THE NOISE COMPONENT IN MODEL-BASED CLUSTERING

Thesis submitted to the
Faculty of Science, University of London
for the degree of Doctor of Philosophy

by Pietro Coretto

Department of Statistical Science — University College London

April 2008

I, Pietro Coretto, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Pietro Coretto

"There are two possible outcomes: if the result confirms the hypothesis, then you have made a measurement. If the result is contrary to the hypothesis, then you have made a discovery."

— Enrico Fermi

Acknowledgements

I would like to express my deep gratitude to my supervisor Dr Christian Hennig, who has been so patient and always generous with his time. His guidance, insight and kindness have been invaluable. Above all, I want to thank him for having shared his keen and deep scientific thinking, and his original ideas about statistics and data analysis.

I am also grateful to Professor A. Philip Dawid for his encouragement which convinced me to continue my studies when my self-confidence seemed to be irrecoverable.

I am indebted to all those close to me. In particular, I am thankful for my wife, Assunta, who has been with me, to help and sustain me through all the difficulties I have faced. I am grateful to my family who take such pride and interest in my academic studies and achievements

Abstract

Model-based cluster analysis is a statistical tool used to investigate group-structures in data. Finite mixtures of Gaussian distributions are a popular device used to model elliptical shaped clusters. Estimation of mixtures of Gaussians is usually based on the maximum likelihood method. However, for a wide class of finite mixtures, including Gaussians, maximum likelihood estimates are not robust. This implies that a small proportion of outliers in the data could lead to poor estimates and clustering. One way to deal with this is to add a "noise component", i.e. a mixture component that models the outliers. In this thesis we explore this approach based on three contributions.

First, Fraley and Raftery (1993) propose a Gaussian mixture model with the addition of a uniform noise component with support on the data range. We generalize this approach by introducing a model, which is a finite mixture of location-scale distributions mixed with a finite number of uniforms supported on disjoint subsets of the data range. We study identifiability and maximum likelihood estimation, and provide a computational procedure based on the EM algorithm.

Second, Hennig (2004) proposed a sort of model in which the noise component is represented by a fixed improper density, which is a constant on the real line. He shows that the resulting estimates are robust to extreme outliers. We define a maximum likelihood type estimator for such a model and study its asymptotic behaviour. We also provide a method for choosing the improper constant density, and a computational procedure based on the EM algorithm.

The third contribution is an extensive simulation study in which we measure the performance of the previous two methods and certain other robust methodologies proposed in the literature.

Contents

List of Tables

List of Figures

CHAPTER 1

An Introduction to Finite Mixture Models

In this chapter we will give a brief overview of finite mixture models. After having introduced basic definitions and properties we shall give some interpretation of samples arising from finite mixture disributions. We introduce the identifiability problem and its solution for some classes of models. As a final step we will review some of the estimation methods and clustering techniques.

§1.1. Overview

Statistical analysis via finite mixture models is a widely used tool employed in many scientific fields. Their obvious application is where there is some known group-structure in the population of interest, or when the main goal is to reveal such group-structure in the data as in cluster analysis. Due to their flexibility mixture models are also being extensively exploited as a semiparametric device to model unknown distributional shapes. Fields where mixture models have been successfully applied include economics, astronomy, biology, genetics, medicine, marketing, engineering, etc. In many areas of statistics such as cluster and discriminant analysis, pattern recognition and survival analysis, methods related to mixture models have been a substantial part of the research conducted.

Finite mixture models are models where the distribution function of the probability measure underlying the process which generates the data is assumed to be a convex combination of distribution functions of some parametric family. By this, it is clear that such models are helpful when some degree of heterogeneity characterizes the data. In model-based cluster analysis each group of observations is treated as a mixture component, in econometrics given the population of interest different sub-populations are represented as a mixture of several components. Since any continuous distribution can be approximated arbitrarily well by a finite mixture of continuous distributions, mixture models also provide a valid semiparametric tool to model unknown distributions. For example Priebe (1994) showed that with a sample size of 10,000 observations, a log-normal distribution can be well approximated by a convex combination of about 30 normal distributions while a kernel density estimator would consist of a mixture of 10,000 normal distributions.

The very first analysis involving the use of a finite mixture was due to Pearson (1894, 1895). He had to analyse biological data which consisted of the ratio of forehead to body length of 1000 crabs sampled from the Bay of Naples. The data presented a positive skewness with thin tails, so that fitting a normal distribution was unsatisfactory. Pearson (1894) had the intuition to fit the distribution of the data by using a convex combination of two Gaussian densities with unknown unequal variances. Since the maximum likelihood estimate for such a model does not have a tractable analytical form, Pearson computed the estimated parameters by the method of moments. After Pearson's introduction of mixture models the scientific interest in them has been mitigated by the computational difficulties implied by any estimation procedure. The method of moments leads to closed form estimators only in particular cases. With the revolution of computational capability the computation of estimators of finite mixture parameters has become easier than before. In fact, with the development of the EM algorithm [1], the estimation of the the maximum likelihood estimator for the parameter of many mixture models has became feasible in cases.

§1.2. Basic definitions

Let $\mathscr{F} := \left\{F(x;\theta): \quad x \in \mathbb{R}^q, \theta \in \mathbb{R}^k\right\}$ be a family of one dimensional distribution functions over $\mathbb{R}^q$ indexed by a point θ in a Borel subset of $\mathbb{R}^k$ such that $F(x;\theta)$ is measurable on $\mathbb{R}^q \times \mathbb{R}^k$. Let $G \in \mathscr{G}$ be a k-dimensional distribution function. Let $\mathscr{H}$ be a family of distribution functions. We consider the function $H(x) = \int_{\mathbb{R}^k} F(x;\theta)dG(\theta)$. $H(x)$ is a q-dimensional distribution function called the mixture (or G-mixture) of $\mathscr{F}$ with G as the mixing distribution. When G is

[1] We will review the EM algorithm in extensive form in the next few chapters

discrete with finite support, the set of all finite mixtures $\mathscr{H}$ of the family $\mathscr{F}$ is simply the convex hull of $\mathscr{F}$, i.e.

$$\mathscr{H} := \left\{ H(x) : H(x) = \sum_{j=1}^{s} \pi_j F(x; \theta_j), s \geq 1, \pi_j > 0, \sum_{j=1}^{s} \pi_j = 1, F(x; \theta_j) \in \mathscr{F} \right\}.$$

The parameter s can be understood as fixed or free. In this work we will always deal with mixtures where the number of components s is fixed and known. However in many situations one needs to estimate s, which can be done via Bayes Information Criterion (BIC) (see McLachlan and Peel, 2000a). In this work we will mainly be interested in finite mixtures; these are mixture distributions where the number of components s is finite and the mixing distribution is discrete.
Let $\underline{X_n}$ be a sample of size n, that is $\underline{X_n} := \{X_1, X_2, \ldots, X_n\}$ where $X_i \in \mathbb{R}^q$ is a random variable for $i = 1, 2, \ldots, n$. Where possible a realization of the sample $\underline{X_n}$ is indicated as the n-tuple of random vectors $\underline{x_n} := \{x_1, x_2, \ldots, x_n\}$, i.e. x_i is the realization of the random vector X_i. We assume that X_i is distributed according to a probability measure with distribution function in $\mathscr{F}$. We will also assume that a distribution function $F \in \mathscr{F}$ has a representation in terms of a density function f. Thus we will usually write a mixture model as the density function

$$f(x_i; \eta) = \sum_{j=1}^{s} \pi_j f(x_i; \theta_j), \tag{1.1}$$

with $\eta = (\pi_1, \pi_2, \ldots, \pi_s, \theta_1, \theta_2, \ldots, \theta_s)$, $0 < \pi_j < 1$ for every $j = 1, 2, \ldots, s$, and $\sum_{j=1}^{s} \pi_j = 1$. The quantities $\{\pi_j\}_{j=1}^{s}$ are called mixing proportions or weights. The distribution function $F(x; \theta_j)$ and its density $f(x; \theta_j)$ is usually referred to as the jth mixture component. The model in (1.1) is called an *s-components mixture*.

§1.3. Interpretation of mixture models

There are several cases where modelling via mixture models is reasonable. Mixture models are frequently employed to explain data which exhibit heterogeneity or group-structure. They are also useful to model data where multi-modality or skewness is present. The interpretation of the mixture model differs depending on the nature of the particular data at hand, and on the scope of the analysis. In this section we will also introduce how to simulate artificial samples from mixture distributions.

1.3.1 — Mixtures as a tool to fit a distribution

When the exploration of the data suggests the existence of a multi-modal or skewed structure, mixture models can be a useful and flexible semiparametric device to model the underlying distribution. Mixture distributions are often used in the context of density estimation. Here we are in a situation where the main goal of the analysis is to fit a probability distribution to the data. At one extreme of this we have a nonparametric kernel density estimator which can be seen as the estimate of a mixture density where the number of components is equal to the number of observations and all mixing proportions are assumed to be equal. In fact given an observed sample $\underline{x_n}$ drawn from an unknown distribution, a kernel density estimate at x_i is the estimate of a density

$$\tilde{f}(x_i) = \frac{1}{nh}\sum_{j=1}^{n} k\left(\frac{x_i - x_j}{h}\right),$$

for an appropriate choice of the constant h and the function $k(\cdot)$ which can be itself a density function. It is easy to see that $\tilde{f}(x_i)$ is similar to (1.1) when we set $\pi_j = 1/n$, $s = n$ and replace $f(x_i;\theta)$ with $h^{-1}k((x_i - x_j)/h)$. This shows that for $1 < s < n$, mixture models can be viewed as a semiparametric tool that allows: (i) to gain flexibility with the respect to the fully parametric model (s=1), and (ii) it also allows to keep the dimension of the parameter space finite which avoids many problems in the theoretical analysis.

Let us assume that the observed sample $\underline{x_n}$ is an i.i.d. drawn from a probability distribution represented by (1.1). For example, suppose we have observed the sample for which we produced the histogram in Figure 1.1 or Figure 1.2. In these cases fitting a standard parametric distribution could be unsatisfying and the flexibility of mixture models can improve the fit. Here, interest is not in the group-structure of the data, so that we can interpret each x as draw independently from some distribution having a density function as in (1.1). No particular interpretation of the weights $\{\pi_j\}_{j=1}^{s}$ is relevant here. The proper way to simulate a sample of n i.i.d. observations from a mixture distribution would be to compute the inverse of the distribution function of the mixture -provided that it exists- and then to compute it on a drawn of n i.i.d. numbers from a Uniform(0,1) (this is just the standard application of the uniform probability theorem). There exists another approach to simulation which will be clear in the next few sections.

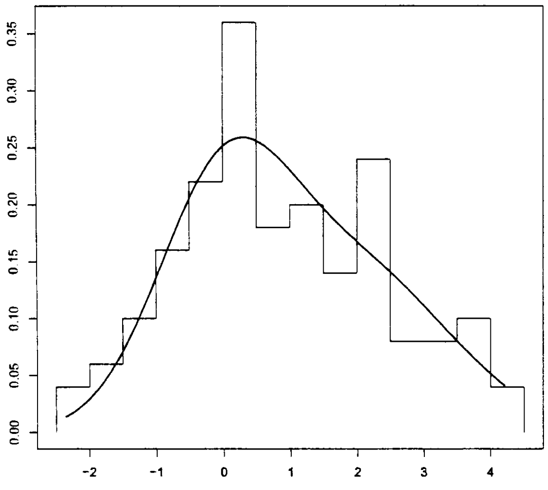

Figure 1.1: Histogram for 100 observations drawn from a distribution 0.5Normal$(0, 1)$ + 0.5Normal$(2, 2)$. The thicker line represents the density function of the true underlying distribution.

1.3.2 — Samples with group-structures

In cluster analysis and discriminant analysis, the interest of the researcher is to understand the group structure of the population under study. In this case each component represents a group in the population with its own behaviour. Here we want to physically identify the s mixture components in (1.1) with s existing groups composing the population under study. We assume that there are s populations and the number of units drawn from each group is not fixed. In this case we can still make sense of the model represented by (1.1). Let Z_i be a categorical random variable taking on the values in $\{1, 2, \ldots, s\}$ with probabilities $\pi_1, \pi_2, \ldots, \pi_s$; and suppose that the conditional density of $X_i|Z_i = j$ is $f_j(x_i)$ where $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, s$. The unconditional density (i.e. the marginal density) of X_i is given by $f(x_i)$ in (1.1). In this case we are writing a probabilistic model for the data generating process where the unknown *component membership* is not taken into account. How to simulate such a sample? First let us work with an s-dimensional component label vector Z_i instead of a single categorical random variable. Now Z_i is such that its jth element, Z_{ij}, takes value one if x_i is originated by the jth mixture component (i.e. X_i is a realization of the distribution represented by $f(x, \theta_j)$), and zero otherwise. Thus Z_i is distributed according to a multinomial distribution consisting of a single draw on s distinct categories with probability parameters given by $\pi = (\pi_1, \pi_2, \ldots, \pi_s)$;

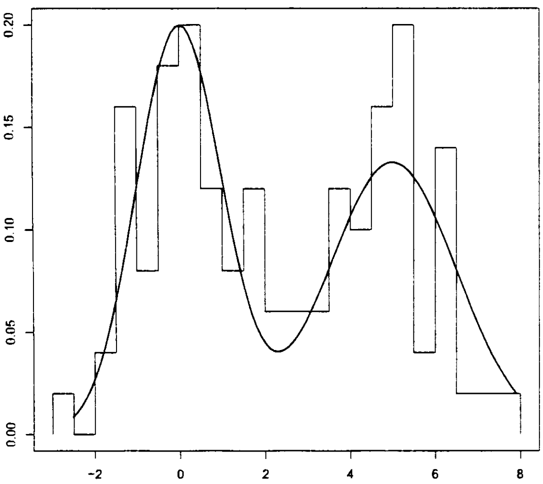

Figure 1.2: Histogram for 100 observations drawn form a distribution 0.5Normal(0, 1) + 0.5Normal(5, 2). The thicker line represents the density function of the true underlying distribution.

that is

$$\Pr\{Z_i = z_i\} = \prod_{j=1}^{s} \pi_j^{z_{ij}}, \tag{1.2}$$

hence

$$Z_i \sim \text{Multinomial}_s(1, \pi), \quad \pi = (\pi_1, \pi_2, \ldots, \pi_s). \tag{1.3}$$

As we shall see in the following chapters this formalization will be useful to write down the likelihood function for such a probabilistic model. Moreover this formalism also gives us a proper method to simulate samples from a mixture distribution. In order to simulate a sample of n observations from (1.1), where the latter is interpreted as the marginal density of X_i, first we draw n observations $\{z_i\}_{i=1}^{n}$ from Multinomial$_s(1, \pi)$. Let $[z_i]_j$ be the jth coordinate of the vector z_i; then let

$$n_j = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left\{[z_i]_j = 1\right\},$$

and draw n_j observations from the distribution represented by $f(x_i, \theta_j)$ for $j = 1, 2, \ldots, s$.

In some applications of mixture models the vector X_i can be physically associated with some of the mixture components. In these situations the vector X_i is known together with the label vector Z_i, and we say that X_i is classified with respect of component memberships. In many situations we have a sample of i.i.d. observations from a mixture distribution as represented by (1.1) and the vectors $\{z_i\}_{i=1}^n$ are not known. This is the case of unclassified data. There are also intermediate situations where the component memberships are partially known to the researcher. The rest of this work is devoted to problems related to unclassified samples.

§1.4. Multimodality and Shapes of Mixture Distributions

Mixture distributions are not necessarily represented by multimodal densities. This is shown in the Figure 1.2 and Figure 1.3 where we plotted the graph of two different densities of mixtures of Gaussian distributions. Models based on mixtures are often used to fit distributions with multiple modes in the density. On the other hand one should be aware that even in the case where multimodality does not seem to occur in the data, a mixture structure can still be a feature of the underlying data generating process. In some restricted cases the relation between the parameters of the mixture distribution and the geometric properties of the representing density function are well known.

In the forthcoming chapters we will be interested mainly in mixtures of univariate normal distributions. Hence it will be worthwhile to mention some of the findings about this particular class. In the case of the homoscedastic univariate normal mixtures (i.e. all the component mixtures have the same variance), many of the geometric properties have been studied. The multimodality of such mixture distributions depend on the separation between the means of adjacent components scaled by the variance. Let us consider the mixture density

$$f(x) = \pi_1 \phi(x; \mu_1, \sigma) + (1 - \pi_1)\phi(x; \mu_2, \sigma),$$

where $\phi(x; a, b)$ is the density function at a point x, of a normal distribution with mean a and standard deviation b. The presence of multimodality in such a mixture depends on the distance

$$\Delta = \frac{|\mu_1 - \mu_2|}{\sigma};$$

which is the Mahalanobis distance between homoscedastic components. To see how the value of Δ affects the shape of such a mixture density, let us look at the plots in Figure 1.3. In such a situation we have two modes when $\Delta > 2$. The

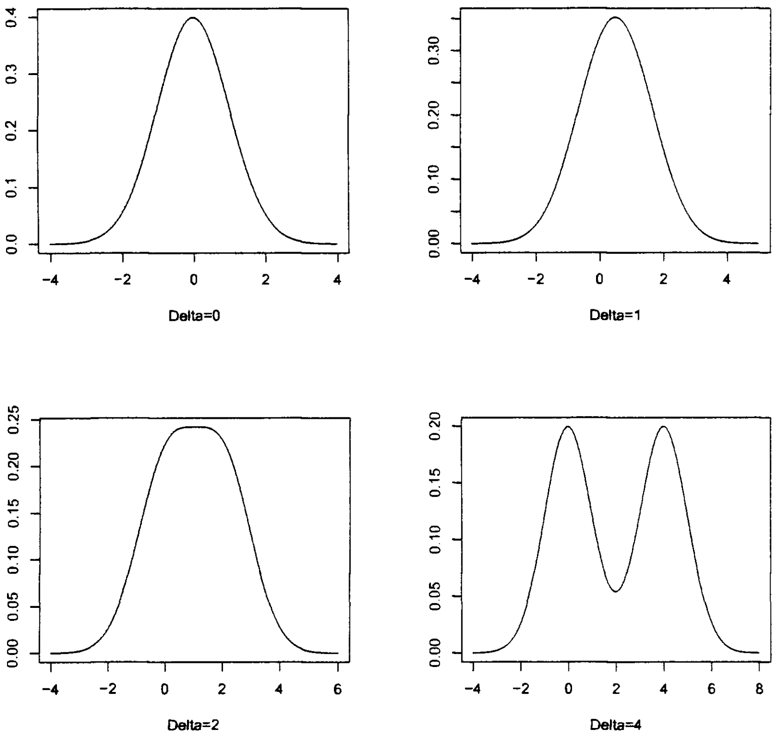

Figure 1.3: Plot of the density functions of mixtures consisting of two univariate normal components with equal proportions, variance $\sigma^2 = 1$, and means $\mu_1 = 0$ and $\mu_2 = \mu_1 + \sigma\Delta$ for $\Delta = 0, 1, 2, 4$.

case when $1 < \Delta < 2$ is also interesting because the density presents a nearly flat region at the maximum. For more details on the exact conditions under which the two-component normal mixture is bimodal, see Robertson and Fryar (1968). Lindsay and Roeder (1992) also studied some mathematical transformation of the mixture which is able to establish the presence of multiple components in the univariate normal case. Shaked (1980) studies the geometric properties of density functions of univariate mixtures from an exponential family in a setup similar to that of Robertson and Fryar (1968), generalizing some of their results. Ray and Lindsay (2005) proposed analytical methods borrowed from topography in order to study the geometry of multivariate normal mixtures.

Despite the effort of many researchers in studying the relations between parameters' value and geometrical properties of the mixture models (mainly normal mixtures), there is no clear and general answer to these issues. However the topic is of extreme relevance because with data at hand we would like to be able to understand the probabilistic structure under the data generating process. In particular this happens in cluster analysis where the main interest is to discover

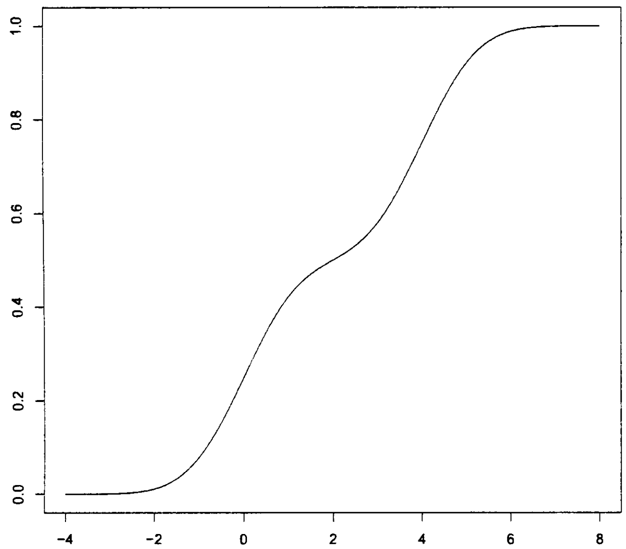

Figure 1.4: Plot of the distribution function of a mixture consisting of two univariate normal components with equal proportions, variance $\sigma^2 = 1$, and means $\mu_1 = 0$ and $\mu_2 = 4$.

group structures in data.

Studies about shapes of distributions are mainly based on density functions. But what happens if we look at the distribution function of a mixture distribution? If the components are reasonably separated the distribution function of a normal mixture has a shape which alternates regions of concavity and convexity, an example being given in Figure 1.4. In Figure 1.5 we represented the density and the distribution function of a two components homoscedastic normal mixture with two modes. An histogram and the empirical distribution function on a sample of 250 units drawn from the same distribution is also presented. We notice that while the separation of the components is such that the density function has two modes, the corresponding distribution function does not look clearly different from a distribution function associated with a symmetric unimodal density. At reasonable graphical resolution of the plot of the graph, the true distribution functions do not seem to suggest that its density has multiple well separated maxima. The presence of the two modes is not clear from the histogram as well, unless we tune ad-hoc the number of cells of the histogram. The empirical distribution function seems to approximate the true distribution function quite well, and thus it does not reveal any presence of a multi-modal structure. These simple examples show that modelling via mixture distributions should be carefully handed. The conclusion here is that there are many cases where the exploration of the

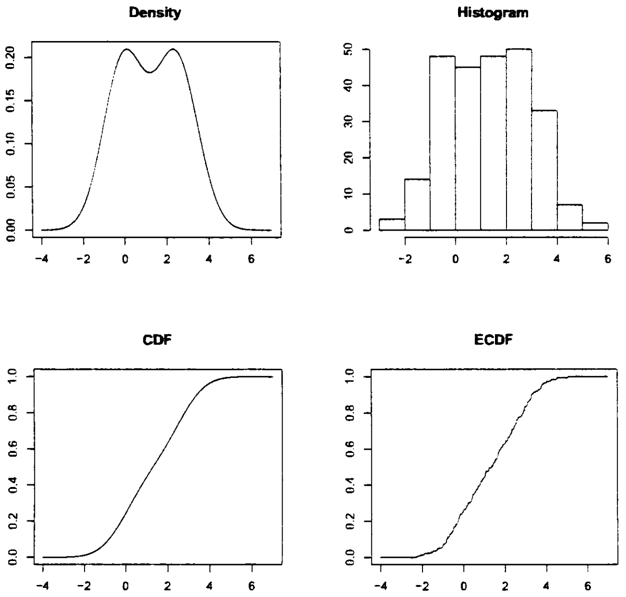

Figure 1.5: Plot of the density and cumulative distribution function of a mixture consisting of two univariate normal components with equal proportions, variance $\sigma^2 = 1$, and means $\mu_1 = 0$ and $\mu_2 = 2.5$. The empirical distribution function is also represented for a sample size of 250 from the same distribution.

data does not immediately lead to the correct conclusion about the presence of multiple modes in the distribution of the data.

§1.5. Identifiability of mixture distributions

Let $\mathcal{P}_\theta$ be a family of probability measures indexed by some unknown – finite or infinite dimensional – index $\theta \in \Theta$ which we call the parameter. We observe an experiment generated by some member of $\mathcal{P}_\theta$. The main problem of statistical inference is to infer θ based on observed data. Instead, identification is a pre-inferential problem which is devoted to assess whether with enough data at hand it is possible to state that different parameter values correspond to different probability measures $P \in \mathcal{P}_\theta$, where the meaning of the word "different" has to be specified. Roughly speaking indentifiability means that there exists a sort of one-to-one[2] correspondence between the indexes $\theta \in \Theta$ and $P \in \mathcal{P}_\theta$. The first account of identification of mixture models was given by Feller (1943) and since then many results extended that work in several directions (we shall review those results in the following paragraphs).

[2] The wording "one-to-one" has not to be taken with its strict mathematical meaning.

The standard notion of identifiability found in intermediate books (e.g. Casella and Berger, 1990, page 511, definition 11.2.1) is that: given a statistical model represented by some distribution function $F(x;\theta)$, we say that the model is identified if $\theta \neq \theta^*$ implies that $F(x;\theta) \neq F(x;\theta^*)$[3]. However identifiability is a general concept that has to be carefully defined depending on the context.
The very first definition of identifiability for finite mixtures was formalized by Teicher (1961). Let us recall the notation used in previous sections. $\mathscr{F} := \{F(x;\theta): \quad x \in \mathbb{R}^q, \theta \in \mathbb{R}^k\}$ is a family of distribution functions over $\mathbb{R}^q$ indexed by a point θ in a Borel subset of $\mathbb{R}^k$ such that $F(x;\theta)$ is measurable on $\mathbb{R}^q \times \mathbb{R}^k$. Let $G \in \mathscr{G}$ be a k-dimensional distribution function with the underlying measure assigning total mass to $\mathbb{R}^k$. Let $\mathscr{H}$ be a family of distribution functions. We consider a map $Q : \mathscr{G} \longrightarrow \mathscr{H}$, where its image is defined as $Q(G) = H$, $H(x) = \int_{\mathbb{R}^k} F(x;\theta)dG(\theta)$. Following Teicher (1961), the mixture model generated by the family $\mathscr{F}$ with mixing distribution in $\mathscr{G}$ is said to be identifiable if given $F \in \mathscr{F}$, then Q is a one-to-one map of $\mathscr{G}$ onto $\mathscr{H}$. As we have already noticed, when G is discrete, the set of all finite mixtures $\mathscr{H}$ of the family $\mathscr{F}$ is simply the convex hull of $\mathscr{F}$. Identifiability of the mixture models means that the convex hull $\mathscr{F}$ has a unique representation property, which can be translated into the following:

Definition 1.1. Let $\mathscr{H}$ be the class of finite mixtures generated by the class $\mathscr{F}$ with discrete mixing distribution. Given

$$H(x,\eta) = \sum_{j=1}^{s} \pi_j F(x;\theta_j), \quad \pi_j > 0, \theta_j \neq \theta_r \quad \forall j,r = 1,2,\ldots,s, \quad j \neq r,$$

and

$$H(x,\eta^*) = \sum_{i=1}^{z} \pi_i^* F(x;\theta_i^*), \quad \pi_i^* > 0, \theta_i^* \neq \theta_k^* \quad \forall i,k = 1,2,\ldots,z, \quad i \neq k;$$

$H(\cdot,\eta) = H(\cdot,\eta^*)$ implies that $s = z$, and there is some permutation $\bar{j}$ of the indexes $j = 1,2,\ldots,s$ such that $\pi_j = \pi^*_{\bar{j}}$ and $\theta_j = \theta^*_{\bar{j}}$, then we say that $\mathscr{F}$ generates identifiable finite mixture distributions.

The definition above has been used to study the identification of a number

[3] We replaced density functions in the definition given by Casella and Berger (1990) with distribution functions. The reason is that changing a density function on a set of points which have zero measure with respect of the underlying measure, provide us with unchanged probabilities on the support.

of models. Feller (1943) studied models based on mixtures of gamma densities. Teicher (1961) formalized the definition of identifiability for general mixture models. He extended the results in Feller (1943) showing the identifiability of finite mixtures generated by Poisson distributions. He also showed that models based on mixtures of uniform and binomial distributions are not identifiable. Teicher (1963) gave a sufficient condition for identifiability of a general class of finite mixture models and showed that mixtures based on univariate Gaussian distributions are identifiable. Yakowitz and Spragins (1968) defined identifiability for classes of finite mixtures (Definition 1.1) and gave a necessary and sufficient condition for the identifiability of such models. The main theorem in Yakowitz and Spragins (1968) states that given a discrete mixing distribution the class $\mathscr{F}$ generates identifiable mixtures if and only if $\mathscr{F}$ is a linearly independent set over the field of the real numbers. They apply their theory showing that exponential distributions, multivariate Gaussian distributions, Cauchy distributions and negative binomials generate identifiable mixture models. Atienza et al. (2006) weakened the assumption of the sufficient conditions given by Teicher (1963) and showed that mixtures of Log-Gamma distributions and mixtures of Lognormal, Gamma and Weibull distributions are identifiable with the respect to the Definition 1.1.

1.5.1 — Identifiability and estimation

Identifiability is relevant for estimation. To see why this is case, let us consider an example. Let us assume that we are estimating the parameters of a non identifiable distribution. In this case once we estimated our parameters we don't know which distribution we have estimated.

Let us go back to the Definition 1.1, and let $\eta = (\pi_1, \pi_2, \ldots, \pi_s, \theta_1, \theta_2, \ldots, \theta_s)$, $\eta \in \Gamma$ where $\Gamma \subset (0,1)^s \times \mathbb{R}^{ks}$. If a model is identifiable with respect to Definition 1.1 this means that for every $\eta \in \Gamma$ there exists a set of parameters Γ^* such that every $\eta^* \in \Gamma^*$ give the same value for the mixture distribution function, and each vector $\eta^* \in \Gamma^*$ has the same components as η but permuted according to the Definition 1.1. We just look for the existence of a permutation $\bar{j}$ where $\pi_j = \pi^*_{\bar{j}}$ and $\theta_j = \theta^*_{\bar{j}}$; this means that we identify the distribution up to "component labels switching". This last point is worth to be made as precise as possible. By the component label we mean the index j in the previous expressions. This definition implies that if we have the mixture $H_1(x) = .3F(x;\theta_1) + .7F(x;\theta_2)$ and we "switch the labels" obtaining $H_2(x) = .7F(x;\theta_2) + .3F(x;\theta_1)$, in the definition given they represent the same point of the space $\mathscr{H}$, i.e. they "identify" the same mixture distribution.

The label switching can be seen as a limitation of identifiability. In fact it means that a finite number of different parameter vectors define the same distribution. Thus when we estimate the parameter of a mixture distribution the question is which distribution we have estimated. But for each parameter we are able to list all possible parameter vectors that give the same distribution. Thus we can construct a rule for restricting the parameter space such that it contains only the vectors of parameters which give different distributions. The usual practice is to construct a function L which maps the vector η into a vector with permuted indexes $j = 1, 2, \ldots, s$ according to some specified rule (e.g. a lexicographic order). For example suppose to consider an s-component normal mixture, so that $\theta_j = (\mu_j, \sigma_j)$, where μ_j and σ_j are the mean and standard deviation parameters respectively. Let us consider the assignment rule $L(\eta) = \bar{\eta}$ $= (\pi_{\bar{1}}, \pi_{\bar{2}}, \ldots, \pi_{\bar{s}}, \theta_{\bar{1}}, \theta_{\bar{2}}, \ldots, \theta_{\bar{s}})$ where $\{\bar{1}, \bar{2} \ldots \bar{s}\}$ is a permutation of the indexes $j = 1, 2, \ldots, s$ according to the lexicographic ordering $\theta_{\bar{1}} \prec \theta_{\bar{2}} \prec, \ldots, \prec \theta_{\bar{s}}$ where $\theta_i \prec \theta_k$ if and only if $\mu_i < \mu_k$, or $\mu_i = \mu_k$ and $\sigma_i < \sigma_k$. We can suppose to look at a region of $\bar{\Gamma} \subset \Gamma$ where $\eta \in \bar{\Gamma}$ is such that $L(\eta) = \eta$. When frequentist estimation of the parameter η is considered, this approach does not cause problems. However this is not the case in the Bayesian framework where posterior simulation is used to make inference (see McLachlan and Peel, 2000a, Chapter 4). These problems are not of main interest in this work, so we will not analyze them in further details. It is however worth to mention (and in fact it will be useful in the next few chapters), that the other approach to solve the problem of label switching is to look at a parameter space alternative to Γ. That is, we can look at the topological quotient space of Γ, say $\tilde{\Gamma}$, obtained with respect to the equivalence class over elements if Γ induced by Definition 1.1. This means that all point of Γ giving an equivalent mixture distribution with respect to Definition 1.1 are collapsed to a single point in $\tilde{\Gamma}$. This is the approach used by Redner (1981) to show the strong consistency of the maximum likelihood estimator for mixture distributions. Notice that Redner's approach to consistency applies for all possible non identifiable distributions.

In our definition of mixtures we considered the case where proportions are all positive and the number of components is fixed. If this was not the case we should consider other sources of lack of identifiability. For example we can have a $(s-1)$-components mixture equivalent to an s-components mixture in two cases: (i) one of the mixing proportion in the s-components mixture is zero; (ii) two or more components in the s-components mixture are the same. Again, this type of identifiability problem can be overcome just by restricting the parameter space. In this thesis these problems are solved by restricting the class of finite mixture to the case when the number of components is fixed and the proportion parameters are all positive. And this is not a limit for our analysis.

§1.6. Estimation

Over the years, a variety of methods have been developed to estimate parameters of finite mixture distributions. In this work we will not treat the case of non-parametric estimation, but we will be interested in the cases when the model (1.1) is known up to the parameter η. Parametric estimation methods for mixture models consist of the method of moments, maximum likelihood and many minimum distance methods. The main reason for such a huge literature about mixture estimation is probably to be attributed to the computational complexity associated with it. For instance, it is well known that even in the case of the simplest homoscedastic two-normal mixture the maximum likelihood method does not lead to a closed form expression for the estimator. In this section we will briefly review some of the main methods. Maximum likelihood will be discussed extensively in the next chapter.

Pearson (1894) derived the formula for the estimate of the parameters of a homoscedastic two-normal mixture by the method of moments. Recently moment based methods for mixture distributions estimation received renewed attention after the work of Lindsay and Basak (1993) and Furman and Lindsay (1994). Quandt and Ramsey (1978) introduced a generalization, the moment generating function method, and this can be seen as a minimum distance method. That is, the mixture parameter is estimated by minimizing the square distance between the moment generating function under the model and the empirical moment generating function. The moment generating function is defined on the real line for any given parameter value, so the method above requires the choice of the points of the argument of the moment generating function over which it is computed. This choice seems to be critical, and though it does not affect consistency of the estimator, it does affect its efficiency. However Hosmer (1978) noticed that the moment generating function can be a good alternative to the maximum likelihood estimator in small samples. Kiefer (1978) argued that the moment generating function method introduced by Quandt and Ramsey (1978) performs better than the method of moment introduced by Pearson (1894). Kumar et al. (1978)

suggested to apply a minimum distance method which makes use of the characteristic function instead of the moment generating function. Recently there has been renewed interest in these methods. A complete review is given by Yu (2004).

There are several other ways to estimate the parameter of mixture models. Titterington et al. (1985) offer a comprehensive review of many of them with particular emphasis on minimum distance methods based on the distribution function. That is, the mixture parameter is estimated as the minimizer of some distance measure between the empirical distribution function and the distribution function under the model. These estimators are attractive in situations where they lead to a closed form estimator.

Beyond consistency and efficiency considerations, the main problem of estimation of mixture distributions is computability. For instance, Karlis and Xekalaki (2003) showed that the moment generating function estimator by Quandt and Ramsey (1978) is easily computable for two-normal homoscedastic mixture, while in the case of a three-normal homoscedastic mixture numerical routines hardly produce a sensible estimate. Even with the revolution of high speed computers and the fast growth of research in the field of numerical optimization, usually it is not easy to compute mixture parameter estimates easily. The maximum likelihood method is still the most popular method for estimation of finite mixtures. One of the reasons for its popularity is probably the relative ease of computations. In the case of maximum likelihood, in fact, the possibility to apply the EM algorithm (see Dempster et al., 1977) offers a great computational device where, for many mixture distributions, at each iteration of the algorithm an exact computation is performed without involving any numerical optimization routine.

§1.7. Model based clustering

Cluster analysis is mainly interested in understanding group structures in the data. Sometimes such structures are suggested by the physical or social meaning of the data, sometimes these structures are not evident and the aim of the statistical analysis is to discover them using clustering techniques. Banfield and Raftery (1993) used the wording *model based clustering* to name an approach where a group in a population under study is identified with a probability distribution and the whole population is modelled as arising from a mixture distribution. McLachlan and Basford (1988) showed the usefulness of mixture models as a way of proving an effective clustering for data from many experimental designs.

In the model based clustering approach each group in the data is assigned to a mixture component. The population of interest is assumed to have a distribution represented by (1.1) for some choice of s and the component densities. Once the parameters in (1.1) are estimated via some of the methods available the fitted model is used to establish to which component mixture each datum belongs. This approach consists in associating each group (cluster) to a component mixture, and the goal of the analysis is to establish component memberships of data. The way to do this will be explained in the next few paragraphs. Of course clustering is relevant only in the presence of unclassified data.

