2004

Faculteit Wetenschappen

Statistical Modelling Strategies for Reliability Data on Physical Components with Possibly Multiple Causes of Failure

Proefschrift voorgelegd tot het behalen van de graad van
Doctor in de Wetenschappen, richting Wiskunde,
te verdedigen door

Ellen ANDRIES

Promotor : Prof. dr. G. Molenberghs
Co-promotor : Prof. dr. L. De Schepper

INSTITUUT VOOR MATERIAALONDERZOEK

PARTNER IN DE UNIVERSITEIT LIMBURG

Voor Papa

Een woord van dank ...

Vier jaar geleden ben ik aan een uitdaging begonnen, vandaag is er dit doctoraat. Laat ik het maar clichématig stellen, omdat het gewoon zo is, zonder de steun en hulp van een aantal mensen was dit er zeer zeker niet gekomen. Een dankjewel is zelfs veel te weinig ...

Boven aan mijn lijst staan twee mensen die dit werk mee vorm gegeven hebben. Een betere promotor dan Prof. Dr. Geert Molenberghs kan ik mij niet bedenken. Bedankt Geert, voor de vrijheid van werken die ik kreeg, de voortdurende stimulans om het anders te bekijken, de talrijke tips en raadgevingen, maar bovenal voor de motivatie die je me telkens gaf, vooral in de laatste twee moeilijke jaren.

Dr. Kristof Croes heeft me aangemoedigd om dit werk te maken. Alleen heeft hij zich dit volgens mij reeds lang beklaagd. Ik vrees dat ik hem ondertussen ook een nieuw stel oren mag kopen. Bedankt Kristof, voor het luisteren, het lezen, de talrijke discussies, en nog zoveel meer.

Mijn copromotor Prof. Dr. Luc De Schepper wil ik bedanken voor het gegeven vertrouwen in mijn werk en de mogelijkheid om dit op het IMO te kunnen uitvoeren.

De overige leden van mijn doctoraatscommissie, Prof. Dr. Gilbert Knuyt en Prof. Dr. Ward De Ceuninck, alsook Prof. Dr. Geert Verbeke, zou ik willen bedanken voor de getoonde interesse doorheen deze vier jaren.

Ik heb me op het IMO altijd thuis gevoeld en daar hebben mijn collega's een grote verdienste aan. Ik was zeker niet altijd even aangenaam en heb

meer dan genoeg geklaagd. Bij deze dus een welgemeende sorry! Een aantal diensten op het IMO heb ik veelvuldig bezocht: Eric moet ik bedanken voor het geduld met de probleempjes van mijn computer en de vrouwen van het secretariaat voor het oplossen van alle mogelijke problemen waar ik nogal vaak mee kwam aandraven.

Ik kon bij veel vrienden en familie terecht om te ontspannen. Eén plaatsje in de Tierstraat heeft er voor gezorgd dat ik mij van in het begin heb thuis gevoeld in het verre limburg. Bedankt Krista en co!

Ik weet dat papa trots zou zijn, maar ook dat mama het nu voor twee is. Jullie hebben me altijd ten volle gesteund en mijn keuzes aanvaard. Bedankt voor alles papa, mama, Karen en Paskal!

Meer dan wie ook weet je hoeveel moeite dit me heeft gekost. Je was er altijd, aanvaardde mijn humeur, maakte me terug aan het lachen en zorgde ervoor dat ik dit doctoraat toch afwerkte. Ik hoop dat ik het je ooit allemaal kan teruggeven. Dank je Dieter.

Bedankt iedereen,
Ellen

Contents

Chapter 1

Introduction

In the past, a microelectronic device used to be evaluated on its performance and cost price. A high-quality product was understood to have no defects or systematic failures on the moment it was leaving the manufacturer. Moreover, the reliability of a device, i.e., the ability to perform its function under normal working conditions during a specific period of time (Birolini, 1994), was not a serious issue. Today, the reliability of a device has become an important subject. A high-quality product now also stands for a reliable and safe product. The reasons of this tendency are manifold. Not only has the competition between several manufacturers moved to the level of a demonstrable reliability of their product, but also there is the increasing expectation of the customer, as microelectronic devices have become indispensable in human life. There are the growing costs of maintenance and failure, the tendency to compare products based upon quality standards like ISO9001, the fact that some products should be highly reliable for safety reasons and the use of electronics in "harsh" environments, like the automobile industry and space travel.

On top of that, with the advent of high-tech devices more and more reliability problems are encountered. One of the main reasons of this is the ongoing miniaturization of the integrated circuit (IC) due to the demand

for highly sophisticated and high speed products. The result of this down-scaling reveals itself both in an increase in the number of components of an IC per unit of volume and a decrease of the dimension of these components. Where components with a dimension in the order of μm were innovative in the eighties, these days they are already old-fashioned. The consequences of device scaling with respect to reliability are clear. A higher reliability of the components is required to obtain the same reliability requirements for an IC, physical mechanisms not playing a role in the aging process of older technologies become important in new, down-scaled, technologies and new materials, like Cu, high-K and low-K, have to be introduced.

In summary, the performance of reliability experiments has become an essential issue in the development of microelectronic devices. Hereby, reliability comes mainly back at three places: during the design of a chip, during the qualification of a production process and during the production monitoring. In essence, this work deals with the statistical modelling of data generated by means of a reliability experiment. This kind of data is often complex and requires for their processing advanced statistical techniques. In some cases the microelectronic industry is faced with a lack of adapted software. In addition, many statistical techniques used in other areas cannot be applied in a straightforward manner due to the specific nature of reliability data.

1.1 Aspects of a reliability analysis

The concept of "reliability analysis" is often indicated as a pure statistical matter. Nevertheless, statistics are only a part of it. On the whole, it includes all kinds of investigations, experiments, analyses, ..., in order to determine various aspects of the reliability of an item. A good overview of what is understood under reliability analysis in microelectronics can be found in Birolini (1994) and Ohring (1998). One of the important parts in

microelectronics, is the process to understand the working and failure of a specific component. For a basic component this means the physical understanding of the aging process, while for a more complex device this means the relation between its failure and the failure of its basic components. Thanks to this understanding, many models used in the statistical analysis can be warranted. As this project deals with the statistical aspects of reliability analysis, these models will not be questioned.

Basic concept and definitions

There exists a broad class of reliability experiments each with its own specific goals. For example, there are controlled laboratory experiments with basic components to assess their initial reliability, field experiments with devices already on the market to compare results with those obtained from designed laboratory tests, burn-in experiments to remove early failures from a population of components, ... Within this class of experiments, a substantial part is carried out with the intention of assessing product reliability. In other words, the goal of the experiment is to predict the reliability of the *device under test* (DUT).

The *reliability* of a device is defined as the probability that the device under normal working conditions and during a specific period of time will carry out its function. Moreover, if τ is the failure-free operating time of a device (under normal working conditions) and considered to be a random variable with distribution function $\mathrm{F}(t) = \mathrm{P}(\tau \leq t)$, then reliability can be quantified by means of the *reliability function* $\mathrm{R}(t) = 1 - \mathrm{F}(t)$, i.e., the probability that the device has not failed at time t (Birolini, 1994). In general, an item is qualified as being reliable when it satisfies certain reliability requirements. The latter are commonly expressed in terms of an $x\%$ percentile or $t_{x\%}$ quantile, i.e., the time at which $x\%$ of the total population of tested items has failed, with x a predefined percentage which is usually very small. This $t_{x\%}$ quantile is related to the distribution of τ through

$F(t_{x\%}) = P(\tau < t_{x\%}) = x$. It explains the importance of the estimation of low quantiles in reliability analysis.

Event time data Reliability data collected from experiments carried out to assess reliability, are so-called *time-to-failure* or *failure time* data. This means that for each DUT an event time t^e is registered. Failure time data belong to the class of event time data. Examples of the latter from other areas are survival data in biology and medicine, and event time data in social sciences. Still, the specific nature of failure time data often requires another approach, in particular when compared to survival data. Main differences between these two groups of even time data are:

- Reliability experiments deal with items that are "dead" material, in contrast to survival experiments. As a result, many reliability experiments can be carried out in controlled circumstances, which is not the case for medical or biological studies concerning humans or animals.
- Experiments in survival analysis can easily last a few years, contrary to reliability experiments in microelectronics that should be carried out fast and for a relatively small amount of money. This is due to both a competition in price and performance between different manufacturers.
- While in survival analysis non- or semi-parametric statistical methods are extensively used, in reliability analysis mostly parametric techniques are applied. Partly, this is related to the different objectives aimed at with both analyses. In particular, in a reliability analysis often a lot of extrapolation is required. Another reason is that reliability data are usually more appropriate for parametric techniques due to the controlled environment of the experiment.

So, although there are a lot of statistical techniques to analyze event time data, which is for a large part due to the extensive research in survival analysis, it is not straightforward to apply them on failure time data. Moreover,

it is difficult to carry out a reliability analysis with most commercial statistical software packages, as they include only a limited number of functions for the (parametric) analysis of survival data.

The observed event time t^e of a DUT is defined as the time span between the start of the experiment for this DUT and the occurrence of an event. The latter is either a failure or a removal of the DUT. In case the observed event is a failure, the event time is the *failure time* or observed *failure free operating time* t of the DUT, i.e., the time span between the initial operation and failure of the device. For each device, the definition of failure is specified in advance and depends on the reliability requirements for the device under test. For example, the failure of a light bulb can be defined as the moment that the bulb does not burn anymore, but equally well the bulb can be defined as having failed when the amount of light it diffuses is decreased with a certain percentage. If the DUT is removed from the experiment before it failed, the event time, i.e., the time point of removal, is a *censoring time* and the observation is referred to as *right censored*. The reason for this removal can be a defect of the measuring system, end of the test, ... There exist another kind of censoring where the DUT has already failed on the moment that it is removed from the experiment. In this case, the observed event time is *left censored*. An example of this situation is when the DUT has failed before the first time point of inspection. A special case of censoring is when the exact event time of a device cannot be observed, but only a time interval in which the event occurred. The observed event time is then *interval censored*. This situation typically happens when the DUT cannot be monitored continuously, i.e., the device is inspected for a failure or a removal only at specific points in time.

Type of samples Depending on the nature of the observed event times, several kinds of a failure time sample can be collected from a reliability experiment. Some important examples are:

Complete sample. The experiment is finished when all components have failed. Either all failure times are observed or they are all interval censored. Due to a lack of time, this kind of samples does not often occur in real terms. However, they are the basis for developing statistical techniques.

Type I singly right censored sample. The experiment is stopped at a predefined point in time. Either a device has failed or it is right-censored at the last measured time point. The number of failed devices is not known in advance, in contrast to the total length of the experiment. Since the total test time is known, this type of experiment is popular.

Multiple right censored sample. A device can be right censored at any time during the experiment. This means that censoring times can be smaller than failure times (in contrast to a type I singly censored sample). This sample can be the result, for example, of an experiment where some measurement systems of devices brake down, of an experiment where failed devices are replaced by new devices or of a group of experiments stopped at different points in time.

A lot of experiments carried out in a controlled environment will be type I singly censored, while field data will be more frequently multiple censored. Many other forms of samples exist, an overview can be found, for example, in Lawless (1982, Chap. 1).

Failure time distributions

In the simplest setting of an experiment, items from one population are put into operation under the same conditions. Although the items are produced in the same way, they are not identical due to physical differences caused by the production process. As a result, there is a spread present on the observed failure times. These failure times are a realization of the positive random variable τ. Its (failure time) distribution reflects the variability

in failure of an item caused through the differences in the population. There are many distributions that can be used to model the variability between the failure times, a description of most of them can be found in Meeker and Escobar (1998, Chap. 4) and Lawless (1982, Chap. 1). Note that a normal distribution is usually not suitable and as such rarely applied, since its domain is the real axis and a failure time is a priori non negative.

In practice, only a few distributions are used as failure time distribution. The distributions commonly considered are the exponential, lognormal, weibull and gamma distribution. Attempts to use other distributions are seldom found in literature. Of these four, the lognormal and Weibull distribution are the most popular ones. The lognormal distribution function $\mathrm{F}_{LN}(t)$ is given by:

$$\mathrm{F}_{LN}(t) = \Phi\left(\frac{\ln(t) - \ln(\eta)}{1/\beta}\right), \tag{1.1}$$

with $\Phi(x)$ the standard normal distribution function, $\eta > 0$ a scale and $\beta > 0$ a shape parameter. Often, the location parameter $\mu = \ln(\eta)$ and the scale parameter $\sigma = 1/\beta$ are used instead of η and β. This originates from the fact that if a random variable T is lognormally distributed with parameters η and β, then the logarithm of T, i.e., $Y = \ln(T)$, is normally distributed with parameters μ and σ. The Weibull distribution function $\mathrm{F}_W(t)$ is given by:

$$\mathrm{F}_W(t) = 1 - \mathrm{e}^{\left[-\left(\frac{t}{\eta}\right)\right]^{\beta}} = \Phi_{SEV}\left(\frac{\ln(t) - \ln(\eta)}{1/\beta}\right), \tag{1.2}$$

with $\Phi_{SEV}(x) = 1 - \mathrm{e}^{-\mathrm{e}^{x}}$ the standard smallest extreme value (SEV) or Gumbel distribution (Gumbel, 1958), $\eta > 0$ a scale and $\beta > 0$ a shape parameter. Also here holds that if a random variable T is Weibull distributed, then $Y = \ln(T)$ will have a SEV distribution with parameters $\mu = \ln(\eta)$ and $\sigma = 1/\beta$.

There are two main reasons why the lognormal and Weibull distribution are often applied to reliability data. First, both distributions are

log-location-scale distributions, i.e., their distribution function can be written as $F(t) = F_0\left(\frac{ln(t)-\mu}{\sigma}\right)$ with $F_0(y)$ a standard function independent of any parameter, μ a location and σ a scale parameter. For these distributions, statistical theory and mathematics are relatively simple. Second, both distributions have properties that make them appropriate to model failure times. Extreme value theory suggest the suitability of the Weibull distribution and the central limit theorem of the lognormal distribution for many applications (Crowder et al., 1991; Birolini, 1994). Apart from this, the Weibull distribution allows both a monotonically increasing ($\beta > 1$) and decreasing ($\beta < 1$) hazard function. An increasing hazard is suitable for components that are subjected to wear out or fatigue, while a decreasing hazard can be used for components with an initial weakness (Birolini, 1994). Further, for some typical components of an IC, the failure time distribution is considered to be Weibull or lognormal based on either a theoretical development or a longterm experience (for example, Lloyd, 1979; Lloyd and Kitchin, 1990; Degraeve, 1998).

The exponential distribution is the most simple failure time distribution and a special case of the Weibull distribution (i.e., $\beta = 1$). In the past, it was used extensively, partly due to the availability of easy statistical methods for this function. Still, the exponential distribution turned out to be inappropriate for many current devices due to its property of a constant hazard or no memory (Lawless, 1982, pp. 14-15). This implicitly assumes that devices do not age, wear out or have no weaknesses. It can be used for robust devices that do not start to wear out for a long period. At present, however, most high-tech devices are not robust. The gamma distribution is not often applied, mainly because mathematically it is not attractive. Also it is no log-location-scale distribution, although it does allow both monotonically increasing and decreasing hazard functions.

Lifetime models

In order to assess the reliability of an item, it would be sufficient to test a sample of the population at normal working conditions. If one would last long enough, eventually all items of the sample will have failed. The failure time distribution and so also the reliability function of the failure free operating time τ for this item, could then be estimated from the observed failures times. There is, however, one crucial problem with this procedure: the experiment would generally last ages (in the order of years). Since in the microelectronic industry new devices and technologies are developed in quick succession, this procedure is not suitable to test new devices. The technique generally applied to shorten experiments such that is still possible to estimate the reliability of a device at working conditions is called *accelerated aging*. Based on the physical mechanisms that are responsible for the aging of a device, physical factors are searched for that can accelerate the aging of a device through an increase or decrease of their normal value. These factors are referred to as *stress factors*. Typical stress factors are temperature, current density, voltage, electrical field, ... A *stress level* is a value of a stress factor. If an experiment is then carried out at elevated stress levels, devices will fail much earlier in time. Usually, it is assumed that the applied stress will not have an influence on the type of failure time distribution, but will just modify the parameters (Birolini, 1994). An *acceleration* or *lifetime model* is then used to extrapolate the failure time distribution from higher stress levels to normal working conditions. This model relates the median failure free operating time τ of a device to the stress factors applied and assumes that the shape of the failure time distribution is independent of the applied stress. This last assumption can be easily adapted through the incorporation of a model for the shape parameter. A lifetime model is mainly built upon both theoretical and experimental physical evidence. Some are broadly accepted, while for others a large disagreement exist.

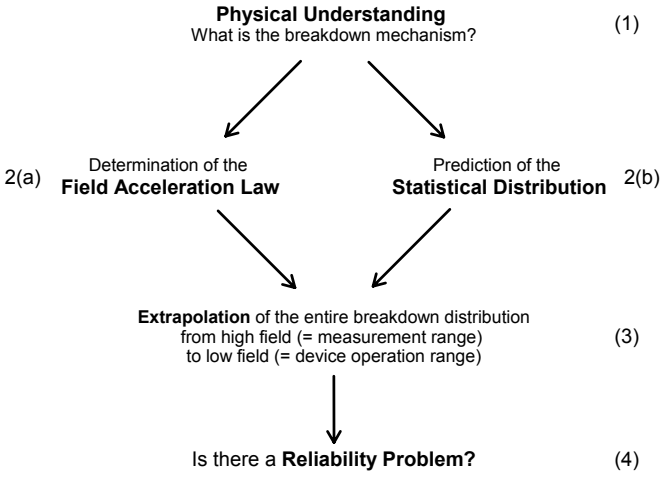


Figure 1.1: Schematic outline of how to approach the reliability problem for an oxide (Degraeve, 1998).

Example 1: reliability of an oxide To summarize, we give an example of how the reliability analysis for an oxide can be handled. The outline of how to approach its reliability problem is taken from Degraeve (1998) and given in Figure 1.1. An oxide is a dielectric, which is part of a capacitor and of a transistor. Devices which are on their turn important components of an IC. Oxides become constantly thinner and lately also new materials are tested. Each time a "new" oxide is developed or an old one is improved, its reliability has to be studied. The outline in Figure 1.1 reflects the different steps which have to be carried out in a reliability analysis. The first step is the work of engineers or physicists and is essential for the development of the models as used in the next step. Obviously, a lot of research and initial experiments are carried out. In the second step, two models are searched for based upon results from (1) and possible prior knowledge. In step 2(b), the type of failure time distribution for τ is chosen such that it reflects the

variability of failure within the population of oxides. For this kind of device, typically the Weibull distribution is taken. In step 2(a), it is decided how the lifetime model (or Field acceleration law) should look like. For an oxide, mainly two stress factors are important in its aging process: temperature and electrical field. Possible values for these factors in reliability requirements are: E = 4MV/cm and T = 100 °C. By an increase of the level of one of these two factors, the oxide will age faster and so fail earlier. When the only stress factor in the experiment is the electrical field, one of the two following acceleration models are commonly used:

$$\eta = C\mathrm{e}^{(\gamma \mathrm{E})} \tag{1.3}$$

$$\eta = C\mathrm{e}^{(G/\mathrm{E})} \tag{1.4}$$

with η the median lifetime of the oxide, E the value of the electrical field and C, γ and G parameters of the model depending on the specific oxide. Clearly, conclusions about the reliability of an oxide will depend on the acceleration model used. Large disagreements are found in literature concerning the use of one of these two models (Degraeve et al., 1998; Martin et al., 1998). In step (3) the models of step (2) are combined in order to extrapolate the failure time distribution of τ at normal working conditions. Although the scale parameter of the Weibull distribution is equal to the 63% quantile and not the 50% quantile or median, it is generally substituted in the lifetime model. The underlying idea is that the parameter η in model (1.3) or (1.4) could represent any quantile of the failure time distribution. Based on reliability data collected from an experiment carried out at elevated values for the electrical field, the models of step (2) can be estimated. From this, the failure time distribution (and so the reliability function) at normal working conditions can be estimated. In step (4), the obtained results are compared with the stated (in advance) reliability requirements. Note that the outline given in Figure 1.1 can be generalized to many other basic components of an IC or even more complex devices.

Parameter estimation

In microelectronics, there is no general agreement on the choice of the estimation method. Many different techniques are considered, from which some of them are even questionable. For example, a popular method to estimate the parameters of a lognormal or Weibull distribution is least squares estimation based on the corresponding probability plot. Nevertheless, the resulting estimated standard errors are incorrect and mostly by far too small. Also there is a lack of adapted software (or easy to use software). It is common practice to consider the most simple method and not necessarily the most adapted one. Although there are standards available from JEDEC of how to analyze certain experiments (for example, the standard JESD37 to estimate the parameters of a lognormal distribution in case of censored and singly right censored samples), many of the proposed methods are not up-dated, and in addition not always followed.

However, under the force of circumstances, there is the tendency to pay more attention to the use of suitable methods. Next to least squares estimation, which is still used both in appropriate and inappropriate situations, the maximum likelihood (ML) method is applied when feasible. Although other techniques are described in advanced statistical (reliability) literature, like Meeker and Escobar (1998), they are rarely used.

Example 2: a reliability sample from an accelerated test Figure 1.2 shows on a lognormal probability plot, a reliablity sample collected from an accelerated test. This failure time sample is obtained from a temperature storage experiment on commercial metal film resistors carried out at the Institute for Materials Research (IMO). The only stress factor used for this accelerated test is temperature. The applied stress levels are 120 °C, 145 °C and 155 °C. The temperature at normal working conditions is 90 °C. At each level, 126 components were tested. Apart from 1 right censored observation for stress level 155 °C at the end of the experiment, all components failed. For

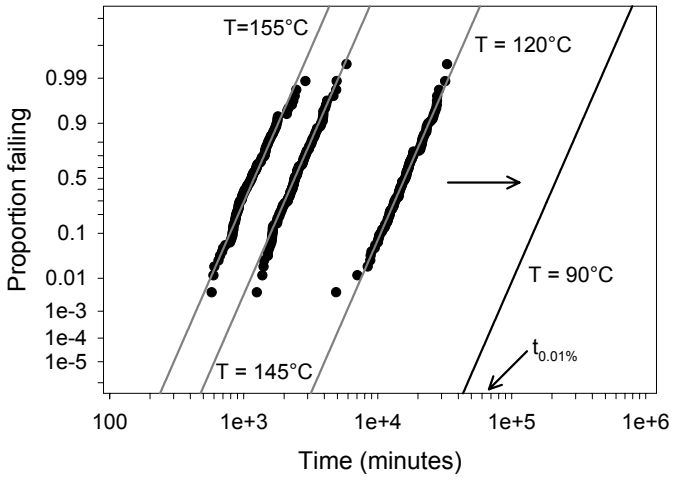
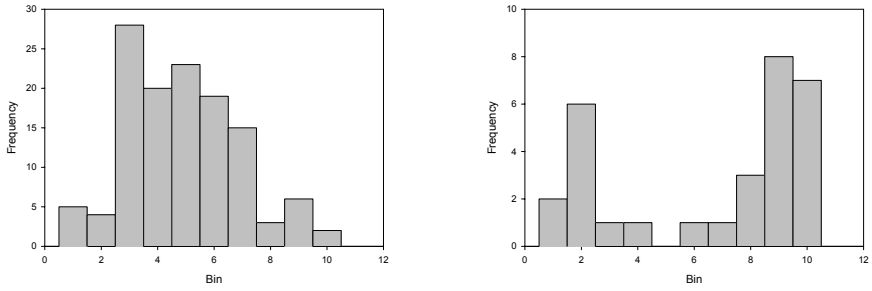


Figure 1.2: Lognormal probability plot of the failure time sample of example 2. The maximum likelihood estimate of the failure time distribution at all stress levels as well as at the working condition is shown.

this type of device, often the lognormal distribution is considered as failure time distribution and the following Arrhenius model as lifetime model:

$$\eta = C\mathrm{e}^{\left(\frac{E_a}{k_B \mathrm{T}}\right)}, \tag{1.5}$$

with η the median lifetime, T the absolute temperature in degrees K, k_B the Boltzmann constant ($=8.6\,\text{e-}5\,\mathrm{eV/K}$) and C and E_a two parameters of the model. The acceleration factor E_a is commonly referred to as the activation energy. Since the scale parameter of the lognormal distribution is also the median of the distribution, it can be substituted in model (1.5) for η. Based on physical experience, the shape parameter of the lognormal distribution is assumed to be equal for all stress levels. The following model is then fitted



(a) Sample 1.

(b) Sample 2.

Figure 1.3: Histograms of (logarithmic) failure time samples.

to the sample by means of ML estimation:

$$F(t; C, E_a, \beta) = \Phi\left(\frac{\ln(t) - \ln(\eta_{\mathrm{T}})}{1/\beta}\right)$$
$$\eta_{\mathrm{T}} = C\mathrm{e}^{\left(\frac{E_a}{k_B \mathrm{T}}\right)}. \tag{1.6}$$

The fits of the ML estimate of the failure time distribution at the stress levels and at the temperature of 90 °C are also shown in Figure 1.2. From this, $t_{x\%}$ quantiles can be estimated and compared with reliability requirements. For example, a possible reliability specification could be that the $t_{0.01\%}$ quantile (at T=90 °C) should be larger than 1 month ($\approx$ 4.4e+4 minutes). Since the ML estimate for the $t_{0.01\%}$ quantile is 6.62 e+4 minutes, with a 95% asymptotic confidence interval given by [5.48 e+4; 7.89 e+4] minutes, this requirement would be achieved.

1.2 Multimodal failure time data

In a reliability analysis, quite often not much attention is paid to the choice of the failure time distribution for τ. Mostly, its type is rather

chosen a priori with the belief that a simple (one or) two parameter distribution, like the Weibull or lognormal distribution, is sufficient to describe the variability in failure within the population of interest. In other words, it is implicitly assumed that τ has a *homogeneous* distribution. Although this used to hold for many devices, more and more it seems that for some of them these "basic" distributions do not capture anymore all features present on the spread of the failure times. To illustrate what is meant, Figure 1.3 displays two histograms. Both depict the logarithmic failure times of a complete failure time sample obtained from an experiment on a certain device, carried out in the simplest setting. As noted, in histogram 1.3a the failure times are more or less grouped, in contrast to histogram 1.3b which shows the presence of two groups among the failure times. It turns out that the distribution of the sample in 1.3a can be adequately described by a lognormal distribution, while for the sample in 1.3b none of the homogeneous distributions are appropriate. They all lack the possibility to model two (or more) groups of failure times.

The sample depicted in histogram 1.3b is a typical example of a so-called *multimodal*, in particular bimodal, sample. The failure time sample is heterogeneous, in the sense that the failure times can be divided into a number of groups with each group related to a different failure behavior. Physically, this means that within the population several failure mechanisms act, which cause "the same kind of devices" to fail differently. This explains the term *multimodal*, referring to the several failure causes or mechanisms present in the population. It ought to be mentioned that these multimodal failure time samples are not new (Kao, 1959; Joyce et al., 1976). Only, in the past apparently the microelectronic industry experienced no reliability problems when a simple failure time distribution was used to model the distribution of the failure time τ of these multimodal populations. Today, however, not only a more adapted "heterogeneous" model is required due to the fact that more reliability problems are encountered, but also multi-

modal failure time samples occur much more frequently due to the advanced technology (Møltoft, 1983; Atakov et al., 1994; Ogawa et al., 2001).

1.2.1 Types of heterogeneous failure time distributions

Physically, a different failure behavior for the same kind of devices is usually the result of one of the following two situations. On the one hand it occurs that each device of the population can only fail due to one kind of mechanism, but within the population there are several groups of devices with each having a different failure behavior. These differences can be the result of a different failure mechanism between the groups or of the same mechanism which started to act on a different point in time. A typical example of this situation is the production of a device through several machines or through a production process that can lead to essential differences (with respect to the failure behavior) within a device (Fisher et al., 2000). On the other hand it happens that each device can fail due to multiple "competing" failure mechanisms, but the device only fails due to the mechanism which acts first. Again the population can be divided into several groups of devices with a different failure behavior. Only, where for the first situation theoretically it would be possible to distinguish the groups before the devices are put into operation, here a subdivision can only be made after all devices failed. A typical example is given by a device that can fail due to either weaknesses in the material caused by the production process or to wear out through its aging process. The former is referred to as an *extrinsic* failure, the latter as an *intrinsic* failure (Degraeve, 1998).

For these multimodal populations a *heterogeneous* failure time distribution, built up of homogeneous distributions, is required to adequately describe the variability within the failure times. From a statistical point of view, the first situation is modeled through a *finite mixture distribution* and the second trough a *competing risks or minimum type* model. Both kind of heterogeneous distributions are very popular in certain domains of statistics

and the statistical literature concerning these models is huge.

1.2.2 The use of heterogeneous distributions in reliability

In spite of the need to use complex failure time distributions, they are rarely applied in practice. The main reason for this is the lack of adapted software in the microelectronic industry. But also the fact that organizations, like JEDEC, which prescribe standard methods for the implementation of reliability analyses, do not even consider methods to analyze multimodal failure time samples, will not stimulate a more advanced statistical analysis of these samples. This situation is not surprising given the rather limited amount of (reliability) literature concerning the application and estimation of heterogeneous failure time distributions. Although many textbooks on reliability analysis mention the importance of these distributions, they often do not consider estimation tools for them. Nevertheless, this is remarkable given that the earliest articles dealing with multimodal failure time samples, appeared already in the fifties (Acheson and McElwee, 1951; Kao, 1959).

Literature Articles in the domain of reliability, involved with the analysis of multimodal populations can be roughly divided into two groups. In the first one, new devices with a multimodal failure behavior are discussed. The main purpose is the physical understanding of the different failure causes or mechanisms. Often, multimodal failure time samples are shown, discussed and fitted, but mainly one has to guess the estimation method carried out (Sichart and Vollertsen, 1991; Fisher et al., 2000; Ogawa et al., 2001). Hereby, the main group of articles deals with bimodal populations. Literature about this subject has grown a lot recently and is still growing (Møltoft, 1983). The second group includes articles that are related to the statistical analysis of multimodal failure time samples. Many of them consider relatively simple estimation methods, from which graphical methods are the most popular (Kao, 1959; Joyce et al., 1976; Møltoft, 1983; Jiang and

Murthy, 1995). Also, (new) estimation procedures are sometimes proposed in order to avoid the more complicated and computational unattractive ML method (Ling and Pan, 1998; Mu et al., 2000). Note that the latter, however, should be treated with extreme caution since mostly properties of obtained estimators are rarely known and quite often any statistical or physical relevance is lacking. Still others, introduce different heterogeneous distributions which do not rely on any physical evidence, in order to allow a graphical estimation procedure (Zhang and Ren, 2002). Only a few articles deal with more advanced estimation techniques, like the ML method (Mendenhall and Hader, 1958; Chan and Meeker, 1999). Even so, many of them use the methods without considering the pitfalls.

Software Many up-to-date applied software packages (used in the microelectronic industry) can estimate models that assume a basic failure time distribution, like (1.6), including several forms of censoring. Quite often, however, they are not able to estimate a heterogeneous failure time distribution. To our knowledge, WEIBULL and FAILURE, are two of the few packages, if not the only ones, which do allow the estimation of certain heterogeneous failure time distributions. While FAILURE makes only use of the ML method, in WEIBULL also a least squares technique is available. Nevertheless, there seems to be some serious problems involved with their ML estimation procedure. Namely, when using FAILURE for the ML estimation of a complex failure time distribution, starting or initial values for the parameters have to be supplied in order to obtain ML estimates. Therefore, the package is rather useless in practice, since software is demanded which allows the "automatic" estimation of models. Although this requirement is fulfilled by WEIBULL for some heterogeneous distributions, there is still another problem of even more concern. Apparently, both packages can lead to different ML estimates when estimating the same model to the same sample. Clearly, this gives rise to different results and as such there is no unified ML

approach. Furthermore, it is still the question whether both, one or none of these two estimates have the desired properties of an ML estimate and whether even more solutions would be possible. Importantly, no one seems to be aware of these problems or take them seriously.

1.3 Objective

Too often somewhat simple graphical procedures are still applied as only estimation tool in the few cases that a heterogeneous distribution is considered for the analysis of multimodal failure time samples. But, even if the more advanced ML method is used, it appears that its correctness cannot be guaranteed. Distributions are estimated rather blindly with too much respect for the ML method and without any justification of its use. What about the appropriateness of the ML method for these distributions, the multiple solutions, the statistical properties of the obtained solutions, the initial values which have to be supplied, ...

In spite of this, more and more there is not only the demand to analyze multimodal failure time samples with appropriate models, but also to use, due to its increasing popularity, the maximum likelihood method for their estimation. The main reason which prevents this method from being used is the absence of a software package which allows the estimation of these heterogeneous distributions, in a manner of speaking, with a press of the button.

The aim of this project has its origin in these problems associated with maximum likelihood estimation. Namely, how to estimate certain heterogeneous distributions to multimodal and in particular bimodal samples, with a principled ML approach such that it is workable in practice. Importantly, this project could only be established through a strong cooperation with other domains of statistics. It demonstrates the universal character of statistics. It is the hope that in the future the interaction between the

different domains will become more pronounced.

1.4 Outline

From a modelling point of view, there are several ways to combine simple homogeneous distributions to a heterogeneous distribution. Nevertheless, only two models are omnipresent in the statistical literature, i.e., the finite mixture and the competing risks model. Their popularity can be mainly attributed to the fact that for many problems, as also for the problem at hand, these models are the most natural and logical ones. In this work, we will focus on the mixture model. Reasons are related to the amount of knowledge already available about this model and the fact that, so far, apparently it is the model most commonly considered in reliability problems. Chapter 3 introduces the finite mixture model. Some definitions and other notions are given as well as the form of the mixture required to model the distribution of multimodal failure time samples, i.e., the general finite mixture model. Also, we discuss the ML estimation of this model and its related problems.

As this project is built upon the aim to analyze multimodal failure time samples, some of them are used as examples throughout this work. They are obtained from experiments carried out at IMO, from other companies or from literature. Chapter 2 introduces these examples. To illustrate that this work should not be restricted to the domain of reliability only, a few other samples present in the literature are used as examples as well. Although, they are not the result of a reliability experiment, they can be modeled by means of a general finite mixture distribution.

In Section 1.2.2, we referred to several problems encountered during the ML estimation of heterogeneous distributions. Essentially, these problems boil down to the non-existence of the classical ML estimate and the large number of maxima of the likelihood function. In Chapter 4, both

problems will be handled in detail for the general finite mixture model. Some (standard) solutions dealing with the nonexistent ML estimator are compared to the alternative likelihood estimator. It is discussed how the so-called spurious maxima and the many maxima can add something to the likelihood analysis. In essence, we give our perception of how to deal with the (maximum) likelihood estimation of finite mixtures in practice.

The main concern for the microelectronic industry are not the theoretical problems associated with ML estimation, but the practical problem of how to obtain these estimates. We developed a starting value method that automatically calculates the required estimates. In addition, it can be used as an exploration tool. This procedure is introduced in Chapter 5. Further, its performance as starting value procedure is evaluated and compared to other methods and we consider some of its additional features.

To conclude, two case studies are discussed in Chapter 6. The first handles the analysis of a bimodal failure time sample obtained from an accelerated test at 3 stress levels. The effect on the reliability conclusions whether or not the bimodal failure behavior is taken into account, is illustrated, as well as how appropriate likelihood estimates can be easily obtained for a heterogeneous failure time distribution. In the second example, the galaxy sample is considered. For this sample, the specific number of mixture components is unknown. We show how the concept of spurious maxima and the use of the starting value method can contribute to the discussion of the number of mixture components.

In the final Chapter, main conclusions are stated as well as some ideas for future research.

Chapter 2

Key examples

For the moment, this chapter only contains the QQ-plots of the samples used as an example throughout this work. It will contain a brief description of the samples: background, failure time distributions usually considered, ...

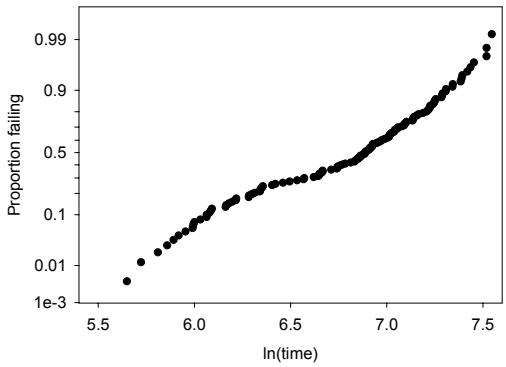


Figure 2.1: Lognormal QQ-plot of the resistor sample.

2.1 Failure time samples

2.1.1 The resistor sample

2.1.2 The interconnect sample.

2.1.3 Two laser samples

The laser A sample

The laser B sample

2.1.4 Two electromigration samples.

The EM1 sample

The EM2 sample

2.1.5 Appliance failure sample

(taken from Lawless (1982, p. 256))

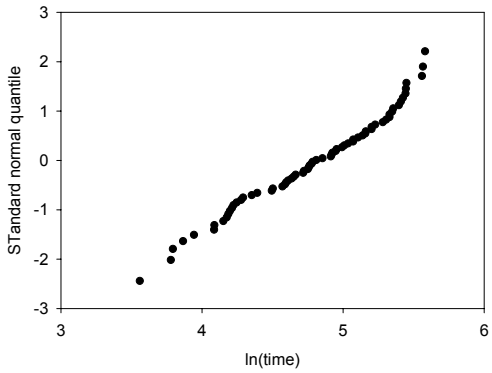
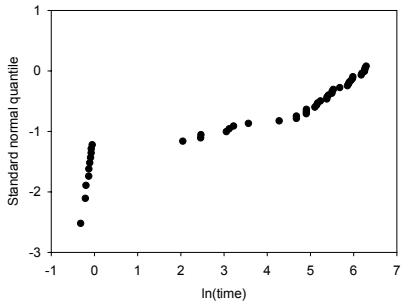
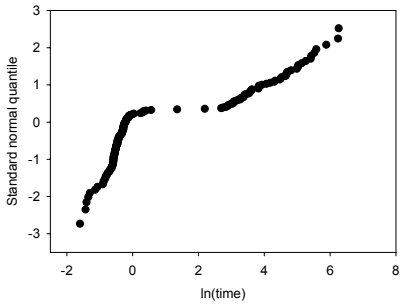


Figure 2.2: Lognormal QQ-plot of the interconnect sample.

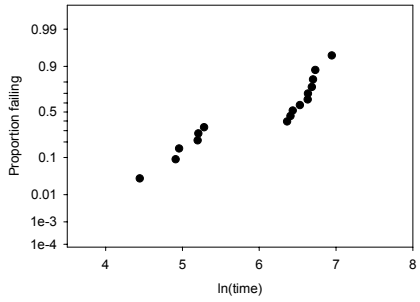


(a) Lognormal QQ-plot of the laser A sample.

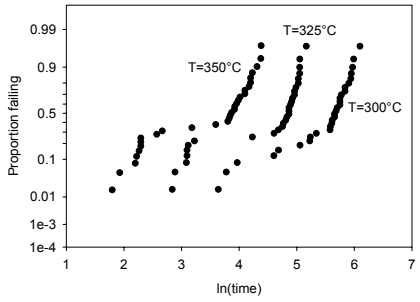


(b) Lognormal QQ-plot of the laser B sample.

Figure 2.3: QQ-plots of two laser samples.

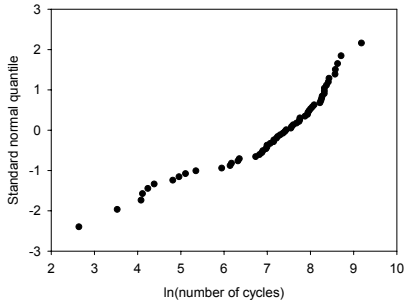


(a) Lognormal QQ-plot of the EM1 sample.

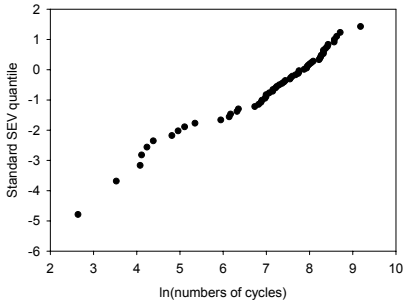


(b) Lognormal QQ-plot of the EM2 sample.

Figure 2.4: QQ-plots of electromigration samples.



(a) Lognormal QQ-plot.



(b) Weibull QQ-plot.

Figure 2.5: QQ-plots of the appliance failure sample.

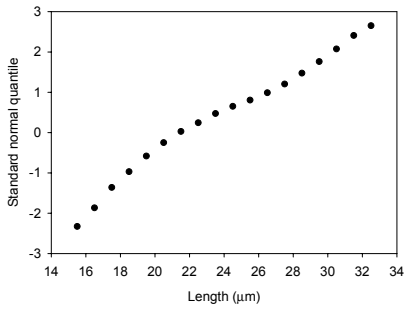


Figure 2.6: Normal QQ-plot of the Pearson sample.

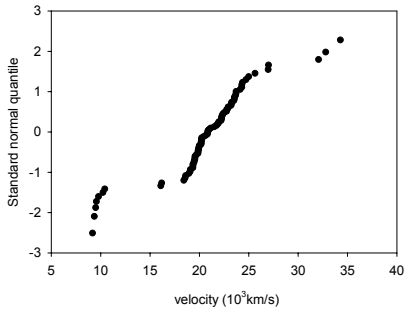


Figure 2.7: Normal QQ-plot of the galaxy sample.

2.2 Other multimodal samples

2.2.1 The Pearson sample

(taken from Everitt and Hand (1981, p. 46))

2.2.2 The galaxy sample

(taken from Aitkin (2001))

Chapter 3

Finite mixtures

So far, we related the finite mixture model to the distribution of multimodal failure time samples. In general, however, this model can be situated in a much broader context. Not only it is applied in many other situations, but also it belongs to the immensely rich class of mixture distributions. The latter exist in a wide range of forms, are used in a wide class of applications and although the related literature is huge already, it is still growing. Well-known examples of areas where mixtures arise, are random effects models, clustering, repeated measurement models, latent class models, empirical Bayes estimation, smoothing, ... An overview of the different situations leading to the use of a mixture distribution, can be found, for example, by Lindsay (1995, chap. 1).

With the article of Pearson in 1894, finite mixtures appear to be the first kind of mixture models encountered in practice. Given that finite mixtures can be regarded as the most natural derivation of a mixture model, this is not illogical. In particular, finite mixtures arise naturally in the context of modelling heterogeneous populations that can be subdivided into a finite number of homogeneous populations or *components* (Lindsay, 1995, p. 2). Today, finite mixtures are immensely popular and used in all kind of applications. Many examples of the latter are given in Titterington et al.

(1985). Literature concerning this topic is huge and covers a broad range of problems from which a lot of them are considered in Everitt and Hand (1981). A more up-to-date overview, with main focus on multivariate mixtures, can be found in McLachlan and Peel (2000).

In the following Section 3.1, some definitions and terminology concerning mixtures in general are given. The kind of finite mixture model, which will be of main interest in this work, i.e., the mixture adapted to the case of multimodal failure time samples, is presented in Section 3.2. Next, section 3.3 deals with the theoretical and numerical identifiability of finite mixtures. The concept of well or poorly separated mixtures, as well as their graphical representation, is discussed in Section 3.4. In the final section, we take a closer look at the maximum likelihood estimation of finite mixtures, in particular to the related problems of which some were indicated yet in Section 1.2.2.

3.1 Definitions and terminology

The density function $h(y)$ of a (univariate) mixture in its most general form is given by:

$$h(y; \mathrm{G}, \boldsymbol{\beta}, \boldsymbol{x}) = \int f(y|\boldsymbol{\phi}; \boldsymbol{\beta}, \boldsymbol{x}) \mathrm{dG}(\boldsymbol{\phi}). \tag{3.1}$$

The density $f(y)$ is called the *(mixture) component density* and can take on any form. Quite often it is a member of the exponential family, especially the normal distribution is frequently used. The parameter $\boldsymbol{\phi}$ is the *component parameter* or *latent variable* and can be a vector. The latter also holds for the parameter $\boldsymbol{\beta}$. This is a *nuisance* parameter, which is only optional. Further, a possible multivariate covariate $\boldsymbol{x}$ could belong to the model as well. The distribution G is referred to as the *mixing* or *latent* distribution. It can be considered as the distribution for the unobservable "random variable" $\boldsymbol{\phi}$. Depending on the specified distribution for G, several kinds of mixtures are

distinguished. If the distribution of G is not specified, i.e., if no parametric family for its distribution is assumed, the mixture model (3.1) without $\boldsymbol{\beta}$ is referred to as a *nonparametric mixture model* and with $\boldsymbol{\beta}$ as a *semiparametric mixture* model. These distributions lend themselves extremely well as a model for unobserved population heterogeneity (Böhning, 2000, Chap. 1) and also appear in the context of empirical Bayes estimation (Maritz and Lwin, 1989). In contrast, *parametric mixture* models are obtained when the distribution of G belongs to a certain parametric family. In case a continuous distribution is assumed, the mixture model is called a *mixed model*. These models are frequently used and arise, for example, in the context of models for longitudinal data (Verbeke and Molenberghs, 1997, Chap. 3), repeated measurements models (Davidian and Giltinan, 1995) and hierarchical Bayes models (Carlin and Louis, 1996). On the other hand, when G is a discrete distribution, the resulting mixture is a *finite mixture* distribution. As mentioned previously, such a model is appropriate when the heterogeneity present in the sample can be quantified into a (known) finite number of homogeneous subsamples.

Thus, for a finite mixture the mixing distribution G is a discrete probability measure. This means that the support of G corresponds to a limited number of points $\boldsymbol{\phi_j}$, $j = 1 \ldots M$, i.e., the different values that $\boldsymbol{\phi}$ can take on. These points are referred to as *support points*. Further, at each support point G puts *mass* π_j, i.e., $\mathrm{P}(\boldsymbol{\phi} = \boldsymbol{\phi_j}) = \mathrm{G}(\boldsymbol{\phi_j}) = \pi_j$. The latter are called *probability masses* or more commonly *proportion parameters*. The density $f_M(y)$ of an *M-component finite mixture* can then be written as:

$$\begin{aligned} f_M(y;\boldsymbol{\theta}) &= \sum_{j=1}^{M} \pi_j f(y;\boldsymbol{\phi_j},\boldsymbol{\beta}) \quad \text{with} \\ \boldsymbol{\theta} &= (\boldsymbol{\beta},\boldsymbol{\phi_1},\ldots,\boldsymbol{\phi_M},\pi_1,\ldots,\pi_M), \quad \sum_{j=1}^{M} \pi_j = 1. \end{aligned} \tag{3.2}$$

Generally, the component density $f(y)$ is a one or two parameter distribu-

tion. As a result, the support points $\boldsymbol{\phi_j}$ are parameters in case of a one-parameter component density or a two-parameter component density and a nuisance parameter β present. The latter will also be referred to as the *common* parameter among the mixture components. The support points will be a two-dimensional vector if there is no nuisance parameter and the component density has two parameters. If there is no common parameter β among the mixture components, the mixing is over all component parameters and the mixture will be termed a *general finite mixture*.

3.2 A model for multimodal failure time samples

Most (univariate) multimodal failure time samples have a number of characteristics in common. Of these, three important aspects are given below. They mainly describe the requirements for a heterogeneous failure time distribution.

1. For many reliability experiments carried out, reliability engineers generally know whether there is more than one failure mechanism involved. Mostly, however, the specific failure reason for each DUT is unknown, since at the end of the experiment it is either too difficult or too expensive (in terms of both money and time) to determine the failure reason of each DUT.

2. The subsamples of a multimodal failure time sample, representing the groups with a different failure behavior, are simply homogeneous failure time samples.

3. For many devices there is a priori no reason to assume a relation between the groups with a different failure behavior (for example, Joyce et al., 1976; Fisher et al., 2000).

The first condition implies the suitability of a finite mixture model (under the assumption that the different failure mechanisms are non competing).

