR(k)}$, and the second being the parametric bootstrap (a general reference about the bootstrap can be found in Efron and Tibshirani (1993)). Then, our modification of the basic MIR method is a Monte-Carlo approach, with the following general computational scheme:

a) For $2 \leq k \leq 4$, *estimate the parameters by* $\hat{q}_k$, *compute* $MIR(k)$ *as in the estimation step of basic MIR and evaluate* $a_k = \frac{MIR(k+1)}{MIR(k)}$.

b) For $k = 2$, generate 99 bootstrap samples from the k-component model with parameters $\hat{q}_k$, and compute $\tilde{a}_k$ for each one of them.
c) If a_k is "atypically large" as compared to the bootstrapped $\tilde{a}_k$'s, that is, a_k is larger than 94 values of $\tilde{a}_k$, increase k by 1 and repeat steps b and c (the maximum value of k is 4); otherwise choose the present k as the solution and stop.

We call this method the *Modified MIR* procedure (Polymenis, 1997), and we present some simulation results, in order to compare it to the basic MIR of Windham and Cutler. Table 2.3 gives the results of Modified MIR, based on 100 replications of data from the three-component distribution used previously.

From Table 2.3, one can see that there is a distinct improvement in using the Modified MIR. Indeed, for every value of σ, the frequency of detection of the true underlying model by Modified MIR is larger than that of basic MIR. For a large spread, corresponding to poor separation of the data ($\sigma = 1.5$), more than half of the times one obtains the right detection, and there is a fairly quick increase in the frequency of correct detection, as σ decreases (as σ decreases from 1.5 to 1.33, the correct detection frequency increases from 54 to 67). In the case of basic MIR, this increase is slower (from 42 to 47). On the other hand, for the smaller spreads (corresponding to better separation of the components), one can be confident that Modified MIR will almost always detect the right model. This is very encouraging, especially when one considers the case $\sigma = 1$ where, as noticed previously, the separation into clusters is not clear-cut.

	$k=2$	$k=3$	$k=4$	$k=5$
$\sigma = 1.50$	45	54	1	0
$\sigma = 1.33$	29	67	4	0
$\sigma = 1.00$	3	94	3	0
$\sigma = 0.67$	0	94	6	0

Table 2.3. Frequencies of Identification of the Number of Components using Modified MIR.

In summary, the Modified MIR procedure improves on basic MIR. This technique provides a new approach for estimating the number of components in a mixture, based on matrices of information ratios (see also Titterington (1997b)). Furthermore, it has another nice feature, in that it is easy to implement for multivariate data.

We have, until now, applied this procedure to the example used by Windham and Cutler, for the purpose of comparing their method to ours; in that example, the parameters were the component means and the mixing weights; we can certainly apply this method, considering also the component variances as parameters to estimate. In order to apply this concept in practice, we implemented the Modified MIR procedure on the following example from Marron and Wand (1992).

The data arise from the following well separated univariate mixture distribution

$$f(x) = 0.5N(-1.5, (0.5)^2) + 0.5N(1.5, (0.5)^2),$$

where the parameters are the mixing proportions, the component means and the component variances. Then, generating 50 replications of data with sample sizes 100 and 250, the Modified MIR procedure detects the correct model

in 48 of these replicates, for both cases. Thus, we see that the method performs very well in this example.

For the rest of this chapter, the example used to illustrate the different methods is the three-mixture model used in Windham and Cutler (1992).

2.2.3 Some different approaches

Since this procedure is based on the idea that the value of MIR slumps suddenly when the true number of components is increased by 1, one could wonder what should happen if, instead of considering the ratio a_k, corresponding to a relative decrease in the MIR value, one considers its absolute value, that is, taking a_k to be $MIR(k+1)$. Then a procedure similar to the Modified MIR yields the results of Table 2.4. Comparison with Table 2.3 shows superiority of the Modified MIR, especially in the poorly separated component cases corresponding to $\sigma = 1.5$ and 1.33; in the case $\sigma = 1.5$, the performance of this method is even worse than that of basic MIR, since on a clear majority of occasions a two-component model is selected. Thus, the method does not work well, and is inappropriate. However, one should note that, conversely, for the two other values of σ, corresponding to much better separated components, the method performs as well as Modified MIR.

Until now, we have used these techniques for cases where the underlying distribution of the data is a true mixture, in the spirit of Windham and Cutler's paper. Suppose now that we want to include the possibility that the data could arise from a single-component distribution. Then, from Table 2.4, it

	$k=2$	$k=3$	$k=4$	$k=5$
$\sigma = 1.50$	65	32	2	1
$\sigma = 1.33$	47	47	3	3
$\sigma = 1.00$	1	93	5	1
$\sigma = 0.67$	0	95	2	3

Table 2.4. Frequencies of Identification of the Number of Components using the Absolute Decrease Criterion.

seems that these techniques do not work well in that case. Indeed, considering the absolute decrease criterion and incorporating in it the case of a single-component model, we obtain for the poorly separated cases (which are the cases of interest in a test for a single distribution), an overwhelming amount of single-component validations, and very few three-mixture detections: these results are reported in Table 2.5. We conclude that this method is inappropriate. Another approach that combined estimation of the single-component distributions as before with estimation of the true mixtures by Modified MIR led to approximately the same results as those of Table 2.5.

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
$\sigma = 1.50$	85	12	3	0	0
$\sigma = 1.33$	64	13	23	0	0
$\sigma = 1.00$	13	2	82	3	0
$\sigma = 0.67$	0	0	98	2	0

Table 2.5. Frequencies of Identification of the Number of Components using the Absolute Decrease Criterion, when allowing for the Possibility that the Underlying Distribution is not a Mixture.

2.3 Combining the bootstrap likelihood ratio with the Modified MIR

2.3.1 On bootstrapping the likelihood ratio: methodology and simulation results

Following the ideas of McLachlan (1987) and Feng and McCulloch (1996), we use here a Monte-Carlo procedure similar to that of the Modified MIR, but in the log-likelihood context, including the identification of single distributions. Denoting the likelihood ratio statistic by $T_k^{k+1}(q|X) = 2[L(\hat{q}^{k+1}|X) - L(\hat{q}^k|X)]$, where $\hat{q}^k$ is the parameter estimate derived for the k-component fitted model, we present now this procedure:

a) For $1 \leq k \leq 4$, *evaluate the parameters* $\hat{q}^k$, *and compute* $T_k^{k+1}(\hat{q}^k|X) = 2[L(\hat{q}^{k+1}|X) - L(\hat{q}^k|X)]$.
b) For $k = 1$, *generate* 99 *bootstrap samples from the* k*-component model with parameters* $\hat{q}^k$, *and compute* $\tilde{T}_k^{k+1}(\tilde{q}^k|\tilde{X})$ *for each one of them, where* $\tilde{X}$ *and* $\tilde{q}^k$ *stand, respectively, for a bootstrap sample and a bootstrap parameter estimate under the hypothesis of a* k*-component model.*
c) If $T_k^{k+1}(\hat{q}^k|X)$ *is "atypically large" as compared to the* 99 *bootstrap values of* $\tilde{T}_k^{k+1}(\tilde{q}^k|\tilde{X})$, *that is, the former is larger than* 94 *values of the latter, increase* k *by* 1 *and repeat steps b and c (the maximum value of* k *is* 4*); otherwise choose the present* k *as the solution and stop.*

We call this method the *Bootstrap Likelihood Ratio* (*BLR*) procedure. We applied the BLR to the three-component model used for basic and Modified MIR, and the results are presented in Table 2.6. The method gives excellent

results, in the sense that it detects the right number of components almost always, and for all degrees of separation corresponding to the four values of σ.

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
$\sigma = 1.50$	0	0	95	5	0
$\sigma = 1.33$	0	0	96	4	0
$\sigma = 1.00$	0	0	99	0	1
$\sigma = 0.67$	0	0	95	5	0

Table 2.6. Frequencies of Identification of the Number of Components using the BLR Procedure.

2.3.2 Using the BLR inside the Modified MIR procedure

Since the techniques based on MIR can only be applied when the data arise from a true mixture, and, in view of the results of Table 2.6, one might use, as a first step, the bootstrap likelihood to identify the single distributions, and as a second step, the Modified MIR to identify the mixtures, all in the same procedure. Thus, if the underlying distribution is a three-component mixture, one should expect similar results to Table 2.3. Furthermore, this good behaviour of the bootstrap likelihood has to be the same for any underlying distribution of the data, so that, if the data arise from a single normal distribution, this should be detected before the Modified MIR is used. Indeed, Table 2.7 presents results in an example when the data arise from a single normal distribution; k-component models are fitted to that data, where $1 \leq k \leq 5$. The results show again excellent performance of the method. Thus, we use this combination of bootstrap likelihood and Modified MIR, for Windham and Cutler's (1992) example where the underlying distribution

was a three-component mixture, and the results are shown in Table 2.8; as predicted, these results are very close to those given in Table 2.3.

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
$\sigma = 1.50$	97	3	0	0	0
$\sigma = 1.33$	97	3	0	0	0
$\sigma = 1.00$	95	5	0	0	0
$\sigma = 0.67$	95	5	0	0	0

Table 2.7. Frequencies of Identification of the Number of Components using the BLR Procedure, when the Real Distribution is a Normal.

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
$\sigma = 1.50$	0	47	51	2	0
$\sigma = 1.33$	0	37	62	1	0
$\sigma = 1.00$	0	1	97	2	0
$\sigma = 0.67$	0	0	93	7	0

Table 2.8. Frequencies of Identification of the Number of Components using the Combined BLR and Modified MIR Procedure.

Thus, in the case $k \geq 2$, BLR must be preferred to the Modified MIR, especially for small sample sizes. However, for larger sizes, it would seem that the performance of these methods is very similar, since simulation results show that in the case where the sample size equals 300, MIR detects the three-component distribution 97 per cent of the times for $\sigma = 1.50$, and 98 per cent of the times for $\sigma = 1.33$. In the case that the possibility $k = 1$ is included, again BLR does better than the combined BLR and MIR for small sample sizes, but for larger sizes these two methods should be very similar.

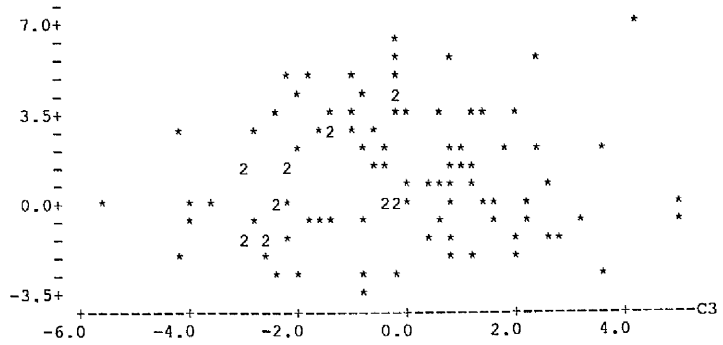

Figure 2.1. Data for $\sigma = 1.5$.

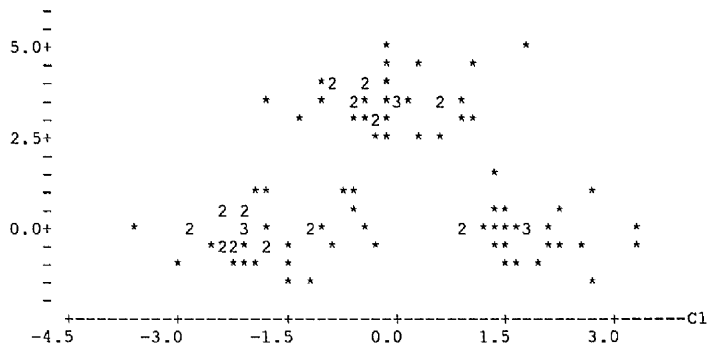

Figure 2.2. Data for $\sigma = 0.67$.

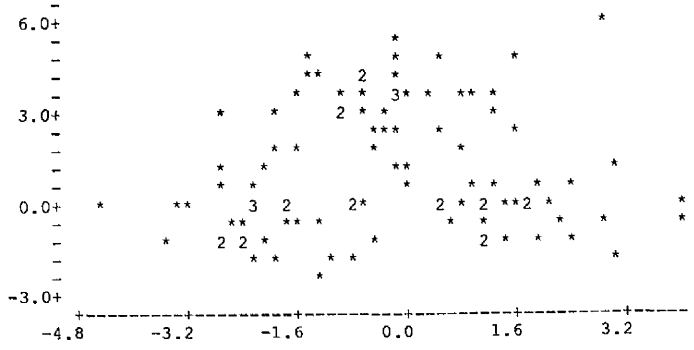

Figure 2.3. Data for $\sigma = 1$.

Chapter 3

On a Test for the Number of Components based on the Stochastic EM Algorithm

3.1 The stochastic EM algorithm

3.1.1 Generalities

We study, in this chapter, a stochastic version of the EM algorithm, called *Stochastic EM* or *SEM* (Celeux and Diebolt (1985), Celeux and Diebolt (1986a), Celeux and Diebolt (1988), Diebolt and Ip (1996)). We also describe some technical assumptions underlying an asymptotic theorem, and provide theoretical justification of the reasons why a test proposed in the literature cannot formally be considered as a "standard" test.

The main idea of the SEM methodology is to insert a stochastic step S between the E-step and the M-step of the EM algorithm. In the sequel, the symbol r will denote the SEM iterations. The SEM algorithm can then be described as follows.

Let us define an upper bound K for the unknown number of components, and a threshold $c(N)$ lying between 0 and 1 (where N is the sample size). The SEM iteration $q^r \rightarrow q^{r+1}$ is

E-step: for $k = 1, ..., K$ *and* $i = 1, ..., N$ *compute* $t_k^r(x_i)$ *as before.*
S-step: for every observed x_i*, draw the pseudo-complete sample* $y_i = (x_i, z_i)$*, by replacing each missing quantity* z_i *by a value drawn at random according to the probabilities* $t_k^r(x_i)$*. This amounts to drawing a single multinomial observation* $z^r(x_i)$ *with probabilities* $(t_k^r(x_i), k = 1, ..., K)$*. If* $\frac{1}{N}\sum_{i=1}^{N} z_k^r(x_i) < c(N)$*, draw at random new values of* z_i *from a preassigned distribution on* $\mathbf{Z}$*, such that* $\frac{1}{N}\sum_{i=1}^{N} z_k^r(x_i) \geq c(N)$ *and go to the M-step.*
M-step: compute the ML estimates q^{r+1} *based on the pseudo-complete sample constructed at the S-step. This amounts to computing*

$$p_k^{r+1} = \frac{1}{N}\sum_{i=1}^{N} z_k^r(x_i).$$

The estimation of the θ_k'*s depends on the nature of the underlying mixture density.*

Remarks:

1) We give the main idea that underlies the S-step. This step creates a partition of the data sample into K classes. The idea is then that, if there exists

a class k such that the number of observations that fall into it is "small", then this would mean that the parameter has reached the boundary of the parameter space (by "small" it is meant that the number of observations is smaller than $Nc(N)$; if the data vector is of dimension d, we typically choose (Celeux and Diebolt, 1985) $c(N) = c(N, d) = \frac{d+1}{N}$).

2) In practice, instead of going through the tedious procedure generated by the S-step, the alternative suggested (Celeux, 1987) in the case that the class k contains very few elements is to delete the k^{th} component and run the algorithm on the basis of the remaining $K - 1$ components. This approach gives very good results for large sample sizes.

So, the main difference with EM is that, instead of maximising an expected log-likelihood, the SEM algorithm simulates the missing data and, in this way, creates a pseudo-complete sample whose log-likelihood is then directly maximised to yield the next estimator. Since, in general, the expression of the complete-data log-likelihood can be put into a closed form, the M-step is easy to implement; in these situations, the SEM algorithm is very attractive.

Points a, b and c below give a quick comparison between the EM and the SEM algorithms.

a) The SEM algorithm allows misspecifications of the number of components in a mixture model: indeed, one need only know an upper bound of this number; the SEM algorithm will always find the exact number provided the sample size is large enough. This is a general property of this algorithm (Celeux and Diebolt, 1988), and one can easily see this idea from the practical implementation of the S-step.

b) The random imputation principle implemented at the S-step deletes a nice feature that one finds in EM, namely that the observed log-likelihood is increased at each iteration; however, at the same time, in contrast to EM, it allows SEM to avoid saddle-points and local maxima.
c) The initial values of the parameters are not important any more. Provided that the samples are big enough, the sequence generated by SEM, will converge in distribution to a stationary distribution approximately concentrated around the MLE. However, for small samples, the results depend on the initial values, and it would be more appropriate, then, to use a variant of SEM, the so-called *Simulated Annealing EM* (*SAEM*) algorithm. If q^r_{SAEM} is the parameter estimate at iteration r of the SAEM algorithm, then the $(r+1)^{th}$ iteration step of SAEM can be written as $q^{r+1}_{SAEM} = \gamma_r q^{r+1}_{SEM} + (1-\gamma_r) q^{r+1}_{EM}$, where q^{r+1}_{SEM} and q^{r+1}_{EM} are the respective parameter estimates using SEM and EM, and (γ_r) is a sequence of positive numbers decreasing slowly from 1 to 0. The main reference for this method can be found in Celeux and Diebolt (1992), but the method is also discussed in other papers (for example Robert (1992), or Celeux, Chauveau and Diebolt (1995)).

The conclusion that can be drawn from points a, b and c, is that, in practice, for a reasonably large sample size, SEM improves EM; the latter should be preferred only in cases where the components of the mixture model are well-separated, and the number of components known beforehand.

It is important to emphasise that a new factor (complication) introduced by the SEM algorithm is that we have now two probability spaces, as follows. 1) The sample space (Ω,A,P) where $\Omega=(\mathbb{R}^d)^N$ (d is the dimension of the sample data), A = Borel sets associated with the product topology of Ω,

P=$\prod_{i\in N} F_i$, where F is the distribution of the mixture under consideration.
2) The space of random drawings $(\xi, B(\xi), \pi)$: when we fix the sample $(X_1,...,X_N)$, we get at each iteration $j = 1, ..., r$ of SEM, N independent drawings from $t_k(x_i)$ with $i = 1, ..., N$. The probability space of this sequence of drawings is called *the space of random drawings* (Celeux and Diebolt, 1986b).

3.1.2 Some theory in the one-parameter case

Let us first consider the one-parameter case, that is, where only the true mixing weight p in a two-component mixture is unknown.

If we denote by p_N^r the successive iterates of the SEM algorithm for the mixing weight parameter, p, and by p_N the maximum likelihood estimate of p, the SEM iteration $(r) \to (r+1)$ can be written

$$p_N^{r+1/2} = T_N(p_N^r) + V_N(p_N^r, z_r),$$

where T_N and V_N are defined below.

Define now $c(N)$ to be a sequence of thresholds such that $G^N = [c(N), 1 - c(N)]$. Then we have
if $p_N^{r+1/2} \in [c(N), 1 - c(N)]$ then $p_N^{r+1} = p_N^{r+1/2}$,
otherwise, draw p_{r+1}^N from a preassigned distribution supported in $[c(N), 1 - c(N)]$ and go to the E-step (this is a consequence of the S-step).

The deterministic part of the above equation, due to the EM methodology, is

$$T_N(p_N^r) = \frac{1}{N}\sum_{i=1}^{N} t_1^r(x_i).$$

Note that, in the calculation of $t_1^r(x_i)$, only the proportion parameter is updated.

Also,

$$V_N(p_N^r, z_r) = \frac{1}{N}\sum_{i=1}^{N}(z_1^r(x_i) - t_1^r(x_i)).$$

$V_N(p_N^r, z_r)$ can be written as $N^{-1/2}s_N(p_N^r)\eta_N^{r+1}(p_N^r, z_r)$ with

$$\eta_N^{r+1}(p_N^r, z_r) = \sum_{i=1}^{N}(z_1^r(x_i) - t_1^r(x_i))[\sum_{i=1}^{N} t_1^r(x_i)(1 - t_1^r(x_i))]^{-1/2},$$

and s_N is a function defined on $[0, 1]$ by

$$s_N^2(p) = \frac{1}{N}\sum_{i=1}^{N} t_1(x_i)(1 - t_1(x_i))$$

for $p \in [c(N), 1 - c(N)]$; $s_N^2(p)$ is a non-negative constant if $x \notin [c(N), 1 - c(N)]$.

It has been shown that

$$\lim_{N\to\infty} s_N(p_N) = s,$$

with $s \neq 0$, and on the other hand that $s_N(0) = s_N(1) = 0$ (Celeux and Diebolt, 1986b).

3.1.3 The general case

We consider now the general case where all the mixture parameters are unknown.

Let q=(p_1,...,p_{K-1},θ_1,...,θ_K) in $\mathbb{R}^p$, with $p = K - 1 + lK$, be the vector parameter that we want to estimate. Let q_N be the asymptotically convergent solution of the EM algorithm under the assumptions of Redner and Walker's theorem.

We need now introduce some assumptions as in Celeux and Diebolt (1986b):

(A1) The assumptions of Redner and Walker's Theorem (1984) hold.
(A2) For $h = 1/N > 0$ consider a decreasing family G_h of Borel sets of $\mathbb{R}^p$ such that the following holds:
let $G = \cup_{h>0} G_h$ be a fixed Borel set of $\mathbb{R}^p$; there exists a real number b, with $0 < b \leq 1$, such that the ball B of $\mathbb{R}^p$ with centre 0 and radius b is included in all Borel sets G_h, for h small enough.