It is clear that model based clustering is beyond exploratory data analysis. Since statistical inference is used to discover group structures, these techniques allow to apply inferential methods to test the validity of the conclusions about the structures discovered. In fact by the use of testing techniques this approach gives us the possibility to assess in a rigorous way whether apparent clusters are due to random fluctuations in the data. In this thesis we will not be interested in testing; our aim is to estimate clusters via the statistical procedure we will describe in this section.

We assume that $\underline{X_n} := \{X_1, X_2, \ldots, X_n\}$ is a sample from a mixture distribution represented by (1.1), the n-tuple $\underline{x_n} := \{x_1, x_2, \ldots, x_n\}$ is the realization of the random variables in $\underline{X_n}$. Let $\delta : \underline{x_n} \longrightarrow \{1, 2, \ldots, s\}$ be an assignment function, that is, $\delta(x_i) = j$ means that the ith observation is assigned to the jth component mixture. The optimal Bayes assignment function is defined by

$$\delta(x_i) := \arg\max_{h \in \{1,2,\ldots,s\}} \tau_h(x_i), \tag{1.4}$$

where $\tau_j(x_i)$ is the probability that the ith observation has been drawn from the jth component. In the framework of the notation used in Section 1.5, $\tau_j(x_i) = \Pr\{Z_{ij} = 1 | x_i\}$. When the mixture density is (1.1), then $\tau_j(x_i) = \pi_j f(x_i, \theta_j) / f(x_i, \eta)$. Thus the optimal Bayes assignment function consists of assigning the the observation x_i to the component j if $\pi_j f(x_i, \theta_j) \geq \pi_k f(x_i, \theta_k)$ for all $k \in \{1, 2, \ldots, s\} \setminus \{j\}$. We notice that $\delta(x_i)$ is not always a singleton. In this case the observation can be arbitrarily assigned to one of the elements in $\delta(x_i)$ unless other restrictions are imposed. By doing this we assume that the cost of misallocation with respect to each component is equal. Given the model (1.1) the optimal Bayes assignment function can be estimated by the plug-in rule. That is, once the parameter η is estimated by $\hat{\eta}$, the assigment can be done taking an element of

$$\hat{\delta}(x_i) := \arg\max_{h\in\{1,2,\ldots,s\}}\hat{\pi}_h f(x_i, \hat{\theta}_h). \tag{1.5}$$

This approach has been extensively applied with success in many practical situations and it will be used in our investigations in the next few chapters.

§1.8. Conclusions

In this introduction we gave some basic definition for mixture models. We also tried to give a brief presentation of the many ways a mixture distribution can be used to model a given dataset. Of course our presentation was not complete in any sense. The theory of mixture distributions has been extensively studied in the last century as well as problems of identification and inference related to them. For a comprehensive introduction to mixtures with references see for example Titterington et al. (1985) or McLachlan and Peel (2000a). If the theoretical investigation of mixture models produced a huge amount of research and literature, the production of applied works is even bigger. With the revolution of high speed electronic computers the estimation of high-dimensional parameters becomes easier and applications of mixture models become more and more popular. In the previous section we described one of the many uses of mixture models, i.e. clustering. McLachlan and Peel (2000a) is a good source where a considerable number of applications are presented and references given. Beyond cluster analysis there are many fields of statistical science where mixture models are successfully used: regression theory, neural networks, hidden Markov models and many others. The two cited references are good sources to get a comprehensive overview about finite mixture models.

CHAPTER 2

Maximum Likelihood Estimation of Finite Mixtures

In this chapter we shall review the relevant literature about estimation of mixture distributions via the maximum likelihood method. Asymptotic properties of the maximum likelihood estimator will be discussed in detail. We introduce the general formulation of the EM algorithm and provide a description of its implementation in the case of finite mixtures. Convergence of the EM algorithm is also discussed.

§2.1. Introduction

In Chapter 1 we introduced some of the methods of estimation for parameters of finite mixtures with a fixed number of components where the mixing distribution is unknown and component densities are known up to their parameters. In the past, maximum likelihood estimation of finite mixtures has received limited attention due to its computational complexity. When the model satisfies the Cramér-Rao regularity conditions[1], the maximum likelihood estimator is derived solving a system of possibly nonlinear equations. In the case of finite mixtures in most cases these equations have no closed form solutions.

Baker (1940), Rao (1948) and Mendenhall and Hader (1958) derived iterative procedures to solve the maximum likelihood equations for two-normal component mixtures in the univariate case. Rao (1948) used the method of scoring,

[1] This wording basically means that the assumptions set by Cramér (1946) in order to get consistency and asymptotic normality are satisfied. This also means that assumptions that allow us to write the Fisher information in terms of first derivatives of the log-likelihood functions are also satisfied.

while Mendenhall and Hader (1958) relied on Newton-Raphson methods. Despite the successful implementation in the case of a two-normal mixture, these iterative methods for solving equations were unsatisfying in cases. As computer power become available with less constraints, maximum likelihood estimation was explored for a variety of mixture models. Mixture of Gaussian distributions were the most studied. Hasselblad (1966) studied the the maximum likelihood estimator for mixtures with an arbitrary but finite number of components. A two-normals multivariate mixture with common but unknown covariance matrix was first studied by Day (1969) and then by John (1970). The case with an arbitrary number of multivariate Gaussian components was studied by Wolfe (1970) and in more detail by Peters and Walker (1978). Redner (1981) proved the strong consistency for non-identifiable distributions for a class of estimators which contains the maximum likelihood as a special case. The work by Redner (1981) applies to a general class of distributions which includes finite mixtures. Hathaway (1985) proposed a constrained maximum likelihood estimator for univariate finite mixtures of Gaussian distributions and studied the strong consistency of such an estimator.

Gaussian mixtures occupied a considerable space in the literature about mixture distributions. The book by McLachlan and Peel (2000a) gives an overview of many applications where the maximum likelihood method is applied to non-normal mixture models. In the most part of the literature cited above the maximum likelihood estimator is studied for experimental designs with unclassified observations. Hosmer (1978) investigated the maximum likelihood estimator for many experimental designs where classified, unclassified and also partially classified data are available. Finally we note the paper by Laird (1978) where it is shown that under various regularity assumptions the maximum likelihood estimate of a mixture with possibly an infinite number of components is actually a finite mixture density.

§2.2. Maximum likelihood estimates

In this section we will define maximum likelihood estimates for mixture distributions. Let us assume that X_i, $i = 1, 2, \dots, n$, are random variables defined on $\mathbb{R}^q$, $\underline{X_n} = \{X_1, X_2, \dots, X_n\}$ is an i.i.d. sequence from a distribution having density

$$f(x, \eta) = \sum_{j=1}^{s} \pi_j f_j(x, \theta_j), \tag{2.1}$$

where f_j is a density on $\mathbb{R}^q$, $s < +\infty$, $0 < \pi_j \leq 1$, $\sum_{j=1}^{s} \pi_j = 1$, $\theta_j \in \mathbb{R}^{p_j}$ for some $p_j \geq 1$. The number of components s is fixed and known. We set $\eta = (\pi_1, \pi_2, \ldots, \pi_s, \theta_1, \theta_2, \ldots, \theta_s)$, with $\eta \in \Gamma$. Let $\underline{x_n} = \{x_1, x_2, \ldots, x_n\}$ an observed sample, i.e. a realization of the sequence of random variables $\underline{X_n}$. The likelihood function associated with the sample is given by

$$L(\eta, \underline{x_n}) = \prod_{i=1}^{n} \sum_{j=1}^{s} \pi_j f_j(x_i, \theta_j). \tag{2.2}$$

The maximum likelihood estimator is defined as

$$\hat{\eta}_n := \arg\max_{\eta \in \Gamma} L(\eta, \underline{x_n}). \tag{2.3}$$

The likelihood above is suited for all those situations where we have unclassified data coming from populations of which the distributions are represented by the densities f_j $j = 1, 2 \ldots, s$. The maximum likelihood estimator is equivalent to

$$\hat{\eta}_n := \arg\max_{\eta \in \Gamma} l(\eta, \underline{x_n}),$$

where $l(\eta, \underline{x_n})$ is the log-likelihood function, i.e.

$$l(\eta, \underline{x_n}) = \sum_{i=1}^{n} \log \left[\sum_{j=1}^{s} \pi_j f_j(x_i, \theta_j) \right]. \tag{2.4}$$

The maximum likelihood estimator is probably the most popular statistical tool and a huge amount of literature has been devoted to it over the years. Particular attention has been given to the study of the maximum likelihood estimator for mixtures where a number of issues arise relating to analytical and computational problems.

The first problem is that we cannot always rely on the fact that the estimator defined in (2.3) exists. The log-likelihood function may be unbounded over Γ so that $l(\eta, \underline{x_n})$ does not achieve a global maximum. The simplest example of such a problem is when the density in (2.1) is made up to $s > 1$ univariate normal components. In this case setting one of the component's mean equal to one of the observations, say $\mu_1 = x_1$, and letting the corresponding standard deviation $\sigma_1 \longrightarrow 0$, makes the likelihood $L(\eta, \underline{x_n}) \longrightarrow +\infty$ given all the other parameters fixed. There are a number of solutions to this problem. In the case of normal mixtures DeSarbo and Cron (1988) proposed a constrained maximum likelihood estimator where the standard deviations of the component mixtures are larger than a positive constant, i.e. they considered constraint of the type $\sigma_j > c_0 > 0$ for $j = 1, 2, \ldots, s$ and some fixed and known c_0. Hathaway (1985) studied a different type of constraints on the variances with $\min_{i,j}(\sigma_i/\sigma_j) = c_0 > 0$. He showed that the corresponding maximum likelihood estimator is strongly consistent, i.e. it converges to the true parameter almost surely. While the constraints proposed by DeSarbo and Cron (1988) do not lead to scale equivariant maximum likelihood estimates, the constraints studied by Hathaway (1985) do. Recently another approach to deal with the unboundness of likelihood function has been introduced by Ciuperca et al. (2003). The authors proposed a penalized maximum likelihood estimator and showed that it is consistent and asymptotically normal. We remark that in the case of normal mixtures the existence of the maximum likelihood estimate occurs in the case of previously classified data.

In the case that the number of components s is fixed and the mixture model is defined with mixing proportions being larger than zero as in (2.1), then there is still a sort of lack of identifiability due to label switching. If the distribution represented by (2.1) is identifiable with the respect to definition 1.1 in chapter 1, each parameter η obtained by permuting the pairs (π_j, θ_j) gives us an equivalent distribution. If the likelihood is maximized at the true parameter value, say η^0, the maximum is not unique because for every permutation of the pairs (π_j^0, θ_j^0) we get a parameter vector giving the same likelihood value of η^0.

Multiple maxima occur also for reasons not related to lack of identifiability. Usually the likelihood surface of a mixture has many local maxima and flat regions. When the estimator has no closed form solution, numerical computational methods have to be used. These methods are usually able to find a stationary point for the likelihood and it is not always the case that one of these points is the largest local maxima.

The traditional approach to find the maximum likelihood estimator, is to first derive a system of *likelihood equations* which have to be satisfied by the estimates. If the mixture density is continuous and differentiable on the parameter space and under other regularity conditions, the estimator defined in (2.3) should satisfy the system of equations

$$\nabla_\eta l(\hat{\eta}; \underline{x_n}) = 0 \tag{2.5}$$

where $\nabla_\eta l(y; \underline{x_n})$ is the vector of partial derivatives of l with respect of η computed at point $\eta = y$ for a given observed sample $\underline{x_n}$. The conditions in the following analysis are derived in Peters and Walker (1978) and they come from the application of Kuhn-Tucker's sufficient conditions for a constraint maximum (see Bazaraa et al., 2006, Chapter 4). Since the mixing proportions are constrained to be in (0,1) and $\sum_{j=1}^{s} \pi_j = 1$, the likelihood equations for the mixing proportions should also satisfy the condition

$$\nabla_\pi l(\hat{\eta}; \underline{x_n})^{\mathrm{T}}(\pi - \hat{\pi}) \leq 0 \qquad \forall \pi \neq \hat{\pi},$$

where $\pi = (\pi_1, \pi_2, \ldots, \pi_s)$, $\sum_{j=1}^{s} \pi_j = 1$, $\pi_j \geq 0$ for all $j = 1, 2 \ldots, s$, and $\hat{\pi}$ is the corresponding maximum likelihood estimator. Let u_j be an s-dimensional vector that has all components equal to zero unless the jth component which is one. The condition above implies that

$$\nabla_\pi l(\hat{\eta}; \underline{x_n})^{\mathrm{T}}(u_j - \hat{\pi}) \leq 0; \qquad \text{all} \quad j = 1, 2, \ldots, s; \tag{2.6}$$

with equality for all those j for which $\hat{\pi}_j > 0$. The latter is equivalent to

$$\frac{1}{n} \sum_{i=1}^{n} \frac{f_j(x_i; \hat{\theta}_j)}{f(x_i; \hat{\eta})} \leq 1; \qquad \text{all} \quad j = 1, 2, \ldots, s;$$

with equality for all those j for which $\hat{\pi}_j > 0$. Multiplying each side of the expression above by $\hat{\pi}_j$ yields likelihood equations in the form

$$\hat{\pi}_j = \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{\pi}_j f_j(x_i; \hat{\theta}_j)}{f(x_i; \hat{\eta})}; \qquad \text{all} \quad j = 1, 2, \ldots, s; \tag{2.7}$$

The latter equation plays an important role especially in context of the EM algorithm which we will review in the following chapters. By considering the matrix of the second partial derivatives of $l(\eta; \underline{x_n})$ with respect of π (provided that they exist), it is easy to verify its concavity for any fixed values of $\theta_1, \ldots, \theta_s$. It follows that for any fixed value for the parameters $\theta_1, \ldots, \theta_s$, the equations in (2.6) are necessary and sufficient for a maximum with respect to the proportion parameters. The set of equations in (2.5) however are necessary but in general not sufficient for the maximum with the respect to all parameters.

In general likelihood equations for maximum likelihood estimates of mixture parameters have no closed form solution. This happens for example in the case of Gaussian mixtures. Numerical routines are used to compute maxima of $l(\eta; \underline{x_n})$ and in many cases numerical methods are only able to provide a local maximum.

However, depending on the shape of the likelihood surface there could be a huge number of local maxima, and each time the local maximum giving the largest likelihood value has to be chosen.

§2.3. Asymptotic properties for maximum likelihood estimates

Maximum likelihood estimates are usually strongly consistent and asymptotically normal. By strong consistency we mean that the estimator converges almost sure to the true parameter which generated the sample. There are two useful approaches to show strong consistency of the maximum likelihood estimator. One approach invented by Cramér (1946) is to assume some restriction about derivatives and moments of the log-likelihood function and then to use these conditions to show strong consistency and asymptotic normality by use of a Taylor expansion of the log-likelihood function. The other approach followed by Wald (1949) is not to make any assumption about the differentiability of the log-likelihood function and to show strong consistency under set of very general conditions mainly involving the expected value of the log-likelihood. Wald's approach is not able to lead to statements about the convergence in distribution of the estimator. Here we will review the two approaches since this will be useful afterward. The treatment we will give is quite general and does not only apply to finite mixture models. We will continue to maintain the notation as in the previous sections, but it is understood that the following statements are valid for any maximum likelihood estimator for which the assumptions given here with respect to the density and log-likelihood function are valid.

First we will consider the case where differentiability assumptions are made. The notation η_0 will indicate the parameter which generated the sample $\underline{x_n}$ according to the model described in (2.1). Also $F(x;\eta)$ is the distribution function of (2.1). Sometimes, when not ambiguous, we will indicate $f_0 = f(x;\eta_0)$, $F_0 = F(x;\eta_0)$

Assumption 2.1. The support of f does not depend on the parameter value, i.e. the set $S_f := \{x : \quad f(x;\eta) > 0\}$ does not depend on $\eta \in \Gamma$.

Assumption 2.2. The parameter space Γ is open in $(0,1)^s \times \mathbb{R}^{\sum_{j=1}^{s} q_j}$.

Assumption 2.3. With $\mathscr{N}_\varepsilon(\eta') := \{\eta \in \Gamma : \quad \|\eta' - \eta\| \leq \varepsilon\}$, for any $\varepsilon > 0$,

$$\sup_{\eta \in \mathscr{N}_\varepsilon(\eta^0)} \left\| \frac{\partial f(x;\eta)}{\partial \eta} \right\| \leq \infty,$$

$$\sup_{\eta \in \mathcal{N}_\varepsilon(\eta^0)} \left\| \frac{\partial^2 f(x;\eta)}{\partial \eta \partial \eta^{\mathsf{T}}} \right\| \leq \infty.$$

Assumption 2.4. The function $l(\eta; \underline{x_n})$ is bounded above on Γ, it is differentiable up to the third order, with continuous first and second order derivatives.

Assumption 2.5. For all $\eta \in \Gamma$,

$$\left| \frac{\partial \log f(x;\eta)}{\partial \eta_i} \right| \leq t_i(x), \left| \frac{\partial^2 \log f(x;\eta)}{\partial \eta_i \partial \eta_j} \right| \leq t_{ij}(x), \left| \frac{\partial^3 \log f(x;\eta)}{\partial \eta_i \partial \eta_j \partial \eta_k} \right| \leq t_{ijk}(x),$$

for $t_i(x), t_{ij}(x), t_{ijk}(x)$ integrable functions for all $i, j, k = 1, 2, \ldots, s$.

Assumptions 2.4 and 2.5 are needed in order to exchange derivatives and integrals (by applying the dominated convergence theorem). Under the assumptions above the Fisher information at the true parameter is well-defined. This means that $I(\eta_0) < +\infty$ where

$$I(\eta) := \int \left[\frac{\partial \log f(x;\eta)}{\partial \eta} \right] \left[\frac{\partial \log f(x;\eta)}{\partial \eta} \right]^{\mathsf{T}} dF_0. \tag{2.8}$$

Theorem 2.1. *Under assumptions 2.1-2.5, as $n \longrightarrow \infty$, there exists a sequence $\{\eta_n^*\}_{n \geq 1}$ of solutions of the likelihood equations* (2.5) *such that (i) $\eta_n^* \xrightarrow{as} \eta_0$; (ii) $\sqrt{n}(\eta_n^* - \eta_0) \xrightarrow{\mathcal{L}} \text{Normal}(0, I(\eta_0)^{-1})$.*

We will not prove the previous theorem because it is part of the standard statistical literature at an advanced level. This same theorem was proved in Cramér (1946) for general smooth maximum likelihood estimation, and then extended by Chanda (1954) to the multidimensional case. This theorem is only local, in the sense that it states that there exists a sequence of solutions of the likelihood equations which converges to the true parameter with probability one as the size of the sample gets infinitely large. Also, a $\sqrt{n}$–scaled version of this sequence is asymptotically normally distributed. The set of Assumptions above are fulfilled in many situations, for instance for Gaussian mixtures.

Wald (1949) approached the issue of the convergence of the maximum likelihood estimator in a more general setup looking for a global characterization of the problem. The first question to answer is whether the sequence of roots of log-likelihood equations converges to the maximum likelihood estimates. Moreover Theorem 2.1 does not say anything about whether the limit point of the sequence of solutions of the likelihood equations is unique. However Wald's theory is not applicable when the distribution which generates the sample is not globally identifiable. As we noted in Chapter 1 identifiability is a concept of equivalence referred to some family of probability distributions. The concept of identifiability can be defined depending on the class of problem we are considering. Suppose that we have a statistical experiment produced by drawing observations from a distribution function $F(x;a)$, where $a \in A$ and F belongs to some family of distributions $\mathcal{F}$. By "global identifiability" we mean the following: if $a \neq a^*$ this implies that $F(x;a) \neq F(x;a^*)$ for at least one value of x in the support of F. It is clear that the definition in 1.1 is not about global identifiability. In fact, as already noted, in the case of the finite mixture distribution represented by (2.1), any permutations of the pairs (π_j, θ_j) will give the same value for the distribution. Redner (1981) extended Wald's approach to cases where global identifiability does not hold.

Here we present the theorem by Redner (1981) which is useful to establish strong convergence for maximum likelihood sequences of estimates of many mixtures models. Before giving the main theorem we need to state the set of assumptions needed and some new notation. If not stated, the notation used in the previous presentation is used. Again, we adapt the presentation to our finite mixtures setup, however the following theory is applicable to more general models where the following assumptions hold. Let $\mathcal{N}_r(\eta') = \{\eta \in \Gamma : \quad \|\eta - \eta'\| \leq r\}$ be a closed ball of radius $r > 0$. For $\eta \in \Gamma$ and $r, s > 0$ we denote

$$f(x,\eta,r) = \sup_{\eta' \in \mathcal{N}_r(\eta)} f(x,\eta'); \qquad f^*(x,\eta,r) = \max\{1, f(x,\eta,r)\},$$

$$h(x,s) = \sup_{\eta \notin \mathcal{N}_s(\eta_0)} f(x,\eta); \qquad h^*(x,s) = \max\{1, h(x,s)\}.$$

Assumption 2.6. Γ is a compact set.

Let $\tilde{\Gamma}$ be the quotient topological space obtained by collapsing the set $C(\eta') := \{\eta \in \Gamma : F(\cdot;\eta) = F(\cdot;\eta')\}$ in a point $\tilde{\eta}' \in \tilde{\Gamma}$.

Assumption 2.7. For each $\eta \in \Gamma$ and sufficiently small r and sufficiently large s,

$$\int \log f^*(x,\eta,r) dF_0(x) < \infty \tag{2.9}$$

$$\int \log h^*(x,s) dF_0(x) < \infty \tag{2.10}$$

Assumption 2.8. For any sequence $\{\eta_n\}_{n\geq 1}$, if $|\eta_n| \longrightarrow \infty$, then $f(x;\eta_n) \longrightarrow 0$ at any x, except perhaps on a set X which has zero measure according to F_0, and does not depend on the sequence $\{\eta_n\}_{n\geq 1}$.

Assumption 2.9. For any sequence $\{\eta_n\}_{n\geq 1}$ if $\eta_n \longrightarrow \eta$, then $f(x;\eta_n) \longrightarrow f(x;\eta)$ at any x, except perhaps on a set X which has zero measure according to F_0, and it may depend on the limit point η but not on the sequence $\{\eta_n\}_{n\geq 1}$.

Assumption 2.10. $\int |\log f(x;\eta_0)|\, dF_0(x) < \infty$.

Theorem 2.2 (Wald, Redner). *If assumptions 2.6-2.10 are satisfied, then the sequence $\{\hat{\eta}_n\}_{n\geq 1}$ of maximum likelihood estimates is strongly consistent, i.e. $\hat{\tilde{\eta}}_n \xrightarrow{as} \tilde{\eta}_0$.*

Roughly, Wald-Redner's theorem above ensures that under some condition the maximum likelihood estimates is strongly consistent for the set of parameters $C(\eta_0)$ as defined in assumption 2.6. However, in order to get the previous result the sequence of the maximizers of the likelihood has to exist, i.e. the likelihood has to have a maximum. For some popular mixture distributions this does not happen, for instance the case of mixtures of univariate Gaussian distributions when the parameter set is not suitably restricted. In these cases a constrained maximum likelihood estimate has to be considered.

Asymptotic results are only meaningful for large samples. Hence in order to make sense of these results in practical applications we need to understand what happens when n is not large enough. In small samples simulation results showed that when the components are not well separated the maximum likelihood method does not provide very accurate estimates. Redner and Walker (1984) presented a simulation study where the maximum likelihood estimates are computed for a model with two univariate homoscedastic normal components with proportion parameter $\pi = 0.3$, standard deviation $\sigma = 1$ and $\mu_1 - \mu_2 = 1$. They showed that a sample size of 10^6 observations is needed to ensure that the estimated standard deviation of each component of the estimated vector of parameters $(\pi_1, \mu_1, \mu_2, \sigma)$ is less than or equal to 0.1. Several authors addressed the issue of the poor performance of the maximum likelihood estimates when the sample size is small. Hosmer (1973) argued that with poor separation of the components and with a sample size of about hundred observations, maximum likelihood estimates have to be handled with extreme care. On the other hand Day (1969) and Hasselblad (1966) found that the maximum likelihood estimator performs better than the method of moments estimator especially when the components are poorly separated. Hosmer (1978) argued that the moment generating function method proposed by Quandt and Ramsey (1978) outperforms the maximum likelihood estimator in small samples.

§2.4. The EM algorithm

The expectation maximization algorithm, also known as EM algorithm, is a widely used device to compute maximum likelihood estimates in the case of incomplete data. In many situations the algorithm provides sequence of steps at which exact computations are made, and the iterative procedure leads to a local maximum of the likelihood function. The algorithm was first developed for a number of special cases and the first unifying theory w$H(\phi, \phi') := \int_{\mathcal{Y}(x)} [\log k(y|x;\phi)] k(y|x;\phi') dy,$ (1977). Here we present the general theory and then we will give an overview of some issues about its application to finite mixtures models.

Let $\mathcal{Y}$ be a measurable space called the "complete data space", $y \mapsto x(y)$ is a measurable map of $\mathcal{Y}$ onto a measurable space $\mathcal{X}$ called the "incomplete data space". Let $f(y;\phi)$ be a density function over $\mathcal{Y}$, with $\phi \in \Omega$ an indexing parameter for f. The density $g(x;\phi)$ is the density induced by $f(y;\phi)$ through the map $x(y)$. The ultimate goal here is to have a maximum likelihood estimate of ϕ. For a given $x \in \mathcal{X}$ the aim of the EM algorithm is to maximize the "incomplete data log-likelihood" $L(\phi) = \log g(x;\phi)$ over Ω. Let $\mathcal{Y}(x) := \{y \in \mathcal{Y} : x(y) = x\}$. The conditional density $k(y|x;\phi)$ on $\mathcal{Y}$ is given by

$$k(y|x;\phi) = \frac{f(y;\phi)}{g(x;\phi)}, \quad \text{for} \quad x(y) \in \mathcal{X}.$$

For any $\phi, \phi' \in \Omega$ we c$Q(\phi, \phi') := \int_{\mathcal{Y}(x)} [\log f(y;\phi)] k(y|x;\phi') dy$

$$Q(\phi, \phi') := \int_{\mathcal{Y}(x)} [\log f(y;\phi)] k(y|x;\phi') dy, \tag{2.11}$$

and

$$H(\phi, \phi') := \int_{\mathcal{Y}(x)} [\log k(y|x;\phi)] k(y|x;\phi') dy, \tag{2.12}$$

provided that the integrals above are well-defined. By properties of logarithms it easy to see that $L(\phi) = Q(\phi, \phi') - H(\phi, \phi')$.

Let $t = 0, 1, ..T$ denote the iteration index, and $\phi^{(t)}$ the value of ϕ derived at the tth iteration. Let $\phi^{(0)}$ be an arbitrarily fixed initial value for the parameter of interest. The algorithm is as follows:

1. fix $\phi^{(0)} \in \Omega$;

2. For all $t = 1, 2, \ldots$ do the following up to convergence:

(a) E–step: determine $Q(\phi, \phi^{(t)})$;

(b) M–step: choose $\phi^{(t+1)} = \arg\max_{\phi \in \Omega} Q(\phi, \phi^{(t)})$.

In order to be well-defined the algorithm needs that the maximization in the M–step to be well-defined, i.e. we need to assume that $\arg\max_{\phi \in \Omega} Q(\phi, \phi')$ exists for any choice of $\phi' \in \Omega$. The key fact about the EM algorithm lies in the fact that at each iteration the likelihood is increased, i.e. $L(\phi^{(t+1)}) \geq L(\phi^{(t)})$, $t = 0, 1, \ldots$. This is because the M–step ensures $Q(\phi^{(t+1)}, \phi^{(t)}) \geq Q(\phi^{(t)}, \phi^{(t)})$ and by Jensen's inequality $H(\phi^{(t+1)}, \phi^{(t)}) \leq H(\phi^{(t)}, \phi^{(t)})$. This fact implies that the sequence of log-likelihood values at each step $\{L^{(t)}\}_{t \geq 0}$ is monotonically increasing. This is a fundamental property that will be of central importance to show convergence of the algorithm.

2.4.1 — Convergence of the EM algorithm

In this section we ask ourself whether convergence of the algorithm is achieved, and what can be said about $\{\phi^{(t)}\}_{t \geq 0}$. Convergence of the algorithm has been established first by Dempster et al. (1977). Their paper contained some mistakes. Wu (1983) corrected those mistakes and presented the theory that we will review here. In general it is not true that the EM algorithm converges to a point of the parameter space where the log-likelihood function is globally maximized. As we have already pointed out, usually the log-likelihood function has several local maxima. It can also have flat regions where the log-likelihood has very tiny variations. Since the algorithm should stop when two successive iterations differ by less than a small value it is likely that it stops at some point which might not even be a local maximum.

Redner and Walker (1984) gave a portmanteau theorem about convergence which we will report here. This theorem essentially summarizes the paper of Wu (1983). We assume that (2.11) and (2.12) are well-defined for every $\phi, \phi' \in \Omega$ and that $\arg\max_{\phi \in \Omega} Q(\phi, \phi')$ exists for any choice of $\phi' \in \Omega$.

Theorem 2.3. *For a given $\phi^{(0)} \in \Omega$, let $\{\phi^{(t)}\}_{t \geq 0}$ be a sequence in Ω satisfying*

$$\phi^{(t+1)} \in \arg\max_{\phi \in \Omega} Q(\phi, \phi^{(t)}) \qquad t = 1, 2, \ldots;$$

Then $\{L(\phi^{(t)})\}_{t \geq 0}$ increases monotonically to a limit L^* (possibly infinite). Furthermore, denoting by $\mathcal{L}$ the set of limit points of $\{\phi^{(t)}\}_{t \geq 0}$ in Ω, one has the following:

1. *$\mathcal{L}$ is closed in Ω;*

2. *If $\left\{\phi^{(t)}\right\}_{t\geq 0}$ is contained in a compact subset of Ω, then $\mathcal{L}$ is compact;*

3. *If $\left\{\phi^{(t)}\right\}_{t\geq 0}$ is contained in a compact subset of Ω and $\left\|\phi^{(t)} - \phi^{(t-1)}\right\| \to 0$ as $t \to \infty$ for a suitable norm $\|\cdot\|$ on Ω, then $\mathcal{L}$ is connected as well as compact;*

4. *If $L(\phi)$ is continuous in Ω and $\mathcal{L} \neq \emptyset$, then L^* is finite and $L(\hat{\phi}) = L^*$ for $\hat{\phi} \in \mathcal{L}$;*

5. *If $Q(\phi, \phi')$ and $H(\phi, \phi')$ are continuous in ϕ and ϕ' and differentiable at $\phi = \phi' = \hat{\phi} \in \mathcal{L}$, then L^* is a stationary point of L and the likelihood equations $\nabla_\phi L(\phi) = 0$ are satisfied at $\hat{\phi} \in \mathcal{L}$.*

The theorem above characterizes the set of all limit points of the sequence $\left\{\phi^{(t)}\right\}_{t\geq 0}$ provided by the EM algorithm. However, the theorem does not guarantee that L^* is the global maximum of the log-likelihood function when it exists. The theorem developed above only says that under some regularity conditions the EM algorithm converges to a point of stationarity of the log-likelihood function. This means that the question of whether a converging sequence of iterations leads to the maximum likelihood estimate remains unanswered. Wu (1983) highlighted that the question can only be answered in particular cases that are difficult to check. For example when $L(\phi)$ is unimodal in Ω and continuity assumptions over Q are satisfied, then the sequence $\left\{\phi^{(t)}\right\}_{t\geq 0}$ produced by the EM algorithm converges to the unique global maximum of L which coincides with the maximum likelihood estimator. Also when $Q(\phi^{(t+1)}, \phi^{(t)}) > Q(\phi^{(t)}, \phi^{(t)})$ at each iteration it is possible to show that L^* in the theorem 2.3 is at least a local maximum. However the condition $Q(\phi^{(t+1)}, \phi^{(t)}) > Q(\phi^{(t)}, \phi^{(t)})$ is not easily verifiable and in general it is not easy to assess whether the sequence provided by the EM algorithm converges to the global maximum of the log-likelihood.

2.4.2 — EM algorithm for mixture distributions

Computation of the maximum of the likelihood of a sample of unclassified data when the population is assumed to be distributed according to a finite mixture distribution is not easy. We interpret the sample of un-classified data as an incomplete data vector where the component label is the missing information. We are interested in the situation described in Section 1.3.2 in Chapter 1, and all the notations remain the same if not otherwise specified.

The observed vector x_i is being viewed as incomplete because the associated vector of component-labels z_i is missing. The complete data is given by the $2n$-tuple of vectors $y_c = (x_1, x_2, \dots, x_n, z_1, z_2, \dots, z_n)$. Here the n-tuple $z = (z_1, z_2, \dots, z_n)$ are considered as realizations of the random variables $Z_1, Z_2, \dots, Z_n$ previously described. As we already noticed, when $X_1, X_2, \dots, X_n$ are i.i.d., we will be willing to assume that $Z_1, Z_2, \dots, Z_n$ are distributed according to multinomial distribution (see (1.3)). We denote the n-tuple of incomplete data with $y_o = (x_1, x_2, \dots, x_n)$. The corresponding complete data likelihood is given by

$$L_c(\eta) := \prod_{i=1}^{n} \prod_{j=1}^{s} (\pi_j f_j(x_i, \eta_j))^{z_{ij}} ;$$

thus the complete data log-likelihood is given by

$$l_c(\phi) = \log L_c(\eta) = \sum_{j=1}^{s} \sum_{i=1}^{n} z_{ij} \left(\log \pi_j + \log f_j(x_i, \theta_j)\right). \qquad (2.13)$$

The definition of the EM algorithm only requires to derive the Q function (E–step) being maximized at each iteration (M–step). In the E–step we just have to take the conditional expectation of the complete data log-likelihood given the observed data and the current value of the parameter at iteration t. Hence

$$Q(\eta, \eta^{(t)}) = \mathrm{E}_{\eta^{(t)}} \left[l_c(\eta | y_o) \right]. \qquad (2.14)$$

The notation $\mathrm{E}_{\eta^{(t)}}$ means that the expectation is taken using $\eta^{(t)}$. Now $\mathrm{E}_{\eta^{(t)}}[Z_{ij} | y_o] = \mathrm{Pr}_{\eta^{(t)}} \{Z_{ij} = 1 | y_o\} = \tau_j(x_i; \eta^{(t)})$. Here $\tau_j(x_i; \eta^{(t)})$ is the posterior probability that the observation x_i has been produced by the jth component mixture with the posterior probability evaluated at the current approximation of the parameter $\eta^{(t)}$; that is

$$\tau_j(x_i; \eta^{(t)}) = \frac{\pi_j^{(t)} f_j(x_i; \theta_j^{(t)})}{f(x_i; \eta_j^{(t)})}, \qquad (2.15)$$

for $j = 1, 2, \dots, s$ and $i = 1, 2, \dots, n$. Using (2.13), (2.14) and (2.15) we have that

$$\begin{aligned}
Q(\eta, \eta^{(t)}) &= \mathrm{E}_{\eta^{(t)}}\left[\sum_{j=1}^{s}\sum_{i=1}^{n} z_{ij}(\log \pi_j + \log f_j(x_i, \theta_j))|y_o\right] \\
&= \sum_{j=1}^{s}\sum_{i=1}^{n} \mathrm{E}_{\eta^{(t)}}[z_{ij}|y_o]\,(\log \pi_j + \log f_j(x_i, \theta_j)) \\
&= \sum_{j=1}^{s}\sum_{i=1}^{n} \tau_j(x_i; \eta^{(t)})\,(\log \pi_j + \log f_j(x_i, \theta_j)) \qquad (2.16)
\end{aligned}$$

Given the current value of the parameter $\eta^{(t)}$, at the M–step we choose $\eta^{(t+1)}$ as the maximizer of $Q(\eta, \eta^{(t)})$. For most finite mixture distributions θ_j only affects the density f_j so that the M–step consists of performing a number of separate maximizations. The $t+1$st M–step for a general finite mixture distribution has solution

$$\pi_j^{(t+1)} = \frac{1}{N}\sum_{i=1}^{n} \tau_j(x_i, \eta^{(t)}), \qquad (2.17)$$

and

$$\theta_j^{(t+1)} = \arg\max_{\theta_j} \sum_{i=1}^{n} \tau_j(x_i, \eta^{(t)}) \log f_j(x_i; \theta_j), \qquad (2.18)$$

$j = 1, 2, \ldots, s$. Equation (2.17) ensures that at each step the proportions computed are non-negative and sum up to one. This is not surprising because the expression in (2.17) is exactly the same as (2.7). The latter was derived solving the likelihood equations, and taking into account the constraints over the mixing proportion parameters. At each iteration the proportion relative to the jth component is computed as the "average" posterior probability, that the observations x_i, $i = 1, 2, \ldots, n$, belong to the jth component mixture. The solution of the maximization problem in (2.18) has also a nice interpretation. In fact, $\theta_j^{(t+1)}$ in (2.18) can be seen as a weighted maximum likelihood of θ_j when the whole population is distributed according to F_j. Thus, $\theta_j^{(t+1)}$ has usually a closed form expression which makes the EM algorithm appealing.