The second suggests that a homogeneous failure time distribution, like the lognormal or Weibull, is most appropriate as component density. The last characteristic points to a general finite mixture, as there is no reason to assume a common (or nuisance) parameter among the mixture components. Apart from this, a failure time sample will often be censored in some way. Until now, finite mixtures are not much used in combination with a censored sample. As such, the methods considered in the following should also be applicable for censored samples.

In spite of the huge (statistical) literature dealing with finite mixtures, most models considered are not appropriate as a heterogeneous failure time distribution. Indeed, the finite mixture model which is most commonly used, is characterized through a normal component density and a common scale parameter σ among all components. The popularity of this model is partly due to the fact that estimation techniques, especially ML estimation, are simplified a lot by mixing only over one parameter instead of two (Section 3.5.2). Other important aspects are its increasing use in smoothing applications and its relation to the nonparametric maximum likelihood estimate (NPMLE). The latter is obtained as the MLE of a nonparametric mixture model. It has been proven that for the specific case of a nonparametric normal mixture model with common fixed scale parameter, a unique discrete NPMLE exist (Lindsay, 1983a,b). The resulting NPMLE is thus a finite normal mixture with common scale parameter.

Other (continuous) distributions that occasionally appear as component density, are often members of the exponential family and lately also of the t-distribution family. However, general mixtures with a two-parameter distribution as component density are not frequently used. Moreover, if the component density has two parameters, then often one of the two is common. This also holds in the few cases that a Weibull mixture is considered. There, either the shape parameter is taken to be common (Jewell, 1982) or fixed (Rider, 1962). Note that an exponential mixture is a Weibull mixture

with a common fixed (=1) shape parameter.

So, main interest is in general finite mixtures with a lognormal or Weibull distribution as component density. Nevertheless, throughout this work general normal and SEV mixtures will often be considered. Given the equivalence between a normal (SEV) mixture and a lognormal (Weibull) mixture, this will not change the objectives. Indeed, if the distribution of a random variable T is a finite lognormal (Weibull) mixture with density $\sum_{j=1}^{M} \pi_j f_{LN}(y; \eta_j, \beta_j)$, then the distribution of $Y = \ln(T)$ is a finite normal (SEV) mixture with density $\sum_{j=1}^{M} \pi_j f_N(y; \mu_j, \sigma_j)$ and $\mu_j = \ln(\eta_j)$ and $\sigma_j = 1/\beta_j$. To put it differently, through fitting a normal (SEV) mixture to a logarithmically transformed sample, a lognormal (Weibull) mixture is fitted to the untransformed sample. Therefore, all results obtained for normal and SEV mixtures do equally well hold for lognormal and Weibull mixtures. The preference for the normal mixture is logical seen the existing knowledge about this model. The choice for the SEV mixture is related to the fact that the relation between the SEV and Weibull mixture is similar to the relation between the normal and lognormal mixture. In addition, the SEV distribution is, like the normal distribution, a location-scale distribution. Unless stated explicitly, the general mixtures considered in the following will have a (log)location-scale distribution as component density. In addition, we will mostly refer to the location and scale parameters of the location-scale distribution, and not to the corresponding scale and shape parameter.

3.3 Identifiability

Before the estimation of any model can be considered, the question of identifiability has to be answered. Generally, for a mixture distribution, the aim is to identify the mixing distribution $\mathrm{G}(\boldsymbol{\phi})$ based on observations y from the mixture distribution with density $h(y)$. This problem is often regarded as a "missing data" problem since no realizations of $\boldsymbol{\phi}$ are observed,

but of the random variable Y (with mixture density $h(y)$). Given a mixture model, the question is then whether it is meaningful to search for an estimate of G. Information concerning the identifiability of nonparametric and finite mixtures can be found in Teicher (1961, 1963) and Yakowitz and Spragins (1968). Results about the identifiability of other mixtures can be found, for example, in Maritz and Lwin (1989, chap. 2). We briefly summarize the main results.

Definition (Teicher, 1961) A class of mixtures $\{h\}$, with respect to a certain family of component densities $f(\boldsymbol{\phi})$, $\boldsymbol{\phi} \in \Re^m$, induced by a set of distributions $\{\mathrm{G}\}$, is called *identifiable* if

$$h(x; \mathrm{G}_1) = h(x; \mathrm{G}_2) \Rightarrow \mathrm{G}_1 = \mathrm{G}_2, \quad \forall\, \mathrm{G}_1, \mathrm{G}_2 \in \{G\}. \tag{3.3}$$

Given that $\{G\}$ is the class of all possible distributions, then the class of nonparametric mixture models is identifiable in case the component density belongs to the (continuous) one-parameter exponential family or the two-parameter exponential family with a common, fixed scale parameter. However, this class is not identifiable in case the component density is a member of the two-parameter exponential family or a (log)location-scale distribution. Also the semiparametric mixture model is in that case not identifiable (Lindsay, 1995, pp. 53-54).

On the contrary, if $\{G\}$ is the class of all discrete distributions, i.e., $\{G\} = \bigcup_{k=1}^{\infty} G_k$ with G_k the class of distributions with a positive mass to exactly k points, then the following result can be derived from one of Teichers theorems (Teicher, 1963): the class of general finite mixtures with a (log)location-scale distribution as component density is identifiable. Consequently, the mixture models of main interest in this work are identifiable.

Still, this *theoretical* identifiability will not always be sufficient to obtain meaningful estimates. For some samples, the ML estimation problem is *numerically* not identifiable. In other words, the sample size is too small

or the sample contains not enough information to distinguish one particular model of a certain family. While this is mostly not an issue for the estimation of simple one or two-parameter distributions, this is an important issue when considering the maximum likelihood estimation of general finite mixtures. Throughout this work, this topic will return several times.

3.4 Identifying the mixture components

Although (general) finite mixtures are identifiable, not all finite mixtures are clearly recognizable as a mixture distribution. Specifically, for some mixtures it will be easy to identify or recognize its different mixture components, while for others this will not be the case. For mixtures with a common scale parameter this comes down to how well the mixture components are separated in location. In this section, we discuss how this concept can be adapted to the case a of general finite mixture and which graphical tools will be considered to recognize general mixtures, in particular two-component mixtures.

3.4.1 Separation of the mixture components

For normal mixtures with a common scale parameter, the shape of the density function, in particular its number of modes, depends highly on how well the components are separated in location, i.e., how far the different location parameters are situated off each other. For example, for the two-component normal mixture with common scale parameter σ and proportion parameter equal to 0.5, the following subdivision exists (McLachlan and Peel, 2000, pp. 9-10):

- $\Delta = \frac{|\mu_1 - \mu_2|}{\sigma} > 2$. The density function is bimodal. The larger the value for Δ is, the more pronounced the bimodality will be and the better the components can be identified from the mixture (based on

the density plot). Mixtures with a clear bimodal density function are referred to as *well separated* mixtures.

- $\Delta = 2$. The density function has a kind of plateau, i.e., the only maximum value of the density function is reached in several adjoining points.

- $\Delta < 2$. The density function is unimodal. It is not possible to identify the two components of the mixture based on the density plot, or to even identify the distribution as being a mixture. These mixtures are often referred to as *poorly separated* mixtures.

In case of an unequal proportion parameter, the mixture density will no longer be symmetrical, but for each value of π_1 a similar scheme can be obtained with another borderline value for Δ. For example, for $\pi_1 = 0.2$, a sufficient condition to have a unimodal mixture density, is a value of Δ smaller than about 2.7. Further, this concept can be easily adapted to the case of an M-component normal mixture with common scale parameter. Also, it can be extended to any location-scale distribution, like the SEV distribution, as component density. So, M-component mixtures with a common scale parameter have clearly identifiable components if the location parameters are sufficiently different (with respect to the size of the common scale parameter). These mixtures are referred to as well separated mixtures. They can be recognized through their density function which has M modes.

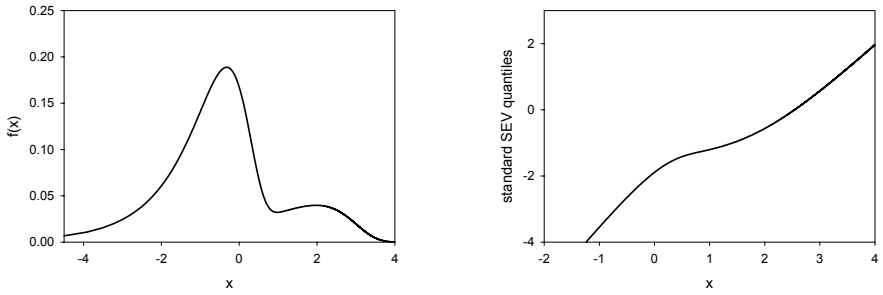
A generalization of this concept to the situation of general finite mixtures is not straightforward. Two things are involved. First, there is the fact that the modality of the density function cannot be reduced anymore to the value of one single quantity Δ. The ratio of the different scale parameters becomes also important. Although there have been some attempts to derive some rules, a general rule is difficult to obtain. A small overview can be found in Everitt and Hand (1981, pp. 27-30). One of the sufficient conditions to obtain a unimodal density for a two-component normal mix-

ture, irrespective of the value of π_1, is given trough $|\mu_2 - \mu_1| \leq 2\min(\sigma_1, \sigma_2)$ (Behboodian, 1970). A necessary condition, however, depends also on the value of π_1 and its relation with $\min(\sigma_1, \sigma_2)$. Second, the separation of mixture components cannot be related anymore merely to the shape of the density function or to the difference between the location parameters. If the mixture components have similar location parameters, but scale parameters which are sufficiently different, then the mixture components are still clearly separated. However, the separation is in "scale" and not in location. Obviously, the mixture density will be unimodal, highly skewed and in addition, difficult to recognize as a mixture on a density plot. This kind of separation, although often ignored, can equally well lead to mixtures with clearly identifiable components (Section 3.4.2).

In the following, we will refer to poorly or well separated mixtures, when the component mixtures are clearly separated in location, in scale or in both. The larger the difference between the location parameters or scale parameters, the better the mixture components are separated or can be identified from the mixture. Note that the value of the proportion parameters will also have an influence on how well the mixture components are separated. For some values, the location parameters or scale parameters have to be further apart than for other values, to obtain a well separated mixture.

3.4.2 Graphical representation of mixtures

When mixture distributions are considered, a histogram is one of the graphical tools most frequently used to visualize a sample. This graphic can be regarded as an empirical counterpart of a density plot. Still, it is rather difficult to use it for the recognition of samples with a mixture distribution. There are two main reasons involved. On the one hand, the density function of a mixture, in particular a general mixture, will not always reveal the presence of a mixture, i.e., its density function will not always contain as many modes as components. Nevertheless, the appropriateness of a mixture



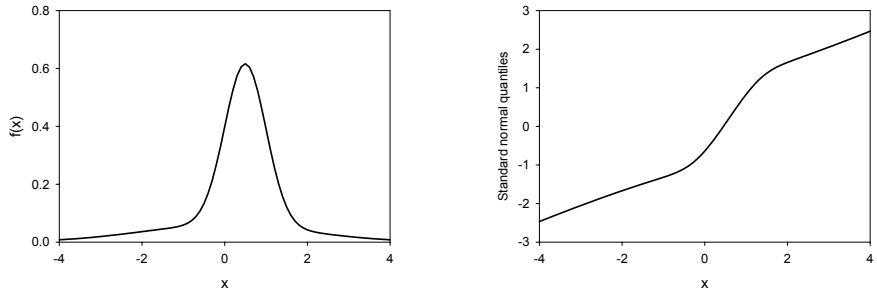
(a) Density plot.

(b) Cdf plot on SEV probability scales.

Figure 3.1: A well separated two-component SEV mixture. Parameter values are $\pi_1 = 0.2,\ \mu_1 = 0,\ \sigma_1 = 0.5,\ \mu_2 = 2.7,\ \sigma_2 = 0.67$

distribution for a sample will be judged on the apparent number of modes of a histogram. On the other hand, it has been indicated already several times that the number of apparent modes in a histogram highly depends on the number of bins or classes used (Everitt and Hand, 1981, p. 109; Haughton, 1997; Aitkin, 2001). As such, by changing the latter, the number of apparent modes of a histogram can be easily changed.

For two-component (general) mixtures, Fowlkes (1979) indicates that there do exist other kind of plots which make it possible to detect more easily the presence of a general mixture distribution. One of them is the quantile-quantile or QQ-plot (Section 5.2.1). These plots are often used in a reliability analysis to depict a sample. Their main feature is that they allow the recognition of many distributions which are based on a (log)location-scale distribution. This also holds for two-component general finite mixtures with a (log)location-scale distribution as component density. Although the detection of mixtures with more than two components is more difficult, these plots still allow an easy comparison between the observed sample and the



(a) Density plot.

(b) Cdf plot on normal probability scales.

Figure 3.2: A well separated two-component normal mixture. Parameter values are $\pi_1 = 0.3,\ \mu_1 = 0,\ \sigma_1 = 2,\ \mu_2 = 0.5,\ \sigma_2 = 0.5$

fitted distribution. Throughout this work, we will mostly use this QQ-plot to depict the samples.

The theoretical counterpart of a QQ plot, is a plot of the cumulative distribution function (cdf) of a distribution on certain probability scales (Section 5.2.1). For a mixture distribution, the scales used are such that the plot of the cdf of its component distribution is a straight line. For example, normal (SEV) probability scales, are given by x for the x-axis and $\Phi^{(-1)}(y)$ $(\ln(-\ln(1-y)))$ for the y-axis. For lognormal or Weibull probability scales, the scale on the x-axis is $\ln(x)$. Based on the shape of the cumulative distribution function on these scales, two-component mixtures can be classified as being well or poorly separated. In particular, the more its curve deviates from a straight line, the better the components of the mixture are separated or the better the mixture components can be identified. As an illustration, Figure 3.1 shows the density plot and the cdf plot of a two-component SEV mixture with components well separated in location, while Figure 3.2 gives the same plots for a two-component normal mixture with

components well separated in scale. In Section 5.2.1 we will discuss how for (well separated) two-component mixtures, the different components can be identified from this kind of plots. To recognize well separated M-component mixtures, each couple of two components has to be considered separately as a two-component mixture. When each of those mixtures is well separated, the M-component mixture will be well separated too (Section 5.5.2).

3.5 Maximum likelihood estimation of general finite mixtures

Through the years, a considerable number of estimation techniques are developed. Most of them also found their way to finite mixture models. Before the advent of the computer in the early nineties, usually either graphical techniques or the method of moments were applied as estimation method for finite mixtures. Examples of graphical methods are given by Harding (1948) and Preston (1952) for the two-component normal mixture model, Kao (1959) for a two-component Weibull mixture and Bhattacharya (1967), for grouped samples of a normal mixture model. The method of moments was probably first used by Pearson (1894) for a two-component general normal mixture. This method of moments was adopted by many, but often with the restriction of a common scale parameter to simplify the estimation procedure. A concise overview of the evolution of this method to finite normal mixtures can be found in Redner and Walker (1984). Further, it was used by Rider (1962) for a Weibull mixture with a common shape parameter. From about the early sixties on, the maximum likelihood (ML) appeared in literature as an estimation method for finite mixtures. Probably, Rao (1948) was one of the first using the ML method for the estimation of a two-component normal mixture with common scale parameter. From then on, the ML method was a preferred estimation method, mainly because of the apparent superior properties of its estimators compared to the esti-

mators of both graphical methods and the method of moments. Nowadays, it is still a popular estimation tool together with Bayesian estimation (for example, Redner et al., 1987; Aitkin, 2001). Sometimes, the two methods are even combined (Aitkin and Rubin, 1985). More information about these and other estimation methods for finite mixture models, in particular normal mixtures, can be found in Everitt and Hand (1981) and Redner and Walker (1984). Here, we focus on the ML estimation of general finite mixtures.

In the following Section 3.5.1, the ML estimation of finite mixtures is briefly handled. Some of the main methods to obtain the ML estimates are introduced. In Section 1.2.2, we pointed to some problems when ML estimation is applied using some reliability software. Apparently, these problems are no coincidence. Moreover, although the ML estimation of finite mixtures seems feasible, without too many difficulties, this does not hold for the ML estimation of general finite mixtures with a (log)location-scale distribution as component density. There are two serious problems involved, which make up the main topics of this project. They are introduced in Sections 3.5.2 and 3.5.3. As a result of these problems, a general framework for the ML estimation of general finite mixtures is still lacking, in spite of the fact that the estimation of the parameters of a general two-component normal mixture is one of the oldest problems in statistical literature.

3.5.1 Calculation of the maximum likelihood estimate

One of the main reasons for the popularity of the maximum likelihood method are the good statistical properties of its estimators. In particular, under suitable regularity conditions, maximum likelihood estimators (MLEs) are *consistent* and *asymptotically efficient* and *normally distributed*. In the classical sense, the MLE is obtained as the global maximum of the likelihood function. For many distributions among which a finite mixture model, the maxima of the likelihood function can be obtained through solving the likelihood equations (LEQs). Specifically, given a (complete) sample

$\boldsymbol{y} = (y_1, \ldots, y_n)$ of sample size n, from a finite mixture with density (3.2), then the likelihood function $L(\boldsymbol{\theta}; \boldsymbol{y})$ is given by:

$$L(\boldsymbol{\theta}; \boldsymbol{y}) = \prod_{i=1}^{n} f_M(y_i; \boldsymbol{\theta}). \tag{3.4}$$

The LEQs are derived through equating to zero the partial derivatives of the log likelihood function (i.e., $\ln L(\boldsymbol{\theta}; \boldsymbol{y})$):

$$\left.\begin{aligned} \frac{\partial \ln L(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \theta_1} &= 0 \\ &\vdots \\ \frac{\partial \ln L(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \theta_p} &= 0 \end{aligned}\right\} \quad \text{LEQs}, \tag{3.5}$$

with p the number of parameters of the mixture model. Further, the covariance matrix of the asymptotic normal distribution of the MLE $\hat{\boldsymbol{\theta}}$, is estimated by:

$$\widehat{\Sigma}_{\hat{\boldsymbol{\theta}}} = \widehat{I(\boldsymbol{\theta})}^{-1} = \left[\sum_{i=1}^{n} \frac{\partial^2 \ln L(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \right]^{-1}, \tag{3.6}$$

with $\widehat{I(\boldsymbol{\theta})}$ the observed Fisher information matrix.

While for some distributions, like the members of the exponential family, a closed form solution exists for the LEQs, mostly for finite mixtures they will have to be solved by means of an iterative procedure. Such a method starts from an initial guess or *starting value* of the unknown parameters, and updates these values during a number of cycles until convergence. The most popular and according to simulations (Everitt, 1984) also the most adequate methods for mixtures, are the Expectation-Maximization(EM)-algorithm and the Newton-Raphson(NR) method, as well as some of its variations.

The EM-algorithm of Dempster et al. (1977) is developed to handle maximum likelihood estimation in case of missing data problems. Each

cycle of the iterative procedure consists of two steps: an expectation step, calculating the expected complete data likelihood, given both the observed data and the current parameter values, and a maximization step, maximizing this expected complete data likelihood to obtain updated parameter values. Given the fact that the estimation of a finite mixture distribution can be viewed as a missing data problem (Section 3.3), the EM-algorithm is quite suitable to handle ML estimation of finite mixtures. A detailed discussion about the aspects of the EM-algorithm in case of finite (normal) mixtures can be found in Redner and Walker (1984). Note that before the introduction of the EM-algorithm in 1977, this iterative procedure was already derived by some authors for the finite mixture problem without regarding it as a missing data problem. Namely, Hasselblad (1969) for finite mixtures with a member of the one-parameter exponential family as component density, and Wolfe (1970) and Day (1969) for (general) finite normal mixtures, worked out, amongst others, an iterative procedure based on the LEQs that appeared to be the EM-algorithm.

In contrast to the EM-algorithm, the NR method is a general iterative procedure intended to solve any set of equations. For a simple one-parameter equation $g(x) = 0$, one cycle into the NR procedure takes the form $x_{k+1} = x_k - \frac{g(x_k)}{dg(x_k)/dx}$. While for this single equation, $\left(\frac{dg(x)}{dx}\right)^{-1}$ is usually easy to obtain, for a multi-parameter problem this involves the inversion of a matrix. Variations of the NR-method, quasi-Newton methods, are intended to simplify the calculation of this matrix inversion. More information on the NR-method and some of its variations, in case of finite normal mixtures, can be found in Everitt (1984) and Redner and Walker (1984).

It will be noted that within this project the EM-algorithm is often preferred to the NR-method. The reasons for this will be explained where necessary, but are mainly related to the instability of the NR-method and the monotonicity property of the EM-algorithm. We implemented a version of the EM-algorithm in the statistical language GAUSS. For the use of the

NR-method, we considered the package CML of GAUSS. Hereby, the CML procedure is used with an analytical description of the gradient function.

3.5.2 The nonexistence of the classical maximum likelihood estimate

In spite of the fact that the LEQs, previously given, can be solved for most finite mixtures, the classical ML method seems to break down for general finite mixtures with a (log)location-scale distribution as component density. Moreover, for these mixtures the likelihood function is unbounded at some points, also referred to as *singularities*, on the edge of the parameter space. As a result, a classical MLE, defined as the global maximum of the likelihood function does not exist. For example, take $(y_1, \cdots, y_n)$, a sample from a two-component normal mixture with likelihood $L(\boldsymbol{\theta}, \mathbf{y})$ given by:

$$L(\boldsymbol{\theta}, \mathbf{y}) = \prod_{i=1}^{n} \left\{ \frac{\pi_1}{\sqrt{2\pi}\sigma_1} \mathrm{e}^{\left[-\frac{1}{2}\left(\frac{y_i-\mu_1}{\sigma_1}\right)^2\right]} + \frac{(1-\pi_1)}{\sqrt{2\pi}\sigma_2} \mathrm{e}^{\left[-\frac{1}{2}\left(\frac{y_i-\mu_2}{\sigma_2}\right)^2\right]} \right\}. \qquad (3.7)$$

It is then easily seen that the likelihood goes to infinity whenever $\mu_1 = y_i$ and σ_1 approaches zero, with the other parameters having arbitrary values. Clearly, these "maxima" are pathological and do not correspond to useful mixtures. Moreover, they are inconsistent estimates and cannot be regarded as maximum likelihood estimates, since due to its unboundedness, the likelihood function does not have a global maximum (Lehmann, 1980).

The main problem, however, for the ML estimation of general finite mixtures is not this nonexistent MLE, but the fact that there is a huge difference in approach and no ambiguity in the way this ML problem is tackled. For example, constraints on the parameters are incorporated, the construction of the likelihood function is adapted, other existing (local) maxima of the likelihood function are chosen as estimate, ..., but rarely methods are questioned or compared to each other. In Chapter 4, we not only try to clarify this situation, but also give our point of view. Among other things,

we will discuss and compare some important different approaches, look at the behavior of the largest local maximum and discuss the importance of considering the surface of the likelihood function.

3.5.3 The multiple root problem of the likelihood equations

It is well-known that the likelihood function of a mixture can have multiple maxima. In particular, the LEQs for a general finite mixture model usually have a lot of roots. The reason for this is related to the specific nature of a general mixture model, as it models groups within a sample whether these groupings are "real" or purely random. Nevertheless, in most cases, interest is only in one specific root of the LEQs, which will often be the root corresponding to the global or largest local maximum of the likelihood function. However, most iterative procedures used to solve the LEQs have no guaranteed global convergence. Moreover, it appears that the root obtained highly depends on the starting values used. This explains the situation encountered previously with the two (reliability) software packages: other starting values were used, which resulted in different maxima.

As such, the choice of the starting values, in case of a general finite mixture, primarily determines whether the intended maximum will be identified or not. In spite of this, only relatively little research efforts have been devoted to the search for good starting values. Main focus remains directed towards the improvement of the iterative procedures, especially the EM-algorithm (Ueda et al., 2000; Celeux et al., 2001). We believe, however, that a lot can be gained already when started with good, well-reasoned starting values. As discussed in Chapter 5, they not only improve the performance of iterative procedures, but also make simulations and bootstrap procedures feasible (in terms of both time and unambiguity) and importantly make it possible to fit mixtures in real terms (software, industry). In chapter 5, we introduce a starting value method that allows a well-founded (maximum) likelihood estimation of general finite mixtures in practice.

Chapter 4

Likelihood estimation of general finite mixtures

The aim of this chapter is to clarify the situation concerning the maximum likelihood estimation of general finite mixtures. As discussed, the latter is problematic. Although in literature some solutions are proposed, there is no general agreement of how to deal with it. There is too much ambiguity in the approach of obtaining "adapted" ML estimates. The idea is to set up a framework around the ML estimation of general finite mixtures. This is done through investigating and comparing some existing techniques, whether or not yet accepted. Quite likely, the solutions given will not be the only ones, but at least the ones considered are checked in detail and their pros and cons are known. It allows, for the first time, a sensible approach for the estimation of general finite mixtures, based on the likelihood function.

The two main problems for the ML estimation of general finite mixtures with a (log)location-scale distribution as component, are clear:

1. A classical MLE, i.e., the global maximum of the likelihood function, does not exist due to the unboundedness of the likelihood function.

2. The LEQs contain usually a large number of roots.

Although at first sight, the second problem is only a technical one, related to the calculation of the MLE, it is nevertheless the direct cause of the presence of the so-called spurious maxima. It will be handled in Section 4.3.

The main core of methods, proposed in literature to satisfactorily tackle the unbounded likelihood problem, try to regularize the problem by removing the unboundedness. Either the likelihood function is adapted or the parameter space is restricted. While the latter is frequently used, we believe that it suffers from some important drawbacks. Section 4.1 reviews and discusses some standard methods to deal with an adapted or restricted ML estimation of general finite mixtures. An alternative method, ignoring the unboundedness of the likelihood function and referred to as *likelihood estimation*, is based on the fact that there exists a local maximum of the likelihood function with good statistical properties. This theory is rooted in the literature, acknowledged by some authors, but still not often applied. By going back to Cramér, we will review in Section 4.2 that well-behaved estimates as a solution of the likelihood equations (LEQs) do exist for general finite mixtures, despite the non-existence of the MLE. In particular, it will be discussed that for the mixtures considered in this project, the largest local maximum of the likelihood corresponds to these well-behaved estimates. In Section 4.4, we have a closer look at some sample properties of the largest local maximum.

In spite of this, not everyone agrees on the use of the latter: McLachlan and Peel (2000), amongst others, argue that one should first skip some spurious maxima before selecting the largest local one. Although, such spurious maxima are indeed an issue, the way they are handled so far, seems to be flawed. In addition, it is often overlooked that they also appear for most other adapted ML methods. The result aimed at here is to put an end to the myth that the likelihood estimate, defined as the largest local maximum of the likelihood, cannot be used. In Section 4.3, we give our perception on this problem, explain the presence of spurious maxima, relate it to the sample

size and importantly gives some guidelines of how to deal with it in practice. Finally, in Section 4.5, two estimation techniques, based on the ML method, are compared with respect to what really concerns, namely inference. In particular, three adapted likelihood estimators and the likelihood estimator are compared.

4.1 Removing the unboundedness

The unboundedness of the likelihood function, in case of a general finite mixture, is the immediate cause of the non-existence of a global maximum, and so also of the (classical) MLE. As a result, many techniques proposed to overcome the problem of a nonexistent MLE, try to remove in some way this unboundedness in order to still obtain a, perhaps modified, MLE. The most common approaches are based either on adapting the likelihood (Section 4.1.1) or on restricting the parameter space (Section 4.1.2).

4.1.1 Adaptation of the likelihood function

Cox and Hinkley (1974, Chap. 9) pointed out that the anomaly of an infinite likelihood function would disappear if one would take into account the inherent grouped nature of the data. In practice, all observations are discrete and therefore a continuous model is only a theoretical concept. Similarly, Aitkin (2001) states that the unboundedness of the likelihood arises from its approximation to the actual grouped data likelihood.

From their point of view, the infinity problem results from a misspecification of the likelihood. As such, the problem could then be solved through a more principled construction of the likelihood function. It should be built as:

$$L(\boldsymbol{\theta}, \mathbf{y}) = \prod_{i=1}^{n} \left[F_M(y_i + \delta/2) - F_M(y_i - \delta/2) \right], \tag{4.1}$$

with $F_M(x) = \sum_{m=1}^{M} \pi_m F(x|\mu_m, \sigma_m)$, the cumulative distribution function (cdf) of the mixture, $F(x)$ the cdf of the mixture component and δ the grouping interval or the measurement instrument's precision with which y_i is measured. As a consequence, this likelihood is bounded between 0 and 1. Moreover, if a global maximum exists, it corresponds to a consistent MLE. Note that, given a value for δ, the observations y_i are often recorded with an accurateness larger than allowed by δ. Usually, these observations are used in the likelihood function (4.1) without adapting them. We take the view that this way of handling should be avoided. Instead, we prefer one of the two following options:

1. The observations y_i are rounded off according to δ (i.e., to $y_{i\delta}$) and used in (4.1) instead of y_i. The resulting estimator will be referred to as MLEδ.

2. The likelihood function, defined in (4.1), is replaced through

$$L(\boldsymbol{\theta}, \mathbf{y}) = \prod_{i=1}^{n} \{F_M(\lceil y_i \rceil) - F_M(\lfloor y_i \rfloor)\}, \tag{4.2}$$

   with $\lceil y_i \rceil$ ($\lfloor y_i \rfloor$) the observation y_i rounded up (down) to the precision δ. The estimator will be referred to as MLEδ^*.

Another approach, based on the same argument that all observations are discrete, is to bin the sample into a number of classes (m_c), mostly with equal width h (McLachlan and Peel, 2000, p. 101). The likelihood function is then defined as

$$L(\boldsymbol{\theta}, \mathbf{y}) = \prod_{j=1}^{m_c} [F_M(b_j) - F_M(a_j)]^{n_j}, \tag{4.3}$$

with a_j (b_j) the lower (upper) limit of the j^{th} class with $b_j - a_j = h$ and n_j the number of observations within the class. Again, the likelihood function is bounded and a consistent MLE (referred to as MLE$_b$) is obtained if a global maximum exists.

Apart from the fact that numerically, it can be demanding to manipulate a likelihood composed of differences of cumulative distributions instead of densities, these adapted methods seem to be useful alternatives for the (classical) ML method. Still, there are some drawbacks involved and some important comments are in place. First, some authors state that the argument of discrete data does not necessarily get to the real issue. "Whether or not it is possible in practice, it is still legitimate to suppose that the observations are intrinsically continuously distributed and that discreteness is the approximation" (Cheng and Iles, 1987, p. 98). Further, the original likelihood (in the continuous case) was composed of density contributions $f(x;\theta)$, derived from probability elements $\mathrm{P}(x \in dx) = f(x,\theta)dx$ (Cramér, 1946, chap. 32), obviating the need to be bounded above by 1. Also, despite the fact that the infinite spikes of the likelihood do not yield useful estimates, the infinity is not counter-intuitive. Indeed, if one of the variances goes to zero, the corresponding component of the mixture becomes discrete with a contribution to the joint distribution that will be "infinitely" greater than a continuous one (e.g., you cannot better fit a point than by assigning the entire mass to it).

Second, even if one considers (4.1) (or (4.2)) as the correct specification of the likelihood, how should one then choose the precision δ? In rare cases, this value is known as being the precision of the measurement system. But for most cases, the value of δ is unknown and one would be unable to choose it without an unacceptably high amount of arbitrariness. Similarly, this holds true for the number of classes m_c in case a binned likelihood is assumed. Nevertheless, the parameters δ and m_c can be regarded as smoothing parameters. From this point of view, these methods could be valuable in a kind of sensitivity analysis (Section 4.5).

Third, the global maximum of the likelihood function, irrespective of the adapted form used, does not always exist. In particular, the adapted likelihood function is discontinuous, like the classical likelihood function,

at some points just outside the parameter space (for example at $\sigma_1 = 0$, $\mu_1 = y_1 - \delta/2$ and the other parameters having arbitrary values). Due to these discontinuities, an adapted likelihood function sometimes attains its supremum in a point situated outside the parameter space. A supremum, however, which is not a global maximum since this point does not belong to the parameter space. As such, an adapted MLE does not exist either. Basically, this problem is similar to the unbounded likelihood problem. In Sections 4.2 and 4.5.3, we come back on this issue. Examples are given in Sections 4.3 and 4.5.3.

In summary, adapted MLEs could be an option, but in essence they mostly suffer from the same problem as the classical MLE. These methods are not equivalent to classical ML estimation but to ML estimation adapted for grouped data. A distinction which also exist for other distributions.

4.1.2 Restriction of the parameter space

Another way of bounding the likelihood is by restricting the parameter space. The singularities of the likelihood are situated on the edge of the parameter space. Hence, by constraining the latter such that problematic points are excluded, a bounded likelihood over the restricted space can be obtained. However, one still has to prove the existence of a consistent MLE for this kind of restricted likelihood problems.

One of the most popular forms of constraints, although it is often not recognized as such, is the imposition of an equal scale parameter among the different components of the mixture. In this case, the singularities of the likelihood $L(\boldsymbol{\theta}, \mathbf{y})$ disappear (if the sample has a size larger than 1 and not all values are equal), the likelihood becomes bounded and a consistent MLE exists (Everitt and Hand, 1981). But, although this restrictive assumption may be justified in some cases, in general it not a satisfactory solution (Fisher et al., 2000; Joyce et al., 1976).

For a general mixture model, one can prevent the scale parame-

ters of the different components to become zero, by interrelating them. In this way, a constrained parameter space without singularities is obtained. Moreover, Quandt and Ramsey (1978), amongst others, noted the existence of a consistent MLE in case a relationship between the standard deviations of the true mixture normal component densities was known and incorporated as constraints. A possibility, for example, are constraints of the form $\sigma_i = k_{ij}\sigma_j$, with k_{ij} known constants. Exact knowledge of the constants, however, is rare.

Alternatively, inequality constraints can be imposed as is done by Hathaway (1985). He introduced the following inequalities:

$$\sigma_i \geq c\sigma_{i+1}, i = 1, \ldots, (M-1); \sigma_M \geq c\sigma_1, \text{with } c \in]0, 1]. \tag{4.4}$$

For this restricted likelihood problem in case of a normal mixture model, Hathaway proved the existence of a global maximum of the likelihood regardless of the value of c. Further, he showed the consistency of such a global maximum if the constrained parameter space contained the true parameter. In other words, a consistent MLE exists if the true parameter is in the restricted parameter space. Hathaway and Bezdek (1986) also adapted the EM-algorithm to incorporate restrictions.

Another way to restrict the parameter space and exclude singularities is to work directly with compact subsets. For these likelihood problems, Redner and Walker (1984) proved the existence, as sample size goes to infinity, of a consistent MLE for the normal mixture problem over any compact subspace containing the true parameter.

In spite of the fact that most of the methods mentioned are reasonable, we do not consider them as an option. The major problem with this kind of approach are the restrictions imposed on some of the parameters, while a priori for most problems the parameters of a general finite mixture can take on any value. Also, all results concerning the consistency of the MLE are based upon the assumption that the constrained parameter space

contains the true parameter, although the latter is unknown. So, the choice of the value of c and the choice of the compact subset, without knowledge of the true parameter, is rather problematic. In addition, often restrictions are too limiting. For example, imposing equality of the scale parameter is an easy way to proceed, but in a lot of cases it is implausible. For the estimation of a general finite mixture model, restricting the parameter space is essentially circumventing the real problem.

4.2 An alternative: likelihood estimation

One of the main reasons for the popularity of the maximum likelihood method are the good statistical properties of the corresponding estimators. Namely, they are consistent, asymptotically efficient and asymptotically normally distributed under suitable regularity conditions. Although the classical MLE does not exist for the general finite mixture model, as noted before its likelihood function has many local maxima. Often, one of these maxima and in particular the largest local one, was considered instead, as it was understood to have the same properties as the MLE. In the course of time, evidence appeared in literature for the normal mixture model, justifying the approach tacitly followed. This provided a way to avoid the search for a global maximum. First, empirical evidence was found, for example by Quandt (1978) and Duda and Hart (1973), that a local maximum, more specifically the largest one, corresponds to reasonable parameter estimates. Later, Sundberg (1974), for incomplete data from an exponential family and Kiefer (1978), for a switching regression model and Lehmann (1983), for general situations, provided a solid basis for such an approach. They all proved the existence of a consistent sequence of roots of the likelihood equations for their particular problem or provided some regularity conditions.

In what follows a review of this theory is given (Section 4.2.1), together with our perception of how it can be used to obtain parameter

estimates with good statistical properties in case of general finite mixtures with a (log)location-scale distribution as component density and without any restriction on the parameters (Section 4.2.2).

4.2.1 Review

Cramér (1946, chap. 32-33) discussed the method of maximum likelihood for one-parameter distributions. Although he first defines the MLE as the value which renders the likelihood as large as possible, his final definition of an MLE is different from the classical one. Moreover, he states: "**Any** *solution of the likelihood equation will be called a maximum likelihood estimate of the unknown parameter*". With this definition in mind, he proves that under certain general conditions as the sample size goes to infinity the likelihood equation has **a** (but not any) solution that converges in probability to the true parameter value, hence is consistent. Further, this solution is also asymptotically efficient and normally distributed. In other words, Cramér proved the existence of a solution of the likelihood equation with good statistical properties and called it an MLE.

In 1948, Huzurbazar showed, under the same conditions as Cramér, that with probability going to one as the sample size goes to infinity, such a consistent root is unique and corresponds to a local maximum of the likelihood. Thus, if a density satisfies the conditions of Cramér, a local maximum of the likelihood exists , which possesses the required statistical properties. This result provides a useful alternative to the condition of global maxima only.

Wald (1949) gave a proof of consistency of the classical MLE, i.e., with the usual meaning attached to the global maximum of the likelihood. His proof was based on totally different and more demanding assumptions compared to Cramér's conditions. In essence, Wald does not use differentiability assumptions; even the LEQs do not have to exist. In addition, Wald notes that Cramér is only proving the consistency of a local maxi-

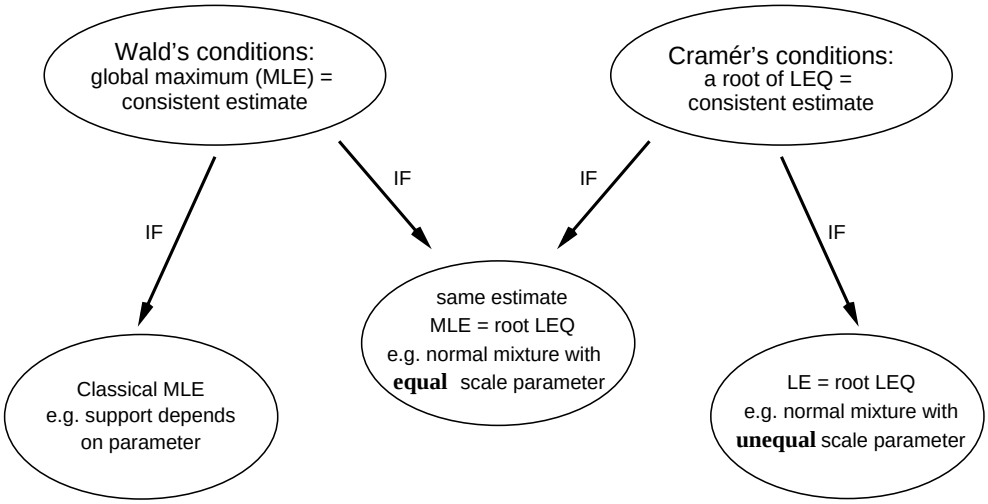


Figure 4.1: Likelihood estimation

mum, in contrast to his proof of the consistency of a global maximum. A concise overview of both sets of conditions can be found in Section A.1 of the appendix. Note that Cramér's results were for the one-parameter case (in contrast to Wald). Aitchison and Silvey (1958) generalize his results to the multi-parameter case, whereas Chanda (1954) (proven by Tarone and Gruenhage 1975) extends the uniqueness theorem of Huzurbazar. The conditions are straightforward extensions of the one-parameter case.

Nowadays, the classical meaning of the MLE is still used in combination with the conditions of Cramér. Sometimes, as is the case for general finite mixtures, this leads to the problem of a seeming failure of the maximum likelihood method. Nevertheless, there would be no problem if the correct conditions would be considered. Figure 4.1 gives an overview of these different approaches of (maximum) likelihood estimation. On the one hand, there are the conditions of Wald ensuring the consistency of a global maximum, if this maximum exists. On the other hand, there are the conditions of Cramér guaranteeing the existence of a consistent local maximum. Depend-

ing on the assumptions a parametric family fulfills, we have the following three possibilities:

Both conditions hold. They refer to the same estimate, i.e., the global maximum. It can be found as a solution of the LEQs. This is the case for a lot of two-parameter distributions such as the normal, Weibull, gamma, ..., but also for a finite mixture with common scale parameter.

Only Wald's conditions hold. If the global maximum exists, the classical MLE is consistent. This is typical for distributions that do not have a derivative at some points in the parameter space. An example is when the support depends on some parameter, like the uniform distribution.

Only Cramér's conditions hold. Even if the global maximum exists, it does not necessarily have good statistical properties, but at least one local maximum does. It corresponds to a root of the LEQs. This is often the case for distributions with singularities situated on the edge of the parameter space, such as for many general M-component mixtures.

When applied to the problem at hand, it appears that the conditions of Cramér are fulfilled for most general finite mixtures with a (log)location-scale distribution as component density. In particular for normal mixtures, it has been shown by Sundberg (1974) and Kiefer (1978), that this parametric family satisfies the necessary conditions. A comparable proof can be used, to demonstrate that the conditions are also fulfilled for SEV (or Weibull) mixtures. No additional difficulties are encountered when considering an adapted likelihood function instead of the classical density likelihood. In the same way, it can be shown that also in this case the conditions of Cramér are satisfied (if minimum 3 adjacent intervals are available for each component of the mixture).

Already in 1956, Kiefer and Wolfowitz pointed out that a consistent MLE did not exist for the general normal mixture model. They also indicated that this was not only due to the "technical" problem of a nonexistent global maximum. The problem was more profound as also the conditions of Wald were not satisfied. In particular, there are problems with the integrability assumption (Section A.1). The same holds true for general SEV mixtures. But, while it is taken for granted that problems are solved through a discretization or grouping of the sample, this is not necessarily the case. It is not because the likelihood becomes bounded that a global maximum or even a consistent MLE would exist. Apparently, for the adapted likelihood functions, there are some problems too both with the existence of a global maximum and the assumptions of Wald. Moreover, we did not found a way to compactify the parameter space such that the continuity assumption is fulfilled. In contrast, through restricting the parameter space a compactification of the latter is possible and as such the conditions of Wald are often satisfied in the restricted space. This result is already indicated by Sundberg (1974) for incomplete data problems from an exponential family, like a normal mixture model. Sundberg states that in the situation of loss of information (such as grouping or mixing), the conditions of Wald are usually much too strong and that results can only be obtained if the parameters are restricted to compact subsets.

Thus, whether or not an adapted ML approach is followed, mostly a consistent (classical) MLE does not exist for the general (log)normal, SEV or Weibull mixture. However, a consistent local maximum does exist. Importantly, for the finite mixture model figure 4.1 shows that whether we either work with a mixture with common or non-common scale parameter, in essence the same kind of estimate is obtained from the likelihood equations, in spite of the convention of terminology to only call the first an MLE. The latter will be referred to as a likelihood estimate (LE).

4.2.2 Multiple roots

It is not sufficient to know that a likelihood estimate exists. The latter also has to be identified to be a useful estimate. This, however, is not an obvious task for a general finite mixture model. Not only its LEQs usually have a large number of roots, but also Cramér's theory only states the existence of a consistent root of the LEQ. No results are available on which root to specify.

Basically, for finite mixtures, there are two types of roots. In the first place, there are multiple roots caused by the non-identifiability of the parameters in the model. Indeed, although the family of general finite mixtures is identifiable (Section 3.3), the parameters are not due to the arbitrariness of the numbering of components of the mixture. Moreover each permutation of the component labels provides another root, resulting in at least $M!$ roots for the likelihood equations. Nevertheless, this problem is not of great concern and can be avoided, for example, by ordering the sizes of the different means or by introducing an equivalence relation in the parameter space making the true parameter identifiable relative to its equivalence class. On the other hand, a second class of roots is of more concern. Day (1969) stated that any pair, triplet,... of distinct observations sufficiently close together, would generate a local maximum of the likelihood, resulting in several roots for the likelihood equations. But his comment that therefore ML estimation breaks down is not warranted as observed previously. These roots are fundamentally different from each other and inherently due to the nature of a finite mixture model.

According to theory, the LEQs contain a "unique consistent" root. Note that unique here refers to a unique equivalence class (Redner, 1981) and that there is also a certain ambiguity in this uniqueness statement (Perlman, 1983). The problem of identifying an unique consistent root is not related to (maximum) likelihood estimation only. Other domains where it appears are, for example, estimating equations and classical least squares estimation.

Small et al. (2000) give a good overview of the problem in general and of several methods dealing with the identification of a consistent sequence of roots. Among the several options, the following three procedures are most appropriate for the problem at hand:

1. If a consistent MLE exists, the global maximum corresponds to the consistent root.

2. Given a consistent estimate, then the root closest (with closest defined through some distance measure) to this estimate is consistent (Lehmann, 1983, p. 421).

3. Given a consistent estimate, the root obtained by using this estimate as starting value for an iterative procedure, is consistent (Small et al., 2000).

Eventually, for the sample size sufficiently large, all procedures will lead to the same root due to the uniqueness of the consistent sequence. For finite normal mixtures with a common scale parameter all three methods are workable. Not only a consistent MLE exists, but also another consistent estimator, namely the moment estimator, can be obtained quite easily (Section 5.1). As well-known, the first method is the most popular one. Nevertheless, sometimes the moment estimator is used as a starting value. It is then understood that the maximum obtained (with method 2 or 3) is the MLE, which is not necessarily true, although the obtained root is consistent too.

At first sight, none of the methods are usable for the general finite mixture model. This is obvious for the first method. But also method 2 and 3 are not option since it is rather problematic to find another consistent estimator (Section 5.1). Still a criterion similar to method 1 can be applied, based on the consistency of the largest local maximum. Indeed, for many general finite mixtures and in particular those considered here, the

root corresponding to the largest (finite) local maximum of the likelihood is consistent. For the normal mixture model, this can be shown in several ways using results described in Section 4.1.2. Both the propositions of Hathaway (1985) and Redner and Walker (1984) on the consistency of the global maximum in a constrained parameter space, imply the consistency of the largest local maximum. Similarly, it can be shown that it also applies for SEV mixtures. For the adapted likelihood methods, the results of Sundberg (1974) for the normal mixture model, imply the existence of a consistent root, which, at least asymptotically, maximizes the adapted likelihood function in every compact subset of the parameter space. In general, this last result holds true for any distribution satisfying both Cramér's conditions over the entire parameter space and Wald's conditions over any compact subspace containing the true parameter.

We preferred this "largest local" criterion to identify a consistent root. In the following, we will refer to the largest local maximum as the LE. This choice is founded on the connection with the classical maximum likelihood method and the fact that it seems one of the few methods which is feasible in practice. As Small et al. (2000) point out: "the multiple root problem of the LEQs has one big advantage as opposed to other problems, in that roots can be compared relatively based on their likelihood value." So, why not using this property. In spite of this, in the next chapter it will become clear that the search for this largest local maximum is not obvious.

In summary, for general finite mixtures with a (log)location-scale distribution as component density the likelihood estimate corresponds to the largest local maximum of the likelihood function. It has the same properties as the MLE and coincides with the MLE when this estimate exists and is consistent. Apparently, the same holds for most adapted likelihood methods.

4.3 The problem of spurious maxima

In spite of the results concerning the LE, McLachlan and Peel (2000, chap. 2-3) argue, amongst others, that this largest local likelihood criterion cannot be followed since a so-called spurious maximum can be chosen as LE. In particular, it was noted, when estimating a general finite normal mixture, that for some samples the largest local maximum of the likelihood could correspond to a maximum with implausible values for the parameters, i.e., a spurious maximum. Note that the presence of "implausible" maxima in the likelihood function for these mixtures was already observed by Day (1969).