We digress here to say that the variables and functions used in the sequel are centred, for simplicity. We then have

$$X_r^h = q_N^r - q_N$$

and, for $x \in G$,

$$T_h(x) = T_N(x + q_N) - q_N$$

$$s_h(x) = s_N(x + q_N)$$

$$\eta_{r+1}^{h}(x,z) = \eta_{N}^{r+1}(x+q_N, z).$$

(A3) We consider real numbers $A(h)$ with $0 \leq A(h) < 1$ such that

$$lim_{h\to 0} A(h) = 1,$$

where $A(h)$ is a decreasing function.

(A4) The functions $T_h(x)$ are such that $\forall x \in G_h$ we have

$$|T_h(x)| \leq A(h)|x|,$$

where $|.|$ is the norm on $\mathbb{R}^p$ introduced in Redner and Walker's theorem.

(A5) For $x \notin G_h$, $T_h(x) = t_h$, where t_h is a constant and $t_h \in A(h)G_h$.

(A6) There exists a real w, with $0 < w < 1$, a matrix $a \in M_p(\mathbb{R})$ (i.e the set of square $p \times p$ matrices with real entries) and, for all positive h, a matrix $a_h \in M_p(\mathbb{R})$ such that:
1) For all positive h, $||a_h|| \leq w$ where $||.||$ is the operator norm associated with the norm of (A4).
2)

$$\lim_{h\to 0} a_h = a.$$

$$x \in B \Rightarrow |T_h(x) - a_h x| \leq c|x|^2,$$

where c is a constant.

(A7) Let s_h be the mapping: $\mathbb{R}^p \to M_p(\mathbb{R})$. Then, for all positive h and for all $x \notin G_h$, we have that $s_h(x) = v_h$ is a constant matrix.

(A8) $\sup(||s_h(x)||, h > 0, x \in \mathbb{R}^p) = C$ with $C < \infty$.

(A9) There exists a matrix $s \in M_p(\mathbb{R})$ such that

1)

$$\lim_{h \to 0} s_h(0) = s$$

2)

$$x \in B \Rightarrow ||s_h(x) - s_h(0)|| \leq c|x|,$$

where c is a constant.

(A10) For r a positive integer, $h > 0$ and $x \in \mathbb{R}^p$, let the r.v's $\eta_r^h(x, z)$ and the normal r.v.'s $\epsilon_r(z)$ on $(\xi, B(\xi), \pi)$ take values in $\mathbb{R}^p$; we have then that, for $h > 0$ and $x \in \mathbb{R}^p$ both fixed, the r.v's $\eta_r^h(x, .)$ (resp. $\epsilon_r(z)$) are i.i.d.

(A11)

1)

$$E_z(\eta_r^h(x, z)) = E_z(\epsilon_r(z)) = 0.$$

2)

$$E_z((\eta_r^h)^2) = E_z((\epsilon_r(z))^2) = 1.$$

3) $\eta_r^h(x, z)$ converges in distribution to $\epsilon_r(z)$ as $h \to 0$.

The Markov chain $(SEM)_h$ can be written as

$$X_{r+1/2}^h = T_h(X_r^h) + h^{1/2} s_h(X_h^r) \eta_{r+1}^h(X_r^h, z).$$

Then, $X^h_{r+1} = X^h_{r+1/2}$, if $X^h_{r+1/2} \in G_h$, and $X^h_{r+1} = \Gamma_{r+1}$ otherwise, where Γ_{r+1} is a realization of a r.v. drawn independently of the r.v.'s $X^h_0, ..., X^h_r$ and of $\eta^h_{r+1}(X^h_{r+1}, .)$, according to a preassigned distribution γ_h with support in the Borel set $A(h)G_h$; this results from the S-step.

(A12)
X^h_0 and $\eta^h_1(x, .)$ are independent.
(A13)

$$E_z((X^h_0)^2) < \infty.$$

(A14) The support of the distribution of X^h_0 is a subset of G_h.

(A15) The associated Markov chain $(Z^h_r; r \geq 0)$ is defined as

$$Z^h_{r+1} = a_h Z^h_r + s_h(0)\epsilon_{r+1}(z)$$

and

$$Z^h_0 = Y^h_0.$$

This is a linear AR(1) with normal white noise, and it is assumed that, for small h, the chain $(Z^h_r; r \geq 0)$ is ergodic.

(A16) The associated Markov chain $(Z_r; r \geq 0)$ is defined as

$$Z_{r+1} = aZ_r + s\epsilon_{r+1}(z).$$

This also is a linear AR(1) and it is assumed that the chain $(Z_r; r \geq 0)$ is ergodic. We denote its stationary distribution by λ. Thus, λ is the normal distribution on $\mathbb{R}^p$ with mean vector zero and non-singular covariance matrix (from (A16)), defined by

$$\sum_{l=0}^{\infty} a^l s s^T (a^T)^l,$$

where s^T and a^T are the transposes of s and a respectively. This sum is convergent according to (A6).

Remark: the main restriction here is that each function T_h has a unique fixed point which in this case is taken to be the point zero (because the variables are centred), for simplicity. However, this restriction does not seem to be a problem since, as h goes to zero, the consistent estimator becomes prominent (Diebolt and Celeux, 1992).

The main asymptotic theorem is the following (Celeux and Diebolt (1986), Celeux (1987)).

SEM THEOREM:

Let ϕ^h be the stationary distribution of the chain $(SEM)_h$, and let ψ^h be the stationary distribution of the normalized chain $(h^{-1/2} X_r^h; r \geq 0)$. Then, under assumptions (A1) to (A16) and under the condition $lim_{h \to 0} h^b (1 - A(h))^{-1} = 0$, where $b = \frac{\alpha}{2(1+\alpha)}$ for any $\alpha \in]0, 1]$, ψ^h converges to λ as h tends to 0.

This amounts to saying that, if X_N is a r.v. defined on G^N and distributed according to the stationary distribution ψ_N of the Markov chain generated by SEM, then, under assumptions (A1) to (A16), we have that $N^{1/2}(X_N - q_N) \to N(\mathbf{0}, \Gamma)$ in distribution as $N \to \infty$, and the matrix Γ can be expressed in terms

of the true parameter q. This theorem is the equivalent, for the SEM algorithm, of Redner and Walker's theorem for the EM algorithm.

3.2 Problems when considering a test for the number of components based on SEM

3.2.1 Construction of the test

It has been proposed (Celeux, 1987) to use the SEM algorithm in order to assess the quality of an estimate given by some classification method; for that purpose, a test based on the SEM iterates is proposed, and its distribution is derived; this test is also used by Soromenho (1994).

We assume here that the estimate is given by the EM algorithm, and construct the test in the following way:
Let us consider the uncentred r.v's q_N^r that we denote by q_r for notational simplicity. In addition, let us denote by $\hat{q}_N$ the MLE given by EM and let q_N be the unique strongly consistent solution of the likelihood equations whose existence is guaranteed thanks to Redner and Walker's theorem.
The null hypothesis (H_0) is that the correct number of components is known and that $\hat{q}_N = q_N$.

Let us consider now the above $\hat{q}_N$ as a starting point, and, from this position, run r iterations of the SEM algorithm and consider the empirical mean $\bar{q}_r = \frac{1}{r}\sum_{j=1}^{r} q_j$ and covariance matrix $M_r = \frac{1}{r}\sum_{j=1}^{r}(q_j - \bar{q}_r)(q_j - \bar{q}_r)^T$.

Then, it has been claimed in Celeux (1987) that, thanks to (H_0), the stationary state has been reached at iteration zero of the SEM algorithm, and so, using the SEM Theorem, the sequence of SEM estimates (q_r) of the mixture parameters can be considered as a sequence of independent realizations of a normal distribution with mean q_N and some covariance matrix $N^{-1/2}\Gamma$; therefore, the statistic defined as $T_r^2 = \frac{1}{r+1}(\hat{q}_N - \bar{q}_r)^T M_r^{-1}(\hat{q}_N - \bar{q}_r)$ can be considered to be distributed as a Hotelling's statistic with $(r-1)$ degrees of freedom, and the statistic $F = \frac{T_r^2}{r-1}\frac{r-p}{p}$ is distributed as a Fisher $F(p, r-p)$; p is the number of parameters and r the number of SEM iterations, and $p < r$.

The main purpose of this work is to show the following:
1) it is not true that the r.v's generated by SEM are mutually independent; indeed, they show a "weak dependence" on each other and constitute asymptotically (as the sample size N tends to infinity), a linear autoregressive Markov chain of order one (AR(1)), so that, from a mathematical viewpoint, the above T_r^2 statistic is not a Hotelling's statistic;
2) it is possible to derive the limiting distribution (as $r \to \infty$) of T_r^2.
The remainder of this chapter deals only with the first point; the next chapter will deal with the second point.

3.2.2 On some properties of the SEM iterates

First, when considering the SEM Theorem, it is not correct to speak in terms of independent r.v's since, in the one-parameter case, for example, we would obtain by assumption (A16) that the covariance of the stationary measure λ of the chain $(Z_r; r \geq 0)$ would be, for $i < j$, $cov(Z_i, Z_j) = a^{j-i}\frac{s^2}{1-a^2}$ which is not equal to 0.

As for the technical assumptions underlying the SEM Theorem, we now need to centre the r.v.'s for simplicity. Furthermore, as we did for q_N^r, we will denote the centred variables X_i^h by X_i, where $i = 1, ..., r$, for simplicity. Thus, $X_i = q_i - q_N$ for $i = 1, ..., r$, and the null hypothesis (H_0) becomes $X_0 = 0 = q_N$. For $X_{r+1/2} \in G_h$ the Markov chain generated by the SEM algorithm satisfies the following recurrence equation:

$$X_{r+1} = T_h(X_r) + h^{1/2} s_h(X_r) \eta_r^h(X_r, z).$$

With the change of variables $x = h^{1/2}y$ we use Lemma 2 of Celeux and Diebolt (1986b), which is as follows.

For all $x = h^{1/2}y$ and all α with $0 < \alpha \leq 1$, the following two inequalities hold:

$$|h^{-1/2} T_h(h^{1/2}y) - a_h y)| \leq c\, h^{\alpha/2} |y|^{1+\alpha}$$

$$||s_h(h^{1/2}y) - s_h(0)|| \leq c\, h^{\alpha/2} |y|^{\alpha},$$

where c is a constant.

These are proved using assumptions (A6) and (A9).
The normalized chain Y_r is constructed as follows, for $X_r \in G_h$:

$$h^{1/2} Y_{r+1} = T_h(h^{1/2} Y_r) + h^{1/2} s_h(h^{1/2} Y_r) \eta_{r+1}^h(h^{1/2} Y_r, z).$$

We multiply both sides by $h^{-1/2}$ to obtain

$$Y_{r+1} = h^{-1/2} T_h(h^{1/2} Y_r) + s_h(h^{1/2} Y_r) \eta_{r+1}^h(h^{1/2} Y_r, z),$$

so that, for small h, we have

$$h^{-1/2}T_h(h^{1/2}Y_r) \approx a_h Y_r.$$

Moreover, by (A6), $\lim_{h\to 0} a_h = a$.
On the other hand, for small h, we have

$$s_h(h^{1/2}Y_r) \approx s_h(0).$$

Moreover, by (A9), $\lim_{h\to 0} s_h(0) = s$.
Thus, using assumption (A11), we obtain asymptotically (as $h \to 0$) that the normalized chain Y_r satisfies the recurrence equation

$$Y_{r+1} = aY_r + s\epsilon_{r+1}.$$

Since $h^{1/2}$ is just a normalizing value, we can multiply both sides of the above equation by it to get back to our original chain X_r so that, asymptotically,

$$X_{r+1} = aX_r + h^{1/2}s\epsilon_{r+1}.$$

On the other hand, the subsets G_h of G increase towards G as h goes to 0, and, asymptotically, $X_r \in G$ so that the above equation is valid everywhere.

Consequently, the null hypothesis (H_0) that we consider will be as follows.
As the sample size N tends to infinity, the sequence ($X_r \in \mathbb{R}^p$) of the estimates of the parameters of the mixture density can be considered as a linear AR(1) Markov chain, satisfying

$$X_{r+1} = aX_r + N^{-1/2} s\epsilon_{r+1}, \tag{3.1}$$

for $r > 0$, and $X_0 = 0$, where $a \in M_p(\mathbb{R})$, $||a|| < 1$ and $||.||$ is the operator norm associated with the norm of assumption (A3); in the univariate case this norm is just the absolute value. Also, $s \in M_p(\mathbb{R})$, s is non-singular and $\epsilon_r \sim N(\mathbf{0}, I)$, where $\mathbf{0}$ is the null vector of $\mathbb{R}^p$ and I is the $p \times p$ identity matrix.

Furthermore, equation 3.1 shows a weak dependence among the SEM iterates and therefore it is not straightforward to see how the proposed T_r^2 should approximate a Hotelling's statistic. An important fact that we have to emphasise is that, when N is fixed, the Markov chain $(SEM)_h$ is ergodic. This result is stated in Celeux and Diebolt, 1986b. Hence the Markov chain satisfying equation 3.1 is ergodic and therefore has a stationary distribution. General results about Markov chains can be found in Doob (1953), Kemeny and Snell (1974) and Grimmett and Stirzaker (1992) among others.

In the univariate case there is also a result by Broniatowski and Diebolt (1987) concerning the general AR(1):
$X_{r+1} = T(X_r) + \epsilon_r$ with $T : \mathbb{R} \to \mathbb{R}$. They proved that, in the linear case where $T(y) = ay$, the necessary and sufficient condition for this chain to be ergodic is that $|a| < 1$, and, if this is true, the chain X_r has a stationary distribution.

Another important result is that the chain $(SEM)_h$ is uniformly strongly mixing (Diebolt and Celeux, 1992). A similar result has also been proved by Athreya and Pantula (1986) for an autoregressive process, where the ϵ_i's need

not be Gaussian any more, under some technical assumptions.

3.3 Connection with the theory of stationary random variables

3.3.1 An extension of the central limit theorem

We consider in the sequel the univariate case. We will show then that the results here generalize, for our case, the central limit theorem (CLT) for stationary random variables (Theorem 18.5.3 of Ibragimov and Linnik, 1971), and we have to stress the fact that, as for this theorem, the main factor underlying the result in our case is a strong mixing property.

We recall first the definition of strongly and uniformly mixing processes. A process (X_r) is said to be strongly(α) mixing (resp. uniformly(ϕ) mixing) if, defining by F_r^n the σ-algebra $\sigma(X_j; r \leq j \leq n)$, we have for all $A \in F_0^r$ and $B \in F_{r+n}^{\infty}$ that

$$\alpha(n) = sup_{A \in F_0^r, B \in F_{r+n}^{\infty}} |P(A \cap B) - P(A)P(B)| \to 0$$

and respectively

$$\phi(n) = sup_{A \in F_0^r, B \in F_{r+n}^{\infty}} |P(B|A) - P(B)| \to 0$$

as $n \to \infty$.

Some theoretical results concerning these processes are given in Athreya and

Pantula (1986), Bradley (1986), Davydov (1973) and Kolmogorov and Rozanov (1960).

Remark: a $\phi - mixing$ process is $\alpha - mixing$ (Athreya and Pantula (1986), Bradley (1986)).

These definitions are valid for stationary and non-stationary processes. In our case, the process defined by equation 3.1, is defined as *stable* in the sense given by Anderson (1959); however, since its autocovariance function depends on r, it is not stationary in "full strength". Nevertheless, we will see that a central limit theorem still applies. We state now the central limit theorem (CLT). There is a version of the CLT concerning $\alpha - mixing$ stationary processes and a version concerning $\phi - mixing$ stationary processes (Ibragimov and Linnik, 1971). They both say roughly that, under some conditions on the absolute moments and the mixing coefficients, the quantity $\sigma^2 = E(X_0^2) + 2\sum_{j=1}^{\infty} E(X_0 X_j)$ is finite, where $E(X_0^2)$ is the common variance and $E(X_0 X_j)$ the covariance of the centred variables. Also, if $\sigma \neq 0$, then

$$\sigma^{-1} r^{-1/2} \sum_{j=1}^{r} X_j \rightarrow N(0,1)$$

in distribution, as $r \rightarrow \infty$.

By 'stationarity', we mean that we consider a sample X_1,...,X_r, of the sequence of SEM iterates, once the stationary state has been reached (we suppose that X_1 is the first such iterate). Then, $X_i = \frac{s}{\sqrt{N}} \sum_{j=0}^{\infty} a^j \epsilon_{i-j}$, for $i = 1, ..., r$, whose distribution is the stationary probability measure of our Markov chain $(SEM)_h$. Thus, the expressions that we compute in the stationary case are

derived from our stable case, by taking r to tend to infinity. This, will allow comparison between asymptotic results obtained for our case and for the stationary case.

At stationarity, the SEM iterates then satisfy the assumptions of these CLT's, and we obtain

$$E(X_0^2) = \lim_{r\to\infty} \frac{s^2(1-a^{2r})}{N(1-a^2)} = \frac{s^2}{N(1-a^2)},$$

and

$$E(X_0 X_j) = \lim_{r\to\infty} a^j \frac{s^2(1-a^{2r})}{N(1-a^2)} = \frac{s^2}{N(1-a^2)} a^j.$$

Then

$$\sum_{j=1}^{\infty} E(X_0 X_j) = \frac{s^2 a}{N(1-a^2)(1-a)},$$

and we obtain

$$\sigma^2 = \frac{s^2}{N(1-a)^2} < \infty.$$

Hence

$$\frac{\bar{X}}{\sigma/\sqrt{r}} \to N(0,1)$$

in distribution as $r \to \infty$.

The assumptions of the CLT being satisfied, we will have at stationarity that $\frac{\bar{X}}{\sigma/\sqrt{r}} \to N(0,1)$ in distribution as $r \to \infty$.

We will obtain an analogous result in the case of the chain (X_r); indeed, this result is a univariate specialization of the lemma presented in the next chapter, so that we can consider that the above CLT's extend to our case.

3.3.2 Definition of the test statistic

If we define $\bar{X} = \frac{1}{r}\sum_{j=1}^{r} X_j$, and $M_r = \frac{1}{r}\sum_{j=1}^{r}(X_j - \bar{X})(X_j - \bar{X})^T$, as before, the statistic that will be used to evaluate the solution given by the EM algorithm is defined as

$$T_r^2 = (X_0 - \bar{X})^T M_r^{-1}(X_0 - \bar{X}).$$

This is in fact equivalent to the statistic used in Celeux (1987) and Soromenho (1994), and mentioned before, with the difference that the r.v.'s considered here are centred. Since $X_0 = 0$, T_r^2 can be written as

$$T_r^2 = \bar{X}^T M_r^{-1} \bar{X}.$$

For finite r, T_r^2 obviously cannot be a standard Hotelling. Thus, the above comments, and the facts that, in proving the SEM Theorem, r is considered to tend to infinity as N tends to infinity and that in our case stationarity is reached as $r \to \infty$, suggest trying the case where r is large.
Again, let us consider the univariate case; then

$$T_r^2 = \frac{\bar{X}^2}{\frac{1}{r}\sum_{i=1}^{r}(X_i - \bar{X})^2}.$$

The denominator can also be written as $\frac{1}{r}\sum_{i=1}^{r} X_i^2 - \bar{X}^2$.

In the case where r is large, algebraic calculations show that the distributions of $\frac{1}{r}\sum_{i=1}^{r} X_i^2$ and $\frac{s^2}{N(1-a^2)r}\sum_{i=1}^{r}\epsilon_i^2$ are not the same; this is the subject of the next section.

3.4 T_r^2 is not a standard Hotelling statistic

We consider here, the case where the parameter is one-dimensional. The asymptotic calculations performed in this section assume that the sample size N is fixed, and that the number of iterations r is large. Let us now compute the first two moments of the expression $\frac{1}{r}\sum_{i=1}^{r} X_i^2$. In the independence case this expression can be written as $\frac{V}{r}$, where V is distributed like a $\chi^2(r)$, and the variance of this expression equals two times its expectation divided by r; in our case, some algebraic calculation yields

$$\frac{1}{r}\sum_{i=1}^{r} X_i^2 = \frac{s^2}{N(1-a^2)r}[\sum_{i=1}^{r}\epsilon_i^2 - D_r + 2\sum_{i<j} a^{j-i}(1-a^{2r-2j+2})\epsilon_i\epsilon_j],$$

where

$$D_r = a^{2r}\epsilon_1^2 + ... + a^4\epsilon_{r-1}^2 + a^2\epsilon_r^2.$$

Now let us denote by B_r the r.v.

$$\frac{s^2}{N(1-a^2)r}[2\sum_{i<j} a^{j-i}(1-a^{2r-2j+2})\epsilon_i\epsilon_j - D_r].$$

We have that

$$E(B_r) = -\frac{s^2}{N(1-a^2)}E(D_r) = -\frac{a^2(1-a^{2r})s^2}{N(1-a^2)^2 r},$$

which yields

$$E(\frac{1}{r}\sum_{i=1}^{r} X_i^2) = \frac{s^2}{N(1-a^2)} - \frac{a^2(1-a^{2r})s^2}{N(1-a^2)^2 r} \approx \frac{s^2}{N(1-a^2)},$$

for large r. Also,

$$var(B_r) = \frac{s^4}{N^2(1-a^2)^2 r^2}[var(2\sum_{i<j} a^{j-i}(1-a^{2r-2j+2})\epsilon_i\epsilon_j) + var(D_r)$$

$$-2cov(2\sum_{i<j} a^{j-i}(1-a^{2r-2j+2})\epsilon_i\epsilon_j, D_r)].$$

It is easily seen that

$$cov(\sum_{i<j} a^{j-i}(1-a^{2r-2j+2})\epsilon_i\epsilon_j, D_r) = 0,$$

and that

$$\frac{s^4}{N^2(1-a^2)^2 r^2} var(D_r) \sim \frac{1}{r^2},$$

(since $var(D_r) = \frac{2a^4(1-a^{4r})}{1-a^4}$). We now calculate

$$\frac{s^4}{N^2(1-a^2)^2 r^2}[var(2\sum_{i<j} a^{j-i}(1-a^{2r-2j+2})\epsilon_i\epsilon_j)],$$

and, after some algebra, we find it to be

$$\frac{4a^2 s^4}{N^2(1-a^2)^3 r^2}[(1-a^2)^2(1-a^{2r-2}) + (1-a^4)^2(1-a^{2r-4}) + ...$$

$$+(1-a^{2r-2})^2(1-a^2)] \sim \frac{1}{r},$$

since every term inside the square brackets is greater than $(1-a^2)^3$. We conclude that $var(B_r) \sim \frac{1}{r}$.