In the case that $f_j(x; \theta_j)$ is the density of a Gaussian distribution for all $j = 1, 2, \ldots, s$; with $\theta_j = (\mu_j, \sigma_j)$ where μ_j is the mean parameter and σ_j the standard deviation, the M–step with respect of the Gaussian parameters is simple to compute

$$\mu_j^{(t+1)} = \frac{1}{\sum_{i=1}^{n} \tau_j(x_i, \eta^{(t)})} \sum_{i=1}^{n} x_i \tau_j(x_i, \eta^{(t)}), \tag{2.19}$$

and

$$\sigma_j^{(t+1)} = \sqrt{\frac{1}{\sum_{i=1}^{n} \tau_j(x_i, \eta^{(t)})} \sum_{i=1}^{n} (x_i - \mu_j^{(t+1)})^2 \tau_j(x_i, \eta^{(t)})}. \tag{2.20}$$

As explained in the previous section the EM algorithm does not always converge to the global maximum of the log-likelihood (provided that it exists). A convergent sequence of estimates produced by the EM algorithm can either converge to a local maximum, to a global maximum or to a point at which the log-likelihood is flat. Very often the log-likelihood function has several local peaks due to the random fluctuations in the sample at hand. One should always run the EM algorithm for a set of different starting values, and then choose the EM solution which gives the largest log-likelihood value. Seidel et al. (2000) showed that different strategies for starting values can lead to quite different estimates in the context of mixtures of exponential distributions. In their book McLachlan and Peel (2000a) argued that in the case of multivariate normal mixtures the definition of appropriate starting values for the proportion parameters and variance matrices is of primary importance in order to get reasonably good estimates. Several methods of fixing starting values have been proposed in the literature. McLachlan and Peel (2000a) review most of them. Karlis and Xekalaki (2003) conducted a simulation study in order to compare the estimates obtained by the EM algorithm using different existing strategies to fix starting values. They also propose a new method based on a partitioning strategy of the data. Their simulation study is about mixtures of two and three univariate Gaussian components with equal variances. One of the main findings in the paper by Karlis and Xekalaki (2003) is that even for the case of a two-component homoscedastic Gaussian mixture, some of the strategies commonly used in practice (e.g. random starting values) can lead to estimates which are quite far from the underlying true parameter even for large sample size. They also showed that in the case of a three components homoscedastic Gaussian mixture all the existing methods including the one they proposed have difficulties when the proportions are not equal across the components. This is to stress that convergence of the EM algorithm is not about global convergence to the global maximizer of the likelihood. Different starting values can produce different estimates so that computations and conclusions have to be made with extreme care.

§2.5. Conclusions

In this chapter we gave a summary of the estimation theory for finite mixture models when the number of components is fixed and known. We discussed several estimation methods that have been proposed in the literature and we highlighted the important role of the maximum likelihood method. We stressed the importance of identifiability and we discussed some of the important issues that relate identifiability to estimation. We presented the maximum likelihood theory assuming differentiability of the log-likelihood function. We also stated the strong consistency of the maximum likelihood estimator in a more general setup without differentiability assumptions. Even if the maximum likelihood estimator has very good asymptotic properties, in many situations the researcher must take care of the sample size. In fact, several authors have shown that this estimator can exhibit poor performance when the sample size is not large enough. We presented the theory of the EM algorithm for finite mixtures and we also highlighted several issues about how a wrong choice of the starting values can affect its performance.

CHAPTER 3

Identification and Estimation of Location-Scale Mixtures with Uniform Noise

In this chapter we shall study a finite mixture model where uniform components are mixed with distributions belonging to some location-scale family. This class of models has been introduced by Fraley and Raftery (1998) in cluster analysis and have been suggested as a device to achieve robustness. Hennig (2004) studied robustness properties of several mixture models including the one we will study when there is just one uniform component. We will study identifiability, maximum likelihood estimation and computation via the EM algorithm.

§3.1. Introduction

Maximum likelihood estimation of mixtures of Gaussian distributions is a widely used technique in cluster analysis, classification and density estimation. In cluster analysis maximum likelihood for Gaussian mixtures was first studied by Wolfe (1967) and Day (1969); a comprehensive source on the topic is the book by McLachlan and Peel (2000a). Banfield and Raftery (1993) introduced the term "model based cluster analysis" to identify all those methods where groups in a population under study are associated with the components of a mixture distribution. Banfield and Raftery (1993) contains an extensive summary on the subject. Applications to density estimation and discrimination are discussed in Hastie and Tibshirani (1996) and Roeder and Wasserman (1997). However, the maximum likelihood estimator of the parameters of a finite normal mixture with fixed number of components is not robust against outliers. In fact, the estimator of the mean of each component is a weighted sum of observations and for each observation the weights sum up to 1 (see chapter 2 sec. 2.2 and sec. 2.4); that means that at least one of these parameters can become arbitrarily large if a single extreme point is added to the dataset.

A number of approaches has been proposed to overcome this problem. Fraley and Raftery (1998) proposed a model where a component accounting for "noise" is added to a mixture of Gaussian distributions. The noise is modelled as a uniform distribution on the convex hull (the range in one dimension) of the data. Another strategy is to model the data via finite mixtures of t-distributions instead of Normals, and the motivation is that t-distributions have heavy tails so that they are better in fitting data points distant from the center of the distribution. However, t-distributions are not able to fit extreme outliers (see Hennig, 2004). Maximum likelihood estimation of mixtures of t-distributions is extensively treated in McLachlan and Peel (2000a). A further approach is to use robust estimators such as Huber (1964, 1981) M-estimators. These correspond to maximum likelihood estimators for finite mixtures of Huber's least favorable distributions (see Huber, 1964). M-estimators are considered in Campbell (1984), McLachlan and Basford (1988) and Kharin (1996), amongst others. These methods have been shown to have better robustness properties (see Banfield and Raftery, 1993; McLachlan and Peel, 2000a). Kharin (1996) and Hennig (2004) have studied theoretical foundations of such statistical procedures. Kharin (1996) studied the case of finite Gaussian mixtures with fixed number of components. He showed that as the sample size goes to infinity and under certain assumptions on the speed of the convergence of the proportion of contamination to 0, Huber's M-estimator performs better than the maximum likelihood estimator. Hennig (2004) defined robustness measures for cluster analysis and studied robustness properties of the maximum likelihood estimator for mixtures of a general class of location-scale models including cases where the presence of outliers is taken into account by the addition of a "noise" component. Hennig (2004) also studied robustness of an estimator defined as the maximizer of an improper log-likelihood where a fixed constant (improper) density on the real line is added to a mixture of location-scale distributions. He showed that while the estimator with improper density is robust against extreme outliers, this is not the case for the maximum likelihood estimator of the model with uniform noise. However the model with uniform noise has good robustness properties when the outliers are not extremely large. In Chapter 4 we will show this by a Monte Carlo experiment.

The main interest of this chapter is to study identifiability and maximum likelihood estimation of a model with uniform noise. The case of the fixed improper density will be analyzed in the next chapter. Banfield and Raftery (1993) proposed a model where one uniform distribution is mixed with a finite number of normal distributions (from now on this will be labeled "Gaussian-uniform mixture model"). In this chapter we will study the general case when a finite number of uniform distributions is mixed with a finite number of distributions belonging to some family satisfying a given set of assumptions. The Gaussian-uniform mixture model will be a particular case of such a general class of models. For this class we will study identifiability, maximum likelihood estimation and computation of the maximum likelihood estimates.

§3.2. The model

In this section we introduce the notation and the main assumptions about the general model under study. Let $0 < s < \infty$ be the number of components in our mixture distribution, and let q be the number of uniform components $0 < q < s$ in the mixture. Let X be a real valued random variable distributed according to the following distribution function:

$$G(x;\eta) = \sum_{k=1}^{q} \pi_k U(x;\theta_k) + \sum_{l=q+1}^{s} \pi_l \Phi(x;\theta_l), \qquad (3.1)$$

where $\eta = (\pi, \theta)$, $\pi = (\pi_1, \pi_2, \ldots, \pi_s)$, $0 < \pi_j < 1$, $\sum_{j=1}^{s} \pi_j = 1$. Here $\theta = (\theta_1, \theta_2, \ldots, \theta_s)$, where $\theta_k = (a_k, b_k)$, a_k and b_k take values on the real line, and $-\infty < a_k < b_k < +\infty$ for each $k = 1, 2, \ldots, q$. Thus $\pi \in (0,1)^s$, $\theta_k \in \Theta_1 := \mathbb{R}^{2q}$ for $k = 1, 2, \ldots, q$. The parameter θ_l lies in some finite dimensional space Θ_2 for each $l = q+1, q+2, \ldots, s$. Furthermore the parameter space is denoted by $\Gamma := (0,1)^s \times \mathbb{R}^{2q} \times \Theta_2^{s-q}$. U is the uniform distribution function, i.e.

$$U(x;\theta_k) = \frac{x - a_k}{b_k - a_k} \mathbf{1}_{[a_k, b_k]}(x) + \mathbf{1}_{(b_k, +\infty)}(x),$$

$k = 1, 2, \ldots, q$, with $\mathbf{1}_A$ being the indicator function of the set A. The distribution function U has the density

$$u(x;\theta_k) = \frac{\mathbf{1}_{[a_k, b_k]}(x)}{b_k - a_k}.$$

The distribution function Φ belongs to a family of distributions satisfying

Assumption 3.1. $\Phi(x;\theta)$, $\theta \in \Theta_2$, is absolutely continuous with respect to the Lebesgue measure. It has density $\phi(x;\theta), \theta \in \Theta_2$, which is continuous both with respect to $x \in \mathbb{R}$ and $\theta \in \Theta_2$.

For notational convenience we will often rewrite the model in (3.1) as

$$G(x,\eta) := \sum_{j=1}^{s} \pi_j F_{v_j}(x;\theta_j) \tag{3.2}$$

where $v_j = \{1,2\}$ for $j = 1,2,\ldots,s$, when $v_j = 1$ then $F_{v_j} = U$, whenever $v_j = 2$ then $F_{v_j} = \Phi$. Moreover $g(x;\eta)$ will denote the density of $G(x;\eta)$.

§3.3. Identifiability

3.3.1 — Identifiability of "heterogeneous" mixtures

In section 1.5 (chapter 1) we introduced the identifiability problem for general finite mixtures. From now onward we will refer to definition 1.1 as "single family identifiability". This wording will become clear in the next few paragraphs. In this section we define and study the identifiability of a class of models which consists of a mixture of distributions coming from different families. We are in a situation where a finite number of distributions belonging to a general class of continuous distributions is mixed with a finite number of uniform distributions. We call such a mixture distribution "heterogeneous". Here the term heterogeneous mixtures means that the components in the mixture belong to different families of distributions. Such a statistical model can be very attractive in all those situations where the underlying heterogeneity in the data generating process is strong enough to let us consider that groups of observations come from populations with completely different features. In fact the uniform distribution here is introduced as a probabilistic model for noise, while Φ should represent the probabilistic structure of the clusters under study. Here we do not require that the number of components is known, nor do we require that the number of components belonging to each of the families of distributions is known. This situation is more general than that of the model proposed by Fraley and Raftery (1998), where the number of uniform components is considered as fixed and known. In fact in the uniform–normal mixture model proposed by Fraley and Raftery (1998) the number of uniform components is fixed to be one.

We now refer to section 1.5 (chapter 1) where we presented the definition of identifiability as given by Teicher (1961). Let us assume that $\mathscr{F}_k$, with

$k = 1, 2, \ldots, m$, are all families of probability distribution functions. For each k, $F_k(x; \theta) \in \mathscr{F}_k$, $\theta \in \Theta_k$, and Θ_k is some finite dimensional parameter space. A general element of the set of finite mixtures generated by the class $\mathscr{E} = \cup_{k=1}^{m} \mathscr{F}_k$ will be called heterogenous mixture distribution. Teicher's definition of identifiability does not require that the number of components in each family is fixed. However this is relevant in a situation where membership to different population components have different meaning. In our model for example we want to distinguish between noise components and non-noise components, and we want that the number of distributions belonging to each of the family composing the mixture is identified. To see why definition 1.1 does not take into account the identifiability of family memberships let us consider some results in the paper by Atienza et al. (2006). The authors studied the identifiability a model proposed by Marrazzi et al. (1998) in the context of fitting the length of stay in a hospital; the model is a mixture of three components: one Lognormal, one Gamma and one Weibull distribution. Atienza et al. (2006) gave a new sufficient condition for identifiability of finite mixtures following Teicher's definition, and based on this they showed the identifiability of the afore–mentioned class of mixtures. However, following the proof of their Theorem 3 it is clear that for some values of the parameters, a component having Gamma distribution cannot be distinguished from a component having Weibull distribution. Thus the number of components belonging to each family cannot be identified.

Here we will give a definition of identifiability which is similar to the one given by Teicher (1961) but adds some more restrictions so that family membership of components is taken into account in the sense explained above. It should now be clear why we named the identifiability defined by Teicher as "single family identifiability". Before we give our definition, let us introduce some more notation.

We will consider the set of all heterogenous finite mixtures generated by $\mathscr{E}$ with a discrete mixing distribution. Let $s < +\infty$ be the number of components of the heterogenous mixture, and let $c = (n_1, n_2, \ldots, n_m)$ be a vector of natural numbers where n_k, $k = 1, 2, \ldots, m$, indicates the number of distributions belonging to $\mathscr{F}_k$ being present in the mixture. From now on it is understood that c is finite, and of course it must be $s = \sum_{k=1}^{m} n_k$. We will call c the "composition" index. $\mathscr{H}$ is the family of all the finite mixtures generated from $\mathscr{E}$ with a discrete mixing distribution. A general element of $\mathscr{H}$ will be $H_c(x; \eta) = \sum_{j=1}^{s} \pi_j F_{k_j}(x; \theta_j)$, where $k_j \in \{1, 2, \ldots, m\}$ for $j = 1, 2, \ldots, s$, expresses the "family membership" of the jth component (e.g. $k_2 = 1$ means that the distribution of the second mixture component belongs to $\mathscr{F}_1$). The parameter η lies in the parameter set Ω, and $\eta = (\pi_1, \dots, \pi_s, \theta_1, \dots, \theta_s)$. We will consider the following definition:

Definition 3.1 (Global Identifiability). Let $\mathscr{H}$ be the class of finite mixtures generated by the class $\mathscr{E}$. Let $\mathscr{H}^\star \subseteq \mathscr{H}$, and $H_c \in \mathscr{H}^\star$. Given

$$H_c(x, \eta) = \sum_{j=1}^{s} \pi_j F_{v_j}(x; \theta_j), \pi_j > 0, \theta_j \neq \theta_r \quad \forall j, r = 1, 2, \dots, s, \quad j \neq r,$$

and

$$H_{c^*}(x, \eta^*) = \sum_{j=1}^{z} \pi_j^* F_{v_j}(x; \theta_j^*), \pi_j^* > 0, \theta_j^* \neq \theta_k^* \quad \forall j, k = 1, 2, \dots, z, \quad j \neq k;$$

if $H_c(\cdot, \eta) = H_{c^*}(\cdot, \eta^*)$ implies $s = z$, and that there exists a permutation $\bar{j}$ of the indexes $j = 1, 2, \dots, s$ such that $\pi_j = \pi^*_{\bar{j}}$, $\theta_j = \theta^*_{\bar{j}}$, $k_j = k_{\bar{j}}$, for $k_j, k_{\bar{j}} \in \{1, 2, \dots, m\}$, and $c = c^*$, then we say that $\mathscr{E}$ generates globally identifiable finite mixture distributions in $\mathscr{H}^\star$.

As highlighted before, we use the wording *global identifiability* to make a distinction between the notion of identifiability given in definition 3.1 with the one given in definition 1.1. The latter refers to Teicher's definition which we named "single family identifiability". With reference to definition 3.1, we require that the permutation of the component label (the index j) is constructed so that for each family $\mathscr{F}_k$ we identify the parameters, obtaining $\pi_j F_{k_j}(x; \theta_j) = \pi^*_{\bar{j}} F_{k_{\bar{j}}}(x; \theta^*_{\bar{j}})$, and at the same time we require that the number of distributions identified in the family $\mathscr{F}_k$ is consistent with the composition index c. To see the relevance of this argument let us refer to the model proposed by Banfield and Raftery (1993). In that case we require that not only the uniform parameters, the Gaussian parameters and all proportions are identified but we also require that it is possible to identify the number of noise components and Gaussian components, and which parameter belongs to which family. Notice that we have defined identification for a subclass $\mathscr{H}^\star$ of the class of finite mixtures generated by $\mathscr{E}$. This allows to study identifiability also when we restrict some of the families $\mathscr{F}_k$. For example in our situation we have to restrict the family of one-dimensional uniform distributions.

3.3.2 — The identifiability of the model with uniform noise

First, we will introduce some notation and assumptions and we also reconcile the exposition here with the notation used in the previous sections. We consider the model defined in Section 3.2 with the addition of the following definitions: (i) $\mathscr{F}_1$ is the family of all uniform distributions with support on an interval; (ii) $\mathscr{F}_2$ is a family of one dimensional distributions satisfying assumption 3.1. Finite mixtures generated by $\mathscr{F}_2$ are assumed to be identifiable in the sense of definition 1.1.

Let $n_1 = q$, $n_2 = s - q$ and let $c = (q, s - q)$ be the composition index. $\mathscr{H}$ is the family of finite mixtures generated by $\mathscr{E} = \mathscr{F}_1 \cup \mathscr{F}_2$, obtained by mixing q distributions from $\mathscr{F}_1$ and $s - q$ distributions from $\mathscr{F}_2$. The function $g_c(x;\eta)$ will denote the density of the distribution function $G_c(x;\eta)$ as defined in Section 3.2. $G_c(x;\eta)$ is an element of $\mathscr{H}$, with $\eta \in \Gamma$. $\mathscr{H}^\star \subset \mathscr{H}$ is the set of mixtures generated by $\mathscr{E}$ such that if $G_c(x;\eta)$ belongs to $\mathscr{H}^\star$, then $[a_t, b_t] \cap [a_r, b_r] = \emptyset$ for all $r, t = 1, 2, \ldots, q$ and $r \neq t$.

To show identifiability here we will make use of arguments based on derivatives so that it is necessary to introduce some more notation before we can state and prove the next result. The reason is that the uniform parameters coincide with the extreme of the uniform supports, and at these points the distribution function is not differentiable. By identifying the set of points where the distribution function is not differentiable we have identified the uniform components. The density $g_c(x;\eta)$ is discontinuous at a finite number of points, namely at $x \in W := \{a_1, b_1, a_2, b_2, \ldots, a_q, b_q\}$. Thus by properties of the Riemann integral, $dG_c(x;\eta)/dx = g_c(x;\eta)$ at all $x \in \mathbb{R}\backslash W$. However, right and left derivatives of G_c at all points in W exist and can be found by taking right and left limits of derivative quotients. The notation $D_y^-(\eta)$ and $D_y^+(\eta')$ stands for the left and right derivative of G_c respectively, and these derivatives are evaluated at a point y when the parameter vector is η', i.e.

$$D_y^-(\eta') = \lim_{t\uparrow 0} \frac{G_c(y+t;\eta') - G_c(y;\eta')}{t},$$

$$D_y^+(\eta') = \lim_{t\downarrow 0} \frac{G_c(y+t;\eta') - G_c(y;\eta')}{t}.$$

Computing these derivatives for the model (3.2) and for $h = 1, 2, \ldots, q$ will give us

$$D_{a_h}^-(\eta) \quad = \quad \sum_{l=q+1}^{s-q} \pi_l \phi(a_h;\theta_l);$$

$$D^{+}_{a_h}(\eta) = \frac{\pi_h}{b_h - a_h} + \sum_{l=q+1}^{s-q} \pi_l \phi(a_h; \theta_l);$$

$$D^{-}_{b_h}(\eta) = \frac{\pi_h}{b_h - a_h} + \sum_{l=q+1}^{s-q} \pi_l \phi(b_h; \theta_l);$$

$$D^{+}_{b_h}(\eta) = \sum_{l=q+1}^{s-q} \pi_l \phi(b_h; \theta_l).$$

Proposition 3.1. *The class $\mathscr{E} = \mathscr{F}_1 \cup \mathscr{F}_2$ generates globally identifiable heterogeneous mixtures in $\mathscr{H}^\star \subset \mathscr{H}$.*

Proof. Let us assume that $G_c(x;\eta) = G_{c^*}(x;\eta^*)$, i.e.

$$\sum_{j=1}^{s} \pi_j F_{v_j}(x;\theta_j) = \sum_{j=1}^{z} \pi_j^* F_{v_j}(x;\theta_j^*), \tag{3.3}$$

for every x, $v_j \in \{1,2\}$ and $j = 1,2,\ldots,s,\ldots z$, i.e. without loss of generality we assume that $s \leq z$. For a given function $f(y,z)$ differentiable at least on a subset of its own domain, we define the set

$$S_f(z) := \left\{ y : \frac{\partial^-}{\partial y} f(y,z) \neq \frac{\partial^+}{\partial y} f(y,z) \right\};$$

provided that all at points in $S_f(z)$ left and right partial derivatives of f exist. The assumption that $G_c(x;\eta) = G_{c^*}(x;\eta^*)$ implies that $S_{G_c}(\eta) = S_{G_{c^*}}(\eta^*)$. If $\#(A)$ stands for the cardinality of the set A, then $\#(S_{G_c}(\eta)) = \#(S_{G_{c^*}}(\eta^*)) = 2q$ which means that the number of the uniform components q is uniquely identified. Given a finite set $A := \{y_1, y_2, \ldots, y_n\}$, with $y_i \in \mathbb{R}$ all $i = 1,2,\ldots,n$, $\mu(A) \in \mathbb{R}^n$ denotes a vector where the components are all the elements of A. Furthermore $\bar{\mu}(A)$ is defined as $\bar{\mu}(A) = (y_{(1)}, y_{(2)}, \ldots, y_{(n)})$ where $y_{(i)}$ is such that $y_{(i)} \leq y_{(i+1)}$ all $i = 1,2,\ldots,n-1$. Now, $\bar{\mu}(S_{G_c}(\eta)) = (x_{(1)}, x_{(2)}, \ldots, x_{(2q)}) = \bar{\mu}(S_{G_{c^*}}(\eta^*)) = (x^*_{(1)}, x^*_{(2)}, \ldots, x^*_{(2q)})$. We recall that $\mathscr{H}^\star \subset \mathscr{H}$ is the set of mixtures generated by $\mathscr{E}$ such that if $G_c(x;\eta)$ belongs to $\mathscr{H}^\star$, then $[a_t, b_t] \cap [a_r, b_r] = \emptyset$ for all $r,t = 1,2,\ldots,q$ and $r \neq t$. Thus, we take a set of pairwise different indexes

$r_i \in \{1, 2, \ldots, s\}$ with $i = 1, 2, \ldots, q$ and we fix

$$\begin{aligned} \theta_{r_1} &= (a_{r_1}, b_{r_1}) = (x_{(1)}, x_{(2)}), \\ \theta_{r_2} &= (a_{r_2}, b_{r_2}) = (x_{(3)}, x_{(4)}), \\ &\vdots \\ \theta_{r_q} &= (a_{r_q}, b_{r_q}) = (x_{(2q-1)}, x_{(2q)}). \end{aligned}$$

Let us take another set of pairwise different indexes $t_i \in \{1, 2, \ldots, s\}$ with $i = 1, 2, \ldots, q$ and we fix

$$\begin{aligned} \theta^*_{t_1} &= (a^*_{t_1}, b^*_{t_1}) = (x^*_{(1)}, x^*_{(2)}), \\ \theta^*_{t_2} &= (a^*_{t_2}, b^*_{t_2}) = (x^*_{(3)}, x^*_{(4)}), \\ &\vdots \\ \theta^*_{t_q} &= (a^*_{t_q}, b^*_{t_q}) = (x^*_{(2q-1)}, x^*_{(2q)}). \end{aligned}$$

$\bar{\mu}(S_{G_c}(\eta)) = \bar{\mu}(S_{G_{c^*}}(\eta^*))$ implies that $\theta^*_{r_i} = \theta^*_{t_i}$ for all $i = 1, 2, \ldots, q$. Let us consider the equation

$$\left(D^+_{a_{r_i}}(\eta) - D^-_{a_{r_i}}(\eta)\right)(b_{r_i} - a_{r_i}) = \left(D^+_{a^*_{t_i}}(\eta^*) - D^-_{a^*_{t_i}}(\eta^*)\right)(b^*_{t_i} - a^*_{t_i}),$$

for all $i = 1, 2, \ldots, q$. By applying derivatives' formula in page 51, these equations give $\pi_{r_i} = \pi^*_{t_i}$ all $i = 1, 2, \ldots, q$. Hence, we have that there exists a permutation $\bar{j}$ of the indexes $j = 1, 2, \ldots, s, \ldots z$ such that if $j = r_i$ then $\bar{j} = t_i$, for which $\theta_j = (a_j, b_j) = \theta^*_{\bar{j}} = (a^*_{\bar{j}}, b^*_{\bar{j}})$, $\pi_j = \pi^*_{\bar{j}}$ and $v_j = v_{\bar{j}} = 1$. By this we have identified the number of uniform components, and all their parameters. Without loss of generality let us assume that $r_i = t_i$ for all $i = 1, 2, \ldots, q$ and that $r_1, r_2, \ldots, r_q = 1, 2, \ldots, q$.

For $j = q + 1, q + 2, \ldots, s, \ldots, z$ all the mixture components belong to $\mathscr{F}_2$ and q is identified as well. We consider the one-to-one transformation $\tilde{\pi}_j = \pi_j/(1 - \sum_{j=1}^{q} \pi_j)$ for $j = q + 1, q + 2, \ldots, s, \ldots, z$; and $\tilde{\pi}^*_j$ is defined analogously. Note that the denominator of $\tilde{\pi}_j$ is identified, in fact it depends on $\pi_1, \pi_2, \ldots, \pi_q$ which has been already identified. By (3.3) and the previous results we can write

$$\sum_{j=q+1}^{s} \tilde{\pi}_j F_{v_j}(x, \theta_j) = \sum_{j=q+1}^{z} \tilde{\pi}^*_j F_{v_{\bar{j}}}(x, \theta^*_j). \qquad (3.4)$$

By assumption the class of finite mixtures over $\mathscr{F}_2$ is identifiable with respect to definition 1.1, thus we have that: (i) $s = z$ and these indexes are identified; (ii) there exists some permutation $\bar{j}$ of indexes $j = q+1, q+2, \dots, s$ such that $\tilde{\pi}_j = \tilde{\pi}^*_{\bar{j}}$ and $\theta_j = \theta^*_{\bar{j}}$. But, $\tilde{\pi}_j = \tilde{\pi}^*_{\bar{j}}$ implies $\pi_j = \pi^*_{\bar{j}}$. Thus the $s - q$ components belonging to $\mathscr{F}$, their parameters and their mixing proportions are identified. The proof is completed by noting that having identified q and s it also results that $c = c^*$. ■

Given the proposition above we can easily get the next result.

Corollary 3.1. *Let $\mathscr{F}_2$ be the class of Gaussian distributions, then the class $\mathscr{F} = \mathscr{F}_1 \cup \mathscr{F}_2$ generates globally identifiable mixtures in $\mathscr{H}$.*

Proof. The result follows easily by noting that: (i) Gaussian distributions satisfy assumption 3.1; (ii) they are single family-identifiable by theorem 3 in Yakowitz and Spragins (1968). ■

We defined $\mathscr{F}_1$ so that it contains uniform distributions having not intersecting support and this can be explained with an example. Let us assume that $\mathscr{F}_1$ contains all uniform distribution with support on a real interval, and let us consider the following mixture distribution

$$\frac{1}{3}U(x; 0, 2) + \frac{1}{3}U(x; 2, 4) + \frac{1}{3}F(x; \theta),$$

where F is some distribution function satisfying assumption 3.1. We notice that $1/3U(x; 0, 2) + 1/3U(x; 2, 4) = 2/3U(x; 0, 4)$ so that identifiability does not hold. In fact, not only the parameters of the uniform distributions are not identifiable but also the composition index referring to the uniform components would not be identifiable.

In this section we studied the identifiability for some class of mixture distributions. We gave a new definition of identifiability which takes into account heterogeneity in the mixture, and we showed that a wide class of mixtures with uniform components are identifiable. In the literature the model consisting of a Gaussian-uniform mixture has been proposed by Fraley and Raftery (1998). This model has also been used to overcome problems of robustness. In this proposal the number of uniform components is fixed to be one, but we could be interested in determining whether the number of noise components is zero, one or even more than one. The estimation of the number of noise components requires that this number is identified. We extended this model to the case when more then one uniform components is added to a mixture of a class of continuous location-scale mixtures and we showed that the resulting mixture is identifiable. Mixtures with more than one uniform component are interesting because if we have more than one outlier which are well separated we could fit these outliers by more than one uniform components. For instance the empirical distribution of financial market risk measures (e.g. the so called market beta) often presents a small proportion of extreme points in both tails, and these points could be fitted using two uniform components.

§3.4. Maximum likelihood estimation

In this section we will study maximum likelihood estimation of the distribution in (3.1) when s and q are fixed and known. We consider the problem of existence of the maximum likelihood estimate, as well as its asymptotic properties. We will show that under some constraint on the parameter space the maximum likelihood estimate exists. Furthermore we will show that this estimate is strongly consistent, i.e. the sequence of maximum likelihood estimates converges almost surely to the true parameter as the sample size becomes arbitrarily large.

3.4.1 — Existence

In his classical work, Day (1969) studied finite mixtures of normal distributions. He highlighted several issues including the problem of the unboundness of the likelihood function. Let us assume that for a given sample $\underline{X_n} := \{X_1, X_2, \dots, X_n\}$ is an i.i.d. sequence of random variables distributed according to a finite mixture of m Gaussian distributions. The log-likelihood function associated with a realization $\underline{x_n} := \{x_1, x_2, \dots, x_n\}$ of $\underline{X_n}$ is given by

$$L_n(\xi) := \sum_{i=1}^{n} \log p(x_i; \xi)$$

where $p(x; \xi)$ is the density of a finite mixture of m Gaussian densities. Here $\xi = (\pi_1, \dots, \pi_m, \mu_1, \sigma_1, \dots, \mu_m, \sigma_m)$, with π_j being the proportion of the jth component, and μ_j and σ_j being the mean and standard deviation of the jth component respectively; $j = 1, 2, \dots, m$. If we fix $\mu_j = x_j$ and take σ_j arbitrarily close to 0 then $L_n(\xi) \longrightarrow +\infty$. This means that a global maximum fails to exist and numerical optimization algorithms would fail. As noted by Tanaka and Kawakami (2007), this problem also affects the wider class of location-scale mixtures. To overcome the unboundeness of the likelihood function two different constrained maximum likelihood estimators have been proposed.

DeSarbo and Cron (1988) studied the case of finite mixtures of normal distributions and they constrained the parameter space requiring that $\sigma_j \geq c > 0$, for all $j = 1, 2, \ldots, m$. These constraint lead to a well-defined optimization program for which a maximum of the likelihood function exists and it is shown to be strongly consistent (see Redner, 1981). However the choice of the constant c is critical. If c is chosen large enough such that for some j the true $\sigma_j < c$ the maximum likelihood estimator is obviously not consistent. This issue has been studied by Tanaka and Kawakami (2007) for general location-scale finite mixtures. Let σ_j now be the scale parameter of the jth location-scale component, Tanaka and Kawakami (2007) considered constraints of the type: $\sigma_j \geq c_n$, $c_n = c_0 \exp(-n^d)$, $c_0 > 0$, $0 < d < 1$, $j = 1, 2, \ldots, m$. As n (the sample size) goes to infinity the sequence of constraining constants $\{c_n\}_{n \geq 1}$ converges to 0. Under this type of constraints the authors showed that a sequence of maximum likelihood estimates is strongly consistent. A drawback of this kind of restriction is that maximum likelihood estimators are no longer scale equivariant because the scale of the data can be made smaller than the fixed constraining constant by multiplying all the observations by a real number.

In the case of normal mixtures Day (1969) noted that spurious maximizers of the likelihood function, corresponding to parameter points having some component standard deviations very small relative to others, are generated by small number of sample points grouped sufficiently close together. Dennis (1981) proposed to constrain the parameter space imposing that $\min_{i,j} \sigma_i / \sigma_j \geq c$ for a constant $c \in (0, 1]$, $i, j = 1, 2, \ldots, s$. Hathaway (1985) showed that the aforementioned set of constraints leads to a well posed optimization program and that the corresponding sequence of maximum likelihood estimates are strongly consistent. These types of constraints have the advantage that the constrained maximum likelihood estimator will be scale equivariant.

Here we study the existence of the maximum likelihood estimates of the model (3.1) under some additional assumptions about Φ. We do not treat estimation of the number of components. From now onward it is assumed that s and q are fixed and known. Let us go back to the notation fixed in Section 3.2 and let us denote $\theta_l = (\mu_l, \sigma_l)$, $\mu_l \in \mathbb{R}$, $\sigma_l \in \mathbb{R}_+ \backslash \{0\}$ for each $l = q+1, q+2, \ldots, s$, $\theta_l \in \Theta_2 := \mathbb{R}^{s-q} \times \mathbb{R}_+^{s-q} \backslash \{0\}$ for $l = q+1, q+2, \ldots, s$. The parameter space is now denoted by $\Gamma := [0, 1]^s \times \mathbb{R}^{s-q} \times \mathbb{R}_+^{s-q} \backslash \{0\}$.

Assumption 3.2. The density ϕ belongs to the location-scale family, i. e.

$$\phi(x;\theta) = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right).$$

Furthermore $\phi(x;\theta)$ is continuous with respect to θ at all x, and it is measurable for every θ; $\lim_{|z|\to\infty}\phi(z) = 0$, $\phi(0) < \infty$, $\lim_{\sigma\downarrow 0}\phi(x;\theta) = 0$ for all $x \neq \mu$, moreover $\phi(x;\theta)$ is monotonically increasing for each $x \in (-\infty, \mu]$ and it is monotonically decreasing for each $x \in [\mu, +\infty)$.

Remark 3.1. This assumption implies that if $x = \mu$ than $\lim_{\sigma\downarrow 0}\phi(x;\theta) = +\infty$. Let us take two points: $x = \mu$ and $y \neq \mu$. By the Assumption 3.2, as $\sigma \downarrow 0$ then $\phi(x;\theta) \longrightarrow +\infty$, $\phi(y;\theta) \longrightarrow 0$; but $\phi(y;\theta)\phi(x;\theta) \longrightarrow 0$. This means that as σ goes to zero $\phi(y;\theta)$ converges to zero at speed faster than that of $\phi(x;\theta)$ going to infinity.

Now we go back to model (3.1). Let $\underline{X_n} := \{X_1, X_2, \ldots, X_n\}$ be a sequence of i.i.d. random variables with distribution function $G(x;\eta)$. Let $\underline{x_n} := \{x_1, x_2, \ldots, x_n\}$ be a realization of $\underline{X_n}$ with associated log-likelihood function

$$L_n(\eta) := \sum_{i=1}^{n} \log g(x_i;\eta). \tag{3.5}$$

Notice that if for some i and k we take $x_i = a_k$, a choice of b_k arbitrarily close to a_k will make the likelihood unbounded. The same happens in the normal mixture case, if for some i and j we choose $x_i = \mu_j$ and we take σ_j close to zero. The two effects are related by noting that taking a_k close to b_k means that the variance of the kth uniform component becomes arbitrarily small. In fact, we could have parameterized the uniform components in terms of the mean and scale parameter, however in order to show the next results the parametrization adopted is more convenient.

We denote $v_j = \sigma_j$ for $j = q+1, \ldots, s$ and $v_j = (b_j - a_j)/\sqrt{12}$ for $j = 1, 2, \ldots, q$ (which is the standard deviation for the jth uniform component). We define the constrained parameter set as

$$\Gamma_c := \left\{\eta \in \Gamma : \min_{t,r} \frac{v_t}{v_r} \geq c > 0, \quad c \in (0,1]\right\}. \tag{3.6}$$

Remark 3.2. This constraint implies that if one of the scale parameters gets arbitrarily small all the other scale parameters have to converge to zero at the same rate. Let us take $v_{min} := \min\{v_j; j = 1, \ldots, s\}$ and $v_{max} := \max\{v_j; j = 1, \ldots, s\}$, the constraint above implies that $v_{min} \geq cv_{max}$. This implies that the parameters of the kth uniform components have to be such that $b_j - a_j \geq \sqrt{12}cv_{max}$. Therefore, if all the other variances are kept fixed, the constraint above puts a bound on the minimum length of the support of the uniform components.