To make more clear what is meant with a spurious maximum, we look at an example given by McLachlan and Peel (2000, pp. 100-101). They generated a sample of size 100 from a two-component normal mixture with parameter values $\mu_1 = 0$, $\sigma_1 = \sigma_2 = 1$, $\mu_2 = 2$ and $\pi_1 = 0.5$. A normal QQ-plot of this sample is shown in Figure 4.2a. Two maxima of the likelihood function were located. In Table 4.1, which gives the parameter values of several maxima of this likelihood, they are referred to as maximum 3 and 6, respectively. Clearly, of these two, maximum 3 has the largest likelihood value. But, it also has a value for π_1 which is about $2/100$ and a very small value for the scale parameter σ_1. Moreover, the first component of the mixture (corresponding to maximum 3) is related to a subgroup of only two successive data points of the ordered sample. As such, it is highly unlikely that one would consider that maximum 3 reflects the "truth". This maximum is related to a pure random cluster of data points in the sample and therefore it is called spurious. Further, the other maximum found (i.e., maximum 6 in Table 4.1) was considered to be the LE, due to the fact that its parameter values are much more plausible. In addition, when the sample was binned into 7 intervals of equal width, apparently the parameter values of the MLE then obtained are close to the parameter values of maximum 6, confirming their conclusion that maximum 6 was the LE. Hereby, binning the sample was regarded as a procedure to remove spurious maxima since

Table 4.1: Some local maxima of the likelihood function of the simulated sample from McLachlan and Peel.

maximum	μ_1	σ_1	μ_2	σ_2	π_1	lnL
1 (LE)	-0.83	0.00040	1.06	1.33	0.020	-163.89
2	2.52	0.00065	0.99	1.34	0.020	-165.53
3	**-2.16**	**0.0085**	**1.09**	**1.28**	**0.020**	**-165.94**
4	0.91	1.41	1.71	0.47	0.86	-170.25
5	0.96	1.39	1.62	0.28	0.90	-170.25
6	**-0.70**	**0.95**	**1.38**	**1.11**	**0.17**	**-170.56**
MLE	1.03	1.34				-171.29

Note: The first 3 maxima are the largest local maxima; the maxima in bold are obtained by McLachlan and Peel. The last row gives the MLE obtained for a normal distribution.

the occurrence of these maxima in the likelihood function was attributed to the continuous nature of the data.

We also scanned the whole parameter space for solutions of the LEQs. The way this is carried out is explained in Chapter 5. A lot more maxima than indicated by McLachlan and Peel, are found. Some of these are given in Table 4.1. The first 3 maxima are the largest local maxima of the likelihood function, the last 3 are the only maxima found which have plausible parameter values. Between maximum 3 and 4 more than 20 other maxima are situated. Clearly, the largest local maximum was not obtained by McLachlan and Peel and the likelihood function contains many more maxima than two. However, maximum 1 and 3 are similar in nature: both are truly spurious. As such, there is still the problem that the LE is not believable or reflecting the truth. But, as noted from Table 4.1, there is also no a priori reason to take maximum 6 as the LE. Why not choosing maximum 4 or 5? Indeed, both have a larger likelihood value and their parameter values also seem plausible. The only motivation for choosing maximum 6 as LE, is that it is the maximum *closest* to the true values, with closest defined by some distance measure. However in real examples, one

does not know the true values, which underscores that there are no good grounds to choose maximum 6 as the LE.

The argument that spurious maxima are due to the continuous nature of data is not warranted either. Indeed, binning the sample into a number m_c of intervals with equal width (or equivalently introducing a measurement error δ for the data), will not solve the problem of spurious and multiple maxima of the likelihood function. The presence of these maxima is related to the specific nature of a general finite mixture, as it models clusters within a sample, whether these clusters are real or random. Of course, the number of maxima of the likelihood, and so also of spurious maxima, found will decrease when m_c becomes smaller, since clusters of the sample with a small within variation will be smoothed out. But, how far can we decrease m_c without smoothing out the "real" subdivision of the sample?

To illustrate this, we binned the sample shown in Figure 4.2a into 80, 50, 20 and 7 classes of equal width. In Table 4.2, some "maxima" of the likelihood function for each binned sample are given. Several things can be inferred from this table. First, regardless of the number of classes used, there are problems with the "global" maximum. In each case, the likelihood function seems to attain its largest value in several points with approximately the same values for the proportion parameter and the parameters of the second component, while the values for the parameters of the first component are different. Some of these apparent maxima are tabulated. They can be recognized through the ? sign behind the values for the parameters of the first component. These points are no consistent estimates and even no maxima. The reason is that only two adjacent intervals are used to estimate the parameters of the first component, while at least 3 intervals are required to obtain a consistent estimate (Sundberg, 1974). Second, distinct spurious maxima do not necessarily disappear. Often, as the number of classes reduces, they change into problematic maxima, before they fade away (for example, the maxima related to maximum 3). We will give a more

Table 4.2: Local maxima of the likelihood function for several binned samples from the simulated sample of McLachlan and Peel.

m_c	max	μ_1	σ_1	μ_2	σ_2	π_1	lnL
	3	-2.217 ?	0.00369 ?	1.118	1.238	0.0286	-416.718
	3	-2.213 ?	0.0117 ?	1.118	1.238	0.0286	-416.718
	3	$]-2.218,-2.137[$	$\to 0$	1.085	1.276	0.0191	-418.439
80	4	1.690	0.0344	0.988	1.367	0.0517	-420.025
	1	-0.513	0.0108	1.082	1.331	0.0371	-420.343
	5	0.955	1.391	1.650	0.288	0.902	-421.192
	6	-0.621	0.962	1.422	1.092	0.195	-421.450
	3	-2.161 ?	0.0184 ?	1.118	1.246	0.0275	-371.314
	3	-2.166 ?	0.00808 ?	1.118	1.246	0.0275	-371.314
50	1	-0.492	0.0170	1.098	1.328	0.0431	-373.192
	5	0.976	1.376	1.705	0.110	0.928	-373.517
	6	-0.609	0.925	1.431	1.097	0.197	-374.505
		-2.022 ?	0.0408 ?	1.126	1.240	0.0306	-280.776
	3	$]-2.300,-1.975[$	$\to 0$	1.113	1.252	0.0265	-280.861
20	4-5	0.934	1.368	1.965	0.273	0.910	-282.847
	6	-0.610	0.931	1.434	1.094	0.199	-283.145
7	6?	-1.326 ?	0.173 ?	1.326	1.124	0.109	-178.326
		-1.369 ?	0.00963 ?	1.326	1.124	0.109	-178.326

Note: $]a,b[$ and $\to 0$ refer to a maximum that would be attained in the points $\sigma_i = 0$ and $\mu_i \in]a,b[$. The second column refers to the labels of the maxima in Table 4.1.

detailed discussion of these first two points for the adapted ML methods in Section 4.5.3. Third, we found no maximum, as so no consistent root, in case of 7 classes. Only apparent maxima with the same likelihood value and for which one of the components corresponds to only two adjacent intervals are identified. If we would bin the sample in another way, other results are possible. Fourth, binning the sample does not solve the problem of which maximum to choose as the LE. Not only a global maximum is often not an option, but also up to and including a number of 20 classes, there are at least, two maxima with plausible parameter values.

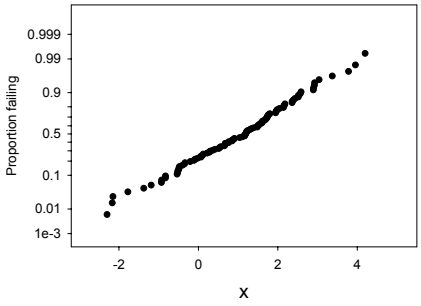
In summary, estimation procedures that pick out a maximum of the likelihood function with plausible parameter values or look for a maximum with parameter values that are close to the parameter values of the LE obtained from a binned sample, are subjective, will not lead to a consistent sequence of estimators and make the inference results unreliable. As such, we do not recommend them. Nevertheless, we cannot neglect that there is sometimes a problem with the LE, in the sense that it does not reflect the true parameter values. The idea here is to clarify this situation and to pass some well-founded means of how to handle this difficulty in practice. We will first consider the problem of these "spurious" maxima in its entirety. Therefore, a picture of the global problem is drawn through a classification of the samples (Section 4.3.1). Then, it is discussed how the appearance of a spurious LE and the known statistical properties of the LE can go together (4.3.2). To end, we give our perception, by means of some guidelines, of how to deal with spurious maxima in practice (4.3.3).

4.3.1 Stability of a sample

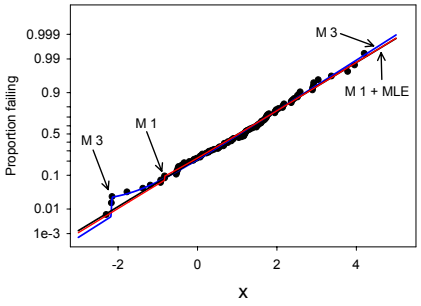
Highly unstable samples

Consider again the sample shown in Figure 4.2a and discussed previously. The problem for this sample was that the parameter values of the LE were implausible or not reflecting the truth. The reason for this spurious LE has to be searched for in the lack of information available within the sample in order to fit a two-component mixture. In other words, although this particular finite mixture model (i.e., the mixture with parameter values $\mu_1 = 0$, $\mu_2 = 2$, $\sigma_1 = \sigma_2 = 1$ and $\pi_1 = 0.5$) is theoretically identifiable (Section 3.3), numerically for this sample it is not. This can be observed in several ways:

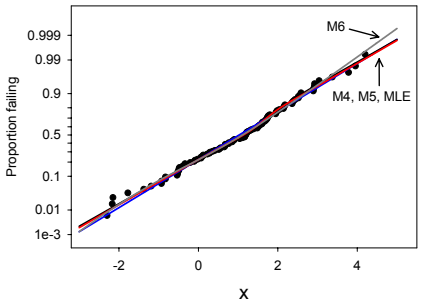
- A single normal distribution can be used to model the sample satisfactorily. A test of normality does not reject the null hypothesis for



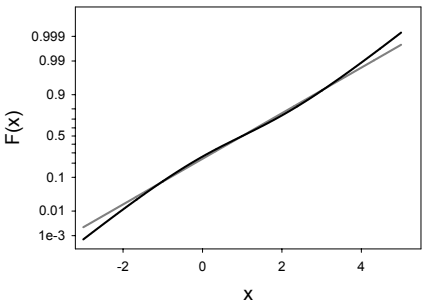
(a) Normal QQ-plot of the sample.



(b) M1-M3: fit of the 1^{st} and 3^{th} maximum of Table 4.1, MLE: fit of a normal distribution.



(c) MLE: fit of a normal distribution, M4-M5-M6: fit of the $4^{th}, 5^{th}$ and 6^{th} maximum of Table 4.1.



(d) Cdf of the true mixture and cdf of a single normal distribution with the same mean and standard deviation of the mixture.

Figure 4.2: Simulated sample of size 100 from McLachlan and Peel (2000, pp. 100). The true parameter values are $\mu_1 = 0$, $\sigma_1 = \sigma_2 = 1$, $\mu_2 = 2$ *and* $\pi_1 = 0.5$.

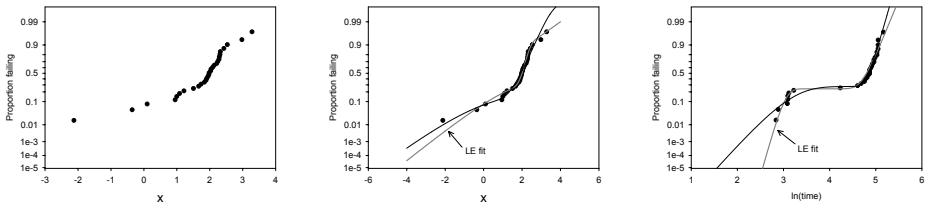
$\alpha = 0.005$. The correlation between the sample quantiles $x_{(i)}$ and $\Phi^{-1}[(i - 0.375)/(n + 0.25)]$ is 0.9957. Moreover, without the prior knowledge that this sample is simulated from a two-component mixture, no one would fit a two-component mixture to this sample.

- In Figure 4.2b the fit of the MLE obtained from a single normal distribution and the fits of the two spurious maxima 1 and 3 (Table 4.1) are shown. Apart from a small deviation in a small number of data points, the fits can hardly be distinguished. While one of the two components of the mixtures (corresponding to these spurious maxima) fits exactly 2 or 3 data points, the other component, for which the parameter values of the location and scale parameter resembles the parameter values of the MLE, has to fit the rest of the sample. Figure 4.2c shows also the fit of the MLE of a single normal distribution, but now with the fits of the three plausible maxima given in Table 4.1. Again, the distinction between all 4 fits is minimal, certainly within the range of data.

- Figure 4.2d depicts (on normal probability scales) the cumulative distribution function (cdf) of the true two-component normal mixture with the cdf of a normal distribution with the same mean and standard deviation of the mixture. As can be observed, except for the extreme tail ends, these two distributions can hardly be distinguished.

Thus, unless the sample size is unduly large, a single normal distribution can equally well be used to fit a sample generated from this particular two-component normal mixture. As a result, solutions of the LEQs for which one of the two components of the mixture, fits exactly 2 or 3 data points, will correspond to maxima at the top of the likelihood function, i.e., maxima with a large likelihood value.

This sample is a typical example of what we define as a *highly unstable* sample with respect to the (general) two-component normal mixture. This means that the largest local maximum of the likelihood function can



(a) Normal QQ-plot of a simulated sample of size 30 from the normal mixture $\mu_1 = 0$, $\sigma_1 = 1$, $\mu_2 = 2$, $\sigma_2 = 0.5$, $\pi_1 = 0.2$.

(b) LE fit of the sample of Figure 4.3a and fit of the 2^{nd} largest maximum (max. 2 in Table 4.3).

(c) Lognormal QQ-plot of a real failure time sample of size 29 with the LE-fit and fit of the 2^{nd} largest maximum.

Figure 4.3: A simulated and a real unstable sample.

be altered through some minor perturbations in the sample, that there are several maxima of the likelihood function with about the same large likelihood value (Table 4.1), and that there is no maximum which dominates the likelihood function.

Unstable samples

Figure 4.3a shows the normal QQ-plot of a sample of size 30, simulated from the two-component normal mixture with parameter values $\mu_1 = 0$, $\sigma_1 = 1$, $\mu_2 = 2$, $\sigma_2 = 0.5$ and $\pi_1 = 0.2$. The 5 largest maxima of the likelihood function (for a general two-component normal mixture) are given in Table 4.3. At first sight, there seems to be no problem. The parameter values of the LE are credible and its fit is acceptable (Figure 4.3b). Nevertheless, the parameter values of the LE are not at all in the neighborhood of the true values, while those of the 2^{nd} largest maximum are *closest* to the true values. This means that the LE is also spurious, i.e., its parameter

Table 4.3: The 5 largest local maxima of the likelihood of the simulated sample of size 30 *shown in Figure 4.3a.*

maximum	μ_1	σ_1	μ_2	σ_2	π_1	lnL
1 (LE)	1.22	1.40	2.09	0.22	0.43	-32.94
2	0.17	1.46	2.03	0.50	0.17	-34.57
3	1.62	1.08	2.28	0.032	0.85	-38.79
4	1.98	0.0037	1.70	1.05	0.062	-40.06
5	2.32	0.0048	1.68	1.04	0.061	-40.31
MLE	1.72	1.02				-43.23

Note: The last row shows the MLE obtained for a normal distribution.

values do not reflect the truth, in spite of the fact that it is not possible to derive it from the parameter values itself.

As a result, the problem is the same as for the highly unstable sample in Figure 4.2a, only less pronounced. Also the reason for this spurious LE is the same: the sample size n is too small to distinguish the true distribution. Moreover, it is not possible to numerically identify only one, i.e., the true, two-component normal mixture. While for the previous sample, the mixture could not be distinguished from a single normal distribution, here it is clear from the QQ-plot in Figure 4.3a that a normal distribution would not be appropriate, i.e., a straight line will not fit the sample satisfactorily. However, there are two solutions of the LEQs for which the corresponding two-component mixture models are difficult to distinguish within the range of data. Outside this range, differences become marked. Consequently, conclusions drawn will depend highly on which of the mixtures (i.e., which of the two maxima at the top of the likelihood) is chosen. For example, the null hypothesis of equal scale parameters versus the alternative of unequal scale parameters would be rejected with the likelihood ratio test (LRT) if the first maximum was taken (LRT-value = 6.461), but accepted if the second maximum was considered (LRT-value = 3.195) on a 95% level. Also the difference in estimation of the low quantiles could influence the decision

whether a physical component is accepted as reliable or not.

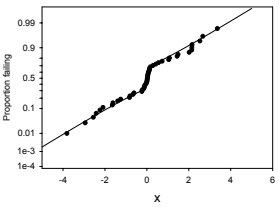
This sample is an example of an *unstable* sample (with respect to a general two-component normal mixture), i.e., a few maxima, mostly with plausible parameter values, are at the top of the likelihood function (Table 4.3). The corresponding mixtures have a similar fit within the range of data. Small perturbations in the sample can alter the largest maximum of the likelihood into one of the other maxima at the top of the likelihood.

An example of a real unstable sample (with respect to a two-component lognormal mixture) is given on a lognormal QQ-plot in Figure 4.3c. The LE-fit of a two-component lognormal mixture is depicted, together with the fit of the 2^{nd} largest maximum. As noted, the difference, between the two mixtures, with regard to the estimation of low quantiles, will be large.

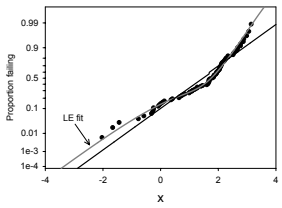
Stable samples

Figure 4.4a depicts the normal QQ-plot of a simulated sample of size 50 from a two-component normal mixture with parameter values $\mu_1 = 0$, $\sigma_1 = 0.1$, $\mu_2 = 0$, $\sigma_2 = 2$ and $\pi_1 = 0.5$. Table 4.4 gives the 4 largest local maxima of the likelihood function. The LE has plausible parameter values and does reflects the true values. It is also the maximum closest to these true values. Most other maxima found have implausible values for the parameters. Apparently, the sample is large enough to distinguish one specific two-component mixture. Moreover, this sample is an example of a *stable* sample, i.e., the likelihood function is dominated by one maximum. Other maxima are pushed into the background. Small perturbations in the sample will not alter the largest local maximum of the likelihood function. Further, the difference in value of the likelihood between the first and the second maximum is large.

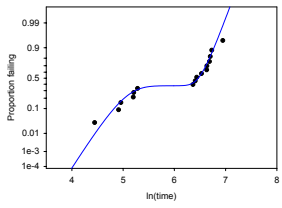
Another example of a stable sample is shown in Figure 4.4b. It is a sample of size 80, simulated from a two-component normal mixture with



(a) Normal QQ-plot of a simulated sample of size 50 with LE fit. True parameter values are $\mu_1 = 0$, $\sigma_1 = 0.1$, $\mu_2 = 0$, $\sigma_2 = 2$, $\pi_1 = 0.5$.



(b) Normal QQ-plot of a simulated sample of size 80 with fits of the LE and the 2^{nd} largest maximum. True parameter values are $\mu_1 = 0$, $\sigma_1 = 1$, $\mu_2 = 2$, $\sigma_2 = 0.5$, $\pi_1 = 0.4$.



(c) Lognormal QQ-plot of a real failure time sample of size 16 with LE-fit.

Figure 4.4: Simulated and real stable samples.

parameter values $\mu_1 = 0$, $\sigma_1 = 1$, $\mu_2 = 2$, $\sigma_2 = 0.5$ and $\pi_1 = 0.4$. The 4 largest local maxima of the likelihood function are tabulated in Table 4.5. Here, the largest local maximum is not so dominant as for the previous sample, but it is resistant to small perturbations. Further, the difference in likelihood value between the 1^{st} and 3^{th} maximum is considerable and the 2^{nd} maximum has implausible parameter values with a fit that differs a lot from the LE fit (Figure 4.4b). No other maximum with reasonable

Table 4.4: The 4 largest local maxima of the likelihood of the simulated sample of size 50 *shown in Figure 4.4a.*

maximum	μ_1	σ_1	μ_2	σ_2	π_1	lnL
1 (LE)	0.010	0.098	-0.035	1.79	0.34	-73.71
2	0.038	0.00032	-0.022	1.49	0.040	-82.35
3	2.15	0.0058	-0.15	1.40	0.058	-82.43
4	0.019	0.0017	-0.022	1.49	0.038	-85.63

Table 4.5: The 4 largest local maxima of the likelihood of the simulated sample of size 80 *shown in Figure 4.4b.*

maximum	μ_1	σ_1	μ_2	σ_2	π_1	lnL
1 (LE)	0.36	1.08	2.07	0.51	0.37	-111.62
2	1.68	3.33e-005	1.44	1.14	0.025	-112.39
3	1.68	0.0027	1.43	1.15	0.046	-116.29
4	1.68	0.0014	1.43	1.15	0.036	-116.86

parameter values is found at the top of the likelihood function. Also here, the LE is the maximum closest to the true values.

Obviously, for these two examples, the LE can be trusted. Although many more maxima with mostly implausible parameter values are present in the likelihood function, none of them bother. The samples contain enough information, i.e., the sample size is large enough, to numerically distinguish the underlying distribution. To conclude, Figure 4.4c gives a lognormal QQ-plot of a real failure time sample together with the LE fit of a two-component lognormal distribution. It is an example of stable sample encountered in practice.

4.3.2 Discussion

Previous examples made clear the problem, touched upon already by some, but never treated in detail. Namely, when estimating a general two-component normal mixture to a sample, for certain samples the LE does not reflect the true parameter values, but rather a random grouping within the sample. In other words, the LE is unreliable or *spurious*. Sometimes, this is clear from the parameter values of the LE itself, but equally well it may not be. The former was true for highly unstable samples, the latter for unstable samples.

Cause of spurious likelihood estimate

We indicated that for (highly) unstable samples, a two-component mixture was numerically not identifiable. The likelihood function contained several local maxima with a large likelihood value and for which the corresponding mixtures had a resembling fit within the range of data. The reason for this non-identifiability problem, resulting in an unreliable LE, is a too small sample size (under the assumption that the true model is a two-component mixture). The cause is twofold. First, there is the fact that the "consistency" property of the LE is an asymptotic concept. This means that only for a sufficiently large sample size the estimators looked at will approach the true parameter values. For small sample sizes, on the contrary, nothing is known about the performance of a consistent estimator. Second, there would be no problem if the LEQs had only one root. However, due to the nature of the mixture model itself, the LEQs contain many roots.

Relation to efficiency of the likelihood estimator It has already been pointed out that for small sample sizes the MLE could be unreliable in case of a finite normal mixture model (Hosmer, 1973). However, this was not related to the occurrence of spurious maxima at the top of the likelihood function, but to the fact that for these samples the MLE behaved rather poorly. Especially, estimates were not accurate enough to be useful estimates. Redner and Walker (1984) and Behboodian (1972), amongst others, indicated that for poorly separated mixtures a huge sample size is required not only to obtain efficient but also accurate estimates. In particular, Behboodian calculated the Fisher information matrix $I(\boldsymbol{\theta})$ approximately for several two-component normal mixtures. He noted that this matrix goes to a singular matrix, i.e., the condition number $\|I(\boldsymbol{\theta})\|$ of $I(\boldsymbol{\theta})$ becomes infinite, as the mixture components come closer or the proportion parameter goes to 0 or 1. Redner and Walker states that one can expect only a limited accuracy for the estimates in case of ill-conditioned problems. The latter refers

to problems where the solution is very sensitive to perturbations in the data. The maximization of the likelihood for unstable samples is an example of an ill-conditioned problem. Both authors infer that for mixtures with a large condition number for $I(\boldsymbol{\theta})$, i.e., poorly separated mixtures, the sample size has to be huge to obtain precise estimates (see also Table 4.10).

Nevertheless, although the numerical identifiability of the problem is related to the accurateness (and efficiency) of the estimates, it was not related to the appearance of spurious maxima at the top of the likelihood function. Still, both features are a direct result of a too small sample size. Moreover, spurious maxima are neglected, i.e., a plausible maximum is searched for instead, while spurious maxima turn out to be the best option to recognize too small sample sizes in case of a multiple root problem (Section 4.3.3). From this point of view, it is interesting to observe the (dis)similarity between the surfaces of the likelihood for small and large sample sizes in case the LEQs have multiple roots and in case they only have one root. On the one hand, for a small sample size, the likelihood function will have a (very) flat curvature in case the LEQs have a unique root. The flatness of the surface of the likelihood in case the LEQs have multiple roots, is expressed through several maxima which are at the top of the likelihood. One could think of a bumpy surface (see also Figure 4.5a). On the other hand, the likelihood function will have a sharp curvature for a large sample size in the one root case, while for the multiple root case the sharpness of the likelihood is expressed through one dominating maximum (see also Figure 4.5b). This means that, while in the one root case, a small sample size can be noticed through the large value of the standard errors (which are in relation to the curvature of the likelihood function), this is not entirely true for the multiple root case. There, it is important to not only focus on the largest maximum, but to obtain an overall view of the surface of the likelihood function. This is the only way to obtain information about the credibility of the LE.

Definitions and general comments We introduced the notion of stability of a sample for a two-component normal mixture. In a similar way it can be defined for any other distribution. Only the difference between a highly unstable and an unstable sample is specific to the case of a general M-component mixture. For the former, the largest local maximum is truly spurious, i.e., at least one of the components of the mixture is related to a subgroup of only a few data points. Such maxima are referred to as *distinct spurious*.

The problem of an unreliable LE is not related merely to the case of likelihood estimation or to the mixture model. It is inherent to all consistent estimators obtained as a solution of the LEQs and where these equations have multiple roots. As such, it can just as well happen in case of classical ML estimation or as will be illustrated further on for adapted likelihood estimation. In the following paragraph an example is given of a distribution where a spurious MLE can occur.

Apparently, for small sample sizes, the property of consistency for a likelihood estimator when multiple roots are present in the LEQs, is not enough to guarantee that the estimator is meaningful. As suggested a couple of times, a spurious maximum can, on a purely theoretical basis, be defined as any maximum not *closest* to the true values, with *closest* defined by some distance measure. As such, for each sample, there is only one proper maximum. Importantly, for some sample size n on, this maximum will be equal to the LE or MLE due to their consistency property. Note that any other method proposed to obtain a consistent root of the LEQs would suffer from the same problem of spurious maxima (Section 4.2.2).

Example of a spurious MLE

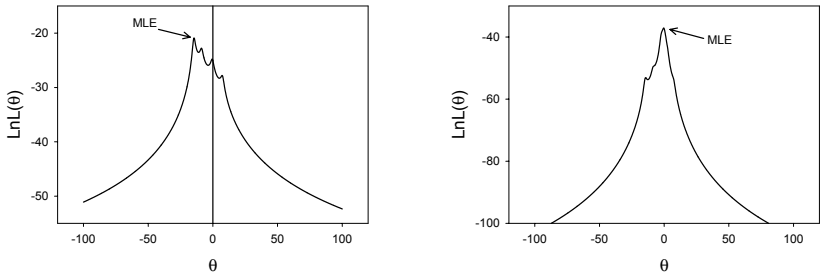
A simple example to illustrate that the appearance of spurious maxima at the top of the likelihood function could also occur in case a consistent MLE exists, is given by the one-parameter Cauchy location distribution

(Barnett, 1966; Reeds, 1985). Its density function is:

$$f(x) = \frac{1}{\pi[1 + (x - \theta)^2]} \quad (-\infty < x, \theta < \infty), \tag{4.5}$$

with θ a location parameter. This parametric family fulfills both the conditions of Cramér and Wald (Perlman, 1983). Therefore, the MLE exists, is consistent and can be found as a root of the LEQ. This equation, however, has usually more than one root or the likelihood function has more than one maximum. Here, the presence of multiple maxima is related to the absence of finite moments for the Cauchy location distribution. In particular, Reeds (1985) showed that anomalous local maxima are related to outlying values of the sample which arrive frequently due to the heavy tails of the Cauchy distribution. Similar to the case of the mixture model, it is not possible to distinguish an anomalous root (i.e., a spurious root) from a proper one in case the sample size is too small.

As an example, Figure 4.5a depicts the logarithm of the likelihood function of a sample of size 5, generated from the Cauchy location distribution with location parameter $\theta = 0$. As noted, the likelihood function has 4 maxima. The MLE corresponds to an anomalous root, since its parameter value is quite far from 0 and it is the maximum farthest from the true value. According to the definition of stability, this is an unstable sample (with respect to the Cauchy location distribution). Indeed, leaving out only one data point, will easily switch the global maximum of the likelihood function into one of the other 3 maxima. If the sample size of this sample is increased to 9, however, the sample becomes stable as shown in Figure 4.5b. One maximum, i.e., the one closest to the true value, dominates the likelihood function. Clearly, the same behavior is observed as for the examples discussed in Section 4.3.1. Namely, for small sample sizes, the MLE cannot be trusted, while for large samples the MLE is reliable. Importantly, the value of "small" and "large" depends highly on the distribution used. For the one-parameter Cauchy distribution, a small sample size means a value



(a) Simulated sample of size 5. (b) Simulated sample of size 9.

Figure 4.5: Logarithm of the likelihood function of simulated samples from the Cauchy location distribution with true parameter value $\theta = 0$.

not larger than about 6, while a large sample size is from about 10 onwards. As such, in practice for this distribution there will be no problem with the MLE, since usually the sample size will be larger than 10. For the finite general mixture model, on the contrary, for some mixtures a size of 50 will be large enough, while for others 1000 or even 10000 will not be sufficient. Consequently, the credibility of the LE is an important issue there.

Required sample size

Quite likely the value of $\|I(\boldsymbol{\theta})\|$ determines not only the sample size required to obtain accurate estimates, but will also be in relation with the sample size needed to have a non spurious or reliable LE. Moreover, it is expected that the sample size required will depend highly on the true finite mixture model, especially on how well its components are separated. To demonstrate this, we carried out a small simulation study for the general two-component normal mixture model.

Samples are generated from a two-component normal mixture model with 12 different sets of parameter values divided into 3 groups of 4. In each group, one parameter is varied in order to study one aspect of the identifiability of the mixture components, i.e., how well the two component distributions can be identified from the mixture or how well they are separated. This is related to mainly three aspects of the mixture: the difference in location parameter of the two component distributions, the size of the ratio of the two scale parameters and to a lesser degree the size of the proportion parameter. The cumulative distribution functions of these 12 mixtures are displayed in Figures 5.7a, 5.7b and 5.7c of Chapter 5. In the first group, the location parameter of the second component, μ_2, is varied. It takes the values $1, 2, 3,$ and 4. The values of the other parameters are $\mu_1 = 0$, $\sigma_1 = \sigma_2 = 1$ and $\pi_1 = 0.5$. The larger the value of μ_2, the better the component distributions are separated (Figure 5.7a). For the second group, all parameters, except the scale parameter of the first component, are kept fixed. The values for the parameters here are $\mu_1 = \mu_2 = 0$, $\sigma_2 = 2$, $\pi_1 = 0.5$ and $\sigma_1 = 0.1,\ 0.2,\ 0.5,\ 1$. In spite of the common location parameter, the components of the mixture can still be clearly identified if the ratio of the two scale parameters deviates sufficiently from 1 (Figure 5.7b). In the last group, the proportion parameter is altered from a small value (0.2) over two average values (0.4 and 0.6) to a large value (0.8). The values for the other parameters are $\mu_1 = 0$, $\sigma_1 = 1$, $\mu_2 = 2$ and $\sigma_2 = 0.5$. It is clear from Figure 5.7c that also the value of π_1 has an influence on the identifiability of the mixture components.

For each set of parameter values, sample sizes of $20, 50, 100, 200, 300,$ $400, 500$ and 1000 are used, with 1000 simulations in each case. Results are summarized in Table 4.6. The possible spurious nature of the LE is assessed through a comparison with the maximum closest to the true values. Here, *closest* is defined by the Euclidean distance, but with the scale and proportion parameters rescaled such that their domain is the same as for the location parameters. The tabulated value k is then the number of times

Table 4.6: The number of times out of 1000 (k) that the largest local maximum of the likelihood is a spurious maximum.

	μ_2			
n	1	2	3	4
20	951	933	842	589
50	988	971	794	279
100	996	972	587	35
200	997	970	213	0
300	998	964	51	0
400	997	942	12	0
500	999	888	2	0
1000	1000	642	0	0

(a) First group: separation in location.

	σ_1			
n	1	0.5	0.2	0.1
20	943	845	435	208
50	964	622	60	10
100	970	227	1	0
200	910	14	0	0
300	812	0	0	0
400	744	0	0	0
500	616	0	0	0
1000	151	0	0	0

(b) Second group: separation in scale.

	π_1			
n	0.8	0.6	0.4	0.2
20	906	803	691	643
50	927	712	408	287
100	901	475	101	64
200	787	100	25	28
300	637	23	13	15
400	428	10	9	17
500	271	3	5	10
1000	9	0	3	4

(c) Third group: varying the proportion parameter.

out of 1000 that the LE is spurious. The value of k should go to 0 when n increases.

Clearly, for all sets of parameter values, except for one set with results tabulated in the first column of Table 4.7a, the value of k shows finally a decreasing trend. The dependency between how well the mixture components are separated and the sample size required such that the LE is the maximum closest to the true values, is evident from the tables. Moreover, the value of k goes relatively fast to 0 for mixtures that have clearly identifiable components (i.e., the last one or two columns in each table). For some sets of parameter values a sample size of 100 or even lower would be sufficient, while for others a sample size of 200 is required. But, for some poorly identifiable mixtures, although k shows at last a decreasing trend, 0 is not reached for even a sample size of 1000. The worst case is the mixture with parameter values $\mu_1 = 0$, $\mu_2 = 1$, $\sigma_1 = \sigma_2 = 1$ and $\pi_1 = 0.5$ (1^{st} column in Table 4.7a), where k does not show at all a decreasing trend before a sample size of 1000. It even gets worse as n increases. For example, for $n = 1000$, in none of the generated samples, the LE was equal to the maximum closest to the true values. The reason is clear: this specific mixture can be hardly distinguished from a single normal distribution. As seen in Figure 5.7a, the cdf of this mixture is practically a straight line. It is doubtful that any sample of this particular mixture distribution will be ever identified as coming from a mixture.

In summary, from some sample size onwards, the LE will be a good estimator. But the sample size required depends highly on how well the components of the true mixture are separated. For some mixtures, a (very) small sample size will be sufficient, but for others even a huge sample size will not do.

4.3.3 Guidelines

Although in theory the definition of a spurious maximum sounds nice, in practice there is one big problem: the "truth" is not known. It is not possible to search for the maximum closest to the true values. It is possible, however, to search for the LE. As shown, if the sample size is large enough, the LE will be the maximum closest to the true values. In other words, it will not be spurious. Still, the sample size required is not known either. Fortunately, the stability of a sample gives an excellent idea whether the sample size is large enough, i.e., whether the LE can be trusted. We derived some easy to use but important guidelines. They are based on the fact that not only the LE has to be looked at, but also other maxima of the likelihood function.

- The sample is *highly unstable*, i.e., many maxima from which a lot have implausible parameter values are at the top of the likelihood function. If the true distribution is a two-component mixture distribution the sample size is far too small to detect this mixture. One can select from several options, apart from proceeding with the LE or any other maximum: look for prior information (like physical background), increase sample size or use a simpler model. For example, for the sample shown in Figure 4.2a, a normal distribution would equally well fit this sample. Moreover, an increase of sample size would not help in this case, unless it would be huge.

- The sample is *unstable*, i.e., a few maxima which have mostly credible parameter values, are dominating the likelihood function. Generally, if the true distribution is a two-component mixture distribution, the sample size is somewhat too small to distinguish between several two-component mixtures. Often, a worst-case scenario can be used: based on the few maxima dominating the likelihood function, several analyses are carried out. The one with worst results (with respect to what is

asked) is taken. Again prior information or an increase of the sample size could help. For example, for the real sample shown in Figure 4.3c, information of other experiments led to the 2^{nd} maximum (and not the LE) as the proper one.

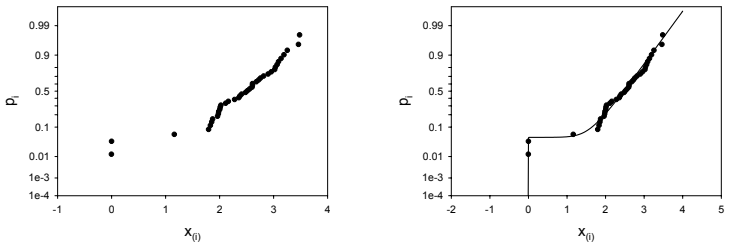
- The sample is *stable*, i.e., one maximum dominates the likelihood function or the largest maximum is followed by distinct spurious maxima. In such a case, there is nothing suggesting that the LE cannot be trusted.

The stability of a sample (with respect to any model) tells a lot about the credibility of the LE or MLE. Obviously, the likelihood function will have to be scanned for local maxima in a well-reasoned way. In Chapter 5, we explain how we dealt with this. Further, an extension of these guidelines to mixtures with more than 2 components or other component distributions is evident. Also, they can be just as well used for the adapted likelihood methods.

To conclude this discussion, note that not all maxima with a very small value for the proportion parameter are spurious. This occurs, for example, in case the sample has a small group of outliers. In this situation, the guidelines represented above, can also be used. As an example, consider the sample shown on a normal QQ-plot in Figure 4.6a. This sample of size 40 is simulated from a two-component normal mixture with parameter values $\mu_1 = 0$, $\sigma_1 = 0.005$, $\mu_2 = 2.5$, $\sigma_2 = 0.5$ and $\pi_1 = 0.06$. As observed, the sample has a subgroup of two outlying data points. Table 4.7 gives the 3

Table 4.7: The 3 largest local maxima of the likelihood of the simulated sample of size 40 *shown in Figure 4.6.*

maximum	μ_1	σ_1	μ_2	σ_2	π_1	lnL
LE	-0.0026	0.0014	2.50	0.53	0.045	-27.45
2	2.61	0.00015	2.37	0.77	0.050	-37.10
3	2.61	0.0036	2.36	0.78	0.068	-40.98



(a) Normal QQ-plot.

(b) LE fit of a 2-component mixture.

Figure 4.6: Simulated sample of size 40 *from a two-component normal mixture with parameter values* $\mu_1 = 0$, $\sigma_1 = 0.005$, $\mu_2 = 2.5$, $\sigma_2 = 0.5$ *and* $\pi_1 = 0.06$

largest local maxima of the likelihood function. Based on the difference in likelihood value between the first and the second maximum, this sample is stable. So, in spite of the small value of the proportion parameter, the LE seems reliable, and indeed it reflects the true parameter values.

4.4 Sample properties of the likelihood estimator

In the previous sections, we obtained several results related to the distribution of the LE in case of a general finite mixture with a (log)location-scale distribution as component density. They can be summarized as follows:

- The LE is consistent. Its asymptotic distribution is normal with a variance equal to $I(\boldsymbol{\theta})^{-1}/n$, i.e., the LE is asymptotically efficient.

- For small sample sizes n, with small depending on how well the components are separated, the LE is often spurious.

- For large n, the LE is the maximum closest to the true values.
- For poorly separated mixtures, a large sample size is required to obtain accurate or precise estimates, i.e., to obtain a rather small asymptotic variance.

To link these results and to complete the overall picture on the properties of the distribution of the likelihood estimator, both for small and large sample sizes, we carried out a (small) Monte Carlo simulation. In particular, the distribution of the LE for several two-component normal mixtures at various sample sizes is simulated and compared to its asymptotic distribution and to the (simulated) distribution of the maximum obtained when using the true values as starting values. When applying the EM-algorithm as iterative procedure, this last distribution is, from a moderate sample size onwards, equal to the distribution of the maximum closest to the true parameter values. Only in case of a (very) small sample size, for a few samples the maximum obtained with the true parameter values will not be the maximum closest to the true parameter values.

We considered the same 12 mixtures as in Section 4.3.2. For each combination of parameter values and sample size, 1000 samples were generated. A simulated distribution for the LE and the maximum derived from the true parameter values, is obtained, as well as a distribution for the estimators of the corresponding variance-covariance matrices. To visually evaluate the simulation results, we mainly used normal QQ-plots to look at the (univariate) distribution for the estimator of each parameter and scatter plots to consider the relation between the estimators of all parameters. Further, some sample statistics are calculated: the mean of the estimates ($\overline{\hat{\theta}}$) with estimated precision, the absolute bias, the variance of the estimates ($S^2_{\hat{\theta}}$) and also the mean of the estimated variances ($\overline{\hat{\sigma}^2_{\hat{\theta}}}$) with estimated precision.

Table 4.8 presents some sample statistics for the well separated mixture with parameter values $\mu_1 = 0$, $\sigma_1 = 0.1$, $\mu_2 = 0$, $\sigma_2 = 2$, $\pi_1 = 0.5$, while Table 4.9 gives results for the poorly separated mixture with parameter

values $\mu_1 = 0$, $\sigma_1 = 1$, $\mu_2 = 2$, $\sigma_2 = 0.5$, $\pi_1 = 0.8$. The number between brackets is an estimate of the precision, if available. The parameter σ_θ^2 is the variance of the asymptotic distribution for the different parameters. Further, "corr." refers to the correlation between the sample quantiles $\hat{\theta}_{(i)}$ and $\Phi^{-1}[(i-0.375)/(n+0.25)]$. A value very close to one is an indication for a normal distribution. These two examples illustrate quite well the tendency which was noticed for all mixtures considered. Specifically, the following conclusions can be drawn.

First, the small sample properties of the LE are not all comparable with its large sample or asymptotic properties, i.e., the distribution of the LE for small sample sizes cannot be modeled by the asymptotic distribution. For example, in Table 4.8a at sample size 20 and in Table 4.9a at sample sizes 100 and 200, it can be observed that:

- Estimates are biased. Especially, in case of the poorly separated mixture, the bias can be large. For example, at $n = 100$ the bias of the 5 parameters varies from a value of 0.151 to 1.329, while at $n = 20$ for the parameters of the well separated mixture its value is between 0.0147 and 0.164.

- The distribution of the estimator of most parameters is clearly not normal. This can be seen, for example, from the values of the correlation parameter which are not close enough to one or from the normal QQ-plots in Figures 4.7, 4.8a and 4.8c. Mostly, their shape does not reveal a straight line.

- The variance of the distribution of the LE for each parameter (i.e., $S_{\hat{\theta}}^2$) is much larger compared to the asymptotic variance (i.e., σ_θ^2). For example, the ratio $S_{\hat{\theta}}^2/\sigma_\theta^2$ varies from 1.22 to 93 at $n = 20$ for the parameters of the well separated mixture and from 1.36 to 47.7 for the poorly separated mixture at $n = 200$.

n	Sample stat	μ_1	σ_1	μ_2	σ_2	π_1
	$\overline{\hat{\theta}}$	0.0189	0.0853	-0.0423	1.836	0.480
		(0.0109)	(0.00171)	(0.0227)	(0.0162)	(0.00543)
	bias	0.0189	0.0147	0.0423	0.164	0.0202
20	$S^2_{\hat{\theta}}$	0.120	0.00292	0.517	0.262	0.0295
	$\overline{\hat{\sigma}^2_{\hat{\theta}}}$	0.00123	0.000741	0.405	0.209	0.0137
		(6.51e-5)	(3.95e-5)	(0.00913)	(0.00480)	(1.19e-4)
	$\sigma^2_{\hat{\theta}}$	0.00129	0.000880	0.400	0.214	0.0161
	corr.	0.481	0.920	0.992	0.998	0.975
	$\overline{\hat{\theta}}$	0.00106	0.0973	-0.0103	1.962	0.505
		(0.00117)	(0.000663)	(0.0127)	(0.00935)	(0.00269)
	bias	0.00106	0.00273	0.0103	0.0382	0.00487
50	$S^2_{\hat{\theta}}$	0.00137	0.000440	0.160	0.0873	0.00722
	$\overline{\hat{\sigma}^2_{\hat{\theta}}}$	0.000536	0.000393	0.166	0.0880	0.00638
		(9.99e-6)	(1.58e-5)	(0.00198)	(0.00104)	(3.42e-5)
	$\sigma^2_{\hat{\theta}}$	0.000515	0.000352	0.160	0.0857	0.00646
	corr.	0.731	0.986	0.998	0.999	0.993
	$\overline{\hat{\theta}}$	0.000701	0.0986	-0.00242	1.971	0.502
		(5.14e-4)	(4.39e-4)	(0.00884)	(0.00680)	(0.00183)
	bias	0.000701	0.00135	0.00242	0.0289	0.00151
100	$S^2_{\hat{\theta}}$	0.000264	0.000193	0.0781	0.0462	0.00336
	$\overline{\hat{\sigma}^2_{\hat{\theta}}}$	0.000264	0.000188	0.0805	0.0429	0.00322
		(2.87e-6)	(3.20e-6)	(6.69e-4)	(3.54e-4)	(8.10e-6)
	$\sigma^2_{\hat{\theta}}$	0.000258	0.000176	0.0801	0.0429	0.00323
	corr.	0.999	0.996	0.999	0.999	0.999

(a) Sample properties of the largest local maximum.

n	Sample stat	μ_1	σ_1	μ_2	σ_2	π_1
	$\overline{\hat{\theta}}$	-0.0000413	0.0951	-0.0310	1.880	0.508
		(0.00125)	(0.00117)	(0.0220)	(0.0154)	(0.00399)
	bias	0.0000413	0.00488	0.0310	0.120	0.00840
20	$S^2_{\hat{\theta}}$	0.00156	0.00137	0.485	0.236	0.0159
	$\overline{\hat{\sigma}^2_{\hat{\theta}}}$	0.00145	0.000899	0.422	0.219	0.0151
		(5.91e-5)	(4.11e-5)	(0.00864)	(0.00456)	(7.97e-5)
	corr.	0.997	0.960	0.999	0.997	0.999
	$\overline{\hat{\theta}}$	0.00185	0.09753	-0.0107	1.963	0.506
		(0.000762)	(0.000643)	(0.0127)	(0.00932)	(0.00261)
	bias	0.00185	0.00247	0.0107	0.0374	0.00560
50	$S^2_{\hat{\theta}}$	0.000581	0.000414	0.160	0.0868	0.00683
	$\overline{\hat{\sigma}^2_{\hat{\theta}}}$	0.000538	0.000396	0.166	0.0881	0.00640
		(9.94e-6)	(1.58e-5)	(1.97e-3)	(1.03e-3)	(3.32e-5)
	corr.	0.999	0.988	0.998	0.999	0.999

(b) Sample properties of the maximum attained with the true values.