On the other hand, $var(\frac{1}{r}\sum_{i=1}^{r} X_i^2)$ can be written as

$$var(\frac{s^2}{N(1-a^2)}\frac{1}{r}\sum_{i=1}^{r}\epsilon_i^2) + E(B_r^2) + 2E[(\frac{s^2}{N(1-a^2)r}\sum_{i=1}^{r}\epsilon_i^2 - \frac{s^2}{N(1-a^2)})B_r].$$

After some algebra we find that the third term of the above sum can be written as $-\frac{2a^2(1-a^{2r})s^4}{N^2(1-a^2)^3r^2} \sim \frac{1}{r^2}$, and the first term is, as in the independence case, $\frac{2s^4}{N^2(1-a^2)^2r}$. The second term is of order $\frac{1}{r}$. Thus, for large r, we obtain

$$var(\frac{1}{r}\sum_{i=1}^{r} X_i^2) \approx \frac{2s^4}{N^2(1-a^2)^2r} + O(\frac{1}{r}).$$

Since this variance is not equal to two times the expectation divided by r, $\frac{1}{r}\sum_{i=1}^{r} X_i^2$ has not the same distribution as in the independence case. The main purpose of the next chapter, will then be to find a limiting (as $r \to \infty$) approximation of T_r^2.

Chapter 4

On the Limiting Distribution of T_r^2

4.1 Introduction

We consider here the case where the parameter is a vector. It is well known that if the random vectors of dimension p in hand were independent, rT_r^2 would follow a $\chi^2(p)$ distribution, as $r \to \infty$, and thus, for r large, one can consider that the rescaled T_r^2 is a $\chi^2(p)$ divided by r. The main interest of this chapter will then be to derive the distribution of T_r^2 for r large. Some simulation results will also be produced that will demonstrate the performance of T_r^2.

As derived previously (equation 3.1), the null hypothesis that we consider is as follows: the Markov chain $(SEM)_h$ satisfies the autoregressive equation

$$X_{r+1} = aX_r + N^{-1/2} s\epsilon_{r+1}, \tag{4.1}$$

where $X_r \in \mathbb{R}^p$, a is a $p \times p$ matrix with spectral radius $r(a) < 1$, s is a non singular $p \times p$ matrix, $\epsilon_r \in \mathbb{R}^p$ and $\epsilon_r \sim N(\mathbf{0}, I)$ where $\mathbf{0}$ is the vector with p coordinates equal to 0 and I is the $p \times p$ identity matrix.

4.2 The limiting result

4.2.1 A result for the sample variance-covariance matrix

The statistic that will be used to test for the number of components in a mixture will be

$$T_r^2 = \bar{X}^T M_r^{-1} \bar{X},$$

where $M_r = \frac{1}{r} \sum_{i=1}^{r} (X_i - \bar{X})(X_i - \bar{X})^T$. We are interested in the asymptotic behaviour of M_r. The main result here is based on the following theorem (Theorem 5.5.2., Anderson, 1971).

THEOREM

If X_r is defined by (4.1), with a having eigenvalues less than 1 in absolute value, if the ϵ_r 's are i.i.d. with $E(\epsilon_r) = \mathbf{0}$ and $E(\epsilon_r \epsilon_r^T) = \Sigma$, then

$$\frac{1}{r} \sum_{i=1}^{r} X_i X_i^T \to \sum_{l=0}^{\infty} a^l \Sigma (a^T)^l$$

in probability as $r \to \infty$.

The assumptions of this theorem are satisfied in our case; indeed, on the one hand, the matrix norm defined by $r(a)$ is the norm used by Redner and

Walker in their theorem, so that the eigenvalues of a are less than 1 in absolute value, and on the other hand the conditions on the moments of the ϵ_r 's are also satisfied. The matrix Σ is $\frac{ss^T}{N}$ in our case, and according to the technical assumptions underlying the SEM Theorem, the limiting constant matrix $\sum_{l=0}^{\infty} \frac{a^l ss^T (a^T)^l}{N}$ exists and is non-singular. We remark that, as in Brockwell and Davis (1991), p. 408, convergence in probability of a random matrix will mean convergence in probability of all the components of the matrix.
On the other hand, let us compute the variance of $\bar{X}$ for r large. The following lemma can then be derived.

LEMMA

Under (H_0), *for large* r, *the variance-covariance matrix of the vector* $\bar{X}$ *can be written as*

$$Var(\bar{X}) = (I-a)^{-1} ss^T (I-a^T)^{-1} \frac{1}{Nr} + o(\frac{1}{r}).$$

Proof

$$\bar{X} = \frac{1}{\sqrt{N}r}(I-a)^{-1}[(I-a^r)s\epsilon_1 + ... + (I-a)s\epsilon_r]$$

and $E(\bar{X}) = \mathbf{0}$ because $E(\epsilon_i) = \mathbf{0}$ for $i = 1, ..., r$. Also,

$$Var(\bar{X}) = E(\bar{X}\bar{X}^T) = \frac{1}{Nr^2}(I-a)^{-1}[(I-a^r)sE(\epsilon_1\epsilon_1^T)s^T(I-a^r)^T + ...$$

$$+(I-a)sE(\epsilon_r\epsilon_r^T)s^T(I-a)^T](I-a^T)^{-1}.$$

Since $E(\epsilon_i \epsilon_i^T) = I$, we obtain

$$Var(\bar{X}) = \frac{1}{Nr^2}(I-a)^{-1}[(I-a^r)ss^T(I-(a^T)^r) + ...$$

$$+(I-a)ss^T(I-a^T)](I-a^T)^{-1}.$$

After some algebra, the term inside the square brackets can be put in the form

$$(r-1)ss^T + NVar(X_r) - ss^T a^T(I-a^T)^{-1}(I-(a^T)^r) - a(I-a)^{-1}(I-a^r)ss^T.$$

Now, according to the technical assumptions, $\frac{\sum_{k=0}^{r} a^k ss^T (a^T)^k}{N}$ converges to a constant matrix C as r goes to infinity. We may therefore consider that, for large r, $Var(X_r) = C$. On the other hand, since the spectral radius of a is less than 1, it follows that $lim_{r\to\infty} a^r = [0]$ so that we consider $a^r = (a^T)^r = [0]$ for large r. We therefore obtain

$$Var(\bar{X}) = \frac{1}{Nr^2}(I-a)^{-1}[(r-1)ss^T + NC - ss^T a^T(I-a^T)^{-1} - a(I-a)^{-1}ss^T](I-a^T)^{-1}.$$

The lemma follows. Thus, approximately,

$$\bar{X} \sim N(\mathbf{0}, \frac{1}{Nr}(I-a)^{-1}ss^T(I-a^T)^{-1}).$$

Some algebra shows that this result is equivalent to Theorem 5.5.8. of Anderson (1971), which is as follows, using our notations.

THEOREM

If X_r is defined by equation (4.1) where the eigenvalues of a are less than 1 in absolute value, and if the $N^{-1/2}s\epsilon_i$'s are i.i.d. with expectation equal to the

vector $\mathbf{0}$ *and variance-covariance matrix equal to* $N^{-1/2}SS^T$, *then* $\sqrt{r}\bar{\bar{X}}$ *has a limiting normal distribution with mean* $\mathbf{0}$ *and variance-covariance matrix* $C(I-a^T)^{-1}+(I-a)^{-1}C-C$.

Now, on the one hand, the convergence of $\bar{\bar{X}}$ to $\mathbf{0}$ in probability is equivalent to the convergence of $[\sum_{i=1}^{p}(\bar{\bar{X}}^i)^2]^{1/2}$ to 0 in probability by Proposition 6.1.2 of Brockwell and Davis (1991), where $\bar{\bar{X}}^i$ are the components of the vector $\bar{\bar{X}}$, and on the other hand, applying Markov's inequality, we obtain

$$P(|\bar{\bar{X}}^i|^2>\epsilon)\leq\frac{1}{\epsilon}E((\bar{\bar{X}}^i)^2)$$

and since $E((\bar{\bar{X}}^i)^2)=var(\bar{\bar{X}}^i)$, we obtain, using the previous Lemma, that for each $\epsilon>0$, $P(|\bar{\bar{X}}^i|^2>\epsilon\to 0$ for each $i=1,...,p$. It results that

$$\bar{\bar{X}}\to\mathbf{0}$$

in probability as $r\to\infty$. (Another way is to say that the expectations of every component of $\bar{\bar{X}}$ are equal to 0, and the corresponding variances tend to 0 by the previous Lemma, so that by Proposition 6.2.4. of Brockwell and Davis (1991), each component of $\bar{\bar{X}}$ tends to 0 in probability.)
We first use Lemma 3.2.1 of Anderson (1958), stating that

$$M_r=\frac{1}{r}\sum_{i=1}^{r}X_iX_i^T-\bar{X}\bar{X}^T.$$

We need now to derive the limit in probability of $\bar{X}\bar{X}^T$; for that purpose we use the following Proposition 6.1.4 of Brockwell and Davis (1991).

PROPOSITION

If (X_r) *is a sequence of p-dimensional random vectors such that* $X_r \to X$ *in probability and if* $g : \mathbb{R}^p \to \mathbb{R}^m$ *is a continuous mapping, then* $g(X_r) \to g(X)$ *in probability.*

Thus, every linear combination of elements of $\bar{X}$ tends in probability to 0, and every element of the matrix $\bar{X}\bar{X}^T$ tends in probability to 0. It follows that $\bar{X}\bar{X}^T$ tends in probability to the matrix with elements equal to 0. (Another way of obtaining this result is the following: every component of $\bar{X}$ and subsequently $\bar{X}^T$ tends to 0 in probability and thus, from Proposition 6.1.1. of Brockwell and Davis (1991), so does any product of these components.)

In order to conclude, we use now the following Proposition 6.3.8 of Brockwell and Davis (1991).

PROPOSITION

If (X_r) *and* (Y_r) *are sequences of random p-vectors such that* $X_r \to X$ *in distribution and* $Y_r \to b$ *in distribution, where* b *is a constant, then* $X_r + Y_r \to X + b$ *in distribution.*

In our case, the elements of the matrix M_r tend, thus, in probability, to the corresponding elements of the matrix C, and hence $M_r \to C$ in probability.

It follows that

$$M_r^{-1} \to C^{-1}$$

in probability as $r \to \infty$.

4.2.2 Deriving the limiting distribution of T_r^2

From the previous paragraph, we have that for large r, rT_r^2 can be approximated in distribution by $r\bar{X}^T C^{-1}\bar{X}$ which we call V. Then, V can be written as $(\sqrt{r}C^{-1/2}\bar{X})^T(\sqrt{r}C^{-1/2}\bar{X})$, where $\sqrt{r}C^{-1/2}\bar{X}$ is a normal vector, which we call W, with mean $\mathbf{0}$ and variance-covariance matrix $\Gamma_r = rC^{-1/2}Var(\bar{X})C^{-1/2}$. Now, Γ_r being symmetric, there is an orthogonal matrix Q such that $\Delta = Q^T\Gamma_r Q$ is diagonal where δ_l will denote the $(l,l)^{th}$ element, for $l = 1, ..., p$, and since Γ_r is positive definite, the elements of $\sqrt{\Delta}$ are positive. Setting $A = Q\sqrt{\Delta}$, we obtain $\Gamma_r = AA^T$, with rank of A equal to p. Hence, there is a matrix A and a standardized normal vector Z such that $W = AZ$. We have then

$$W^T W = Z^T A^T AZ$$

where $A^T A = \Delta$, so that we obtain

$$W^T W = Z^T \Delta Z = \delta_1 Z_1^2 + ... + \delta_p Z_p^2,$$

where $Z_l^2 \sim \chi_1^2$ for $l = 1, ..., p$.

Now, Δ is the matrix of eigenvalues of Γ_r or, equivalently, those of $C^{-1}Var(\sqrt{r}\bar{X})$. The following theorem then obtains.

THEOREM

Under the null hypothesis H_0, for r large, the statistic T_r^2, used in order to estimate the number of components in a mixture model, is approximately distributed as $\frac{\sum_{l=1}^{p}\delta_l Z_l^2}{r}$, where the Z_l^2's are i.i.d. $\chi^2(1)$ variates and the δ_l's are eigenvalues of $C^{-1}Var(\sqrt{r}\bar{X})$.

4.3 Another approach for deriving the limiting distribution of T_r^2

4.3.1 Another form for the statistic T_r^2

We use here first, an algebraic approach similar to the one used by Anderson (1958), in order to derive the distribution of T^2. So, let us consider an orthogonal matrix $Q(p,p)$ such that
$q_{1l} = \frac{\bar{X}_l}{\sqrt{\bar{X}^T\bar{X}}}$ for $l = 1, ..., p$. Let us set $U = Q\bar{X}$ and $B = rQM_rQ^T$ where, as before, $M_r = \frac{1}{r}\sum_{i=1}^{r}(X_i - \bar{X})(X_i - \bar{X})^T$. Then we have

$$u_1 = \sum_{i=1}^{p} q_{1l}\bar{X}_l = \frac{\bar{X}_1^T}{\sqrt{\bar{X}^T\bar{X}}}\bar{X}_1 + ... + \frac{\bar{X}_p^T}{\sqrt{\bar{X}^T\bar{X}}}\bar{X}_p = \frac{\bar{X}^T\bar{X}}{\sqrt{\bar{X}^T\bar{X}}} = \sqrt{\bar{X}^T\bar{X}}$$

and for $j \neq 1$ we have

$$u_j = \sum_{l=1}^{p} q_{jl}\bar{X}_l = \sum_{l=1}^{p} q_{jl}q_{1l}\sqrt{\bar{X}^T\bar{X}} = \sqrt{\bar{X}^T\bar{X}}\sum_{l=1}^{p} q_{jl}q_{1l} = 0,$$

since Q is orthogonal.

Thus we have that

$$T_r^2 = \bar{X}^T M_r^{-1}\bar{X} = rU^T B^{-1}U,$$

where $U = (U_1, 0, ..., 0)^T$ and

$$B^{-1} = \begin{pmatrix} b^{11} & b^{12} & \dots & \dots & b^{1p} \\ b^{21} & b^{22} & \dots & \dots & b^{2p} \\ \dots & \dots & \dots & \dots & \dots \\ b^{p1} & b^{p2} & \dots & \dots & b^{pp} \end{pmatrix}.$$

Thus

$$\frac{T_r^2}{r} = U_1^2 b^{11}.$$

Now, set B to be the inverse matrix of B^{-1}, write $B = (b_{ij})$ and let us partition B as

$$B = \begin{pmatrix} b_{11} & (b_{(1)})^T \\ b_{(1)} & B_{22} \end{pmatrix},$$

where $b_{(1)} = (b_{12}, ..., b_{1p})^T$ and

$$B_{22} = \begin{pmatrix} b_{22} & \dots & \dots & b_{2p} \\ b_{32} & \dots & \dots & b_{3p} \\ \dots & \dots & \dots & \dots \\ b_{p2} & \dots & \dots & b_{pp} \end{pmatrix}.$$

We partition B^{-1} in the same way:

$$B^{-1} = \begin{pmatrix} b^{11} & (b^{(1)})^T \\ b^{(1)} & B^{22} \end{pmatrix}.$$

From Theorem 8.2.1, result 1(d) of Graybill (1969) we obtain

$$b_{11} - b_{(1)}^T B_{22}^{-1} b_{(1)} = (b^{11})^{-1}.$$

If we write $b_{11.2,...,p} = \frac{1}{b^{11}} = b_{11} - (b_{(1)})^T B_{22}^{-1} b_{(1)}$, our statistic becomes

$$T_r^2 = \frac{\bar{X}^T \bar{X}}{\frac{b_{11.2,...,p}}{r}}.$$

On the other hand

$$B = rQM_rQ^T = \sum_{i=1}^{r} (Q(X_i - \bar{X}))(Q(X_i - \bar{X}))^T.$$

Setting then $V_i = Q(X_i - \bar{X})$, we obtain

$$B = \sum_{i=1}^{r} V_i V_i^T.$$

Now, partition X_i and V_i into two sub-vectors with 1 and $(p-1)$ components, respectively, so that $X_i = (X_i^{(1)}, (X_i^{(2)})^T)^T$ and $V_i = (V_i^{(1)}, (V_i^{(2)})^T)^T$, with $X_i^{(1)} \in \mathbb{R}$, $X_i^{(2)} \in \mathbb{R}^{p-1}$, $V_i^{(1)} \in \mathbb{R}$, $V_i^{(2)} \in \mathbb{R}^{p-1}$.
Since $b_{11} \in \mathbb{R}$ and $((V_1^{(1)})^2 + ... + (V_r^{(1)})^2) \in \mathbb{R}$ we have

$$b_{11} = \sum_{i=1}^{r} (V_i^{(1)})^2.$$

We compute now the term $b_{(1)}^T B_{22}^{-1} b_{(1)}$.
For this purpose we set $b_{(1)}^T B_{22}^{-1} = G$ and $B_{22} = H$ and we have

$$b_{(1)}^T B_{22}^{-1} b_{(1)} = (b_{(1)}^T B_{22}^{-1})(B_{22})(B_{22}^{-1} b_{(1)}) = GHG^T$$

(since B_{22} is symmetric).
(H is a $(p-1) \times (p-1)$ matrix and G a row vector of dimension $p-1$.)
Now, $G = \sum_{i=1}^{r} V_i^{(1)} (V_i^{(2)})^T H^{-1}$ and $H = \sum_{i=1}^{r} V_i^{(2)} (V_i^{(2)})^T$.
Defining $V^{(1)} = (V_1^{(1)}, ..., V_r^{(1)})$ and

$$V^{(2)} = \begin{pmatrix} V_{11}^{(2)} & \dots & V_{r1}^{(2)} \\ \dots & \dots & \dots \\ V_{1,p-1}^{(2)} & \dots & V_{r,p-1}^{(2)} \end{pmatrix},$$

we then obtain

$$GHG^T = V^{(1)} (V^{(2)})^T H^{-1} V^{(2)} (V^{(1)})^T.$$

Now, we use Theorem 6, appendix 1 of Anderson (1958), which states that, *if* C *is a* $p \times p$ *positive semidefinite matrix of rank* $r(\leq p)$, *then there is a non-singular matrix* A *such that*

$$ACA^T = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix},$$

where the identity is of order r.

Consequently, in our case, we can find a non-singular matrix F such that $FHF^T = I_{(p-1,p-1)}$ (here H is of full rank), and so

$$(F^T)^{-1}H^{-1}F^{-1} = I_{(p-1,p-1)}.$$

Let $E_2 = FV^{(2)}$ so that $V^{(2)} = F^{-1}E_2$; we have

$$E_2(E_2)^T = FV^{(2)}(V^{(2)})^T F^T = F(\sum_{i=1}^{r} V_i^{(2)}(V_i^{(2)})^T)F^T$$

$$= FHF^T = I_{(p-1,p-1)}.$$

Thus, the $(p-1)$ rows of E_2 are orthogonal and their norm is equal to 1 (E_2 is a $(p-1) \times r$ matrix).

We now use Lemma 2, appendix 1 of Anderson(1958), which states that, *if* A *is an* $n \times m$ *matrix (with* $n > m$*) such that* $A^TA = I$, *then there exists an* $n \times (n-m)$ *matrix* B *such that* $(A\,B)$ *is orthogonal.*

Hence, in our case, it is possible to find an $(r-p+1) \times r$ matrix E_1 such that

$$E = \begin{pmatrix} E_1 \\ E_2 \end{pmatrix},$$

and E is an $r \times r$ orthogonal matrix.

Let us now set $K_i = \sum_{\beta=1}^{r} e_{i\beta} V_\beta$ (where the $e_{i\beta}$'s are the entries of the matrix E) so that, setting $K^{(1)} = (K_1^{(1)}, ..., K_r^{(1)})$, we have

$$K_1^{(1)} = e_{11} V_1^{(1)} + ... + e_{1r} V_r^{(1)}, ..., K_r^{(1)} = e_{r1} V_1^{(1)} + ... + e_{rr} V_r^{(1)},$$

so that

$$(K^{(1)})^T = E(V^{(1)})^T.$$

Thus $K^{(1)} = V^{(1)} E^T$, which is equivalent to $V^{(1)} = K^{(1)} E$ since E is orthogonal, and

$$GHG^T = K^{(1)} E (E_2)^T (F^{-1})^T H^{-1} F^{-1} E_2 E^T (K^{(1)})^T.$$

However, on the one hand $(F^{-1})^T H^{-1} F^{-1} = I_{(p-1,p-1)}$, and on the other hand $E_1 (E_2)^T$ is the $(r - p + 1) \times (p - 1)$ matrix $[0]$ because E is orthogonal and $E_2 (E_2)^T = I_{(p-1,p-1)}$. Thus we obtain

$$GHG^T = K^{(1)} ([0]^T, I)^T ([0]^T, I) (K^{(1)})^T,$$

where

$$K^{(1)} ([0]^T, I)^T = (K_{r-p+2}^{(1)}, ..., K_r^{(1)})$$

and

$$([0]^T, I)(K^{(1)})^T = (K_{r-p+2}^{(1)}, ..., K_r^{(1)})^T.$$

We obtain finally

$$GHG^T = (K_{r-p+2}^{(1)}, ..., K_r^{(1)})(K_{r-p+2}^{(1)}, ..., K_r^{(1)})^T$$

$$= (K_{r-p+2}^{(1)})^2 + ... + (K_r^{(1)})^2 = \sum_{i=r-p+2}^{r} (K_i^{(1)})^2.$$

Now, since $K_i = \sum_{\beta=1}^{r} e_{i\beta} V_\beta$, we use Lemma 3.3.1 of Anderson (1958), stating that if $C = (c_{\alpha\beta})$ is orthogonal then $\sum_{\alpha=1}^{r} X_\alpha X_\alpha^T = \sum_{\alpha=1}^{r} Y_\alpha Y_\alpha^T$, where $Y_\alpha = \sum_{\beta=1}^{r} c_{\alpha\beta} X_\beta$.