We define the constrained maximum likelihood estimator as

$$\hat{\eta}_n := \arg\max_{\eta \in \Gamma_c} L_n(\eta) \tag{3.7}$$

The existence of $\hat{\eta}_n$ is not immediate. The constrained parameter set Γ_c is not compact, and moreover the log-likelihood function $L_n(\eta)$ it is not continuous on Γ and Γ_c. The function $L_n(\eta)$ has infinitely many points of discontinuity. To see this let us assume, without loss of generality, that there is just one uniform component (i.e. $q = 1$), and let us simplify the notation by imposing $a_1 = a$ and $b_1 = b$. The notation $x_{(i)}$ stands for the ith order statistic of the observed sample; i.e. $x_{(i-1)} \leq x_{(i)} \leq x_{(i+1)}$ for each $x_i \in \underline{x_n}$, $i = 2, \ldots, n-1$. For every vector η such that $x_{(i-1)} \leq a < x_{(i)}$ $L_n(\eta)$ is discontinuous at all points η such that $b = x_{(i)}, x_{(i+1)}, \ldots, x_{(n)}$. Similarly for all vector η such that $x_{(i)} < b \leq x_{(i+1)}$ $L_n(\eta)$ is discontinuous at all points η such that $a = x_{(1)}, x_{(2)}, \ldots, x_{(i)}$.

In order to show that $L_n(\eta)$ achieves its maximum on Γ_c we will give some intermediate Lemmas before we state and prove the main proposition. The next remark will be useful throughout the proofs of the following lemmas and proposition.

Remark 3.3. Let $\eta \in \Gamma_c$ be such that the uniform parameters of the jth uniform component fixed to be $a_j = x_p$ and $b_j = x_t$, for some $j = 1, \ldots, q$ and $p, t = 1, \ldots, n$, with $p \neq t$ and $x_p < x_t$. Let $N_\varepsilon^-(x_p) \equiv [x_p - \varepsilon, x_p)$ and $N_\xi^+(x_t) \equiv (x_t, x_t + \xi]$, where ε and ξ are positive real numbers fixed so that $N_\varepsilon(x_p)$ and $N_\xi(x_t)$ do not contain any data point. If $\eta' \in \Gamma_c$ coincides with η except that $a'_j \in N_\varepsilon(x_p)$ and $b'_j \in N_\xi(x_t)$, it follows that $L_n(\eta') < L_n(\eta)$. In fact, in order to maximize the log-likelihood function with respect to the parameters of the jth uniform component, we need to choose the parameters so that the length of the support of the jth uniform density is minimized for any given number of data points contained in it.

Lemma 3.1. *Let $\underline{x_n}$ contain at least $s+1$ distinct points. Then, under Assumption 3.2, $\sup_{\eta \in \Gamma_c} L_n(\eta) = \sup_{\eta \in \bar{\Gamma}_c} L_n(\eta)$, where $\bar{\Gamma}_c$ is a compact set contained in Γ_c.*

Proof. We want to show that in order to maximize the value of $L_n(\eta)$, the choice of $\eta \in \Gamma_c$ can be bounded. First let us fix some notations: we denote $m_n :=$

$\min\{x_i, \quad i = 1, \ldots, n\}$ and $M_n := \max\{x_i, \quad i = 1, \ldots, n\}$. The proof is now divided in four parts.

Part A. Let us take $\eta' \in \Gamma_c$ with $\mu'_j \leq m_n$ for some $j = q+1, \ldots, s$. We also consider the vector $\eta'' \in \Gamma_c$ which is equal to η' except that $\mu''_j = m_n$. This implies that $L_n(\eta') \leq L_n(\eta'')$. By analogy we take $\eta' \in \Gamma_c$ with $\mu'_j \geq M_n$ for some $j = q+1, \ldots, s$. Then we consider the vector $\eta'' \in \Gamma_c$ which is equal to η' except that $\mu''_j = M_n$. This implies that $L_n(\eta') \leq L_n(\eta'')$.

Part B. Now let us take $\eta' \in \Gamma_c$ such that $a'_k = m_n$ and $b'_k = M_n$ for some $k = 1, \ldots, q$. The parameter $\eta'' \in \Gamma_c$ is equal to η' except that the parameters of the kth uniform component are such that $a''_k \leq m_n$ and $b''_k \geq M_n$. By the arguments given in Remark 3.3 it follows that $L_n(\eta'') \leq L_n(\eta')$.

Part C. Unboundness of the summands of the likelihood function can happen when the scale parameter of a location-scale density gets arbitrarily small and the corresponding location parameter is fixed to be equal to one of the data points. Unboundness also happens when the support of a uniform density becomes so small that it collapses on a single data point. By Remark 3.2 we know that if one of the scale parameters gets arbitrarily close to zero, all the others also get arbitrarily close to zero at the same rate. Let us take a sequence $\{\eta_t\}_{t\geq 1}$ such that $v^t_j \downarrow 0$ for all $j = 1, \ldots, s$ while all the other parameters are fixed. For each $t \geq 1$, let us fix (without loss of generality) $a^t_j = x_j$ for all $j = 1, \ldots, q$ and $\mu^t_j = x_j$ for all $j = q+1, \ldots, s$. Furthermore $v^t_j \downarrow 0$ means that $\sigma^t_j \downarrow 0$ for all $j = q+1, \ldots s$ and $b^t_j \downarrow x_j$ for all $j = 1 \ldots q$. By assumption the vector $\underline{x_n}$ contains at least $s+1$ points. Without loss of generality let as assume that $\underline{x_n}$ contains just $s+1$ points, the case when $m > s+1$ goes along the same lines. The log-likelihood $L_{s+1}(\eta^t)$ can be rewritten as

$$L_{s+1}(\eta^t) = \log\left(g(x_{s+1}; \eta^t) \prod_{i=1}^{s} g(x_i; \eta^t) \right). \tag{3.8}$$

By Assumption 3.2 as $t \longrightarrow \infty$ then $g(x_i; \eta^t) \longrightarrow +\infty$ for each $i = 1, \ldots, s$, and $g(x_{s+1}; \eta^t) \longrightarrow 0$. However by Remark 3.1, we know that $g(x_{s+1}; \eta^t)$ converges to zero at a speed faster than of $g(x_i; \eta^t)$ diverging to $+\infty$, $i = 1, \ldots, s$. The latter means that as $t \longrightarrow \infty$ then $L_{s+1}(\eta^t) \longrightarrow -\infty$.

Part D. If one of the scale parameter gets arbitrarily large, v_j becomes arbitrarily large for all $j = 1, \ldots, s$. Let us take a sequence $\{\eta_t\}_{t\geq 1}$, where η_t is such that $\sigma^t_j \longrightarrow +\infty$ for all $j = q+1, \ldots, s$. The latter implies that $b^t_j - a^t_j \longrightarrow +\infty$ for all $j = 1, \ldots, q$. By Assumption 3.2, as $t \longrightarrow \infty$ then $L_n(\eta^t) \longrightarrow -\infty$.

By the results in A–D we can conclude that $\sup_{\eta\in\Gamma_c} L_n(\eta) = \sup_{\eta\in\bar{\Gamma}_c} L_n(\eta)$; where

$\bar{\Gamma}_c := [0,1]^s \times \bar{\Theta}_1 \times \bar{\Theta}_2$, with

$$\bar{\Theta}_1 := \{\theta_k \in \Theta_1 : m_n \leq a_k < b_k \leq M_n, k = 1, \ldots q\}, \tag{3.9}$$

and

$$\bar{\Theta}_2 := \{\theta_j \in \Theta_2 : m_n \leq \mu_j \leq M_n, \underline{\sigma} \leq \sigma_j \leq \overline{\sigma}, j = q+1, \ldots, s\}, \tag{3.10}$$

for some choice of the constants $\underline{\sigma}$, and $\overline{\sigma}$ such that $0 < \underline{\sigma} < \overline{\sigma} < \infty$. The sets $\bar{\Theta}_1$ and $\bar{\Theta}_2$ are now compact as well as the set $\bar{\Gamma}_c$. ■

Lemma 3.2. *Let $\underline{x_n}$ contain at least $s+1$ distinct points, and let $\eta^* \in \bar{\Gamma}_c$ be a local maximum for $L_n(\eta)$. Then η^* is such that for all $k = 1, 2, \ldots, q$, (a_k^*, b_k^*) either coincides with a pair of distinct points in $\underline{x_n}$, or (a_k^*, b_k^*) is such that $b_k^* - a_k^* = \sqrt{12} c v_{max}$, where $v_{max} \equiv \max\{v_j^*, j = 1, \ldots, s\}$ and the interval $[a_k^*, b_k^*]$ contains at least one data point.*

Proof. Under the assumptions of the Lemma above η^* is a local or a global maximum for the log-likelihood function over Γ_c. Going back to the proof of Lemma 3.1 (Part B) we recall that $m_n \leq a_k^* \leq b_k^* \leq M_n$ for all $k = 1, \ldots, q$, where m_n and M_n are the minimum and the maximum data point respectively. Let us denote $v_{max} = \max\{v_j^*; j = 1, \ldots, s\}$. Under the assumption that $\underline{x_n}$ contains at least $s+1$ distinct points, because of Part C in the proof of the Lemma 3.1 and Remark 3.2, it follows that $b_k^* - a_k^* \geq \sqrt{12} c v_{max} > 0$ for each $k = 1, \ldots, q$. From now onward $\eta^*(a_k, b_k) \in \Gamma_c$ denotes the parameter vector with all components equal to those of η^* except the parameters of the kth uniform component which are set to be a_k, b_k. Also if y is a data point, then $N_\varepsilon^-(y) \equiv [y - \varepsilon, y)$ and $N_\varepsilon^+(y) \equiv (y, y+\varepsilon]$, where $\varepsilon > 0$ is such that $N_\varepsilon^-(y)$ and $N_\varepsilon^+(y)$ do not contain any data point. Let $\{\tilde{x}_{(1)}, \tilde{x}_{(2)}, \ldots, \tilde{x}_{(n)}\}$ be the set of all distinct points of $\underline{x_n}$ such that $\tilde{x}_{(i)} < \tilde{x}_{(i+1)}$ all $i = 1, 2, \ldots, n-1$. Let us consider two pairs of distinct data points, $\tilde{x}_{(d)}$ and $\tilde{x}_{(e)}$, with $d, e = 1, \ldots, n$, $d < e$, and $\tilde{x}_{(e)} - \tilde{x}_{(d)} \geq \sqrt{12} c v_{max}$. There are three cases: (i) the interval $(\tilde{x}_{(d)}, \tilde{x}_{(e)})$ contains at least a pair of distinct data points; (ii) the interval $(\tilde{x}_{(d)}, \tilde{x}_{(e)})$ does not contains any data point; (iii) the interval $(\tilde{x}_{(d)} \tilde{x}_{(e)})$ contains just one data point.

Case (i). We assume that the interval $(\tilde{x}_{(d)}, \tilde{x}_{(e)})$ contains more than one distinct data points. Let us consider the points $\tilde{x}_{(d+1)}$ and $\tilde{x}_{(e-1)}$, with $d+1 < e-1$. Now there are two further cases: (i.a) $\tilde{x}_{(e-1)} - \tilde{x}_{(d+1)} \geq \sqrt{12} c v_{max}$; or (i.b) $\tilde{x}_{(e-1)} - \tilde{x}_{(d+1)} < \sqrt{12} c v_{max}$.

Case (i.a). First we assume that $\tilde{x}_{(e-1)} - \tilde{x}_{(d+1)} \geq \sqrt{12} c v_{max}$. By Remark 3.3

we conclude that for any possible $\varepsilon, \xi > 0$ and any $a_k \in N_\varepsilon^-(\tilde{x}_{(d)})$ and $b_k \in N_\xi^+(\tilde{x}_{(e)})$, $L_n(\eta^*(a_k, b_k)) < L_n(\eta^*(\tilde{x}_{(d)}, \tilde{x}_{(e)}))$. Applying the same argument as above: for any possible $\varepsilon, \xi > 0$ and any $a_k \in N_\varepsilon^-(\tilde{x}_{(d+1)})$ and $b_k \in N_\xi^+(\tilde{x}_{(e-1)})$, $L_n(\eta^*(a_k, b_k)) < L_n(\eta^*(\tilde{x}_{(d+1)}, \tilde{x}_{(e-1)}))$. This means that either $\eta^*(\tilde{x}_{(d+1)}, \tilde{x}_{(e-1)})$ and $\eta^*(\tilde{x}_{(d)}, \tilde{x}_{(e)})$ are candidates for a local maximum. Case (i.b). We now assume the case when $\tilde{x}_{(e-1)} - \tilde{x}_{(d+1)} < \sqrt{12}cv_{max}$. As before, for any possible $\varepsilon, \xi > 0$ and any $a_k \in N_\varepsilon^-(\tilde{x}_{(d)})$ and $b_k \in N_\xi^+(\tilde{x}_{(e)})$, it follows that $L_n(\eta^*(a_k, b_k)) < L_n(\eta^*(\tilde{x}_{(d)}, \tilde{x}_{(e)}))$. Now, the pair $(a_k, b_k) = ((\tilde{x}_{(d+1)}, \tilde{x}_{(e-1)}))$ is not contained in the constrained parameter space since the constraint does not hold. Let us take any (a_k', b_k') such that $b_k' - a_k' = \sqrt{12}cv_{max}$ and $a_k' \leq \tilde{x}_{(d+1)} < \tilde{x}_{(e-1)} \leq b_k'$. Notice that the corresponding parameter η' now lies on the boundary of Γ_c. By the same argument as before, for any possible $\varepsilon, \xi > 0$ and any $a_k \in N_\varepsilon^-(a_k')$ and $b_k \in N_\xi^+(b_k')$, it follows that $L_n(\eta^*(a_k, b_k)) < L_n(\eta^*(a_k', b_k'))$. Which means that either $\eta^*(\tilde{x}_{(d)}, \tilde{x}_{(e)})$ and $\eta^*(a_k', b_k')$ are candidates for a local maximum.
Case (ii). We assume that the interval $(\tilde{x}_{(d)}, \tilde{x}_{(e)})$ does not contain any data point. We can apply the same argument as before and show that $\eta^*(\tilde{x}_{(d)}, \tilde{x}_{(e)})$ is a local maximum. Case (iii). We assume that the interval $(\tilde{x}_{(d)}, \tilde{x}_{(e)})$ contains just a single distinct data point. By applying the same argument as in part (i.b), we conclude that either $\eta^*(\tilde{x}_{(d)}, \tilde{x}_{(e)})$ and $\eta^*(a_k', b_k')$ are candidates for a local maximum; where a_k' and b_k' are such that $b_k' - a_k' = \sqrt{12}cv_{max}$ and $\tilde{x}_{(d)} \leq a_k'$ and $b_k' \leq \tilde{x}_{(e)}$.
Parts (i.a)–(iii) complete the proof. ■

Notice that the Lemma above only concentrates on a single uniform component. When we deal with all the q uniform components we have to take into account that we are assuming that all uniform supports are disjoint interval. Hence it must be that $a_1^* < b_1^* <, \ldots, < a_q^* < b_q^*$.

Proposition 3.2. *Let Lemmas 3.1 and 3.2 hold. Under Assumption 3.1, $L_n(\eta)$ achieves its maximum over Γ_c.*

Proof. Let η^* be a local maximum, and $v^* := \max\{v_1^*, v_2^*, \ldots, v_s^*\}$. We recall that Lemma 3.1 implies that v_{max} is contained in a closed and bounded real interval. For all possible values of v_{max} (as defined in the previous lemma) we know all possibile values of the uniform parameters for which the corresponding η is a candidate for a local maximum. This can be done applying Lemma 3.2. Let $\bar{L}_n(\bar{\eta})$ be the log-likelihood function when all uniform parameters are fixed in order to get a local or global maximum. The new vector of parameters is now $\bar{\eta} = (\pi_1, \ldots, \pi_s, \theta_{q+1}, \ldots, \theta_s)$, while all the uniform parameters are set as defined in the previous Lemma. The parameter $\bar{\eta}$ lies in $\in [0,1]^s \times \bar{\Theta}_2$, where Θ_2 corresponds to the set defined in (4.8). By Assumption 3.1, $\bar{L}_n(\bar{\eta})$ is continuous on the compact set $[0,1]^s \times \bar{\Theta}_2$, and hence has a maximum. Applying this argument for all possible values of v_{max} we can find all possible local maxima of $L_n(\eta)$ on Γ_c, and hence among these we get the global maximum. ■

3.4.2 — Asymptotic analysis

In this section we will study the asymptotic properties of the estimator defined in (3.7) and under some additional assumptions we shall show that it is strongly consistent for the true parameter. The technique usually used to show consistency and asymptotic normality for the sequence of the maximum likelihood estimates consists of assuming differentiability of the likelihood function, plus other regularity conditions about continuity and integrability of derivatives of the likelihood function up to the third order. The afore-mentioned set of assumptions is also known as Cramér–Rao regularity conditions. The standard asymptotic analysis for maximum likelihood estimators can not be used here. We have several problems: (i) the model (3.1) implies a likelihood function with infinitely many discontinuity points; (ii) in order to achieve a global maximum for the log-likelihood we need to restrict the parameter space to a set which is not compact; (iii) the distribution we want to estimate is identifiable only up to label switching.

Wald (1949) studied a general class of estimators of which the maximum likelihood is a particular case, and he showed strong consistency under general conditions not involving derivatives of the likelihood function. However, in Wald's approach it is assumed that the parameter space is compact and that the model is fully identifiable. This is not the case in our situation. In fact, the parameter space is not compact because we allow each $\sigma_l > 0$ for all $l = q+1, \ldots, s$ and $a_k < b_k$ $k = 1, \ldots, q$. Furthermore, full identifiability is not achieved. In fact by Proposition 3.1 we are able to distinguish two distributions in $\mathscr{F}$ only up to components label switching. Redner (1981) extended the results in Wald (1949). First he defined consistency for sequences of estimates of parameters of non-identifiable distributions, and then he showed the consistency of sequences of maximum likelihood estimates for such distributions. However, Redner's theory deals with compact parameter spaces. Kiefer and Wolfowitz (1956) studied the class of estimators introduced by Wald (1949) in the case when the parameter space is not compact. On the other hand the authors assume full identifiability of the model as in Wald (1949).

Hathaway (1985) studied the strong consistency of the maximum likelihood sequence for finite mixtures of Gaussian distributions on a constrained set of the same kind as (3.6). The author used the theory of Kiefer and Wolfowitz (1956) with an approach similar to that employed by Redner (1981). Here we will adopt a similar approach. We will give some additional notation, and after that we shall state and prove some intermediate lemmas.

From now onward $\eta_0 \in \Gamma_c$ will denote the true parameter, i.e. $G(x.\eta_0)$ is the distribution which generated the sample $\underline{X_n}$. As in Kiefer and Wolfowitz (1956) we define a metric δ on Γ:

$$\delta(\eta, \eta_*) := \sum_{j=1}^{3s} \left|\arctan \eta^j - \arctan \eta_*^j\right| ,$$

for all $\eta, \eta_* \in \Gamma$ with η^j being the jth component of the vector η. We complete the set Γ_c with all limits of its Cauchy sequences. That is, $\bar{\Gamma}_c$ is the set Γ_c along with the limits of its Cauchy sequences in the sense of δ. As in Hathaway (1985) we will show that sufficient conditions given by Kiefer and Wolfowitz (1956) hold. In some cases this set of sufficient conditions is not easy to show for the density of one observation. In Section 6 of Kiefer and Wolfowitz (1956) it is argued that in some cases it is easier to work with the joint density of a vector of observations. This strategy is discussed in more detail by Perlman (1972).

Let $Y = (X_1, X_2, \dots, X_m)$ be a vector of m random variables independently distributed according to $G(x; \eta)$. Let $g_m(y; \eta)$ the joint density of the component of Y.

Lemma 3.3. *We assume that $\{\eta^t\}_{t\geq 1}$ is a sequence in $\bar{\Gamma}_c$ and $\eta^* \in \bar{\Gamma}_c$. For every sequence $\eta^t \longrightarrow \eta^*$, $g_m(y; \eta^t) \longrightarrow g_m(y; \eta^*)$; except perhaps on a set $E \subset \mathbb{R}^m$ which depends on η^* and whose Lebesgue measure is zero.*

Proof. Since ϕ is continuous (Assumption 3.1) we only have to take care of the discontinuities introduced by the uniform components. Let us take a sequence $\{\eta^t\}_{t\geq 1}$ converging to η^* in $\bar{\Gamma}_c$. If the point $y \in \mathbb{R}^m$, $y = (x_1, x_2, \dots, x_m)$ is such that $x_i \neq a_k^*$ and $x_i \neq b_k^*$ for all $i = 1, 2, \dots, m$ and $k = 1, 2, \dots, q$ than it easy to see that the thesis of the statement holds because $\mathbf{1}_{[a_k^t, b_k^t]}(x_i) \longrightarrow \mathbf{1}_{[a_k^*, b_k^*]}(x_i)$ for all k, i. This is not the case for all points $y' \in E$ where for some k and i there is some $a_k^* = x_i'$ and/or $b_k^* = x_i'$. Thus the statement above holds, in fact the set E depends on the limit point η^* and has zero Lebesgue measure. ■

The joint density of m observations $g_m(y;\eta)$ is itself a mixture of s^m components each one having the form

$$\bar{g}(y;\gamma,\eta) = \prod_{r=1}^{m} f_{p_r}(x_r;\theta_{j_r}),$$

where the parameter γ denotes the vector of indexes $(j_1, j_2, \ldots, j_m)$ with $j_r \in \{1, 2, \ldots, s\}$ for all $r = 1, 2, \ldots, m$. Moreover, $p_r = 1$ if $j_r \in \{1, \ldots, q\}$, and $p_r = 2$ if $j_r \in \{q+1, \ldots, s\}$. As indicated before $f_1 = u$ and $f_2 = \phi$. Also, $\theta_{j_r} = (a_{j_r}, b_{j_r})$ for $j_r \in \{1, 2, \ldots, q\}$ and $\theta_{j_r} = (\mu_{j_r}, \sigma_{j_r})$ for $j_r \in \{q+1, \ldots, s\}$. For any $\gamma = (j_1, j_2, \ldots, j_m)$ let us denote

$$\bar{\pi}(\gamma) = \prod_{r=1}^{m} \pi_{j_r}.$$

The joint density of the vector y can be written as

$$g_m(y;\eta) = \sum_{h=1}^{s^m} \bar{\pi}(\gamma_h)\bar{g}(y;\gamma_h,\eta)$$

where $\gamma_1, \ldots, \gamma_{s^m}$ are all possible vectors γ obtained by combining the sets of indexes $\{j_1, j_2, \ldots, j_m\}$ with $j_r \in \{1, 2, \ldots, s\}$ for all $r = 1, 2, \ldots, m$.

Now we prove two intermediate Lemmas which will be useful to show that Kiefer-Wolfovitz sufficient conditions for the consistency of the maximum likelihood estimator are satisfied for the joint density of m observations, with $m > s$. From now onward E will denote the expectation operator under the distribution G; $\mathrm{E}_{\eta'} f$ stands for the expectation of the function f under the distribution G with the parameter η'. Before to state and prove the next Lemma we consider the following

Assumption 3.3. For some $j = q+1, \ldots, s$, $\mathrm{E}_{\eta_0} \log \phi(x;\mu_j^0, \sigma_j^0) > -\infty$, where μ_j^0, σ_j^0 are components of η_0

Lemma 3.4. *Let Assumption 3.3 hold. Then for any $m > s$ $E_{\eta_0} \log g_m(y;\eta_0) > -\infty$.*

Proof. Let us choose h^* such that: $\gamma_{h^*} = \{j^*, j^*, \ldots, j^*\}$, $j^* \in \{q+1, \ldots, s\}$, and j^* is such that $\mathrm{E}_{\eta_0} \log \phi(x;\mu_{j^*}^0, \sigma_{j^*}^0) > -\infty$. The existence of such a γ_{h^*} is ensured by Assumption 3.3. The following chain of inequalities completes the proof:

$$\mathrm{E}_{\eta_0} \log g_m(y;\eta_0) = \mathrm{E}_{\eta_0} \log \sum_{h=1}^{s^m} \bar{\pi}(\gamma_h) g(y;\gamma_h,\eta_0) \geq \mathrm{E}_{\eta_0} \log \pi(\gamma_{h^*}) \bar{g}(y;\gamma_{h^*},\eta_0) \geq$$

$$\log \pi(\gamma_{h^*}) + \mathrm{E}_{\eta_0} \log \bar{g}(y;\gamma_{h^*},\eta_0) \geq \log \pi(\gamma_{h^*}) + \mathrm{E}_{\eta_0} \sum_{r=1}^{m} \log \phi(x_r;\mu^0_{j^*},\sigma^0_{j^*}) \geq$$

$$\log \pi(\gamma_{h^*}) + \sum_{r=1}^{m} \mathrm{E}_{\eta_0} \log \phi(x_r;\mu^0_{j^*},\sigma^0_{j^*}) > -\infty$$

■

In order to prove the next Lemma a further Assumption is needed.

Assumption 3.4. Let X and Y be two random variables independently distributed according to G, and let E_{η_0} denote the expectation under G, then

$$\mathrm{E}_{\eta_0} \sup_{(\mu,\sigma)\in\mathbb{R}\times\mathbb{R}_+} \log \frac{1}{\sigma^t} \phi(x;\mu,\sigma)\phi(y;\mu,\sigma) < +\infty, \tag{3.11}$$

for any finite $t \geq 1$.

The Assumption above is fulfilled when ϕ is Gaussian, this will be stated in a Corollary 3.2.

Lemma 3.5. *Let Assumptions 3.2 and 3.4, hold, then for any $m > s$,*

$$E_{\eta_0} \sup_{\eta\in\bar{\Gamma}_c} \log g_m(y;\eta) < +\infty.$$

Proof. The thesis of the Lemma is true if

$$\mathrm{E}_{\eta_0} \sup_{\eta\in\bar{\Gamma}_c} \log \bar{g}(y;\gamma,\eta) < +\infty \tag{3.12}$$

holds for all possible indexes γ. In order to show (3.12) we introduce a convenient parametrization of the uniform components in G in terms of their means and standard deviations[1]. For all $k = 1, 2, \ldots, q$ we fix $\mu_k = (a_k + b_k)/2$, $\sigma_k = (b_k - a_k)/\sqrt{12}$. Now $u(x;\theta_k) = u(x;\mu_k,\sigma_k)$ with

$$u(x;\mu_k,\sigma_k) = \frac{1}{\sqrt{12}\sigma_k} \mathbf{1}_{[\mu_k - \sqrt{3}\sigma_k;\mu_k+\sqrt{3}\sigma_k]}.$$

[1] Here the parametrization of the uniform components in terms of their means and standard deviations eases the notation.

We now assume that $m = s+1$. For each index γ all the factors of $\bar{g}(y;\gamma,\eta)$ are bounded over $\bar{\Gamma}_c$ unless we take σ_{p_r} close to 0 and $\mu_{p_r} = x_r$ for some $r = 1,2,\dots,s+1$ and $p_r \in \{1,2\}$. Let $\bar{g}_m(y;\gamma,\eta)$ be such that $s-1$ of its components are set such that their location parameters are equal $s-1$ of the components in y. Hence for some indexes $h,t \in \{1,2,\dots,s+1\}$ and $z \in \{1,2,\dots,s+1\}$ we can write

$$\sup_{\eta\in\bar{\Gamma}_c} \log \bar{g}(y;\gamma,\eta) \leq \sup_{\eta\in\bar{\Gamma}_c} \log Q\frac{1}{\sigma_z^{s-1}} f_{p_h}(x_h;\mu_z,\sigma_z) f_{p_t}(x_t;\mu_z,\sigma_z), \tag{3.13}$$

where Q is some finite constant.

We now consider the above inequality in three possible cases: (i) $p_h = p_t = 2$; (ii) $p_h = p_t = 1$; (iii) $p_h = 1$ and $p_t = 2$.

Case (i). If $p_h = p_t = 2$, then $f_{p_h} = f_{p_h} = \phi$, applying the operator E_{η_0} on both the left and right-hand side of (3.13), by Assumption 3.4 the condition (3.12) holds proving the statement.

Case (ii). If $p_h = p_t = 1$, then $f_{p_h} = f_{p_h} = u$. Let us introduce the function

$$\Delta_1(x_t,x_h;\mu_z,\sigma_z) = \log \frac{Q_1}{\sigma_z^{s+1}} \mathbf{1}_{[\mu_z-\sqrt{3}\sigma_z;\mu_z+\sqrt{3}\sigma_z]}(x_h) \mathbf{1}_{[\mu_z-\sqrt{3}\sigma_z;\mu_z+\sqrt{3}\sigma_z]}(x_t),$$

with Q_1 a finite constant. We note that $\Delta_1(x_t,x_h;\mu_z,\sigma_z) < T < +\infty$ for some T and any choice of μ_z and σ_z at any x_h and x_t. Whence

$$\mathrm{E}_{\eta_0} \sup_{\eta\in\bar{\Gamma}_c} \Delta_1(x_t,x_h;\mu_z,\sigma_z) < T < +\infty.$$

This means that the condition (3.12) holds proving the statement.

Case (iii). Let us now assume that $p_h = 1$ and $p_t = 2$ then $f_{p_h} = u$ and $f_{p_t} = \phi$. We introduce the function

$$\Delta_2(x_t,x_h;\mu_z,\sigma_z) = \log \frac{Q_2}{\sigma_z^{s+1}} \mathbf{1}_{[\mu_z-\sqrt{3}\sigma_z;\mu_z+\sqrt{3}\sigma_z;]}(x_h)\phi(x_t;\mu_z,\sigma_z),$$

where Q_2 is some constant. By Assumption 3.2

$$\Delta_2(x_t,x_h;\mu_z,\sigma_z) = \begin{cases} \log \frac{Q_2}{\sigma_z^{s+1}}\phi(\frac{x_t-\mu_z}{\sigma_z}), & \text{if } x_t,x_h \in [\mu_z-\sqrt{3}\sigma_z;\mu_z+\sqrt{3}\sigma_z;]; \\ -\infty, & \text{otherwise.} \end{cases}$$

We observe that for some T', $\Delta_2(x_t,x_h;\mu_z,\sigma_z) \leq T' < +\infty$ for any choice of (μ_z,σ_z) at any x_h and x_t except when $x_t = x_h$. In fact when $x_t = x_h$ we can take $\mu_z = x_t = x_h$ and $\sigma_z \downarrow 0$ making $\Delta_2(\mu_z,\sigma_z)$ approaching to $+\infty$. Notice that the set of points where $x_h = x_t$ has zero Lebesgue measure in $\mathbb{R}^2$. Whence

$$\mathrm{E}_{\eta_0} \sup_{\eta \in \tilde{\Gamma}_c} \Delta_2(x_t, x_h; \mu_z, \sigma_z) =$$

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \sup_{\mu_z, \sigma_z} \Delta_2(x_t, x_h; \mu_z, \sigma_z) g(x_t; \eta_0) g(x_h; \eta_0) dx_h dx_t \leq T' < +\infty,$$

which implies that (3.12) holds, and this proves the statement. The proof is completed by noting that any $m > s + 1$ would not change the the cases (i)–(iii). ■

Wald (1949) and Kiefer and Wolfowitz (1956) assumed that the model is fully identifiable. In our context the definition of identifiability allows for components label switching. The approach used by Redner (1981) and Hathaway (1985) to overcome this difficulty is to work with a properly defined quotient topological space of the parameter set. We define the set

$$C(\eta') := \left\{ \eta \in \Gamma_c : \int_{-\infty}^{x} g(t; \eta) dt = \int_{-\infty}^{x} g(t; \eta') dt \quad \forall x \in \mathbb{R} \right\}.$$

Let $\tilde{\Gamma}_c$ be the quotient topological space obtained from Γ_c by identifying $C(\eta')$ to a point $\tilde{\eta}' = \eta'$. As in Redner (1981) by using the theory of Kiefer and Wolfowitz (1956) it is possible to show strong consistency of the sequence of maximum likelihood estimates on the quotient space $\tilde{\Gamma}_c$. Before we state and prove the next result we need to set up some more notations:

$$\mathscr{N}_\varepsilon(\eta') := \left\{ \eta \in \Gamma_c : \forall \eta^\star \in C(\eta') \quad \delta(\eta, \eta^\star) \leq \varepsilon \right\},$$

where $\varepsilon > 0$.

Proposition 3.3. *Under the assumptions of Lemmas 3.3, 3.4 and 3.5, for any $\varepsilon > 0$ there exists $h(\varepsilon) \in (0, 1)$ such that*

$$\Pr \left\{ \lim_{n \to \infty} \sup_{\eta \in \Gamma_c \setminus \mathscr{N}_\varepsilon(\eta_0)} \frac{\prod_{i=1}^{n} g(x_i; \eta)}{\prod_{i=1}^{n} g(x_i; \eta_0)} < h(\varepsilon)^n \right\} = 1 \qquad (3.14)$$

Proof. Assumptions 3.1 and 3.2 fulfill Assumptions 1 and 3 in Kiefer and Wolfowitz (1956) for the joint density if $m > s$. On the other hand, Lemmas 3.3, 3.4 and 3.5 fulfill Assumptions 2 and 5 in Kiefer and Wolfowitz (1956) for the joint density if $m > s$. This implies that the result (2.12) in Kiefer and Wolfowitz (1956) holds (see also comments in Section 6 in Kiefer and Wolfowitz (1956) and the paper by Perlman (1972)) and hence the equation 3.14 above holds proving the statement. ■

The result above implies convergence of the maximum likelihood estimator on the quotient space. The sequence of estimators defined in (3.7) is strongly consistent for $\tilde{\eta}_0$, i.e. $\hat{\tilde{\eta}}_n \xrightarrow{\text{as}} \tilde{\eta}_0$. By Proposition 3.1 this means that whenever n is infinitely large the sequence of estimates $\hat{\tilde{\eta}}_n$ converges almost surely to a point $\tilde{\eta}_0$ which coincides with η_0 up to permutation of the pairs $(\pi_{j,0}, \theta_{j,0})$ via permutation of the indexes $j = 1, 2, \dots, s$. The case where ϕ is a Gaussian density is easily obtainable.

Corollary 3.2. *Let $\phi(x; \mu, \sigma)$ be the Gaussian density with mean parameter μ and standard deviation σ, then Proposition 3.3 holds.*

Proof. When $\phi(x; \mu, \sigma)$ is the Gaussian density with mean parameter μ and standard deviation σ Assumptions 3.1–3.3 are fulfilled. The proof now rest on the verification of Assumption 3.4. Since $\phi(x; \mu, \sigma)$ is the Gaussian density then

$$\log \frac{1}{\sigma^t}\phi(x; \mu, \sigma)\phi(y; \mu, \sigma) = B(\mu, \sigma) = \log \frac{1}{2\pi\sigma^{t+2}} \exp\{-\frac{1}{2\sigma^2}[(x-\mu)^2 + (y-\mu)^2]\}$$

for some $t \geq 1$. The maximum of $B(\mu, \sigma)$ exists on $\mathbb{R} \times \mathbb{R}_+$, this can be verified along the same line of the arguments given in proof of the lemma 3.1 (parts A, C). Since $B(\mu, \sigma)$ is continuous and differentiable on $\mathbb{R} \times \mathbb{R}_+$, we derive the first order conditions for a maxima and we obtain that these are satisfied at

$$\mu^* = \frac{x+y}{2}, \quad \text{and} \quad \sigma^* = \frac{|x-y|}{\sqrt{2(t+2)}}.$$

By simple calculations we get

$$B(\mu^*, \sigma^*) = \log \frac{2(t+2)^{\frac{t+2}{2}} \exp\{-(t+2)/2\}}{|x-y|^{t+2}} = \log \frac{T}{|x-y|^{t+2}}$$

for $0 < T < +\infty$, where the constant T depends on t. Therefore

$$\mathrm{E}_{\eta_0} \sup_{(\mu,\sigma) \in \mathbb{R} \times \mathbb{R}_+} \log \frac{1}{\sigma^t}\phi(x; \mu, \sigma)\phi(y; \mu, \sigma) = \mathrm{E}_{\eta_0} B(\mu^*, \sigma^*) \tag{3.15}$$

But

$$\mathrm{E}_{\eta_0} B(\mu^*, \sigma^*) = \log T + (t+2) \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \log|x-y| g(x; \eta_0) g(y; \eta_0) dx dy < +\infty.$$

The last inequality proves the statement. ■

While we have shown that the sequence of maximum likelihood estimates converges with probability one to the true parameter on the quotient space we do not provide any asymptotic normality result. There are two main reasons for this: (i) while probability statements such as strong or weak consistency on the quotient space are easily interpretable, this is not the case for the convergence in distribution; (ii) asymptotic normality is usually obtained approximating the log-likelihood function using Taylor expansion and then applying some central limit theorem. In the case under study the log-likelihood function is not even continuous and derivatives–based approximation is impossible.