Table 4.8: Simulation results for the well-separated mixture with parameter values $\mu_1 = 0$, $\sigma_1 = 0.1$, $\mu_2 = 0$, $\sigma_2 = 2$, $\pi_1 = 0.5$.

n	Sample stat	μ_1	σ_1	μ_2	σ_2	π_1
	$\overline{\hat{\theta}}$	0.340 (0.00706)	1.169 (0.00423)	0.671 (0.0404)	0.0449 (0.00426)	0.951 (0.00258)
	bias	0.340	0.169	1.329	0.455	0.151
100	$S^2_{\hat{\theta}}$	0.0498	0.0179	1.631	0.0181	0.00663
	$\overline{\hat{\sigma}^2_{\hat{\theta}}}$	0.0159 (2.20e-4)	0.00796 (8.11e-5)	0.00232 (3.71e-4)	0.00102 (1.33e-4)	0.000747 (7.35e-5)
	$\sigma^2_{\hat{\theta}}$	0.0637	0.0264	0.0701	0.0324	0.0141
	corr.	0.937	0.921	0.981	0.614	0.641
	$\overline{\hat{\theta}}$	0.287 (0.00733)	1.151 (0.00425)	0.853 (0.0410)	0.0975 (0.00604)	0.937 (0.00327)
	bias	0.287	0.151	1.147	0.403	0.137
200	$S^2_{\hat{\theta}}$	0.0536	0.0180	1.676	0.0364	0.0106
	$\overline{\hat{\sigma}^2_{\hat{\theta}}}$	0.0100 (2.69e-4)	0.00479 (8.60e-5)	0.00404 (4.41e-4)	0.00166 (1.33e-4)	0.000934 (8.39e-5)
	$\sigma^2_{\hat{\theta}}$	0.0318	0.0132	0.0351	0.0162	0.00704
	corr.	0.925	0.918	0.962	0.747	0.760

(a) Sample properties of the largest local maximum.

n	Sample stat	μ_1	σ_1	μ_2	σ_2	π_1
	$\overline{\hat{\theta}}$	0.0300 (0.0111)	0.953 (0.00433)	1.779 (0.0249)	0.457 (0.00550)	0.775 (0.00337)
	bias	0.0300	0.0466	0.221	0.043	0.025
100	$S^2_{\hat{\theta}}$	0.122	0.0187	0.619	0.0302	0.0113
	$\overline{\hat{\sigma}^2_{\hat{\theta}}}$	0.0733 (0.00590)	0.0216 (0.000867)	0.0983 (0.00758)	0.0285 (0.00135)	0.0215 (0.00225)
	corr.	0.918	0.998	0.782	0.998	0.984
	$\overline{\hat{\theta}}$	-0.0194 (0.00699)	0.973 (0.00350)	1.923 (0.0127)	0.483 (0.00445)	0.779 (0.00278)
	bias	0.0194	0.0270	0.0769	0.0172	0.0212
200	$S^2_{\hat{\theta}}$	0.0488	0.0122	0.161	0.0198	0.00774
	$\overline{\hat{\sigma}^2_{\hat{\theta}}}$	0.0380 (0.00185)	0.0124 (3.65e-4)	0.0550 (0.00417)	0.0183 (0.00104	0.0103 (6.54e-4)
	corr.	0.938	0.998	0.756	0.998	0.993

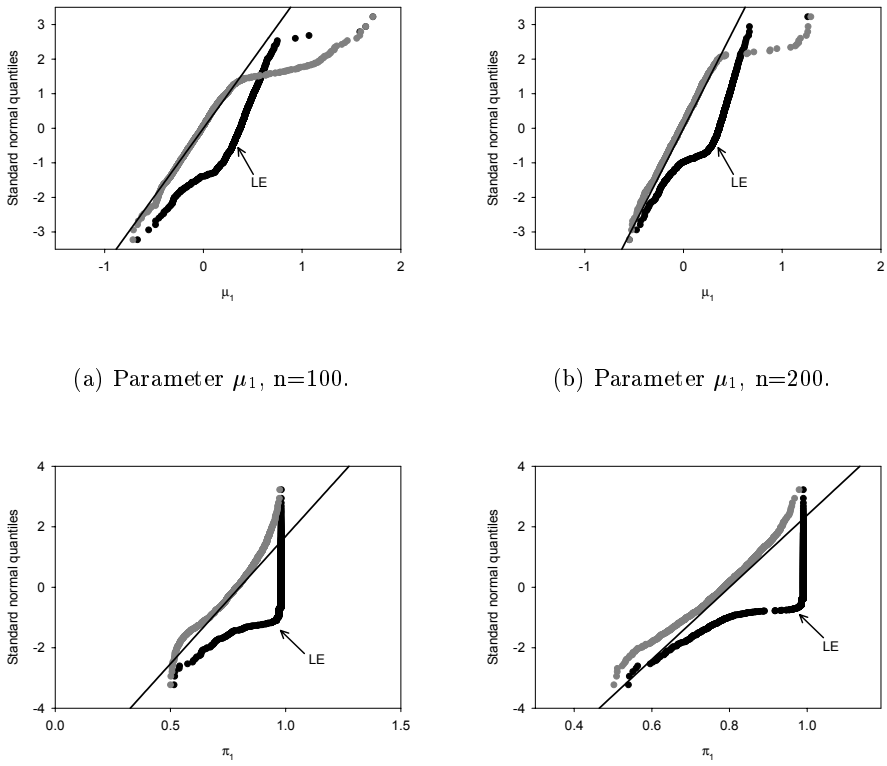
(b) Sample properties of the maximum attained with the true values.

Table 4.9: Simulation results for the mixture with poorly separated components. Parameter values are $\mu_1 = 0$, $\sigma_1 = 1$, $\mu_2 = 2$, $\sigma_2 = 0.5$, $\pi_1 = 0.8$.

Further, the estimated variance of the LE ($\overline{\hat{\sigma}^2_{\hat{\theta}}}$) is much too small in comparison with the variance of the distribution of the LE ($S^2_{\hat{\theta}}$) (for example, at $n = 20$ for the well separated mixture, the ratio $S^2_{\hat{\theta}}/\overline{\hat{\sigma}^2_{\hat{\theta}}}$ varies from 1.25 to 97.6 and from 2.25 to 703 at $n = 100$ for the poorly separated mixture). As a result, in case of a small sample size, not only the LE will be biased and asymptotic normal confidence intervals cannot be used, but also its estimated standard error (when based on the observed information matrix) is highly unreliable.

Second, for small sample sizes, the distribution of the maximum obtained with the true values is in general (much) closer to the asymptotic distribution than the distribution of the LE. This can be noticed, for example, through comparing the sample statistics in Table 4.8a and Table 4.8b at sample size 20 and Table 4.9a and Table 4.9b at both sample sizes. Still, as can be observed from the tables as well, also for this maximum it is mostly not appropriate to use the asymptotic properties for small sample sizes, in particular this holds true for poorly separated mixtures. To illustrate both aspects, Figures 4.8a and 4.8c give the normal QQ-plot of the simulated distribution of the LE and the maximum closest to the true values for the parameters σ_1 and σ_2 of the mixture with well separated components at sample size 20. Figure 4.7 gives similar plots for the parameters μ_1 and π_1 of the poorly separated mixture at sample sizes 100 and 200. As observed, the distribution of the maximum obtained with the true values is in general closer to the asymptotic distribution, but also is mostly not normal. Further, note the large difference between the two simulated distributions for the poorly separated mixture.

Third, in general the distribution of the LE can be approximated satisfactorily by its asymptotic distribution, if the largest local maximum is also the maximum closest to the true parameter values. Moreover, for the mixtures considered, a sufficiently large sample size corresponds to those sizes for which Table 4.6 indicates that the LE is mostly not a spurious max-



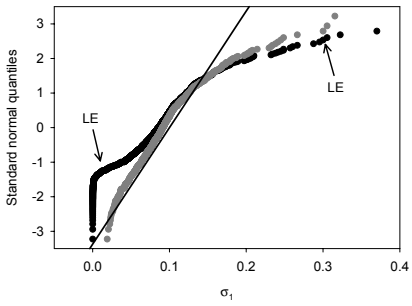
(a) Parameter μ_1, n=100.

(b) Parameter μ_1, n=200.

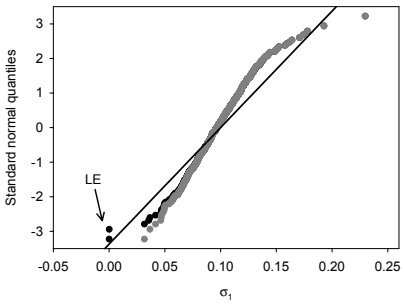
(c) Parameter π_1, n=100.

(d) Parameter π_1, n=200.

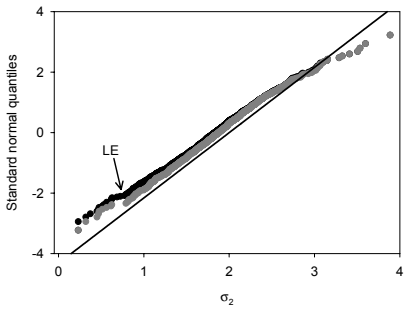
Figure 4.7: Normal QQ-plots of the simulated distribution of the LE (indicated, black) and the maximum obtained with the true values (grey) for some parameters of the poorly separated mixture ($\mu_1 = 0$, $\sigma_1 = 1$, $\mu_2 = 2$, $\sigma_2 = 0.5$, $\pi_1 = 0.8$). The straight line is the cumulative distribution function of the asymptotic distribution.



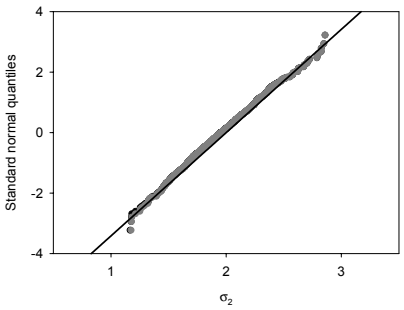
(a) Parameter σ_1, n=20.



(b) Parameter σ_1, n=50.



(c) Parameter σ_2, n=20.



(d) Parameter σ_2, n=50. Almost no distinction visible between both simulated distributions.

Figure 4.8: Normal QQ-plots of the simulated distribution of the LE (indicated, black) and the maximum obtained with the true values (grey) for some parameters of the well separated mixture ($\mu_1 = 0$, $\sigma_1 = 0.1$, $\mu_2 = 0$, $\sigma_2 = 2$, $\pi_1 = 0.5$). The straight line is the cumulative distribution function of the asymptotic distribution.

imum. For example, at $n = 100$ for the well separated mixture, of the 1000 simulated samples no spurious LE was found while the distribution of the LE resembles its asymptotic distribution (Figure 4.9). On the other hand, at $n = 50$ still a small number of spurious LE were identified while for some parameters the distribution of the LE cannot be modeled by its asymptotic distribution (Figures 4.8b and 4.8d). Importantly, this also seems to hold for the distribution of the maximum obtained with the true values. Namely, it appears that although the distribution of the maximum derived from the true values is generally closer to the asymptotic distribution, the approximation will be adequate if the distribution of the LE has become equal or close to the distribution of the maximum closest to the true values. For example, in Figure 4.8 it can be observed that in approaching the asymptotic distribution, both simulated distributions "first" approach each other. Note that in Table 4.8b no sample statistics are given for the sample size 100, since at this size the simulated distribution of the LE and the maximum obtained with the true values were equal. To illustrate the relation with the asymptotic distribution at this sample size, Figure 4.9 shows the normal QQ-plot of the simulated distribution of the LE for all 5 parameters together with the cumulative distribution function of the asymptotic distribution. As noted, for all parameters the simulated distribution can be adequately modeled through the asymptotic distribution.

In summary, as long as the LE is not the maximum closest to the true values, its asymptotic properties cannot be guaranteed. Small sample properties of the LE highly depends on the specific mixture used and should be simulated when required. For real samples, the question whether the sample size is large enough to use asymptotic properties reduces again to the question whether the LE is spurious or not. An answer which can be found through looking at the stability of the sample or the nature of the surface of the likelihood. In other words, the guidelines given in the previous section about the credibility of the LE do equally well hold for the use of its asymp-

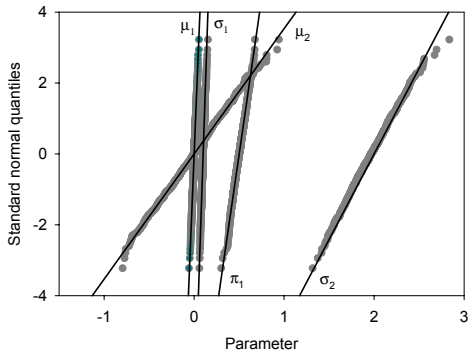


Figure 4.9: Normal QQ-plot of the simulated distribution of the LE at n=100 for all parameters of the mixture with parameter values $\mu_1 = 0$, $\sigma_1 = 0.1$, $\mu_2 = 0$, $\sigma_2 = 2$, $\pi_1 = 0.5$. *The straight line is the cumulative distribution function of the asymptotic distribution.*

totic properties. Although in case of an unstable sample, it is preferable to simulate the distribution of the LE as its asymptotic properties are still questionable. Further, it appears that the sample size required at which the distribution of the maximum closest to the true values is sufficiently close to the asymptotic distribution, is essentially the same as for the distribution of the LE.

We conclude this section with an important remark. In case of ML estimation, the sample size required to obtain precise estimates is usually derived from the variance of the asymptotic distribution. While for most distributions for which the likelihood function has a single maximum it is taken for granted that at this sample size the asymptotic properties can be used as well, this is mostly not true in case of likelihood estimation of general finite mixtures. To be precise, if n is large enough such that the LE is reliable, then in general estimates will be precise as well. The reverse, however, does not hold. To illustrate this, Table 4.10 gives the sample size required to obtain an asymptotic variance for all parameters which is at most

Table 4.10: Minimum sample size needed such that the variance of the asymptotic distribution of all parameters is less than 0.01.

	Group 1: μ_2				Group 2: σ_1				Group 3: π_1			
	1	2	3	4	1	0.5	0.2	0.1	0.8	0.6	0.4	0.2
n	100000	1000	90	30	200	85	80	80	70	100	160	400

0.01, for the 3 groups of parameter values. When compared with the results of Table 4.6, it can be seen that in many cases this calculated sample size is mostly not sufficient to obtain also a non spurious LE. As a result, when using general mixtures, in sample size calculations the spurious nature of the LE should be considered rather than its accuracy.

4.5 The (maximum) likelihood estimator versus the (maximum) likelihood estimator adapted for measurement error

In statistics, there has always been and, probably will always be, the discussion on the nature of measured data. Namely, should measured data be considered as "continuous" or rather as "discrete". The importance of this has to be looked for, among other things, in the use of the maximum likelihood method to fit a certain model to a sample. In applying the classical ML method, the likelihood function is composed of density contributions $f(x, \underline{\theta})$, with f the density of the model considered. Implicitly, in using this *density representation* for the likelihood, measurements are assumed to be continuous. Alternatively, to incorporate the discrete nature of measurements, an adapted ML method was proposed. Hereby, the density contributions of the likelihood are replaced through certain differences of the cumulative distribution function (Section 4.1.1). The likelihood has then a *cdf representation*. Due to the numerical complicatedness of this adapted ML method as compared to the classical version and the belief in

this classical ML method, the discussion faded into the background. Nevertheless, for the general finite mixture model, it was thrown up again by some authors with the argument that a more principled construction of the likelihood would solve the problems encountered with the classical method. As noted previously, this argument cannot be warranted.

In spite of this, the question whether these methods really give rise to other conclusions, is still left open. Namely, what are the differences between obtained parameter estimates, estimated standard errors and results of hypothesis tests. Are the differences large enough to influence the decision making process? Therefore, without entering into the discussion whether measurements have to be considered as continuous or discrete, we will compare these methods with respect to the most important point, i.e., inference results. Instead of confining oneself to only one of these two methods, we belief that one can confirm conclusions and turn this situation to its advantage. First, in Section 4.5.1 the objectives and the procedure followed are discussed. Main results are given in Section 4.5.2 in case the MLE exists and in Section 4.5.3 if not. We end this section with some conclusions.

4.5.1 Outline

The classical (maximum) likelihood method and several versions of the adapted (maximum) likelihood method will be compared for a normal mixture model. A distinction is drawn between the case of a common and a non-common scale parameter. Table 4.11 gives a sketch of the several estimators considered. In situation A, both the conditions of Wald and Cramér are fulfilled. Nearly always, a consistent global maximum exists, irrespective of the representation used for the likelihood function. For the density representation this leads to the MLE, while for the cdf representation, we introduced in Section 4.1.1 the estimators MLEδ and MLEδ^*. A third adapted maximum likelihood estimator, namely MLEδ_s, is added. For this estimator, the contributions to the likelihood function are given by

Table 4.11: Different estimators considered for the normal mixture model.

Representation	Estimator	
of likelihood	A: Common scale	B: Non-common scale
density	MLE	LE
cdf	MLEδ	LEδ
	MLEδ^*	LEδ^*
	MLEδ_s	LEδ_s

$F(y+\delta/2) - F(y-\delta/2)$ (i.e., likelihood function (4.1) is used). In contrast to the other two adapted estimators, the intervals $[y+\delta/2, y-\delta/2]$ are symmetrical around the observations. Its use will be explained further on. For situation B, attention is replaced from the global maximum to the largest local maximum. Mostly, the latter exists and is consistent. For the density representation the LE is obtained, while for the cdf representation we will refer to the estimators as LEδ, LEδ^* and LEδ_s. The latter are based on the same construction of the likelihood function as used for MLEδ, MLEδ^* and MLEδ_s. Note that situation A in combination with the cdf representation is essentially considered as "the" classical ML method. Although situation B is already an adjustment of the classical version, its density representation is still considered as a classical method.

Objective

Generally, it is assumed that the density representation has one major advantage as opposed to the cdf representation. Namely, numerically it is much easier to maximize a likelihood function with a density representation. We did not pay specific attention to this aspect, although it appeared directly from the simulations carried out. We will mainly focus on the differences between the estimators and will only discuss some computational issues when it was felt to be a problem. Stated below are the main questions to be answered.

- Is there an essential difference between situation A (i.e., existence of consistent MLE) and situation B (i.e., no consistent MLE) in comparing the classical (maximum) likelihood method and the adapted (maximum) likelihood methods? In other words, are the differences between the density representation and the cdf representation in case of a finite mixture with a common scale parameter comparable to the case of a general finite mixture?

- When are the two (maximum) likelihood methods, i.e., the classical and the adapted, comparable with regard to the decision making process (both in situation A and B)?

  - Is the obtained maximum with both methods essentially different or is it only a matter of precision? Does this depends on the δ-value used?

  - Is there a range of δ-values for which the estimates and estimated standard errors of the parameters are the same (upon a certain precision) and for which the outcome of hypothesis tests is similar. If so, are there any guidelines possible?

  - Is it possible to derive a general rule for all samples.

- What is the influence of the value of δ?

Procedure

Both a theoretical and an empirical approach are considered. The former is used to identify some quantities which could be related to possible differences between the methods. It is based on the fact that the density representation can be seen as the limit of the cdf representation for $\delta \to 0$,

namely

$$\lim_{\delta \to 0} \frac{F(y_\delta + \delta/2; \boldsymbol{\theta}) - F(y_\delta - \delta/2; \boldsymbol{\theta})}{\delta} = f(y; \boldsymbol{\theta}) \tag{4.6}$$

$$\lim_{\delta \to 0} \frac{F(\lceil y \rceil; \boldsymbol{\theta}) - F(\lfloor y \rfloor; \boldsymbol{\theta})}{\delta} = f(y; \boldsymbol{\theta}) \tag{4.7}$$

The practical method consists in comparing the results of the estimation methods for a number of samples where the parameter δ is varied from a small to a large value. The aim is to check some theoretical results, to state some guidelines and to identify some additional relations. It is carried out for a two-component normal mixture model both with a common and non-common scale parameter. For each model, estimates of the model parameters (μ_1, μ_2, σ, and π_1) or ($\mu_1, \mu_2, \sigma_1, \sigma_2$, and π_1) and the quantiles $t_{0.01}, t_{0.001}$ as well as their estimated standard errors are compared. The quantiles are studied due to their importance in reliability analysis. For the general normal mixture model, this comparison includes also the value of the likelihood ratio test statistic (LRT) to test the null hypothesis of a common scale parameter against the alternative of an unequal scale parameter. Maximization is performed using the EM-algorithm. It is followed by a few steps with the NR-method to derive standard errors. The tolerance used is 1e-8 on the gradient of the logarithm of the likelihood function. We are quite confident to have a guaranteed convergence to a global or largest local maximum. This result is based on the use of specific starting values and explained in detail in the next chapter. Note that differences between the density and cdf representation with regard to computational aspects, reveal itself in the use of the EM-algorithm. The latter is due to its double iterative character for the cdf representation much slower than for the density representation (Section 5.5.1).

4.5.2 The maximum likelihood method versus the maximum likelihood method adapted for measurement error

Theoretical aspects

If the adapted likelihood function is composed of "symmetrical" cdf contributions, the classical ML method can be regarded as an approximation of the adapted ML method. Namely, given any density function $f(y;\boldsymbol{\theta})$ and corresponding cdf $F(y;\boldsymbol{\theta})$, the following Taylor expansion exists:

$$\begin{aligned}
&F(y+\delta/2;\boldsymbol{\theta}) - F(y-\delta/2;\boldsymbol{\theta}) \\
&= F(y;\boldsymbol{\theta}) + \frac{\delta/2}{1!}f(y;\boldsymbol{\theta}) + \frac{(\delta/2)^2}{2!}f'(y;\boldsymbol{\theta}) + \frac{(\delta/2)^3}{3!}f''(y;\boldsymbol{\theta}) + \ldots \\
&- \left(F(y;\boldsymbol{\theta}) - \frac{\delta/2}{1!}f(y;\boldsymbol{\theta}) + \frac{(\delta/2)^2}{2!}f'(y;\boldsymbol{\theta}) - \frac{(\delta/2)^3}{3!}f''(y;\boldsymbol{\theta}) + \ldots\right) \\
&= \delta f(y;\boldsymbol{\theta}) + 2\frac{(\delta/2)^3}{3!}f''(y;\boldsymbol{\theta}) + 2\frac{(\delta/2)^5}{5!}f''''(y;\boldsymbol{\theta}) + \ldots
\end{aligned} \tag{4.8}$$

Based on this, it can be concluded that the approximation

$$\frac{F(y+\delta/2;\boldsymbol{\theta}) - F(y-\delta/2;\boldsymbol{\theta})}{\delta} \approx f(y;\boldsymbol{\theta}) \tag{4.9}$$

holds if the higher order terms in (4.8) are negligible. If so, the classical ML method and the adapted ML method with a "symmetrical" cdf representation, will usually give rise to similar parameter estimates. The question is then when these higher order terms can be ignored.

For a finite mixture model with a location-scale distribution as component density and a common scale parameter, it is not sufficient to take $\delta/2$ smaller then one, to have a workable approximation at each point in the parameter space. Namely, for values of the scale parameter smaller than δ, the terms $f'(y;\boldsymbol{\theta})$, $f''(y;\boldsymbol{\theta}), \ldots$ cannot be neglected. For example, for a two-component mixture model, the Taylor expansion (4.8) can be rewritten

as:

$$\pi_1 \left[\Phi\left(\frac{y-\mu_1}{\sigma} + \frac{\delta/2}{\sigma}\right) - \Phi\left(\frac{y-\mu_1}{\sigma} - \frac{\delta/2}{\sigma}\right)\right] + (1-\pi_1)\left[\Phi\left(\frac{y-\mu_2}{\sigma} + \frac{\delta/2}{\sigma}\right) - \Phi\left(\frac{y-\mu_2}{\sigma} - \frac{\delta/2}{\sigma}\right)\right]$$
$$= \pi_1 \left[\left(\frac{\delta}{\sigma}\right)\phi\left(\frac{y-\mu_1}{\sigma}\right) + \frac{1}{2^2\,3!}\left(\frac{\delta}{\sigma}\right)^3 \phi''\left(\frac{y-\mu_1}{\sigma}\right) + \ldots\right] + (1-\pi_1)\left[\left(\frac{\delta}{\sigma}\right)\phi\left(\frac{y-\mu_2}{\sigma}\right) + \frac{1}{2^2\,3!}\left(\frac{\delta}{\sigma}\right)^3 \phi''\left(\frac{y-\mu_2}{\sigma}\right) + \ldots\right], \tag{4.10}$$

with $\Phi(x)$ the standard cdf of the location-scale distribution and $\phi(x)$ its corresponding density function. Since all the derivatives of $\phi(y-\mu/\sigma)$ goes to zero for $\sigma \to 0$, it follows from this derivation that the quantity δ/σ marks out the regions in the parameter space where the approximation will work and where not. Namely, for values of σ such that δ/σ is sufficiently smaller than one, the other terms in the expansion can be mostly neglected. But, in case δ/σ is not smaller than one, certainly higher order terms of the Taylor expansion are required to obtain a satisfactory approximation.

As a result, it is likely that the quantity $\delta/\hat{\sigma}^{MLE}$, with σ^{MLE} the MLE of the common scale parameter, determines whether the estimates MLE and MLEδ_s and other inference results are comparable or not. Indeed, if this quantity is too close to or larger than one, the approximation (4.9) does not hold in the neighborhood of the MLE, while the opposite is true for values sufficiently smaller than one.

Still, this is not entirely correct for the adapted ML methods, giving rise to the estimators MLEδ and MLEδ^*. The derivation leading to the Taylor expansion (4.10) is for the "symmetrical" differences $F(y+b)-F(y-a)$, with y the exact midpoint of the interval $[y-a, y+b]$. In contrast, the intervals for the estimators MLEδ and MLEδ^* are mostly not symmetrical around the observations y. The only exception is when the given observations

are already rounded off to the value of δ. Then, the estimators MLEδ and MLEδ_s will be equal and MLEδ^* cannot be derived. For a "non-symmetrical" interval $[y-a, y+b]$, the Taylor expansion of a cdf likelihood contribution becomes:

$$\begin{aligned} & F(y+b;\boldsymbol{\theta}) - F(y-a;\boldsymbol{\theta}) \\ & = \delta f(y;\boldsymbol{\theta}) + \frac{1}{2!}(b-a)\delta f'(y;\boldsymbol{\theta}) + \frac{a^3+b^3}{3!} f''(y;\boldsymbol{\theta}) + \ldots \end{aligned} \tag{4.11}$$

As noted, compared to the expansion (4.8), a term of the second degree is present. Although, it disappears as δ decreases, smaller values of δ/σ are required (in comparison with the symmetrical case), to sufficiently approximate the likelihood with cdf representation through the likelihood with density representation. As such, it is expected that the differences between the estimators MLE and MLEδ (or MLEδ^*) will be somewhat larger than between the estimators MLE and MLEδ_s at the same value of δ.

Empirical results

Several samples, both real and simulated, are fitted to a two-component normal mixture model with common scale parameter, using the different estimation methods for a whole range of δ-values. Only the results for one sample, i.e., the most problematic one, are shown in Table 4.12. It tabulates the largest absolute difference found between estimates and estimated standard errors as well as the parameter where this value is reached. The estimator MLEδ_s is added to look at the differences between the use of a symmetrical and a non-symmetrical interval around the observations for the adapted ML methods.

The sample used for Table 4.12 has size 50 and is simulated from a normal distribution with parameter values $\mu = 0$ and $\sigma = 0.5$. It is one of the few samples found for which the likelihood function for a two-component normal mixture has more than one maximum. Indeed, for most samples the likelihood function (irrespective of its representation) has, except for some

Table 4.12: Maximum absolute difference between parameter estimates (first table) and estimated standard errors (second table) of the different methods. The parameter where the difference is largest, is indicated between brackets.

δ	$\frac{\delta}{\hat{\sigma}^{MLE}}$	MLE - MLEδ	MLE - MLEδ^*	MLEδ - MLEδ^*	MLE - MLEδ_s
1e-6	1.94e-6	2.10e-7 ($t_{0.01}$)	4.57e-7 ($t_{0.001}$)	6.59e-7 ($t_{0.001}$)	1.27e-10 ($t_{0.01}$)
1e-5	1.94e-5	2.04e-6 (μ_1)	6.08e-6 (μ_1)	8.12e-6 (μ_1)	2.92e-11 ($t_{0.001}$)
1e-4	1.94e-4	1.69e-5 (μ_1)	5.59e-5 ($t_{0.001}$)	7.26e-5 ($t_{0.001}$)	1.33e-9 ($t_{0.001}$)
1e-3	1.94e-3	3.98e-4 ($t_{0.01}$)	3.76e-4 (μ_1)	7.56e-4 (μ_1)	1.33e-7 ($t_{0.001}$)
1e-2	1.94e-2	3.48e-3 (μ_1)	5.47e-3 (μ_1)	8.95e-3 (μ_1)	1.33e-5 ($t_{0.001}$)
0.1	0.194	5.21e-2 ($t_{0.01}$)	1.61e-2 ($t_{0.001}$)	6.21e-2 ($t_{0.001}$)	1.33e-3 ($t_{0.001}$)
0.2	0.389	7.63e-2 (μ_1)	9.25e-2 (μ_1)	0.169 (μ_1)	5.35e-3 ($t_{0.001}$)
0.4	0.777	1.55 (μ_1)	5.52e-2 ($t_{0.01}$)	1.56 (μ_1)	2.16e-2 ($t_{0.001}$)
0.6	1.17	1.61 (μ_1)	0.690 (μ_1)	1.17 (μ_2)	4.95e-2 ($t_{0.001}$)
1	1.94	0.202 ($t_{0.01}$)	-	-	0.152 ($t_{0.001}$)

δ	# I	MLE - MLEδ	MLE - MLEδ^*	MLEδ - MLEδ^*	MLE - MLEδ_s
1e-6	50	2.95e-8 (μ_1)	1.58e-7 ($t_{0.01}$)	1.45e-7 ($t_{0.01}$)	2.02e-10 (μ_1)
1e-5	50	1.45e-6 ($t_{0.01}$)	3.06e-6 (μ_1)	3.99e-6 (μ_1)	4.49e-11 (μ_1)
1e-4	50	1.06e-5 ($t_{0.01}$)	2.12e-5 ($t_{0.01}$)	3.18e-5 ($t_{0.01}$)	9.01e-10 ($t_{0.01}$)
1e-3	50	2.01e-4 (μ_1)	2.11e-4 ($t_{0.01}$)	3.86e-4 (μ_1)	9.01e-8 ($t_{0.01}$)
1e-2	44/42	2.31e-3 (μ_1)	4.50e-3 (μ_1)	6.82e-3 (μ_1)	9.01e-6 ($t_{0.01}$)
0.1	20	2.19e-2 (μ_1)	2.18e-2 (μ_1)	1.35e-2 ($t_{0.01}$)	8.94e-4 ($t_{0.01}$)
0.2	12/14	2.63e-2 ($t_{0.01}$)	9.71e-2 (μ_1)	0.114 ($t_{0.01}$)	3.48e-3 ($t_{0.01}$)
0.4	8	0.561 ($t_{0.01}$)	4.82e-2 (μ_1)	0.551 ($t_{0.01}$)	1.23e-2 ($t_{0.01}$)
0.6	6	0.589 ($t_{0.01}$)	0.901 (μ_1)	1.43 (μ_1)	2.13e-2 ($t_{0.01}$)
1	5/**4**	0.189 (μ_1)	-	-	4.28e-2 (μ_1)

Note: # I is the the number of different intervals in the sample when using the adapted ML methods related to the estimators MLEδ and MLEδ^*. In case of an unequal number between the two methods, the first number refers to the sample corresponding to the estimator MLEδ.

boundary solutions corresponding to a single normal distribution, only one maximum. The likelihood function for this sample has two maxima, both situated at the top. The global maximum has likelihood value -41.97, the other maximum -42.86. The ML estimates are $\hat{\mu}_1 = -1.66$ (0.620), $\hat{\mu}_2 = 0.110$ (7.46e-2), $\hat{\sigma} = 0.515$ (5.52e-2) and $\hat{\pi}_1 = 0.0200$ (2.18e-2). From this table, several things can be noted.

- For small values of $\delta/\hat{\sigma}^{\mathrm{MLE}}$, there is a negligible difference between all methods, regarding estimates and estimated standard errors. Differences are only a matter of precision (i.e., the number of significant figures which is equal among estimates). It improves as δ decreases in value.

- When the value of $\delta/\hat{\sigma}^{\mathrm{MLE}}$ is smaller than about 0.5, differences between ML estimates (standard errors) and the estimates MLEδ_s (standard errors) are unimportant. Once this quantity becomes larger, the global maxima of the two likelihood functions are still situated in the same region, but differences between estimates are not negligible anymore.

- For the two adapted ML methods with a non-symmetrical contribution to the likelihood function, the value of $\delta/\hat{\sigma}^{\mathrm{MLE}}$ has to be smaller than about 0.1, to obtain estimates (and estimated standard errors) which are equal to the MLEs for at least the first two figures. The influence on the precision of the additional second degree term in the Taylor expansion is clear, as this precision for the adapted ML method with symmetrical contributions is noticeably larger at the same value of δ. Once $\delta/\hat{\sigma}^{\mathrm{MLE}}$ is larger than 0.1, differences become marked. In addition, sometimes the global maxima of the likelihood functions are essentially different, i.e., they are not situated anymore in the same region of the parameter space. For example, for $\delta = 0.4$, the two maxima of the adapted likelihood function corresponding to MLEδ,

have changed from order. As a result, MLEδ for $\delta = 0.2$ and MLEδ for $\delta = 0.4$ correspond to totally different maxima. For a value of $\delta = 0.4$ inference results based on the MLE (or MLEδ^*) and on the MLEδ can be seriously different.

- The number of different intervals in the discretised samples has mostly no effect on the difference with the classical ML method. Except when the number of adjacent intervals reduces to 4 or lower, there exist no consistent adapted ML estimator anymore.

The results for another sample with a large value of $\hat{\sigma}^{\mathrm{MLE}}$ are given in Table B.6 in Section B.3 of the appendix. In general, results for all other samples are similar, apart from one thing. Namely, for samples for which the likelihood function has only one maximum, no change of maxima will arrive for the adapted ML methods. As δ increases, differences enlarges, but global maxima stay in the same region (like the behavior of MLEδ_s for the sample discussed). The results can be easily extended to an M-component mixture or to other (log)location-scale distributions as component density.

4.5.3 The likelihood method versus the likelihood method adapted for measurement error

Surface of an adapted likelihood function

Before we compare the several estimation methods, we first take a closer look at the estimation problems encountered when using the adapted ML methods. As mentioned, these problems are similar in nature as those of the classical ML method. Namely, for a general finite mixture model a consistent adapted MLE does mostly not exist and spurious maxima, in particular distinct spurious maxima, do not automatically disappear through the introduction of a measurement error.

Table 4.13 gives for a range of δ-values, some maxima of the adapted likelihood function corresponding to LEδ for a stable sample, in case of a

Table 4.13: Some maxima of the adapted likelihood function corresponding to LEδ in case of a two-component normal mixture model, for the large stable resistor sample (n=125). The first row at each value of δ gives always the largest "maximum" found.

δ	$\hat{\mu}_1$	$\hat{\sigma}_1$	$\hat{\mu}_2$	$\hat{\sigma}_2$	$\hat{p}_1$	lnL
1e-8	5.649	$\to 0$	6.786	0.450	0.00800	-2367.02
	7.022	0.251	6.164	0.236	0.714	-2368.69
	6.745	*3.55e-5*	*6.777*	*0.464*	*0.0159*	*-2375.13*
1e-6	6.164	0.236	7.022	0.251	0.286	-1793.04
	5.649	$\to 0$	6.786	0.450	0.00800	-1795.97
	6.745	*3.55e-5*	*6.777*	*0.464*	*0.0159*	*-1799.48*
1e-3	6.164	0.236	7.022	0.252	0.286	-929.575
	6.745	$\to 0$	6.777	0.463	0.0151	-939.765

Note: The symbol $\to 0$ for one of the scale parameters refers to a maximum attained outside the parameter space. Maxima in italic are distinct spurious.

general two-component normal mixture. It allows to form an idea of the surface of the adapted likelihood function. Among the maxima tabulated, the largest one, whether attained in or outside the parameter space, is always given. Tables 4.14a and 4.14b report the same but for highly unstable samples and for other adapted likelihood functions. Although, for each sample, only one of the three adapted ML methods is used, results are similar if one of the two other adapted ML methods is considered. Further, the examples given, illustrate quite well the situation encountered for most other samples. Below, our main findings are summarized. They hold for any of the three adapted likelihood estimators given in Table 4.11.

- If δ is sufficiently small, then for each sample, regardless of its stability, the adapted likelihood function attains a maximum, outside the parameter space, in the points $\mu_1 \in]a_i, b_i[$, $\sigma_1 = 0$, $\pi_1 = n_i/n$ and with μ_2 and σ_2 the adapted MLEs when fitting the sample without the i^{th} interval to a single component distribution. Hereby, a_i and b_i are the limit points of the intervals of the discretised sample and n_i

the original number of duplicates in the continuous sample. For all samples looked at, an upper limit for δ was found so that the adapted likelihood in case of a smaller value for δ attained its supremum in one of these points or "singularities". Since these points do not belong to the parameter space, no global maximum exists at these δ-values. But even if they would belong to the parameter space, they are no consistent estimates. For example, for the sample in Table 4.14a this is true for δ = 1e-4 or smaller, for the sample in Table 4.14b this holds for δ = 1e-6 and for the sample in Table 4.13 a value as small as 1e-8 is required. Note that these "maxima" are equivalent to the singularities of the density likelihood. Only here, μ can take on a range of values and the likelihood value of such a singularity is bounded. But, like the singularities of the density likelihood, also here they should be ignored.

- For small values of δ distinct spurious maxima are present as many as for the likelihood function with density representation. For larger values, they are still present but their number decreases. Note that only a few of them are included in the tables.

- As δ increases in value, the singularities fade way, i.e., their likelihood value diminish in comparison to the likelihood value of other maxima, or they disappear. This can be observed, for example, by considering the position of the largest maximum at the smallest value for δ, for the larger values of δ. Still, other difficulties appear as both distinct spurious maxima and, latter on, maxima with one of the scale parameters quite small, can change into new problematic points. In particular, there are two kinds of problematic points for the adapted likelihood function (in contrast to only one for the density likelihood). On the one hand, there are the "maxima" equivalent to the infinities of the density likelihood. A distinct spurious maximum will turn into such a maximum if the few data points belonging to one component,

will be situated in the same interval of the discretised sample. On the other hand, these few data points could also belong to two adjacent intervals of the discretised sample. As such, they cause an increase of the likelihood for μ_i going to the common point of these two intervals and σ_i going to 0, but often the maximum value of the likelihood is not reached within any point (in or outside the parameter space). This situation becomes problematic when it leads to the supremum of the likelihood. Not only, no global maximum then exists, even not outside the parameter space, but also due to the limits on the numerical accuracy, apparent maxima will be identified. This was noticed, for example, for the binned sample in Table 4.2. In the following, for reasons of simplicity, we will use for both kind of situations the term singularity and maximum attained outside the parameter space. No distinction will be made anymore, since both do not lead to a consistent root and should be discarded. Due to the fact that these singularities are sometimes present at the top of the adapted likelihood function, a classical adapted MLE does not exist.

- For stable samples, from a certain value for δ on, the consistent root of the LEQs corresponds mostly to the global maximum, until the number of intervals in the discretised sample has become to small. Although, distinct spurious maxima and other singularities are still present, they are not situated at the top of the likelihood function (Table 4.13).

- For highly unstable samples, distinct spurious maxima at the top of the adapted likelihood function for small values of δ, become often a singularity where the likelihood attains its supremum for larger values of δ. For example in the Tables 4.14b and 4.14a, the second maximum at the smallest value for δ, becomes the largest maximum and a singularity at one value larger for δ. It depends on the value of δ and the way of constructing the adapted likelihood function, whether

there exists a global maximum, and which maximum is the largest one (Tables 4.14b and 4.14a). Unstable samples behave rather like stable samples, although for large values of δ there could be a change of maxima at the top of the adapted likelihood.

Also for these adapted ML methods there exists an alternative of looking at the largest local maximum instead of the global maximum to find a consistent estimate. The only difference with the density likelihood is that the latter is always unbounded and therefore the global maximum never exists. In contrast, for stable and many unstable samples, from some value of δ onwards, the adapted likelihood function will mostly have a global maximum. In this case, it equals the largest local maximum, but this does not change the strategy of searching for the largest local maximum.

Comparing different estimators

Similarly as for the two-component mixture with common scale parameter, it can be derived that $\Delta = \delta / \min(\sigma_1, \sigma_2)$ is an important quantity to look at when the classical likelihood and the adapted likelihood methods are compared for the general two-component mixture model. The only difference in the Taylor expansion (4.10) is the occurrence of a different scale parameter in the first and second component. As a result, we expect that for $\hat{\Delta} = \delta / \min(\hat{\sigma}_1^{\mathrm{LE}}, \hat{\sigma}_2^{\mathrm{LE}})$ sufficiently smaller than one, the likelihood method and the adapted likelihood method with symmetrical contributions will lead to similar conclusions. For the other two adapted likelihood methods quite likely a smaller value of $\hat{\Delta}$ is needed to obtain negligible differences.

To check these findings, we fitted a general two-component normal mixture model to a series of samples using all three adapted likelihood methods and this for different values of δ. Here, the results of two samples are given, i.e., a stable and a highly unstable sample. Further, in Chapter 6, a comparison for a general M-component normal mixture with more than two components is worked out.

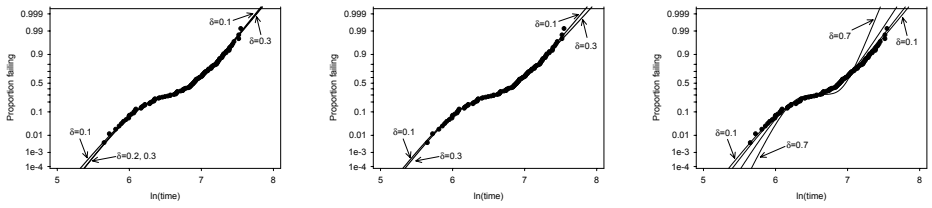
δ	$\hat{\mu}_1$	$\hat{\sigma}_1$	$\hat{\mu}_2$	$\hat{\sigma}_2$	$\hat{p}_1$	lnL
1e-4	4.222	$\to 0$	4.099	0.433	0.0768	-242.082
	4.560	*0.000299*	*4.071*	*0.413*	*0.0764*	*-245.950*
	3.600	0.285	4.340	0.213	0.314	-250.614
	3.953	0.384	4.524	0.0869	0.728	-250.659
1e-3	4.560	$\to 0$	4.071	0.413	0.0765	-185.659
	4.222	$\to 0$	4.099	0.433	0.0761	-186.804
	3.209	*0.0150*	*4.182*	*0.342*	*0.0757*	*-189.271*
	3.601	0.285	4.340	0.213	0.314	-190.753
	3.952	0.385	4.524	0.0872	0.727	-190.801
1e-2	*3.210*	*0.0147*	*4.181*	*0.342*	*0.0757*	*-129.365*
	4.555	$\to 0$	4.073	0.413	0.0724	-130.277
	3.621	0.295	4.346	0.209	0.328	-130.837
	3.952	0.384	4.522	0.0879	0.727	-130.928
0.1	3.942	0.386	4.519	0.0816	0.713	-70.987
	4.499	0.00619	4.018	0.417	0.190	-71.799
	3.200	$\to 0$	4.179	0.347	0.0726	-71.925

(a) Adapted likelihood function corresponding to LEδ^* for a small highly unstable failure time sample (n=26).

δ	$\hat{\mu}_1$	$\hat{\sigma}_1$	$\hat{\mu}_2$	$\hat{\sigma}_2$	$\hat{p}_1$	lnL
1e-6	3.558	$\to 0$	4.802	0.484	0.0147	-977.339
	5.329	*8.22e-5*	*4.767*	*0.502*	*0.0293*	*-980.646*
	5.340	0.146	4.611	0.446	0.237	-985.049
1e-4	5.327	$\to 0$	4.767	0.502	0.0293	-667.490
	3.558	$\to 0$	4.802	0.484	0.0147	-668.792
	5.445	*0.00281*	*4.756*	*0.495*	*0.0405*	*-671.409*
	5.340	0.146	4.611	0.446	0.237	-671.898
1e-2	5.445	$\to 0$	4.755	0.495	0.0412	-357.829
	5.340	0.146	4.611	0.446	0.237	-358.746
	5.329	$\to 0$	4.770	0.502	0.0252	-360.834
	3.558	$\to 0$	4.802	0.485	0.0144	-360.225
0.1	5.340	0.144	4.611	0.445	0.237	-202.170
	5.385	0.0416	4.710	0.483	0.110	-202.929
	5.444	$\to 0$	4.743	0.491	0.0593	-204.139

(b) Adapted likelihood function corresponding to LEδ_s, for the highly unstable interconnect sample (n=68).

Table 4.14: Some maxima of an adapted likelihood function for two highly unstable samples in case of the two-component normal mixture model. The first row at each value of δ gives always the largest "maximum" found.



(a) LEδ fits for $\delta = 0.1$, 0.2 and 0.3.

(b) LEδ^* fits for $\delta = 0.1$ and 0.3.

(c) LEδ_s fits for $\delta = 0.1$, 0.3, 0.5 and 0.7.

Figure 4.10: QQ-plot of the resistor sample supplemented with several fits.

A stable sample The sample considered is the resistor sample introduced in Section 2.1.1. It is a large stable sample of size 125. One maximum is dominating the (density) likelihood function. The likelihood estimates are $\hat{\mu}_1 = 6.164$ (0.0638), $\hat{\sigma}_1 = 0.236$ (0.0460), $\hat{\mu}_2 = 7.022$ (0.0354), $\hat{\sigma}_2 = 0.251$ (0.0268), $\hat{\pi}_1 = 0.286$ (0.0538). The sample is estimated with all three adapted likelihood methods for a range of δ values. Table 4.15a compares the parameter estimates and Table 4.15b the estimated standard errors, while Table 4.15c gives the values of the likelihood ratio test (LRT) statistic and corresponding p-value to test the null hypothesis $\mathrm{H}_0 : \sigma_1 = \sigma_2$ against the alternative $\mathrm{H}_A : \sigma_1 \neq \sigma_2$. A number of conclusions can be drawn.

- If $\hat{\Delta}$ is smaller than about 0.5-1, differences among estimates, estimated standard errors and p-values are negligible between the classical likelihood method and the "symmetrical" adapted likelihood method. The more this value decreases, the smaller the differences become. For larger values, differences cannot be neglected anymore, but final conclusions are mostly comparable. Only the precision diminishes. Note that the estimates $\hat{\sigma}_1^{\mathrm{LE}_s}$ and $\hat{\sigma}_2^{\mathrm{LE}_s}$ go to zero for increasing values of δ. This is illustrated in Figure 4.10c.

δ	$\hat{\Delta}$	LE - LEδ	LE - LEδ^*	LEδ - LEδ^*	LE - LEδ_s
1e-6	4.23e-6	1.74e-7 ($t_{0.001}$)	4.36e-8 (μ_2)	2.09e-7 ($t_{0.001}$)	2.00e-11 (σ_2)
1e-5	4.23e-5	1.20e-6 (μ_1)	8.42e-7 (μ_1)	2.04e-6 (μ_1)	4.91e-11 ($t_{0.001}$)
1e-4	4.23e-4	3.78e-6 (μ_1)	2.25e-5 ($t_{0.001}$)	2.30e-5 ($t_{0.001}$)	4.76e-9 ($t_{0.001}$)
1e-3	4.23e-3	3.78e-5 (μ_1)	1.45e-4 ($t_{0.001}$)	1.66e-4 ($t_{0.001}$)	4.76e-7 ($t_{0.001}$)
1e-2	4.23e-2	6.78e-4 (μ_2)	8.72e-4 (μ_1)	1.08e-3 (μ_2)	4.76e-5 ($t_{0.001}$)
0.1	0.423	2.17e-2 (μ_1)	1.42e-2 (μ_1)	3.57e-2 (μ_1)	4.78e-3 ($t_{0.001}$)
0.2	0.846	2.26e-2 ($t_{0.001}$)	1.59e-2 ($t_{0.001}$)	3.86e-2 ($t_{0.001}$)	1.9e-2 ($t_{0.001}$)
0.3	1.27	5.83e-2 ($t_{0.01}$)	6.16e-2 (μ_1)	2.66e-2 ($t_{0.001}$)	4.5e-2 ($t_{0.001}$)
0.5	2.12	-	-	-	0.14 ($t_{0.001}$)

(a) Maximum absolute difference between parameter estimates. The parameter where the difference is largest, is indicated between brackets.