In our case, $K_\alpha^{(1)} = \sum_{i=1}^{r} e_{\alpha i} V_i^{(1)}$, where the $K_\alpha^{(1)}$'s and the $V_i^{(1)}$'s are real numbers, so that

$$\sum_{i=1}^{r} (V_i^{(1)})^2 = \sum_{i=1}^{r} (K_i^{(1)})^2,$$

for any r.

Finally we obtain

$$b_{11.2,...,p} = b_{11} - b_{(1)}^T B_{22}^{-1} b_{(1)}$$

$$= \sum_{i=1}^{r} (K_i^{(1)})^2 - \sum_{i=r-p+2}^{r} (K_i^{(1)})^2 = \sum_{i=1}^{r-p+1} (K_i^{(1)})^2$$

so that the statistic T_r^2 takes the form

$$T_r^2 = \frac{\bar{X}^T \bar{X}}{\frac{1}{r} \sum_{i=1}^{r-p+1} (K_i^{(1)})^2}.$$

Note that T_r^2 is a generalisation of the one-dimensional case with $p = 1$ and $K_i^{(1)} = X_i - \bar{X}$.

4.3.2 The limiting result

From the previous paragraph, it results that

$$rT_r^2 = \frac{r\bar{X}^T\bar{X}}{\frac{1}{r}\sum_{i=1}^{r-p+1}(K_i^{(1)})^2}.$$

Now, since p is finite, the orders of magnitude of $E(\frac{1}{r}\sum_{i=1}^{r-p+1}(K_i^{(1)})^2)$ and $var(\frac{1}{r}\sum_{i=1}^{r-p+1}(K_i^{(1)})^2)$ are the same for every p. But for $p = 1$,

$$E(\frac{1}{r}\sum_{i=1}^{r}(X_i - \bar{X})^2) \to \frac{s^2}{N(1-a^2)}$$

and it can be shown that

$$var(\frac{1}{r}\sum_{i=1}^{r}(X_i - \bar{X})^2) \to 0.$$

It follows that

$$\frac{1}{r}\sum_{i=1}^{r-p+1}(K_i^{(1)})^2 \to \frac{1}{b}$$

in probability, as $r \to \infty$, where b is the $(1,1)^{th}$ element of the matrix $QC^{-1}Q^T$. A similar argument as the one used in the previous section leads then to the following result: if Γ_r' is the variance-covariance matrix of $\sqrt{b}\sqrt{r}\bar{X}$, and Δ' the matrix of eigenvalues of Γ_r', there exists a standardized normal vector J such that T_r^2 is approximately distributed as

$$\frac{\sum_{l=1}^{p}\delta_l' J_l^2}{r}$$

for r large. This result is then equivalent to the Theorem derived in the previous section.

4.4 Simulation results

In this section we give some simulation results based on simple examples in order to highlight the performance of T_r^2. We then compare T_r^2 to a statistic T^2 obtained by letting $a = [0]$, in equation 4.1. The statistic T^2 is then based on independent random variables. We notice that, for reasonably large sample sizes, these statistics validate (or invalidate) a model in the same way.

We first explain how T^2 can be computed in practice. From equation 4.1, we can construct a new process (X_r'), based on the SEM iterates, in the following way: for $1 \leq i \leq r$, let us denote by X_i' the random variables $X_1' = X_1 = \frac{s}{\sqrt{N}}\epsilon_1$, $X_2' = X_2 - aX_1 = \frac{s}{\sqrt{N}}\epsilon_2$,...,$X_r' = X_r - aX_{r-1} = \frac{s}{\sqrt{N}}\epsilon_r$.
We estimate a by $\hat{a_r} = \sum_{i=1}^{r} X_{i+1}X_i^T(\sum_{i=1}^{r} X_iX_i^T)^{-1}$ (Anderson, 1959; Hall and Heyde, 1980), which is the maximum likelihood estimate of a; some discussion is also given by Hurwitz (1950), and the consistency of $\hat{a_r}$ in the unstable case is discussed by Rubin (1950) and some asymptotic results are derived by Anderson (1959).
The empirical mean is

$$\bar{X}' = \frac{1}{r}\sum_{i=1}^{r} X_i',$$

and the empirical variance-covariance matrix is

$$M' = \frac{1}{r}\sum_{i=1}^{r}(X_i' - \bar{X}')(X_i' - \bar{X}')^T.$$

Then, the statistic T^2 used to test for $X_0' = X_0 = 0$ can be written as follows

$$T^2 = (\bar{X}')^T(M')^{-1}\bar{X}'.$$

As in Celeux (1987), the general criterion is that small values of T_r^2 (resp. T^2), suggest that the method has found the correct model. This is because, in the case that H_0 is satisfied, the chain generated by SEM will always stay around $\mathbf{0}$ thus producing low values of T_r^2, whereas in the opposite case, the chain will deviate from the starting point, thus producing high values of T_r^2.

In the sequel we describe the experiments and compare these two statistics under the null hypothesis (i.e. EM has found the MLE); the SEM algorithm always starts from the solution given by EM.

We first consider samples of sizes 50, 100 and 500 from the well-separated mixture model $1/2N(0,1) + 1/2N(4,1)$ denoted by Two-mixture(1) in Table 4.1, and from the moderately-separated mixture model $1/2N(0,1) + 1/2N(3,1)$ denoted by Two-mixture(2) in Table 4.1, where the parameters are the mixing proportions and the means. We then fit to those data a mixture of two normals, taking the true values of the parameters as starting points; in these examples, we run 3000 iterations of EM and 200 iterations of SEM.

We then consider the case where the data arise from a $N(0,1)$ distribution and we fit a mixture of two components to those data, with starting points mean1=0, mean2=3, for sample sizes 50 and 100, and mean1=0, mean2=2.75, for sample size 500, and proportions $p1 = p2 = 0.5$. For sample size $N = 50$ we run 50 EM iterations and 15 SEM iterations; for $N = 100$ we run 200 EM iterations and 30 SEM iterations, and for $N = 500$ we run 5000 EM iterations and 60 SEM iterations.

The results are reported in Table 4.1. In both cases, the EM algorithm finds the right answer, and the values of the two statistics are small and close to each other.

H_0	N=50	N=100	N=500
Normal	$T^2 = 0.867$ $T_r^2 = 0.821$	$T^2 = 0.248$ $T_r^2 = 0.192$	$T^2 = 3.652$ $T_r^2 = 3.896$
Two-mixture(1)	$T^2 = 0.143$ $T_r^2 = 0.110$	$T^2 = 9.03 \times 10^{-2}$ $T_r^2 = 7.26 \times 10^{-2}$	$T^2 = 9.48 \times 10^{-3}$ $T_r^2 = 7.52 \times 10^{-3}$
Two-mixture(2)	$T^2 = 0.102$ $T_r^2 = 7.60 \times 10^{-2}$	$T^2 = 1.57 \times 10^{-3}$ $T_r^2 = 1.26 \times 10^{-3}$	$T^2 = 0.108$ $T_r^2 = 9.64 \times 10^{-2}$

Table 4.1. Values of T^2 and T_r^2.

Table 4.1 suggests that in the case that the sample size is small, both statistics give satisfactory results. We consider now an even smaller size, for instance, the case of 25 data arising from a $N(0,1)$ distribution to which we fit a mixture of two normal distributions with starting points mean1=0, mean2=0.5, $p1 = p2 = 0.5$. Then, running 1000 EM iterations from these starting points and 37 SEM iterations, we obtain $T_r^2 = 1.214$ and $T^2 = 0.985$; the values of these statistics are small and close to each other, thus validating the result obtained by EM, which is mean1=0.338, mean2=0.338 (duplicates), $p1 = 0.492$, $p2 = 0.507$. Hence, these methods work quite well in practice even for small sizes.

It is worthwhile noting that some care is needed in choosing the number of SEM iterations: indeed, considering the previous example where $N = 25$, and starting the EM algorithm at mean1=0 and mean2=3, one cannot perform more than 6 SEM iterations, because of numerical singularities that occur when one of the proportions gets close to 0; in this case $T_r^2 = 48.077$ and

$T^2 = 72.954$, and the high values of these statistics can lead to misleading conclusions.

Therefore, one has to ensure that enough SEM iterations are run before these statistics are obtained. We should also add that this problem always occurs when the model is overfitted, independently of the sample size.

Considering again the case where (H_0) corresponds to a $N(0,1)$, let us suppose that the algorithm used to estimate the parameters (either EM or any other method) does not find the solution, so that the starting points for SEM are far from the true parameters, for example, mean1=-0.365, mean2=0.491, and weights $p1 = 0.606$, $p2 = 0.393$; the sample size considered is 500. Then, running 155 SEM iterations we obtain $T_r^2 = 12.789$ and $T^2 = 10.370$; we see that both statistics are large, suggesting that we should reject (H_0), and different from each other.

Finally, we outline the danger that one can run if SEM does not find the solution, and this is shown in the following example. We fit a mixture of two normal distributions to 500 data arising from a mixture of three normal distributions, for example, $1/3N(-2,1)+1/3N(0,1)+1/3N(2,1)$; then SEM (like EM) points wrongly towards the two-mixture model. However, in the same way as before, we run 1000 EM and 500 SEM iterations and obtain $T_r^2 = 0.030$ and $T^2 = 0.032$, so that these statistics are very close to each other and have small magnitudes, thus validating the two-mixture model.

In consequence, one has to make sure that SEM has found the right answer before using any of these tests. A way of doing this is to fit the highest-component model believed to be compatible with the data, since SEM works

well when an upper bound of the actual number of components is available, and a large sample size is used.

Chapter 5

On the Distribution of the Likelihood Ratio Test Statistic when the Mixture Proportions are known

5.1 Introduction

5.1.1 Generalities

Following Goffinet et al (1992), we study, in this chapter, the asymptotic behaviour of the Likelihood Ratio Test Statistic (LRTS) under the null hypothesis of a single-component distribution versus the hypothesis of a mixture of two components in the particular case where the mixing proportions are known a priori. In fact, for cases where the component parameters are known,

the problem simplifies considerably: the authors mention a paper by Durairajan and Kale (1982), where a local test for testing for a single-component distribution against a mixture of two components is derived. In the same context, the asymptotic distribution of the LRTS, under the null hypothesis, is derived in Titterington et al (1985), and a generalization of this result is given in Chen and Cheng (1992). (This has been mentioned in the introductory part.) For the case that we study here, the known parameters are no longer those that specify the component densities, but those specifying the mixing proportions. The general problem here, then, is to find the asymptotic distribution of the LRTS under the null hypothesis when testing between a single-component distribution and a mixture. The authors mention that the problem studied here is similar to the one faced by Aitkin and Rubin (1985), which we discussed in the introduction, and which has been theoretically established by Quinn et al (1987). Because of this similarity, we mention here the nature of that problem. Under the null hypothesis, Quinn et al (1987) proved that the Fisher information matrix has positive probability of not being positive definite, and therefore the regularity assumptions are not fulfilled for the standard asymptotic theory to apply. We shall first verify that, here too, the Fisher information matrix is singular under the null hypothesis (as done in Goffinet et al, 1992) and therefore the usual theorem mentioned in the introductory part does not apply for the LRTS to be asymptotically distributed as a $\chi^2(1)$.

We now define the problem in a more formal way, in order to state the main theorem derived by Goffinet et al (1992) concerning the asymptotic distribution of the LRTS. Let $X = (X_1, ..., X_N)$ be N independent univariate r.v.'s with common probability density function

$$h(x;\theta_1,\theta_2,\sigma) = pf(x;\theta_1,\sigma) + (1-p)f(x;\theta_2,\sigma),$$

where $f(x;\theta_i,\sigma) \sim N(\theta_i,\sigma)$ and p is assumed to be known.
We study here, under the null hypothesis H_0: $\theta_1 = \theta_2$ or equivalently $h(x;\theta_1,\theta_2,\sigma) = f(x;\theta_1,\sigma)$, the behaviour of the likelihood ratio test statistic (LRTS)

$$T(X) = 2[sup_{\theta_1,\theta_2,\sigma} \sum_{i=1}^{N} \log h(X_i;\theta_1,\theta_2,\sigma) - sup_{\theta_1,\sigma} \sum_{i=1}^{N} \log f(X_i;\theta_1,\sigma)].$$

Goffinet et al's result is as follows.

THEOREM
Under H_0, the limiting distribution of $T(X)$ is:
1) a $\chi^2(1)$ distribution if σ is unknown and $p \neq 0.5$;
2) $0.5\chi^2(0) + 0.5\chi^2(1)$ otherwise.

This theorem gives two possibilities for the asymptotic distribution of $T(X)$ depending on some assumptions on the parameters. Our main objective will be to investigate the form of the distribution of $T(X)$ in cases where $p \neq 0.5$ but p is close to 0.5, for sample sizes N which are useful in practice. We study the behaviour of $T(X)$ in those cases, theoretically and by simulation results.

In the sequel, we follow the calculations of Goffinet et al, since we use the same type of argument. We also highlight some inaccuracies in the proof of their theorem.

5.1.2 Failure of the standard regularity conditions

The first step is to show, following Goffinet et al's reasoning, that the Fisher information matrix under H_0 is singular. The statistic $T(X)$ being invariant under translation and linear transformations of X, we can take for simplicity the true value of θ_1 to be 0, and that of σ to be 1. We use the notation

$$K_{lm}(X_i; a, b) = \partial^{l+m} \log h(X_i; \theta_1, \theta_2, \sigma)/\partial a^l \partial b^m,$$

where a and b can be θ_1, θ_2, σ or some functions of these parameters. When the density of X_i is $f(x; \theta_1, \sigma)$ with $\theta_1 = 0$ and $\sigma = 1$, we obtain

$$K_1(X_i; \theta_1) = pX_i,$$

$$K_1(X_i; \theta_2) = (1-p)X_i$$

and

$$K_1(X_i; \sigma) = X_i^2 - 1.$$

Indeed, we can write

$$K_1(X_i, \theta_1) = \frac{p}{\sigma^2}(X_i - \theta_1)\frac{f(X_i; \theta_1, \sigma)}{h(X_i; \theta_1, \theta_2, \sigma)};$$

however, under H_0, $h(X_i; \theta_1, \theta_2, \sigma) = f(X_i; \theta_1, \sigma)$ so that

$$K_1(X_i; \theta_1) = \frac{p}{\sigma^2}(X_i - \theta_1).$$

Since $\theta_1 = 0$ and $\sigma = 1$, we obtain

$$K_1(X_i, \theta_1) = pX_i.$$

In the same way

$$K_1(X_i;\theta_2) = \frac{1-p}{\sigma^2}(X_i-\theta_2)\frac{f(X_i;\theta_2,\sigma)}{h(X_i;\theta_1,\theta_2,\sigma)}.$$

However, under H_0, $\theta_1 = \theta_2$, so that

$$K_1(X_i;\theta_2) = \frac{1-p}{\sigma^2}(X_i-\theta_2),$$

and, for $\theta_1 = 0$ and $\sigma = 1$, we obtain

$$K_1(X_i;\theta_2) = (1-p)X_i.$$

The same type of calculation yields $K_1(X_i;\sigma)$.

We have then

$$E(K_1(X_i;\theta_1)K_1(X_i;\theta_1)) = p^2E(X_i^2) = p^2,$$

$$E(K_1(X_i;\theta_1)K_1(X_i;\theta_2)) = p(1-p)E(X_i^2) = p(1-p),$$

$$E(K_1(X_i;\theta_1)K_1(X_i;\sigma)) = pE(X_i^3) - pE(X_i) = 0,$$

$$E(K_1(X_i;\theta_2)K_1(X_i;\sigma)) = (1-p)E(X_i^3) - (1-p)E(X_i) = 0,$$

$$E(K_1(X_i;\sigma)K_1(X_i;\sigma)) = E(X_i^4) - 2E(X_i^2) + 1 = 2,$$

$$E(K_1(X_i;\theta_2)K_1(X_i;\theta_2)) = (1-p)^2E(X_i^2) = (1-p)^2.$$

We obtain finally

$$I = \begin{pmatrix} p^2 & p(1-p) & 0 \\ p(1-p) & (1-p)^2 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

and $det(I) = 0$. Thus, the information matrix is singular and hence the classical development leading to the asymptotic distribution of the LRTS is not applicable.

5.2 Case 1: σ known

5.2.1 Main results

This case has been thoroughly investigated by Goffinet et al (1992). Nevertheless, we will give some quick proofs. We recall that the result in this case concerns the second part of their theorem.

In order to prove this, the authors made the following change of variables:

$$\begin{cases} \mu = p\theta_1 + (1-p)\theta_2 \\ \delta = \theta_1 - \theta_2. \end{cases}$$

For $\theta_1 = \theta_2 = 0$ we have $\mu = 0$ and $\delta = 0$; let us compute $K_1(X_i; \delta)$ at the point $\mu = 0$, $\delta = 0$.

From the above system of equations, we obtain

$$\begin{cases} \theta_1 = \mu + (1-p)\delta \\ \theta_2 = \mu - p\delta \end{cases}$$

and thus we obtain

$$K_1(X_i;\delta) = \partial \log h(X_i;\mu+(1-p)\delta,\mu-p\delta,\sigma)/\partial\delta.$$

After some algebra, this quantity can be written in the form $\frac{Num(\mu,\delta)}{Den(\mu,\delta)}$, where Num denotes a numerator and Den a denominator, such that, at $\mu=\delta=0$, we obtain $N(0,0)=0$ and $D(0,0)=\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2\sigma^2}X_i^2)$. Thus, at $\mu=\delta=0$, $K_1(X_i;\delta)=0$. (For reasons of clarity we do not write out Num and Den in detail.)

Hence,

$$E[K_1(X_i;\delta)K_1(X_i;\delta)]=0,$$

and

$$E[K_1(X_i,\mu)K_1(X_i;\delta)]=0,$$

so that the information matrix is singular.