§3.5. Computations via the EM algorithm

In Chapter 2 Section 2.4 we introduced the general theory about the EM algorithm. Here we will discuss the algorithm for the likelihood function based on the mixture model with uniform noise. We rewrite the density of the distribution (3.2) as

$$g(x, \eta) := \sum_{j=1}^{s} \pi_j f_{v_j}(x; \theta_j), \tag{3.16}$$

where all the notations of Section 3.2 are maintained, f_{v_j} is the density of F_{v_j}. Therefore for $j = 1, \dots, q$, $v_j = 1$ and $f_{v_j} = u$; for all $j = q+1, \dots, s$, $v_j = 2$ and $f_{v_j} = \phi$. The EM algorithm we study here is intended to seek a maximum for the log-likelihood function $l_n(\eta) = \sum_{i=1}^{n} \log g(x_i; \eta)$ over the constrained set Γ_c. Let the index $t = 1, 2, \dots$ be the iteration index of the algorithm, and let us introduce the following notations

$$w_{i,j}^{(t)} = \frac{\pi_j^{(t)} f_{v_j}(x_i; \theta_j^{(t)})}{g(x_i; \eta^{(t)})};$$

$$Q(\eta, \eta^{(t)}) = \sum_{j=1}^{s} \sum_{i=1}^{n} w_{i,j}^{(t)} \log \pi_j + \sum_{j=1}^{s} \sum_{i=1}^{n} w_{i,j}^{(t)} \log f_{v_j}(x_i; \theta_j);$$

the quantity $w_{i,j}^{(t)}$ can be interpreted as the estimated posterior probability at the iteration t that the observation x_i has been drawn from the jth mixture component. For a given choice of $c \in (0, 1]$, the algorithm is as follows:

1. fix $\eta^{(0)} \in \Gamma_c$;

2. For all $t = 1, 2, \dots$, up to convergence do the following:

(a) E–step: determine $Q(\eta, \eta^{(t)})$;

(b) M–step: choose $\eta^{(t+1)} = \arg\max_{\eta \in \Gamma_c} Q(\eta, \eta^{(t)})$.

In the case of our finite mixture the M-step at iteration t is simply to compute

$$\pi_j^{(t+1)} = n^{-1} \sum_{i=1}^{n} w_{i,j}^{(t)}, \qquad j = 1, 2, \ldots, s, \tag{3.17}$$

$$\theta_j^{(t+1)} = \arg\max_{\theta_j} \sum_{i=1}^{n} w_{i,j}^{(t)} \log f_{v_j}(x_i; \theta_j) \qquad j = 1, 2, \ldots, s. \tag{3.18}$$

Wu (1983) established the theory of convergence of the EM algorithm under the assumption that the function Q computed in the E–step is continuous and differentiable at any iteration in all its arguments. Theorem 4.1 in Redner and Walker (1984) offers a summary of the results in Wu (1983). Under the assumption that ϕ is continuous with respect of θ_j the M-step is clearly well-defined for all $j = q+1, \ldots, s$. However the discontinuities introduced by the uniform components create some inconvenience. Fraley and Raftery (1998) proposed the model with one uniform component plus a finite number of Gaussian components. They fixed the uniform parameters such that uniform support is the range of the data[2]. This approach was also followed by Dean and Raftery (2005) where they used a one-dimensional normal-uniform mixture model for differential gene expression detection for cDNA microarrays. In the next proposition we show that this approach leads to an inconsistent estimate. Before we do that, let us introduce some notations. Given the sample $\underline{\mathrm{X}}^n$ we define two functions: for a constant $h \in \mathbb{R}$, $m_n(h) := \min\{x_i \in \underline{\mathrm{X}}^n \,:\, x_i \geq h\}$ and $M_n(h) := \max\{x_i \in \underline{\mathrm{X}}^n \,:\, x_i \leq h\}$.

Proposition 3.4. *For $j = 1, 2, \ldots, q$ let $\theta_j^{(0)}$ with $-\infty < a_j^{(0)} < b_j^{(0)} < +\infty$ be the initial values for the uniform parameters. Suppose that the interval $[a_j^{(0)}, b_j^{(0)}]$ contains at least two data points. Let n be fixed and finite. Then at any iteration $t = 1, 2, \ldots$ an EM solution is such that $a_j^{(t)} = m_n(a_j^{(0)}) < b_j^{(t)} = M_n(b_j^{(0)})$ for all $j = 1, 2, \ldots, q$.*

Proof. At iteration $t + 1$ the computation of the uniform parameters is done by solving the M-step for the uniform component, which is

$$(a_j^{(t+1)}, b_j^{(t+1)}) := \arg\max_{(a,b) \in \Theta_{1,c}} \sum_{i=1}^{n} w_{i,1}^{(t)} q_i(a_j, b_j);$$

[2] In their paper Fraley and Raftery (1998) used the convex hull of data since they treated the the multidimensional case.

where the set Θ_1 corresponds to (4.7) and $\Theta_{1,c} := \{\theta \in \Theta_1 : a < b\}$,

$$w_{i,j}^{(t)} := \pi_j^{(t)} \frac{\mathbf{1}_{[a_j^{(t)}, b_j^{(t)}]}(x_i)}{(b_j^{(t)} - a_j^{(t)})} \frac{1}{g(x_i; \eta^{(t)})};$$

$$q_i(a_j, b_j) := \log \frac{\mathbf{1}_{[a_j, b_j]}(x_i)}{b_j - a_j}.$$

From now onward every argument is intended to be valid for all $j = 1, 2, \ldots, q$. Let us consider the first iteration, i.e. $t = 1$. Any $a_j^{(t)} < a_j^{(0)}$ and $b_j^{(t)} > b_j^{(0)}$ is not a solution in the the M-step above. In fact for all i such that $x_i \notin \mathbb{R} \backslash [a_j^{(0)}, b_j^{(0)}]$ we have $w_{i,j}^{(0)} = 0$, while for all i such that $x_i \in [a_j^{(0)}, b_j^{(0)}]$ it results that $(b_j^{(1)} - a_j^{(1)})^{-1} < (b_j^{(0)} - a_j^{(0)})^{-1}$. The latter implies that for every $i = 1, 2, \ldots, n$ $w_{i,j}^{(0)} q_i(a_j^{(1)}, b_j^{(1)}) < w_{i,j}^{(0)} q_i(a_j^{(0)}, b_j^{(0)})$. Therefore the solution for the M-step has to be searched in $[a_j^{(0)}, b_j^{(0)}]$. For all i such that $x_i \in [a_j^{(0)}, b_j^{(0)}]$, $w_{i,j}^{(0)} > 0$. If $x_i \notin [a_j^{(1)}, b_j^{(1)}]$ it follows that $q_i(a_j^{(1)}, b_j^{(1)}) = -\infty$. Hence, the optimal solution is thus to take the smallest interval containing all $x_i \in [a_j^{(0)}, b_j^{(0)}]$, therefore $a_j^{(1)} = m_n(a_j^{(0)})$ and $b_j^{(1)} = M_n(b_j^{(0)})$. If we assume that $a_j^{(0)}$ and $b_j^{(0)}$ are two data points, then it is easy to see that $a_j^{(1)} = a_j^{(0)}$ and $b_j^{(1)} = b_j^{(0)}$. Now since $m_n(a_j^{(0)})$ and $M_n(b_j^{(0)})$ are two data points, taking $t = 2$ and applying the same argument would lead us to conclude that $a_j^{(t)} = m_n(a_j^{(0)})$ and $b_j^{(t)} = M_n(b_j^{(0)})$ at any iteration $t = 1, 2, \ldots$. Notice that since the parameter has to lie in Γ_c, then a choice of initial values such that $m_n(a_j^{(0)}) < M_n(b_j^{(0)})$ completes the proof. ■

Let us assume that η_n^{EM} is the limit point[3] of an EM run for a given set of starting values when the sample is $\underline{X_n}$. We consider a random sequence $\left\{\eta_n^{(0)}\right\}_{n \geq 1}$ of starting values, for each n we initialize the algorithm with $\eta_n^{(0)}$ and we compute the limit point η_n^{EM}. We assume that the sequence of starting values for the uniform parameters $\left\{(a_{j,n}^{(0)}, b_{j,n}^{(0)})\right\}_{n \geq 1}$ converges in probability to the finite limits $(a_{j,\infty}^{(0)}, b_{j,\infty}^{(0)})$ for all $j = 1, 2, \ldots, q$. We recall that $a_{j,0}$ and $b_{j,0}$ are the true uniform parameters.

Proposition 3.5. *If $(a_{j,\infty}^{(0)}, b_{j,\infty}^{(0)}) \neq (a_{j,0}, b_{j,0})$ for all $j = 1, 2, \ldots, q$, the sequence $\{\eta_n^{EM}\}_{n \geq 1}$ does not converge in probability to η_0.*

Proof. It follows easily from Proposition (3.4). We showed that for n fixed an EM run will lead us to $a_{j,n}^{EM} = m_n(a_{j,n}^{(0)})$ and $b_{j,n}^{EM} = M_n(b_{j,n}^{(0)})$. But $m_n(a_{j,n}^{(0)}) \xrightarrow{\text{P}} a_{j,\infty}^{(0)}$

[3] Up to this point we did not discuss about convergence of the EM algorithm for this model. We take convergence of the algorithm for granted at this point.

and $M_n(b_{j,n}^{(0)}) \xrightarrow{\text{P}} b_{j,\infty}^{(0)}$ for all $j = 1, 2, \ldots, q$. ■

If we have one uniform component and we always fix $a_n^{(0)} \leq \min\{\underline{\text{X}}_n\}$ and $b_n^{(0)} \geq \max\{\underline{\text{X}}_n\}$ as Fraley and Raftery (1998) proposed, it is easy to see that for n large the estimate computed by the EM algorithm are $m_n(a_n^{(0)}) = \min\{\underline{\text{X}}_n\} \xrightarrow{\text{as}} -\infty$ and $M_n(b_n^{(0)}) = \max\{\underline{\text{X}}_n\} \xrightarrow{\text{as}} +\infty$. Which means that the uniform component estimator vanishes as n gets large.

Proposition 3.4 above simply says that whatever is the pair of initial values for the parameters of each of the uniform components, if they coincide with a pair of distinct data points, then the EM algorithm will not update them. If these initial values do not coincide with pairs of distinct data points then they are updated only at the first step with pairs of data points. Proposition 3.4 together with Proposition 3.3 suggest a way to implement the EM algorithm which eventually lead to a maximum likelihood estimate. The latter is just to initialize the algorithm with the uniform components initialized for all pairs of data points, then for each initialization we compute a solution and we choose the one associated with the largest likelihood.

Let us put it formally. The strongly consistent constrained maximum likelihood estimator defined in (3.7) is such that the estimates of $\hat{a}_{j,n}$ and $\hat{b}_{j,n}$ coincide with data points for all n and $j = 1, 2, \ldots, q$.. Let D_n be the set of all pairs of distinct points of the observed sample $\underline{x_n}$ such that if $(x_r, x_p) \in D_n$ then $x_r < x_p$, for $p, r = 1, 2, \ldots, n$ and $r \neq p$. Assuming that there are $z \leq n$ distinct points in $\underline{x_n}$, the cardinality of D_n will be $d = z(z-1)/2$. Let Y_n be the set of all possible unordered q-tuples of elements of D_n, assuming that $d \geq q$. The cardinality of Y_n will be $h = (d+q)!/(d!q!)$. We denote an element of Y_n with $y_r = ((a_{1,r}, b_{1,r}), (a_{2,r}, b_{2,r}), \ldots, (a_{q,r}, b_{q,r}))$; the vector η_r is the parameter $\eta \in \Gamma$ with the uniform parameters set to be pairs of distinct data points: $\eta_r = (\pi_{1,r}, \ldots, \pi_{s,r}, y_r, \theta_{q+1,r}, \ldots, \theta_{s,r})$. As before the iteration index is t while $\varepsilon > 0$ is a fixed constant. Moreover for $r = 1, 2, \ldots, h$

$$w_{i,j,r}^{(t)} = \frac{\pi_{j,r}^{(t)} u(x_i; a_{j,r}, b_{j,r})}{g(x_i; \eta_r^{(t)})}, \qquad j = 1, 2, \ldots, q;$$

$$w_{i,j,r}^{(t)} = \frac{\pi_{j,r}^{(t)} \phi(x_i; \mu_{j,r}^{(t)}, \sigma_{j,r}^{(t)})}{g(x_i; \eta_r^{(t)})}, \qquad j = q+1, q+2, \ldots, s;$$

and

$$Q(\eta_r, \eta_r^{(t)}) = \sum_{j=1}^{s}\sum_{i=1}^{n} w_{i,j,r}^{(t)} \log \pi_{j,r} + \sum_{j=1}^{s}\sum_{i=1}^{n} w_{i,j,r}^{(t)} \log f_{v_j}(x_i; \theta_{j,r}).$$

We now give the afore-mentioned implementation of the EM algorithm:

```
1. For $r = 1, 2, \dots, h$ fix $\pi_{1,r}^{(0)}, \pi_{2,r}^{(0)}, \dots, \pi_{s,r}^{(0)}$ and $\theta_{q+1,r}^{(0)}, \theta_{q+2,r}^{(0)}, \dots, \theta_{s,r}^{(0)}$,

2. For all $r = 1, 2, \dots, h$ do {
   for all $t = 1, 2, ...,$ up to convergence compute the following:
      (a) E–step: determine $Q(\eta_r, \eta_r^{(t)})$;
      (b) M–step: choose $\eta_r^{(t+1)} = \arg\max_{\eta_r \in \Gamma_c} Q(\eta_r, \eta_r^{(t)})$.
   }
   Store $l_n(\eta_r^*)$, where $\eta_r^*$ is the parameter for which convergence in the condition above occurs.

3. Compute $\eta^* = \arg\max_r \{l_n(\eta_r^*)\}_{r=1}^{h}$
```

That is, in the M-step we get rid off the maximization over the uniform parameters. In fact, as noted before by Lemma 3.2, if we set the uniform parameters to distinct pairs of points we find a local maximum for the likelihood function. This follows from the fact that the log-likelihood function is at a local maximum each time the uniform parameters are set to be a pair of distinct points. Let us denote $\Theta_{2,c} := \{\theta \in \Theta_2 : \min_{i,j} \sigma_i \backslash \sigma_j \geq c\}$, the M-step for each $r = 1, 2, \dots, h$ is just to compute

$$\pi_{j,r}^{(t+1)} = n^{-1} \sum_{i=1}^{n} w_{i,j,r}^{(t)}, \tag{3.19}$$

$$(\mu_{j,r}^{(t+1)}, \sigma_{j,r}^{(t+1)}) = \arg \max_{(\mu_j, \sigma_j) \in \Theta_{2,c}} \sum_{i=1}^{n} w_{i,j,r}^{(t)} \log \phi(x_i; \mu_j, \sigma_j). \tag{3.20}$$

When ϕ is the Gaussian density, this is just the constrained problem studied in Hathaway (1986). Notice that the constraint with respect of the uniform parameters is embodied in the construction of the set Y_n. Convergence of the afore-mentioned EM algorithm follows easily:

Proposition 3.6. *For each $r = 1, 2, \dots, h$ the revised EM algorithm converges to a point $\eta_r^* \in \Gamma_c$ and the solution $\eta^* = \arg\max_{\eta \in \Gamma_c} l_n(\eta)$ exists.*

Proof. By Proposition 3.2 the maximum of the log-likelihood function exists on Γ_c. By the same argument in Theorem 4.1 in Redner and Walker (1984) the log-likelihood function increases at each step. By the the same arguments in the proof of 3.2 the M-step has a solution at each t for any r. Applying the proof of Theorem 4.1(v) in Redner and Walker (1984) we get the result. ■

Depending on the number of observations and the number of uniform components the algorithm described can be computationally infeasible. However, when we have more than one uniform component the computational complexity can be reduced under some assumptions on the true model. In practice we need to consider some selection rule which allows us to reduce h. This will be seen in more details within the simulation study which we will explore in Chapter 5.

§3.6. Conclusions

In this chapter we defined a mixture model where a finite number of uniform densities is mixed with a finite number of location-scale densities. We estimate such a model by maximum likelihood. This model is suggested when outliers are mostly concentrated on a certain region of the data range. We defined and showed identifiability for this model as well as we developed the estimation theory. The asymptotic for such an estimator is not trivial. The log-likelihood function is not continuous and in order to get an equivariant estimator we need to restrict the parameter space by non-smooth constraints. However, we have been able to show that the sequence of maximum likelihood estimates is strongly consistent. We also developed the EM algorithm for such an estimator and we highlighted some practical problems when the number of uniform components is larger than one.

CHAPTER 4

Robust Improper Maximum Likelihood Estimator

In this chapter we study an estimator which is defined as the maximizer of the sample mean of the logarithm of an improper density. We provide a justification for it in terms of robustness, we develop the related asymptotic theory, and we derive a computational method based on the EM algorithm.

§4.1. Introduction

In this chapter we introduce a robust method to do model-based cluster analysis. We will introduce an estimator which is the maximizer of the sample mean of the logarithm of an improper density. The idea and the motivations for such an estimator has been presented in the work of Hennig (2004). The author built the robustness theory for model-based cluster analysis based on the concept of breakdown behaviour (see Hampel, 1971; Donoho and Huber, 1983). He showed that the maximum likelihood estimates for a wide family of finite mixtures of location-scale distributions are not robust against outliers. The family under consideration also includes finite mixtures of Gaussians, Gaussians plus uniform noise supported on the range of the data (see Fraley and Raftery, 1998), as well as finite mixtures of t-distributions. The afore-mentioned paper also proposed the following robust methodology: given the finite location-scale mixture under consideration, the author suggested to add a component represented by a constant density over the real line (an improper density). This improper density has the role to catch even extreme outliers. The maximizer of the log-likelihood function associated with the improper density is breakdown-robust even in the presence of extreme outliers. For mixtures with fixed number of components this methodology exhibits the best breakdown behaviour amongst those considered in the afore-mentioned paper. In this chapter we will refer to such estimator as the "robust improper maximum likelihood estimator". In Hennig (2004) the value of this constant improper density is considered to be fixed. The author also gives some guidelines about how to fix it in real situations. The paper by Hennig (2004) does not contain any estimation theory about the proposed methodology, nor it does contain empirical evidence of performance.

The main contribution of this chapter is the following: (i) we construct an estimation theory for the robust improper maximum likelihood estimator (Section 4.2 and 4.3); (ii) we provide a feasible computational method to compute the resulting estimator based on the EM algorithm (Section 4.4); (iii) we argue that the choice of the constant improper density is crucial and not trivial. We also propose a data-dependent methodology for choosing the improper density (Section 4.5). We conducted an extensive simulation study where the robust improper maximum likelihood estimator has been compared against other robust alternatives under a variety of data generating processes. The simulation study will be presented in the next chapter.

§4.2. Model and estimation

4.2.1 — Data generating process and model

Let $(\Omega, \mathcal{A}, \mathcal{P})$ be a probability space. Let X be a real valued random variable defined on the afore-mentioned space. In this thesis we are mainly interested in one dimensional random variables. However, the results of the theory presented here are easily applicable to multidimensional random variables without any restriction. The random variable X has a proper distribution function Q which represents the measure $\mathcal{P}$. Let $\underline{X_n}$ be a random sample from Q, i.e. $\underline{X_n} := \{X_1, X_2, \dots, X_n\}$, where all X_i are independent replicas of X, $i = 1, 2, \dots, n$. The vector of observations $\underline{x_n} := \{x_1, x_2, \dots, x_n\}$ is a realization of $\underline{X_n}$.

The distribution Q is the true data generating process. In other words Q is the distribution behind the collected data. The researcher does not know Q and his or her aim is to infer some of its features. Our ultimate goal here is not to provide an estimate of Q, we do not even attempt to estimate it. While the approach presented here can be adopted in more general situations – we will come back to this point later – the class of problems we are mainly interested in, are of the type of clustering and classification.

In our context, we are interested in those situation where Q generates data with group structure. As highlighted before, we are also interested in situations where outliers are present in the data. We want to be able to infer the group structure of the data and summary statistics for the sub-populations composing the sample. For each $x_i \in \underline{x_n}$ we want to infer to which of the sub-population x_i belongs, and we to produce statistics for location and scale measures for all the sub-populations. The strategy presented here is the following: (i) the researcher assumes that there are s sub-populations; (ii) we model the data as coming from a finite mixture with components represented by densities belonging to some parametric family; (ii) we add a component represented by a constant density on the real line, i.e. an improper density. The latter component has the role of capturing noise. The resulting mixture density representing the population is improper; (iii) we maximize the associated "improper" sample log-likelihood function in order to estimate location, scale and proportion parameters; (iv) such estimate is used to assign data points to groups via the Bayes' rule.

There is an important issue to be stressed before we introduce notation and assumptions. It should be clear now that it is not assumed that the data are generated by a probability measure represented by an improper density. The true (and unknown) data generating process is Q. Instead we define an improper density as a device to estimate some of the features of the population. In classical parametric maximum likelihood estimation, we assume that the family of probability models under which we build the likelihood function is the same family of models to which belongs the distribution generating the data. Here the perspective is different, we do not assume any model for the data generating process, and we built a "pseudo-model" which consists in an improper density function with the role of capturing some features of the population under study. The term pseudo-model is due to the fact the we want to distinguish this approach with the classical parametric set-up described above. In some particular case the pseudo-model can be interpreted as model, but this will be clearer thereafter and we shall discuss that in Section 4.4. The idea presented here – and the related theory that we are going to develop – is applicable in more general situations beyond classification and cluster analysis.

We now describe our pseudo-model in more details. Let $\Phi(x; \mu, \sigma)$ be a distribution function on the real line parameterized in terms of a location parameter $\mu \in \mathbb{R}$ and a scale parameter $\sigma \in R_+$. $\Phi(x; \mu, \sigma)$ is represented by a density function $\phi(x; \mu, \sigma)$. More assumptions about ϕ will be given to prove the following statements. The role of Φ is to catch the structure of the sub-populations, while the pseudo-model for the main population is given by a finite mixture where each component is distributed according to Φ. We are in a situation where the researcher knows how many sub-populations are present. Hence, from now onwards the number of groups in the population under study is fixed to be equal to s, with $1 < s < \infty$. This means that s ought not to be estimated from the data. Our main aim is to take into consideration the presence of outliers or noise. As before in this thesis the term noise means observations that are not supposed to be coming from any of the sub-populations. We account for noise by introducing a further mixture component with a density which is constant over the real line, let us say that it is $c \in (0, \bar{c}]$. Therefore, we account for the presence of noise by introducing an improper constant density. We name the latter the "noise component". The value of c is fixed throughout the rest of the chapter. Later we will discuss criteria to choose suitable value for the c. The pseudo-model we are going to consider is the following

$$\lambda_c(x; \eta) = \pi_1 c + \sum_{j=2}^{s} \pi_j \phi(x; \mu_j, \sigma_j); \tag{4.1}$$

where $\eta = (\pi_1, \ldots, \pi_s, \mu_2, \ldots, \mu_s, \sigma_2, \ldots, \sigma_s)$, $0 < \pi_j < 1$ for $j = 1, \ldots, s$, and $\sum_{j=1}^{s} \pi_j = 1$. The ultimate goal is to estimate the parameter $\eta \in \Gamma$, with $\Gamma := [0,1]^s \times \mathbb{R}^{s-1} \times \mathbb{R}_+^{s-1} \backslash \{0\}$. Our λ_c is not a proper density. The improper density c can be thought as the approximation of the density of the noise component. The latter can be everything, but here the model is built following some simple considerations based on the meaning that we attach to the noise component: (i) the noise is understood here as a component having a large support and a small density value, (ii) the noise component can cause extreme outliers, so that a noise component has to attach positive probabilities even to events corresponding to extreme values.

When an estimate of η is available we then have statistics about location and scale measures of the sub-populations, and we can also assign observations to the sub-populations. Let $\hat{\eta}_{c,n}$ be an estimate of η based on the observed sample $\underline{x_n}$, than we can compute $\hat{\tau}_{c,n}(i,j) = \hat{\pi}_j^{c,n} \phi(x_i; \hat{\mu}_j^{c,n}, \hat{\sigma}_j^{c,n}) / \lambda_c(x_i; \hat{\eta}_{c,n})$ for all $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, s$. The quantity $\hat{\tau}_{c,n}(i,j)$ can be interpreted as the "improper posterior probability" (conditioning on the observed sample) that the observation x_i comes form the jth group in the population. The term "improper" is due to the fact that $\lambda_c(x_i; \hat{\eta}_{c,n})$ is an improper density. In particular $\hat{\tau}_{c,n}(i,1) = \hat{\pi}_1^{c,n} c / \lambda_c(x_i; \hat{\eta}_{c,n})$ defines the "improper posterior probability" that the ith observation is generated by the noise component. In same situations $\hat{\tau}_{c,n}(i,1)$ can also be interpreted as posterior probability, and we shall discuss these things in the following sections.

The estimated vector η can be used to classify points. First we compute

$$\hat{\tau}^*_{c,n}(i) := \arg \max_{j=1,\dots,s} \hat{\tau}_{c,n}(i,j), \quad i = 1, 2, \dots, n; \tag{4.2}$$

based on this, the ith observation is assigned to the the jth component if $\hat{\tau}^*_{c,n}(i) = j$. In the next paragraph we are going to describe the method of estimation used to compute $\hat{\eta}_{n,c}$.

4.2.2 — Robust improper maximum likelihood estimator

The estimation of the parameter η is obtained via the maximization of the "improper" sample log-likelihood function associated with $\lambda_c(x, \eta)$. Before we introduce the new estimator we need to set some additional assumptions on ϕ.

Assumption 4.1. The density ϕ belongs to the location-scale family, i. e.

$$\phi(x; \theta) = \frac{1}{\sigma} \phi\left(\frac{x - \mu}{\sigma}\right).$$

Furthermore $\phi(x; \theta)$ is continuous with respect to θ at all x, and it is measurable for every θ; $\lim_{|z| \to \infty} \phi(z) = 0$, $\phi(0) < \infty$, $\lim_{\sigma \downarrow 0} \phi(x; \theta) = 0$ for all $x \neq \mu$, moreover $\phi(x; \theta)$ is monotonically increasing for each $x \in (-\infty, \mu]$ and it is monotonically decreasing for each $x \in [\mu, +\infty)$.

The Assumption above includes a wide variety of models including those considered by Hennig (2004). Notice that if ϕ is the Gaussian density, it is easily shown that it satisfies Assumption 4.1. We now introduce the target function that we maximize in order to obtain our estimator:

$$l_{c,n}(\eta) := \frac{1}{n} \sum_{i=1}^{n} \log \lambda_c(x_i, \eta). \tag{4.3}$$

We call $l_{c,n}(\eta)$ the "improper log-likelihood function". Our estimate of η would be a maximizer of $l_{c,n}(\eta)$. However the function $l_{c,n}(\eta)$ is not bounded over Γ. In fact – as already highlighted in the previous chapters – if we take $\mu_j = x_i$ for some $i = 1, \dots, n$ and $j = 1, \dots, s$ then for $\sigma_j \downarrow 0$ we get $l_{c,n}(\eta) \longrightarrow +\infty$.

In order to obtain a maximum we need to restrict the parameter space. The kind of scale-equivariant constraints studied in Hathaway (1985) and in the previous chapter are not feasible in this situation. Those constraints work in situations when summands of the log-likelihood function go to zero at a certain speed when the scale parameters get arbitrarily close to zero and the location parameters are different from data point values. This argument does not apply in this situation because the presence of the improper term in λ_c causes $\lambda_c(x, \eta) \geq \pi_1 c > 0$ for all $x \in \mathbb{R}$. The restricted parameter space we adopt here is $\Gamma_t := \{\eta \in \Gamma : \sigma_j \geq t > 0, j = 2, \dots, s\}$.

Unfortunately the function $l_{c,n}(\eta)$ cannot be expected to have a unique maximum. If we take the vector η and we permute some of the triples (π_j, μ_j, σ_j) we still obtain the same value for $l_{c,n}$. This is a label-switching problem which also occurs in classical mixture estimation as seen in previous chapters. While in classical mixture model estimation we know that the only source of multiple maxima is the label-switching problem, this is not the case here. Together with label-switching there could be other causes for multiple maxima. Thus we define the improper maximum likelihood estimator (RIMLE) as a member of the set of maxima of the target function, that is

$$\hat{\eta}_{c,n} \in \hat{\Delta}_{c,n} := \arg\max_{\eta \in \Gamma_t} l_{c,n}(\eta). \tag{4.4}$$

The choice of $\eta_{c,n}$ among the members of $\Delta_{c,n}$ is irrelevant to prove the next lemmas. Moreover, the RIMLE is not scale-equivariant. In fact for a given fixed t if we multiply all the observations for some constant, it can be that the scale of one of the sub-population becomes smaller than t and thus it is not contained in the constrained set anymore. The next Lemma proves that under such a constraint a global maximum for the improper log-likelihood function does exist.

Proposition 4.1. *Under Assumption 4.1 $l_{c,n}(\eta)$ achieves its maximum on Γ_t*

Proof. The proof goes along the same lines as in Lemma 3.1 in Chapter 3. First we show that there exists a compact set $\bar{\Gamma}_t \subseteq \Gamma$, such that $\sup_{\eta \in \Gamma_t} l_{c,n}(\eta) = \sup_{\eta \in \bar{\Gamma}_t} l_{c,n}(\eta)$. Let us fix some notations: we denote $m_n := \min\{x_i, \quad i = 1, \dots, n\}$ and $M_n := \max\{x_i, \quad i = 1, \dots, n\}$. The proof is now divided into

two parts.
Part A. Let us take $\eta' \in \Gamma_t$ with $\mu'_j \leq m_n$ for some $j = 2, \ldots, s$. We also consider the vector $\eta'' \in \Gamma_t$ which is equal to η' except that $\mu''_j = m_n$. This implies that $l_{c,n}(\eta') \leq l_{c,n}(\eta'')$ because the density $\phi(x; \mu, \sigma)$ is monotonically increasing for each $x \in (-\infty, \mu)$ by Assumption 4.1. By analogy we take $\eta' \in \Gamma_t$ with $\mu'_j \geq M_n$ for some $j = 2, \ldots, s$. Then we consider the vector $\eta'' \in \Gamma_t$ which is equal to η' except that $\mu''_j = M_n$. This implies that $l_{c,n}(\eta') \leq l_{c,n}(\eta')$.
Part B. Let us take a sequence $\{\eta_k\}_{k\geq 1}$, where η_k is such that $\sigma_j^k \longrightarrow +\infty$ for some $j = 2, \ldots, s$. The vector $\eta' \in \Gamma_t$ is equal to η_k except that all scales depending on k are now fixed. As k gets large enough the component densities depending on k will take values close to zero for all finite x, this mens that $\lim_{k\to\infty} l_{n,c}(\eta_k) \leq l_{n,c}(\eta')$.
By the results in parts A and B we can conclude that $\sup_{\eta\in\Gamma_t} l_{c,n}(\eta) = \sup_{\eta\in\bar{\Gamma}_t} l_{c,n}(\eta)$; where $\bar{\Gamma}_t := [0,1]^s \times M_n^{s-1} \times \Sigma_{t,n}^{s-1}$, with

$$M_n := \{\mu \in \mathbb{R} : m_n \leq \mu \leq M_n\}, \tag{4.5}$$

and

$$\Sigma_{t,n} := \{\sigma \in \mathbb{R} : 0 < t \leq \sigma \leq \bar{\sigma}\}, \tag{4.6}$$

for some choice of the constant $\bar{\sigma} < +\infty$. The sets M_n and $\Sigma_{t,n}$ are now compact as well as the set $\bar{\Gamma}_t$. By Assumption 4.1 $l_{c,n}(\eta)$ is continuous on $\bar{\Gamma}_t$ which implies that it achieves its maximum. The latter completes the proof. ■

§4.3. Strong Consistency

In classical parametric maximum likelihood theory we assume that there is a "true" parameter value which generated the statistical experiment under study. In this setup consistency means convergence in probability or almost sure convergence to the true parameter. Here we do not have a true parameter and consistency means convergence in probability or almost sure convergence of a sequence of estimators to some point in the parameter space. We want to study conditions under which for n large enough a sequence $\{\hat{\eta}_{c,n}\}_{n\geq 1}$ approaches some value $\eta_* \in \Gamma_t$ with probability one. Under some regularity conditions if for n large the sequence of functions $\{l_{n,c}(\eta)\}_{n\geq 1}$ converges uniformly to some function $l_c(\eta)$, and η_* is the unique maximizer of $l_c(\eta)$ then we can show that the sequence of maximizers of $\hat{\eta}_{c,n}$ converges to η_*. This approach dates back to the seminal work of Jennrich (1969), and it is part of the more recent developments of the empirical processes theory (see Van der Vaart and Wellner, 1996). Unfortunately the function $l_c(\eta)$ cannot be expected to have a unique maximum in this setup. Under some regularity conditions we shall show that as n gets large then

$$l_{c,n}(\eta) \xrightarrow{\text{as}} l_c(\eta) = \int \log \lambda_c(x, \eta) dQ,$$

uniformly over Γ_t. Let us suppose that η_* is a global maximum for $l_c(\eta)$. If we take the vector η_* and we permute some of the triples $(\pi_j^*, \mu_j^*, \sigma_j^*)$ we still obtain the same integral $l_c(\eta_*)$. Again this is due to the label-switching problem. Again, there could be other causes for the existence of multiple maxima, and the set of maximizers of $l_c(\eta)$ could even not be countable. We avoid the problem of multiple maxima of the asymptotic target function studying convergence on a quotient topological space of Γ. This method will be clearer later. Before we procede with the next Lemmas we need to introduce a further

Assumption 4.2. For every μ and σ such that $\eta \in \Gamma_t$ there exists a measurable function $t(x)$ such that $|\log \phi(x; \mu, \sigma)| \leq t(x)$ for every x, and $t(x)$ is integrable with respect to Q.

Lemma 4.1. *Under Assumption 4.2, there exists a function T integrable with respect to Q such that $|\log \lambda_c(x; \eta)| \leq T(x)$ for every x and η.*

Proof. We define the following sets: $A := \{x \in \mathbb{R} : \lambda_c(x; \eta) \geq 1\}$, and $B := \{x \in$

$\mathbb{R} : \lambda_c(x) < 1\}$. The following chain of inequalities proves the statement

$$\begin{aligned}
|\log \lambda_c(x)| &= \left|\log\left(\pi_1 c + \sum_{j=2}^{s} \pi_j \phi_j(x)\right)\right| \mathbf{1}_A(x) + \\
&\quad + \left|\log\left(\pi_1 c + \sum_{j=2}^{s} \pi_j \phi_j(x)\right)\right| \mathbf{1}_B(x) \\
&\leq \left|\pi_1 c + \sum_{j=2}^{s} \pi_j \phi_j(x)\right| \mathbf{1}_A(x) + \\
&\quad + \left[|\log \pi_1 c| + \sum_{j=2}^{s} |\log \pi_j \phi_j(x)|\right] \mathbf{1}_B(x) \\
&\leq \left[\pi_1 c + \sum_{j=2}^{s} \pi_j \phi_j(x)\right] \mathbf{1}_A(x) + \\
&\quad + \left[|\log \pi_1 c| + \sum_{j=2}^{s} \pi_j t_j(x)\right] \mathbf{1}_B(x) = T(x)
\end{aligned}$$

By Assumption 4.2 $T(x)$ is integrable with respect to Q. ■

The next Proposition establishes that the functional version of our target function achieve its maximum over Γ_t.

Proposition 4.2. *Under Assumptions 4.1 and 4.2, $l_c(\eta)$ achieves its maximum over Γ_t.*

Proof. By the Lemma 4.1 and by the dominated convergence theorem, for any sequence $\{\eta_k\}_{k\geq 1}$,

$$\lim_{k\longrightarrow\infty} l_c(\eta_k) = \int \lim_{k\longrightarrow\infty} \log \lambda(x; \eta_k) dQ;$$

the latter will be useful throughout the proof. Now we want to show that there exists a compact set $\bar{\Gamma}_t \subseteq \Gamma_t$ such that $\sup_{\Gamma_t} l_c(\eta) = \sup_{\bar{\Gamma}_t} l_c(\eta)$. This is showed in two steps along the same lines of the proof of the Proposition 4.1.