δ	# I	LE - LEδ	LE - LEδ^*	LEδ - LEδ^*	LE - LEδ_s
1e-6	125	3.64e-8 (μ_1)	3.37e-8 (μ_1)	2.41e-8 ($t_{0.01}$)	8.12e-12 ($t_{0.001}$)
1e-5	126	1.42e-7 ($t_{0.01}$)	1.97e-7 (μ_1)	2.14e-7 (μ_2)	7.85e-12 ($t_{0.001}$)
1e-4	125	3.98e-6 (μ_1)	6.08e-6 ($t_{0.001}$)	4.45e-6 ($t_{0.001}$)	7.78e-10 ($t_{0.001}$)
1e-3	118/121	2.03e-5 (μ_1)	8.11e-5 (μ_1)	6.50e-5 ($t_{0.001}$)	7.78e-8 ($t_{0.001}$)
1e-2	87/84	1.31e-4 (μ_1)	6.99e-4 (μ_1)	8.30e-4 (μ_1)	7.78e-6 ($t_{0.001}$)
0.1	20	1.08e-2 (μ_1)	1.80e-3 (μ_1)	9.09e-3 (μ_1)	7.89e-4 ($t_{0.001}$)
0.2	11/10	1.59e-2 (μ_1)	2.86e-3 ($t_{0.001}$)	1.52e-2 (μ_1)	3.3e-3 ($t_{0.001}$)
0.3	7/8	1.76e-2 ($t_{0.01}$)	6.56e-3 (μ_1)	2.86e-2 (μ_1)	8.0e-3 ($t_{0.001}$)
0.5	**5/5**	-	-	-	3.0e-2 ($t_{0.001}$)

(b) Maximum absolute difference between estimated standard errors. The parameter where the difference is largest, is indicated between brackets.

		LRT-value				p-value			
δ	$\hat{\Delta}$	LE	LEδ	LEδ^*	LEδ_s	LE	LEδ	LEδ^*	LEδ_s
1e-6	4.23e-6	0.0571	0.0571	0.0571	0.0571	0.811	0.811	0.811	0.811
1e-5	4.23e-5	0.0571	0.0571	0.0571	0.0571	0.811	0.811	0.811	0.811
1e-4	4.23e-4	0.0571	0.0571	0.0573	0.0571	0.811	0.811	0.811	0.811
1e-3	4.23e-3	0.0571	0.0575	0.0562	0.0571	0.811	0.810	0.813	0.811
1e-2	4.23e-2	0.0571	0.0573	0.0494	0.0571	0.811	0.811	0.824	0.811
0.1	0.423	0.0571	5.99e-3	0.112	0.0571	0.811	0.938	0.738	0.811
0.2	0.846	0.0571	0.0930	0.131	0.0565	0.811	0.760	0.717	0.812
0.3	1.27	0.0571	0.0261	0.186	0.0534	0.811	0.872	0.666	0.817
0.5	2.12	0.0571	-	-	0.0217	0.811	-	-	0.883

(c) Value of the LRT statistic and corresponding p-value ($H_0 : \sigma_1 = \sigma_2$, $H_A : \sigma_1 \neq \sigma_2$). The LRT statistic is assumed to have a χ^2 distribution with 1 df.

Table 4.15: Comparison of the different likelihood methods for the resistor sample.

- For the other two adapted likelihood methods, the value of $\hat{\Delta}$ certainly has to be smaller than 0.1, to obtain unimportant differences with the classical likelihood method. Nevertheless, although faster than for the "symmetrical" adapted likelihood method, also for these methods only the precision diminishes. As an example, Figures 4.10a and 4.10b give some fits of the adapted likelihood estimates for values of $\hat{\Delta}$ larger than 0.1. As noted, all fits are still comparable.

In summary, for all methods and values of δ only one maximum dominates the adapted likelihood function. It is situated in the neighborhood of the dominating maximum of the density likelihood. The findings drawn here, can be extended to most other stable samples and even to many unstable samples with only a few maxima, not distinct spurious, at the top of the likelihood function. Note that the results are similar to the case where the MLE exists.

A (highly) unstable sample Table 4.16 summarizes the results for the interconnect sample (Section 2.1.2). This sample of size 68 is highly unstable. The likelihood estimate is a distinct spurious maximum: $\hat{\mu}_1 = 4.767(5.84\text{e-}5)$, $\hat{\sigma}_1 = 0.502(4.12\text{e-}5)$, $\hat{\mu}_2 = 5.329(0.0618)$, $\hat{\sigma}_2 = 8.22\text{e-}5(0.0437)$, $\hat{\pi}_1 = 0.971\ (0.0205)$. Note that inference results (of the likelihood method) are not reliable here due to the instability of the sample. However, one can still compare the different estimation methods. As observed from Table 4.16, for a value of $\hat{\Delta}$ not larger than about 0.1, results from all methods are comparable. Again differences between the classical method and the symmetrical adapted likelihood method are considerably smaller as compared to the other adapted likelihood methods. In addition, as long as $\hat{\Delta}$ is not larger than about 1, differences are only related to a diminishing precision, i.e., the largest maximum of the adapted likelihood functions is similar to the largest maximum of the classical likelihood function. Once $\hat{\Delta}$ is larger than about one, the dissimilarity between likelihood estimates and adapted likelihood

δ	$\hat{\Delta}$	LE - LEδ	LE - LEδ^*	LEδ - LEδ^*	LE - LEδ_s
1e-6	1.22e-2	3.67e-7 (π_1)	4.86e-7 (π_1)	8.53e-7 (π_1)	5.07e-10 (σ_2)
1e-5	0.122	4.75e-6 (π_1)	3.78e-6 (π_1)	8.53e-6 (π_1)	5.07e-8 (σ_2)
5e-5	0.608	1.23e-5 (π_1)	1.71e-5 (μ_2)	2.50e-5 (μ_2)	1.31e-6 (σ_2)
1e-4	1.22	3.02e-5 (π_1)	0.26 (μ_2)	0.26 (μ_2)	5.91e-4 (σ_2)
1e-3	12.2	0.117 (μ_2)	0.117 (μ_2)	1.48e-4 ($t_{0.001}$)	0.942 (π_1)
1e-2	122	0.210 (π_1)	0.210 (π_1)	1.22e-3 (μ_2)	0.208 (π_1)
0.1	1.22e+4	0.228 (π_1)	0.183 (π_1)	6.46e-2 (μ_2)	0.208 (π_1)

(a) Maximum absolute difference between parameter estimates. The parameter where the difference is largest, is indicated between brackets.

δ	$\hat{\Delta}$	LE - LEδ	LE - LEδ^*	LEδ - LEδ^*	LE - LEδ_s
1e-6	68	1.54e-6 (μ_1)	2.03e-7 (μ_1)	3.57e-7 (μ_1)	2.54e-10 (σ_1)
1e-5	68	1.99e-6 (μ_1)	1.58e-6 (μ_1)	3.57e-6 (μ_1)	2.55e-8 (σ_1)
5e-5	68	5.18e-6 (μ_1)	5.18e-6 (μ_1)	3.32e-6 (μ_2)	6.72e-7 (σ_1)
1e-4	68	1.25e-5 (μ_1)	8.01e-4 (μ_2)	8.01e-4 (μ_2)	3.32e-6 (σ_1)
1e-3	65	4.51e-3 (π_1)	4.52e-3 (π_1)	5.58e-5 (μ_1)	0.143 ($t_{0.001}$)
1e-2	57/61	0.151 ($t_{0.001}$)	0.151 ($t_{0.001}$)	5.36e-4 (π_1)	0.310 ($t_{0.001}$)
0.1	18/20	0.156 ($t_{0.001}$)	0.150 ($t_{0.001}$)	4.07e-2 (π_1)	0.314 ($t_{0.001}$)

(b) Maximum absolute difference between estimated standard errors. The parameter where the difference is largest, is indicated between brackets.

		LRT-value				p-value			
δ	$\hat{\Delta}$	LE	LEδ	LEδ^*	LEδ_s	LE	LEδ	LEδ^*	LEδ_s
1e-6	1.22e-2	10.854	10.864	10.840	10.854	9.86e-4	9.816e-4	1.26e-3	9.86e-4
1e-5	0.122	10.854	10.721	10.963	10.854	9.86e-4	1.06e-3	9.29e-4	9.86e-4
5e-5	0.608	10.854	11.220	11.221	10.854	9.86e-4	8.09e-4	8.09e-4	9.86e-4
1e-4	1.22	10.854	10.079	8.361	10.861	9.86e-4	1.50e-3	3.83e-3	9.82e-4
1e-3	12.2	10.854	2.970	2.820	3.064	9.86e-4	8.48e-2	9.31e-2	8.00e-2
1e-2	122	10.854	2.100	2.165	2.046	9.86e-4	0.147	0.141	0.153
0.1	122	10.854	1.301	3.056	2.047	9.86e-4	0.254	8.00e-2	0.153

(c) Value of the LRT statistic and corresponding p-value ($H_0 : \sigma_1 = \sigma_2$, $H_A : \sigma_1 \neq \sigma_2$). The LRT statistic is assumed to have a χ^2 distribution with 1 df

Table 4.16: Comparison of the different likelihood methods for the highly unstable interconnect sample.

estimates is no longer a matter of precision only. Maxima at the top of the adapted likelihood functions are not necessarily in the neighborhood of the largest local maximum of the classical likelihood function. For example:

- At δ = 1e-4, LEδ^* corresponds to a totally different maximum than LE, LEδ and LEδ_s.

- At δ = 1e-3, all adapted likelihood functions have a largest maximum which largely differs from the maximum corresponding to the LE (Table 4.16a).

The observations for this sample do hold for any other highly unstable sample and for some unstable samples. Importantly, inference results of the adapted likelihood methods depend heavily on the value of δ used and the sample remains (highly) unstable. No unambiguous conclusions can be drawn. This demonstrates once more that it is dangerous to rely on inference results for highly unstable samples, regardless of the likelihood method used. Adapting the likelihood function will not improve the situation. Moreover, numerically it can be difficult to distinguish the largest local maximum among all singularities and apparent maxima.

4.5.4 General conclusions

No essential differences are observed between the situation where the MLE exists and where not, when applied to finite normal mixtures. If $\hat{\Delta}$ equals $\delta/\hat{\sigma}^{\mathrm{MLE}}$ in case of a common scale parameter and $\delta/\min(\hat{\sigma}_1^{\mathrm{LE}}, \hat{\sigma}_2^{\mathrm{LE}})$ for the general mixture model, then the following summary can be made of the comparison between the classical (maximum) likelihood method and the adapted "non-symmetrical" (maximum) likelihood methods:

$\hat{\Delta} < 0.1$

For all samples:

-Unimportant to negligible differences between estimates and estimated standard errors.

-If δ decreases, differences diminish.

$\hat{\Delta} > 1$

Stable samples:

-Differences are due to a diminishing precision.

-Comparable largest maximum for all likelihoods.

Highly unstable samples:

-Important differences due to non comparable largest maxima for the different likelihoods, not a matter of precision.

-Largest maximum of the adapted likelihood depends on the value of δ.

-Different conclusions.

Unstable samples:

-Mostly classified by the stable samples, but for some samples differences between methods evolve like for a highly unstable sample.

Similar results hold for the "symmetric" adapted likelihood method. Only, the boundaries on $\hat{\Delta}$ can be taken less stringent. For $\hat{\Delta} < 1$, differences between the two methods are quite smaller than the differences between the classical likelihood method and the non-symmetrical adapted likelihood methods. They are often even acceptable. Note that the boundary values used (i.e., 0.1 and 1), are only approximate values. Furthermore, the results can be easily extended to (log)normal mixtures with more than two components and to SEV or Weibull mixtures.

The parameter δ is a kind of smoothing parameter. The larger it value is, the more small groupings of data points, whether random or not, will disappear in the discretised sample. Through the use of a large measurement error, a large part of random groupings can be removed, but possibly also the "true" one. Importantly, for most samples the stability does not depend on the value of δ used. If the exact value of δ is known, a simple comparison with the likelihood estimates of the scale parameters, can reveal whether an analysis with an adapted likelihood method will add something. If the

value of δ is unknown, a kind of sensitivity analysis can be carried out to verify how consistent the results are with those obtained from the likelihood method. It should be realized, however, that its introduction is useless in the impossible search for a proper likelihood estimate for a highly unstable sample.

Chapter 5

An automatic starting value procedure

In the previous chapter, we have compared for a general finite mixture model some important estimation methods, based on the maximum likelihood method. Until now, in the examples and simulations given, we did not explain how the estimates were calculated. Moreover, it was taken for granted that the estimate, i.e., the largest local maximum of the likelihood function under consideration, was found. In this chapter we discuss how to obtain such estimates and in particular how we were able to detect the largest local maximum with a certainty of almost 100%. In essence, it all amounts to the choice of the starting values. Specifically, for the finite mixture model, estimates can be found as a root of the LEQs corresponding to the largest local maximum of the likelihood function. However, to solve these equations an iterative procedure as well as a set of initial or starting values for the parameters, are required (Section 3.5.1). While for most classical one or two parameter distributions the choice of starting values is not important since their LEQs contain a single root, for the mixture problem apparently good starting values are essential.

The key problem for the general mixture model, primarily due to the usually large number of roots for the LEQs, is that the root obtained may be highly sensitive to the starting values used (Fowlkes, 1979; Redner and Walker, 1984; Böhning, 2000). In other words, a poor choice of starting values will not necessarily give rise to the required largest local maximum, i.e., such starting values can converge to an improper root. An example of where this can lead to, apart from wrong point and interval estimates, is given by Böhning (2000, pp. 66-70). There, the null distribution of the likelihood ratio test statistic in case of the null hypothesis of a simple exponential distribution against the alternative of a two-component exponential mixture, is simulated. It is demonstrated that the simulated distribution quantiles of the likelihood ratio statistic can differ a lot depending on the starting value procedure used. Note that in this example only a Weibull mixture with a common known shape parameter is used, in which case the number of roots are considerably less compared to a general Weibull mixture. As such, one can expect that things will become even worse for general mixtures. But there are more problems involved with poor starting values. Namely, they can lead, in particular for the EM-algorithm, to very slow convergence and to a failure of convergence, especially in combination with the NR-method.

On the other hand, there are multiple reasons to spend some time in searching *good* starting values, i.e., parameter values in the neighborhood of a (largest) local maximum of the likelihood function. Namely, such starting values are the best way to speed up the algorithm (Furman and Lindsay, 1994), in particular the EM-algorithm. They are the best option to arrive at the largest local maximum with an iterative procedure and they considerably reduce the chance of failure to converge, mainly in case of a NR based iterative method. Also, given a set of good starting values an indication of the stability of the sample can be obtained. In addition, if the method calculating these starting values, i.e., the starting value method, allows automation, then simulations and bootstrap procedures are feasible (in terms

of both time and unambiguity) and it becomes possible to fit mixtures in real terms (software, industry).

Nevertheless, only relatively little research efforts have been devoted to these starting values. Instead, a lot of attention is paid to the improvement of the weaknesses of some iterative procedures, in particular the EM-algorithm. Ironically, in case of finite mixtures, a lot can be gained already through the choice of a series of well-reasoned starting values. In addition, as Böhning (2000, p. 67) states: "the starting value problem is mostly treated in an ad hoc manner and the impression is left that the choice of starting values is not crucial". Often, it is assumed that in case of univariate mixtures relatively few problems are met in order to arrive at the proper maximum (Everitt and Hand, 1981, p. 47), despite the indication of the opposite by others (Fowlkes, 1979). It is believed that as long as the sample size is large enough and the mixture components quite well separated, then any reasonable starting value will end in the largest local maximum. However, no one really seems to know what is meant with "large enough", "quite well separated" and "reasonable". Also, what to do in cases where the estimation of a mixture is required with poorly separated components (for example, when the null distribution of a likelihood ratio statistic has to be simulated) or where the sample size is too small? There, it is of importance that also if a distinct spurious maximum is the largest local one, that it can be found. All too often, it is taken for granted that by trying a number of starting values and picking out the largest maximum (Everitt and Hand 1981, pp. 41-42; Hastie et al. 2001, p. 239), the MLE or LE is obtained, without any further justification.

Here, the starting value problem for general finite (log)normal, SEV and Weibull mixtures is handled. In Section 5.1, we discuss first the rather limited amount of literature about starting values. An overview is given of the existing starting value methods, as well as some of their major shortcomings. We specify how, to our opinion, a starting value method should

look like or should perform. A new developed starting value method to obtain the LE for two-component mixtures, fulfilling some of the most important requirements is introduced in Section 5.2. Although good starting values are required, irrespective of the fact of an LE or an adapted LE is searched for, main focus is first on a starting value procedure for the LE. The reason is that in essence the adapted LEs discussed in the previous chapter can be viewed as special cases of the LE. Once a good method is obtained for the LE, it should be possible to extend it to the case of adapted LEs (Section 5.5.1). Furthermore, the starting value method is essentially developed for a two-component mixture. The reason for this is twofold: it is the simplest general finite mixture model, although already complicated, and the most commonly used one in reliability analysis. In Section 5.5, among other extensions, it is discussed how the developed method can be generalized to the case of more than two components. A simulation study, described in Section 5.3, demonstrates the excellent performance of the proposed starting value method and compares it with the performance of some other methods. Further, it is explained how the method can be extended to find almost always the LE, also in case of small sample sizes. Next to this, it is illustrated in Section 5.4 by means of some examples, how some features of the developed starting value method can be used to obtain an idea about the amount of information available in the sample with respect to a finite mixture model.

5.1 Literature review

We begin with an overview of the most important methods, present in literature and (experimental) software packages, to obtain starting values. Some of these methods are specific to the mixture model case, others are primarily used for other problems but could serve as well for the mixture problem.

At random procedures. Starting values are generated by means of a random process. Roughly these methods can be divided into two groups: those derived specific for the finite mixture model and those methods that apply for more general problems. Within the last group, the most "simple" method is to randomly choose a value for each parameter. Finch et al. (1989) developed this technique further, by introducing the concept of probabilistic measures of adequacy. It is then possible to determine the chance that a global maximum was not found yet with a set of random starting points. Hereby, each parameter has a certain distribution attached to it, from which to sample its values. Another example of a general procedure is the bootstrap root search method, proposed by Markatou et al. (1998) and Markatou (2000), in a weighted likelihood context. The idea is to detect as many roots of the LEQs as possible and the global maximum in particular. Briefly, B bootstrap samples of size m are generated from the original sample. For each bootstrapped sample, moment estimates of the parameters are calculated and used as starting values. Based on the technique of Finch, it is suggested that for their particular problem, 100 bootstrap samples are usually sufficient to obtain with a high chance the global maximum. The main difference with the method of Finch is that the starting values, although obtained by a random process, are data driven. Note that this method will be difficult to use due to a problem with the moment estimates in case of general finite mixtures as explained further on.

McLachlan and Peel (1998) (see also McLachlan and Peel, 2000) suggest two random methods for the finite mixture case, implemented in their software program EMMIX. The first procedure consists of randomly dividing the sample into M subgroups, with M the required number of mixture components. The EM-algorithm is then started from the M-step given a complete data likelihood, based on the divi-

sion, and not from the E-step given starting values for the parameters. To remove effects of the central limit theorem on the starting values, it is suggested to first take a subsample of the data which is then randomly divided (instead of the whole sample) and used in the first M-step of the EM-algorithm. A second method randomly generates g mean values from a multivariate normal distribution with the sample mean and sample variance as parameters. Another random method specific to the finite mixture case is given by Finch et al. (1989) for a two-component normal mixture with common scale parameter. A starting value for the proportion parameter is generated from the standard uniform distribution. Starting values for the other parameters are derived from a division of the sample into two subgroups based upon the generated value of the proportion parameter. Note that for these last two random methods maxima, corresponding to mixtures for which the components are primarily separated in scale, will be rarely detected.

Although random methods are easy to use, they usually have one major shortcoming in that they do not deliver well-reasoned starting values. Certainly, this holds true for the methods of McLachlan and Finch. As such, the main advantage of speeding up the iterative procedure and reducing the number of failures to convergence is lost, since starting values can equally well be poor choices. Furthermore, not much is known about the performance of these methods in case of finite mixtures, i.e., the number of starting values required to obtain with a high probability the largest maximum. According to Finch et al. (1989), the relatively simple problem of a two-component normal mixture with common scale parameter is already rather difficult to handle with random starting values, in case of a mixture with poorly separated components.

Multivariate starting value methods. In a model-based clustering ap-

proach, multivariate finite (normal) mixtures are frequently used to model clusters within the data. Conversely, many clustering methods, which are rather based upon heuristics, but intuitively reasonable procedures (Fraley and Raftery, 2000), can be used to obtain starting values for the finite normal mixture model. Examples are k-means clustering methods, hierarchical clustering methods (Johnson and Wichern, 1998, chap. 12) and the fuzzy c-means algorithm (Bezdek and Dunn, 1975).Although these methods seem to work fine in case of multivariate mixtures or univariate mixtures with a common scale parameter, in our experience they perform quite unsatisfactorily as starting value method for the univariate general mixture model. Its major shortcoming is that maxima corresponding to mixtures for which the components are mainly separated in scale, will rarely be found. But there are more disadvantages. Not much is known about their performance: it has not been checked whether these starting values give rise to the largest local maximum. For the fuzzy c-means algorithm, Hathaway and Bezdek (1986) show in the univariate case that estimates obtained with this algorithm are not necessarily consistent, a property which is desirable for starting values. These clustering methods also depend on some extra (algorithmic) parameters. Sometimes even starting values are required. Also it is unknown how they will perform for mixtures with other component densities than the normal distribution. Another possibility to obtain starting values for multivariate mixtures is the use of principle component analysis (McLachlan, 1988). This, however, is useless in the univariate case.

Consistent estimation methods. A consistent estimator, used as a starting value, converges to the consistent root of the LEQs, i.e., to the MLE or LE if the sample size is large enough (Section 4.2.2). As such, any consistent estimator will be a good starting value. Well-known examples of consistent estimators are moment estimators and estima-

tors obtained from the moment generating function. Also the true parameter values, though useless in real applications, have this desirable property. There is, however, one major problem blocking these methods from being used as a starting value method. Namely, for the general finite mixture model it is rather problematic to find or to calculate such consistent estimators. Indeed, Bowman and Shenton (1973) establish for general (log)normal finite mixtures that moment estimators are not unique and importantly do not always exist. Besides, apparently there is no simple way to calculate them: for a general two-component normal mixture, Pearson's famous 9-nonic has to be solved (Pearson, 1894). Note that this problem does not appear in case of a finite (log)normal mixture with a common scale parameter. There, Furman and Lindsay (1994) proved the existence of a unique moment estimator which could relatively easy be calculated, irrespective of the number of mixture components. Further, not much is known about moment estimators for SEV or Weibull mixtures. When moment estimation is applied, the shape parameter is mostly taken to be common (Rider, 1962). In general, also an iterative procedure is required for their own calculation. Similar remarks hold for estimators based on the moment generating function (Quandt, 1978; Quandt and Ramsey, 1978). These estimators are a generalization of moment estimators.

Techniques for a normal mixture with a common scale parameter. The LEQs for mixtures with a common scale parameter contain considerably less roots than in case of a general mixture model. Still, starting values are of importance, especially for small sample sizes or when too many mixture components are considered (for example, Finch et al. 1989 and Böhning 2000, pp. 67-70). In literature, some suggestions and methods are present to derive starting values for this specific model. There are the unique, easy to calculate, moment estimators (Furman and Lindsay, 1994). Further, Finch et al. (1989) uses starting values

based upon an estimate of Engelman and Hartigan (1969) for the proportion parameter. This estimate is obtained by maximizing the ratio of the between sum of squares to within sum of squares among all possible splits of the sample. According to Finch this estimate would only miss the global maximum in 3% of the samples, when the latter are generated from a single normal distribution. Lindsay (1995, pp. 65-66) indicates that the NPMLE also can be used to derive starting values. Indeed, given a value of the common scale parameter, a unique, discreet NPMLE exists and can be calculated (Böhning, 1995). This estimate suggests then the maximum number of mixture components. But, as mentioned too by Lindsay, this method would be hard to use in simulations and practice. Although most of these methods work fine for this specific mixture model, they can be hardly adopted to the situation of a general mixture. The case of the moment estimator is discussed previously. The Engelman-Hartigan estimate of the proportion parameter could also be derived for a SEV mixture with common scale parameter, but when generalized, mixtures with components separated mainly in scale, will be overlooked. Finally, the theory of the NPMLE does not hold for general finite mixtures with a (log)location-scale distribution as component density, mainly due to the non-identifiability problem of general nonparametric mixtures (Section 3.3).

Graphical procedures. As the name suggests, these methods utilize graphics. Often the latter will be a probability or QQ-plot (Section 5.2.1). A large part of those methods make use of the naked eye and were a popular estimation tool before the introduction of the computer, mainly due to the unappealing moment estimates (Fowlkes, 1979). Graphical estimates were quite rough, but sufficient at that time. According to Fowlkes, one of the best graphical methods originates from Harding (1948), relating the shape of the configuration of a normal QQ-plot to the parameters of a general two-component normal mixture. Note

that this kind of graphical procedures are still sometimes used (in a reliability context) to estimate general finite mixtures (Møltoft, 1983; Jiang and Murthy, 1995). Jiang and Murthy (1995) consider the estimation of two-component Weibull mixtures. It ought to be mentioned that their technique, though laborious, is one of the few that takes into account also the estimation of mixtures with components separated in shape. Bhattacharya (1967) developed a method for grouped normal data not based on a probability plot, quite useful in an exploratory analysis. While it is often possible to derive useful starting values for mixtures with clearly separated components, these methods are not an option as a starting value method. Indeed, they are subjective and allows hardly ever automation. On the other hand, there are more objective graphical methods using a computerized approach. Examples are least squares estimation based upon a QQ-plot and the starting value method of Fowlkes (1979). The former is not considered as an option since it also involves a set of equations that has to be solved. The latter formalized in essence the approach of Harding (1948). This method works fine in many cases and leads to well-reasoned starting values, but is difficult to computerize, since starting values are again needed. In addition, the method does not take into account the possibility of a mixture where the components are separated in scale.

Other procedures. There exist some methods to find the global maximum of a function without the need to specify a starting value, i.e., the so-called global optimization procedures (Cetin et al., 1993; Battiti and Tecchiolli, 1996; Chelouah, 2000). For the general finite mixture case, they are not really an option due to the singularities of the likelihood. In addition, the few which exist, are very slow and do require in turn bounds of the parameter space. Further, there have been some attempts to turn the EM-algorithm into a global convergence algorithm (Celeux et al., 1996, 2001). In other words, to make the algorithm less

sensitive to the starting values used. We prefer the search for good starting values, as not only the largest local maximum is of interest, but also the stability of the sample. In spite of this, the performance of these adapted EM-algorithms is never properly investigated. Most simulation studies were not only very small, but also compared the result of the adapted algorithm with the outcome of for example, the EM-algorithm. In none of the cases, a comparison was made with the true largest local maximum.

From this survey, it is not only clear that most starting value procedures have some important shortcomings when applied to the general finite mixture model, but also that we expect a lot from a starting value method. If we had the choice, the "perfect" starting value method should at least have the following two properties:

- For any sample, one of the produced starting values has to converge (with an iterative procedure) to the root corresponding to the required maximum.
- The method should include the ability to automate or should be *computerizable*.

The first requirement implies having a good starting value method. Nevertheless, mainly due to the nature of a general finite mixture model, it is not possible to obtain this for small sample sizes with any starting value procedure, unless the whole parameter space would be scanned. Instead, we replace this requirement by the property of *consistency*. Hereby, we term a starting value method *consistent* when one of the produced starting values converge to the required maximum, at least for sufficiently (reasonable) large samples. The second property implies that the algorithm should allow automation, should be easy to carry out, avoiding complex or lengthy calculations, and ought to be fast, since slow algorithms cannot be used in

Table 5.1: Properties of possible starting value methods for the finite general (two-component) mixture model.

	Mcl1/Mcl2	Clus	MM	E-H	Fow
Original model	general $M > 2$	general $M > 2$	common $M > 2$	common $M = 2$	general $M = 2$
Complete	Yes/No	No	No	No	No
Extension possible to					
SEV/Weibull	Yes	?	Yes	Yes	Yes
Censored data	?	?	No	?	Yes
$M > 2$	Yes	Yes	Yes	Yes	?
Consistency	?	No	Yes	No	No
Computerizable	$\pm$	Yes	No	Yes	No
More than 1 starting value	Yes	Yes	No	No	No

Explanation of the abbreviations: Mcl1: Randomly subdividing the sample; McL2: Generating means from a multivariate normal distribution; MM: Method of moments; Clus: Clustering procedures; E-H: The estimates of Engelman-Hartigan; Fow: The method of Fowlkes

simulations and in practice. In addition, it includes the use of good starting values, improving the performance of the iterative procedure.

Apart from these two requirements, which actually any starting value method for any model should meet, we demand something more specific to the problem at hand. Namely, it should be possible to extend the method for censored data due to the nature of reliability data, for grouped (binned) data in order to obtain adapted likelihood estimates and preferably also for more than two mixture components. Further, in Section 4.3.3 we stressed the fact that it was important to look at the stability of the sample or to obtain an idea about the surface of the likelihood. Therefore, the method should not lead to one well-reasoned starting value, but to a (small) number of starting values, which allow to have an idea about the stability of the sample.

To conclude, Table 5.1 summarizes the main properties of those starting values methods which, at least in theory, can be extended to a

method applicable for general two-component normal mixtures. Some comments on the table are in place:

- The original model consisted in all cases of a normal component density.

- The property *complete* refers to whether all kinds of mixtures can be reached, i.e., if also starting values can be obtained that converge to a maximum corresponding to a mixture with components separated mainly in scale.

- The extension to SEV/Weibull component densities and censored samples are required to have a useful method for reliability purposes. Also, it is preferable that the method can be used for a mixture model with more than two components.

- There is only one method, namely the method of moments, which is consistent by theory, if the estimates exist. The main problem for all other methods is that not much is known about their consistency. There is even no empirical evidence. An additional problem, mostly neglected, is that given a starting value, it is difficult to determine whether the root corresponds to the required largest local maximum. Far too often it is taken for granted that the largest maximum is obtained if none of the other algorithms in the experiment lead to a larger maximum (Furman and Lindsay, 1994; Ueda et al., 2000) or when the parameter values of the root look reasonable. Although, the results obtained could be good, none of these techniques are tested to see whether they really work, i.e., whether the required maximum is indeed obtained for sufficiently large sample sizes. Sometimes the obtained root is compared (with a certain distance measure) to the true values or to the root obtained with the true values as justification that the MLE is obtained (Furman and Lindsay, 1994; Celeux et al., 2001).

Although this root is also an example of a consistent estimate, it is not necessarily the MLE or LE.

- The two random methods by Mclachlan are not quite considered as being computerizable since they do not lead to well-reasoned starting values.

- There are some strong indications that most clustering methods, the estimates of E-H and the method of Fowlkes are consistent in case of a normal mixture with common scale parameter. But since they are not complete for the general mixture model, they are not consistent for the general mixture model.

To our opinion, none of these methods, except perhaps the complete random method, are acceptable as a starting value method for a general finite mixture model since each of them lack at least one essential property. The idea was to develop a new alternative technique which should a least be complete, computerizable with a number of well-reasoned starting values and consistent. In addition, it should allow an easy extension to censored data. The resulting method is introduced in the next section.

5.2 The tangent-rico method

In spite of the fact that most existing graphical procedures are found to be unsuitable as a starting value method, they have the interesting feature of delivering data driven and often well-reasoned starting values. For that reason, we have opted for a graphical approach in developing a starting value method. As basic graphical tool the QQ-plot is considered. Compared to other plots that visualize distributions, this plot allows much more easily the recognition of finite mixtures, in particular two-component mixtures (Section 3.4). Note that the choice of this kind of plot entails that the starting value method will be difficult to extend to mixtures with a component

distribution not belonging to a (log)location-scale family. Nevertheless, this is a not an issue for the mixtures considered here as well as in many other reliability situations.

The algorithm is based on the relation between the shape of the QQ-plot of a sample and a split of this sample into two subgroups, corresponding to the two components of the mixture. It makes use of the nature of a mixture as a model for populations with multiple groups present. In Section 5.2.1, we discuss the different shapes of a QQ-plot for a two-component mixture. The algorithm is explained in Section 5.2.2 and the automated method is introduced in Section 5.2.3.

5.2.1 Quantile-quantile plot of a general two-component mixture

Given a complete ordered sample $(x_{(1)}, x_{(2)}, \ldots, x_{(n)})$ and a distribution with cumulative distribution function (cdf) F, then a *probability* or *quantile-quantile(QQ)-plot* of this sample for the distribution F, is defined as a plot of the sample quantiles $x_{(i)}$ versus the theoretical quantiles $F^{-1}(p_i)$. For a (log)location-scale distribution, a QQ-plot of a sample is generally constructed through plotting $x_{(i)}$ $(\ln(x_{(i)}))$ versus the theoretical quantiles $F_0^{-1}(p_i)$, with $F_0(x) = F(x|\mu = 0, \sigma = 1)$ the cdf of the standard location-scale distribution (D'agostino and Stephens, 1986, pp. 25, 464). If the underlying distribution of the sample is also a (log)location-scale distribution, then the shape of its QQ-plot should resemble a straight line. The probabilities p_i are referred to as quantile probabilities or more commonly as *plotting positions*. There is no unique way to determine these points and in the literature there is a lot of discussion about the best choice for them. In general, the "best" choice depends on the application of the QQ-plot and even then there is no general agreement. More information can be found in Meeker and Escobar (1998, chap. 6).

One possibility for the plotting positions is given by $\frac{i-c}{n-2c+1}$, with

$0 \leq c \leq 1$, $1 \leq i \leq n$ and n the sample size (D'agostino and Stephens, 1986, p. 25). For $c = 0.5$ or $c = 0$, this leads to the popular choices $\frac{i-0.5}{n}$ or $\frac{i}{n+1}$. However, we have opted for the positions $1 - e^{-\frac{1}{2}(S_i+S_{i-1})}$, with $S_i = \sum_{j=1}^{i} \frac{1}{n-j+1}$ the empirical cumulative hazard function, introduced by Nelson (1982). They are based upon the plotting positions $1 - e^{-S_i}$, which are a natural extension of the hazard plotting positions S_i. Only, the mid-points $\frac{1}{2}(S_i + S_{i-1})$ are considered, as it is argued that these points would agree better with a distribution fitted by maximum likelihood (Nelson, 1982). Nevertheless, as will be discussed further on, for the intended application the choice of the plotting positions appears to be rather unimportant.

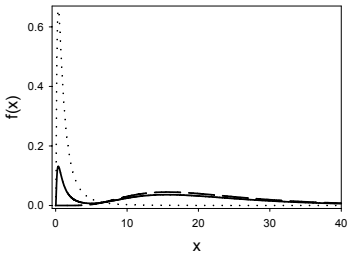
A *theoretical QQ-plot* of a distribution with cdf G versus a distribution with cdf F is defined as a plot of the theoretical quantiles $G^{-1}(p)$ versus the theoretical quantiles $F^{-1}(p)$. In addition, x can be plotted against $F^{-1}(G(x))$. If a sample is distributed according to G, then the shape of the QQ-plot for F of this sample, should resemble the shape of the theoretical QQ-plot of G against F. When F is the cdf of a standard location-scale distribution (i.e., F_0) and G the cdf of a mixture with a component distribution belonging to the same parametric family as F_0, then the theoretical QQ-plot of G versus F_0 is nothing else than a plot of the cdf of the mixture distribution on appropriate probability scales (i.e., scales according to the distribution of F_0). For a mixture with a log-location-scale distribution as component, this kind of QQ-plot is constructed through plotting $\ln(x)$ against $F_0^{-1}(G(\ln(x)))$. In the following, unless stated explicitly, the term theoretical QQ-plot of a mixture will refer to these specific plots. Also when there is no ambiguity possible whether either an (empirical) QQ-plot or a theoretical QQ-plot is meant, the term theoretical will be left out.

As well-known, the shape of the QQ-plot of a (log)location-scale distribution against the standard (log)location-scale distribution is a straight line. In contrast, the shape of the QQ-plot of a two-component mixture will be a curve. It can now be proven that in case the component has a

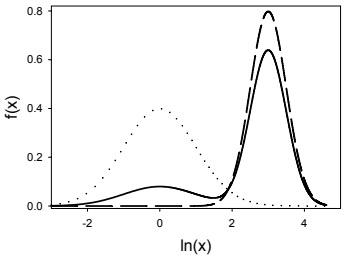
SEV (Weibull) distribution, this theoretical QQ-plot will have 3 *inflection points*, i.e., points where the second derivative of the curve is zero. This number reduces to 1 if the scale parameter is common (Jiang and Murthy, 1995). The same can be shown to hold true for a (log)normal component distribution. Nevertheless, depending on both the separation in location (scale) and in scale (shape) of the mixture components and, to a lesser degree, the size of the proportion parameter, not all inflection points will be visible on a theoretical QQ-plot. Moreover, two different configurations can be distinguished: a steep-flat-steep form (Figure 5.1c) and a flat-steep-flat form (Figure 5.2c). The recognition of one of these two shapes on an (empirical) QQ-plot is the basis of the starting value method.

Steep-flat-steep Form

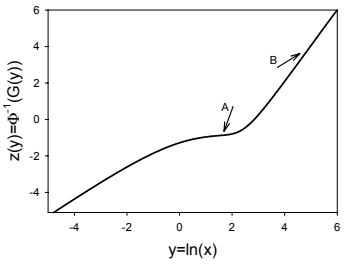
When the two component densities of a two-component mixture distribution are placed one after the other on a density plot, as shown in Figures 5.1a–5.1b, the shape of the QQ-plot of this mixture (versus its component distribution) will be steep-flat-steep (Figure 5.1c). In this case, the difference between the location (scale) parameters of the mixture components is more important than the difference between their scale (shape) parameters. Only the right tail of the component density with the smallest location (scale) parameter overlaps with the left tail of the other component density in the density plot. The further the components of the mixture are separated (in location), the smaller this overlap becomes and the more pronounced the steep-flat-steep form will be. The QQ-plot is characterized through the presence of one clear inflection point, referred to as point A in Figure 5.1c. It corresponds to a minimum for the first derivative of the QQ-plot as shown in Figure 5.1d and 5.1e. Still another inflection point (B) corresponding to a maximum for the first derivative of the QQ-plot is present. Only, it is hardly visible and as such not important. Moreover, due to a limited numerical accuracy, the QQ-plot of is lacking a third inflection



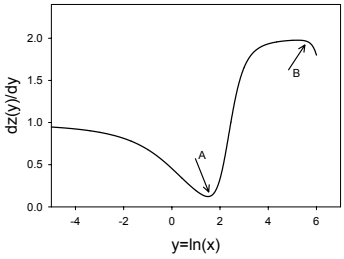
(a) Lognormal densities of 1^{st} (dotted) and 2^{nd} component (dashed), and mixture (solid line).



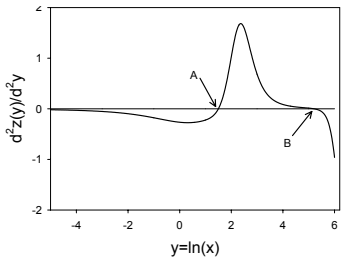
(b) Densities for the equivalent normal mixture.



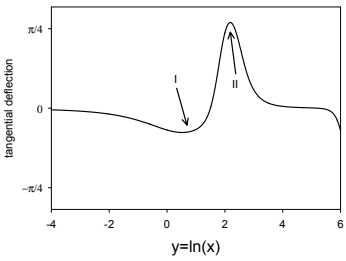
(c) Lognormal QQ-plot of the mixture with inflection points A and B.



(d) Derivative of QQ-plot.



(e) Second derivative of QQ-plot.



(f) Tangential deflection of QQ-plot.

Figure 5.1: A two-component lognormal mixture with a steep-flat-steep form. Parameter values are $\pi_1 = 0.2$, $\mu_1 = ln(\eta_1) = 0$, $\sigma_1 = 1/\beta_1 = 1$, $\mu_2 = 3$, $\sigma_2 = 0.5$.

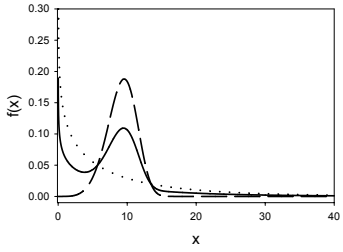
point. This point (C) would be situated after point B for the example in Figure 5.1c, since $\sigma_1 > \sigma_2$. In case $\sigma_1 < \sigma_2$, the order of the inflection points would be C, B and A. In case of a common scale (shape) parameter, point A is the only inflection point left.

For (log)normal mixtures with $\pi_1 = 0.5$ and $\sigma_1 = \sigma_2$, it can be proven that π_1 equals $\Phi(y_I)$, with y_I the y coordinate of inflection point A (Fowlkes, 1979). For other proportions and unequal scale (shape) parameters, $\Phi(y_I)$ is still a good approximation for π_1 as long as the overlap between the component densities is not too large. The approximation improves when the overlap becomes smaller. Importantly, inflection point A indicates roughly where to situate the different components of the mixture: the first component before this point and the second one after it.

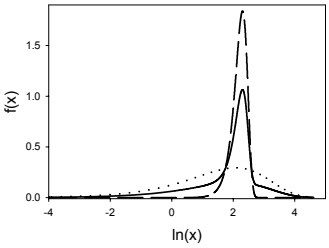
Flat-steep-flat Form

If the overlap between the two component densities becomes too large, one of these two densities will become fully enclosed by the other density (Figures 5.2a – 5.2b). The QQ-plot then has a flat-steep-flat form (Figure 5.2c). Contrary to the previous form, the shape of the QQ-plot is now dominated by the difference between the scale (shape) parameters of the mixture components. For Weibull or lognormal mixtures, the component density with the largest shape parameter (β) is surrounded by the one with the smallest shape parameter. Consequently, the tail ends in the QQ-plot are dominated by the component density with the smallest shape parameter. For normal or SEV mixtures, the component density with the largest scale (σ) parameter will dominate the tail ends of the QQ-plot.

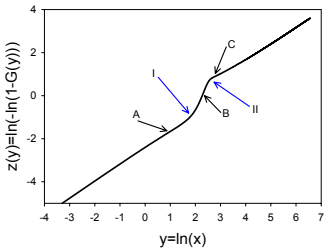
As shown in Figure 5.2e, which gives the second derivative of the QQ-plot, three inflection points, referred to as A,B and C, are now present on the QQ-plot (Figure 5.2c). Points A and C correspond to a minimum for the derivative of the QQ-plot, while B corresponds to a maximum (Figure5.2d). Further points A and C roughly mark out where each component of the mix-



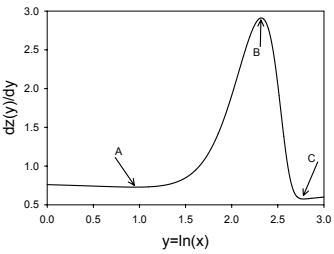
(a) Weibull densities of 1^{st} (dotted) and 2^{nd} component (dashed), and mixture (solid line).



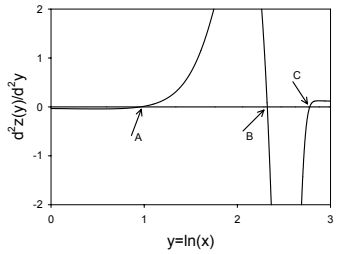
(b) Corresponding densities for the SEV mixture.



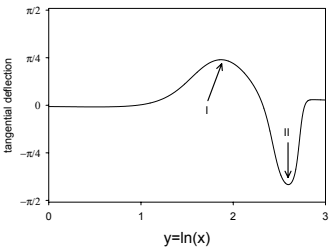
(c) Weibull QQ-plot of mixture with inflection points (A, B, C) and points of maximal tangential deflection (I, II).



(d) Derivative of the QQ-plot.



(e) Second derivative of the QQ-plot.



(f) Tangential deflection of the QQ-plot.

Figure 5.2: A two-component Weibull mixture with a flat-steep-flat form. Parameter values are $\pi_1 = 0.5,\ \eta_1 = 8,\ \beta_1 = 0.8,\ \eta_2 = 10,\ \beta_2 = 5$.

ture is dominating: before A and after C, the component with the largest scale parameter (or smallest shape parameter) has the upper hand over the other; while the reverse is true between these two inflection points. Unfortunately, as can be noticed from Figures 5.2c and 5.2d, none of these two inflection points can be clearly distinguished from the QQ-plot. As such, they will rarely be detected on a QQ-plot of a sample. Although the opposite is true for the inflection point B, it cannot be used to locate the two components of the mixture. Note that with this form, none of the quantities $\Phi(y)$, with y the y-coordinate of one of the inflection points, gives a good approximation of the proportion parameter.

A more distinct feature of a QQ-plot with a flat-steep-flat shape, is the fact that near points A and C, there are points I and II, where the *tangential deflection*, i.e., the angle between two adjoining lines, is larger then anywhere else on the QQ-plot (Figures 5.2c – 5.2f) or where the cosine of the tangential deflection attains a minimum. These points are termed *nodes*. The nodes can be distinguished through the sign of the sine of the tangential deflection: for node I the sine is positive while the sine is negative for node II. Such nodes also exist in case the QQ-plot has a steep-flat-steep shape. One node is then situated before and the other after the inflection point A (Figure 5.1f). Only now the first node is characterized through a negative sine and the second through a positive sine.

Summary

In general, the QQ-plot of a two-component mixture, will be characterized either through the presence of one clear inflection point or through two nodes. Even more than on the QQ-plot these specific points can be identified on two plots derived of the QQ-plot: the derivative plot and the cosine of the tangential deflection plot.

- In case the QQ-plot has a steep-flat-steep shape, its derivative plot will have one clear minimum. The corresponding plot of the cosine of the

tangential deflection will have two minima. The first minimum, having a negative sine, is situated before the inflection point and the other, having a positive sine, is situated after the inflection point.

- In case the QQ-plot has a flat-steep-flat shape, its plot of the cosine of the tangential deflection will have two minima, with the first having a positive sine and the second a negative sine. On the derivative plot, between these two nodes a clear maximum is situated.

The empirical counterparts of these two plots derived of the QQ-plot, will not only be used in the starting value method. Apparently, they are useful as an exploration tool to obtain an idea about the number of possible components and to situate these components (Section 5.4).

5.2.2 Algorithm

The main problem in obtaining starting values for the mixture model, is that we do not know to which component of the mixture each unit of the sample belongs. Or, in terms of reliability, the specific failure reason of each device is unknown. If the membership of each unit would be known, estimates for the parameters of the components densities, i.e., (μ_1, σ_1) and (μ_2, σ_2), could easily be obtained. Indeed, instead of fitting the whole sample to the mixture, each subsample can then be fitted separately to the component distribution. Likewise, an estimate for the proportion parameter π_1 can then simply be determined as the ratio of the size of one subsample to the total sample size.

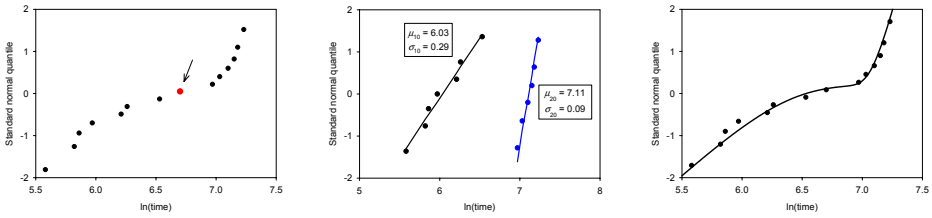
This fact will be used in the algorithm to arrive at starting values for (μ_1, σ_1) and (μ_2, σ_2): given a subdivision of the sample, starting values for the parameters of the component densities will be the MLEs obtained from fitting each subsample separately to the component distribution. The only question left then is how to subdivide a sample. Here, this division is related to the specific shape of the QQ-plot of the sample, i.e., depending

on whether the shape is steep-flat-steep or flat-steep-flat, a split is proposed based on the place of possible inflection points or nodes.

Tangent-rico method for the steep-flat-steep form

A QQ-plot with a steep-flat-steep shape has one clearly visible inflection point A. This point can be seen as a kind of turning point for the domination of the mixture by one of its components. Indeed, before this point A the first component of the mixture, i.e., the component with the smallest location parameter, roughly dominates the mixture, while after A the other component dominates. Hereby, "before" and "after" refer to the x coordinate of point A. In practice, it can be stated that data points situated before this inflection point are more *likely* to belong to the first component, while the reverse is true for the other data points. Based on this, a division of the sample is possible. The most clear-cut one, places all points before the inflection point A in one subsample and all points after in the other subsample. A more refined procedure excludes from this division all points in the *neighborhood* of A, since for these points the *uncertainty* to which group they belong, is largest. From this division, a starting value for the proportion parameter π_1 can be obtained as the ratio of the number of points in the first subsample to the total number of data points in the two subsamples.