Thus, as in Cox and Hinkley (1974, page 304), the authors expand the log-likelihood function that we denote, following Goffinet et al's notation, by $L(X)$ (which, here, is a function of the two variables μ and δ) up to order 4. For this, they rescale the parameters as $\tilde{\mu}=\mu N^{1/2}$ and $\tilde{\delta}=\delta N^{1/4}$. Since $\partial L(X)/\partial\delta=\sum_{i=1}^{N}K_1(X_i;\delta)=0$, and $\sum_{i=1}^{N}K_{11}(X_i;\delta,\mu)=0$, then, under H_0, $L(X)$ can be written as

$$L(X)=L(X;0,0)+\frac{1}{2}N^{-1/2}\sum_{i=1}^{N}K_2(X_i;\delta)\tilde{\delta}^2+N^{-1/2}\sum_{i=1}^{N}K_1(X_i;\mu)\tilde{\mu}$$

$$+\frac{1}{6}N^{-3/4}\sum_{i=1}^{N}K_3(X_i;\delta)\tilde{\delta}^3+\frac{1}{2}N^{-1}\sum_{i=1}^{N}K_{21}(X_i;\delta,\mu)\tilde{\delta}^2\tilde{\mu}+\frac{1}{2}N^{-1}\sum_{i=1}^{N}K_2(X_i;\delta)\tilde{\mu}^2$$

$$+\frac{1}{24}N^{-1}\sum_{i=1}^{N}K_4(X_i;\delta)\tilde{\delta}^4+o_p(1),$$

where $o_p(1)$ denotes a random variable that is $o(1)$ in probability (Cox and Hinkley, 1974, pp. 281-282); in other words, under H_0, the term $o_p(1)$ tends to 0 in probability as the sample size N goes to infinity. Goffinet et al (1992) then derive the following equalities:

$$E[K_{21}(X_i;\delta,\mu)] = -E[K_2(X_i;\delta)K_1(X_i;\mu)],$$

$$E[K_3(X_i;\delta)] = 0,$$

$$E[K_4(X_i;\delta)] = -3E[K_2(X_i;\delta)^2].$$

5.2.2 Proof of the above results

Since these are well established results, we only give an idea of the proofs. We denote, for simplicity, $h(X_i,\theta_1,\theta_2,\sigma)$ by h, the order of the partial derivatives n by (n) in the superscript, and the variables with respect to which we differentiate in the subscript in the usual way. We have then

$$K_{21}(X_i;\delta,\mu) = \frac{h^{(3)}_{\delta^2\mu}}{h} - \frac{h^{(1)}_{\mu}}{h}\frac{h^{(2)}_{\delta^2}}{h},$$

and, at $\mu = \delta = 0$, $\frac{h^{(1)}_{\mu}}{h} = K_1(X_i;\mu)$ and $\frac{h^{(2)}_{\delta^2}}{h} = K_2(X_i;\delta)$.
After some algebra, we obtain

$$\frac{h^{(3)}_{\delta^2\mu}}{h} = p(1-p)(X_i^3 - 3X_i),$$

whose expectation is 0 since $X_i \sim N(0,1)$ under H_0. The first equality follows. For the second equality, some algebra yields

$$K_3(X_i;\delta) = \frac{h^3 h^{(3)}_{\delta^3} + 2h(h^{(1)}_{\delta})^3 - 3h^2 h^{(1)}_{\delta} h^{(2)}_{\delta^2}}{h^4};$$

now, at $\mu = \delta = 0$, $h^{(1)}_{\delta} = 0$ so that

$$K_3(X_i;\delta) = \frac{h^{(3)}_{\delta^3}}{h} = p(1-p)[(3p^2 - 3(p-1)^2)X_i + ((p-1)^2 - p^2)X_i^3],$$

whose expectation is 0. The equality follows.
For the third equality,

$$K_4(X_i;\delta) = \frac{-4h^4 h^{(1)}_{\delta} h^{(2)}_{\delta^2} + 18h^3 (h^{(1)}_{\delta})^2 h^{(2)}_{\delta^2} - 3h^4 (h^{(2)}_{\delta^2})^2 + h^5 h^{(4)}_{\delta^4} - 6(h^{(1)}_{\delta})^4 h^2}{h^6},$$

and under H_0 this can be written as

$$K_4(X_i;\delta) = -\frac{3(h^{(2)}_{\delta^2})^2}{h^2} + \frac{h^{(4)}_{\delta^4}}{h};$$

however, at the point $\mu = \delta = 0$,

$$\frac{h^{(4)}_{\delta^4}}{h} = p(1-p)[-3(p-1)^3 + 3p^3 + (6(p-1)^3 - 6p^3)X_i^2 + (p^3 - (p-1)^3)X_i^4],$$

whose expectation is 0. The equality follows.
Now, the term in $\tilde{\delta}^3$ is $o_p(1)$, since its expectation is 0 and its variance is smaller than that of the terms in $\tilde{\mu}$ and $\tilde{\delta}^2$, and therefore we obtain

$$L(X) = L(X;0,0) + B(\tilde{\mu},\tilde{\delta}^2)^T + (\tilde{\mu},\tilde{\delta}^2)C(\tilde{\mu},\tilde{\delta}^2)^T + o_p(1),$$

where $B = (N^{-1/2}\sum_{i=1}^{N} K_1(X_i;\mu), \frac{1}{2}N^{-1/2}\sum_{i=1}^{N} K_2(X_i;\delta))^T$,

$$C = \begin{pmatrix} \frac{1}{2}N^{-1}\sum_{i=1}^{N} K_2(X_i;\mu) & \frac{1}{4}N^{-1}\sum_{i=1}^{N} K_{21}(X_i;\delta,\mu) \\ \frac{1}{4}N^{-1}\sum_{i=1}^{N} K_{21}(X_i;\delta,\mu) & \frac{1}{24}N^{-1}\sum_{i=1}^{N} K_4(X_i;\delta) \end{pmatrix}$$

and

$$E(C) = -\frac{1}{2}E(B^T B);$$

this result is a direct consequence of the three equalities derived previously. Furthermore, it is straightforward to verify that

$$E(B) = \mathbf{0}.$$

Remarks:

1) The term in $\tilde{\delta}^2\tilde{\mu}$ can also be considered as $o_p(1)$, in the same sense as for the term in $\tilde{\delta}^3$, since $E[K_{21}(X_i;\delta,\mu)] = 0$, but this should not affect the result since this term is not on the diagonal of the matrix C. The same argument as in Cox and Hinkley (1974, page 321), with the constraint $\delta^2 \geq 0$, leads then to the second result of Goffinet et al's theorem.

2) The authors note that in the case $p = \frac{1}{2}$, the same result can be obtained more simply, using the fact that $K_3(X_i;\delta) = 0$ and applying the general results (Case 5) of Self and Liang (1987).

5.3 Case 2: σ unknown and $p = 0.5$

5.3.1 Some theoretical argument

Here, Goffinet et al (1992) obtain the same limiting distribution for $T(X)$, as for case 1.

We shall go through the calculations to show first that the change of variables introduced in the paper is not adequate. For $p = \frac{1}{2}$, we have

$$\begin{cases} \theta_1 = \mu + \frac{\delta}{2} \\ \theta_2 = \mu - \frac{\delta}{2}, \end{cases}$$

so that

$$h = \frac{1}{2\sigma\sqrt{2\pi}} \exp(-\frac{(X_i - (\mu + \frac{\delta}{2}))^2}{2\sigma^2}) + \frac{1}{2\sigma\sqrt{2\pi}} \exp(-\frac{(X_i - (\mu - \frac{\delta}{2}))^2}{2\sigma^2}).$$

In the same way as before, $K_1(X_i; \delta) = 0$ at the point $\mu = 0$, $\delta = 0$, and

$$h^{(2)}_{\delta^2} = \frac{1}{4\sqrt{2\pi}}(X_i^2 - 1) \exp(-\frac{1}{2}X_i^2),$$

so that

$$K_2(X_i; \delta) = \frac{h^{(2)}_{\delta^2}}{h} = \frac{1}{4}(X_i^2 - 1).$$

On the other hand,

$$K_3(X_i; \delta) = \frac{1}{4}(-\frac{3}{4}X_i + \frac{1}{4}X_i^3 + \frac{3}{4}X_i - \frac{1}{4}X_i^3) = 0$$

under H_0, after some algebra; in the same way, under H_0,

$$K_5(X_i;\delta) = \frac{h^{(5)}_{\delta^5}}{h} = \frac{15}{16}X_i - \frac{10}{16}X_i^3 + \frac{1}{16}X_i^5 - \frac{15}{16}X_i + \frac{10}{16}X_i^3 - \frac{1}{16}X_i^5 = 0.$$

Indeed, in general for any k odd,

$$K_k(X_i;\delta) = 0,$$

and this result is proved in Goffinet et al (1992). The change of variables used here is $\epsilon = \frac{1}{4}\delta^2$ instead of δ, since all terms in δ are in powers of δ^2. Denoting $\frac{1}{\sigma}$ by ω, we can prove now that

$$K_1(X_i;\epsilon) = -\frac{1}{2}K_1(X_i;\omega).$$

Note that in their paper, Goffinet et al (1992) do not find exactly the same equality; they find instead, $K_1(X_i;\epsilon) = -K_1(X_i;\omega)$.
We first prove now that, again, the Fisher information matrix is singular. Indeed, from the above change of variables, we obtain

$$\begin{cases} \theta_1 = \mu + \sqrt{\epsilon} \\ \theta_2 = \mu - \sqrt{\epsilon}, \end{cases}$$

so that h becomes

$$h = \frac{1}{2\sqrt{2\pi}\sigma}\exp(-\frac{1}{2\sigma^2}(X_i - (\mu+\sqrt{\epsilon}))^2 + \frac{1}{2\sqrt{2\pi}\sigma}\exp(-\frac{1}{2\sigma^2}(X_i - (\mu-\sqrt{\epsilon}))^2.$$

Taking $\sigma = 1$, for simplicity, we then compute $K_1(X_i;\epsilon)$; this type of calculation is similar to the calculations done previously in this chapter, and

yields

$$K_1(X_i;\epsilon) = \frac{h_\epsilon^{(1)}}{h} = \frac{1}{2}(X_i^2 - 1),$$

for $\mu = \epsilon = 0$.

On the other hand, by replacing σ by $\frac{1}{\omega}$ in the expression for h, we obtain

$$h = \frac{1}{2\sqrt{2\pi}}\omega \exp[-\frac{1}{2}\omega^2(X_i-(\mu+\sqrt{\epsilon}))^2] + \frac{1}{2\sqrt{2\pi}}\omega \exp[-\frac{1}{2}\omega^2(X_i-(\mu-\sqrt{\epsilon}))^2],$$

so that

$$K_1(X_i;\omega) = \frac{h_\omega^{(1)}}{h} = 1 - X_i^2$$

at $\mu = \epsilon = 0$, $\omega = 1$.

We then derive the information matrix

$$I = \begin{pmatrix} E[K_1(X_i;\mu)^2] & E[K_1(X_i;\mu)K_1(X_i;\epsilon)] & E[K_1(X_i;\mu)K_1(X_i;\omega)] \\ E[K_1(X_i;\epsilon)K_1(X_i;\mu)] & E[K_1(X_i;\epsilon)^2] & E[K_1(X_i;\epsilon)K_1(X_i;\omega)] \\ E[K_1(X_i;\omega)K_1(X_i;\mu)] & E[K_1(X_i;\omega)K_1(X_i;\epsilon)] & E[K_1(X_i;\omega)^2] \end{pmatrix}$$

$$= \begin{pmatrix} E[K_1(X_i;\mu)^2] & -\frac{1}{2}E[K_1(X_i;\mu)K_1(X_i;\omega)] & E[K_1(X_i;\mu)K_1(X_i;\omega)] \\ -\frac{1}{2}E[K_1(X_i;\omega)K_1(X_i;\mu)] & \frac{1}{4}E[K_1(X_i;\omega)^2] & -\frac{1}{2}E[K_1(X_i;\omega)^2] \\ E[K_1(X_i;\omega)K_1(X_i;\mu)] & -\frac{1}{2}E[K_1(X_i;\omega)^2] & E[K_1(X_i;\omega)^2] \end{pmatrix}.$$

The determinant of I is found to be 0, so that I is singular.

5.3.2 Solving the problem

A new variable ϕ is then introduced by the authors, using the change of variable $\phi = (\omega-1)-\epsilon$, and the parameters are rescaled as $\tilde{\mu} = \mu N^{1/2}$, $\tilde{\phi} = \phi N^{1/2}$, $\tilde{\omega} = (\omega - 1)N^{1/4}$. We now show that this change of variable does not work properly. Indeed, we obtain the following system of equations:

$$\begin{cases} \theta_1 = \mu + \sqrt{\omega - 1 - \phi} \\ \theta_2 = \mu - \sqrt{\omega - 1 - \phi}, \end{cases}$$

so that

$$h = \frac{1}{2\sqrt{2\pi}}\omega \exp[-\frac{1}{2}(X_i - \sqrt{\omega - 1 - \phi} - \mu)^2] + \frac{1}{2\sqrt{2\pi}}\omega \exp[-\frac{1}{2}(X_i + \sqrt{\omega - 1 - \phi} - \mu)^2].$$

We now prove that the random variable $K_1(X_i;\phi) \neq 0$, and $E[K_1(X_i;\phi)] = 0$. We have that $K_1(X_i;\phi) = \frac{h_\phi^{(1)}}{h}$; using then a Taylor expansion around 0, we find that

$$h_\phi^{(1)} = \frac{1}{\sqrt{2\pi}\sigma}(1 - X_i^2)\exp(-\frac{1}{2}X_i^2),$$

and, on the other hand,

$$h = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2}X_i^2);$$

it results that, under H_0, $K_1(X_i;\phi) = \frac{1-X_i^2}{2} = -K_1(X_i;\epsilon)$, and $E[K_1(X_i;\phi)] = 0$. This is related in fact, to one of the problems that we mentioned in the beginning of this chapter, namely that, in their paper, Goffinet et al (1992) find $K_1(X_i;\phi)$ to be 0, which is not obviously the case according to our calculations. We, then, cannot apply the same arguments as in Case 1 to infer about the distribution of $T(X)$. Therefore, the change of variables used by them is not very interesting. Our proposal is to have μ, δ and ω as variables. Now, since $K_1(X_i;\delta) = 0$, as in the previous case, we consider the expansion of $L(X)$ up to order 4, and the only difference from Case 1 is that we have a third variable ω to take into account in the expansion; it is clear that the fact that $K_k(X_i;\delta) = 0$ for k odd, will not affect the expansion $L(X)$ of case 1. We then rescale to $\tilde{\mu} = \mu N^{1/2}$, $\tilde{\delta} = \delta N^{1/4}$, and $\tilde{\omega} = (\omega - 1)N^{1/2}$, and verify

that the information matrix I is singular. Indeed

$$I = \begin{pmatrix} E[K_1(X_i;\mu)^2] & E[K_1(X_i;\mu)K_1(X_i;\delta)] & E[K_1(X_i;\mu)K_1(X_i;\omega)] \\ E[K_1(X_i;\delta)K_1(X_i;\mu)] & E[K_1(X_i;\delta)^2] & E[K_1(X_i;\delta)K_1(X_i;\omega)] \\ E[K_1(X_i;\omega)K_1(X_i;\mu)] & E[K_1(X_i;\omega)K_1(X_i;\delta)] & E[K_1(X_i;\omega)^2] \end{pmatrix}.$$

At $\mu = \delta = 0$, $\omega = 1$ we obtain, after some algebra,

$$K_1(X_i;\mu) = X_i,$$

and we know from previous calculation that $K_1(X_i;\omega) = 1 - X_i^2$; it results that

$$E[K_1(X_i;\mu)K_1(X_i;\omega)] = E(X_i - X_i^3) = 0,$$

$$E[(K_1(X_i;\delta)K_1(X_i;\omega)] = 0,$$

since $K_1(X_i;\delta) = 0$, and

$$E[K_1(X_i;\delta)K_1(X_i;\mu)] = 0$$

for the same reason. Thus the determinant of I is 0. Again $L(X)$ will have to be expanded up to order 4; we then obtain

$$L(X) = L(X;0,0,1) + N^{-1/2}\sum_{i=1}^{N} K_1(X_i;\mu)\tilde{\mu} + N^{-1/2}\sum_{i=1}^{N} K_1(X_i;\omega)\tilde{\omega}$$

$$+\frac{1}{2}N^{-1/2}\sum_{i=1}^{N} K_2(X_i;\delta)\tilde{\delta}^2 + \frac{1}{2}N^{-1}\sum_{i=1}^{N} K_2(X_i;\mu)\tilde{\mu}^2 + \frac{1}{2}N^{-1}\sum_{i=1}^{N} K_2(X_i;\omega)\tilde{\omega}^2$$

$$+N^{-1}\sum_{i=1}^{N} K_{11}(X_i;\mu,\omega)\tilde{\mu}\tilde{\omega} + \frac{1}{2}N^{-1}\sum_{i=1}^{N} K_{21}(X_i;\delta,\mu)\tilde{\delta}^2\tilde{\mu} + \frac{1}{2}N^{-1}\sum_{i=1}^{N} K_{21}(X_i;\delta,\omega)\tilde{\delta}^2\tilde{\omega}$$

$$+\frac{1}{24}N^{-1}\sum_{i=1}^{N}K_4(X_i;\delta)\tilde{\delta}^4+o_p(1).$$

Thus,

$$L(X)=L(X;0,0,1)+B(\tilde{\mu},\tilde{\delta}^2,\tilde{\omega})+(\tilde{\mu},\tilde{\delta}^2,\tilde{\omega})C(\tilde{\mu},\tilde{\delta}^2,\tilde{\omega})^T+o_p(1),$$

where $B=(N^{-1/2}\sum_{i=1}^{N}K_1(X_i;\mu),\frac{1}{2}N^{-1/2}\sum_{i=1}^{N}K_2(X_i;\delta),N^{-1/2}\sum_{i=1}^{N}K_1(X_i;\omega))^T$, and

$$C=\begin{pmatrix}\frac{1}{2}N^{-1}\sum_{i=1}^{N}K_2(X_i;\mu) & \frac{1}{4}N^{-1}\sum_{i=1}^{N}K_{21}(X_i;\delta,\mu) & \frac{1}{2}N^{-1}\sum_{i=1}^{N}K_{11}(X_i;\mu,\omega)\\ \frac{1}{4}\sum_{i=1}^{N}K_{21}(X_i;\delta,\mu) & \frac{1}{24}N^{-1}\sum_{i=1}^{N}K_4(X_i;\delta) & \frac{1}{4}N^{-1}\sum_{i=1}^{N}K_{21}(X_i;\delta,\omega)\\ \frac{1}{2}N^{-1}\sum_{i=1}^{N}K_{11}(X_i;\mu,\omega) & \frac{1}{4}N^{-1}\sum_{i=1}^{N}K_{21}(X_i;\delta,\omega) & \frac{1}{2}N^{-1}\sum_{i=1}^{N}K_2(X_i;\omega)\end{pmatrix}.$$

As in Case 1, we have that $E(C)=-\frac{1}{2}E(B^TB)$. This is because we have

$$E[K_{11}(X_i;\mu,\omega)]=0=E[K_1(X_i;\mu)K_1(X_i;\omega)],$$

$$E[K_{21}(X_i;\delta,\mu)]=0=-E[K_2(X_i;\delta)K_1(X_i;\mu)],$$

$$E[K_{21}(X_i;\delta,\omega)]=-E[K_2(X_i;\delta)K_1(X_i;\omega)]$$

(since $K_{21}(X_i;\delta,\omega)=\frac{h^{(3)}_{\delta^2\omega}}{h}-\frac{h^{(1)}_{\omega}}{h}\frac{h^{(2)}_{\delta^2}}{h}$, $E(\frac{h^{(3)}_{\delta^2\omega}}{h})=0$, $K_1(X_i;\omega)=\frac{h^{(1)}_{\omega}}{h}$ and $K_2(X_i;\delta^2)=\frac{h^{(2)}_{\delta^2}}{h}$), and, as in the previous case,

$$E[K_4(X_i;\delta)]=-3E[K_2(X_i;\delta)^2];$$

also

$$E[K_2(X_i;\omega)]=-E[K_1(X_i;\omega)^2].$$

Furthermore the expectation of the score vector B is $\mathbf{0}$. The conclusion will

then be the same as for Case 1.

(Remark: we verify that all terms in $L(X)$ are non-zero).

Very recent information that we obtained from Goffinet confirmed our feelings about the change of variable used in Goffinet et al's paper. Indeed, the correct change of variable would be, in that case, to replace in their paper ϵ by $\epsilon/2$ so that we obtain $\phi = (\omega - 1) - \epsilon/2$, yielding $K_1(X_i; \omega) = 0$; (the other parameters are left unchanged). Then, Goffinet et al's theorem is proved using the same argument as for Case 1.

5.4 Case 3: σ unknown and $p \neq 0.5$

5.4.1 Main results

The result in this case concerns the first part of their theorem. In this case, $K_2(X_i; \delta)$ is proportional to $K_1(X_i; \omega)$; indeed, using the same kind of calculation as in the previous cases yields

$$K_1(X_i; \omega) = 1 - X_i^2$$

at $\mu = \delta = 0$ and $\omega = 1$, and, on the other hand,

$$K_2(X_i; \delta) = -p(1-p)(1 - X_i^2) = -p(1-p)K_1(X_i; \omega).$$

Thus, Goffinet et al (1992) consider the following parameters: μ, $\delta = (p - p^2)^{\frac{1}{2}}(\theta_1 - \theta_2)$, $\epsilon = (\omega - 1) - \delta^2$. Then, for the purpose of expanding $L(X)$, they rescale the parameters as $\tilde{\mu} = \mu N^{1/2}$, $\tilde{\delta} = \delta N^{1/6}$, and $\tilde{\epsilon} = \epsilon N^{1/2}$.

We first show that, again, there is a similar problem here as for the previous case, concerning an inaccuracy in Goffinet et al's paper, in the sense that this change of variable does not lead to the result. Indeed, we compute $K_2(X_i;\delta) = \frac{h^{(2)}_{\delta^2}}{h}$; we find it to be equal to $1 - X_i^2$ at $\mu = \delta = \epsilon = 0$. This contradicts Goffinet et al's result stating that $K_2(X_i;\delta) = 0$. This is an important problem since $K_2(X_i;\delta)$ plays a crucial role in the expansion of the log-likelihood function. However, as for the previous case, recent information obtained from Goffinet is that the correct change of variable comes by replacing δ^2 by $\delta^2/2$ in the expression of ϵ so that we obtain $\epsilon = (\omega - 1) - \delta^2/2$ and leave the other parameters unchanged. With this change of variable, it becomes then true that $K_2(X_i;\delta) = 0$.

We give now the basic relationships concerning this case (these are valid when using the new change of variable that we reported hereabove):

$$K_1(X_i;\delta) = K_2(X_i;\delta) = 0,$$

$$E[K_6(X_i;\delta)] = -10E[K_3(X_i;\delta)^2],$$

$$E[K_4(X_i;\delta)] = E[K_5(X_i;\delta)] = E[K_{21}(X_i;\delta,\xi)] = 0,$$

$$E[K_{31}(X_i;\delta,\xi)] = -E[K_3(X_i;\delta)K_1(X_i;\xi)],$$

where ξ stands for μ or ϵ.

The main result of Goffinet et al (1992) in this case is the following.

$L(X) = L(X;0,0,0) + N^{-1/2}\sum_{i=1}^N K_1(X_i;\mu)\tilde{\mu} + N^{-1/2}\sum_{i=1}^N K_1(X_i;\epsilon)\tilde{\epsilon}$
$+\frac{1}{2}N^{-1}\sum_{i=1}^N K_2(X_i;\mu)\tilde{\mu}^2 + \frac{1}{2}N^{-1}\sum_{i=1}^N K_2(X_i;\epsilon)\tilde{\epsilon}^2 + N^{-1}\sum_{i=1}^N K_{11}(X_i;\mu,\epsilon)\tilde{\mu}\tilde{\epsilon}$
$+\frac{1}{6}N^{-1/2}\sum_{i=1}^N K_3(X_i;\delta)\tilde{\delta}^3 + \frac{1}{6}N^{-1}\sum_{i=1}^N K_{31}(X_i;\delta,\mu)\tilde{\mu}\tilde{\delta}^3 + \frac{1}{6}N^{-1}\sum_{i=1}^N K_{31}(X_i;\delta,\epsilon)\tilde{\epsilon}\tilde{\delta}^3$

$+\frac{1}{720}N^{-1}\sum_{i=1}^{N}K_6(X_i;\delta)\tilde{\delta}^6 + o_p(1)$.