Part A. Let us take a sequence $\{\eta_k\}_{k\geq 1}$, where none of the elements of η_k depends on k except some of the the location parameters $\mu_{\bar{j}}^k \longrightarrow \pm\infty$, for some index $\bar{j}$. Without loss of generality we assume that $\bar{j} \in \{2, 3, \ldots, q\}$ for some $q \leq s$. We also consider a vector η', which is the same vector as η_k with all the location parameters fixed to be μ'_j, with $|\mu'_j| < \infty$ for all $j = 2, \dots, q$. By Assumption 4.1 we can write

$$\lim_{k\to\infty} \log\left(\pi_1 c + \sum_{j=2}^{q} \pi_j \phi(x; \mu_j^k, \sigma_j) + \sum_{j=q+1}^{s} \pi_j \phi(x; \mu_j, \sigma_j)\right) \le$$

$$\le \log\left(\pi_1 c + \sum_{j=2}^{q} \pi_j \phi(x; \mu'_j, \sigma_j) + \sum_{j=q+1}^{s} \pi_j \phi(x; \mu_j, \sigma_j)\right)$$

Integrating both sides of the previous inequality against dQ, taking limits and applying the dominated convergence theorem we get $\lim_{k\to\infty} l_c(\eta_k) \le l_c(\eta')$.
Part B. We now assume that $\{\eta_k\}_{k\ge1}$ is a sequence where all the elements of η_k do not depend on k except some of the the scale parameters $\sigma^k_{\bar{j}} \longrightarrow +\infty$, for some index $\bar{j}$. Without loss of generality we assume again that $\bar{j} \in 2, 3, \dots, q$ for some $q < s$. We also consider a vector η', which is the same vector as η_k with all the scale parameters fixed to be $\sigma_j < +\infty$, for all $j = 2, \dots, q$. By Assumption 4.1 we can write

$$\lim_{k\to\infty} \log\left(\pi_1 c + \sum_{j=2}^{q} \pi_j \phi(x; \mu_j, \sigma_j^k) + \sum_{j=q+1}^{s} \pi_j \phi(x; \mu_j, \sigma_j)\right) \le$$

$$\le \log\left(\pi_1 c + \sum_{j=2}^{q} \pi_j \phi(x; \mu_j, \sigma'_j) + \sum_{j=q+1}^{s} \pi_j \phi(x; \mu_j, \sigma_j)\right)$$

By the same argument as above we have that $\lim_{k\to\infty} l_c(\eta_k) \le l_c(\eta')$. By results in part A and B we can conclude that $\sup_{\eta\in\Gamma_t} l_c(\eta) = \sup_{\eta\in\bar{\Gamma}_t} l_c(\eta)$; where $\bar{\Gamma}_t := [0,1]^s \times M^{s-1} \times \Sigma_t^{s-1}$, with

$$M := \left\{\mu \in \mathbb{R} : \underline{\mu} \le \mu \le \bar{\mu}\right\}, \tag{4.7}$$

and

$$\Sigma_t := \{\sigma \in \mathbb{R} : t \le \sigma \le \bar{\sigma}\}, \tag{4.8}$$

for some choice of the constants of $-\infty < \underline{\mu} < \bar{\mu} < +\infty$ and $\bar{\sigma} < +\infty$. The sets M and Σ_t are now compact as well as the set $\bar{\Gamma}_t$.
Part C. For any sequence $\{\eta_k\}_{k\le1}$, such that $\eta_k \longrightarrow \eta$ we have that

$$\lim_{k\to\infty} \log \lambda_c(x; \eta_k) = \log \lambda_c(x; \eta),$$

and the latter is implied by Assumption 4.1. By applying the dominated convergence theorem we get the continuity of $l_c(\cdot)$. The continuity of $l_c(\cdot)$ together with the compactness of $\bar{\Gamma}_t$ proves the desired result. ■

The previous result implies that there exists $\eta^* \in \Gamma_t$ such that $l_c(\eta^*)$ is a global maximum. Notice that by the previous remark the maximum is not unique. We also notice that a maximum for l_c is contained in $\bar{\Gamma}_t \subseteq \Gamma_t$. Thus even if we refer to Γ_t, it is obvious that whenever η^* is maximum for l_c, then η^* is contained in $\bar{\Gamma}_t$. In order to deal with multiple maxima we need to introduce some more notation. We define the set

$$H(\eta') := \left\{ \eta \in \bar{\Gamma}_t : \int \log \lambda_c(x;\eta) dQ = \int \log \lambda_c(x;\eta') dQ \right\}.$$

Let δ a distance on Γ, (e.g. the Euclidian distance), for any $\varepsilon > 0$ we define

$$\mathscr{N}_\varepsilon(\eta') := \left\{ \eta \in \bar{\Gamma}_t : \forall \eta^* \in H(\eta') \quad \delta(\eta,\eta^*) < \varepsilon \right\},$$

and

$$T(\eta',\epsilon) := \bar{\Gamma}_t \backslash \mathscr{N}_\varepsilon(\eta').$$

The set $T(\eta',\epsilon)$ contains all the points of Γ_t that are distant at least an ϵ from all those points of $\bar{\Gamma}_t$ that give a value for the functional version of the target function equal to $l_c(\eta')$. Let us define the equivalence relation $\sim$ as follows:

$$\forall (\eta,\eta') \in \bar{\Gamma}_t, \qquad \eta \sim \eta' \Leftrightarrow l_c(\eta) = l_c(\eta').$$

Let $\tilde{\Gamma}$ the space obtained from $\bar{\Gamma}_t$ by identifying $H(\eta')$ to a point $\tilde{\eta}' = \eta'$. More precisely the space $\tilde{\Gamma}$ is the quotient topological space generated by the equivalence relation $\sim$ on the space $\bar{\Gamma}_t$. It is easy to see that

$$T(\eta',\epsilon) \equiv \tilde{\Gamma}_t \backslash B_\varepsilon(\tilde{\eta}'),$$

where $B_\varepsilon(\tilde{\eta}')$ is an open ball of radius ε centred on $\tilde{\eta}' \in \tilde{\Gamma}$. This will allow us to show consistency on the quotient space, and this strategy is inspired by the work of Redner (1981) about strong consistency of the maximum likelihood estimator for non identifiable distributions.

Proposition 4.3. *Let $X_1, \dots, X_n$ an iid sample from Q, under Assumptions 4.1 and 4.2, for every $\epsilon > 0$ and for any sequence $\{\hat{\eta}_{c,n}\}_{n \geq 1}$ of maximizers for $l_{c,n}$*

$$\mathcal{P}\{\exists n_0 : \forall n \geq n_0 \quad \hat{\eta}_{c,n} \in T(\eta_*,\epsilon)\} = 0 \tag{4.9}$$

Proof. Part A. By Assumption 4.1 $\lambda_c(x;\eta)$ is continuous in η at any x and it is measurable with respect to Q at each η. By Lemma 4.2 $\bar{\Gamma}_t \subseteq \Gamma_t$ is compact and it contains a maximum for l_c. By Lemma 4.1 there exists $T(x)$ which is an integrable function with respect to Q, and such that $|l_c(x;\eta)| \leq T(x)$. Sufficient conditions for Theorem 2 in Jennrich (1969) are satisfied, which implies that $l_{n,c}(\eta)$ converges uniformly to $l_c(\eta)$ with probability one on $\bar{\Gamma}_t$, i.e.

$$\sup_{\eta\in\bar{\Gamma}_t} |l_{c,n}(\eta) - l_c(\eta)| \xrightarrow{\text{as}} 0. \tag{4.10}$$

Part. B. Now we want to show that $l_c(\hat{\eta}_{c,n}) \xrightarrow{\text{as}} l_c(\eta_*)$. Let us consider the following chain of inequalities:

$$\begin{aligned} 0 &\leq l_c(\eta_*) - l_c(\hat{\eta}_{c,n}) \\ &= l_c(\eta_*) - l_{c,n}(\hat{\eta}_{c,n}) + l_{c,n}(\hat{\eta}_{c,n}) - l_c(\hat{\eta}_{c,n}) \\ &\leq l_c(\eta_*) - l_{c,n}(\eta_*) + l_{c,n}(\hat{\eta}_{c,n}) - l_c(\hat{\eta}_{c,n}) \\ &\leq 2\sup_{\eta\in\bar{\Gamma}_t} |l_{c,n}(\eta) - l_c(\eta)| \xrightarrow{\text{as}} 0 \end{aligned}$$

which implies that $l_c(\hat{\eta}_{c,n}) \xrightarrow{\text{as}} l_c(\eta_*)$.

Part C. By continuity of $l_c(\eta)$ and Proposition 4.2 we have that for every $\epsilon > 0$ there exists a $\beta > 0$ such that $l_c(\eta) + \beta < l_c(\eta_*)$ for all $\eta \in T(\eta_*, \epsilon)$. Let us consider two sets

$$A_n := \{\omega \in \Omega : \hat{\eta}_{c,n} \in T(\eta_*, \epsilon)\},$$

and

$$B_n := \{\omega \in \Omega : |l_c(\hat{\eta}_{c,n}) - l_c(\eta_*)| > \beta\}.$$

By construction $A_n \subseteq B_n$ for all n, which means that $\mathcal{P}\{A_n\} \leq \mathcal{P}\{B_n\}$ for all n. By Part B, for large n we have that $\mathcal{P}\{B_n\} = 0$, which implies that for n large enough $\mathcal{P}\{A_n\} = 0$. The latter proves the desired result. ■

The previous Proposition implies strong convergence on the quotient topological space, i.e. $\tilde{\eta}_{c,n} \xrightarrow{\text{as}} \tilde{\eta}^*$.

§4.4. Computation of the RIMLE

In this section we propose an EM algorithm to compute the improper maximum likelihood estimator and we show that the theory of the convergence of the EM algorithm still holds. The reason for this is that the improper density c can be interpreted as the value of a uniform density supported on a subset of $\mathbb{R}$ containing all data points. If this is the case, λ_c can be interpreted as proper density, and implementation of the EM algorithm's machinery follows easily.

Let us give an example to show this interpretation. We consider a collection of $q \leq n$ disjoint intervals such that their union

$$H := \bigcup_{g=1}^{q} [a_g, b_g]$$

contains the data set. We introduce the following density function

$$h(x) = c \sum_{g=1}^{q} \mathbf{1}_{[a_g,b_g]}(x) \quad x \in \mathbb{R}; \tag{4.11}$$

The density h attributes a value c to all those points which are contained in H. The h is a uniform density on a disconnected subset of $\mathbb{R}$. Since c is positive, in order to obtain a proper density function we need that the integral of h over $\mathbb{R}$ is equal to 1. That is

$$\int_{-\infty}^{+\infty} h(x)dx = \sum_{g=1}^{q} \int_{a_g}^{b_g} cdx = c \sum_{g=1}^{q} (b_g - a_g) = 1.$$

Hence, we can choose the intervals $[a_g, b_g]$ such that

$$c = \frac{1}{\sum_{g=1}^{q} (b_g - a_g)},$$

and we obtain a proper density h which has constant value c at all points in H. In particular c is the reciprocal of the Lebesgue measure of the set H. The density h is one of the many possible interpretations of c. We note that if we assume that the distribution Q is represented by a mixture of $s - 1$ ϕ-densities plus a component having density h, then the model underlying the data generating process would coincide with the pseudo-model λ_c. Notice that this interpretation allows us to justify many different values for the c. If we need a small c we can take $q = 1$, a_1 small enough and b_1 large enough so that $[a_1, b_1]$ contains all the data points. In the latter case c can be less than the reciprocal of the range of the data. When ϕ is Gaussian, taking c equal the reciprocal of the range of the data would lead to the model proposed in Fraley and Raftery (1998). If we need a large value of c, it suffices to take the intervals $[a_g, b_g]$ small enough but containing each a single data point. In this case the sum of the lengths of the intervals can be made very small. This means that the value of c would be large.

We must stress that the construction of such a density h depends on the data set. But in practice this helps for the implementation of an EM algorithm which is the aim of this section. The EM algorithm is intended to seek for a maximum of the improper log-likelihood function

$$l_{c,n}(\eta) = \sum_{i=1}^{n} \log \left[\pi_1 c + \sum_{j=2}^{s} \pi_j \phi(x_i; \mu_j, \sigma_j) \right]$$

over the constrained set Γ_t. Let the index $k = 1, 2, \ldots$ be the iteration index of the algorithm. Along the same lines as in previous chapters we introduce the following notations

$$w_{i,1}^{(k)} = \frac{\pi_1^{(k)} c}{\lambda_c(x_i; \eta^{(k)})};$$

$$w_{i,j}^{(k)} = \frac{\pi_j^{(k)} \phi(x_i; \mu_j^{(k)}, \sigma_j^{(k)})}{\lambda_c(x_i; \eta^{(k)})}, \quad j = 2, 3, \ldots, s;$$

$$Q(\eta, \eta^{(k)}) = \sum_{i=1}^{n} \sum_{j=1}^{s} w_{i,j}^{(k)} \log \pi_j + \sum_{i=1}^{n} \left[w_{i,1}^{(k)} \log c + \sum_{j=2}^{s} w_{i,j}^{(k)} \log \phi(x_i; \mu_j, \sigma_j) \right];$$

$$H(\eta, \eta^{(k)}) = \sum_{i=1}^{n} \left[w_{i,1}^{(k)} \log \frac{\pi_1 c}{\lambda_c(x; \eta)} + \sum_{j=2}^{s} w_{i,j}^{(k)} \log \frac{\pi_j \phi(x_i; \mu_j, \sigma_j)}{\lambda_c(x; \eta)} \right];$$

and

$$l_{c,n}(\eta) = Q(\eta, \eta^{(k)}) - H(\eta, \eta^{(k)}).$$

For a given choice of t, the algorithm is as follows:

1. fix $\eta^{(0)} \in \Gamma_t$;
2. For all $k = 1, 2, ...,$ up to convergence do the following:
   (a) E–step: determine $Q(\eta, \eta^{(k)})$;
   (b) M–step: choose $\eta^{(k+1)} = \arg\max_{\eta \in \Gamma_t} Q(\eta, \eta^{(k)})$.

The M-step at iteration k is simply to compute

$$\pi_j^{(k+1)} = n^{-1} \sum_{i=1}^{n} w_{i,j}^{(k)}, \qquad j = 1, 2, \ldots, s, \tag{4.12}$$

$$(\mu_j^{(k+1)}, \sigma_j^{(k+1)}) = \arg \max_{\mu_j, \sigma_j \geq t} \sum_{i=1}^{n} w_{i,j}^{(t)} \log \phi(x_i; \mu_j, \sigma_j) \qquad j = 2, 3, \ldots, s. \quad (4.13)$$

As in the case of the maximum likelihood estimator for the mixture with uniform noise, here the M-steps are simply to compute a weighted maximum likelihood estimator for each of the location-scale components. This reduces the complexity of the computational effort by a considerable margin. Of course when ϕ is Gaussian, the M-step in 4.13 becomes

$$\mu_j^{(k+1)} = \left(\sum_{i=1}^{n} w_{i,j}^{(t)} \right)^{-1} \sum_{i=1}^{n} w_{i,j}^{(t)} x_i$$

$$\sigma_j^{(k+1)} = \sqrt{\left(\sum_{i=1}^{n} w_{i,j}^{(t)} \right)^{-1} \sum_{i=1}^{n} w_{i,j}^{(t)} (x_i - \mu_j^{(k+1)})^2}$$

for each $j = 2, 3, \ldots, s$.

Proposition 4.4. *Under assumption 4.1, the EM algorithm converges to a point $\eta^* \in \Gamma_t$, and the point η^* possibly belongs to* $\arg\max_{\eta \in \Gamma_t} l_{c,n}(\eta)$.

Proof. Since for any data set λ_c can be interpreted as a proper density, the theory about EM algorithm presented in Redner and Walker (1984) applies. By Proposition 4.2 the maximum of the sample improper log-likelihood function exists on Γ_t. By Theorem 4.1 in Redner and Walker (1984) the improper log-likelihood function increases at each step. By the the same arguments in the proof of Proposition 4.2 the M-step has a solution at each step for any c. By Theorem 4.1(v) in Redner and Walker (1984) we get the result. ■

§4.5. Selection of the improper density

Hennig (2004) suggested that the choice of the c could be driven by subject matter considerations. He also illustrated some examples. However this is not an easy task. In the next chapter we will show in detail an experimental study where we compare several robust alternatives to the maximum likelihood estimation for Gaussian mixtures. Monte Carlo experiments suggest that the RIMLE has very attractive properties when we are able to give a reasonable choice of the c. However the choice of c is not trivial. Under a variety of data generating processes we discovered that a bad choice of c can make this estimator the worst both in terms of estimated parameters and clustering performances.

To show the importance of the selection of the c we provide here some experimental results. We consider the following data generating process:

$$0.1\mathrm{U}(17,25) + 0.30\mathrm{N}(1,.5) + 0.25\mathrm{N}(7,2) + 0.35\mathrm{N}(14,1.5).$$

As usual N stands for the Gaussian probability model and U for the uniform model. The Gaussian components are parameterized as $\mathrm{N}(\mu, \sigma^2)$ where μ is the mean and σ^2 is the variance. This model is also considered in the next chapter, we shall discuss it in more detail. It consists of a mixture of three reasonably separated normal components plus a uniform noise component with support on the right hand side of the non-noise components. We define the noise as the set of points coming from the uniform distributions. We estimate the RIMLE with ϕ equal to the Gaussian density, and $s = 4$. Computations are done via the EM algorithm previously defined. We take a fine grid of values for $c \in [0, 0.2]$. The grid consists in 500 equidistant points, say c_h, for $h = 1, 2, ..., 500$ (this means that distance between two successive points is about 4×10^{-4}). For each $n = 50, 200, 500$ we do do the following

1. we draw 1000 samples (replicas).

2. For each sample we compute the RIMLE for each c_h, $h = 1, 2, ..., 500$. Here the Gaussian components are ordered by increasing means (if two components result to have the same mean we order them by increasing variance). The RIMLE is used to classify points via the Bayes rule described in Section 4.2. We obtain the percentage of misclassified points (a point is misclassified if it is not assigned to the component that generated it).

3. For each c_h, for $h = 1, 2, ..., 500$ we compute the mean of the misclassification percentage across the 1000 values obtained in the 1000 replicas.

The results of this experiment are reported in Fig. 4.1. Notice that we did not consider values of $c > 0.2$ because these would produce an average misclassification percentage approximately equal to 100%. We notice the following:

- the methodology can lead to a really small average misclassification percentage: for $n = 50$ the minimum is 4% with $c = 0.0151$, for $n = 200$ it is 2.5% with $c = 0.0175$, and for $n = 500$ we get 1.8% with $c = 0.0150$.

- The behavior of the curve of the average misclassification percentages as a function of c seems to follow a certain path. It monotonically decreases

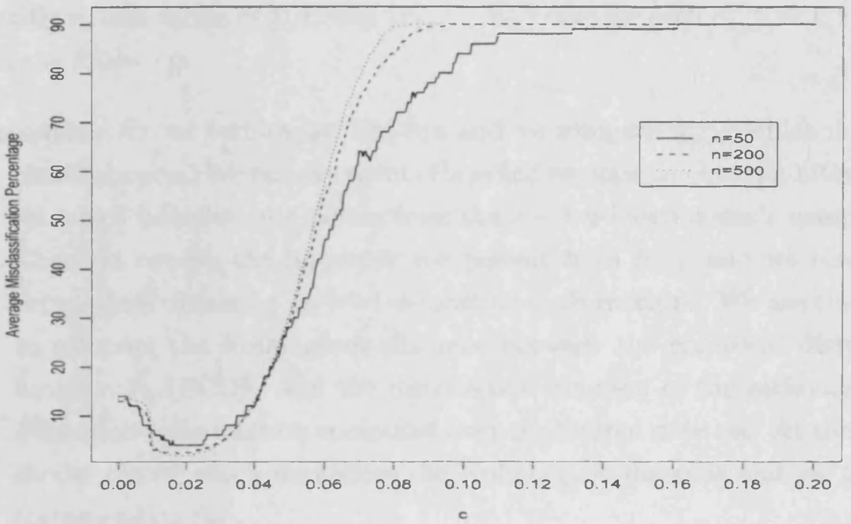

Figure 4.1: Average misclassification percentage vs fixed values of the improper density c for sample size $n = 50, 200, 500$ computed over 1000 repetitions

relatively fast up to a minimum and than monotonically increases relatively fast up to 100% where it stays for large c.

- There exists an interval of values of c, where the average misclassification percentage is small. The curvature of the graph in the region where we get the minimum average misclassification percentage is relatively picked. The latter means that it is easy to end up with a bad choice of c leading to a large misclassification percentage.

- The behaviour described here has been replicated under other data generating processes. In fact, we also explored situations where the noise is generated with distributions other than the uniform, and the non-noise components are not Normals (we shall discuss these models in the next chapter). This does not mean that the arguments given here are of general validity. This is a summary of empirical evidence accumulated by experience with many artificial statistical experiments.

This example, and the large empirical evidence not presented here, convinced us that we need a method to select c based on the data. In particular, we considered different data dependent choices of the c based on a grid of candidates. We describe five alternatives here. For all these alternatives, first we fix a range of possible candidates $[0, \bar{c}]$ (the choice of $\bar{c}$ is discussed afterwards). We take a grid of equally spaced values in $[0, \bar{c}]$, say $\{c_1, \ldots, c_m\}$, and for each c_h, $h = 1, 2, \ldots, m$, we do the following:

Alternative A: we perform an EM run and we compute $\hat{\eta}_{c_h,n}$, which is used to classify points. We remove points classified as noise to obtain a filtered data set which includes only points from the $s-1$ ϕ-location-scale components. Then we remove the improper component from $\hat{\eta}_{c_h,n}$ and we rescale the proportions obtaining an $s-1$ ϕ-location-scale mixture. We use this vector to compute the Kolmogorov distance between the empirical distribution function $\mathbb{F}_n$ (ECDF) and the distribution function of the estimated $s-1$ ϕ-location-scale mixture computed over the filtered data set. At the end we choose the c^* which minimizes the Kolmogorov distance and we take the corresponding $\hat{\eta}_{c^*,n}$.

Alternative B: we perform an EM run and we compute $\hat{\eta}_{c_h,n}$, which is used to classify points. As before, we obtain a filtered data set which includes only points from ϕ-location-scale components. Then we remove the improper component from $\hat{\eta}_{c_h,n}$ and we rescale the proportions obtaining the parameter vector of a proper $s-1$ ϕ-location-scale mixture. We use the latter as initial value to perform a further EM run for an $s-1$ ϕ-location-scale mixture over the filtered data set, we thus obtain the estimate $\dot{\eta}_{s-1,j}$. We compute the Kolmogorov distance between the ECDF and the distribution function of the estimated $s-1$ ϕ-location-scale mixture under $\dot{\eta}_{s-1,h}$ computed over the filtered data set. We choose the c^* which minimizes the Kolmogorov distance and we take the corresponding $\hat{\eta}_{c^*,n}$.

Alternative C: we perform an EM run and we compute $\hat{\eta}_{c_h,n}$. We use this vector to compute the quantity

$$D_h = \max_i |\mathbb{F}_n(x_i) - \Lambda_n(x_i, \hat{\eta}_{c_h,n})|,$$

where:

$$\Lambda_n(x_i, \hat{\eta}_{c_h,n}) = \hat{\pi}_1 c_h (x_i - \min \underline{x_n}) + \sum_{j=2}^{s} \hat{\pi}_j \Phi(x_i; \hat{\mu}_j, \hat{\sigma}_j),$$

and $i = 1, 2, \ldots, n$, $h = 1, 2, \ldots, m$. The quantity $\Lambda_n(x_i, \hat{\eta}_{c_h,n})$ can be interpreted as a pseudo-distribution function associated with the improper density $\lambda_{c_h}(\cdot, \hat{\eta}_{c_h,n})$ computed on x_i. Notice that the latter does not exist when n goes to infinity. We choose the c^* which minimizes the D_h and we take the corresponding $\hat{\eta}_{c^*,n}$.

Alternative D: we perform an EM run and we compute $\hat{\eta}_{c_h,n}$, which is used to classify points. We remove noise points to obtain the filtered data set which includes only points from estimated ϕ-location-scale components. Then we remove the noise component from $\hat{\eta}_{c_h,n}$ and we rescale the proportions obtaining the parameter vector of an $s-1$ ϕ-location-scale mixture. We use the latter as initial value to perform a further EM run for a proper $s-1$ ϕ-location-scale mixture over the filtered data set. We obtain the estimate $\dot{\eta}_{s-1,h}$ and the corresponding sample log-likelihood value $\dot{l}_{s-1,h}$. Thus, we choose the c^* which gives the largest sample log-likelihood value and we take the corresponding $\hat{\eta}_{c^*,n}$.

Alternative E: we perform an EM run and we compute $\hat{\eta}_{c_h,n}$. We use this vector to compute the posterior probability that each observation is not an outlier. Hence we obtain a weighted data set, and based on it we compute the Kolmogorov distance between the corresponding ECDF and the proper distribution function of an $s-1$ ϕ-location-scale mixture. The latter is obtained removing the improper component from $\hat{\eta}_{c_h,n}$ and rescaling the proportion parameters. As usual we choose the c^* which minimizes the Kolmogorov distance and we take the corresponding $\hat{\eta}_{c^*,n}$.

We implemented all these alternatives and we evaluated their performances for several data generating processes. Method B always leads to very attractive performances both in terms of the estimation of η and in terms of classification. The performances of this method seems to not depend on the particular model. Method C performs slightly worse than B, the remaining methods always select a c which is too large or too small, and this often depends on the model. Moreover this methodology is well defined for multidimensional data sets even though the computation of multidimensional distribution function can add some computational complications.

Notice that method B implies a decision about the upper bound of the improper density, that is the $\bar{c}$. However this is not critical. In fact, if $\bar{c}$ is large enough, then we end up with a situation where all data points are classified as noise, and this is not interesting to us. The value $\bar{c}$ could be defined such that the proportion of noise points classified at the beginning of the procedure described under alternative B does not exceed $\alpha\%$. The choice of the α depends on the type of problem we will analyze.

§4.6. Conclusions

Based on the work by Hennig (2004) in this chapter we defined an estimator which is the maximizer of the pseudo-log-likelihood function associated to a mixture of location-scale densities with the addition of a noise component represented by an improper density on the real line (i.e. a constant). We provided the estimation theory showing that, for fixed number of components and fixed value of the improper density, the RIMLE is strongly consistent for the maximizer of the integral of the pseudo-log-density function with respect to the distribution function which generated the data. We advise to use such a method whenever the outliers spread over the entire data-range. The selection of the constant density value is crucial and we also proposed several methodologies to optimally select this value based on the dataset at hand. We advise to use the value of the improper constant density that minimizes the Kolmogorov distance between the distribution function of the estimated model without the noise component and the empirical distribution function; where both distributions are computed over the data not assigned to the noise component. Before we conclude this chapter we want to stress that the robustness theory developed by Hennig (2004) is for fixed value of the improper density. Here we provide a methodology to select the improper density, so that an extension of Hennig's theory to this case is necessary in order to justify this approach. Moreover, convergence theory developed in this chapter also refers to the case when the improper density is fixed. Asymptotic theory for the case when the improper density is estimated by method B could also be developed, but this is not trivial.

CHAPTER 5

Empirical Evidence

In this chapter we present a simulation design to compare six robust alternatives to the maximum likelihood estimator for Gaussian mixtures after which we draw some conclusions.

§5.1. Introduction

In the previous chapters we developed the statistical theory for two robust alternatives for model-based clustering: the maximum likelihood estimator (MLE) for mixtures with uniform noise (see Chapter 3), and the robust improper maximum likelihood estimator (RIMLE, see Chapter 4). In this chapter we compare empirically these methods with Banfield and Raftery's (1993) approach with the uniform noise supported on the data range, McLachlan and Peel's (2000b) t-mixture approach and the standard maximum likelihood estimation for Gaussian mixtures (see Table 5.1). The empirical comparison is made based on simulations of several data generating processes with different features. We compare these methodologies from two points of view: clustering performance and quality of estimates. The empirical results suggest that in the presence of noise the maximum likelihood estimator is seriously affected. Some of the robust alternatives considered in this thesis dramatically improve the standard maximum likelihood method. However, some of these alternatives also should be implemented carefully - particularly when the sample size is not large.

The chapter is organized as follows: in Section 5.2 we describe the estimators and methods being compared; in 5.3 we introduce the data generating processes

Table 5.1: Methods under comparison in the simulation study.

Methods under comparison	**Code**
MLE with uniform noise	G
Banfield and Raftery's approach	R
RIMLE, improper density fixed	IF
RIMLE, improper density optimally selected	IS
McLachlan and Peel's approach, degrees of freedom fixed	TF
McLachlan and Peel's approach, degrees of freedom estimated	TE
MLE for normal mixtures	N

and simulation procedures; in 5.4 we discuss the measures adopted to evaluate the performances of the methods under study; and in 5.6 we provide some conclusions.

§5.2. Estimators and methods

In this section we describe the estimators and methods we propose to compare. The general structure of the simulation study is as follows. First, we consider several data generating processes, and for each we consider a sample size of $n = 50, 200, 500$. For each data generating process and for each sample size we draw 100 samples (replicas), and for each replica we apply seven different estimation methods. For each estimator we perform clustering and compute summary statistics to evaluate relative performance. The data generating processes are described later. Here we describe the estimation methods with particular attention to computational issues. We should stress here that we deal only with the case where the number of mixture components is fixed and known. In particular, each of the estimated models includes s components, one of which is the noise component. The results in the previous chapters are derived mainly for one-dimensional random variables, therefore, in this chapter, we also deal with one-dimensional random variables. The seven methodologies we apply will be identified by the codes: R,G,IF,IS,TF,TE,N.

5.2.1 — Gaussian mixtures with uniform noise (G,R)

In Chapter 3 we introduced the maximum likelihood estimator for uniform-location-scale mixtures. The theory we have developed is rather general and applies to the case where the mixture includes a finite number of uniform components – with disconnected supports – and a finite number of distributions belonging to some location-scale family. The methods described here consist of a mixture of uniform and Gaussian. In particular, we consider the maximum likelihood estimator for a mixture composed of one uniform component (representing noise) and $s-1$ Gaussian components. We do not actually compute the maximum likelihood estimator, but only an approximation. The meaning of the term "approximation" will become clear. We will consider two approaches:

- The methodology proposed by Banfield and Raftery (1993). This consists of fixing the uniform component such that it has support equal to the range of the data. Recall that the resulting solution obtained by applying the EM algorithm eventually provides one of the potentially many local maxima. This methodology provides a maximum likelihood estimator only in some cases, depending on the data generating process (we will say more on this subsequently). This approach will be coded as "R" (meaning "range").
- The second approach is to better approximate the maximum likelihood estimator studied in Chapter 3. This methodology is identified as "G" (meaning "grid").

Implementation of the G-method. As highlighted at the end of Chapter 3, the computational burden introduced by the uniform component makes this estimator difficult to compute. Even with one uniform component the computational complexity is high. To refer back to Section 3.5, if we have only one uniform component, we should initialize the EM algorithm for each possible pair of data points, and for all those pairs of values of the uniform parameters, which are at the border of the constrained set defined in (3.6). This would be unrealistic even for $n = 50$. We could forget the borders of the constrained set defined in (3.6), and initialize only the EM algorithm for each possible pair of data points. But for $n = 50$ this is already a heavy computational load. We adopted a practical solution, which is to consider only pairs of points selected on a grid, on the set of data points. In order to make the computations feasible, for each n we chose a different size of grid of data points. For each replica the methodology was as follows:

1. given the sample size, we define a grid of equi-spaced points on the range of the data. The size of the grid decreases as the sample size increases. For $n = 50$ the grid consists of 20 points, for $n = 200$ the grid consists of 15 points, and for $n = 500$ it consists of 10 points. Then, for each point in the defined grid we take the nearest data point. It can happen

that two points in the selected grid are very close to each other, and the distance between the two points in each pair will determine the variance of the uniform component in the corresponding EM solution. The algorithm described in Section 3.5 makes use of the constraint defined in (3.6). For simplicity we did not implement the numerical routines taking account of such constraints. Instead, we coded the EM algorithm so that the pair of points that define a uniform component with small variance will not be considered. Therefore, one point is eliminated from the grid if the distance from the nearest point is less than 1% of the interquartile range. Thus, for some sample sizes it could be that the grid of data points actually considered has a number of points that is smaller than the prescribed number.

2. We defined the initial values for all other parameters. The initial value of the proportion of the noise (uniform) component is fixed at 0.05. This choice is because in many real life applications it is reasonable to assume that the proportion of noise is small compared to proportions of the other components. The proportions of the other components are initialized at equal value: $0.95/(s-1)$. The variances of the Gaussians are all initialized equal to 1. Of course, in real applications the choice of the variance can be made data dependent (e.g. the choice can be based on the interquartile range), but such a choice should be made depending on the data. The means of the Gaussians are initialized so that given the sample $\underline{x_n}$ the initial means $\mu_1 < \mu_2 <, \ldots, < \mu_{s-1}$ are such that the intervals $[\min(\underline{x_n}), \mu_1]$, $[\mu_{s-1}, \max(\underline{x_n})]$, and $[\mu_j, \mu_{j+1}]$ for all $j = 2, 3, \ldots, s-3$ contain the same proportion of data points. Again, this choice will not always provide reasonable results in real applications.

3. Given the grids of data points selected previously, say $y_1 < y_2 <, \ldots, < y_g$, we define all possible pairs (y_r, y_p) such that $y_r < y_p$. For each of these pairs we run an EM algorithm. In each of these runs the uniform parameters are fixed at equal to the composed pair, while all other parameters are initialized as described above. The resulting procedure is an application of the EM algorithm described in Section 3.5 with $q = 1$ and the number of pairs of data points for the uniform is reduced to a subset of all possible pairs. For variances of the uniform components, we coded the EM algorithm such that if one component reaches a very small variance, i.e. 10^{-3}, the algorithm stops. We checked for whether this ever happened in the simulations and it does not seem to. The EM algorithm stops when either the number of

iterations exceeded 500 or when the difference in the log-likelihood values in two successive iterations is less than or equal to 10^{-6}.

4. For each run of the EM algorithm we compute the resulting value of the log-likelihood function at the point where the EM stopped. We chose the parameter vector that corresponds to the largest log-likelihood value. This means that we estimated the uniform parameters with the corresponding pair of data points for which the EM solution provided the largest log-likelihood value.

5. In Chapter 3 we discussed the identifiability of such a model. We can identify the number of uniform components and their parameters. On the other hand, the Gaussians are identifiable only up to component label switching. This means that in permuting the triples of proportions, means and variances of the Gaussian components, we still have the same estimated distribution. This is relevant when we use clustering and want to compare the performance of this method compared to others. We need to decide which distribution we have estimated in the previous step. To do this we apply lexicographic ordering criteria. Given the triples of Gaussians parameters $\hat{\pi}_j, \hat{\mu}_j, \hat{v}_j$, for $j = 2, 3, \dots, s$, we consider a permutation $\bar{j}$ of the indexes $j = 2, 3, \dots, s$ such that $\hat{\mu}_{\bar{j}} \leq \hat{\mu}_{\bar{j}+1}$, if $\hat{\mu}_{\bar{j}} = \hat{\mu}_{\bar{j}+1}$ then $\hat{v}_{\bar{j}} \leq \hat{v}_{\bar{j}+1}$, and if $\hat{v}_{\bar{j}} = \hat{v}_{\bar{j}+1}$ then $\hat{\pi}_{\bar{j}} \leq \hat{\pi}_{\bar{j}+1}$. Actually. the probability is that $\hat{\mu}_{\bar{j}} = \hat{\mu}_{\bar{j}+1}$ is zero, but this could be due to numerical approximations in the computations. The resulting estimated distribution from the previous step will have the parameter vector

$$(\hat{\pi}_1, \hat{a}, \hat{b}, \hat{\pi}_{\bar{1}}, \hat{\mu}_{\bar{1}}, \hat{v}_{\bar{1}}, \hat{\pi}_{\bar{2}}, \hat{\mu}_{\bar{2}}, \hat{v}_{\bar{2}} \dots \hat{\pi}_{\bar{s}}, \hat{\mu}_{\bar{s}}, \hat{v}_{\bar{s}}),$$

where $(\hat{\pi}_1, \hat{a}, \hat{b})$ are the uniform parameters.

6. The estimated parameter vector is used to classify points via Bayes' rule. The set of noise points is identified with the set of points assigned to the uniform component.

Implementation of the R-method. This is very similar to the method described above . The procedure is the same except that we do not select the grid of points for estimating the uniform parameters. For each replica we run just one EM algorithm and the uniform parameters are estimated as $\min(\underline{x}_n)$ and $\max(\underline{x}_n)$. The initialization of the algorithm, as well as all other details, remain the same as before. This corresponds with what Fraley and Raftery (2002) propose.

5.2.2 — Gaussian mixtures with improper density (IF,IS)

The estimator discussed in this section is the RIMLE as presented in Chapter 5. In this simulation study we consider $s-1$ Gaussian components plus the constant improper density. As discussed in Chapter 5 the choice of c for the improper density value, is critical. We include two approaches: one with the improper density value selected using method B discussed in Section 4.5, which we refer to as "IS" (meaning "improper density selected"); in the other we fix the improper density value, a method that we refer to as "IF" (meaning "improper density fixed"). Notice that with the IS-method the value c depends on the sample, thus it changes in each replica, while with the IF-method, c is fixed for every sample size and replica.

Implementation of the IS-method. For each data generating process and each replica the methodology is as follows:

1. We define the initial values for all parameters. This is done in the way that was used for the G-method. The proportion of the noise (improper) component is fixed at 0.05, and the proportions of the other components are initialized with equal value $0.95/(s-1)$. Means and variances of Gaussian components are initialized in the same way as used for the G-method.

2. The initial vector is used to apply methodology B, described in section 4.5. The EM algorithm is coded so that if one component reaches a variance less than or equal to 10^{-3}, the algorithm stops. The EM algorithm stops either when the number of iterations exceed 500, or when the difference in the log-likelihood values for two successive iterations is less than or equal to 10^{-6}. In none of the simulations did the algorithm reach the limit of 500 iterations.

3. The parameters referring to Gaussian components are ordered using the lexicographic ordering described before. The resulting estimated vector was used to classify points via the Bayes rule. The set of noise points is identified with the set of points assigned to the improper component.

Implementation of the IF-method. Hennig (2005) offers some guidelines for the choice of the constant improper density, on the basis of subject matter considerations. In a simulation study such as ours, we do not have information other than on the drawn sample. We should not take account of knowledge related to the data generating processes in order to adjust tuning constants. Therefore, what value of c should we consider? Under these conditions it is impossible to formulate any reasonable choice for c. The objective of this study is twofold: we want to demonstrate the potential of this methodology compared to the alternatives; and we want to compare the performances of the IS and IF methods. For each data generating process considered, we calibrated the value of c so that the estimator ensures the lowest average misclassification percentage for a sample size of $n = 200$. The rationale for this choice is that we can fix a value of c that allows the best performances of this methodology in terms of clustering, for the mid-sized sample. This merely illustrates the potential of the methodology under study; this procedure is not usable in real situations. In fact, to do this calibration of c requires that the researcher knows the true groups in the data, which is not feasible in reality. Thus, the value of the fixed c and the corresponding performances of the IF method should only be taken as a benchmark. This will tell us how far we can go with the improper density approach once the selection of c becomes optimal.