Of course, any inflection point is unknown for the QQ-plot (for a certain (log)location-scale family) of a sample. As such, given an empirical QQ-plot, somehow an estimate for the inflection point A has to be searched for. This is done using the property that the first derivative of a theoretical QQ-plot of a mixture attains its minimum in point A. Note that although for a continuous derivable curve, a *tangent line* or derivative is unambiguously defined in each point, this notion is far less clear for a discrete curve, as for example an empirical QQ-plot. The definition as well as the calculation of the derivative in a point of a discrete curve is taken from Anderson and Bezdek (1984). Briefly, the derivative in a point (x, y) is defined as the



(a) Lognormal QQ-plot of the whole sample.

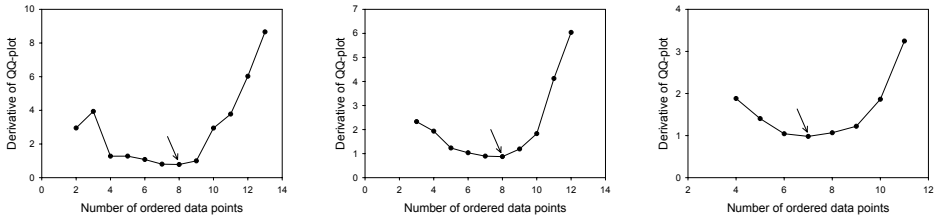
(b) Lognormal QQ-plot of two subsamples and their ML fit.

(c) Likelihood fit of a 2-component lognormal mixture.

Figure 5.3: Deriving starting values for the sample EM1. Division of the sample is based on the data point indicated in Figure 5.3a.

derivate of the *best* line through a number of neighboring points or *neighbors* around (x, y), with "best" in the sense that the sum of the squares of the orthogonal distances between these points to each other line is minimized. So, to calculate the derivative in a point, only the number of neighbors, m, has to be chosen. The larger m is, the "smoother" the derivative of the discrete curve will be.

Candidates for the inflection point A are then those points of the QQ-plot where the (numerical) derivative of the QQ-plot attains a local minimum. Here, a minimum of a row of figures is defined as a number in the row which is smaller than $k-1$ $(k+1)$ earlier and $k+1$ $(k-1)$ following neighbors, with each neighbor having a larger value than the preceding one. Throughout this work, k is taken to be 2. Whereas in theory there is only one inflection point A, in practice the derivative of a discrete QQ-plot usually has multiple local minima. In a lot of cases, the global minimum is the best choice. In the final algorithm, for a given value of m, the two minima with the smallest value for the derivative are taken as candidates for the inflection point. Each minimum will lead to another division of the sample



(a) $m = 3$ (b) $m = 5$ (c) $m = 7$

Figure 5.4: Plots of the derivative of the QQ-plot of Figure 5.3a for several m-values.

and consequently to other starting values.

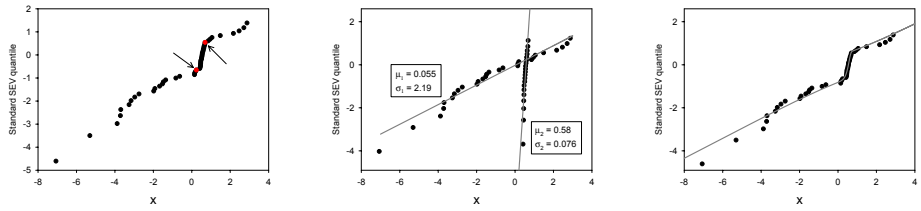
As an example, consider the lognormal QQ-plot of the failure time sample EM1 in Figure 5.3a (Section 2.1.4). The QQ-plot has a clear steep-flat-steep form. For several values of m the numerical derivative of this QQ-plot is calculated and shown in Figures 5.4a, 5.4b and 5.4c. With $k = 2$, the 8^{th} failure time for $m = 3$ or 5 and the 7^{th} failure time for $m = 7$, is selected as the only candidate (given m) for the inflection point A. The subsamples, obtained from a clear-cut division (i.e., only the inflection point is left out for the division) based on the 8^{th} failure time as candidate for the inflection point, are shown in Figure 5.3b on a lognormal QQ-plot. Also, the ML fit of each subsample together with the MLEs, i.e., the starting values for the parameters of the components of the mixture, are shown. The starting value for π_1 is $7/13 = 0.54$. With these starting values, the EM-algorithm converged to the largest local maximum of the likelihood (for a two-component lognormal mixture). Likelihood estimates of the parameters are $\hat{\pi}_1 = 0.58; \hat{\mu}_1 = 6.13; \hat{\sigma}_1 = 0.37; \hat{\mu}_2 = 7.11; \hat{\sigma}_2 = 0.09$. Figure 5.3c depicts the fit.

Tangent-rico method for the flat-steep-flat form

A QQ-plot with a flat-steep-flat form has in theory not one but two inflection points (A and C in Figure 5.2c), marking out the regions where one of the two components of the mixture is dominating. Practically, data points between these two inflection points are more *likely* to belong to the same component, likewise for data points before the first and after the second inflection point. As such, a clear-cut division or a more refined one through exclusion of neighbors of the inflection points, can be obtained. Still, as discussed previously, these inflection points are not useful since they are difficult to detect on an empirical QQ-plot. However, near the inflection points, two clearly visible nodes are situated which can equally well be used to mark out the regions on the QQ-plot where one of the components of the mixture is dominating.

Also here, given the QQ-plot of a sample, these nodes are unknown. Candidates will be searched for based on the property that theoretically both nodes have a minimum value for the cosine of the tangential deflection of the QQ-plot. For a discrete curve, we calculate the tangential deflection of a point (x, y) as the angle between the "best" line through the m consecutive points with (x, y) as endpoint and the best line through the m consecutive points with the neighbor of (x, y) as starting point (Anderson and Bezdek, 1984). Data points for which the cosine of the tangential deflection attains a local minimum, are then candidate nodes. Moreover, if the sine of the tangential deflection is positive (negative), such a data point will be a candidate for the first (second) node. Generally, multiple local minima will be found and multiple combinations will be possible. In the final algorithm for each value of m, the best two combinations of candidates for the nodes will be chosen.

Figure 5.5a depicts the SEV QQ-plot of a simulated sample of a two-component SEV mixture with sample size 50 and true parameter values $\pi_1 = 0.6, \mu_1 = 0, \sigma_1 = 2, \mu_2 = 0.5, \sigma_2 = 0.1$. As noted, it has a flat-



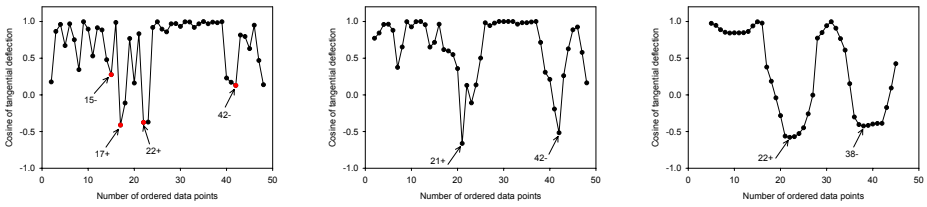
(a) SEV QQ-plot of the whole sample.

(b) SEV QQ-plot of two subsamples and their ML fit.

(c) Likelihood fit of a two-component SEV mixture.

Figure 5.5: Deriving starting values for a simulated sample. Division of the sample is based on the two data points indicated on Figure 5.5a.

steep-flat form. The tangential deflection of the QQ-plot is calculated for several m-values. Plots of the cosine of this tangential deflection are shown in Figure 5.6. Note the smoothing property of m: the more it increases in value, the smoother the curve becomes. Data points where an important local minimum is attained, are indicated together with the sign of the sine of the tangential deflection in these points. For $m = 3$, the local minimum situated at the 15^{th} data point, cannot be a candidate for the second node (negative sine) since no candidates for the first node are situated before this data point. For $m = 5$ (and $k = 2$), the algorithm selected the 21^{th} and 42^{th} data points, as best combination of possible candidates for the first and second node. The SEV QQ-plot of the two subsamples, obtained from a clear-cut division based on these selected data points as candidate nodes, is shown in Figure 5.5b. The subsamples are fitted separately with the ML method. Resulting fits and MLEs, which are the starting values for the parameters of the components of the mixture, are also given in Figure 5.5b. A starting value for π_1 is $28/48 = 0.58$. The result of the EM-algorithm with the use of the these starting values is the largest local maximum. The likelihood



(a) $m = 3$ (b) $m = 5$ (c) $m = 11$

Figure 5.6: Plots of the cosine of the tangential deflection of the QQ-plot of Figure 5.5a for several m-values.

estimates are $\hat{\pi}_1 = 0.58$, $\hat{\mu}_1 = 0.11$, $\hat{\sigma}_1 = 2.14$, $\hat{\mu}_2 = 0.096$, $\hat{\sigma}_2 = 0.57$. The resulting fit is shown in Figure5.5c.

Choice of the plotting positions

Since the tangent-rico method makes use of the QQ-plot of a sample, there is the question whether the choice of the plotting positions will have an important influence on the performance of the method. Moreover, will different choices for p_i change a lot the obtained starting values. The answer is no. The reason behind it, is that the main peculiarities of most samples remain, irrespective of the "reasonable" choice of the plotting positions. Or as Nelson (1972) states it: "differences between the various plotting positions appear to be negligible in comparison with the inherent variability in the sample". To justify this, we carried out some small-scale simulations to check the importance of the choice of p_i. In particular, for sample sizes varying from 20 to 1000, 100 samples were generated from either a normal or a SEV mixture. For each simulated sample, the "best" candidate for the inflection point or couple of nodes obtained with different choices for p_i and different values of m, are compared. Mostly, worst-case scenarios were considered, i.e., generating samples from a distribution and analyzing it ac-

cording to a QQ-plot based on another distribution. For example, simulating a sample from a normal distribution and searching the normal QQ-plot for inflection points or simulating from a SEV mixture and searching the normal QQ-plot for couples of nodes. In Section B.1 of the appendix, the results for some cases are tabulated. We briefly summarize the main findings.

- For standard situations, regardless of the sample size used and the value of m, in almost 90% of the simulated samples, the best candidates found were identical for all choices of plotting positions. For the remaining cases, mostly the candidates were equal up to one data point (in the ordered sample).

- For the worst-case situations, regardless of the sample size used and the value of m, mostly there was a match between the candidates in at least 80% of the cases. For at least 90% of the samples, the candidates were equal up to one data point. Often for larger values of m, less matches are found. The only exception is for the plotting position $1 - e^{-S_i}$. With this choice, the percentages are generally worse than indicated here, i.e., a decrease with about 10%.

- For the samples with no match (up to one data point) between two choices of plotting positions, often the best candidate found with one choice of plotting positions is the second best candidate found with another choice of plotting positions. Rarely, it was found that the candidate data points for one choice of plotting positions were totally different compared to another choice.

5.2.3 Automatic Procedure

Up to here, we explained how to calculate starting values, given the shape of a QQ-plot of a mixture. To automatize the proposed algorithm, the method has to be extended such that this form is no longer required as an input. Therefore, given a sample, for each of the two forms, several

starting values will be computed. Moreover, for each form and for each choice of m, two sets of starting values will be determined. As mentioned previously, m is a kind of smoothing parameter. In the procedure, 3 values for m are implemented, namely 5, 10 and 20 percent of the sample size of the sample considered. In doing so, it is possible to obtain starting values where the proportion parameter has quite a small value regardless of the sample size. At the same time, with the other choices for m, starting values are obtained neglecting the very small subgroups, whether purely random or not, of the sample. So, irrespective of the sample size, at most 12 sets of starting values will be obtained for each sample. In the following section, we will demonstrate that for the proposed choices of m this algorithm works to satisfaction.

Mostly the tangent-rico method will be used in combination with the EM-algorithm (Section 3.5.1). There are several reasons why we prefer this algorithm to NR-based procedures. The most important one is the stability of the EM-algorithm, converging generally to the maximum closest to the starting values used. Since the tangent-rico method produces a set of starting values, each in the neighborhood of a local maximum (not necessarily the largest one), it is important to converge to this maximum and not to jump to another one. Not only this increases the chances to obtain the largest local maximum, but also it allows to obtain a good idea about the stability of the sample (Section 5.4). Another reason is that the EM-algorithm contributes to the feasibility of likelihood estimation in simulations. It is relatively easy to overcome problems with convergence to singularities and points on the edge of the parameter space. We are aware of the main disadvantage of the EM-algorithm, being a slow algorithm. But, unless the mixtures had very poorly separated components, this was not a big issue in the following simulations. Although it ought to be mentioned that problems are more severe for SEV or Weibull mixtures than for (log)normal mixtures, since for the former each cycle in the EM-algorithm is double iterative. Apart

from this, when we used the tangent-rico method in combination with the NR-method, often the resulting maxima where similar to the one obtained with the EM-algorithm. This points to the fact that the starting values used, are already close to a maximum of the likelihood function. In addition, no real convergence problems were encountered.

5.3 A simulation study

The main purpose of developing an alternative starting value method, was to find a method which would at least be consistent, computerizable and complete. Through its design, the tangent-rico method is complete. Since the method gives rise to a set of well-reasoned starting values and allows the ability to be automated, it can be expected that the method is also computerizable. On the other hand, the question whether the method is consistent, ensuring the outcome of the likelihood estimate for sufficiently large sample sizes, cannot be answered directly. Since it is not possible to show this property theoretically, a simulation study is carried out to demonstrate the consistency of the tangent-rico method. Furthermore, this simulation study will confirm both its ability to be automated and, through the choice of the mixtures used, the completeness of the tangent-rico method.

To empirically evaluate the consistency, the largest local maximum of the likelihood function has to be compared to the maxima obtained with the generated starting values for each simulated sample. This requires, however, that the largest local maximum of the likelihood function would be known for each sample. In Section 5.3.1, it is explained how we dealt with this problem. Simulations demonstrating the good performance of our method and comparing it to the performance of other methods, are presented in Section 5.3.2. All simulations are carried out in GAUSS. The EM-algorithm was stopped when the relative difference of all 5 model parameters was smaller than 1e-8. During the whole study no evidence was

found that the EM-algorithm was stopped too early or that no maximum was found.

5.3.1 Search for the largest local maximum

In practice, the global maximum of a function can be found through a global optimization method, i.e., a procedure scanning the whole domain of the function. However, as mentioned earlier on, these methods are very slow and in addition difficult to use due to the unboundedness of the likelihood function for a general mixture. Also, we did not consider as an option the search on a 5-dimensional grid of points for the largest maximum, as it would be to laborious. Instead, we made use of the specific nature of a mixture to describe populations where multiple groups are present. All methods considered, including the proposed starting value method are based on this property that is highly useful in the search for local maxima. Indeed, each maximum of the likelihood function (of a two-component mixture) can be related to a certain division of the sample into two groups. Conversely, some splits of the sample into two subsamples will give rise to a local maximum of the likelihood. Starting values converging to these maxima with an iterative procedure, can be easily derived from these splits in a similar way as explained in the introduction of Section 5.2.2.

As a result, based on all possible divisions of the sample into two groups, one should be able to detect all local maxima of the likelihood function, in particular the largest local maximum. This means that for a sample of size n, $2^{(n-1)} - (n+1)$ starting values will have to be verified. Unfortunately, this number increases exponentially and is even for small samples already huge as can be seen from column I of Table 5.2. As a consequence, this method I, is not useful in practice for finding the largest local maximum of a likelihood function. In fact, it can be regarded as a global optimization procedure adapted to the general (two-component) mixture model.

Apparently, method I generates many useless starting values. In-

Table 5.2: Number of possible starting values for several methods based on a split of a sample of size n into two subgroups.

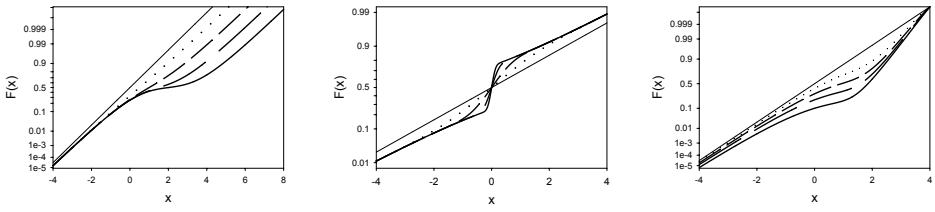
Sample	Method		
size n	I	II	III
10	501	35	$9+12$
20	524267	170	$19+12$
40	$5.498e^{11}$	740	$39+12$
60	$1.153e^{18}$	1710	$59+12$
80	$6.044e^{23}$	3080	$79+12$
100	$6.338e^{29}$	4850	$99+12$
150	$7.136e^{44}$	11025	$149+12$
200	$8.035e^{59}$	19700	$199+12$

deed, a lot of them converge either to the same maximum or to a point situated on the edge of the parameter space (for example, $\pi_1 = 0$ or 1) corresponding to a single distribution instead of a mixture or even to a singularity. According to Day (1969), divisions for which one of the two subsamples consists of data points sufficiently close together, will generate a local maximum. As opposed to this, divisions of the sample for which the two subsamples contain data points that are spread out over the whole ordered sample, will create starting values with quite similar values for the parameters of the first and the second component of the mixture. Such starting values often converge to a maximum that corresponds to a single (component) distribution. Hereby, *close* and *ordered* refer to some distance measure on the observations. Based on this, a simplified method II, for finding the largest local maximum can be derived. Namely, only these splits of the sample for which one of the two subgroups is formed by successive data points of the ordered sample, are considered. As indicated in column II of Table 5.2, the number of possible divisions of the sample decreases dramatically. Moreover, as suggested above and supported by subsequent simulations, at least one of these starting values will give rise to the largest local maximum.

Nevertheless, the number of starting values which has to be checked

remains too large. So, also method II is difficult to use for simulation purposes in case the sample size is larger than about $60 - 80$. The question is then how the number of starting values can be further decreased. For that, we approached the problem in a different way. During the development and testing of the tangent-rico method, it was observed that for those samples where the generated starting values did not converge to the largest local maximum with the EM-algorithm, this largest local maximum was almost always characterized through a very small (or large) value for the proportion parameter. Hereby, the simulated samples were small enough so that method II could be used to identify the largest local maximum. For these samples, irrespective of the value of m, it was usually not possible to derive a starting value that would converge to this largest local maximum. The reason for this is that the largest local maximum of the likelihood corresponded to a split of the sample into one group of all data points, except 2 or 3, and a very small subgroup containing the remaining 2 or 3 successive data points. In other words, a *distinct spurious* maximum was found as largest local maximum. Starting values to arrive at these maxima are related to a division of the sample where one of the two groups contains no more than a few successive data points. In general, these starting values are rarely derived with a starting value procedure.

Apparently, in combining the starting values related to splits of the sample with one of the two subgroups having only two successive data points and those starting values obtained with the tangent-rico method (at most 12), the largest local maximum of the likelihood can be found. Therefore, we considered it as a third method in order to search for the largest local maximum of the likelihood. It is easily seen then that with this method III there are at most $(n-1)+12$ splits of the sample possible. Importantly, as demonstrated hereafter, simulations indicate that this method nearly always leads to the largest local maximum, irrespective of the sample size.



(a) Group 1: dotted line: $\mu_2 = 1$, short dashed: $\mu_2 = 2$, long dashed: $\mu_2 = 3$, solid: $\mu_2 = 4$.

(b) Group 2: dotted line: $\sigma_1 = 1$, short dashed: $\sigma_1 = 0.5$, long dashed: $\sigma_1 = 0.2$, solid: $\sigma_1 = 0.1$.

(c) Group 3: dotted line: $\pi_1 = 0.8$, short dashed: $\pi_1 = 0.6$, long dashed: $\pi_1 = 0.4$, solid: $\pi_1 = 0.2$.

Figure 5.7: Cumulative distribution functions on a normal probability scale for the three groups of parameter values. The straight line is always the cdf of a normal distribution with $\mu = 0$ and $\sigma = 1$.

Results for two-component normal mixtures

Samples are generated from a two-component normal mixture with 12 different sets of parameter values divided into 3 groups of 4. In each group, one parameter is varied to consider one specific aspect of the separation of the mixture components. In the first group all parameters except the location parameter of the second component, μ_2, are kept fixed. The parameter μ_2 is varied to obtain a different degree of separation in location. The parameter values used are: $\mu_1 = 0, \sigma_1 = \sigma_2 = 1, \pi_1 = 0.5$ and $\mu_2 = 4, 3, 2, 1$. The larger the value of μ_2, the better the components of the mixture are separated and the more the theoretical QQ-plot of the mixture deviates from a straight line (Figure 5.7a). For the second group, the scale parameter of the first component, σ_1, is varied. The values for the parameters here are $\mu_1 = \mu_2 = 0, \sigma_2 = 2, \pi_1 = 0.5$ and $\sigma_1 = 0.1, 0.2, 0.5, 1$. Note that the components of the mixture can be clearly identified if the size of the ratio of the two scale parameters is large or small enough (Section 3.4). The more

this ratio deviates from 1, the better the two components are separated in scale (Figure 5.7b). In the last group, the proportion parameter π_1 is altered from a small value (0.2), over two averages values (0.4 and 0.6) to a large one (0.8). The values for the others parameter are $\mu_1 = 0, \sigma_1 = 1, \mu_2 = 2$ and $\sigma_2 = 0.5$. It is clear from Figure 5.7c, that also the size of π_1 has an influence on how well the mixture components are separated.

For each set of parameter values, sample sizes of $10, 20, 50$ and 100 were used, with 1000 simulations in each case. Due to the huge number of starting values, it is not possible, except for size 10, to use method I in practice. Instead, a number of divisions out of the total group of all possible divisions (or equivalently starting values) were randomly chosen. For poorly separated mixtures, we expect the probability to obtain at least one division which leads to the largest local maximum to be quite small. However, if during the simulation starting values are encountered which converge to maxima with a higher likelihood value than the maxima obtained with method II, then there is evidence that this method does not always give rise to the largest local maximum. For the sample sizes $20, 50$ and 100, respectively $1000, 1200$ and 2000 starting values were randomly chosen for each simulated sample.

Results are summarized in Tables 5.3a, 5.3b and 5.3c. For each combination of parameter values and sample size, the maximum with the largest likelihood value obtained with method II is compared to the largest maximum obtained by the other two methods. The first row for each sample size (referred to as $=$), gives the number of times that the maxima of the two methods considered are equal, the second row (referred to as $>$) indicates the number of times that the likelihood value of the maximum obtained with method II is larger than the likelihood value of the maximum obtained with one of the two other methods. Hereby, maxima are considered to be equal when the difference in likelihood value between the two maxima, is smaller than 1e-5, with a tolerance of 1e-8 used in the EM-algorithm. For most sets

of parameter values and choices of sample sizes the simulations last no more than 1 or 2 days. The exceptions are the sample sizes 50 and 100 for the very poorly separated mixtures: $\mu_1 = 1$ in the first group, $\sigma_2 = 1$ in the second group and $\pi_1 = 0.8$ in the third group. The main reason is the evaluation of the more than 5000 starting values in combination with the slowness of the EM-algorithm for this kind of mixtures. Note that in practice it was not possible to carry out simulations for sample sizes larger than 100. In that case method II is not feasible anymore.

The results point to two things. First, the equivalence between method II and III can be assumed. Apart from one or two exceptions in case of a very small sample size, the largest maximum obtained with either method is the same. None of the simulations shown here or performed before, indicated the opposite. Second, there is nothing to suggest that the largest local maximum is not obtained with method II (or method III). Up to now, irrespective of the sample, any maximum obtained from a starting value chosen at random from the group of starting values induced by method I, has no larger likelihood value than the largest maximum obtained with method II. As a result, method III and so also method II, can be seen as a method to obtain the largest local maximum of the likelihood function in case of a two-component (log)normal mixture model.

Results for two-component SEV mixtures

The same kind of simulation study as for normal mixtures has been set up. Three groups of sets of parameter values are considered. For the first and second group the same parameter values as for the normal mixture case are used. In the third group the proportion parameter is also varied, but now the QQ-plot of the mixture has basically a flat-steep-flat shape instead of a steep-flat-steep shape. The parameter values for the third group are $\mu_1 = 0$, $\sigma_1 = 0.5$, $\mu_2 = 1, \sigma_2 = 3$ and $\pi_1 = 0.2, 0.4, 0.6, 0.8$. For this combination of parameter values (i.e., $\sigma_1 < \sigma_2$), the mixture components

		$\mu_2 = 1$		$\mu_2 = 2$		$\mu_2 = 3$		$\mu_2 = 4$	
n		*I*	*III*	*I*	*III*	*I*	*III*	*I*	*III*
10	=	1000	1000	1000	999	1000	1000	1000	1000
	>	0	0	0	1	0	0	0	0
20	=	646	1000	687	1000	681	1000	778	1000
	>	354	0	313	0	319	0	222	0
50	=	229	1000	214	1000	347	1000	784	1000
	>	771	0	786	0	653	0	216	0
100	=	26	1000	95	1000	484	1000	965	1000
	>	974	0	905	0	516	0	35	0

(a) Group 1: separation in location.

		$\sigma_1 = 1$		$\sigma_1 = 0.5$		$\sigma_1 = 0.2$		$\sigma_1 = 0.1$	
n		*I*	*III*	*I*	*III*	*I*	*III*	*I*	*III*
10	=	1000	1000	999 (1)	999 (1)	1000	1000	999 (1)	999 (1)
	>	0	0	0	0	0	0	0	0
20	=	691	1000	733	999	855	1000	935	1000
	>	309	0	267	1	145	0	65	0
50	=	236	1000	537	1000	973	1000	999	1000
	>	764	0	463	0	27	0	1	0
100	=	156	1000	784	1000	1000	1000	1000	1000
	>	844	0	216	0	0	0	0	0

(b) Group 2: separation in scale. The number between brackets is the number of samples where no maximum at all was found.

		$\pi_1 = 0.2$		$\pi_1 = 0.4$		$\pi_1 = 0.6$		$\pi_1 = 0.8$	
n		*I*	*III*	*I*	*III*	*I*	*III*	*I*	*III*
10	=	1000	1000	1000	1000	1000	999	1000	1000
	>	0	0	0	0	0	1	0	0
20	=	779	1000	765	1000	708	1000	689	1000
	>	221	0	235	0	292	0	311	0
50	=	818	1000	718	1000	445	1000	287	1000
	>	182	0	282	0	555	0	713	0
100	=	982	1000	944	1000	575	1000	168	1000
	>	18	0	56	0	425	0	832	0

(c) Group 3: varying the proportion parameter.

Table 5.3: Comparison between the largest maximum obtained with method II and the largest maximum obtained with method I or III.

can be better identified for larger values of π_1 than for smaller values.

Simulations for the SEV-mixture require much more time due to the double iterative character of its EM-algorithm. Compared to the M-step for a normal mixture, there is no closed form solution of the M-step for the SEV mixture. This can slow down the EM-algorithm considerably and makes good starting values even more important since one cycle cost more in time. Therefore, for the simulations with the SEV mixture, we reduced both the number of simulated samples and the sample sizes used. Namely, for each set of parameter values, sample sizes of 20 and 50 are considered, with only 100 simulations in each case. Also, at each sample size only 200 starting values were randomly chosen for each simulated sample, in contrast to the more than 1000 random starting values for the normal case. Tables summarizing the results of the simulation experiment can be found in Section B.2.1 of the appendix. Main findings are similar to the normal mixture case, briefly:

- Methods *II* and *III* are equivalent.
- Method *II* (or *III*) leads generally to the likelihood estimate, irrespective of the sample size.

5.3.2 Consistency of the tangent-rico method

Although method *III* could be used to obtain the likelihood estimate, even for rather large sample sizes, apparently for sufficiently large sample sizes this method, and so its increasing number of starting values, is not necessary. Namely, we will empirically show that the tangent-rico method is consistent. In addition, the dependency between the sample size required to obtain the LE with the tangent-rico method and the separation of the mixture components is studied. Furthermore, we will compare the performance of the proposed starting value method with three other starting value methods, from which two are known to be consistent as well. Of those two, the first method uses the true values as starting values, while

the second makes use of the moment estimates. The former, however, cannot be applied in real terms, while the latter cannot always be used due to the fact that the moment estimates do not always exist (Section 5.1). A last starting value method generates at most 15 starting values, through randomly dividing the sample into two subsamples. This method is actually the method Mcl1 (Table 5.1). While not much is known about its consistency, this simulation study allows also to infer some consistency results. For all four methods, the largest maximum obtained from the EM-algorithm with their starting values, will be compared with the largest maximum obtained with method III.

Results for the two-component normal mixture

The same 12 sets of parameter values as in Section 5.3.1 for the normal mixture case, are used for the simulation study. Note that the groups are put together in such a way that the dependency can be studied between the minimum sample size required to obtain the LE with a certain starting value method and how well the mixture components can be identified on the mixture QQ-plot. Hereby, three aspects of the finite mixture model related to this identifiability should be looked at: the separation in location of the mixture components (group 1), the separation in scale of the mixture components (group 2) and the value of the proportion parameter (group 3).

Results of the simulation experiment are summarized in Tables 5.4, 5.5 and 5.6. For each set of parameter values, sample sizes of $20, 50, 100, 200, 300, 400, 500$ and 1000 were considered with 1000 simulations in each case. The tabulated value k is the number of times out of 1000 that the largest maximum obtained with the starting values of a certain starting value method is equal to the largest maximum obtained with method III. The number in brackets indicates how many times the moment estimates do not exist or that all starting values of a method converge to a singularity. Method A is the proposed starting value method, method B uses the true values as start-

Table 5.4: The number of times (k) out of 1000 that the starting value of a certain method leads to the LE for the sets of parameter values of group 1.

	$\mu_2 = 1$				$\mu_2 = 2$			
n	A	B	C	D	A	B	C	D
20	256 (3)	89	91 (208)		285 (1)	104	91 (163)	
50	153 (1)	19	16 (349)		157	34	35 (178)	
100	48	10 (58)	10 (427)	15 (10)	78	32 (5)	30 (203)	42 (1)
200	20	4 (28)	3 (437)	8 (3)	53	32 (7)	31 (148)	36
300	14	2 (29)	2 (458)	3 (7)	51	36 (1)	36 (127)	40
400	11	5 (26)	5 (455)	7 (7)	73	58	55 (91)	68
500	4	1 (21)	1 (447)	1 (5)	125	112	109 (70)	116
1000	1	0 (18)	0 (466)	1 (3)	368	358	351 (28)	366

	$\mu_2 = 3$				$\mu_2 = 4$			
n	A	B	C	D	A	B	C	D
20	326	177	175 (64)		544	426 (1)	420 (25)	
50	335	209	203 (57)		774	723	717 (11)	
100	470	415	416 (26)	430	970	965	959 (5)	967
200	800	788	784 (6)	794	1000	1000	1000	1000
300	951	949	947 (3)	950	1000	1000	1000	1000
400	989	988	988	989	1000	1000	1000	1000
500	999	998	998	999	1000	1000	1000	1000
1000	1000	1000	1000	1000	1000	1000	1000	1000

ing values, method C uses the moment estimates and method D generates randomly 15 divisions out of the total group of all possible divisions.

From these tables a number of conclusions can be drawn. First, the consistency of all four methods is clear. Indeed, for each set of parameter values, except for one set in the first group, the value of k approaches 1000 when n increases or k shows an increasing trend from a certain sample size on. As can be noted this is not completely true for method C, i.e., the method of moments. For some sets of parameter values the non-existence of the moment estimates is only a problem for small sample sizes (Table 5.4), but apparently for other sets this remains an issue regardless of the sample size (Tables 5.5, 5.6). There are even sets where the moment estimates only exist in no more than 50% of the simulated samples. As such, it is clear that this method cannot be used to obtain starting values for the general finite mixture model. Further, a set of 15 random starting values seems to be sufficient to obtain a consistent random starting value method.

Table 5.5: The number of times (k) out of 1000 that the starting value of a certain method leads to the LE for the sets of parameter values of group 2.

	$\sigma_1 = 1$				$\sigma_1 = 0.5$			
n	A	B	C	D	A	B	C	D
20	275 (4)	95 (138)	86 (356)		404 (2)	222	155 (279)	
50	194	57 (85)	38 (453)		504	400 (1)	284 (311)	
100	135	49 (39)	39 (362)	67	804	772	542 (292)	784
200	133	95 (15)	74 (334)	106 (5)	987	986	632 (358)	987
300	224	188 (4)	111 (320)	198 (2)	1000	1000	619 (379)	1000
400	281	259	164 (329)	263	1000	1000	561 (438)	1000
500	398	384	236 (346)	388	1000	1000	550 (450)	1000
1000	855	850	494 (408)	850	1000	1000	456 (544)	1000

	$\sigma_1 = 0.2$				$\sigma_1 = 0.1$			
n	A	B	C	D	A	B	C	D
20	733 (1)	594	251 (415)		858	809	298 (483)	
50	965	942	587 (355)		999	990	588 (375)	
100	1000	999	652 (343)	1000	1000	1000	656 (336)	1000
200	1000	1000	615 (385)	1000	1000	1000	594 (406)	1000
300	1000	1000	604 (396)	1000	1000	1000	588 (412)	1000
400	1000	1000	581 (419)	1000	1000	1000	571 (429)	1000
500	1000	1000	546 (454)	1000	1000	1000	524 (476)	1000
1000	1000	1000	433 (567)	1000	1000	1000	404 (596)	1000

Second, the better the mixture components are separated, i.e., the more the theoretical QQ-plot of the mixture differs from a straight line (Figure 5.7), the smaller the sample size n has to be such that the LE is reached with one of the starting values. This holds true for any of the starting value methods, as expressed in the Tables 5.4, 5.5 and 5.6. For some mixtures, a sample size of 50 will do, while for others a sample size of 1000 is not even sufficient. Note that for the third group of starting values, there is a dissimilar behavior for values of the proportion parameter larger than 0.5 versus those smaller than 0.5. Indeed, for smaller values of the proportion parameter the sample size n can be taken considerably smaller in order to converge to LE (with one of the constructed starting values) than for larger values of the proportion parameter. The reverse tendency holds in case $\sigma_1 < \sigma_2$.

Third, the proposed starting value method performs at least as good and in many cases even better than the other starting value methods.

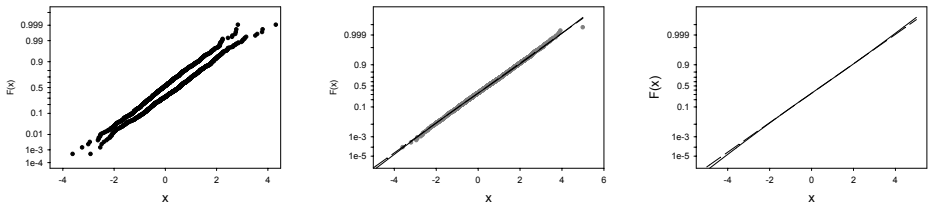
Table 5.6: The number of times (k) out of 1000 that the starting value of a certain method leads to the LE for the sets of parameter values of group 3.

	$\pi_1 = 0.2$				$\pi_1 = 0.4$			
n	A	B	C	D	A	B	C	D
20	551 (6)	393 (118)	355 (201)		501 (3)	349 (13)	337 (83)	
50	814	730	670 (101)		698	603 (1)	524 (153)	
100	984	947	831 (133)	970	953	914	776 (147)	943
200	1000	978	833 (158)	997	1000	985	821 (165)	999
300	1000	987	812 (184)	1000	1000	991	818 (174)	1000
400	1000	987	819 (180)	997	1000	995	826 (167)	1000
500	1000	994	816 (182)	998	1000	997	848 (149)	1000
1000	1000	997	842 (157)	1000	1000	997	843 (154)	1000

	$\pi_1 = 0.6$				$\pi_1 = 0.8$			
n	A	B	C	D	A	B	C	D
20	371	220 (7)	211 (105)		301 (1)	138 (33)	131 (150)	
50	403	296 (1)	261 (146)		200	83 (4)	70 (210)	
100	573	531	452 (171)	550	154	102	78 (221)	106
200	927	908	711 (166)	922	245	217	175 (218)	225
300	989	981	829 (151)	988	380	366	322 (179)	373
400	999	996	845 (147)	999	580	573	476 (195)	577
500	1000	997	867 (129)	1000	736	730	622 (178)	734
1000	1000	1000	901 (99)	1000	993	991	865 (129)	992

In other words, the number of times that the LE is reached when using the starting values of the proposed tangent-rico method, is at least as large as compared to the number of times that the LE is reached when using the true values or moment estimates as starting values. The performance of the at random starting value method is quite similar.

Fourth, there is one set of parameter values in the first group, namely $\mu_1 = 0$, $\sigma_1 = 1$, $\mu_2 = 1$, $\sigma_2 = 1$, $\pi_1 = 0.5$, where k does not finally show an increasing trend. For the given sample sizes it even get worse as n increases. The reason is clear: this mixture distribution is hardly distinguishable from a single normal distribution. The form of its QQ-plot (Figure 5.7a) is almost a straight line. As an example, Figure 5.8a shows both the normal QQ-plot of a simulated sample of size 1000 from this mixture distribution and the normal QQ-plot of a simulated sample of size 1000 of a normal distribution with parameter values $\mu = 0$, $\sigma = 1$. As noted, without prior knowledge, it is not possible to tell which of the two simulated samples



(a) Normal QQ-plot of a sample of size 1000 from the mixture distribution and of a sample of size 1000 form the normal distribution with parameters $\mu = 0$, $\sigma = 1$.

(b) Normal QQ-plot of a sample of size 5000 from the mixture distribution, the ML fit of the normal distribution (dashed line) and the cdf of the true mixture distribution (solid line).

(c) Theoretical QQ-plots of the mixture distribution (solid line) and of the normal distribution $\mu = 0$, $\sigma = 1.12$ (dashed line).

Figure 5.8: Identifiability of the mixture distribution with parameter values $\mu_1 = 0$, $\sigma_1 = 1$, $\mu_2 = 1$, $\sigma_2 = 1$, $\pi_1 = 0.5$.

stems from a mixture distribution. Moreover, one would consider them both coming from a normal distribution. Even a simulated sample of size 5000 of this mixture distribution cannot be recognized as coming from a mixture. As seen in Figure 5.8b, the ML fit of a normal distribution for a sample of size 5000 is almost equal to the cdf of the true mixture distribution. It is doubtful that any sample of this particular mixture distribution will be ever identified as coming from this mixture, without the need for an enormous sample size. The cdf of this mixture is too close to the cdf of the normal distribution with the same mean and standard deviation of the mixture (Figure 5.8c).

Results for the two-component SEV mixture

The same 12 sets of parameter values as those used in Section 5.3.1 for the SEV case, are considered. Compared to the simulation study for the

normal mixture distribution, a reduced number of simulations is carried out. Less sample sizes are handled (i.e., 20, 50, 100 and 500) and in each case only 100 samples are generated. Also, moment estimates are not used as starting values, since it is rather difficult to obtain them. Tables summarizing the results of the simulation study are given in Section B.2.2 of the appendix. The overall picture is the same as for the normal mixture distribution.

In addition, we take a closer look at the number of iterations required to obtain convergence with the EM-algorithm, given a set of starting values. As noted before, for the SEV mixture one cycle in the EM-algorithm is rather time consuming. As such, a smaller number of iterations positively influence the speed of the EM-algorithm. To compare the number of iterations necessary to obtain convergence with the starting values of a certain method, at each combination of sample size and set of parameter values used in the simulation study, the average number of iterations, taken over all simulated samples and starting values used with a certain method, is calculated. These averages are not only derived for the tangent-rico method, the at random method and for the true values, but also for the starting values used to obtain distinct spurious maxima (i.e., the starting values of method III, without the one of the tangent-rico method). The latter is referred to as method E in Table 5.7, which tabulates the resulting averages. Results point to the following conclusions.

- For small to moderate sample sizes, considerably less iterations are required with the starting values of the tangent-rico method, than with the starting values of the at random method. Hereby, the size of small (or moderate) depends on how well the mixture components are separated. In particular for poorly separated mixtures, an advantage is obtained with the tangent-rico method.

- For large sample sizes, this advantage does not exist. Sometimes the starting values of the at random method do require less cycles, while sometimes this is the case for the starting values of the tangent-rico

method. There is an indication that the former holds for mixtures with components mainly separated in scale, while the latter seems to hold for mixtures separated in location.

- Mostly, for small to moderate sample sizes the true values are no better starting values than the one obtained with the tangent-rico method. For large sample sizes, they outperform both the tangent-rico method and the at random method.

- The best starting values (with regard to the number of iterations required) are without any doubt the starting values of method E. Regardless of the sample size and the separation of the mixture components, they require only a limited number of iterations. This illustrates the feasibility, even for moderate samples sizes, of method *III*.

Conclusions

In summary, we believe to have shown the excellent performance of the proposed starting value method. It works as good and mostly even better than using the true values as starting values. For sufficiently large sample sizes, the LE is always reached with one of the starting values (in combination with the EM-algorithm). The required sample size depends on how well the components of the true mixture distribution can be identified or how well they are separated. In addition, we provided a method (*III*), feasible for medium sample sizes, that for any sample give rise to the LE.

Next to this, it is shown that the at random starting value method with a set of 15 starting values is consistent as well. Still, for small to moderate sample sizes these starting values need more iterations until convergence of the EM-algorithm. Especially, in case of poorly separated mixtures, these starting values slow down the algorithm a lot.

Table 5.7: The average number of iterations required until convergence with the starting values of a certain method.

	$\mu_2 = 1$				$\mu_2 = 2$			
n	A	B	D	E	A	B	D	E
20					91	131	143	38
50					233	547	449	39
100					300	434	484	39
500					682	741	847	42
	$\mu_2 = 3$				$\mu_2 = 4$			
n	A	B	D	E	A	B	D	E
20	78	106	126	35	42	37	71	30
50	105	125	182	39	66	62	109	37
100	133	127	203	40	67	55	122	40
500	174	142	261	43	73	50	159	41

(a) Sets of parameter values of group 1.

	$\sigma_1 = 1$				$\sigma_1 = 0.5$			
n	A	B	D	E	A	B	D	E
20					73	125	128	33
50					88	135	177	38
100					106	129	150	39
500					134	105	129	42
	$\sigma_1 = 0.2$				$\sigma_1 = 0.1$			
n	A	B	D	E	A	B	D	E
20	42	34	72	28	38	22	59	24
50	44	38	51	31	37	24	39	26
100	52	37	50	33	51	24	38	28
500	77	35	49	35	61	21	34	29

(b) Sets of parameter values of group 2.

	$\pi_1 = 0.2$				$\pi_1 = 0.4$			
n	A	B	D	E	A	B	D	E
20	89	63	1148	39	56	53	142	31
50	135	114	556	39	86	84	164	34
100	225	121	450	38	94	60	91	37
500	383	110	326	40	131	61	91	40
	$\pi_1 = 0.6$				$\pi_1 = 0.8$			
n	A	B	D	E	A	B	D	E
20	55	61	70	35	100	60	108	33
50	54	52	60	35	49	48	58	35
100	53	44	53	36	45	40	49	40
500	56	40	52	36	39	31	40	37

(c) Sets of parameter values of group 3.

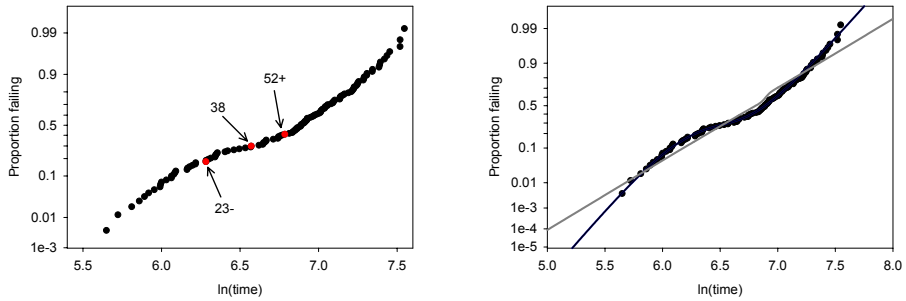
5.4 Additional features of the tangent-rico method

The tangent-rico method was initially developed as a method to construct well-reasoned and data driven starting values. Nevertheless, it turned out that the method itself and some of the plots used in the method add important information to the analysis of a sample when considering a finite mixture distribution. On the one hand, given a specific finite mixture model, the set of maxima obtained with the starting values of the tangent-rico method, give a good indication of the stability of the sample. In some way, these maxima can be viewed as the result of a well-considered scan of the likelihood surface. On the other hand, without specifying the number of components, plots of the derivative and of the cosine of the tangential deflection of the QQ-plot for several values of m, are good exploration tools. They indicate the maximum amount of information available in the sample (with respect to the number of components) and where to situate the different components in the sample. In this section, only the situation of at most two mixture components is handled. The idea behind it, is based on the theoretical counterparts of these plots, when the true distribution is a two-component mixture (Section 5.2.1). A generalization to more than two components is discussed in Section 5.5.2.

In the following, these two features of the tangent-rico method are worked out by means of three examples.

5.4.1 Example 1: the resistor sample

The lognormal QQ-plot of this sample, shown in Figure 5.9, has a pronounced steep-flat-steep shape. It is a classic example of a sample from which its stability (with respect to a two-component lognormal mixture) is already noticed from the QQ-plot itself. Unfortunately, such samples are not often encountered in practice.



(a) Position of the "best" inflection point for $m = 25$ with neighboring nodes.

(b) LE fit of the sample and fit of the other maximum found with the tangent-rico method.

Figure 5.9: Lognormal QQ-plots of the resistor sample.

Applying the tangent-rico method

For a sample of size 125, the values of the smoothing parameter m, used in the tangent-rico method, are 7, 13 and 25 (5%, 10% and 20% of the sample size rounded to the nearest odd integer). For $m = 25$ there is only one appropriate candidate couple of nodes (Figure 5.11b), which results in a total of 11 starting values instead of 12. The best candidate inflection point found for $m = 25$, i.e., data point 38, is indicated on Figure 5.9a. All starting values but one converge to the same maximum with the EM-algorithm. Table 5.8 gives the parameter values and value of the logarithm of the likelihood for these two maxima. As noted, the difference in likelihood value between them is huge. These two facts clearly indicate that the sample is stable. This is confirmed through carrying out method *III*. No larger maximum is detected. Moreover, all other maxima found have a likelihood values which is considerably smaller than the likelihood value of the largest local maximum. To illustrate the latter, the second largest maximum found,

Table 5.8: Local maxima of the likelihood function for the resistor sample in case of a two-component lognormal mixture. Estimated parameters are the location and scale parameters of the log failure times.

method	maximum	$\hat{\mu}_1$	$\hat{\sigma}_1$	$\hat{\mu}_2$	$\hat{\sigma}_2$	$\hat{\pi}_1$	lnL
TR	LE	6.164	0.236	7.022	0.251	0.286	-66.100
TR		6.769	0.474	6.895	0.030	0.937	-78.284
III		6.745	3.55e-5	6.777	0.464	0.0159	-72.545

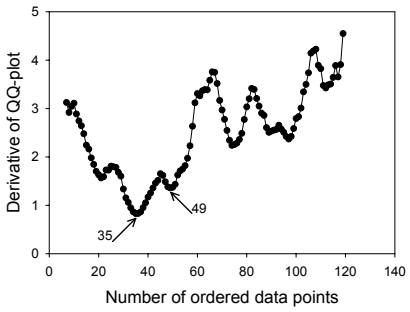
Note: TR refers to the tangent-rico method.

which is distinct spurious, is included in Table 5.8.

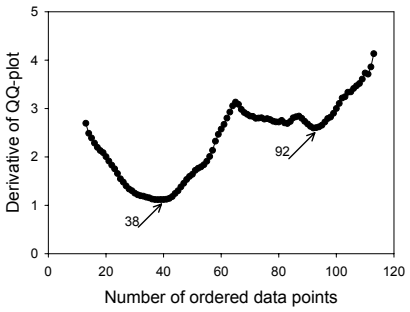
Thus, for this sample no evidence of an unstable likelihood surface is found. It was possible to infer this directly from the maxima obtained with the starting values of the tangent-rico method. There is no reason to mistrust the largest maximum found with the starting values of the tangent-rico method. In conclusion, the fit of the LE is shown in Figure 5.9b. As a comparison the poor fit of the other maximum obtained is also shown.