Some terms do not figure in this expansion; they are considered as $o_p(1)$ for the same reasons as for Case 1.

Thus, $L(X)$ can be written as

$$L(X) = L(X;0,0,0) + B(\tilde{\mu},\tilde{\delta}^3,\tilde{\epsilon})^T + (\tilde{\mu},\tilde{\delta}^3,\tilde{\epsilon})C(\tilde{\mu},\tilde{\delta}^3,\tilde{\epsilon})^T + o_p(1),$$

where $B = (N^{-1/2}\sum_{i=1}^{N}K_1(X_i;\mu), \frac{1}{6}N^{-1/2}\sum_{i=1}^{N}K_3(X_i;\delta), N^{-1/2}\sum_{i=1}^{N}K_1(X_i;\epsilon))$, and

$$C = \begin{pmatrix} \frac{1}{2}N^{-1}\sum_{i=1}^{N}K_2(X_i;\mu) & \frac{1}{12}N^{-1}\sum_{i=1}^{N}K_{31}(X_i;\delta,\mu) & \frac{1}{2}N^{-1}\sum_{i=1}^{N}K_{11}(X_i;\mu,\epsilon) \\ \frac{1}{12}N^{-1}\sum_{i=1}^{N}K_{31}(X_i;\delta,\mu) & \frac{1}{720}N^{-1}\sum_{i=1}^{N}K_6(X_i;\delta) & \frac{1}{12}N^{-1}\sum_{i=1}^{N}K_{31}(X_i;\delta,\epsilon) \\ \frac{1}{2}N^{-1}\sum_{i=1}^{N}K_{11}(X_i;\mu,\epsilon) & \frac{1}{12}N^{-1}\sum_{i=1}^{N}K_{31}(X_i;\delta,\epsilon) & \frac{1}{2}N^{-1}\sum_{i=1}^{N}K_2(X_i;\epsilon) \end{pmatrix}$$

satisfy the relationship $E(C) = -\frac{1}{2}E(B^TB)$, and $E(B) = \mathbf{0}$. The result then follows by the classical development of Cox and Hinkley (1974, pp. 313-314).

5.4.2 Our results for p close to 0.5

In their paper, Goffinet et al (1992) point out that, in view of their simulation results, the convergence of the estimated means of $T(X)$ to the theoretical ones is very slow in the case $p = 0.51$. We also find the same behaviour (see next section), and we give a heuristic explanation for it. Since, for $p = \frac{1}{2}$, $K_k(X_i;\delta) = 0$ for k odd, which is a key factor underlying the asymptotic distribution of $T(X)$, we shall investigate here the case where p is close to $\frac{1}{2}$, that is, $p = \frac{1}{2} + \Delta$, where Δ is small compared to $\frac{1}{2}$, and we shall calculate $K_k(X_i;\delta)$. We shall find $K_k(X_i;\delta) = O_p(\Delta)$, and this result may account for the slow convergence.

From the expressions of μ and δ, we obtain

$$\theta_1 = \mu + (\frac{1-p}{p})^{1/2}\delta,$$

and

$$\theta_2 = \mu - (\frac{p}{1-p})^{1/2}\delta;$$

now, replacing p by $\frac{1}{2} + \Delta$, we can use the Taylor expansion to approximate these expressions, for small Δ, by

$$\begin{cases} \theta_1 = \mu + (1-2\Delta)\delta \\ \theta_2 = \mu - (1+2\Delta)\delta. \end{cases}$$

We verify, thus, that θ_1 and θ_2 are as in Case 2 for $\Delta \to 0$ (also, note that, for $\Delta \to 0$, the δ used here corresponds to taking 2δ in Case 2); hence, we verify that, by letting Δ go to 0, we obtain the same parameters as for Case 2.
Let us now consider the general case

$$\begin{cases} \theta_1 = \mu + (\frac{1-p}{p})^{1/2}\delta \\ \theta_2 = \mu - (\frac{p}{1-p})^{1/2}\delta, \end{cases}$$

for which h becomes

$$h = \frac{p}{\sqrt{2\pi}}(\epsilon + \frac{\delta^2}{2} + 1)\exp[-\frac{1}{2}((\epsilon + \frac{\delta^2}{2} + 1)(X_i - (\mu + (\frac{1-p}{p})^{1/2}\delta)))^2]$$

$$+\frac{1-p}{\sqrt{2\pi}}(\epsilon + \frac{\delta^2}{2} + 1)\exp[-\frac{1}{2}((\epsilon + \frac{\delta^2}{2} + 1)(X_i - (\mu - (\frac{p}{1-p})^{1/2}\delta)))^2],$$

and we are more particularly interested in the behaviour of $K_k(X_i;\delta)$ for k odd.

At $\mu = \delta = \epsilon = 0$, some algebra shows that

$$K_1(X_i;\delta) = p(\frac{1-p}{p})^{1/2} - (1-p)(\frac{p}{1-p})^{1/2} = 0.$$

This is a general result, valid whatever the value of p. We now compute $K_3(X_i;\delta)$ at $\mu = \delta = \epsilon = 0$. Some algebra shows that this can be written as

$$K_3(X_i;\delta) = \frac{h^{(3)}_{\delta^3}}{h} = \frac{(2p-1)}{p^{1/2}(1-p)^{1/2}}(3X_i - X_i^3).$$

For $p = \frac{1}{2}$, we obtain

$$K_3(X_i;\delta) = 0,$$

and $K_3(X_i;\delta) \neq 0$ for $p \neq \frac{1}{2}$. For $p = \frac{1}{2} + \Delta$,

$$K_3(X_i;\delta) = \frac{2\Delta}{(\frac{1}{4} - \Delta^2)^{1/2}}(3X_i - X_i^3).$$

Since Δ is small, using a Taylor expansion of order 1, we obtain

$$K_3(X_i;\delta) \approx 12\Delta X_i - 4\Delta X_i^3.$$

On the other hand,

$$E[K_3(X_i;\delta)] = 0,$$

which is true for any p, and

$$var[K_3(X_i;\delta)] \approx 96\Delta^2,$$

thus yielding that the standard deviation $\sigma[K_3(X_i;\delta)] \approx 4\sqrt{6}\Delta$; it results that

$$K_3(X_i;\delta) = O_p(\Delta).$$

We now turn to $K_5(X_i;\delta)$. Some algebra yields

$$K_5(X_i;\delta) = 0,$$

at $\mu = \delta = \epsilon = 0$, for $p = \frac{1}{2}$, and the same kind of properties as for $K_3(X_i;\delta)$ apply here, yielding, for p close to $\frac{1}{2}$,

$$K_5(X_i;\delta) = O_p(\Delta).$$

Thus, a property similar to the one for Case 2 is found: for k odd, $K_k(X_i;\delta) = 0$ for Case 3 whenever we take $p = \frac{1}{2}$, and, for any small enough Δ,

$$K_k(X_i;\delta) = O_p(\Delta).$$

We make clear that terms that are $O_p(\Delta)$ are larger than terms that are $o_p(1)$ since the latter tend to 0 in probability as the sample size N increases. Thus the asymptotic result holds even for Δ small.

Now, the more Δ increases or equivalently p moves away from $\frac{1}{2}$, the faster the distribution of $T(X)$ approaches that of Case 3. The problem certainly is that, for p close to $\frac{1}{2}$, we may need a huge sample size before the theoretical result on $T(X)$ is useful in practice. In the next section, we shall investigate, using simulations, the form of the distribution of $T(X)$ for reasonable sample sizes.

5.5 Simulation results

We present now some simulation results for cases 2 and 3 (where σ is unknown), in order to investigate in practice the difference in the distribution of $T(X)$, according to how close p is to 0.5; Table 5.1 presents results for the distribution of $T(X)$ under H_0, estimated from 500 simulations for different values of p and N.

We emphasise the result that comes out of our study: we find that, in the cases where p is in the neighbourhood of 0.5 and where $p = 0.5$, the values of the respective estimated expectations of $T(X)$ are very close to each other, and, in the same way, the values of the respective estimated variances of $T(X)$ are very close to each other, independently of the size N. This is natural since $K_k(X_i; \delta) = O_p(\Delta)$ for k odd, independently of N. Since, for p close to 0.5, these values are far from the theoretical values that we expected, and for $p = 0.5$, they are close to the theoretical values that we expected, we verify practically the validation of the Goffinet et al (1992) theorem, for the latter case, and the fact of very slow convergence to the theoretical results in the former case.

We notice that, as N increases, the expectations of $T(X)$ in these two cases tend to 0.5 (for $p = 0.5$) and to a value close to 0.5 (for p close to 0.5); similarly, the variances of $T(X)$ tend to 1.25 (for $p = 0.5$), and to a value close to 1.25 (for p close to 0.5); here again this is because $K_k(X_i; \delta) = O_p(\Delta)$, so that the smaller the Δ, the closer to each other are the two first moments in these two cases. (Certainly, for huge N, this would not be the case any more.) It results that, for values of p close to 0.5, the distribution of $T(X)$ is close

to the distribution of $T(X)$ for $p = 0.5$, and that, the more Δ increases, the more the distribution of $T(X)$ tends asymptotically faster towards a $\chi^2(1)$. Table 5.1 shows the results for three characteristic values of p, used also by Goffinet et al (1992) in their paper, but we performed some more simulations for other values of p, in order to investigate further the intermediate (or transition) stages of the distribution of $T(X)$. The results are given in Table 5.2, where a sample size of 3000 is used; for $p < 0.51$, expectation and variance are closer to the respective expectation and variance for case $p = 0.5$, than to those for case $p = 0.51$; in the case where p increases from 0.501 to 0.65, the expectation and variance of $T(X)$ approach 1 and 2 respectively. Furthermore, in practice, it seems that, for $p \approx 0.65$, the distribution of $T(X)$ is already a $\chi^2(1)$.

Remark about the rate of convergence: we note that, for $p = 0.75$ (which is far from 0.5), a sample size of 500 only is needed to obtain that the estimated expectation of $T(X)$ is close to 1, whereas for $p = 0.5$ a sample of size 3000 has to be considered in order to obtain that the estimated expectation of $T(X)$ gets close to 0.5. As for p close to 0.5, even for a sample size of 3000, the distribution of $T(X)$ is far away from the theoretical one. Therefore, there is quite fast convergence towards the theoretical results in the case $p = 0.75$, a medium convergence in the case $p = 0.5$ and an excruciatingly slow convergence in the case p close to 0.5.

We now draw some histograms of the data based on the 500 simulations that we carried out (Figures 5.1-5.8), for some values of p and for various sample sizes, in order to visualise the distribution of $T(X)$. We are certainly interested in the cases described in Table 5.1, but we focus even more in

those of Table 5.2, since, for p roughly between 0.5 and 0.60, the distribution of $T(X)$ is clearly not a $\chi^2(1)$. We actually find, using our results, that, if the expectation of $T(X)$ is π, then its variance is close to $\pi(3-\pi)$, for a sufficiently large sample (see Table 5.3): these characteristics are those of a $(1-\pi)\chi^2(0)+\pi\chi^2(1)$ distribution.

Figures 5.1 and 5.2 represent data for the case where $p = 0.75$ and for sample sizes $N = 100$ and $N = 500$. Figure 5.3 deals with the case $p = 0.51$ and $N = 3000$. Figure 5.4 deals with the case $p = 0.5$ and $N = 3000$. Figures 5.5, 5.6 and 5.7 deal with the intermediate situations $p = 0.52$, $p = 0.55$, $p = 0.6$ and $N = 3000$ respectively. Figure 5.8 deals with the case where $p = 0.65$ and $N = 3000$. In Figures 5.1 and 5.2 one can roughly recognise the shape of a $\chi^2(1)$ distribution density, whereas, in Figures 5.3, 5.4 and 5.5, there is on the one hand a concentration of values around 0, and on the other hand the $\chi^2(1)$ distribution density: this corresponds roughly to a distribution of $T(X)$ of $\frac{1}{2}\chi^2(0)+\frac{1}{2}\chi^2(1)$. It is obvious from Figures 5.6 and 5.7 that the distribution is not a $\chi^2(1)$ in the corresponding cases (though, in the case of the latter, the distribution is close to a $\chi^2(1)$). Finally, Figure 5.8 shows clear features of a $\chi^2(1)$.

Goffinet et al (1992) choose the number of zeroes to be the number of results smaller than 10^{-9}; this is certainly an ad-hoc approach. We shall proceed in a more objective way: from the data, we calculate 27 α-significance levels; we then plot these estimated levels against the 27 theoretical levels under the $\chi^2(1)$ distribution. The cases considered are those defined previously as cases 2 and 3, corresponding respectively to $p = 0.5$ with σ unknown, and $p \neq 0.5$ with σ unknown. In theory, we should obtain a straight line through the data

covering the whole range of values on the X-axis and on the Y-axis, that is, from $(0,0)$ to $(1,1)$, for the case where $p \neq 0.5$; this is shown in Figures 5.9, 5.10, 5.11, and 5.19 corresponding to values of p far from 0.5; for $p = 0.75$, the line obtained is fairly straight in every one of these plots, including the case where the sample size is only 25. On the other hand, in the case where $p = 0.5$, we should obtain a straight line through the data from $(0,0)$ to $(1,0.5)$, and a vertical line from $(1,0.5)$ to $(1,1)$; this is shown in Figures 5.12 and 5.13. For cases where p is close to 0.5, we do not obtain the theoretical result, that is, the same type of plots as those for $p \neq 0.5$ described hereabove, because of the very slow convergence mentioned previously; we obtain instead the intermediate situations described in Table 5.2, and shown in Figures 5.14, 5.15, 5.16, 5.17 and 5.18, which, then, confirm the conclusion obtained from the histograms and from Table 5.3: since in these cases we obtain a straight line through the data from (0,0) to $(1,\pi)$, and a vertical line from $(1,\pi)$ to (1,1), the distribution of $T(X)$ is of the form

$$(1-\pi)\chi^2(0) + \pi\chi^2(1),$$

and this is certainly true for all cases; for instance, in the case $p = 0.55$ and sample size $N = 3000$, the distribution of $T(X)$ is $0.23\chi^2(0) + 0.77\chi^2(1)$, and the proportion of zeroes is 23 per cent. For practical reasons, this result is then more useful than the theoretical one.

p	N	Th. expect.	Est. expect	Th. var.	Est. var.
0.75	25	1	1.50	2	4.28
	100	1	1.22	2	2.66
	500	1	1.04	2	2.23
0.51	25	1	0.92	2	3.20
	100	1	0.68	2	1.58
	500	1	0.68	2	1.92
	3000	1	0.59	2	1.51
0.50	25	0.5	0.88	1.25	3.09
	100	0.5	0.64	1.25	1.45
	500	0.5	0.64	1.25	1.79
	3000	0.5	0.54	1.25	1.39

Table 5.1. Characteristics of $T(X)$ under H_0 in cases 2 and 3.

p	N	Est. expect.	Est. var.
0.501	3000	0.55	1.40
0.505	3000	0.57	1.45
0.52	3000	0.63	1.59
0.55	3000	0.77	1.85
0.60	3000	0.91	2.03
0.65	3000	1.01	1.97

Table 5.2. Characteristics of $T(X)$ under H_0 for the transition stages between case 2 and case 3.

p	Est. var.	$\pi(3-\pi)$
0.501	1.40	1.35
0.505	1.45	1.39
0.51	1.51	1.42
0.52	1.59	1.49
0.55	1.85	1.73
0.60	2.03	1.90
0.65	1.97	2.01

Table 5.3. Variance of $(1-\pi)\chi^2(0) + \pi\chi^2(1)$. The sample size is 3000.

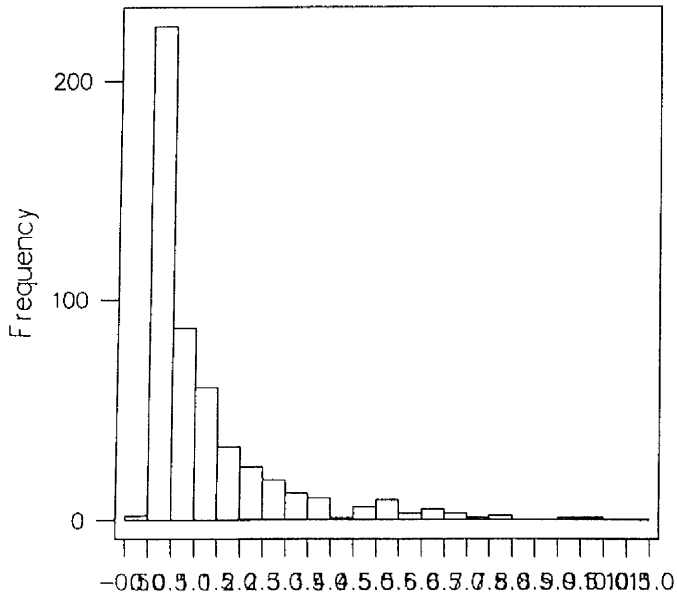

Figure 5.1. Histogram for $p = 0.75$ and $N = 100$.

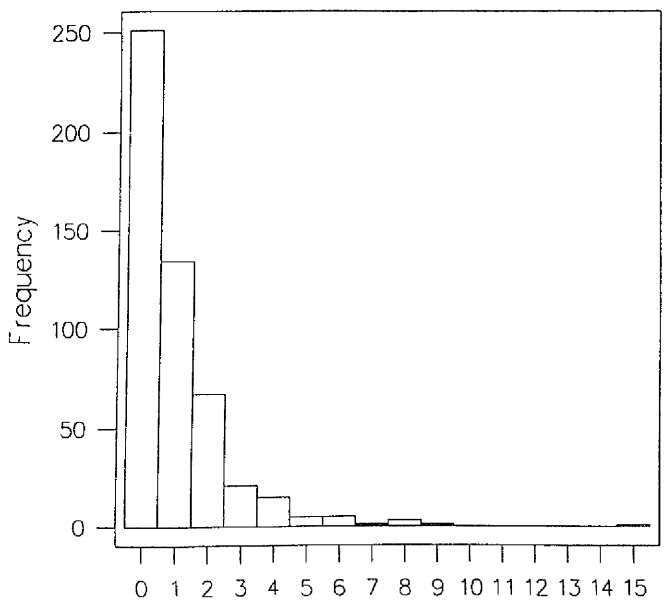

Figure 5.2. Histogram for $p = 0.75$ and $N = 500$.

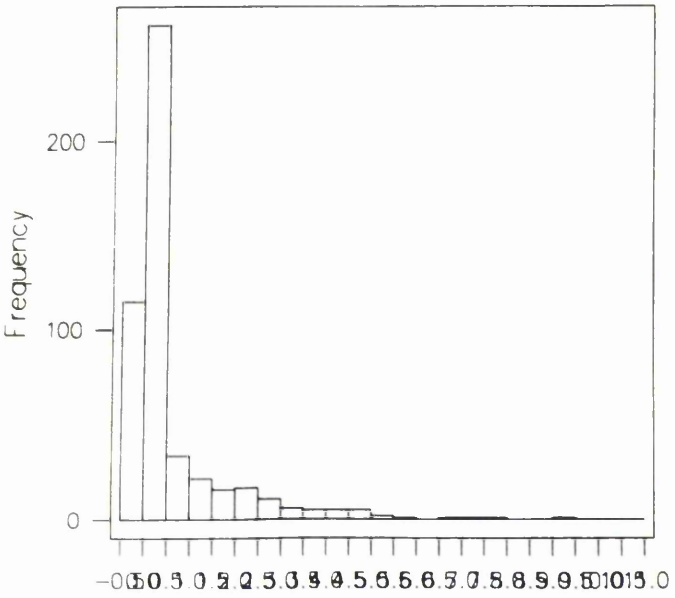

Figure 5.3. Histogram for $p = 0.51$ and $N = 3000$.

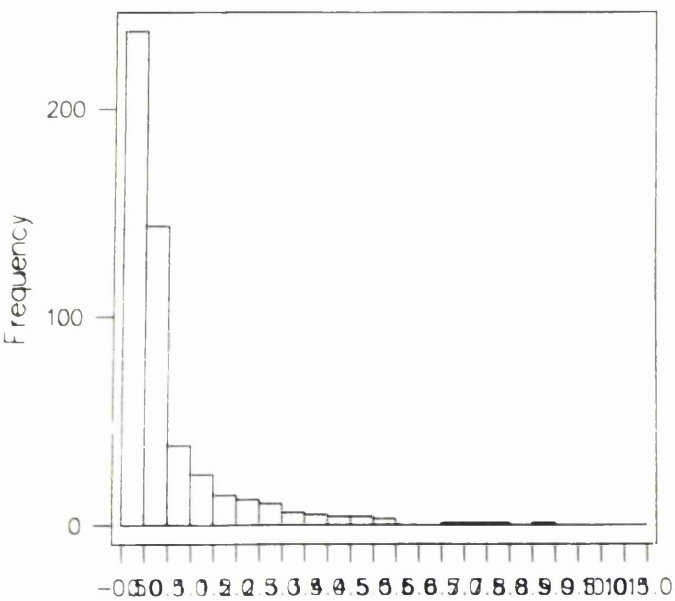

Figure 5.4. Histogram for $p = 0.50$ and $N = 3000$.