The calibration was made using the procedure described in Section 4.5. For each of the data generating processes the RIMLE estimator will consist of $s - 1$ Gaussian components plus the constant improper density. We take a fine grid of values for $c \in [0, 0.2]$. The grid consists of 500 equidistant points, say c_h, for $h = 1, 2, ..., 500$ (this means that the distance between two successive points is about 4×10^{-4}). For a sample size $n = 200$ we do the following:

1. we draw 1000 samples.

2. For each sample we compute the RIMLE for each c_h, $h = 1, 2, ..., 500$. using the EM algorithm described in 4.4. Initialization of the means, variances and proportion parameters is done as for the G-method and R-method. In the same way the parameter vector computed by the EM algorithm is ordered using the same lexicographic ordering. This latter was used to assign data points to the components according to the Bayes rule introduced in Section 4.2. We then obtain the percentage of misclassified points.

3. For each c_h, $h = 1, 2, ..., 500$, we compute the mean of the misclassification percentages across the 1000 replicas.

4. We choose the c^* that achieves the lowest misclassification average percentage.

The resulting values of c^* are used as the fixed improper density value in the IF method. Note that the calibration is not performed on the same samples as used in the comparison. Recall that c^* is fixed for each data generating process and it is maintained as equal across replicas and sample sizes. As highlighted in Chapter 4.2 we do not consider the values of $c > 0.2$ because this would result on average in 100% of the points classified as noise.

Given the value c^*, this method is implemented in the same way as the IS-method except for point 2. In fact, in this case we do not have to select the value c depending on the sample, but can simply use the initial vector to start the EM algorithm described in 4.4 with the improper density fixed at c^*.

5.2.3 — t-mixtures (TF,TE)

Another alternative for robust model-based clustering is to use t-mixtures. McLachlan and Peel (2000b) argue that for elliptical shaped clusters with longer than normal tails or atypical observations, the use of normal components may affect the fit of the data. The strategy proposed by the authors is as follows: (i) to model a sample of iid observations as arising from a finite mixture of t-distributions; (ii) to estimate the parameters via the maximum likelihood method, with estimates computed with the EM algorithm; (iii) to use the vector of estimated parameters to assign data points to components and to isolate noisy observations. The definition of noise in McLachlan and Peel (2000b) is provided later (see Section 5.3.3). The approach developed by the authors does not assume that the sample is drawn from a finite t-mixture. They just make use of finite t-mixtures as a fitting device. Here, we describe the methodology and the related computational procedure.

We consider a finite mixture of univariate t-distributions with $s-1$ components. For $x \in \mathbb{R}$ a finite mixture of t-density is defined as

$$t(x; \theta, u) := \sum_{j=1}^{s-1} \pi_j \psi(x; \mu_i, \sigma_i, g_i) \tag{5.1}$$

$u = (g_1, g_2, \ldots, g_{s-1})$ is the vector of degrees of freedom,

$$\theta = (\pi_1, \pi_2, \ldots, \pi_{s-1}, \mu_1, \mu_2, \ldots, \mu_{s-1}, \sigma_1, \sigma_2, \ldots, \sigma_{s-1})$$

is the vector of proportions, locations and scale parameters; $0 < \pi_j < 1$ for all $j = 1, 2, \ldots, s-1$ and $\sum_{i=1}^{s-1} \pi_j = 1$. The function ψ is the density of a t-distribution with g degrees of freedom:

$$\psi(x;\mu,\sigma,g) = \frac{\Gamma\left(\frac{g+1}{2}\right)}{\sqrt{\pi g}\sigma\Gamma\left(\frac{g}{2}\right)\left(1+\frac{\delta(x,\mu,\sigma)}{g}\right)^{\frac{g+1}{2}}}, \tag{5.2}$$

where

$$\delta(x,\mu,\sigma) := \frac{(x-\mu)^2}{\sigma^2}. \tag{5.3}$$

$\delta(x,\mu,\sigma)$ denotes the Mahalanobis squared distance between x and μ. When a random variable X is distributed according to a t-distribution represented by the density in (5.2), if $g > 1$ then $\mathrm{E}[X] = \mu$, and if $g > 2$ then $\mathrm{Var}[X] = g(g-2)^{-1}\sigma^2$. When the degrees of freedom tend to infinity the density ψ tends to a Gaussian density.

In the strategy proposed by McLachlan and Peel (2000b) the iid sample is modelled as arising from a finite t-mixture with a fixed number of components, say $s-1$. Their aim was to estimate the parameters via the maximum likelihood method. Hence, the estimates considered are defined as the maximizer of the log-likelihood function:

$$(\hat{\theta}_n, \hat{u}_n) = \arg\max_{\theta,u} \sum_{i=1}^{n} \log t(x_i; \theta, u),$$

where $x_i, x_2, \ldots, x_n$ is the observed sample. In many situations the degrees of freedom are considered to be fixed, and only proportions, scale and location parameters are estimated. This considerably simplifies the maximization above. If we want the heaviest tails but still want the means and variances of the subpopulations to be defined, we can fix the degrees of freedom to be equal to 3. But how does this affect the other estimated parameters? There is little discussion in the literature about fixed degrees of freedom vs estimated degrees of freedom. However, we implemented this method with both estimated degrees of freedom and fixed degrees of freedom. When we discuss the experimental results, we will see that this makes a difference.

The maximization above can be done by applying the EM algorithm or its ECM variant. The ECM is an EM variant introduced by Liu (1997) for the case where the degrees of freedom have to be estimated. These computational procedures are described in detail in McLachlan and Peel (2000a).

After estimating the $s-1$ t-components the authors propose a further step in which outliers are identified. Once we obtain the estimates $(\hat{\theta}_n, \hat{u}_n)$ – the method is the same as when the degrees of freedom are fixed – we compute the following quantities:

$$\hat{\tau}(i,j) := \frac{\hat{\pi}_j \psi(x_i; \hat{\mu}_j, \hat{\sigma}_j, \hat{g}_j)}{t(x; \hat{\theta}, \hat{u})} \tag{5.4}$$

for each $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, s-1$. The quantity $\hat{\tau}(i,j)$ is the estimated posterior probability that the ith observation belongs to the jth group. The quantities above are used to compute

$$\hat{\tau}^*(i) := \arg \max_{j=1,2,\ldots,s-1} \hat{\tau}(i,j), \quad i = 1, 2, \ldots, n; \tag{5.5}$$

and based on these, the ith observation is assigned to the jth component if $\hat{\tau}^*(i) = j$.

From these $s - 1$ estimated subpopulations noise is identified using the following method. McLachlan and Peel (2000b) considered the statistic

$$C(i) = \sum_{j=1}^{s-1} \mathbf{1}\{\tau^*(i) = j\} \delta(x_i; \hat{\mu}_j, \hat{\sigma}_j, \hat{g}_j) \tag{5.6}$$

If this statistic is sufficiently large the ith observation is classified as noise. The logic of this is that if the point x_i is far away from the nearest location parameter it is considered to be noise. To decide how large the statistic $C(i)$ must be in order to classify the ith observation as noise, the authors proposed to compare $C(i)$ with the 95th quantile of the Chi-square distribution with 1 degree of freedom. Thus, the point x_i is classified as noise if $C(i) \geq 3.841459$. The authors do not offer a clear explanation of why $C(i)$ should be distributed according to a χ^2_1. Once the statistic $C(i)$ is computed for each $i = 1, 2, \ldots, n$ we can form a further group, the sth group, which includes only noisy observations.

Implementation of the TF and TE method. As pointed out already, in this comparison we consider both estimated and fixed degrees of freedom. For the case with the estimated degrees of freedom we implement the ECM as described in Liu (1997). In each run of the EM/ECM, we initialize the proportions and location parameters as in the G method. The scale parameters are set equal to 1 for all components. More attention has been given to degrees of freedom. During our experiments we noted that the ECM algorithm does not move too far from the starting values with respect to degrees of freedom. This is possibly due to the fact that the log-likelihood surface has many local maxima or has flat regions with respect to the degrees of freedom. We could find no research on this topic. A good practice would be to run the ECM several times, with many possible combinations of initial values for the degrees of freedom, and then select the solution that corresponds to the highest log-likelihood value. However, our simulation study is already very complex to allow to accomodate these further computational complications. The majority of the components in our data generating processes are Gaussians. For a t-distribution, in order to approximate a Gaussian distribution, we need at least 30 degrees of freedom. When the degrees of freedom are estimated they are initialized to be equal to 15 for each component. This number provides a starting distribution that is halfway between Gaussian tails and heavier tails. We think this approach is fair. We will comment more on the estimated degrees of freedom in subsequent sections. For the t-mixture with fixed degrees of freedom we consider them equal to 3 for all components. This choice guarantees the heaviest tails for the population components and also that their means and variances are well-defined. For each data generating process the t-mixture estimator is implemented as follows, for each sample size and each repetition:

1. if the number of groups in the data is s (including noise), we initialize an $s-1$ components t-mixture as described previously.

2. We compute the maximum likelihood estimator of a $s-1$ t-mixture, we order the triples of parameters $(\hat{\pi}_j, \hat{\mu}_j, \hat{\sigma}_j)$ by the same lexicographic ordering described above where μ_j and σ_j play the same role as the mean and standard deviations in the Gaussian based methods previously described. The vector of degrees of freedom (fixed or estimated), is ordered accordingly.

3. The ordered estimated parameter vector is used to assign points to the $s-1$ components and the sth noise component is identified as previously described.

4. In order to make comparisons we compare moments of the simulated sub-populations with moments of the estimated sub-populations. For this method it is not possible to use the parameters of the $s-1$ originally estimated components, to estimate the moments of the s components (including the noise). For each of the s components we compute the means, variances, and proportions of points in each of the groups obtained in the previous clustering. This is to enable us to compare the means, variances and proportions of the estimated groups with the true corresponding values.

5.2.4 — Normal mixtures (N)

Many of the data generating processes under consideration will be chosen as a Gaussian mixture with some contamination. In order to assess the gain introduced by the previous robust alternatives we also consider the maximum likelihood estimator for normal mixtures. This will be denoted as "N-method". When applying this method we do not have a noise component, that is, the number of mixture (normal) components is fixed at $s - 1$. The computations are done via the EM algorithm, as described in Chapter 2.

For each data generating process and each replica the methodology is as follows:

1. We define the initial values for all parameters. The proportions of all components are initialized at a value equal to $1/(s-1)$. Means and variances of the Gaussian components are initialized in the same way as for G-method.

2. The initial vector is used to run an EM algorithm. The EM algorithm is coded such that if one of the components reaches a variance less than or equal to 10^{-3}, the algorithm stops. The EM algorithm stops either when the number of iterations exceeds 500 or when the difference in the log-likelihood values in two successive iterations is less than or equal to 10^{-6}.

3. The parameters that refer to Gaussians components are ordered by the same lexicographic ordering as before. The resulting estimated vector is used to classify points via the Bayes rule. By construction, the percentage of points assigned to noise by this method will be zero.

§5.3. Data generating processes

We considered six different data generating processes which are described and analyzed. In any simulation study the goal of the analysis is twofold: to test how good is a procedure against an alternative even under extreme situations; the simulated data should be realistic, by which we mean likely to occur in the real world. This is a complex trade-off and we have tried to optimize it. Throughout the rest of this chapter, $\mathrm{N}(\mu, v)$ is the Gaussian probability model with mean μ and variance v; $U(a, b)$ is the uniform probability model with support on the interval $[a, b]$. For $g > 2$, $T_g(\mu, v)$ is the non-central T-student probability model with g degrees of freedom, location parameter μ and variance v. Note that we parameterized the T-distribution in terms of variance, assuming that $g > 2$; the reason for this will become clear later. Also, the squared location parameter of a $T_g(\mu, v)$ is given by $v(g-2)/g$.

5.3.1 — Side, inside and wide uniform noise

We considered a class of data generating models in which a number of Gaussian components are mixed with a uniform (noise) component. Noise here is defined as those points drawn from the uniform mixture component. We consider three alternatives, each of which differs from the other two in terms of the position of the uniform support relative to the means of the normals. We refer to these models as side-noise, inside-noise, and wide-noise.

Side-noise. The model is

$$0.1\mathrm{U}(17, 25) + 0.30\mathrm{N}(0, 1.5) + 0.25\mathrm{N}(7, 2) + 0.35\mathrm{N}(14, 1.5). \qquad (5.7)$$

The proportion of the noise is 10%. This is the same for almost all the data generating processes considered. This means that with $n = 50$, on average, we have only 5 points from the noise component. Even if in a real situation the noise could be much more than 10%, we are interested in all those situations where the noise consists of a relatively small proportion of the observed data. In fact we want to assess whether this method can be used to identify noise even when the expected number of noise points is as low as 5 (which is the case when $n = 50$ and the noise proportion is 10%). The noise produced in this model is located on the right of the mean of the largest normal. Fig. 5.1 at the end of this chapter, is a histogram of the sample of 200 points drawn from this model. We also report the density associated with the model computed over the sample points. The uniform noise is located on the right of the histogram making the right tail of the distribution much heavier than the left one. The density clearly has a discontinuity around 17, which is the lower limit of the support of the uniform component.

Note that the Gaussian components are reasonably separated (this can be seen in Fig. 5.1), they have relatively small variances, and their proportions do not deviate much from 0.9/3, which would be as if there were equal proportions for all the Gaussians. Hosmer (1978) showed that when the number of normal components is larger than two and the separation between components is small, the solution provided by the EM algorithm is usually a poor approximation for the maximum likelihood estimate. In our experiments we noted also that this happens particularly when the variances in the underlying normals are relatively large and the proportions deviate considerably from equality. This latter effect was documented by Karlis and Xekalaki (2003). No explanation is provided. However, these problems are not the focus of this research and we want to isolate these phenomena from our study. Hence, the choice of well separated Gaussian components, with relatively small variances and not very dissimilar proportions. For the sake of comparability of results, this choice is maintained for all subsequent models.

Inside-noise. The model is

$$0.1\mathrm{U}(11,19) + 0.30\mathrm{N}(0,1.5) + 0.25\mathrm{N}(7,1.5) + 0.35\mathrm{N}(21,2). \tag{5.8}$$

The model is similar to the previous one with the exception that the uniform noise is now located in the region between the tails of two normals. At the end of this Chapter we show the sample of 200 points drawn from this model together with the density computed over the sample points in (Fig. 5.2).

Wide-noise. The model is as follows:

$$0.1\mathrm{U}(0,21) + 0.45\mathrm{N}(7,2) + 0.45\mathrm{N}(14,1.5). \tag{5.9}$$

Here, the uniform noise spreads over the entire range of the data. The histogram of the sample of 200 points drawn from this model, and the related density computed over the sample points in Fig. 5.3 can be found at the end of the chapter. The R-method is expected to make sense in this situation.

5.3.2 — Outlier process

This model consists of a two-normal mixture plus two extreme outliers, drawn from a uniform distribution in each replica. The outlier process is as follows: in each repetition of the simulation study for sample size equal to n we draw a sample of $n-2$ points from the mixture model

$$0.5\mathrm{N}(0,2) + 0.5\mathrm{N}(5,1.2), \tag{5.10}$$

and then add two outliers from U(20, 25). Of course, this is not precisely a simulation of a mixture because the number of points drawn from the uniform are fixed at two in any replica. The expected number of points from each of the normal components in any repetition is $0.5(n-1)/n$, while the proportion of the noise/outliers points is $2/n$ in all replicas.

We need to explain the term outliers as used in this contest. In general a point is seen as being an outlier if it is far away from the bulk of the data. But how far away must it be to be considered an outlier? Here, the aim is not to give a definition of an outlier; however, it is important to provide some justification for our choice of the outlier generator. Our outlier generator generates points uniformly distributed across the interval $[20, 25]$. With a probability approaching 1, we expect that points less than or equal to 20 will not be generated from the model (5.10). Thus, we categorize the points from $\mathrm{U}(20, 25)$ outliers. In Fig. 5.4, at the end of the chapter, we present a histogram of a sample of 200 points drawn using this process.

5.3.3 — t-noise

Here the noise component is extracted by identifying as noise those points in the tails of the t-components. McLachlan and Peel (2000b), in their paper apply this methodology in a multidimensional setup. In particular, the authors present a data set generated by a mixture of two normals plus a small proportion of noise points generated by a uniform distribution. The example has two dimensions. The uniform noise has support over the entire convex hull of the data. This would correspond in one dimension to the wide-noise data generating process proposed here. Based on this data set the authors apply TE method, but with all the normal components having the same scale matrix and degrees of freedom[1]. The authors claim that this methodology provides good results even though these comments are based on just one artificial data set.

All the data generating processes previously defined are based on Gaussians plus uniform noise. The G, R, IF,and IS methods are all based on estimating a Gaussian mixture plus some noise. However, we also wanted to assess the G, R, IF, IS and N methods with the data generating process based on a finite mixture of t-distributions. In particular, we wanted to consider a data generating process based on a t-mixture where the noise is produced exactly as defined by method TF and TE. In other words, we generate data from a t-mixture and then define the set of noise points as those points belonging to the tails of the t-components. To see how this works, let us go back to the density in (5.1). Suppose we draw an artificial sample from a distribution with the density (5.1). We consider the

[1] This means that the degrees of freedom are estimated under the constraint that they are equal for all components.

following model:

$$0.4\mathrm{T}_3(0,2) + 0.3\mathrm{T}_{10}(6,2) + 0.3\mathrm{T}_{10}(12,1). \tag{5.11}$$

Suppose that the sample is $x_{1,j_1}, x_{2,j_2}, \ldots, x_{n,j_n}$, where $j_i \in \{1, 2, \ldots, s-1\}$, x_{i,j_i} is the ith observation and is generated by the j_ith component. For each observation we compute the quantity $\delta_i = \delta(x_{i,j_i}; \mu_j, \sigma_j, \nu_j)$. Assuming that we can approximate the distribution of δ_i with a Chi-square with 1 degree of freedom, a point x_{i,j_i} will be classified as noise if δ_i is larger than or equal to the 95th percentile of the Chi-square distribution with 1 degree of freedom (which is equal to 3.841459). Finally, we compute the proportions, means and variances of the "true" groups including the noise, to compare with the other methods. The choice of this particular model is founded on the fact that: (i) we wanted a data generating process that produced samples, which, from inspection of the histograms and other tools, appeared similar to the samples produced by some of the previous processes; (ii) based on the arguments put forward previously. we also wanted separation between the mixture components; (iii) we built a process which over 500 replicas produced noise components with an average proportion equal to 10.7%, which is near the expected proportion of noise points considered for some of the other models under study.

5.3.4 — Normal mixture

We want to investigate the behaviour of methods G, R, IF, IS, TF and TE in situations when no noise is present in the data. This is important, to assess whether these methodologies are able to discriminate between presence and absence of noise in the data. Under this data generating process a good method should produce a near-to-zero per cent estimate for the proportion of noise. To this end, we consider the three normals mixture model:

$$0.4\mathrm{N}(0,2) + 0.3\mathrm{N}(6,2) + 0.3\mathrm{N}(12,1) \tag{5.12}$$

Fig. 5.6 provides a representation of the histogram and the density.

§5.4. Evaluation of performances

For each replica we computed the estimated parameters and misclassification percentages. At the end of this process we stored all the information required to compute quantities of interest to make comparative judgements about alternative procedures. The comparisons are based on two aspects: estimated parameters vs true parameters, and misclassification percentages. We deal with these two aspects separately. Each of the data generating processes will generate two tables of output with quantities related to the evaluation of the two aspects.

Misclassification percentage. The main interest in this thesis is model-based clustering analysis. In this context we need to know how many points are misclassified and by which methods. For each replica we compared the true clusters with the estimated clusters. We simulated a data set for each data point recording the component to which it belongs. When the estimated clusters were available we compared the estimated clusters with the true clusters and computed the percentage of points wrongly assigned to each component and the (global) percentage of points wrongly assigned by the method under consideration. Note that both the true parameter vectors and the estimated parameters are ordered using the lexicographic ordering previously defined. For each sample size we averaged these percentages over the 100 replica.

Evaluation of estimated parameters. Recall that in most of the situations under consideration the data generating process does not coincide with the model estimated. For example, consider the side-noise data generating process. When we consider the application of the R and G methods, then the density representing the data generating process and the estimated density belong to the same family and are both indexed over the same parameter space. However, if we consider the TF method, this is not the case. In order to make comparisons we compared the true means, variances, and proportions with the estimated proportions, means and variances. This methodology is well defined because for all the data generating processes considered the means and variances of all the components are well defined.

The evaluation is based on the L_1 distance for vectors of classes of parameters. We recall that if $x, y \in \mathbb{R}^p$, the L_1 distance between x and y is defined as

$$d(x, y) = \sum_{i=1}^{p} |x_i - y_i|.$$

Moreover the evaluation is based on classes of parameters rather than single parameters or the entire parameter vector. The reasons for this are as follows:

- if we consider summary statistics (e.g. means over replicas) of some measure of the distance (e.g. the absolute value difference) between the true parameters and moments and their estimates, in most cases this would involve too many numbers to keep track of.

- On the other hand, we could consider summary statistics (e.g. means over replicas) of the distances (e.g. the L_1 distance) between the true and the estimated parameter vectors. This approach has the drawback that it homogenizes the differences in the different types of parameters in the mixture. To explain this point better, let us consider an example. We assume that for some component j the true proportion is $\pi_j^0 = 0.1$ while the true variance is $v_j^0 = 5$; method A provides the following estimates $\pi_j^A = 0.2$ and $v_j^B = 5$, while method B provides $\pi_j^B = 0.1$ and $v_j^B = 5.1$. Suppose we consider the L_1 distance between the estimated parameter vector and the true parameter vector. We can easily see that the contribution given by the variance and the proportion to the distance of the true and estimated parameter vectors for the jth component are the same for both methods. That is, these contributions are $|\pi_j^0 - \pi_j^A| + |v_j^0 - v_j^A| = |\pi_j^0 - \pi_j^b| + |v_j^0 - v_j^b|$ $= 0.1$. However, the effect on the mixture distribution caused by a change of 0.1 in the proportion of some component[2], is much bigger than the effect of a change of 0.1 in some of the variance. This is because here we are comparing proportions, means and variances which have different domains.

Based on this consideration we made comparisons for classes of parameters or moments. Let us assume that the data generating process consists of s components, including the noise component. From now onward, the noise component is first component. Let $\pi_1^0, \pi_2^0, \dots, \pi_s^0$ be the true proportion parameters, $\mu_2^0, \dots, \mu_s^0$ the true means and $v_2^0, \dots, v_s^0$ the true variances of the non-noise components. Suppose, for instance, that an estimation method A produces the following estimates $\pi_1^A, \pi_2^A, \dots, \pi_s^A$, $\mu_2^B, \dots, \mu_s^B$ and $v_2^A, \dots, v_s^A$. For each replica we consider the L_1 distances of the three different classes of vectors: proportions, means and variances. That is, for each replica we compute

$$d_\pi = \sum_{j=1}^{s} |\pi_j^0 - \pi_j^A|, \qquad d_\mu = \sum_{j=2}^{s} |\mu_j^0 - \mu_j^A|, \qquad d_v = \sum_{j=2}^{s} |v_j^0 - v_j^A|.$$

Note that d_μ and d_v do not contain means and variances for the noise component. This is because not all the methods studied estimate a noise component with a well defined probability distribution with a mean and a variance, and also that in

[2] This means that there is a total change of -0.1 in the proportion of the other components

many applications noise parameters are not interesting. For each sample size the evaluation is based on the 90%-upper-trimmed means (over the 100 replicas) for the three distances above. The 90%-upper-trimmed means are the means that are computed considering only value less than or equal to the 90% quantile. This is because in some replicas the EM algorithm solution can be strongly dependent on the initial values, and in some situations this could cause anomalous values in the above distances. Of course, it would not be reasonable to trim the average distances from below because they are bounded below by zero. Other details in the design of this simulation study are highlighted later.

§5.5. Results

In this section we describe how the methodologies under study perform for each of the data generating processes. There are two tables related to each data generating process: the first reports the average misclassification percentages, and the second reports the upper-trimmed means of the distances for the three classes of parameters described above. In our simulation study we stored information about Monte Carlo expected values for each estimate and the standard errors. For the sake of brevity, we do not report all these numbers although we will comment on some interesting aspects.

5.5.1 — Side-noise and inside-noise

In Tables 5.2 and 5.3 we report the results for the side-noise model. This data generating process consists of one uniform noise component plus three Gaussian components (components 2 to 4). We recall that in this case the G method is only an approximation of the maximum likelihood estimator because this methodology only looks for one of the many local maxima. Also recall that the N method, which is the maximum likelihood for a three normals mixture, does not estimate noise. First, we discuss performances in terms of clustering. Table 5.2 reports the average (across replicas) misclassification percentages. We report the global average percentage, which is the average percentage of points wrongly assigned to the four components. We also report the individual average misclassification percentages. These are average (across replicas) proportions of points wrongly assigned to a certain component. This table structure, reporting clustering performance, is used for all the models considered.

Table 5.2 shows that the best overall performance is achieved by the IS method. Its global average misclassification percentages are 13.8% for n=50, 4.1% for

$n = 200$ and 2.17% for $n = 500$. For $n = 200$ the IF achieves a global average misclassification rate even lower than that of the IS method, but we should remember that we calibrated the fixed c in order to get the best performances with the medium average sample size of $n = 200$. In Table 5.3, the c column reports the average c computed across the replica for the IS method. We also report the c fixed for the IF method, which is constant over each replica for each sample size. We note that the average c selected by the IS method is close to the benchmark c obtained via the calibration for sample size $n = 200, 500$, while for $n = 50$ this is not the case. However, for $n = 50$ the IF method vs IS method performance is fairly similar in terms of global misclassification. The G method comes close to the IS and IF methods only for $n = 500$. This means that we need a certain number of noise points to get a pair of points that reasonably approximate the uniform support. In fact, if we look at the simulation outputs we find that the Monte Carlo expected values for the lower limits of the uniform distribution are 9.97 and 9.98 for $n = 50$ and $n = 200$ respectively, which are far away from the true value (17). The G, on the other hand, does not perform very well for all the sample sizes. The same holds for the TF and TE methods. The overall worst performances resulted from using the N method.

Table 5.3 also provides some important insights. First, performance in terms of clustering reflects the quality of the estimates. By inspecting the upper-trimmed distances of classes of parameters we see that the relative performance of each of the five methods is also confirmed for quality of estimates. One important point that emerges from this table is that for the TF and TE method it seems that poor performance is related to poor performance in terms of variance estimation. This is also the case for some of the subsequent data generating processes. If we look at Table 5.3 we can see that while the behaviour of the estimated proportions and locations does not differ very much from those provided by the other methods, the variances are estimated with a rather large error. We can also see that this phenomenon is stronger when the degrees of freedom are fixed. We recall that in a t-mixture component with degrees of freedom greater than 2, the estimated variance depends on both the degrees of freedom and the squared scale parameter. The problem may be that the EM/ECM algorithm does not provide a good approximation for the maximum likelihood estimates of these parameters. However, this will be evaluated in the case of the t-noise models. In fact, in that case we have a correspondence between the data generating process and the estimated model, and this kind of consideration will be easier. There is another issue related to t-mixtures: the standard errors of the estimated degrees of freedom are huge (when compared with those for the other parameters) for all sample sizes. This is the same for all the data generating processes.

Tables 5.4 and 5.5 show that the behaviour of the inside-noise data generating processes is the same as in the case of side-noise, and the same comments apply. This suggests that wherever the support of uniform noise is placed, the estimators perform analogously.

5.5.2 — Wide-noise

Tables 5.6 and 5.7 refer to the wide-noise model. The R method here, could provide a maximum likelihood estimator as described in previous chapters. In fact, this method always fixes the uniform distribution as having support for the entire data-range. The G method also provides a maximum likelihood estimator when the uniform component takes a pair of points equal to the minimum and maximum of the observations. Of course, since the computations are made via the EM algorithm there is no guarantee that the estimates coincide with the maximum likelihood estimate. If we look at Table 5.6 we can see that for $n = 500$ the performance of the R, IF, IS and TE methods is very similar in terms of the global clustering performance. However, the improper density approach is once again a winner because, for smaller sample sizes, it outperforms the G and all other approaches. However, this result does not imply that RIMLE is better than maximum likelihood. In fact, it should be remembered that unfortunately there is no guarantee that an EM run provides the actual maximum likelihood estimator. We also note that as the sample size increases the G method and the R method get very close to IF and IS. This is because as the sample size increases, the number of noise points in the data set increases and there is a better chance that the G method estimates a uniform component with support over the range of the data. Looking at the output from the simulations we can see that for $n = 500$ in 98% of the replica the support of the uniform component estimated by the G method coincides with range of the data. Hence, in all these cases the G and R methods will yield the same answer.

When the sample size is small ($n = 50$), the G method performs rather badly. In this case, we expect only five noise points in the data set. These are expected to be spread across the entire range of the data. Some of these noise points will be produced in the region where the normal components have the majority of their probability masses. This makes it very difficult to select a pair of data points approximating the uniform noise using this method. With side-noise and inside-noise this difficulty does not emerge, probably because in these cases the uniform points spread out in regions where the normals have limited amounts of probability mass. Relatively speaking, the TE method performs better than the TF approach. Looking at the global average misclassification percentages, we can see that as the sample size increases all the methods tend to converge in terms of performance, with the exception of TF, TE and N. The composition of the global average misclassification percentage is also informative. From the column headed"noise" we can see that overall the IS and IF methods give the lowest average proportion of points originated by non-noise components and assigned to the noise component. This phenomenon is quite strong for small sample sizes. The two methods based on t-mixtures always produce the largest average proportion of points wrongly assigned to the noise component, even for $n = 500$.

Let us now consider Table 5.7. As before performance in terms of clustering is related to performance in terms of estimates. As in the cases studied previously, the optimal c computed in the IS method is a long way below the fixed c values (i.e. 0.043) for small sample sizes.

5.5.3 — The outlier process

In the outlier process the number of noise points is fixed for all the replicas. This means that the true proportion of noise decreases as the sample size increases. That is, the true proportion of noise is 4% for $n = 50$, 1% for $n = 200$ and 0.4% when $n = 500$. If we look at Table 5.8 we can see that the N method would be seriously affected by the presence of extreme outliers. The second worst method is again the TF followed by the TE. It is hard to identify a winner in this case. Despite its behaviour for small samples, the G method achieves the highest global average misclassification percentage for $n = 500$. Moreover, the G method performs rather badly in this situation when the sample size is small. The outliers are placed at the right of the second normal components with a mean equal to 5. The minimum value expected for the outlier is 20, with 25 the maximum expected value. When $n = 50$, placing the uniform support such that it includes points at the right-most normal pays off in terms of the likelihood value. This is demonstrated by the fact that the large average global misclassification percentage is mainly due to points wrongly assigned to the noise component. These points mostly come from the normal distribution having the largest mean. The R method exhibits a global average misclassification percentage, which is near to the optimum for all n. The IF method produces the best numbers in terms of clustering, however it is not so far from the R method. The IS does not perform well for small sample sizes, but for larger n its performance is close to the best. From Table 5.9 we can see that the IS tends to select a c which, on average, is much larger than the fixed. It is interesting that for $n = 50$ the difference between the fixed c and the average selected c is around 0.001, but the difference between the average global misclassification percentages produced by the IF and IS method is about 2.12%. This latter point should convince us that small variations in c can produce fairly wide differences in the final results. This again confirms that a data-driven method to optimally fix the c is needed. From Table 5.9 we note again that with the exception of the N and TF methods, means are estimated with similar precision . On the other hand, we observe more variations in the estimations of variances and proportions.

5.5.4 — t-noise

In this data generating process noise is defined as points arising from the tails of the t-distributions. The model simulated produces on average about 10% of noise (as the processes analysed so far). This time the TF and TE methods are estimators that identify noise as it is defined. As we would expect, these methods are the best performers. Let us look at the Table 5.10. The TE achieves the best global average misclassification percentage for all sample sizes and the TF the second best. From inspection of the stored simulation outputs we can see that while the true degrees of freedom for the three t-components where $(3, 10, 10)$,the Monte Carlo expected values for their estimates are $(99, 88, 114)$ for $n = 50$; (94, 76, 94) for $n = 200$; $(92, 71, 102)$ for $n = 500$. These estimates also have a rather large Monte Carlo standard error. Hence, the EM/ECM algorithm solution does not provide a good answer in terms of estimated degrees of freedom. This is an interesting point because we are not aware of any empirical analyses of performance for the maximum likelihood estimator for t-mixtures. The variance of the t-component is determined by the squared scale parameters multiplied by a term larger than one which depends on the degrees of freedom (see previous sections). If we look at Table 5.11 we see that the TE and TF provide good estimates of the population variances. The solution provided by the EM/ECM algorithm is such that the scale parameters are underestimated and the degrees of freedom are overestimated. The two kinds of bias balance out so that the variability of each component is reasonably matched. This can also happen because of the many local maxima of the likelihood surface. If the variability of the mixture component is fitted, the bias for the degrees of freedom will seriously affect the kurtosis characteristic of the estimated subpopulations distributions. None of the literature documents these anomalies. Future empirical investigations should explore the computational performances of the EM/ECM algorithm.

Referring back to Table 5.10, we can see that method N produces the worst global average misclassification percentage. Method G is worse than N for $n = 50$, it improves for increasing n to match the performance of the R method for $n = 500$. The IF method provides a better answer than the IS, even though both are median performers overall. For a sample size $n = 50$ IS produces the nearest average global misclassification percentage to that of the TE and TF methods. On the other hand, for larger sample sizes the IF method produces a result closest to that for the TE and TF methods. In this case, the IS method has the problem that it selects a c that is too small compared to the fixed one (see Table 5.11). Note that the IS always produces the lowest average percentage of points wrongly assigned to the noise component.

5.5.5 — Normal mixtures

The normal mixture model is introduced to check what happens when there is no noise in the data. If we look at Table 5.12 it is obvious that the best performance is achieved by the N method, which is the maximum likelihood method. What is of interest is whether the IF method produces almost the same performances as the N methods. We can see that the performance of the IF method is close to N and not only for the calibrating sample size $n = 200$. The IS method is also very close to the N method, even though its performance deteriorates moderately as the sample size increases. This is because as n grows we have more and more points from the tails of the normal sub-populations, and many of these points will be assigned to the noise component, which will have a larger improper density value compared with that of the normals in their tails. In fact, if we examine the average misclassification percentages by components, we can see that the greater part of the contribution to the global average percentage comes from points wrongly assigned to the noise component. The G and R methods perform rather badly for small sample sizes, and even for larger sample sizes their performance does not come close to that of the N method. The TF method shows the worst performance. The TE method does better than the TF method even though for large sample sizes its performance is a long way from that of the N method. As we can see from the Table 5.12 the problem with the two methods based on mixtures of t-distributions, is that they wrongly assign too many points to the noise component. This is because these methods assign points to the noise component under the tails of the t-distributions. From Table 5.13 we can see that the selected c is larger than the fixed c for all $n > 50$, while for $n = 50$ it is approximately equal to zero. Again, this is because as n increases there are more points originating from the tails of the normals, which are caught by the improper noise component. From Table 5.13 it can be seen that the ranking of methods in terms of clustering performance is confirmed by the ranking in terms of estimates. As in the previously analysed cases, it seems that most of the variation can be found in the proportions and variances estimates, while the estimation of means among the methods, seems to be rather homogeneous except for small sample sizes.

From an inspection of the outputs of the simulations we can see that if we apply the TE method, the estimation of the degrees of freedom is reasonably accurate. We know from distribution theory that a t-distribution with larger degrees of freedom approximates a standard Gaussian distribution (in practice, degrees of freedom bigger than 30 will provide a good approximation). We report that the Monte Carlo expected values for the estimated degrees of freedom are always bigger than 70 for all n and for all components. However, the Monte Carlo standard errors for the estimated degrees of freedom and scale parameters are large when compared with the Monte Carlo standard errors for the proportions and locations parameters.

This data generating process raises another important issue. If we want to account for the possibility that there is noise in our data set we want to rely on a method that estimates the proportion of the noise at zero when there is no noise in the data. If we look at the Monte Carlo expected values for the noise proportion, we see that only the IS method estimates on average 0% (at the third decimal point) for all sample sizes. The IF method on average estimates a noise proportion of 3%, 4% and 3% for $n = 50, 200, 500$ respectively; while the TE method estimates a noise proportion on average of 5%, 5% and 4% for $n = 50, 200, 500$ respectively. All the other methods estimate an average noise proportion of 10% or more, even for $n = 500$.