Plots derived from the QQ-plot

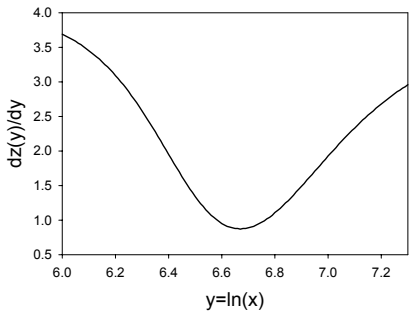
Plots of the derivative and of the cosine of the tangential deflection for the values 13 and 25 are shown in Figures 5.10 and 5.11. The plots for $m = 7$ are similar to the plots for $m = 13$, only less smoothed. Both kind of plots point to the fact that a two-component mixture is appropriate as distribution for the sample. This is most easily seen from the plots for $m = 25$. In particular, the derivative plot has one marked minimum (situated at data point 38). Moreover, there is a good resemblance between the derivative plot of the theoretical QQ-plot of the two-component lognormal mixture with the LEs as parameter values (Figure 5.10c) and its empirical counterpart for $m = 25$ (Figure 5.10d), i.e., with the logarithm of the failure times used as x-coordinate and not the number of ordered data points. Further, the presence of two clear minima for the plot of the cosine of the tangential deflection is not in conflict with the steep-flat-steep shape of the QQ-plot.



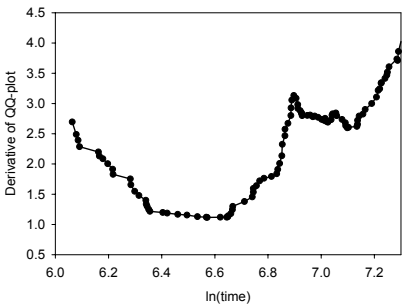
(a) $m = 13$.



(b) $m = 25$.

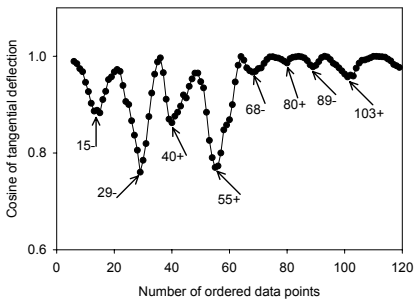


(c) Derivative of the theoretical QQ-plot of the two-component lognormal mixture with as parameter values the LEs.

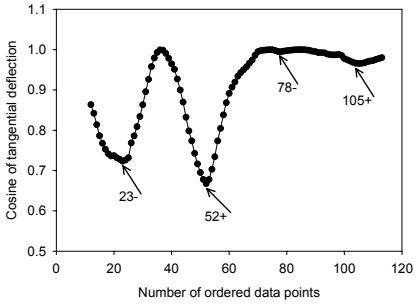


(d) Derivative plot of the QQ-plot for $m = 25$.

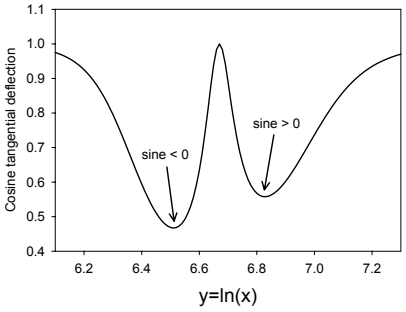
Figure 5.10: Plots of the derivative of the lognormal QQ-plot of the resistor sample.



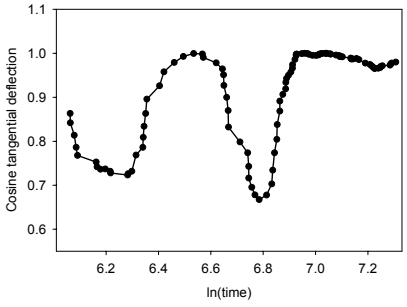
(a) $m = 13$.



(b) $m = 25$.

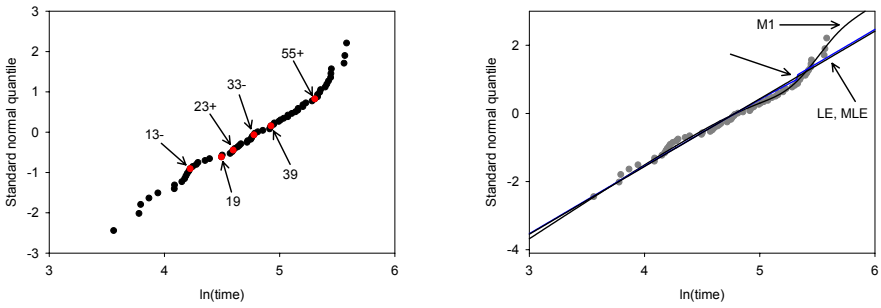


(c) Cosine of the tangential deflection of the theoretical QQ-plot of the two-component lognormal mixture with as parameter values the LEs.



(d) Cosine plot of the tangential deflection of the QQ-plot for $m = 25$.

Figure 5.11: Plots of the cosine of the tangential deflection of the lognormal QQ-plot of the resistor sample. The + refers to a positive sine for the tangential deflection in a point, the − for a negative sine.



(a) Position of the two "best" candidate inflection points and some candidate nodes for $m = 13$.

(b) LE fit, fit of the plausible maximum found (M1) and MLE fit of a single normal distribution.

Figure 5.12: Lognormal QQ-plot of the interconnect sample.

Indeed, the sign of the sine for the first minimum is negative and positive for the second (i.e., the reverse order of for a candidate couple of nodes). These two nodes are situated around the candidate inflection point (Figure 5.9a). Again the empirical plot corresponds nicely to its theoretical counterpart (Figures 5.11c and 5.11d).

Although, for smaller values of m the plots contain more minima and so more suggestions for other candidate inflection points, the general tendency is clear. Namely, these plots suggest the presence of two subsamples and a two-component distribution for the sample with the components separated in location.

5.4.2 Example 2: the interconnect sample

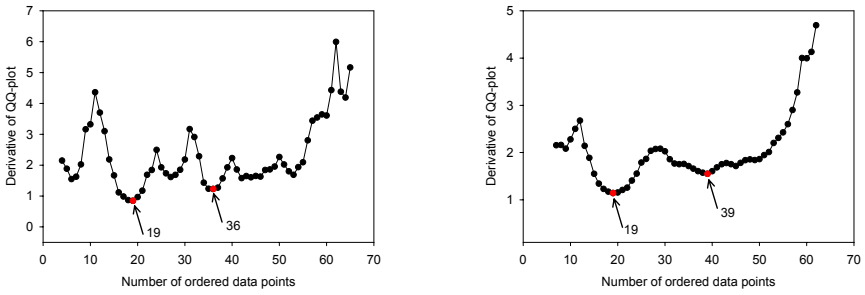
The lognormal QQ-plot of the interconnect sample is shown in Figure 5.12a. Although the shape of this QQ-plot deviates, mainly at the end, from a straight line, it does not reveal either the presence of an additional

failure mode for the devices under study.

Applying the tangent-rico method

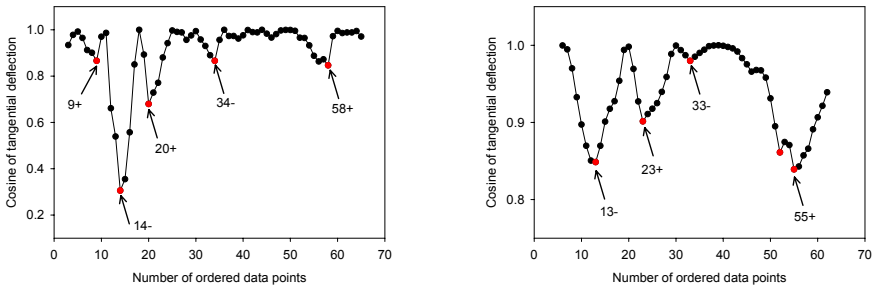
Based on sample size 68, the m-values for the tangent-rico method are 3, 7 and 13. In total, 11 starting values are derived. For $m = 13$, there was only one appropriate candidate couple of nodes available (Figure 5.14b). These starting values converged, with the EM-algorithm, to 4 different maxima, given in Table 5.9. Although the largest maximum found is not distinct spurious, this result points rather to an unstable than a stable sample. It is doubtful whether the largest maximum found is the LE. Not only too many different maxima are obtained (with only 11 starting values), but also there is no large difference in the likelihood values. For this kind of outcome, it is sensible to also apply method *III*. By carrying out the latter, we indeed found some distinct spurious maxima with a larger likelihood value. In particular, the LE is distinct spurious. As such, this sample is unstable. Apart from maximum M1, found with the starting values of the tangent-rico method, most maxima identified are distinct spurious or have a rather small value for the proportion parameter. From this point of view, it is tempting to base inference results on this "plausible" maximum M1. Nevertheless, we do not advise this, since the sample clearly contains not enough information to model two failure modes. In addition, even if it would be the maximum closest to the true values, is asymptotic properties cannot be guaranteed.

In conclusion, Figure 5.12b displays the fits of the LE and of maximum M1, together with the ML fit of a single lognormal distribution. Within the range of data, only at the end there is a substantial difference between the fits of the LE and the MLE on the one hand and the fit of maximum M1 at the other hand. The arrow indicates where the LE fit and the MLE fit differ.



(a) $m = 7$. (b) $m = 13$.

Figure 5.13: Plots of the derivative of the lognormal QQ-plot of the interconnect sample.



(a) $m = 7$. (b) $m = 13$.

Figure 5.14: Plots of the cosine of the tangential deflection of the lognormal QQ-plot of the interconnect sample. The + refers to a positive sine for the tangential deflection in a point, the − for a negative sine.

Table 5.9: Local maxima of the likelihood function for the interconnect sample for a two-component lognormal mixture. Estimated parameters are the location and scale parameters of the log failure times.

method	maximum	$\hat{\mu}_1$	$\hat{\sigma}_1$	$\hat{\mu}_2$	$\hat{\sigma}_2$	$\hat{\pi}_1$	lnL
TR	M1	4.610	0.446	5.340	0.146	0.763	-45.595
TR		4.821	0.496	4.197	0.0244	0.940	-47.588
TR		4.780	0.511	4.919	0.00447	0.970	-48.446
TR		4.784	0.512	4.765	0.0115	0.966	-48.989
III	LE	5.329	8.22e-5	4.767	0.502	0.0293	-41.191
III		5.071	9.41e-5	4.775	0.508	0.0292	-42.319

Plots derived from the QQ-plot

Figures 5.13 and 5.14 depict the plots of the derivative and of the cosine of the tangential deflection of the QQ-plot at the m-values 7 and 13. The plots for $m = 3$ are not shown, they pronounce even more the small random deviations in the sample. From both kind of plots, there is no straight evidence that a two-component mixture would be appropriate as distribution for the sample. Certainly, for $m = 7$ the derivative plot shows too many minima around the same value. Although the two smallest minima remain the same for $m = 13$, there is no minimum which dominates. Note that the starting values derived from the candidate inflection point 39 (at $m = 13$) are close to the parameter estimates corresponding to maximum M1. The same can be said for the plots of the cosine: there is no convincing candidate couple of nodes. Also with each couple of nodes, a large local maximum should correspond for the derivative plot (situated between the two nodes). But for the maxima of the derivative plot the same holds true as for its minima, i.e., there is no clear maximum, except perhaps at the end. The corresponding deviation on the QQ-plot, however, is too small and not distinct enough, to be recognized as being not random.

5.4.3 Example 3: appliance failure sample

Both a lognormal and a Weibull mixture are considered. Figures 5.15a and 5.16a depict the Weibull, respectively the lognormal QQ-plot of this sample. At first sight, differences between these plots are rather small. The shape of both QQ-plots verges to a steep-flat-steep shape.

Applying the tangent-rico method

Regardless of the model under consideration, m takes on the values 3, 7 and 13 in the tangent-rico method. In both cases, a full set of 12 starting values is obtained.

A two-component Weibull mixture The derived starting values converge to 3 different maxima, given in Table 5.10a. From this, the stability of the sample cannot be guaranteed since at least two maxima (M1, M2) seem to be at the top of the likelihood function. Through method *III*, two distinct spurious maxima, among which the LE, are detected with a larger likelihood value than maximum M1. So, this sample is unstable. Figure 5.15b illustrates why a distinct spurious maximum is at the top of the likelihood function. The fits of the LE and of maxima M1 and M2 are shown. For more than 80% of the data (i.e., the data situated within the displayed box), the three fits are comparable. Apparently, the amount of data situated outside the box is not large enough or is still situated to close to the straight line modeling the data inside the box, to guarantee that this deviation is not purely random.

A two-component lognormal mixture Only two maxima are found with the starting values of the tangent-rico method. They are tabulated in Table 5.10b. There is a large difference in likelihood value between these two maxima. The sample seems stable and the largest maximum found is likely to be the largest local maximum. This is confirmed by applying method

method	maximum	$\hat{\sigma}_1$	$\hat{\mu}_1$	$\hat{\sigma}_2$	$\hat{\mu}_2$	$\hat{\pi}_1$	lnL
TR	M1	0.601	4.558	0.713	7.928	0.137	-92.699
TR	M2	1.272	7.227	0.410	8.163	0.598	-94.712
TR		1.010	7.704	0.107	7.113	0.969	-97.378
III	LE	0.00150	8.321	1.025	7.638	0.0481	-90.643
III		0.000606	8.320	1.016	7.657	0.0325	-92.104

(a) Estimation of a two-component Weibull mixture

method	maximum	$\hat{\mu}_1$	$\hat{\sigma}_1$	$\hat{\mu}_2$	$\hat{\sigma}_2$	$\hat{\pi}_1$	lnL
TR	LE	4.728	1.074	7.687	0.693	0.210	-93.834
TR	M1	6.895	1.454	8.312	0.0631	0.878	-101.298
III	M2	8.321	0.00160	7.003	1.447	0.0488	-98.728

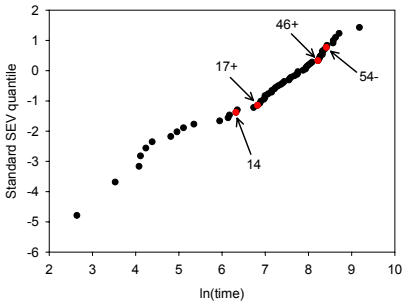
(b) Estimation of a two-component lognormal mixture

Table 5.10: Local maxima of the likelihood function for the appliance failure sample. Estimated parameters are the location and scale parameters of the log failure times.

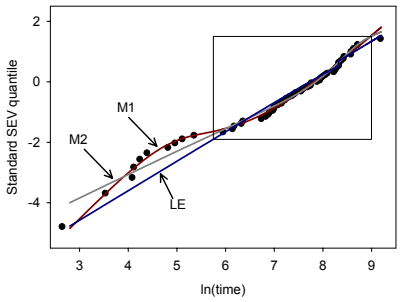
III. No larger maximum is obtained and the likelihood value of the first maximum (M2) that comes after the LE is considerably smaller than the likelihood value of the LE. The fits of the LE and maxima M1 and M2 are shown in Figure 5.16. Here the LE-fit can be distinguished over the whole range of data from the fits of the other two maxima. In contrast to the Weibull case where a single distribution is still an option, this is not the case if the mixture component is assumed to have a lognormal distribution. Clearly, a single lognormal distribution is not appropriate.

Plots derived of the QQ-plot

We now take a closer look at the plots of the derivative and of the cosine of the tangential deflection for both the Weibull and lognormal QQ-plot. All derivative plots are included in Figure 5.17, while the cosine

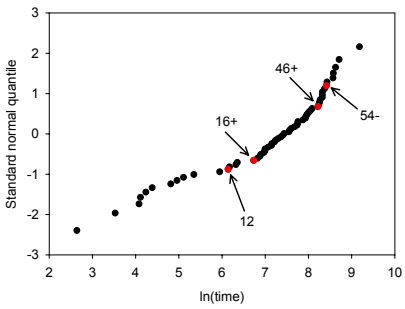


(a) Position of "best" candidate inflection point and some candidate nodes for $m = 13$.

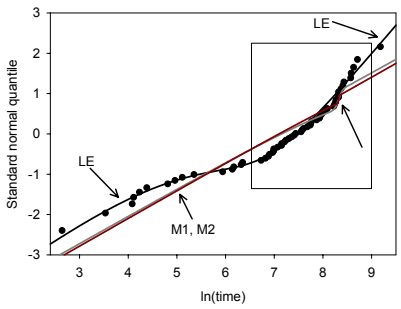


(b) LE fit, fit of maximum M1 and fit of maximum M2.

Figure 5.15: Weibull QQ-plot of the appliance failure sample.



(a) Position of "best" candidate inflection point and some candidate nodes for $m = 13$.



(b) LE fit, fit of maximum M1 and fit of maximum M2.

Figure 5.16: Lognormal QQ-plot of the appliance failure sample.

plots are given in Figure 5.18. The plots derived from the Weibull QQ-plot are quite similar in nature to the ones derived from the lognormal QQ-plot. From both, there is the suggestion that the sample consists of two subgroups. Although, this is more evident for the lognormal case than for the Weibull case. In particular, in both cases the plots of the derivative have a pronounced global minimum which points to an obvious candidate inflection point. This point is situated around data point 12 and related to a couple of minima on the cosine plot situated around data points 6 and 16. But for the lognormal derivative plots, the global tendency after this minimum is increasing, while for the Weibull derivative plots this tendency is less clear. This is illustrated further through the fact that the second candidate inflection point situated around data point 31-34, give rise to another maximum for the Weibull mixture (i.e., maximum M2), while this is not the case for the lognormal mixture.

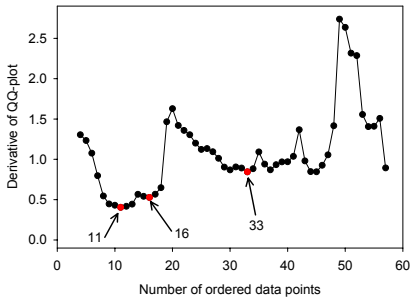
Furthermore, the maximum at the end of the derivative plots corresponds to a candidate couple of nodes (around data points 47 and 54). Based on this candidate couple of nodes, maximum M1 is obtained for the two-component lognormal mixture, while no maximum is found in case of a Weibull mixture. Still, this maximum should be rather looked at as the maximum belonging to the global minimum in case of steep-flat-steep shape (and so corresponding to the couple of nodes around data points 17 and 54), then as a maximum pointing to an additional third component.

5.4.4 Summary

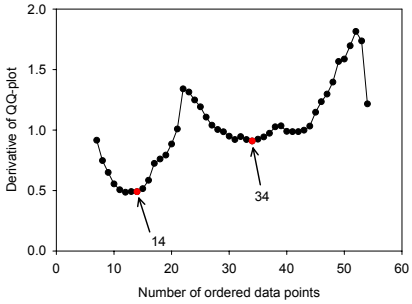
For a general two-component mixture model, mostly it is possible to infer the stability of the sample given the maxima obtained with the starting values of the tangent-rico method. Therefore, one has to look at:

1. The number of different maxima found.

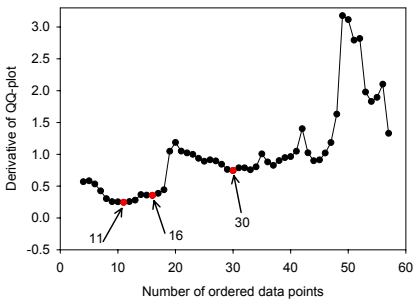
2. The difference in likelihood value between the several maxima found.



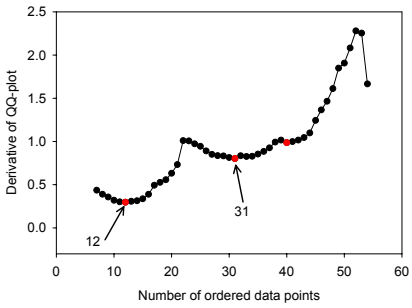
(a) Derivative of Weibull QQ-plot with $m = 7$.



(b) Derivative of Weibull QQ-plot with $m = 13$.

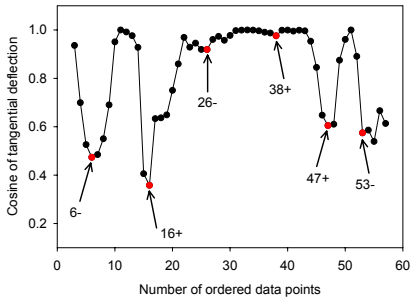


(c) Derivative of lognormal QQ-plot with $m = 7$.

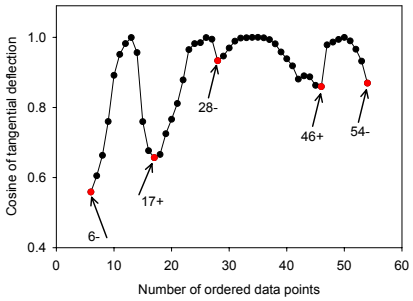


(d) Derivative of lognormal QQ-plot with $m = 13$.

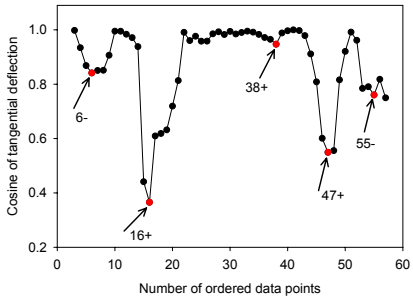
Figure 5.17: Derivative of the QQ-plot of the appliance failure sample.



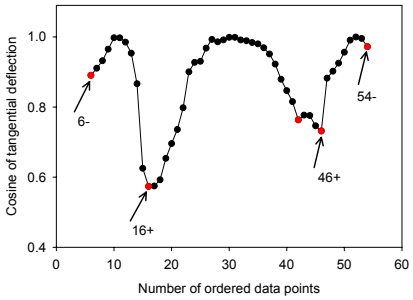
(a) Cosine of tangential deflection of Weibull QQ-plot with $m = 7$.



(b) Cosine of tangential deflection of Weibull QQ-plot with $m = 13$.



(c) Cosine of tangential deflection of lognormal QQ-plot with $m = 7$.



(d) Cosine of tangential deflection of lognormal QQ-plot with $m = 13$.

Figure 5.18: The cosine of the tangential deflection of the QQ-plot of the appliance failure sample.

In general, for a stable sample the largest maximum detected with the starting values of the tangent-rico method, is also the LE. For an unstable sample, method *III* could be carried out to obtain a more complete view of the surface of the likelihood function.

Given a distribution for the mixture components, both the plots of the derivative and of the cosine of the tangential deflection can be simply derived for several values of the smoothing parameter m. Hereby, the choices 5, 10 and 20% of the sample size are sufficient to give a global picture. These plots can be read as follows:

- A distinct (or dominant) global minimum for the derivative plot points to a candidate inflection point, i.e., to a mixing of two distributions separated in location. This candidate indicates more or less where the domination of the first component ends and the domination of the second component begins. On the corresponding plot of the cosine of the tangential deflection, before and after this candidate a clear minimum should be situated; the first having a negative sine, the last a positive sine.

- A distinct (or dominant) global maximum for the derivative plot points to a candidate couple of nodes, i.e., to a mixing of two distributions separated in scale. On the corresponding plot of the cosine of the tangential deflection, before and after this maximum a minimum should be located. This first minimum should have a positive sine, the second a negative sine. At the same time, these two minima form the candidate couple of nodes. Their positions indicate where the different components of the mixture are dominating.

- If there are no pronounced minima and maxima for the derivative plot, quite likely a mixture distribution is not appropriate as a distribution for the sample.

These guidelines are based on the theoretical counterparts of the derivative and cosine plot, given a two-component mixture. They can be extended to the case of more than two components. This and the fact that it is possible to detect any kind of mixture (i.e., also the one separated in scale) on these plots, give them an advantage over other exploration tools.

5.5 Extensions of the tangent-rico method

The tangent-rico method allows an easy extension to many other situations. Here, the most important ones will be discussed. In Section 5.5.1, we consider the extension to censored data problems, which includes the derivation of starting values for the adapted likelihood methods. Further, Section 5.5.2 handles the calculation of starting values for finite mixtures with more than two components. In the last section, the simplified tangent-rico method for the finite mixture distribution with a common location or scale parameter is briefly discussed.

5.5.1 Censoring

One of the main advantages of using a probability plot as the basis of a starting value method, is that this plot can be constructed too in case of censoring. At least, as long as a random censoring mechanism is assumed. As such, to adapt the tangent-rico method to censored samples, only the following two changes has to be carried out:

- The probability plot has to be adjusted to a censored sample.
- The division of the sample into two subsamples has to take into consideration the censored data points.

We will consider into more detail two right censoring mechanisms which often occur in reliability situations. Further, we will handle the case of interval censoring to obtain starting values for the adapted LEs.

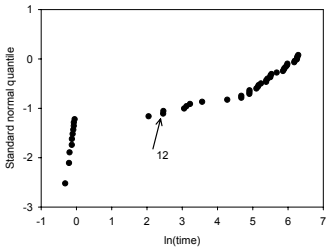
Note that a problem of more concern for censored samples, is the use of the tangent-rico method in combination with the EM-algorithm. In particular, the EM-algorithm becomes for any general mixture with a (log)location-scale distribution as component, a double iterative procedure, as there exists no closed form solution anymore for the M-step. But, in contrast to the double iterative EM-algorithm for complete SEV (Weibull) mixtures, the slowness of the EM-algorithm is now a potential problem. The reason is that censoring is a second kind of missing data, next to the missing group information for each observation. As a result, the EM-algorithm will slow down considerably in case of heavy censored samples. Therefore, for the samples used throughout this work, we also looked at the performance of the NR-method. Although more research is required, it appears that the NR-method could be a useful alternative for censored data problems. Namely, mostly the same group of maxima as with the EM-algorithm, is obtained with the NR-method (from the starting values of the tangent-rico method) and we had almost no convergence problems with the NR-method.

Type I Singly Right censoring

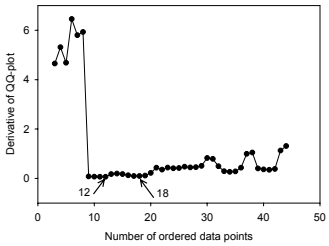
A type I singly right censored sample consists of r failure times and $n-r$ (right) censored observations at the same time point c. If $(y_{(1)}, \ldots, y_{(r)})$ are the ordered failure times, than it holds that $c \geq y_{(r)}$. Although, the probability plot has only $n-r$ points (related to the failures), the plotting positions for these points are the same as in case of a complete sample. So, the probability plot is made up of the points $(y_{(i)}, p_i = 1 - e^{-\frac{1}{2}(S_{i-1}+S_i)})$, $i = 1, \ldots, r$. Given then a candidate inflection point, situated at the s^{th} ordered (failed) observation, the sample is divided in the subsamples $(y_{(1)}, \ldots, y_{(s-1)})$ and $(y_{(s+1)}, \ldots, y_{(r)})$ together with the r censored observations. For a couple of candidate nodes at the s_1^{th} and s_2^{th} data points, the first subsample consists of the failure times $(y_{(1)}, \ldots, y_{(s_1-1)}, y_{(s_2+1)}, \ldots, y_{(r)})$ supplemented with the r censored observations. The second subsample equals

then $(y_{(s_1+1)}, \ldots, y_{(s_2-1)})$.

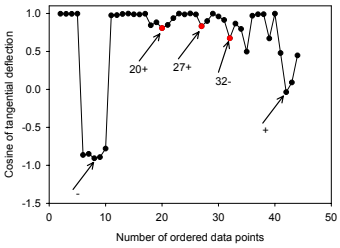
Example The censored laser A sample, introduced in Section 2.1.3, is a type I singly right censored sample, with sample size 85 and 39 censored observations. Its lognormal QQ-plot is shown in Figure 5.19a. It has a pronounced steep-flat-steep shape. For the tangent-rico method, m takes on the values 3, 5 and 9, i.e., 5, 10 and 20% of 46 (the number of points on the QQ-plot). A plot of the derivative for $m = 5$ is given in Figure 5.19b and of the cosine of the tangential deflection in Figure 5.19c. For both, a similar picture is obtained at the other m-values. The derivative plot is dominated through a big jump, which is related to one obvious candidate inflection point. This is confirmed through the presence of one clear minimum with a negative sine, situated just before this jump, on the cosine plot. On both kind of plots, no other distinct features can be observed. Figure 5.19d shows the QQ-plots of the two subsamples based on a division of the best candidate inflection point for $m = 5$, i.e., the 12^{th} data point, and the corresponding ML fits. As noted, the first (uncensored) subsample has an outlying observation. This results in a rather poor ML fit. For $m = 3$, the best candidate was data point 11, resulting in a first subsample without an outlying observation and a much better ML fit. Note that the second subsample is also type I singly right censored. To illustrate the difference between the starting values obtained from a division of the sample based on the 11^{th} and the 12^{th} data point as inflection point, Figure 5.19e shows the fits corresponding to both sets of starting values. Although both fits are good, the one corresponding to the 11^{th} data point is excellent. Moreover, almost no difference exists with the LE-fit for a two-component lognormal mixture. Both sets of starting values lead to the LE. The estimates are $\hat{\mu}_1 = -0.14, \hat{\sigma}_1 = 0.078, \hat{\mu}_2 = 6.46, \hat{\sigma}_2 = 1.81$ and $\hat{\pi}_1 = 0.12$. The sample is clearly stable. The 12 starting values, obtained with the tangent-rico method, converged to 3 different maxima with a huge difference in likelihood



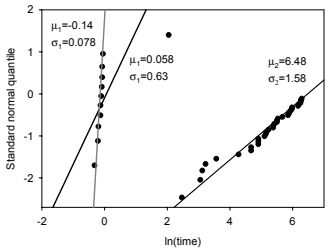
(a) Lognormal QQ-plot.



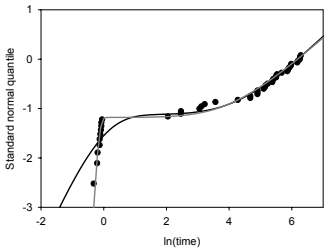
(b) Derivative of the QQ-plot for $m = 5$.



(c) Cosine of the tangential deflection of the QQ-plot for $m = 5$.



(d) Subsamples and ML-fits for a division based on the 12^{th} data point. The grey line is the ML fit for the first subsample without the outlying point.



(e) Fits of the starting values for the best choice for $m = 3$ (grey line) and $m = 5$ (black line).

Figure 5.19: Deriving starting values for the type I singly right censored laser A sample.

value between the largest and second largest maximum found (i.e., -117.436 versus -138.200).

Multiple right censoring

In contrast to the previous situation, censored observations can occur at any point in time. The sample consists of n time points among which r failure times y_i and $n - r$ censoring times c_i. Only failed observations are taken up in the probability plot. If S is the set of all indices of the failure times in the ordered sample, then the plotting positions are defined by $p_i = 1 - e^{-\frac{1}{2}(S_{i-1}+S_i)}$, $i \in S$, with $S_i = \sum_{j \in S, j \leq i} \frac{1}{n-j+1}$. For an inflection point based on the s^{th} ordered failed observation $y_{(s)}$, the sample is divided into the subsamples $(y_{(1)}, \ldots, y_{(s-1)}, c_{(1)}, \ldots, c_{(t)})$, with $c_{(t)} < y_{(s)}$ and $(y_{(s+1)}, \ldots, y_{(r)}, c_{(t+1)}, \ldots, c_{(n-r)})$. Note that early censored observations could perhaps also belong to the second component. However, it is quite impossible to retrieve this information. We only propose one way to deal with multiple censoring. For a couple candidate nodes situated at the s_1^{th} and s_2^{th} ordered failure times, the two subsamples become $(y_{(1)}, \ldots, y_{(s_1-1)}, c_{(1)}, \ldots, c_{(t_1)}, y_{(s_2+1)}, \ldots, y_{(r)}, c_{(t_2+1)}, \ldots, c_{(n-r)})$ and $(y_{(s_1+1)}, \ldots, y_{(s_2-1)}, c_{(t_1+1)}, \ldots, c_{(t_2)})$, with $c_{(t_1)} < y_{(s_1)}$ and $c_{(t_2+1)} > y_{(s_2)}$.

Example Still to work out: deriving starting values for a multiple right censored Weibull sample.

Interval censoring

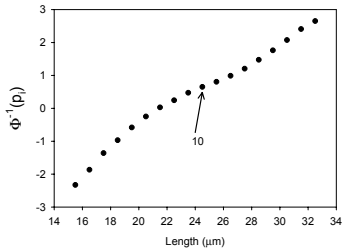
Each observation of an interval censored sample is both left and right censored. No time point is observed, but a time interval in which the event (i.e, a failure or a removal) has occurred. We will only consider grouped or binned samples with non-overlapping intervals. At worst intervals are adjacent, i.e., one interval begins where the previous one ends. This last

situation typically occurs for a large value of the measurement error δ or for a small number of binned intervals.

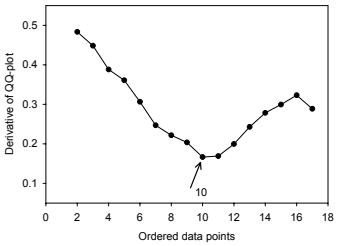
A complete interval-censored sample with n observations, consists of m intervals $]t_{i-1}, t_i]$ with in each interval d_i failed observations ($\sum_{i=1}^{m} d_i = n$). For the construction of a probability plot we follow Meeker and Escobar (1998, pp. 132-134): on the appropriate probability scales the upper endpoints t_i of the intervals are plotted against the nonparametric estimate $\hat{F}(t_i) = \frac{\sum_{j=1}^{i} d_j}{n} = p_i$ as plotting position. Note that with this choice the last point $\hat{F}(t_m) = 1$ cannot be plotted. The division of the sample is handled in a similar way as for a complete sample.

Interval-censored samples can also be right-censored, i.e., the observed event is sometimes a removal. Each interval has then d_j failed observations and r_j censored observations. In this case, the plotting position for the upper endpoint t_i is changed into $1 - \prod_{j=1}^{i}[1 - \frac{d_j}{n_j}]$ with $n_i = n - \sum_{j=0}^{i-1} d_j - \sum_{j=0}^{i-1} r_j$, $r_0 = 0$ and $d_0 = 0$. If each interval contains at least one failed observation, then the division of the sample can be done in the same way as for a complete interval-censored sample. If not, the division of the sample is carried out according to the guidelines, given previously, for right-censored samples.

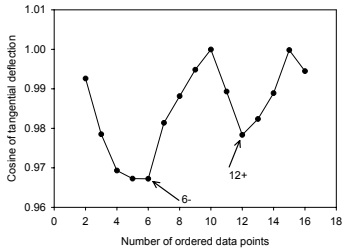
Example A famous complete interval censored sample is the sample considered in 1894 by Pearson (Section 2.2.1). Its normal QQ-plot is shown in Figure 5.20a. The sample has size 1000 with only 19 different adjacent intervals. This results in one m-value (i.e., $m = 3$) for the tangent-rico method. Figures 5.20b and 5.20c give the plots of the derivative and the cosine of the tangential deflection of the QQ-plot. Both plots clearly confirm that a two-component distribution with components separated in location is appropriate for the sample. There is only one one candidate inflection point (corresponding to the 10^{th} ordered interval) and there exist no appropriate couple of candidate nodes. In addition, the global minimum (related to the



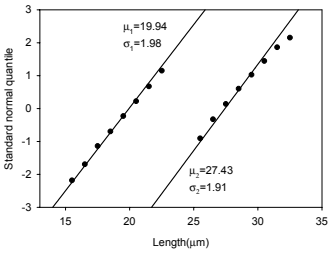
(a) Normal QQ-plot.



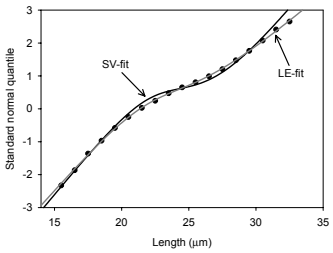
(b) Derivative of the QQ-plot for $m = 3$.



(c) Cosine of the tangential deflection of the QQ-plot for $m = 3$.



(d) Normal QQ-plot of two subsamples based on the 10^{th} interval as inflection point.



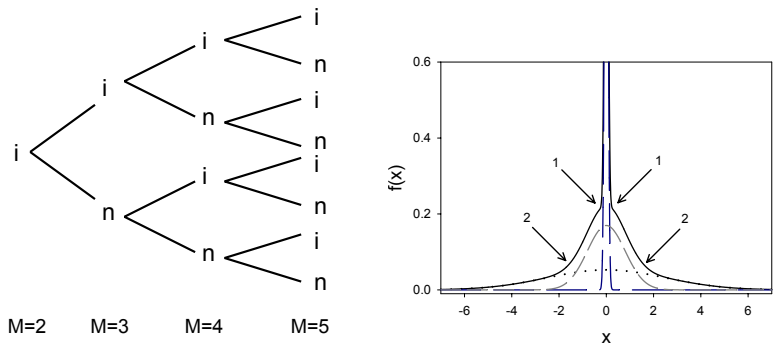
(e) Fit of the best set of starting values (SV-fit) and LE fit.

Figure 5.20: Deriving starting values for the interval censored sample of Pearson.

candidate inflection point) of the derivative plot dominates this plot and it is nicely situated between two minima on the corresponding cosine plot. Figure 5.20d depicts the two subsamples, the fits and values of the corresponding MLEs. The starting value for the proportion parameter equals $\pi_1 = \sum_{i=1}^{9} d_i / (\sum_{i=1}^{9} d_i + \sum_{i=11}^{19} d_i) = 0.73$. The starting values converge to the LE, which is given by $\hat{\mu}_1 = 19.96, \hat{\sigma}_1 = 2.13, \hat{\mu}_2 = 26.16, \hat{\sigma}_2 = 2.74$ and $\hat{\pi}_1 = 0.65$. Clearly, the sample is stable. To summarize, Figure 5.20e shows the small difference between the LE fit and the fit corresponding to the starting values. It illustrates not only the excellent LE fit, but also the good starting values.

5.5.2 More than two components

While a theoretical QQ-plot of a two-component mixture can have two different shapes, there are many more possibilities for the shape of the QQ-plot of a mixture with more than two components. With three components there are at least 6 different forms and this number grows exponentially with the number of components. Nevertheless, without knowing all different shapes, it is possible to extend the tangent-rico method to a general M-component mixture. Namely, most M-component mixtures can be related to a combined series of inflection points and couple of nodes. As such, based on a recursive procedure started from both an inflection point and a couple of nodes, an almost complete extension of the method can be obtained. Figure 5.21a outlines the procedure when started from an inflection point. It depicts the "extension" tree or all possible combinations started from an inflection point. Each inflection point can be followed either through an inflection point or a couple of nodes and similar each couple of nodes can be followed through either an inflection point or a couple of nodes. In addition, there are two ways to combine an inflection point and a couple of nodes and two ways to combine two couples of nodes. The inflection point can be situated between or after a couple of nodes. One couple of nodes can be



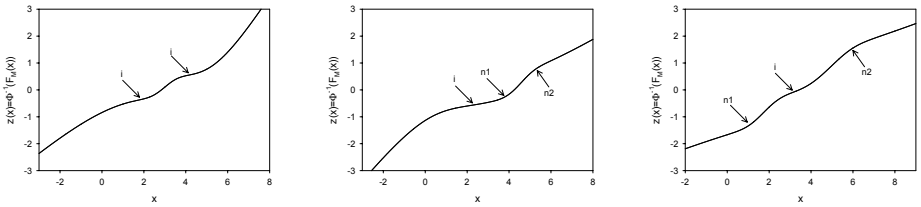
(a) Extension tree started from an inflection point.

(b) A mixture of three components characterized through two couples of nodes with one contained by the other.

Figure 5.21: Extension of the tangent-rico method to more than two components.

situated between another couple of nodes (Figure 5.21b) or after this other couple. As an example, Figure 5.22 gives the QQ-plots for 3 out of the 6 possible combinations for a 3-component mixture. Figure 5.23 shows the corresponding mixture densities. The mixture in Figure 5.22a corresponds to a series of two inflection points. It has three components separated in location. Figures 5.22b and 5.22c show the two possible configurations corresponding to a series of an inflection point and a couple of nodes. In the first, the couple of nodes follows the inflection point, while for the second the inflection point is situated between the two nodes.

Given a sample and a value for the parameter m, then first all candidate inflection points and candidate couples of nodes are searched for and ordered from best candidate to worst. Next, for each combination in the two extension trees, a number of series is built up with the candidate

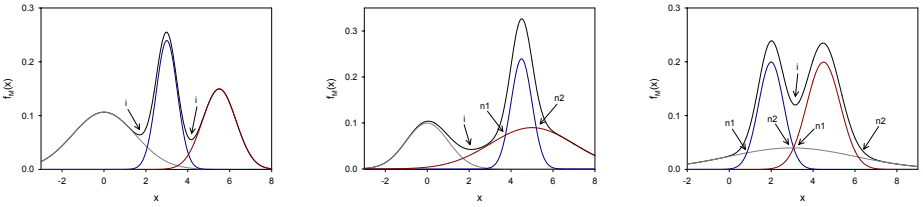


(a) Two inflection points.

(b) An inflection point followed by a couple of nodes.

(c) An inflection between a couple of nodes.

Figure 5.22: Some possible shapes of the QQ-plot for a 3-component normal mixture. The i refers to an inflection point, (n_1, n_2) to a couple of nodes.



(a) Density plot corresponding to the QQ-plot with two inflection points.

(b) Density plot corresponding to the QQ-plot with an inflection point followed by a couple of nodes.

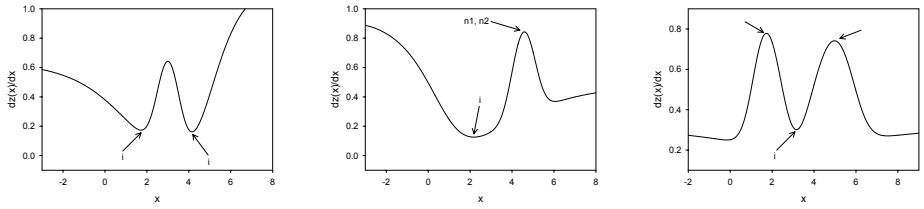
(c) Density plot corresponding to the QQ-plot with an inflection point situated between a couple of nodes.

Figure 5.23: Mixture density and scaled densities (with corresponding proportion parameter) of the mixtures components for the mixture distributions in Figure 5.22.

inflection points and candidate couple of nodes. This number can be varied according to the number of candidates that are used for each position in the combination. Usually, if possible, M candidates are considered. As such, for M=3 up to 6 series of starting values are derived for each combination, resulting in a total of at most 36 sets of starting values for one m-value. Consequently, the number of starting values increases a lot when the number of components increases, but the same holds for the number of maxima of the likelihood function.

In the same way as for a general two-component mixture, the tangent-rico method can be completed to obtain also the largest local maximum in case of small sample sizes (method *III* in Section 5.3.1). Therefore, starting values for distinct spurious maxima have to be constructed. In case of M components, these starting values can be derived by considering one subsample of two successive data points and M-1 other subsamples of the remaining part of the sample. This leads for each couple of successive observations, in contrast to the situation of two components, not to one set of starting values but to a number of sets. For the division of the rest of the sample into M-1 subsamples, the tangent-rico method for M-1 components is used. As a result, it is not possible to guarantee that in any case the largest maximum obtained, is indeed the largest local maximum. Due to the latter and the increasing number of starting values, it is rather difficult to carry out simulations for finite mixtures with more than 3 components.

However, this does not mean that samples cannot be estimated to a general M-component mixture with the likelihood method. By means of the tangent-rico method, a lot of information can be obtained concerning the stability of the sample and the credibility of the largest maximum found. Indeed, if the maxima identified with the starting values of the tangent-rico method, suggest an unstable likelihood surface, then this extension of method *III* can be used to search for maxima with a larger likelihood value. If distinct spurious maxima are found with a larger likelihood value than



(a) Derivative plot of the QQ-plot with two inflection points in Figure 5.22a.

(b) Derivative plot of the QQ-plot with an inflection point followed by a couple of nodes in Figure 5.22b.

(c) Derivative plot of the QQ-plot with an inflection point situated between a couple of nodes in Figure 5.22c.

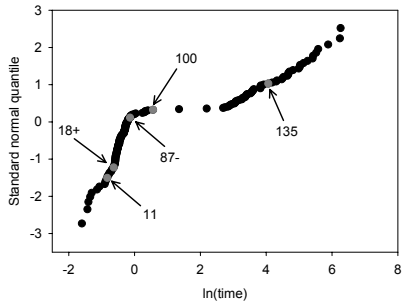
Figure 5.24: Derivative plots of the QQ-plots in Figure 5.22.

maxima obtained with the tangent-rico method or if the largest maximum found is not dominant, then the credibility of the LE, whether it is identified or not, is lost.

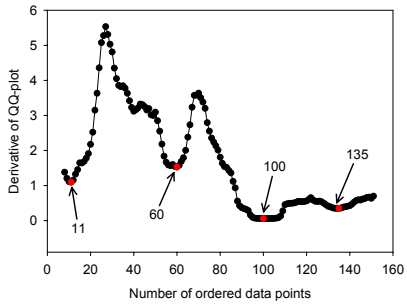
Also here, a useful tool for the exploration of samples, with regard to the possible number of mixture components, are the plots of the derivative and the cosine of the tangential deflection of the QQ-plot. Although they give no clear-cut answer on the possible number of mixture components, they do indicate how many mixture components are relevant and where they can be situated on the QQ-plot. Previously, in Section 5.4.4, we indicated how to interpret these plots when no more than two components were assumed. The case of an unspecified number of components is quite similar. For the derivative plot, usually it can be assumed that it should have one additional pronounced maximum or minimum for each extra component, as this is often the case for its theoretical counterpart. Of course, it has to be taken into account that between any two minima a maximum is situated (and the reverse). The only exception is the case where all components are separated

in scale. There, the derivative plot while have only one sharp maximum, but the cosine plot will reveal several couples of nodes. Further, the cosine plot can be used in the same way as explained earlier on, to confirm the minima and maxima located in the derivative plot. Figure 5.24 gives three examples of derivative plots for a 3-component normal mixture. The derivative plot in Figure 5.24a has two marked minima, in Figure 5.24b one distinct minimum and maximum and in Figure 5.24c two clear maxima.

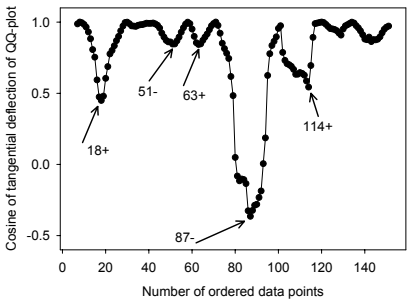
Example The laser B sample is considered (Section 2.1.3). The lognormal QQ-plot of this complete sample of size 158, is given in Figure 5.25a. First, we will discuss the information available from the plots derived from the QQ-plot. Next, we will handle the likelihood estimation of several M-component lognormal mixtures using the tangent-rico method and the extension of method *III*. The m-values used in tangent-rico method are 7, 15 and 31. Plots of the derivative and of the cosine of the tangential deflection of the QQ-plot for $m = 15$ are shown in Figures 5.25b and 5.25c. Basically, the same picture is obtained for other m-values, with more (unimportant) minima for the smaller m-value and a smoother plot for the larger m-value. As observed, the derivative plot has a pronounced global minimum (situated around data point 100) and maximum (between data points 11 and 60). This is line with the picture given by the cosine plot. Namely, it has one obvious candidate couple of nodes (i.e., 18 and 87) related to the maximum and one couple of nodes (i.e., 87 and 114) which is nicely situated around the minimum. Note that the couple (18, 51) is another candidate couple of nodes that can be related to the global maximum of the derivative plot. This suggest that at least a 3-component mixture would be appropriate. Besides these two main features, there is a second maximum (between data points 60 and 100) and minimum (data point 135) that could point to an additional component. This is, however, not likely given that they are not pronounced enough. So, although at first sight the QQ-plot is rather



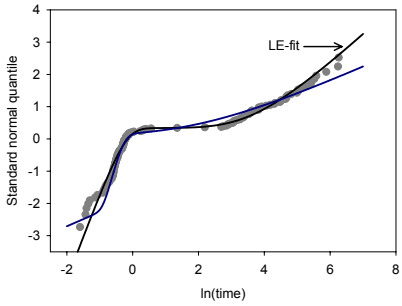
(a) Lognormal QQ-plot with some candidate inflection points and couple of nodes.