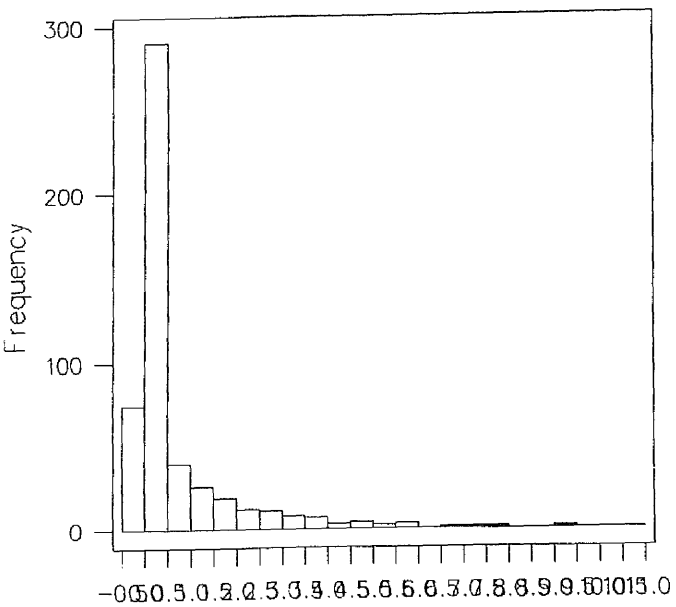

Figure 5.5. Histogram for $p = 0.52$ and $N = 3000$.

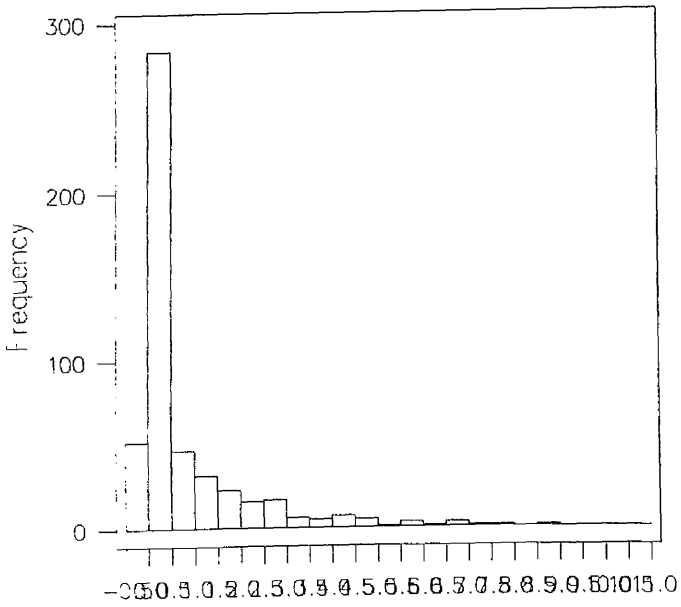

Figure 5.6. Histogram for $p = 0.55$ and $N = 3000$.

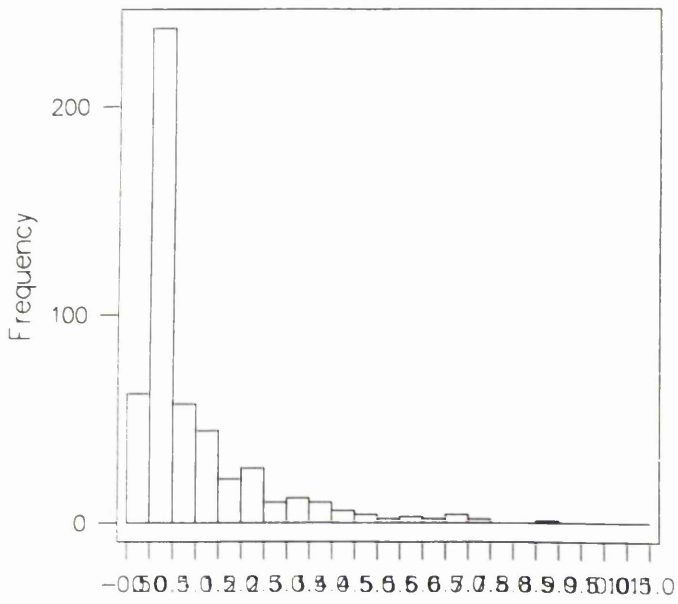

Figure 5.7. Histogram for $p = 0.60$ and $N = 3000$.

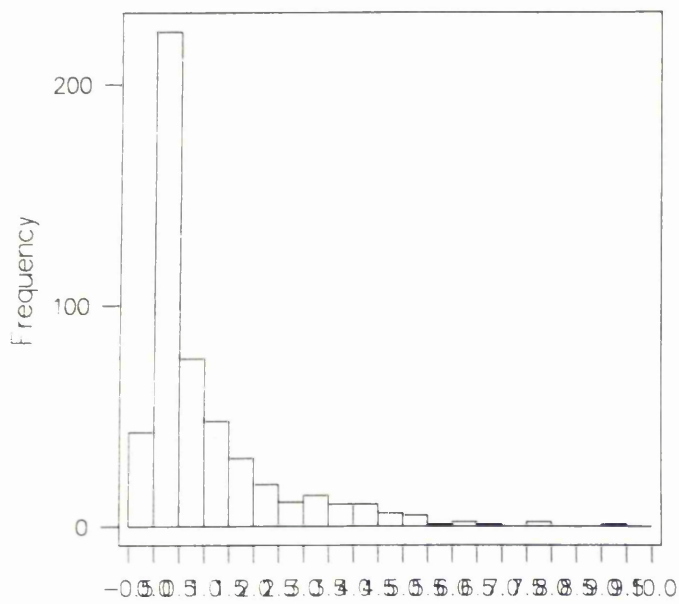

Figure 5.8. Histogram for $p = 0.65$ and $N = 3000$.

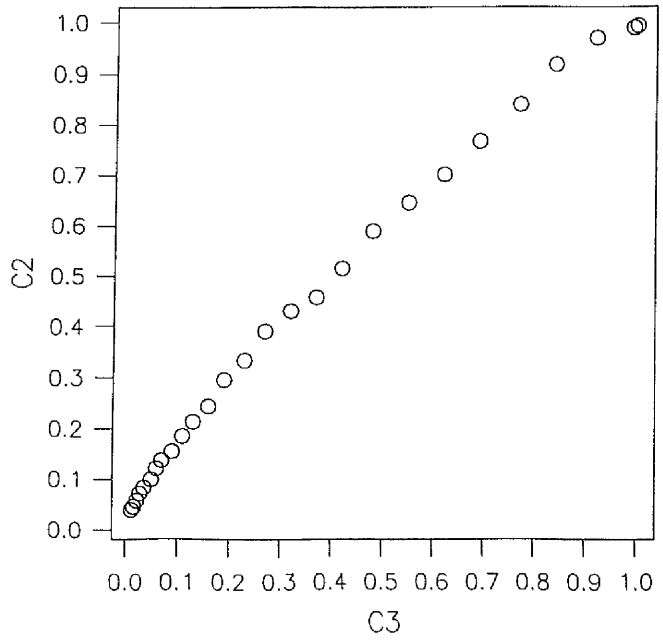

Figure 5.9. Estim./Theor. Significance Levels for $p = 0.75$ and $N = 25$.

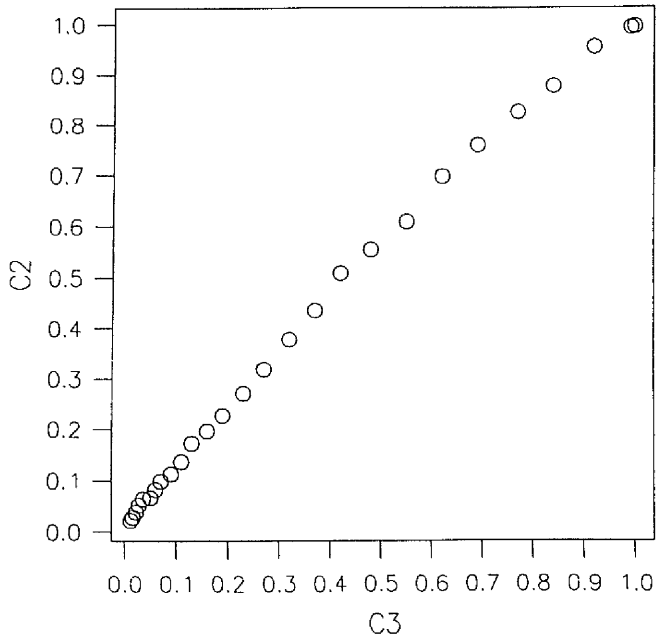

Figure 5.10. Estim./Theor. Significance Levels for $p = 0.75$ and $N = 100$.

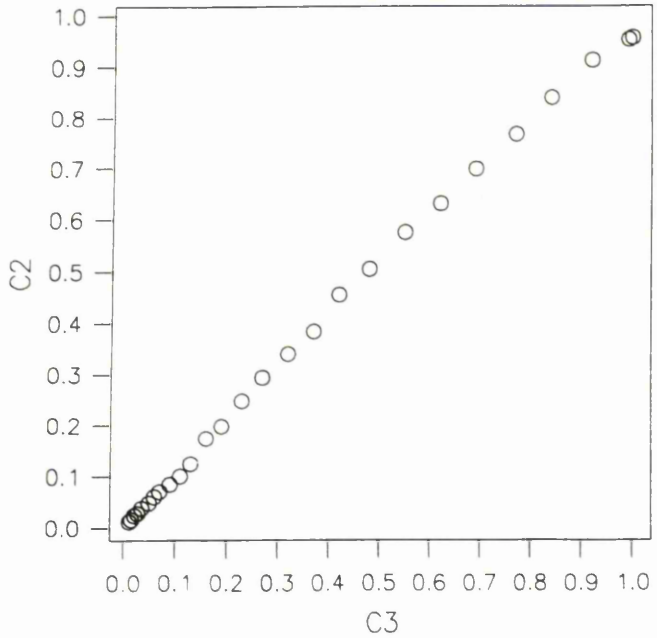

Figure 5.11. Estim./Theor. Significance Levels for $p = 0.75$ and $N = 500$.

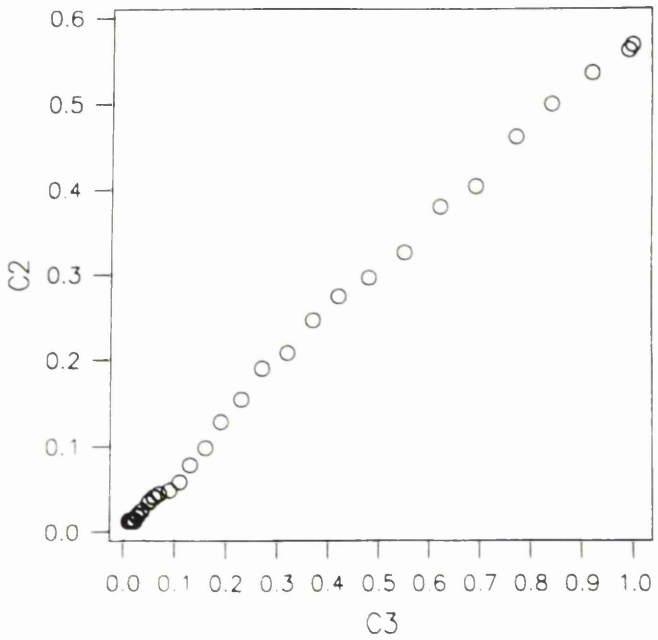

Figure 5.12. Estim./Theor. Significance Levels for $p = 0.50$ and $N = 500$.

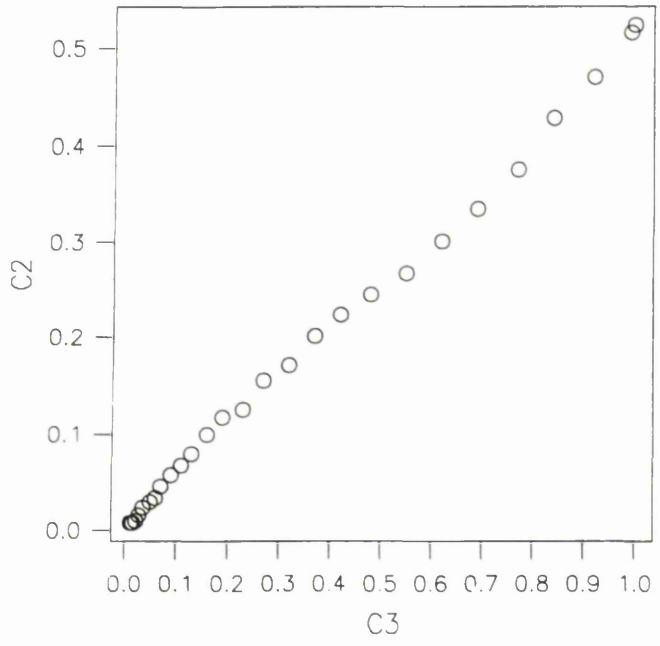

Figure 5.13. Estim./Theor. Significance Levels for $p = 0.50$ and $N = 3000$.

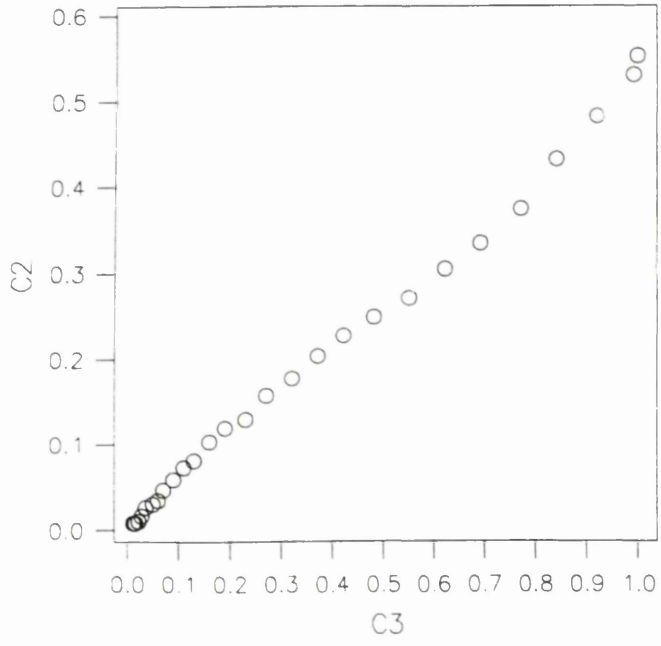

Figure 5.14. Estim./Theor. Significance Levels for $p = 0.501$ and $N = 3000$.

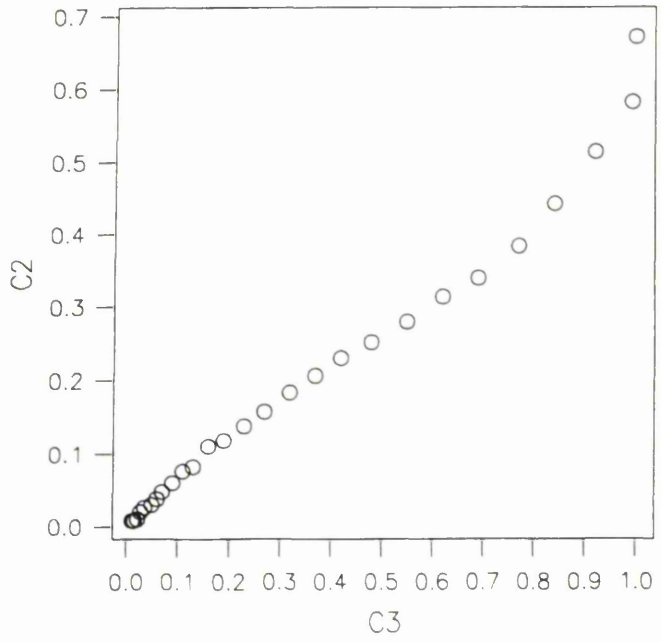

Figure 5.15. Estim./Theor. Significance Levels for $p = 0.505$ and $N = 3000$.

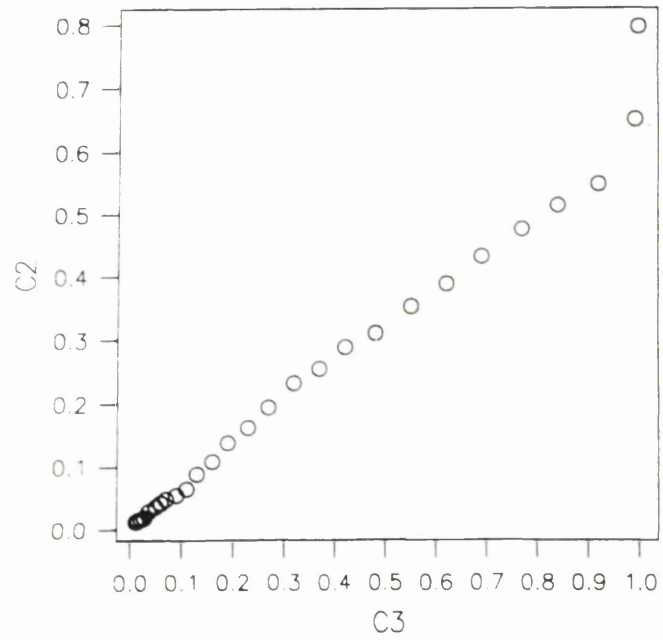

Figure 5.16. Estim./Theor. Significance Levels for $p = 0.51$ and $N = 500$.

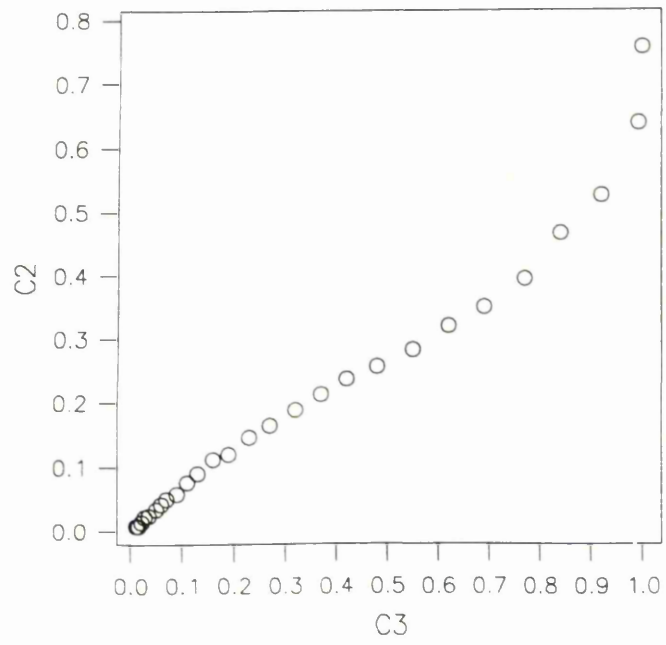

Figure 5.17. Estim./Theor. Significance Levels for $p = 0.51$ and $N = 3000$.

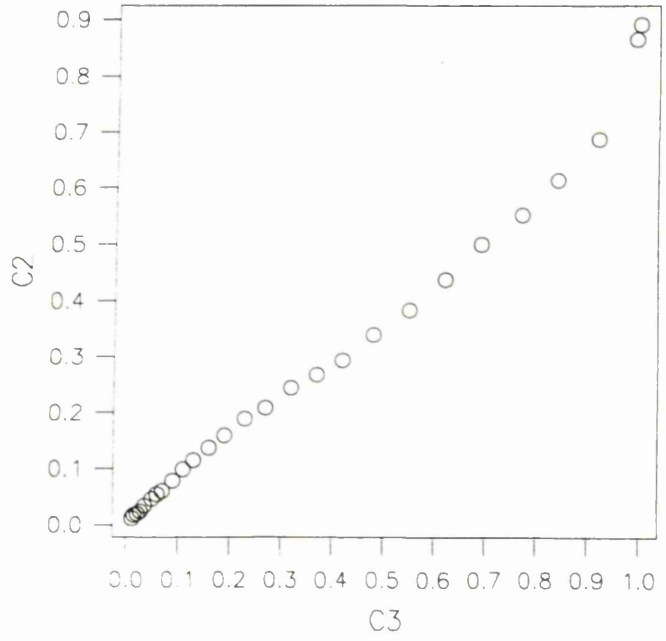

Figure 5.18. Estim./Theor. Significance Levels for $p = 0.55$ and $N = 3000$.

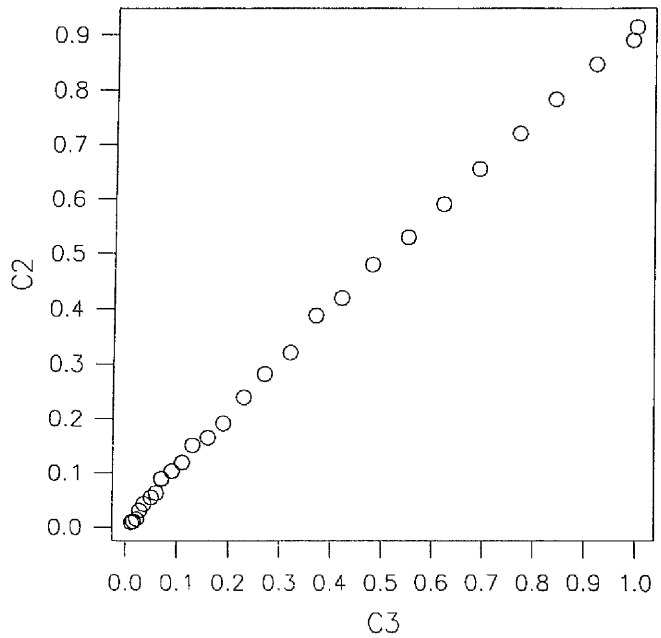

Figure 5.19. Estim./Theor. Significance Levels for $p = 0.65$ and $N = 3000$.