§5.6. Conclusions

In this chapter we developed an extensive simulation Monte Carlo experiment with which we explored empirically the performance of several robust methodologies for model-based cluster analysis. We compared the two estimators we have developed in Chapters 3 and 4 with Banfield and Raftery's (1993) approach with the uniform noise supported on the data range, McLachlan and Peel's (2000b) t-mixture approach and the standard maximum likelihood estimator for Gaussian mixtures (see Table 5.1). From the empirical results it emerges that the performance of the methods under study depends on the underlying data generating process. However, we can say that the RIMLE method with the constant improper density optimally selected is a good safeguard in the sense that it performs rather good in many different situations. In many situations is is either our first best or the second best. When the RIMLE is the second best it results to be very close to the first best.

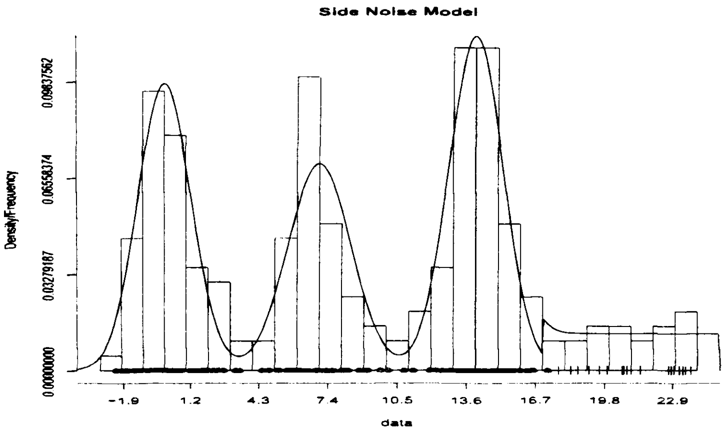

Figure 5.1: Histogram for a sample of 200 points drawn from the side-noise probability model (5.7). The black line represents the density function under the model. Circles on the bottom represent the non-noise points in data set, while strokes represent noise points.

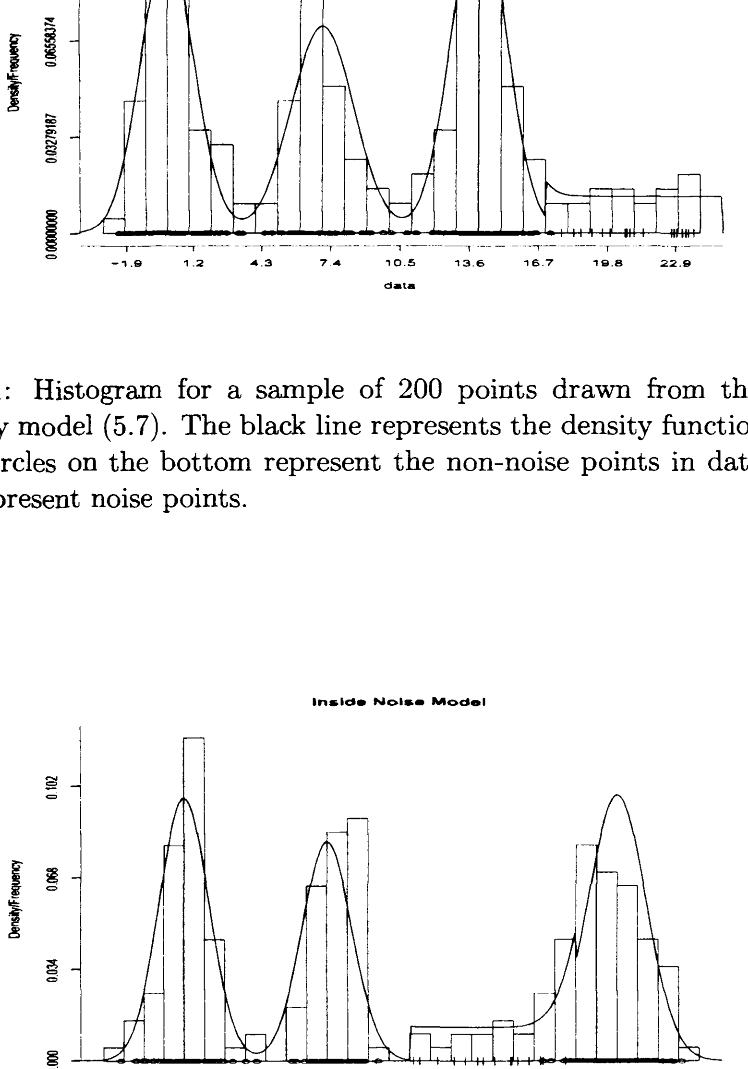

Figure 5.2: Histogram for a sample of 200 points drawn from the inside-noise probability model (5.8). The black line represents the density function under the model. Circles on the bottom represent the non-noise points in data set, while strokes represent noise points.

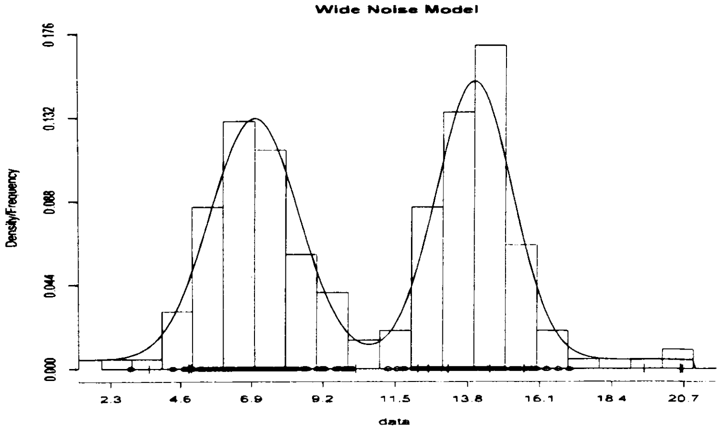

Figure 5.3: Histogram for a sample of 200 points drawn from the wide-noise probability model (5.9). The black line represents the density function under the model. Circles on the bottom represent the non-noise points in data set, while strokes represent noise points.

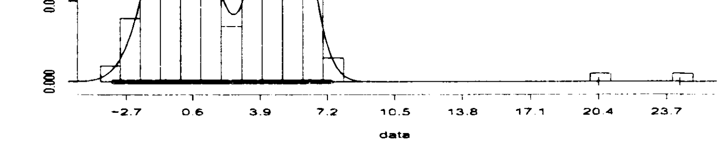

Figure 5.4: Histogram for a sample of 200 points drawn from the outliers process. The black line represents the density function under the model (5.10). Notice that this is the density of the two normal mixtures in (5.10). Circles on the bottom represent the non-noise points in data set, while strokes represent outliers.

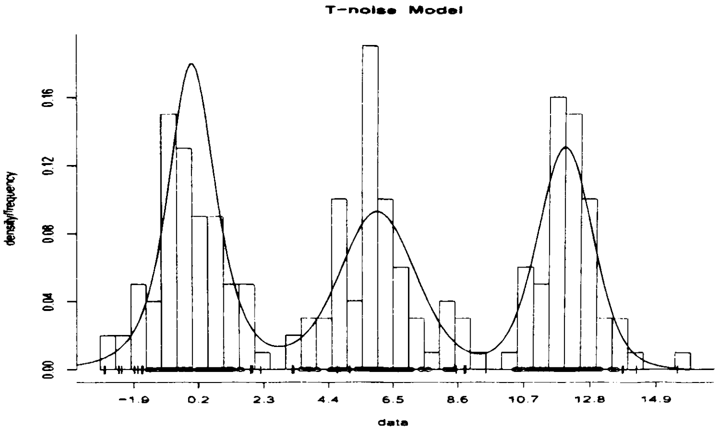

Figure 5.5: Histogram for a sample of 200 points drawn from the T-noise model (5.11). The black line represents the density function under the model. Circles on the bottom represent the non-noise points in data set, while strokes represent noise points.

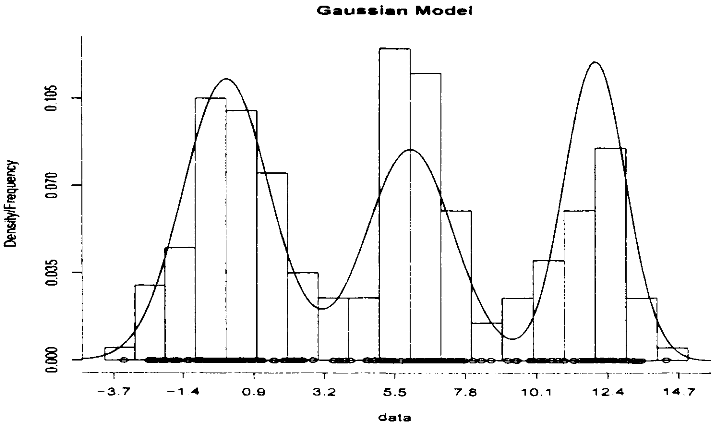

Figure 5.6: Histogram for a sample of 200 points drawn from the Gaussian mixture model (5.12). The black line represents the density function under the model. Single points are reported on the bottom.

Table 5.2: Average misclassification percentages for the "Side noise model". The components' average misclassification percentage is the average percentage of points wrongly assigned to that component. Percentages are computed with respect to n.

N	Method	Global	Component			
			Noise	2	3	4
50	G	20.760	12.440	0.700	5.180	2.440
	R	19.460	14.320	0.620	4.020	0.500
	IF	13.880	3.160	0.680	6.460	3.580
	IS	13.800	3.000	1.140	6.720	2.940
	TE	19.340	3.020	1.920	6.980	7.420
	TF	19.520	6.280	1.940	5.920	5.380
	N	22.300	—	2.060	7.480	12.76
200	G	5.060	4.650	0.100	0.170	0.140
	R	10.390	10.050	0.010	0.310	0.010
	IF	3.790	1.210	0.100	0.460	2.040
	IS	4.080	2.560	0.060	0.770	0.700
	TE	8.450	3.460	0.020	0.010	4.960
	TF	10.650	6.210	0.010	0.950	3.480
	N	13.790	—	0.180	0.460	13.14
500	G	2.690	2.240	0.120	0.200	0.130
	R	8.080	8.070	0.000	0.010	0.010
	IF	2.400	0.860	0.080	0.120	1.340
	IS	2.170	1.480	0.080	0.130	0.480
	TE	7.290	3.240	0.010	0.020	4.020
	TF	9.460	6.140	0.000	0.000	3.310
	N	13.290	—	0.190	0.100	13.00

Table 5.3: Upper-trimmed means of distances for classes of parameters for the"Side noise model". For the methods IF and IS we also report the average value for the improper density c.

N	Method	c	d_π	d_μ	d_v
50	G	—	0.32	1.45	2.45
	R	—	0.46	1.03	2.60
	IF	0.020	0.23	1.15	2.41
	IS	0.012	0.24	1.02	2.09
	TE	—	0.20	1.42	5.48
	TF	—	0.19	1.25	4.61
	N	—	0.34	1.80	13.99
200	G	—	0.17	0.38	1.13
	R	—	0.39	0.40	1.75
	IF	0.020	0.13	0.43	1.16
	IS	0.018	0.16	0.39	1.13
	TE	—	0.11	0.53	2.55
	TF	—	0.13	0.50	2.68
	N	—	0.31	1.02	14.22
500	G	—	0.09	0.23	0.57
	R	—	0.37	0.26	1.48
	IF	0.020	0.11	0.28	0.72
	IS	0.019	0.12	0.25	0.69
	TE	—	0.09	0.41	2.13
	TF	—	0.13	0.39	2.46
	N	—	0.31	1.03	14.31

Table 5.4: Average misclassification percentages for the "Inside noise model". The components' average misclassification percentage is the average percentage of points wrongly assigned to that component. Percentages are computed with respect to n.

N	Method	Global	Component			
			Noise	2	3	4
50	G	21.38	16.74	0.38	2.00	2.26
	R	15.90	12.60	0.62	1.26	1.42
	IF	11.40	3.60	0.72	3.74	3.34
	IS	10.40	2.94	0.98	3.84	2.64
	TE	13.28	3.22	1.70	5.38	2.98
	TF	14.04	6.36	1.66	3.54	2.48
	N	13.62	—	1.78	7.30	4.54
200	G	7.33	5.82	0.07	0.17	1.27
	R	8.87	7.51	0.01	0.01	1.35
	IF	6.54	1.19	0.04	0.85	4.46
	IS	5.76	3.46	0.02	0.03	2.25
	TE	8.54	3.44	0.01	2.02	3.08
	TF	8.88	6.05	0.00	0.44	2.39
	N	10.83	—	0.03	5.03	5.78
500	G	5.68	3.96	0.06	0.30	1.35
	R	7.19	5.72	0.00	0.00	1.46
	IF	5.35	0.88	0.03	0.08	4.35
	IS	4.88	2.53	0.01	0.02	2.32
	TE	7.03	3.19	0.00	0.86	2.98
	TF	8.00	5.60	0.00	0.06	2.33
	N	10.62	—	0.02	4.52	6.08

Table 5.5: Upper-trimmed means of classes of parameters for the"Inside noise model". For the methods IF and IS we also report the average value for the improper density c.

N	Method	c	d_π	d_μ	d_v
50	G	—	0.39	1.77	2.31
	R	—	0.41	1.04	2.29
	IF	0.024	0.23	1.18	2.39
	IS	0.013	0.22	1.11	2.40
	TE	—	0.17	1.44	4.66
	TF	—	0.16	1.21	3.07
	N	—	0.27	1.71	7.82
200	G	—	0.21	0.44	1.09
	R	—	0.32	0.46	1.42
	IF	0.024	0.13	0.66	1.55
	IS	0.026	0.17	0.50	1.21
	TE	—	0.10	0.68	2.02
	TF	—	0.11	0.54	1.50
	N	—	0.23	1.36	7.57
500	G	—	0.16	0.31	0.79
	R	—	0.28	0.33	1.06
	IF	0.024	0.10	0.48	1.19
	IS	0.027	0.15	0.38	0.93
	TE	—	0.07	0.45	1.24
	TF	—	0.10	0.38	1.18
	N	—	0.23	1.23	7.42

Table 5.6: Average misclassification percentages for the "Wide noise model". The components' average misclassification percentage is the average percentage of points wrongly assigned to that component. Percentages are computed with respect to n.

N	Method	Global	Component		
			Noise	2	3
50	G	30.64	23.74	4.14	2.76
	R	14.24	9.58	2.52	2.14
	IF	9.04	0.92	4.64	3.48
	IS	9.78	2.24	4.48	3.06
	TE	10.06	3.70	3.56	2.80
	TF	12.24	6.76	3.12	2.36
	N	10.96	—	5.64	5.32
200	G	9.50	2.98	3.45	3.07
	R	8.10	1.89	3.25	2.96
	IF	7.59	0.46	3.75	3.39
	IS	8.20	1.42	3.60	3.18
	TE	9.54	4.32	2.75	2.46
	TF	11.17	6.56	2.39	2.22
	N	10.22	—	5.10	5.12
500	G	7.58	0.92	3.63	3.03
	R	7.55	0.92	3.60	3.03
	IF	7.51	0.30	3.94	3.27
	IS	7.78	0.71	3.85	3.22
	TE	9.15	4.04	2.82	2.29
	TF	10.84	6.26	2.50	2.08
	N	10.33	—	5.27	5.06

Table 5.7: Upper-trimmed means of classes of parameters for the“Wide noise model”. For the methods IF and IS we also report the average value for the improper density c.

N	Method	c	d_π	d_μ	d_v
50	G	—	0.43	1.03	1.82
	R	—	0.30	0.45	1.23
	IF	0.043	0.17	0.43	1.07
	IS	0.018	0.18	0.42	1.04
	TE	—	0.13	0.42	1.05
	TF	—	0.12	0.43	1.18
	N	—	0.22	0.48	1.77
200	G	—	0.10	0.20	0.48
	R	—	0.09	0.19	0.44
	IF	0.043	0.08	0.19	0.43
	IS	0.032	0.11	0.20	0.54
	TE	—	0.06	0.20	0.66
	TF	—	0.06	0.21	0.91
	N	—	0.20	0.31	1.86
500	G	—	0.05	0.13	0.30
	R	—	0.05	0.12	0.29
	IF	0.043	0.06	0.13	0.32
	IS	0.036	0.08	0.13	0.42
	TE	—	0.04	0.12	0.57
	TF	—	0.04	0.13	0.86
	N	—	0.20	0.29	2.02

Table 5.8: Average misclassification percentages for the "Outliers model". The components' average misclassification percentage is the average percentage of points wrongly assigned to that component. Percentages are computed with respect to n.

N	Method	Global	Component		
			Noise	2	3
50	G	24.92	19.40	3.38	2.14
	R	3.74	0.48	1.66	1.60
	IF	3.14	0.08	1.64	1.42
	IS	5.26	2.36	1.48	1.42
	TE	7.36	3.80	0.38	3.18
	TF	10.08	5.48	2.02	2.58
	N	25.38	—	9.78	15.60
200	G	3.90	1.45	1.02	1.43
	R	2.53	0.07	1.03	1.43
	IF	2.48	0.01	1.03	1.44
	IS	3.36	0.92	0.99	1.45
	TE	7.25	5.62	0.68	0.95
	TF	9.49	8.10	0.53	0.86
	N	22.18	—	0.00	22.18
500	G	2.34	0.00	1.02	1.32
	R	2.35	0.01	1.02	1.32
	IF	2.33	0.00	1.01	1.31
	IS	2.67	0.34	0.99	1.33
	TE	6.77	5.13	0.74	0.90
	TF	10.01	8.88	0.40	0.73
	N	14.20	—	0.01	14.20

Table 5.9: Upper-trimmed means of classes of parameters for the"Outliers model". For the methods IF and IS we also report the average value for the improper density c.

N	Method	c	d_{π}	d_{μ}	d_{v}
50	G	—	0.21	3.41	1.11
	R	—	0.09	0.41	0.90
	IF	0.015	0.06	0.40	0.85
	IS	0.016	0.11	0.41	0.96
	TE	—	0.11	0.42	1.14
	TF	—	0.16	0.51	1.40
	N	—	0.50	0.83	19.59
200	G	—	0.02	0.20	0.38
	R	—	0.03	0.20	0.39
	IF	0.015	0.02	0.20	0.39
	IS	0.024	0.05	0.19	0.45
	TE	—	0.10	0.20	0.85
	TF	—	0.15	0.22	1.06
	N	—	0.51	1.68	10.84
500	G	—	0.01	0.12	0.25
	R	—	0.02	0.12	0.25
	IF	0.015	0.01	0.12	0.25
	IS	0.023	0.03	0.12	0.28
	TE	—	0.10	0.14	0.76
	TF	—	0.17	0.14	1.09
	N	—	0.34	1.39	5.95

Table 5.10: Average misclassification percentages for the "T-noise model". The components' average misclassification percentage is the average percentage of points wrongly assigned to that component. Percentages are computed with respect to n.

N	Method	Global	Component			
			Noise	2	3	4
50	G	24.12	16.10	3.12	2.78	2.12
	R	13.18	9.12	1.40	2.08	0.58
	IF	11.42	4.88	2.12	3.36	1.06
	IS	10.98	1.52	3.58	3.78	2.10
	TE	8.84	1.92	2.08	3.10	1.74
	TF	9.96	4.72	1.12	2.72	1.40
	N	12.44	—	5.22	4.18	3.04
200	G	11.49	5.38	3.02	1.53	1.56
	R	6.66	2.02	2.41	1.22	1.02
	IF	6.25	2.75	1.94	0.90	0.67
	IS	7.89	0.95	3.27	2.10	1.58
	TE	4.77	0.96	1.70	1.24	0.87
	TF	4.82	3.13	0.69	0.76	0.24
	N	11.04	—	5.93	2.31	2.79
500	G	7.76	0.60	3.34	2.09	1.73
	R	7.36	0.22	3.45	2.08	1.60
	IF	4.39	1.65	1.66	0.59	0.50
	IS	6.71	0.29	3.11	1.87	1.45
	TE	3.26	0.44	1.26	1.00	0.56
	TF	3.73	2.76	0.37	0.47	0.14
	N	10.41	—	5.52	2.41	2.48

Table 5.11: Upper-trimmed means of classes of parameters for the"T-noise model". For the methods IF and IS we also report the average value for the improper density c.

N	Method	c	d_π	d_μ	d_v
50	G	—	0.33	0.70	1.15
	R	—	0.28	0.37	0.74
	IF	0.056	0.20	0.36	0.84
	IS	0.010	0.19	0.35	1.08
	TE	—	0.10	0.33	0.71
	TF	—	0.10	0.32	0.61
	N	—	0.20	0.41	1.60
200	G	—	0.15	0.18	0.74
	R	—	0.14	0.15	0.48
	IF	0.056	0.17	0.14	0.40
	IS	0.037	0.10	0.16	0.80
	TE	—	0.06	0.14	0.37
	TF	—	0.05	0.14	0.32
	N	—	0.20	0.20	1.68
500	G	—	0.08	0.11	0.78
	R	—	0.08	0.11	0.73
	IF	0.056	0.16	0.09	0.26
	IS	0.048	0.07	0.10	0.66
	TE	—	0.04	0.09	0.22
	TF	—	0.04	0.08	0.25
	N	—	0.20	0.13	1.68

Table 5.12: Average misclassification percentages for the "Gaussian model". The components' average misclassification percentage is the average percentage of points wrongly assigned to that component. Percentages are computed with respect to n.

N	Method	Global	Component			
			Noise	2	3	4
50	G	29.72	26.70	0.48	1.56	0.98
	R	25.64	24.18	0.22	0.98	0.26
	IF	3.52	0.00	0.76	2.18	0.58
	IE	4.48	1.14	0.70	2.28	0.36
	TE	7.16	4.64	0.58	1.68	0.26
	TF	11.92	10.16	0.40	1.24	0.12
	N	3.48	—	0.76	2.14	0.58
200	G	17.51	16.04	0.69	0.68	0.10
	R	9.82	8.81	0.38	0.60	0.03
	IF	2.08	0.00	0.72	1.19	0.17
	IE	2.71	0.72	0.68	1.17	0.14
	TE	5.95	4.63	0.42	0.84	0.05
	TF	10.62	9.82	0.20	0.57	0.03
	N	2.07	—	0.72	1.18	0.17
500	G	6.33	5.16	0.41	0.62	0.14
	R	4.06	2.73	0.47	0.73	0.12
	IF	1.67	0.00	0.59	0.89	0.18
	IE	2.03	0.39	0.58	0.89	0.18
	TE	5.33	4.33	0.32	0.62	0.06
	TF	10.34	9.81	0.16	0.34	0.03
	N	1.67	—	0.59	0.89	0.19

Table 5.13: Upper-trimmed means of classes of parameters for the"Gaussian noise model". For the methods IF and IS we also report the average value for the improper density c.

N	Method	c	d_{π}	d_{μ}	d_{v}
50	G	—	0.57	1.35	2.61
	R	—	0.62	0.94	2.69
	IF	0.001	0.13	0.75	1.68
	IS	0.000	0.15	0.77	1.77
	TE	—	0.18	0.76	1.91
	TF	—	0.26	0.83	2.15
	N	—	0.13	0.75	1.68
200	G	—	0.36	0.60	1.45
	R	—	0.35	0.41	1.47
	IF	0.001	0.07	0.38	0.80
	IS	0.014	0.10	0.38	0.93
	TE	—	0.12	0.39	1.29
	TF	—	0.20	0.41	1.81
	N	—	0.06	0.38	0.80
500	G	—	0.19	0.31	0.92
	R	—	0.19	0.25	0.87
	IF	0.001	0.04	0.24	0.52
	IS	0.014	0.06	0.24	0.59
	TE	—	0.10	0.26	1.10
	TF	—	0.19	0.27	1.73
	N	—	0.04	0.24	0.52

CHAPTER 6

Concluding Remarks

§6.1. Contributions

Model-based cluster analysis is a statistical tool used to investigate group-structures in data using finite mixture models. Gaussian distributions are a popular device used to model elliptical shaped clusters and the estimation of mixtures of Gaussians is usually based on the maximum likelihood method. In this thesis we focus on mixture models for one-dimensional random variables. Throughout this work the number of components is considered to be fixed and known. For a wide class of finite mixtures, including Gaussians, maximum likelihood estimates are not robust. This implies that a small proportion of outliers in the data could lead to poor estimates and clustering.

One way to deal with this is to add a "noise component", i.e. a mixture component that models the outliers. In this thesis the word "noise" is used to identify all those data points which are extraneous to the sub-populations of interest. We made three main contributions.

In Chapter 3 we introduced a model which is a finite mixture of location-scale distributions mixed with a finite number of uniforms supported on disjoint subsets. We defined and proved the identifiability for such a model. Moreover, we introduced the maximum likelihood estimator and we showed its existence and consistency. We also provided a computational procedure based on the EM algorithm. Because of the computational complexity due to the presence of the uniform components, we also suggest some strategies to handle these problems. In Chapter 5 we explored the properties of this strategy empirically. It turns out that this methodology is particularly suited when there are points from the noise component localized in a certain region of the data-range. In these situations this method is able to achieve very good clustering performances.

Our second theoretical contribution is the development of the robust improper maximum likelihood estimator (RIMLE). Hennig (2004) proposed a pseudo-model in which the noise component is represented by a fixed improper density, that is, a constant on the real line. He showed that the resulting estimates are robust to extreme outliers. The latter is the main motivation for our investigation. We defined a pseudo maximum likelihood estimator for such kind of model and we stated conditions under which we showed that the RIMLE is strongly consistent for the maximizer of the integral of the pseudo-log-likelihood function with respect to the distribution function which generated the data. Furthermore we developed a successful methodology to select the value of the improper density value based on the dataset. This strategy has been also investigated on the basis of the empirical study presented in Chapter 5. The empirical performance of the RIMLE is very encouraging (see the next few paragraphs).

The third contribution of this thesis is an extensive simulation study in which we measure and compare the performance of the previous two methods and certain other robust methodologies proposed in the literature. The RIMLE in general performs better than the other methods. In fact, in all the situations, except the case when data are generated from t-mixtures, the RIMLE performs better than the other methods. But the performance of the RIMLE in the case of data from t-mixtures is not dramatically different from the performance of the maximum likelihood estimator for t-mixtures.

The RIMLE is fairly good at detecting when there is no noise in the data. In particular, this methodology offers the overall best performance when the sample size is small. In fact, when the sample size is small ($n = 50$, in our study) performance in the presence of noise is seriously affected for some of the methodologies. In particular, the maximum likelihood for the mixture model of Gaussians and uniforms — even when the data generating process includes uniform noise — shows poor performance for small sample sizes.

As a last point we would warn that the maximum likelihood estimator for t-mixtures with fixed degrees of freedom can be dangerous. This is demonstrated by the case when the data generating process is based on a mixture of t-distributions.

§6.2. Future works

The next step would be to extend these results to multidimensional random vectors which is of more interest for applications of model-based clustering. Even though the maximum likelihood estimator for mixtures of location-scales with uniforms resulted to be fairly good in some situations, its use in the multidimensional case would be limited by the resulting computational complexity.

The most interesting extension to the multidimensional case seems to be the RIMLE. While the statistical theory we developed in Chapter 4 can be easily adapted to multidimensional random vectors, it has to be assessed whether the selection method for the improper constant density value also works in a multidimensional setup.

It would be interesting to define the selection method as an estimator and to explore its asymptotic properties. The issue of asymptotic distributional properties for such methods is also a challenge.

Bibliography

Atienza, N., J. Garicia-Heras, and J. M. Munoz-Pichardo (2006). A new condition for the identifiability of finite mixture distributions. *Metrika 63*, 215–221.

Baker, G. A. (1940). Maximum likelihood estimation of the ratio of the components of non-homogeneous populations. *Tôhoku Math. J. 47*, 304–308.

Banfield, J. and A. E. Raftery (1993). Model-based gaussian and non-gaussian clustering. *Biometrics 49*, 803–821.

Bazaraa, M. S., H. D. Sherali, and C. M. Shetty (2006). *Nonlinear programming* (Third ed.). Hoboken, NJ: Wiley-Interscience. John Wiley & Sons.

Campbell, N. A. (1984). Mixture models and atypical values. *Math. Geol. 16*, 465–477.

Casella, G. and R. L. Berger (1990). *Statistical Inference* (first ed.). Belmont, California: Duxbury Press.

Chanda, N. K. (1954). A note on the consistency and maxima of the roots of the likelihood equations. *Biometrika 41*, 56–61.

Ciuperca, G., A. Ridofi, and J. Idier (2003). Penalized maximum likelihood estimator for normal mixtures. *Scandinavian Journal of Statistics 30*, 45–59.

Cramér, H. (1946). *Mathematical Methods*. Uppsala: Almqvist and Wiksell.

Day, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika 56*, 463–474.

Dean, N. and A. E. Raftery (2005). Normal uniform mixture differential gene expression detection for cDNA microarrays. *BMC Bioinformatics 6*(173), 1–14.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, (B) 39*(1), 1–47.

Dennis. J. E. J. (Ed.) (1981). *Algorithms for nonlinear fitting*. Cambridge. England. NATO advanced Research Symposium: Cambridge University Press.

DeSarbo. W. S. and W. L. Cron (1988). A maximum likelihood methodology for clusterwise linear regression. *J. Classification 5*, 249–282.

Donoho. D. and P. J. Huber (1983). *The Notion of Breakdown Point*. pp. 157–184. Belmont. CA: Wadsworth.

Feller. W. (1943). On a general class of "contagious" distributions. *The Annals of Mathematical Statistics 14*. 389–399.

Fraley. C. and A. E. Raftery (1998). How many clusters? which clustering method? answers via model-based cluster analysis. *The Computer Journal 41*. 578–588.

Fraley. C. and A. E. Raftery (2002). Model-based clustering. discriminant analysis. and density estimation. *Journal of the American Statistical Association 97*. 611–631.

Furman. W. D. and B. G. Lindsay (1994). Testing for the number of components in a mixture of normal distributions using moment estimators. *Comput. Statist. Data Anal. 17*(5). 473–492.

Hampel. F. R. (1971). A general qualitative definition of robustness. *The Annals of Mathematical Statistics 42*. 1887–1896.

Hasselblad. V. (1966). Estimation of parameters for a mixture of normal distributions. *Technometrics 8*. 431–446.

Hastie. T. and R. Tibshirani (1996). Discriminant analysis by Gaussian mixtures. *Journal of Royal Statistical Society, Ser.B 58*. 155–176.

Hathaway. R. J. (1985). A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *The Annals of Statistics 13*. 795–800.

Hathaway. R. J. (1986). A constrained em algorithm for univariate normal mixtures. *J. Comp. Stat. Simul. 23*. 211–230.

Hennig. C. (2004). Breakdown points for maximum likelihood estimators of locationscale mixtures. *The Annals of Statistics 32*(4). 1313–1340.

Hennig. C. (2005). Robustness of ML estimators of location-scale mixtures. In D. Baier and K. D. Wernecke (Eds.). *Innovations in Classification. Data Science. and Information Systems*, Heidelberg. pp. 128–137. Springer.

Hosmer, D. W. (1978). Comment on: Estimating mixtures of normal distributions and switching regressions. by R. Quandt and J. B. Ramsey. *J. Amer. Statist. Assoc. 73*(364), 730–752.

Hosmer, Jr., D. W. (1973). On MLE of the parameters of a mixture of two normal distributions when the sample size is small. *Comm. Statist. 1*, 217–227.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics 73-101*(35).

Huber. P. J. (1981). *Robust Statistics.* New York: Wiley.

Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *Ann. Math. Statist. 40*. 633–643.

John. S. (1970). On identifying the population of origin of each observation in a mixture of observations of two gamma populations. *Technometrics 12*. 565–568.

Karlis, D. and E. Xekalaki (2003). Choosing initial values for the EM algorithm for finite mixtures. *Comput. Statist. Data Anal. 41*(3-4). 577–590.

Kharin. Y. (1996). *Robustness in Statistical Pattern Recognition.* Dordrecht: Kluwer.

Kiefer, N. M. (1978). Comment on: Estimating mixtures of normal distributions and switching regressions. by R. Quandt and J. B. Ramsey. *J. Amer. Statist. Assoc. 73*(364), 744–745.

Kiefer. N. M. and J. Wolfowitz (1956). Consistency of the maximum likelihood estimation in the presence of infinitely many incidental parameter. *Ann. Math. Statist. 27*(364). 887–906.

Kumar. E. H., H. Nicklin. and A. S. Paulson (1978). Comment on: Estimating mixtures of normal distributions and switching regressions, by R. Quandt and J. B. Ramsey. *J. Amer. Statist. Assoc. 73*(364), 730–752.

Laird. N. (1978). Nonparametric maximum likelihood estimation of a mixed distribution. *J. Amer. Statist. Assoc. 73*(364), 805–811.

Lindsay, B. G. and P. Basak (1993). Multivariate normal mixtures: a fast consistent method of moments. *J. Amer. Statist. Assoc. 88*(422), 468–476.

Lindsay, B. G. and K. Roeder (1992). Moment based oscillation properties of mixture models. Technical Report 92-3, Center for Likelihood Studies, Pennsylvania State University, PA, USA.

Liu, C. (1997). ML estimation of the multivariate t distribution and the EM algorithms. *Journal of Multivariate Analysis 63*, 296–312.

Marrazzi, A., F. Paccaud, C. Ruffieux, and C. Beguin (1998). Fitting the distribuion of lenght of stay by paramteric models. *Medical-Care 36*(6), 915–927.

McLachlan, G. and K. E. Basford (1988). *Mixture Models: Inference and Applications to Clustering.* New York: Dekker.

McLachlan, G. and D. Peel (2000a). *Finite Mixture Models.* New York: Wiley.

McLachlan, G. and D. Peel (2000b). Robust mixture modelling using the t–distribution. *Statistics and Computing 10*(4), 339–348.

Mendenhall, W. and R. J. Hader (1958). Estimation of parameters of mixed exponentially distributed failure time distributions from censored life test data. *Biometrika 45*, 504–520.

Pearson, K. (1894). Contribution to the theory of mathematical evolution. *Philosophical Transactions of the Royal Society of London A 185*, 71–110.

Pearson, K. (1895). Contribution to the theory of mathematical evolution, ii: Skew variation. *Philosophical Transactions of the Royal Society of London A 186*, 343–414.

Perlman, M. D. (1972). On the strong consistency of approximate maximum likelihood estimator. In *Sixth Berkeley Symp. Math. Statist. Probab.*, Volume 1, Usa, pp. 263–282. Univ. of California Press.

Peters, B. C. and H. F. Walker (1978). An iterative procedure for obtaining maximum likelihood estimates of the parameters of mixture of normal distribution. *SIAM Journal of Applied Mathematics 35*, 362–378.

Priebe, C. E. (1994). Adaptive mixtures. *Journal of the American Statistical Association 89*, 796–806.

Quandt, R. E. and J. B. Ramsey (1978). Estimating mixtures of normal distributions and switching regressions. *J. Amer. Statist. Assoc. 73*(364), 730–752.

Rao, R. C. (1948). The utilization of multiple measurements in problems of biological classification. *J. Royal Statist. Soc. Ser. B 10*, 159–193.

Ray, S. and B. G. Lindsay (2005). The topography of multivariate normal mixtures. *Annals of Statistics 33*(5), 2042–2065.

Redner, R. (1981). Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. *The Annals of Statistics 9*, 225–228.

Redner, R. and H. F. Walker (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review 26*, 195–239.

Robertson, C. A. and J. G. Fryar (1968). Some descriptive properties of normal mixtures. *Skand. Aktuarier Tidskr. 52*, 137–146.

Roeder, K. and L. Wasserman (1997). Practical bayesian density estimation using mixtures of normals. *Journal of American Statistical Association 92*, 894–902.

Seidel, W., K. Mosler, and M. Alker (2000). A cautionary note on the likelihood ratio test in mixture models. *Annals of the Institute of Statistical Mathematics 52*, 481–487.

Shaked, M. (1980). On the mixtures from exponential families. *Journal of the Royal Statistical Society, Ser. B 42*, 192–198.

Tanaka, H. and T. Kawakami (2007). C^r strong cell decompositions in nonvaluational weakly o-minimal real closed fields. *Far East J. Math. Sci. (FJMS) 25*(3), 417–431.

Teicher, H. (1961). Identifiability of mixtures. *The Annals of Mathematical Statistics 32*, 244–248.

Teicher, H. (1963). Identifiability of finite mixtures. *The Annals of Mathematical Statistics 34*, 1265–1269.

Titterington, D. M., A. F. M. Smith, and U. E. Makov (1985). *Statistical analysis of finite mixture distributions*. Chichester: John Wiley & Sons Ltd.

Van der Vaart, A. W. and J. A. Wellner (1996). *Weak convergence and empirical processes*. New York: Springer-Verlag.

Wald. A. (1949). Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics 20*. 595–601.

Wolfe. J. H. (1967). Normix: Computational methods for estimating the parameters of multivariate normal mixtures of distributions. Research Memo SRM 68–2. U.S. Naval Personnel Research Activity. San Diego.

Wolfe. J. H. (1970). Pattern clustering by multivariate mixture analysis. *Multivariate Behavioural Research 5*. 329–350.

Wu. C. F. J. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics 11*(1). 95–103.

Yakowitz. S. J. and J. Spragins (1968). On the identifibility of finite mixtures. *The Annals of Mathematical Statistics 39*. 209–214.

Yu. J. (2004. January). Empirical characteristic function estimation and its applications. *Econometric Reviews 23*(2). 93–123.