Chapter 6

Discussion

6.1 Introduction

In this last chapter, we give some impressions about the methods and results that we have presented, as well as as some indications for further work along with some new directions in the area of estimating the number of components. There are certainly open problems in this area; here, we will state some of them that are related to the methods that we presented in this work. We will also comment on a very recent Bayesian approach that has been proposed in the literature. We hope that this discussion will stimulate further research.

We now briefly recap the main points achieved in the present work. We have presented two algorithmic techniques that are based, respectively, on a stochastic variant of the EM algorithm and on information theory. The third technique, is based on a more theoretical approach, to derive the distribution of the likelihood ratio test.

6.2 A Bayesian Approach

In this section, we discuss a Bayesian methodology. Even though this type of approach is not directly related to our work, we believe that it constitutes a substantial breakthrough; indeed, for the first time, this methodology is able to include the estimation of the number of components in a mixture.

In a paper recently read at the Royal Statistical Society, Richardson and Green (1997) presented a new type of Monte Carlo Markov Chain (MCMC) method, the so called "Reversible Jump" MCMC for mixtures, which has the advantage over the usual MCMC, in that it takes into account the fact that the number of components is unknown. In this context, Richardson and Green (1997) derive the Bayes factors $B_{k_1k_2}$ for testing k_1 versus k_2 components, where $B_{k_1k_2} = \frac{p(k_1|X)p(k_2)}{p(k_2|X)p(k_1)}$, ($p(k_1)$ and $p(k_2)$ are the priors on the number of components), and this factor can then be considered as Bayesian information provided by the sample about the number of components, somewhat in the same way as Windham and Cutler's information ratio. We now explain how this factor is used in a practical example.

In a draft version of their paper, Richardson and Green (1996), present the results of a simulation exercise for the identification of the number of components. The true data distribution arise from a two-component mixture and, thus, k_2 is set equal to 2, and k_1 is taken between 1 and 6. The mean Bayes factors $\bar{B}_{12}$,...,$\bar{B}_{62}$ are derived for 50 replications, and the number of components is then estimated by $\hat{k}$, such that $\bar{B}_{\hat{k}2}$ is the highest mean Bayes factor. We briefly report the results here, since they are not quoted on the paper they presented to the RSS. They considered sample sizes of 50 and 250 of data from

the well-separated univariate model that we studied in our second chapter, and from a moderately-separated univariate model with two equally weighted normally distributed components (means=-1 and 1, and variances=4/9). For the first model and for a sample size of 250 of the second model, the results were very encouraging, since the two-component model was preferred to the single- and the three-component models at least 94 per cent of the time. However, in the moderately-separated case with sample size of 50, the two-component configuration was detected only 44 per cent of the time, while the single-component configuration was detected 56 per cent of the time. In this case, therefore, the small sample size has severely influenced the outcome.

It seems then that this methodology yields encouraging results, and is certainly more flexible than the usual MCMC, since it includes the estimation of the crucial number of components. We believe that it is worthwhile to study it further and to generalise it to multivariate mixtures.

We finally note, in passing, that we implemented the BLR technique described in the second chapter, for the moderately-separated model, with sample size 50, in order to compare the results to those provided by Richardson and Green's method. It is interesting to see that BLR detects the correct model 42 per cent of the time, a single component model 56 per cent of the time and a three-component model 2 per cent of the time. We believe then that the BLR performance can be considered as satisfactory.

6.3 A Test based on Stochastic Techniques

The EM algorithm has been, over recent years, a very popular deterministic technique for estimating the MLE in mixture problems, since it allows us to overcome the difficulty of solving non-linear maximum likelihood equations, and has some nice characteristics in addition. However, since at the same time it has some well-known disadvantages, there have been in the literature many attempts for improvement; the stochastic versions are among those, and the SEM algorithm is a typical example of that. It seems that this algorithm corrects some of the problems that one encounters when using EM, but, on the other hand, there are some theoretical complications, in that the successive iterates are realisations of random variables which converge in distribution to the MLE.

Now, our main aim in that context was to study the mathematical properties of a test statistic for the number of components based on the SEM algorithm, which was proposed in the literature (Celeux, 1987). We showed that this statistic cannot be formally considered as an asymptotic Hotelling's statistic, and we derived its actual asymptotic distribution. Then, we proposed a similar type of test statistic derived in a different way from iterates of the SEM algorithm.

These tests measure the stability of a partition of the data in hand into, say, K distinct classes, and, thus, their application differ, as we have seen, from the way they are usually applied in the literature. Comparison of these two tests showed a rather peculiar behaviour: they provided fairly similar results under the null hypothesis. However, since the simulations that we conducted

were used only for univariate mixtures with a small number of parameters, it might be interesting to perform extensive simulations in order to compare these two test statistics in the case of multivariate mixtures. The complexity of the calculations involved in using T^2 for inference is certainly higher, but, nevertheless, in most situations encountered in practice the true number of components is not larger than three, and the data are at most bidimensional. Hence the complexity of the calculations that are involved in using T^2 should not pose a big problem.

6.4 The Information Theory Techniques

We have seen that these techniques are closely related to the rate of convergence of the EM algorithm. It is worthwhile noting that although, as we described in the second chapter, it has been well known for some time in the literature that the rate of convergence of EM is connected to the degree of separation of the mixture components, it is only quite recently (Windham and Cutler, 1992) that this concept came to be used for estimating the number of components in a mixture model.

In Windham and Cutler's MIREV procedure, a validation measure was used to provide an indicator of the method's reliability. We would like to emphasise here what Windham and Cutler point out at the conclusion of their paper, that is, that this measure has not been as yet theoretically justified, and thus there is an open problem as far as this matter is concerned.

Our contribution to the estimation of the number of components by using

this methodology was to apply the information concept in a different way: since, for an overestimated mixture, we are in the presence of singularity problems concerning the information matrix (event E), we consider the minimal number of components K such that event E occurs, and estimate the actual number by $K-1$. For good component separation, the Modified MIR was found to yield very good results, and, for every degree of separation, it was found to compare favourably to Windham and Cutler's basic MIR procedure. As for Windham and Cutler's basic MIR, the Modified MIR is a Monte-Carlo approach. Further research concerning this area would then be to develop a more general theoretical tool that would remove the need for Monte-Carlo work.

6.5 Asymptotic Theory for known Mixing Proportions

A very important class of mixtures are those of two normal components. Goffinet et al (1992) study (Chapter 5) these types of mixture where the mixing proportions are assumed known. We will assume, in this section, that the data in hand arise from these types of mixture, and will look at open problems related to that case.

In the case that the data sample arises from a univariate mixture, we have seen, from Goffinet et al's theorem (1992), how the distribution of the Likelihood Ratio Test Statistic (LRTS) behaves as the sample size tends to infinity. The question then is how the LRTS behaves when the mixture is multivariate; for instance, can the results of the univariate case generalise directly to

multivariate mixtures, or can we at least find the shape of the asymptotic distribution for the LRTS?

In that context, we will report a multivariate result from Goffinet et al (1992); this result concerns the case where the data arise from a bivariate mixture, and is as follows.

Under H_0, the limiting distribution of $T(X)$, if the common variance-covariance matrix is known, is $\frac{1}{2}\chi^2(0) + \frac{1}{2}V^2$, where $V = V_1 + V_2^{1/2}$, and where V_1 is a $N(0,1)$ r.v., V_2 is a $\chi^2(2)$ r.v., and V_1 and V_2 are independent.

From the above result, one can see that the generalisation to the multivariate mixture case is not obtained directly from the results of the univariate mixture case, and that the LRTS distribution might take a more complex form, even for a bivariate mixture with known variance-covariance matrix.
On the other hand, Goffinet et al (1992) state that, when the common variance-covariance matrix is unknown, they have to cope with very complex likelihood equations, and thus could not find any analytical result. Consequently, there are many questions to be answered in that case, concerning the asymptotic distribution of the LRTS, even for a bivariate mixture: for instance, is it still possible to find the asymptotic distribution of the LRTS, and how do the values of p influence the shape of this distribution?
This is certainly an open problem, and some further research could be done to derive this distribution, at least for the bivariate mixture case.

References

Aitkin, M., Anderson, D., and Hinde, J. (1981), "Statistical Modelling of Data on Teaching Styles (with discussion)," J. R. Statist. Soc. A, 144, pp. 419-461.

Aitkin, M. and Rubin, D. (1985), "Estimation and Hypothesis Testing in Finite Mixture Models," J. R. Statist. Soc. B, 47, pp. 67-75.

Anderson, T.W. (1958), "An Introduction to Multivariate Statistical Analysis," Wiley.

Anderson, T.W. (1959), "On Asymptotic Distributions of Estimates of Parameters of Stochastic Difference Equations," Ann. Math. Statist., 30, pp. 676-687.

Anderson, T.W (1971), "The Statistical Analysis of Time Series," Wiley.

Athreya, K, and Pantula, S. (1986), "Mixing Properties of Harris Chains and Autoregressive Processes," J. Appl. Proba., 23, pp. 880-892.

Berdaï, A. and Garel, B. (1996), "Detecting a Univariate Normal Mixture with Two Components," Statistics and Decision, 16, pp.35-51.

Bezdek, J.C. (1981), "Pattern Recognition with Fuzzy Objective Function Algorithms," New-York: Plenum.

Billingsley, P. (1968), "Convergence of Probability Measures," Wiley.

Böhning, D., Dietz, E., Schaub, R., Schlattmann, P., and Lindsay, B.G. (1994), "The Distribution of the Likelihood Ratio for Mixtures of Densities From the One-Parameter Exponential Family," Ann. Inst. Statist. Math., 46, pp. 373-388.

Bosq, D. (1996), "Non-Parametric Statistics for Stochastic Processes," Estimation and Prediction, Lecture Notes in Statistics 110, Springer.

Bozdogan, H., and Sclove, S.L. (1984), "Multi-Sample Cluster Analysis Using Akaike's Information Criterion," Ann. Inst. Statist. Math., 36, Part B, pp. 163-180.

Bradley, R. (1986), "Basic Properties of Stong Mixing Conditions," E. Eberlein and M.S. Taqqu (Eds.), Dependence in Probability and Statistics, pp. 165-192, Birkhäuser.

Brockwell, P.J., and Davis, R.A. (1991), "Time Series: Theory and Methods," Second Edition, Springer-Verlag.

Broniatowski, M., and Diebolt, J. (1987), "Loi Stationnaire et Loi des Fluctuations pour le Processus Autorégressif Général d'ordre un," C. R. Acad.Sci.

Paris, t. 305, pp. 203-206.

Celeux, G. (1987), "Reconnaissance de Mélanges de Densités de Probabilité et Applications à la Validation des Résultats en Classification," Thèse d'Etat, Université Paris 9-Dauphine.

Celeux, G., Chauveau, D., and Diebolt, J. (1995), "Stochastic Versions of the EM Algorithm: An Experimental Study in the Mixture Case," Paper presented at the International Workshop on Statistical Mixture Modelling, at Aussois, France.

Celeux, G., and Diebolt, J. (1985), "The SEM Algorithm: a Probabilistic Teacher Algorithm derived from the EM Algorithm for the Mixture Problem," Comp. Statist. Quart., 2, pp. 73-82.

Celeux, G., and Diebolt, J. (1986a), "L'algorithme SEM: Un Algorithme d'Apprentissage Probabiliste pour la Reconnaissance de Mélange de Densités," Revue de Statistique Appliquée, 34, No. 2, pp. 35-52.

Celeux, G., and Diebolt, J. (1986b), "Etude du Comportement asymptotique d'un Algorithme d'Apprentissage Probabiliste pour les Mélanges de Lois de Probabilité," Rapports de Recherche INRIA, No. 563.

Celeux, G., and Diebolt, J (1988), "A Random Imputation Principle: The Stochastic EM Algorithm," Rapports de Recherche INRIA, No. 901, Programme 5.

Celeux, G., and Diebolt, J. (1992), "A Stochastic Approximation Type EM Algorithm for the Mixture Problem," Stochastics and Stochastics Reports, 41, pp. 119-134.

Chen, J., and Cheng, P. (1992), "A New Approach of Testing the Number of Components in Finite Mixture Models," Technical Report.

Cheng, R.C.H., and Traylor, L. (1995), "Non-Regular Maximum-Likelihood Problems (with discussion)," J. R. Statist. Soc. B, 57, pp. 3-44.

Cox, D.R., and Hinkley, D.V. (1974), "Theoretical Statistics," London: Chapman and Hall.

Davydov, Y. (1973), "Mixing Conditions for Markov Chains," Theory of Probability and its Applications, 18, pp. 312-328.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977), "Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm (with discussion)," J. R. Statist. Soc. B, 39, pp. 1-38.

Diebolt, J., and Celeux, G. (1992), "Asymptotic Properties of a Stochastic EM Algorithm for Estimating Mixing Proportions," Rapports de Recherche INRIA, No. 1591, Programme 5.

Diebolt, J., and Ip, E. (1996), "Stochastic EM: Method and Application," Markov Chain Monte Carlo in Practice, W.R. Gilks, S. Richardson and D.J. Spiegelhalter (Eds.), pp. 259-273, Chapman and Hall, London.

Doob, J. (1953), "Stochastic Processes," Wiley, New-York.

Durairajan, T.M. and Kale, B.K. (1982), "Locally Most Powerful Similar Test for the Mixing Proportion," Sankhyā A, 44, pp. 153-161.

Efron, B., and Tibshirani, R.J. (1993), "An Introduction to the Bootstrap," Chapman and Hall.

Everitt, B., and Hand, D. (1981), "Finite Mixture Distributions," Chapman and Hall.

Feng, Z.D., and McCulloch, C.E. (1996), "Using Bootstrap Likelihood Ratios in Finite Mixture Models," J. R. Statist. Soc. B, 58, pp. 609-617.

Garel, B. (1996), "Asymptotic Theory of Likelihood Ratio Test for the Identification of a Mixture," Technical Report.

Ghosh, J.K., and Sen, P.K. (1985), "On the Asymptotic Performance of the Log-Likelihood Ratio Statistic for the Mixture Model and Related Results," Proc. Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, 2, L.M. Le Cam and R.A. Olshen (Eds.), pp. 789-806, Monterey: Wadsworth.

Goffinet, B., Loisel, P., and Laurent, B. (1992), "Testing in Normal Mixture Models when the Proportions are known," Biometrika, 79, pp. 842-846.

Graybill, F.A. (1969), "Introduction to Matrices with Applications in Statistics," Wadsworth Publishing Company, CA.

Grimmett, G.R., and Stirzaker, D.R. (1992), "Probability and Random Processes," Oxford University Press.

Hall, P., and Heyde C.C. (1980), "Martingale Limit Theory and its Applications," Academic Press, New-York.

Hope, A. (1968), "A Simplified Monte-Carlo Significance Test Procedure," J. R. Statist. Soc. B, 30, pp. 582-598.

Hurwitz, L. (1950), "Least-Square Bias in Time Series," Statistical Inference in Dynamic Economic Models-Cowles Commission Monograph 10, T. Koopmans (Ed.), pp. 365-383, Wiley and Sons, New-York.

Ibragimov, I., and Linnik, Y. (1971), "Independent and Stationary Sequences of Random Variables," Walters-Noordhoff, Groningen.

Kemeny, J., and Snell, J. (1974), "Finite Markov Chains," Springer-Verlag.

Kolmogorov, A.N., and Rozanov, Y. (1960), "On Strong Mixing Conditions for Stationary Gaussian Processes," Theory of Probability and its Applications, 5, pp. 204-208.

Lindsay, B.G. (1983), "The Geometry of Mixture Likelihoods: a General Theory," The Annals of Statistics, 11, pp. 86-94.

Little, R.J.A., and Rubin, D.B. (1987), "Statistical Analysis with Missing Data," Wiley, New-York.

Louis, T.A. (1982), "Finding the Observed Information Matrix when Using the EM Algorithm," J. R. Statist. Soc. B, 44, pp. 226-233.

Marron, J.S., and Wand, M.P. (1992), "Exact Mean Integrated Squared Error," The Annals of Statistics, 20, pp. 712-736.

McLachlan, G.J. (1987), "On Bootstrapping the Likelihood Ratio Test Statistic for the Number of Components in a Normal Mixture," Applied Statistics, 36, pp. 318-324.

McLachlan, G.J., and Basford, K.E. (1988), "Mixture Models: Inference and Applications to Clustering," Marcel Dekker, New-York.

Meng, X.L. (1994), "On the Rate of Convergence of the ECM Algorithm," The Annals of Statistics, 22, pp. 326-339.

Meng, X.L., and Rubin, D.B. (1991), "Using EM to Obtain Asymptotic Variance-Covariance Matrices: The SEM Algorithm," J. Amer. Statist. Assoc., 86, pp. 899-909.

Meng, X.L., and Rubin, D.B. (1992), "Recent Extensions to the EM Algorithm (with discussion)," Bayesian Statistics 4, J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith (Eds.), pp. 307-320, Oxford University Press.

Meng, X.L., and Rubin, D.B. (1994), "On the Global and Componentwise Rates of Convergence of the EM Algorithm," Linear Algebra and its Applications, 199, Special Issue in Honour of Ingram Olkin, pp. 413-425.

Orchard, T., and Woodbury, M.A. (1972), "A Missing Information Principle: Theory and Application," Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, pp. 697-715.

Ostrowski, A.M. (1960), "Solution of Equations and Systems of Equations," New-York: Academic Press.

Peters, B.C., and Walker, H.F. (1978), "An Iterative Procedure for Obtaining Maximum-Likelihood Estimates of the Parameters for a Mixture of Normal Distributions," SIAM J. Appl. Math., 35, pp. 362-378.

Polymenis, A. (1997), "Discussion of the Paper: On Bayesian Analysis of Mixtures with an Unknown Number of Components (by S. Richardson and P.J. Green)," J. R. Statist. Soc. B (to appear).

Quinn, B.J., McLachlan G.J., and Hjort, N.L. (1987), "A Note on the Aitkin-Rubin Approach to Hypothesis Testing in Mixture Models," J. R. Statist. Soc. B, 49, pp. 311-314.

Redner, R.A., and Walker, H.F. (1984), "Mixture Densities, Maximum Likelihood and the EM Algorithm," SIAM Review 26, pp. 195-239.

Richardson S., and Green, P.J. (1997), "On Bayesian Analysis of Mixtures with an Unknown Number of Parameters (with discussion)," J. R. Statist. Soc. B (to appear).

Robert, C.P. (1992), "Discussion of the Paper: Recent Extensions to the EM Algorithm (by X.L. Meng and D.B. Rubin)," Bayesian Statistics 4, J.M. Bernardo, J.O Berger, A.P. Dawid and A.F.M. Smith (Eds.), pp.315-318, Oxford University Press.

Rubin, H. (1950), "Consistency of Maximum-Likelihood Estimates in the Explosive Case," Statistical Inference in Dynamic Economic Models-Cowles Commission Monograph 10, T. Koopmans (Ed.), pp. 356-364, Wiley and Sons, New-York.

Self, S.G., and Liang, K.Y. (1987), "Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions," J. Amer. Statist. Assoc., 82, pp. 605-610.

Silvey, S.D. (1975), "Statistical Inference," Chapman and Hall.

Soromenho, G. (1994), "Comparing Approaches for Testing the Number of Components in a Finite Mixture Model," Computational Statistics 9, pp. 65-78.

Sundberg, R. (1974), "Maximum Likelihood Theory for Incomplete Data from an Exponential Family," Scand. J. Statist., 1, pp. 49-58.

Sundberg, R. (1976), "An Iterative Method for Solution of the Likelihood Equations for Incomplete Data from Exponential Families," Commun. Statist.-Simula. Computa. B, 5, pp. 55-64.

Teicher, H. (1963), "Identifiability of Finite Mixtures," Ann. Math. Statist., 34, pp. 1265-1269.

Thode, H.C., Finch, S.J., and Mendell, N.R. (1987), "Finding the MLE in a Two-Component Normal Mixture," Proc. Statist. Compu. Sec., Amer. Statist. Assoc., Washington DC, pp. 472-475.

Thode, H.C., Finch, S.J., and Mendell, N.R. (1988), "Simulated Percentage Points for the Null Distibution of the Likelihood Ratio Test," Biometrics, 44, 1195-1201.

Titterington, D.M. (1990), "Some Recent Research in the Analysis of Mixture Distributions," Statistics, 21, pp. 619-641.

Titterington, D.M. (1997a), "Mixture Distributions (Update)", Encyclopedia of Statistical Sciences Update, Vol. 1, S. Kotz, C.B. Read and D. Bunks (Eds.), pp. 399-407, Wiley: New-York (to appear).

Titterington, D.M. (1997b), "Discussion of the Paper: On Bayesian Analysis of Mixtures with an Unknown Number of Components (by S. Richardson and P.J. Green)," J. R. Statist. Soc. B (to appear).

Titterington, D.M., Smith, A.F.M., and Makov, U. (1985), "Statistical Analysis of Finite Mixture Distributions," Wiley, New-York.

Wilks, S. (1963), "Mathematical Statistics," Wiley, New-York.

Windham, M.P., and Cutler, A. (1992), "Information Ratios for Validating Mixture Analyses," J. Amer. Statist. Assoc., 87, pp. 1188-1192.

Wolfe, J.H. (1971), "A Monte-Carlo Study of the Sampling Distribution of the Likelihood Ratio for Mixtures of Multinomial Distributions," Technical Bulletin STB 72-2, Naval Personnel and Training Research Laboratory, San Diego, CA.

Wu, J. (1983), "On the Convergence Properties of the EM algorithm," The Annals of Statistics, 11, pp. 95-103.

Yakowitz, S., and Spragins, J. (1968), "On the Identifiability of Finite Mixtures," Ann. Math. Statist., 39, pp. 209-214.

Zacks, S. (1971), "The Theory of Statistical Inference," Wiley.