(b) Derivative of QQ-plot for $m = 15$.



(c) Cosine of the tangential deflection of QQ-plot for $m = 15$.



(d) LE-fit of a two-component lognormal mixture and fit of the 2^{nd} largest maximum.

Figure 5.25: Analyzing the laser B sample.

steep-flat-steep, indicating the presence of only two components, other plots suggest the presence of at least three components.

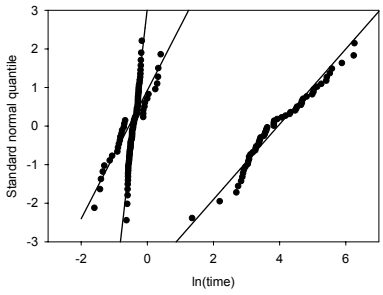
Next, we used the tangent-rico method to fit a two-component lognormal mixture to the sample. The sample seems stable. The LE fit is shown in Figure 5.25d. It has a steep-flat-steep shape with an inflection point situated around data point 100. Also the fit of the second largest maximum found, is given. Its shape is flat-steep-flat with the nodes situated around data points 18 and 87. Its likelihood value (-239.910) is far below the likelihood value of the LE (-228.558). The likelihood estimates are $\hat{\mu}_1 = -0.44, \hat{\sigma}_1 = 0.36, \hat{\mu}_2 = 3.95, \hat{\sigma}_2 = 1.03$ and $\hat{\pi}_1 = 0.63$. Still, in spite of the stability of the sample, the fit of the LE is rather poor at the beginning of the sample.

When fitting a 3-component lognormal mixture to the sample, the stability of the sample seems to be preserved, given the maxima identified with the tangent-rico method. We scanned the likelihood surface for other maxima. Although, it contains a large number of maxima, from which many are plausible, the largest maximum found with the starting values of the tangent-rico method dominates the likelihood function. No distinct spurious maxima are found with a larger likelihood value. The nice fit of the LE is given in Figure 5.26c. A corresponding density plot for the densities of the mixture components is shown in Figure 5.26b. Note that the density functions are rescaled according to the proportion parameter for each component. It can be seen that two components of the mixture are separated in scale (corresponding to a couple of nodes) and that a third component follows which is separated in location with the other two components (corresponding to an inflection point). A possible starting value to identify this maximum corresponds to a subdivision of the sample based on the couple of nodes (18,87) and the inflection point 100. A QQ-plot of the three subsamples with corresponding MLE fits is given in Figure 5.26a. For each of the subsamples, a lognormal distribution seems to fit quite well. The LEs

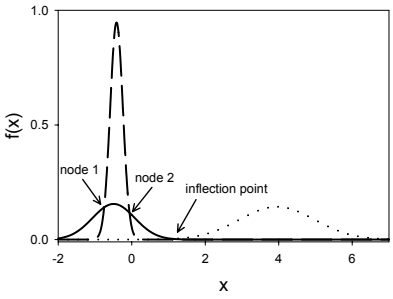
for the parameters are: $\hat{\mu}_1 = -0.49, \hat{\sigma}_1 = 0.57, \hat{\mu}_2 = -0.41, \hat{\sigma}_2 = 0.17, \hat{\mu}_3 = 3.95, \hat{\sigma}_3 = 1.02, \hat{\pi}_1 = 0.22$ and $\hat{\pi}_2 = 0.41$. The likelihood value of the LE is -216.885 (which is considerably larger than the likelihood value of the LE for the two-component mixture). The second largest maximum found was distinct spurious with likelihood value -219.356. The first not distinct spurious maximum identified after the LE, has likelihood value -221.354. It corresponds to a subdivision of the sample according to two inflection points situated around the 11^{th} and the 100^{st} data point.

Subsequently, we fitted a 4-component lognormal mixture to the sample. For this mixture, according to the maxima obtained with the tangent-rico method, the sample was clearly unstable, even highly unstable. With the extension of method *III*, several distinct spurious maxima were found with a likelihood value larger than the largest maximum identified with the tangent-rico method. In addition, several plausible maxima with a small difference in likelihood value were obtained. The fits of the largest maximum identified and of a large plausible maxima found, are shown in Figure 5.26d. The arrow indicates where the fits are deviating from the LE fit of the 3-component lognormal mixture. As observed, the fits are quite similar both to each other and to the LE fit of the 3-component mixture. Likely, the sample is too small to numerically distinguish a 4-component mixture.

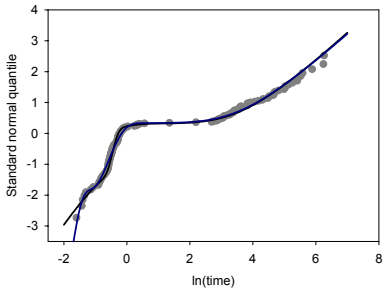
In summary, this sample contains enough information to obtain a reliable LE for a 3-component lognormal mixture. This is suggested by the plots derived of the QQ-plot and confirmed by the stability of the sample with respect to this model. Although 4 or more mixture components could be useful to arrive at a better fit of the sample, these mixtures cannot be used for inference.



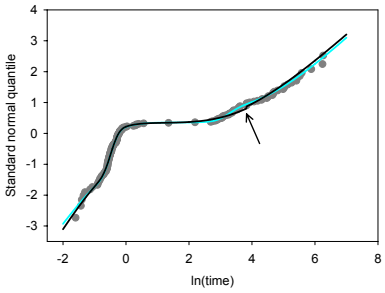
(a) Three subsamples based on a division of the couple of nodes (18, 87) and the inflection point 100.



(b) Scaled density functions of the mixture components of the LE for a 3-component lognormal mixture.



(c) LE-fit of the 3-component lognormal mixture.



(d) Fits of the largest maximum (black) and one large plausible maximum (grey) found for a 4-component lognormal mixture.

Figure 5.26: Analyzing the laser B sample.

5.5.3 A common parameter among the mixture components

Common scale parameter

The tangent-rico method can be simplified a lot when a common scale parameter is assumed among the mixture components. For a two-component mixture, the QQ-plot has always a steep-flat-steep form with only one inflection point (as opposed to the theoretical three inflection points for a general two-component mixture). Moreover, a theoretical QQ-plot of an M-component mixture contain exactly M-1 inflection points, since components can only be separated in location. As such, a plot of the tangential deflection of the QQ-plot is not required to derive starting values. In addition, the number of pronounced minima on the plots of the derivative of the QQ-plot give a clear indication of the number of possible components for the mixture. The calculation of starting values is done in a similar way as for the general finite mixture model, but now with subdivisions of the sample based solely upon candidate inflection points. Note that this procedure can be used for an exponential mixture since the latter is a Weibull mixture with common shape parameter. Further, this method can be considered with a known or unknown common scale parameter.

Common location parameter

Although a mixture with a common location parameter is rarely encountered, it is possible to fit these kind of mixtures. In contrast to a common scale parameter, the shape of the QQ-plot of a two-component mixture is always flat-steep-flat. Only candidate couple of nodes, and thus only plots of the cosine of the tangential deflection, are required to determine starting values. Again the extension to mixtures with more than two components is quite trivial. Namely, mixture components have to be distinguished based on the value of their scale parameter and as such candidate couple of nodes also have to be contained in each other. For example, if (a_1, b_1) and (a_2, b_2)

are the x-coordinates of the two couple of nodes of a 3-component mixture, then the order between these coordinates has to be $a_1 < a_2 < b_2 < b_1$ or the reverse.

Chapter 6

Case studies

6.1 Analyzing a field example: the electromigration sample

The electromigration sample EM2 is an industrial sample (Section 2.1.4). This bimodal failure time sample was passed on with the request for an appropriate bimodal analysis, preferably making use of the maximum likelihood method. For this sample, main interest is in:

- Estimation of the quantile $t_{0.01\%}$ at a temperature of 125 °C.
- Estimation of the electromigration temperature model parameter, i.e., the activation energy E_a, at both modes.
- Estimation of the scale parameter σ at both modes.

Other possible questions could be to assess whether the activation energy and/or scale parameter is equal among the two modes or whether the two modes are present in the sample.

To analyze the sample three methods will be considered. The first is simply to ignore the bimodality and to use the same methods as for

monomodal samples. In spite of the fact that this can lead to wrong (reliability) conclusions or a biased estimate for the activation energy, by experience it is still the preferred way of handling in industry. The monomodal method outlined in the JEDEC standard, JESD63, is a linear (least squares) regression analysis (on the logarithmic transformed failure times). Second, the sample will be split into two subsamples, one for each failure mode. Each of the two subsamples will be analyzed according to monomodal techniques. This is one of the few methods proposed and also used to deal with bimodality. Third, a likelihood analysis using a two-component lognormal mixture model is considered.

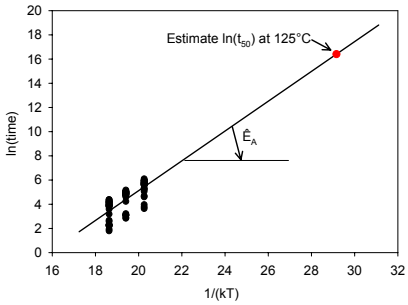
The main question we like to answer is if a bimodal likelihood analysis contributes to the reliability conclusions, i.e., whether it is worth the effort. Moreover, what is the effect of ignoring the failure modes and what are the advantages and disadvantages of simply splitting up the sample into two subsamples compared to the more complicated bimodal likelihood analysis.

6.1.1 A linear regression analysis

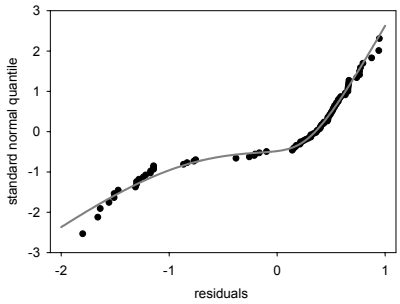
Commonly, for electromigration data, the distribution of the failure times at one stress level is assumed to be lognormal, or the distribution of the logarithm of the failure times to be normal. The linear regression model used, can then be written as:

$$\ln(t_i) = C + E_a x_i + \epsilon_i \qquad \epsilon_i \text{ i.i.d.}, \quad \epsilon_i \sim \mathrm{N}(0, \sigma^2), \quad i = 1...n, \qquad (6.1)$$

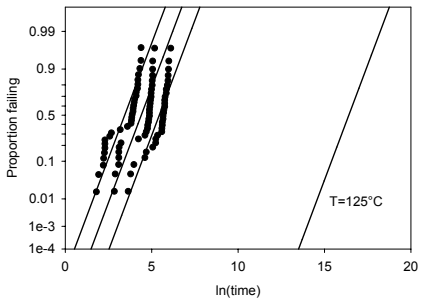
with the covariate $x_i = \frac{1}{k_B \mathrm{T}_i}$ the rescaled temperature, k_B the Boltzmann constant and C, E_a and σ the model parameters. Figure 6.1a shows the estimated least squares regression line on a plot of the covariate x against the logarithm of the failure time $\ln(t)$. Its slope is the estimated activation energy $\widehat{E}_a$. Figure 6.1b gives a QQ-plot of the residuals. Clearly, the assumption of a normal distribution for the error term ϵ_i is violated. This can



(a) Plot of the inverse of the rescaled temperature versus the logarithm of the failure times and corresponding fit.



(b) QQ-Plot of the residuals $t_i - \hat{t}_i$ with LE-fit of a two-component general normal mixture.



(c) QQ-plot of the sample and resulting (MLE) fit of model (6.1), together with the estimated distribution at 125 °C.

Figure 6.1: A lognormal regression model.

also be noticed from Figure 6.1c, which illustrates that a normal distribution gives rise to a poor fit of the failure times. Note that a two-component normal mixture distribution fits the residuals quite well. Consequently, the estimation of low quantiles, under the assumption of a normal distribution, cannot be trusted. Also, even if a common standard deviation (i.e., the scale parameter σ) is assumed among the two failure modes, the estimated standard deviation will be considerably larger than in each of the two modes, i.e., will be biased. Further, if there are indeed two failure modes, likely the estimates of the other model parameters, in particular E_a, will be biased too.

To compare with the methods in the following sections, Table 6.1a gives the estimates and 95% asymptotic confidence intervals for the parameters. Hereby, the maximum likelihood method is used. Although parameter estimates are the same as with least squares regression, estimates of the standard deviation and of the standard errors are slightly different. Through carrying out the ML method, it is easier to compare the results with the likelihood analysis of Section 6.1.3.

6.1.2 Subdividing the sample

Often, if the presence of more than one failure mode is recognized, the part of the sample which is of interest, is selected. In particular, it happens that one of the two failure modes is not considered to be relevant for the intended goals. For example, if the first mode is related to a defect mode, commonly main interest is in the second mode. As such, only estimates of model parameters and quantiles for this second mode are then required. Also, the suggestion has been made several times to only look at the first part of the sample, related to the first failure mode, as main focus is on the estimation of low distribution quantiles.

In these cases, the technique mainly applied is to (visually) split the sample into two subsamples corresponding to each of the two failure modes.

	E_a (eV)	σ	$t_{0.01\%}$ at 325 °C (minutes)	$t_{0.01\%}$ at 125 °C (years)
MLE	1.23	0.782	4.43	1.39
95% CI	[0.986, 1.48]	[0.666, 0.897]	[2.80, 7.02]	[0.122, 15.8]

(a) Maximum likelihood estimation of model (6.1).

Subsample	E_a (eV)	σ	$t_{0.01\%}$ at 325 °C (minutes)	$t_{0.01\%}$ at 125 °C (years)
Sample 1	1.33	0.532	4.15	3.31
	[1.05, 1.61]	[0.395, 0.669]	[2.41, 7.16]	[0.198, 55.4]
Sample 2	1.05	0.156	74.8	4.15
	[0.993, 1.12]	[0.128, 0.184]	[66.9, 83.7]	[2.28, 7.56]

(b) Splitting the electromigration sample into two subsamples: estimation of model (6.1) to both subsamples.

Model	E_a (eV)	σ	$t_{0.01\%}$ at 325 °C (minutes)	$t_{0.01\%}$ at 125 °C (years)
1^{st} comp.	1.36	0.501	4.48	5.09
of M1	[1.05, 1.68]	[0.327, 0.674]	[2.45, 8.19]	[0.230, 113]
2^{nd} comp.	1.06	0.171	69.8	3.98
of M1	[0.985, 1.13]	[0.131, 0.211]	[59.0, 82.6]	[1.90, 8.36]
M1			5.21	4.04
			[2.97, 9.15]	[1.93, 8.47]
M2	1.02	0.353 (σ_1)	7.10	0.296
	[0.942, 1.11]	[0.195, 0.511]	[4.34, 11.6]	[0.120, 0.728]
		0.220 (σ_2)		
		[0.173, 0.266]		
M3	1.06	0.251	9.49	0.53
	[0.971, 1.14]	[0.218, 0.294]	[7.98, 11.24]	[0.226, 1.26]

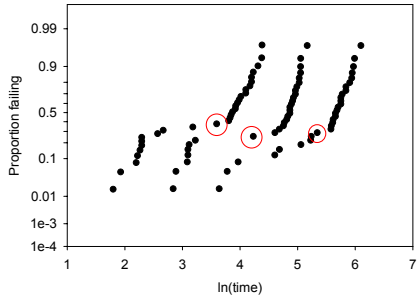
(c) (Maximum) likelihood estimation of model (6.2) and several simplified versions. Model M1: the bimodal regression model (6.2), model M2: (6.2) with common E_a, model M3: (6.2) with common E_a and σ.

Table 6.1: (Maximum) likelihood estimates and 95% asymptotic confidence intervals (CIs) of parameters and some quantiles of several models for the electromigration sample EM2.

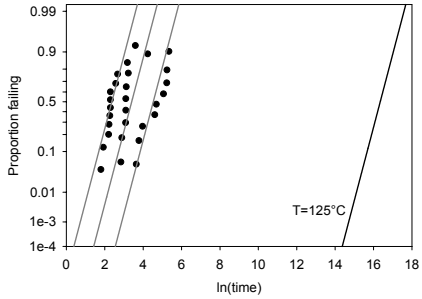
As a result, both subsamples can be analyzed with monomodal techniques and problems seem to be solved.

On the QQ-plot in Figure 6.2a, it is indicated how we subdivided the sample. At each stress level, all data points before the indicated data point and this point itself are allocated to the first subsample, while all data points after are assigned to the second subsample. The lognormal QQ-plots of these two subsamples are given in Figures 6.2b and 6.2c. It is noted that for both subsamples, a lognormal failure time distribution seems to be appropriate at all stress levels, as well as a constant σ over all stress levels. To both samples the linear regression model (6.1) is fitted with the ML method. Estimates and asymptotic confidence intervals for the model parameters and some quantiles are given in Table 6.1b. Note that at both modes, the estimate of the scale parameter is considerably smaller than the estimate obtained in Section 6.1.1 for the whole sample.

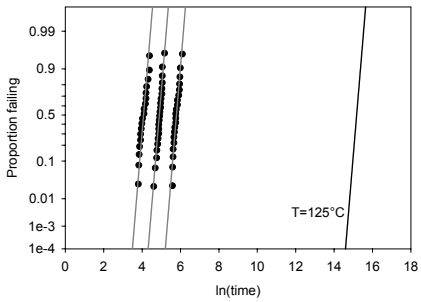
Given that for each device under test the exact failure mode would be known, this method would be fine. However, this is not the case here and the division of the sample will always be subjective. Moreover, through subdividing the sample, it is implicitly assumed in the analysis that there is the knowledge about the exact failure cause of each unit. Although estimates will usually be more efficient than when the uncertainty about the bimodality is taken into account, information is used which does not exist. Consequently, it could lead to wrong results. In addition, for this sample there is the advantage that the split can be carried out relatively easy, but this is not always the case (for example, the clear bimodal resistor sample of Section 2.1.1).



(a) QQ-plot of the sample. The indicated data points are used to split the sample.



(b) QQ-Plot of the first subsample with MLE fits of regression model (6.1).



(c) QQ-plot of the second subsample with MLE fits of regression model (6.1).

Figure 6.2: Analyzing the sample based on a split into two subsamples (division by eye).

6.1.3 A bimodal likelihood analysis

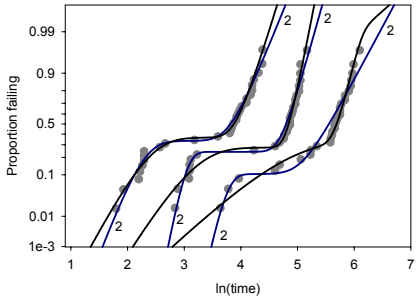
The cumulative distribution function of a "bimodal" regression model, adapted to the electromigration case, is given by:

$$F(t|\boldsymbol{\theta}) = \pi_1 \Phi\left(\frac{\ln(t) - C_1 - E_{a1}x}{\sigma_1}\right) + (1-\pi_1)\Phi\left(\frac{\ln(t) - C_2 - E_{a2}x}{\sigma_2}\right), \tag{6.2}$$

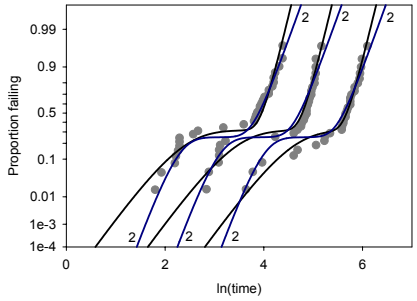
with $x = 1/(k_B T)$. At each stress level, the failure time distribution is a general two-component lognormal mixture. Here, the scale parameters for the two components are assumed to be constant over all stress levels. This model can be simplified either by considering a common scale parameter among the mixture components (i.e., $\sigma_1 = \sigma_2$) or a common parameter for the activation energy (i.e., $E_{a1} = E_{a2}$).

To fit model (6.2) with the likelihood method, starting values to maximize the likelihood function have to be calculated. Therefore, we first analyzed the failure time samples at each stress level separately. For all three stress levels, the sample was unstable (with respect to a two-component lognormal mixture). In each case, two maxima were dominating the likelihood function. Between these two and all other maxima, there was a considerable gap in likelihood value. Figure 6.3a shows for each of the three samples, the fits of these two maxima. At the stress levels 350 °C and 300 °C, the largest local maxima are related to the fits not referred to by 2, while the reverse is true at the other stress level. As noted, the three fits indicated by 2 are similar in shape as well as the other 3 fits. Moreover, they form two groups of 3 similar maxima.

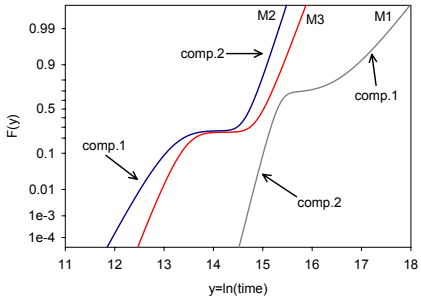
Based on these groups, two sets of starting values are derived. Precisely, for each group, starting values for C_1 (C_2) and E_{a1} (E_{a2}) are obtained from a least squares regression of the 3 estimates for μ_1 (μ_2) on x. For σ_1, σ_2 and π_1 the mean of the 3 parameter values (at the 3 stress levels) is taken. In case of a common scale or common activation energy, the mean of the values at the two modes is considered. For model (6.2) both sets of starting



(a) Fits of the two largest maxima of the likelihood function for a two-component lognormal mixture at each stress level separately.



(b) LE fit of model (6.2) and MLE fit (referred to by 2) of the bimodal regression model with both common scale parameter and activation energy.



(c) Prediction at 125°C of model (6.2) (M1), of the simplified model with common activation energy (M2) and of the model with both common scale parameter and activation energy (M3).

Figure 6.3: A two-component lognormal regression model.

Table 6.2: Comparing different models: values of the log likelihood (lnL) and results of some likelihood ratio tests.

	Monomodal	Bimodal regression model (6.2)			
	model (6.1)	$= \sigma, = E_a$	$= \sigma, \neq E_a$	$\neq \sigma, = E_a$	$\neq \sigma, \neq E_a$
lnL	-103.157	-54.032	-53.437	-52.304	-50.732
$H_0 := \sigma, = E_a$			1.19	3.46	6.60
df, p-value			1, 0.275	1, 0.063	2, 0.0369
$H_0 := \sigma, \neq E_a$					5.41
df, p-value					1, 0.0200
$H_0 :\neq \sigma, = E_a$					3.14
df, p-value					1, 0.0764

Note: the row for H_0 gives the value of the likelihood ratio test statistic, df refers to the corresponding degrees of freedom.

values converge with the NR-method to the same maximum. Further, there was no evidence that the likelihood function contained more than one local maximum (apart from boundary solutions). We also fitted the simplified models with common scale parameter, common activation energy and both parameters common. Figure 6.3b shows the nice LE fit of model (6.2). But also the MLE fit of the model with both a common scale parameter and activation energy, is quite good. The values of the logarithm of the likelihood of these 4 models and of model (6.1) are given in Table 6.2. In addition, the values of some likelihood ratio test statistics are supplemented.

From this, two important observations can be made. First, there is a huge difference in likelihood value between the "monomodal" regression model and any of the bimodal regression models given. Again, this shows that a lognormal distribution is not suitable as failure time distribution at each stress level. Moreover, as illustrated also in Figure 6.3a, a two-component lognormal mixture captures the variability within the failure times quite well. Second, the difference among the several bimodal regres-

sion models is rather small. There is only weak evidence for a different scale parameter between the two mixture components. Moreover, a common activation energy seems appropriate. Note that no further information was received from the company concerning the feasibility of each of those 4 bimodal models. The choice for one of these also has to rely on physical grounds. For example, it can be noted from Figure 6.3c and table 6.1c, that for model (6.2), the two components have changed from order at the stress level T=125 °C (due to the fact that the acceleration factor, i.e., the activation energy, is larger for the first mode than for the second). The consequences with respect to the conclusions drawn for the reliability of the devices under study, are quite large when considering this model compared to the more simplified models. Namely, with model (6.2), the estimated life of the devices of the population is (much) longer (Figure 6.3c).

6.1.4 Discussion

If it is a priori known that there are two (or more) failure modes, there is no reason to not use a so-called bimodal linear regression model. Even if interest is only at one of the two failure modes, the general two-component regression model can be used to obtain estimates and estimated standard errors for the parameters at the different modes (for example, model 6.2). The only difference with the method of Section 6.1.2 is a (slight) loss in efficiency. This can be noted, for example, from the length of the confidence intervals given in Table 6.1b and in Table 6.1c for the parameters of two mixture components of model (6.2). Still, as long as the exact failure cause for the devices is not known, these smaller estimated standard errors should be considered with care. Moreover, the one obtained with the general bimodal regression model do take into account the uncertainty about the failure cause and as such should be preferred. In addition, with this bimodal regression model, several tests can be carried out quite easily and more simplified models can be considered.

The use of the lognormal linear regression model when the presence of two or more failure modes is clear, should be avoided. Not only the estimate of the scale parameter will be biased, but also the estimate of the activation energy is not in line with the results obtained from other models, using a common activation energy at both modes (Table 6.1a versus model M2 and M3 in Table 6.1c). In particular, when a common activation energy is assumed in model (6.2), not only its estimate is considerably smaller but also the length of the obtained confidence interval. To our opinion, there is no reason to not take into account the presence of more than one failure mode. Moreover, when good starting values are available, obtaining a likelihood estimate for bimodal regression models, is a performable task.

6.2 On the number of mixture components for the galaxy sample

As mentioned in Section 2.2.2, the main issue for the galaxy sample is on the modality of the density function of its distribution, i.e., is it multimodal or rather unimodal. In literature, this sample has received already a lot of attention. In general, its distribution is assumed to be a finite normal mixture. As such, the question on the modality of the distribution was turned into a discussion on the number of mixture components. Note, however, that these two problems are not totally equivalent, as a unimodal distribution can be a finite mixture.

Most analyses carried out are of a Bayesian type. Depending on the model used, and in particular the choice of the prior distributions, the resulting answer on the number of components ranges from 3 up to 9. No general agreement is obtained. A brief overview of the Bayesian analyses carried out, is given in Aitkin (2001). Only a few authors consider a (maximum) likelihood analysis. In comparing the results from the Bayesian analysis with a likelihood analysis, Aitkin fits a series (with increasing number of components) of finite normal mixtures both with a common and a non-common scale parameter. Based on results of a bootstrap likelihood ratio test, Aitkin concludes that there is a convincing evidence for three components in case of a normal mixture with common scale parameter or for four components in case of a general normal mixture. Also, there is no convincing evidence for more than these numbers. Hereby, results for the full bootstrap analysis to assess the number of components for the general finite normal mixture model are taken from McLachlan and Peel (2000, pp. 194-196). However, based on the same bootstrap p-values, McLachlan and Peel suggest at a 5% level of significance, 6 components.

In none of the likelihood analyses carried out, the presence of spurious maxima was mentioned. Neither, the surface of the likelihood function

was discussed. In addition, we find a number of 6 components for the general normal mixture model, with a sample size of 82 rather questionable. This can only hold if the 6 components would be very good separated and clearly distinguishable on the QQ-plot. The latter is not the case, as only 3 groups are visible (Figure 2.7). Furthermore, although a value for the measurement error (δ=0.05) is given, it is never accounted for. Apart from this, Aitkin (2001) states that there is a small overstatement in the likelihood when using the density representation in case of a small value of one of the scale parameters (as opposed to the "correct" cdf representation). Still, his likelihood estimates are derived from the density representation and not from the cdf representation.

The objective of the following study is threefold:

- To demonstrate how the methodology introduced in the previous chapters, contributes to the discussion of the number of possible mixture components. In particular, how the analysis of the surface of the likelihood can reveal the maximum amount of information available in the sample to model a certain number of mixture components, without a formal testing procedure.

- To illustrate that results on maxima and likelihood ratio tests should be considered with caution.

- To show the difference on the final result when taken into account the measurement error δ.

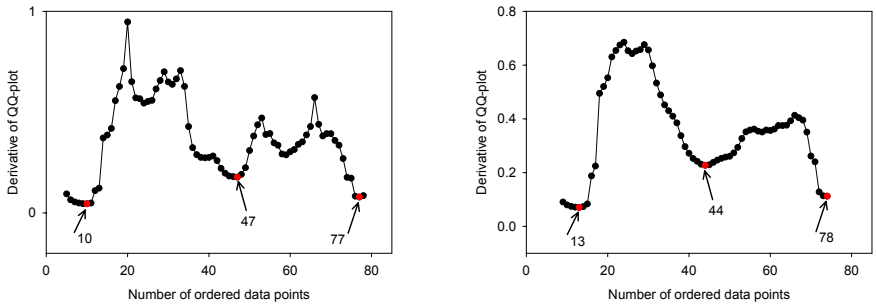
Note that a lognormal distribution for the component density could be more appropriate for the data, given the nonnegative nature of the velocities. Still, we will not consider this, as we want to compare with results obtained in literature.

6.2.1 Likelihood analysis

Interpretation of the derivative and cosine plots

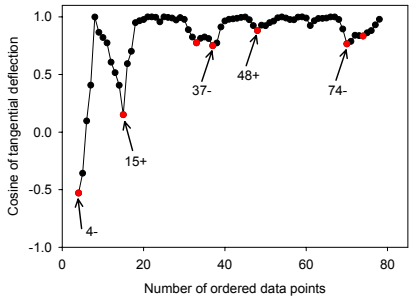
The plots of the derivative and of the cosine of the tangential deflection for m equal to 9 and 17 are given in Figure 6.4. No additional features are observed in plots for smaller values of m. At least two pronounced minima, one situated at the beginning (around data point 10) and one at the end (around 78), are noticed on the derivative plot. For $m = 9$, this first minimum can be related to the couple of nodes (4, 15) at the cosine plot, while for the other minimum only the first node (around data point 74) is visible. These minima correspond to the two outlying groups of data points situated at the beginning and end of the QQ-plot. Either these two groups belong to two different components or to one component which contains the other middle group. For the former the two minima correspond to two candidate inflection points, for the latter these minima can be related to a candidate couple of nodes at the cosine plot (situated around data points 15 and 74). The related maximum on the cosine plot for this couple seems to be divided in two. This could point to an additional (3^{th} or 4^{th}) component situated in the middle of the sample. Either this can be related to the global maximum (corresponding to a candidate couple of nodes situated around the data points 15 and 37) or to a minimum situated around data point 44. Nevertheless, although the middle part of the QQ-plot deviates from a straight line, it is not clear whether this will be enough to be recognized as not random.

In summary, there is a suggestion for at least three components and at most four in case of a general finite mixture model. In case of a common scale parameter, the same proposal can be made for the number of components. Although at the derivative plot, certainly for $m = 9$, there are more than 3 minima, the other minima are no clear indication for an additional component.

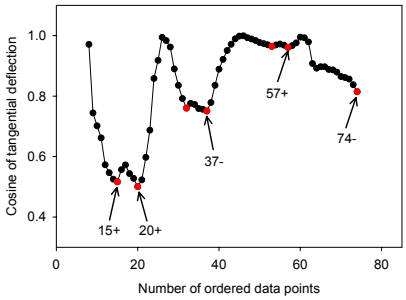


(a) $m = 9$.

(b) $m = 17$.



(c) $m = 9$.



(d) $m = 17$.

Figure 6.4: Plots of the derivative and of the cosine of the tangential deflection of the QQ-plot of the galaxy sample.

Analysis of the likelihood

Non-common scale parameter We scanned the likelihood surface in case of an M-component general normal mixture model, started with one component and stopped at 6 components. Apart from the single normal mixture model, for each number of components the tangent-rico method is first used to obtain an indication of the stability of the sample. In addition, with the extension of method *III*, we also derived specific starting values to search for distinct spurious maxima. Both the EM-algorithm and the NR-method were used as iterative procedure. Although with the same set of starting values, not exactly the same set of maxima was obtained, all different maxima found with one algorithm, were also derived with the other procedure.

Table 6.3 summarizes the main results. Not the maxima itself, but their likelihood values are given to look at the stability of the sample. The numbered maxima are the largest one found (in order). Note that up to three components, we are confident to have found the largest local maxima, i.e., the LE, and the other largest maxima. For more than three components, distinct spurious maxima are found at the top of the likelihood function. Although not likely, it is possible that there exist other distinct spurious maxima with a larger likelihood value. Some comments on the table are provided.

- For M=2, there are two maxima at the top of the likelihood. There is a large gap (in likelihood value) with the following maxima. All maxima shown are found with the starting values of the tangent-rico method. This sample is unstable. Although here the reason is not the sample size, but an unsuitable model as illustrated in Figure 6.5a. The latter shows the poor fits of the LE and the second largest maximum. It can be seen that the LE is related to an inflection point situated around data point 10 (i.e., the best candidate inflection point) and maximum 2 to a couple of nodes situated around the data points 15 and 74

Table 6.3: Value of the log likelihood function (lnL) for several maxima in case of a general M-component normal mixture, with M ranging from 1 to 6.

	Number of components M					
Maxima	1	2	3	4	5	6
(M)LE	**-240.417**	-220.195	**-203.485**	*-196.433*	*-189.951*	***-182.580***
2	-	**-220.362**	-209.837	*-196.856*	*-190.278*	*-184.796*
3	-	-229.071	-212.143	-197.137	*-190.455*	*-185.450*
4	-	-231.733	-212.257	**-197.720**	*-190.875*	*-186.578*
5	-		*-212.810*			
A	-	-236.985		-199.296	-192.235	
B	-				**-196.137**	

Note: Maxima in italic are distinct spurious, maxima in bold are those obtained by Aitkin (2001).

(i.e., one of the two best options for a couple of nodes). The other maxima found can be related to other candidates inflection points or couple of nodes (for example, maximum 3 corresponds to the couple of nodes (15,37) and maximum A to the inflection point 78). Note that between maxima 4 and maximum A several distinct spurious maxima are located.

- For M=3, 78 starting values are derived with the tangent-rico method, resulting in 9 different maxima. The likelihood values of the four largest maxima found, are given in the table. This outcome suggest a rather stable sample. Moreover, with the extension of method III, no larger maximum than these first four maxima, is identified. The largest distinct spurious maximum found, is situated at place 5. Figure 6.5b shows the LE fit. As noted, this maximum is related to inflection points around data points 10 and 78. Figure 6.5c shows the fits of maximum 2 and 3. Maximum 2 corresponds to the inflection point 10 and the couple of nodes (15, 37), while maximum 3 is related to the

couple of nodes (15, 70) and the inflection point situated around data point 44. While the LE captures most features of the sample, apart from perhaps the middle part, maximum 2 and 3 clearly miss either the beginning or the end.

- For M=4, with the 311 starting values of the tangent-rico method, more than 30 maxima are identified. No dominating maxima is found (for example, maximum 3, 4 and A are among the largest maxima identified). This clearly indicates that the sample is unstable. Through applying method III, two distinct spurious maxima with a larger likelihood value are located. The four largest maxima found, all add a fourth component to the LE of the 3-component mixture. In particular, a similar component is related to the beginning and the end of the sample, but they divide the middle part differently. Figure 6.5d shows the QQ-plot restricted to the middle part of the sample and the fits of the four largest maxima. The arrows point to the deviation of the sample related to each maximum. As noted, the difference between all four fits is rather minimal and they are all close to the fit of the 3-component mixture. It illustrates that none of the deviations in the middle part of the sample (in particular, the one corresponding to maximum 4) are pronounced enough to be considered as not at random.

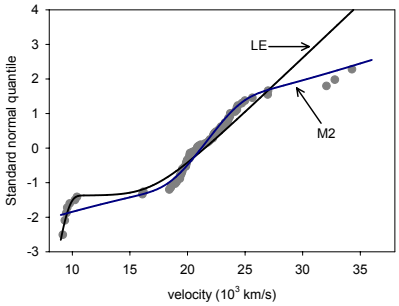
- For M=5 and M=6, respectively 698 and 880 starting values, are derived. While many maxima are found, no maximum is identified which dominates the likelihood. On the contrary, the sample appears to be highly unstable with respect to a 5-component or 6-component mixture model. The largest maximum found at M=6 corresponds to a mixture with two of the components related to two groups of each two isolated data points, as indicated in Figure 6.6. Therefore, some consider this maximum as not being distinct spurious, we does. Not only the fit of

the LE for the 3-component mixture shows that these two groups are not outlying with respect to this model, but also the two additional components has no other goal than exactly fitting these two groups of two points.

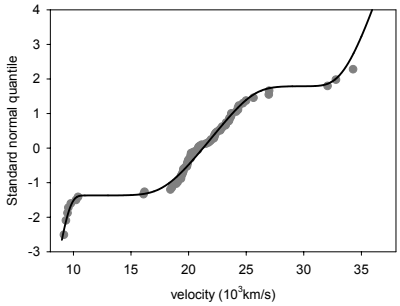
- Maxima indicated in bold are the estimates considered by Aitkin (2001). In 3 out of the 6 cases, among which the case M=2, not the largest maximum is given. Clearly, this influence his results. As he states that the LE for M=6 is not useful (due to its spurious nature), what would have been his decision when the true largest maximum was considered for M=4 and M=5? Importantly, these incorrect estimates also have an effect on the results of a bootstrap likelihood ratio test.

Given the surface of the likelihood at M=4, and in particular 5 and 6, the known results of the bootstrapped likelihood ratio p-values (to test M against M+1 components) should be questioned. Based on which maximum of the likelihood for the M-component mixture, the bootstrap is carried out? In addition, no information is given on which maxima are chosen (in the bootstrap procedure) as estimates. Are they chosen in a consequent way, i.e., the largest local maximum or the maximum obtained with a consistent estimate as starting value? Moreover, when the largest local maximum is searched for in a bootstrap likelihood ratio test, to test M=3 against M=4, it is unlikely that a small p-value will be obtained. Namely, for the bootstrapped samples (based on the LE for the 3-component mixture model), most largest local maxima found for the 4-component mixture will be distinct spurious. As such, quite likely the value of the likelihood ratio statistic for the LE found at M=4, will belong to the middle group.

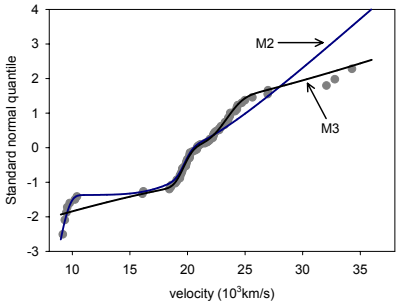
Common scale parameter In general, the normal mixture model with common scale parameter is much easier to handle than the mixture model with non-common scale parameter. The likelihood contains only a few maxima, irrespective of the number of mixture components. As a result, it is



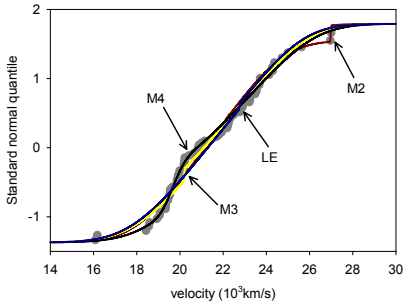
(a) Fits of the LE and maximum 2 (M2) for the two-component mixture.



(b) Fit of the LE for the 3-component mixture.



(c) Fits of maximum 2 (M2) and 3 (M3) for the 3-component mixture.



(d) Fits of the four largest maxima for the 4-component mixture, with focus on the middle part of the sample.

Figure 6.5: Fits of several general M-component normal mixtures for the galaxy sample.

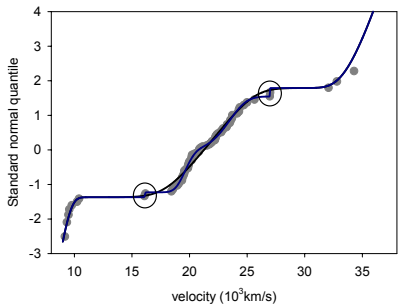


Figure 6.6: Fit of the LE for the 6-component mixture compared to the LE fit of the 3-component mixture.

more feasible to carry out reliable bootstrap likelihood ratio tests. Still, finding the global maximum is not always straightforward and it does happen that a local maximum is taken as the global one.

We used the simplified tangent-rico method to derive starting values. Mixture models up to 6 components are fitted. Table 6.4 give the likelihood values of the largest maxima found. Some observations can be made. First, in all cases, at most two maxima and a number of boundary solutions are identified. Again the MLEs given by Aitkin do not always correspond to the global maximum. This has some consequences for some of its bootstrap likelihood ratio tests (to assess the number of components) carried out. In particular,

- For M=2 no MLE was found. As a result, a bootstrap likelihood ratio test (to test one component against two or two against three) could not be carried out. In spite of the fact that the likelihood function for M=2 has a global maximum.

- For M=5, a local and not the global maximum was found. Consequently, not only the p-value to test 4 against 5 components will be incorrect, but also the bootstrap likelihood procedure to test 5 against

Table 6.4: Value of the log likelihood (lnL) for several maxima in case of an M-component normal mixture with a common scale parameter.

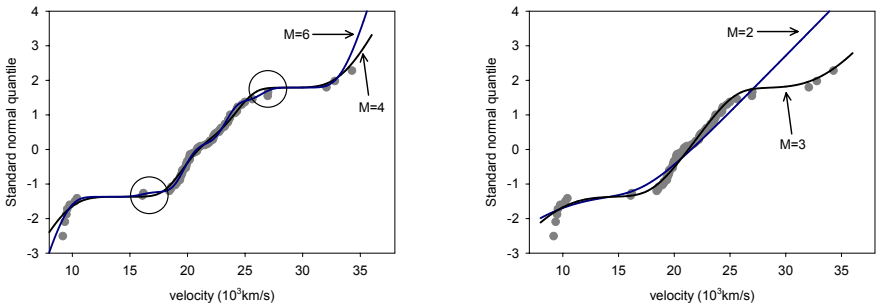
	Number of components M					
Maxima	1	2	3	4	5	6
MLE	**-240.417**	-230.500	**-212.683**	**-208.249**	-205.346	**-197.295**
2	-	-239.371	-	-	**-205.427**	-
A	-	**-240.417**	-230.500	-212.683	-208.249	-205.346

Note: Maxima in bold are those obtained by Aitkin (2001). Maxima A are boundary solutions.

6 components, since simulated samples will be based on this incorrect MLE. For this case, we only expect minor differences when using the correct MLE due to the similar nature of the local and global maximum.

Second, to infer the stability of the sample mainly the difference in likelihood value between the MLE and the MLE of the mixture with one component less has to be looked for. Although, one should be careful with interpretations. For example, for the 6-component mixture the likelihood function has only one maximum, with a likelihood value that is considerably larger than the likelihood value of the global maximum at M=5, i.e., the sample seems to be stable. Still, compared to the MLE at M=4, the two additional components of the MLE at M=6 essentially only fit the two groups of two isolated data points better (Figure 6.7a). Sometimes, as for example at M=5, the instability of the sample is clear from the few maxima at the top of the likelihood.

Third, the difference in likelihood value between the MLE of the two-component and the three-component mixture suggest that a third component is necessary (see also Figure 6.7b). This is confirmed by the result of the likelihood ratio test, to test 1 against 3 components, given by Aitkin (2001). For M larger than 3, mostly the global maximum of the likelihood



(a) MLE fits of the 4 and 6-component mixture model.

(b) MLE fits of the 2 and 3-component mixture model

Figure 6.7: MLE fits of the mixture model with common scale parameter for the galaxy sample.

function does not really dominate this likelihood. Moreover, the additional components only fit some specific features of the middle part of the sample better. As a result, with regard to inference, quite likely no more than 3 components will be accepted. This is in line with the bootstrap likelihood ratio test results of Aitkin. Note that for smoothing applications more than 3 components could be useful.

6.2.2 An adapted likelihood analysis

All observations are rounded to the given value of $\delta = 0.05$. The adapted likelihood estimator looked at is MLEδ. The interval-censored sample consists of 60 different intervals. Its normal QQ-plot is shown in Figure 6.9a. For both the mixture model with common and non-common scale parameter, Table 6.5 give the smallest value among all scale parameters for some of the maxima of the (density) likelihood function. From this, it is possible to deduce the number of mixture components at which differences can be expected between the likelihood analysis and the adapted one. Namely,

in case of an non-common scale parameter, up to M=3 the smallest scale parameter encountered for the largest 3 maxima of the density likelihood, is sufficiently larger than 0.05. As such, conclusions drawn from the likelihood and the adapted likelihood analysis are likely to be similar. On the contrary, from M=4 on, distinct spurious maxima are at the top of the density likelihood function. This gives rise to a value for one of the scale parameters of the largest maximum which is considerably smaller than 0.05. Consequently, different maxima will be at the top of the likelihood and the adapted likelihood function. For a common scale parameter, however, no real differences are expected between both analyses, since up to M = 6 (and even more) the MLE of the scale parameter is large enough compared to 0.05.

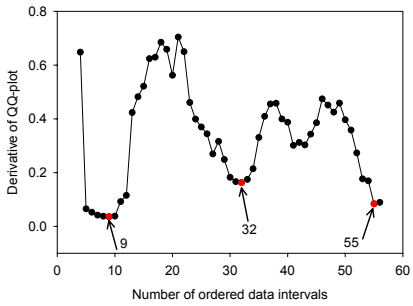
To check these findings, an adapted likelihood analysis, similar to the likelihood analysis, is carried out. In the following, results are briefly described.

Interpretation of the derivative and cosine plot

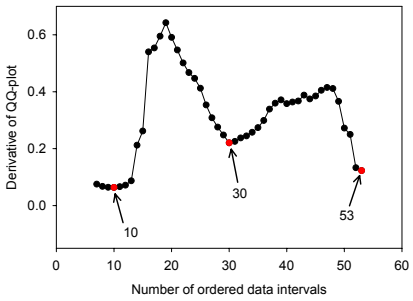
The m-values used in the tangent-rico method are 3, 7 and 13. Derivative and cosine plots for the two largest values are shown in Figure 6.8. It is observed that they resemble those obtained from the uncensored sample (Figure 6.4). As a result, similar conclusions can be drawn.

Analysis of the adapted likelihood

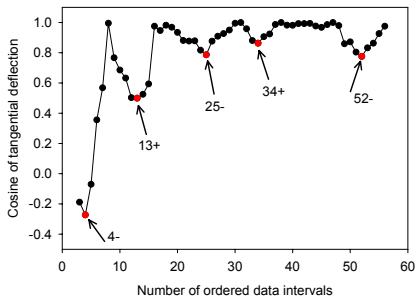
Non-common scale parameter Table 6.6 is the equivalent of Table 6.3 for the adapted likelihood analysis. Except here, maxima are tabulated according to their equivalent in the density likelihood (if found). So, likelihood values are not necessarily placed in an increasing order. Up to 3 mixture components, no problems are encountered. The largest local maxima are easily located. The three largest maxima identified can be compared to those found with the classical likelihood function. Although the difference in likelihood value between two of these maxima can be slightly different in



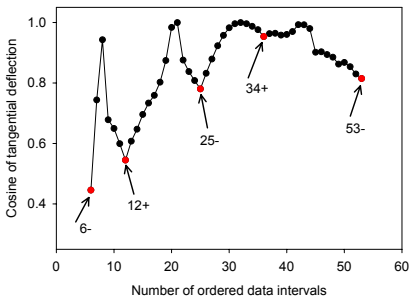
(a) $m = 7$.



(b) $m = 13$.



(c) $m = 7$.



(d) $m = 13$.

Figure 6.8: Plots of the derivative and of the cosine of the tangential deflection of the normal QQ-plot of the interval-censored galaxy sample.

Table 6.5: Smallest value among all scale parameters of a mixture model, for some maxima of Tables 6.3 and 6.4.

	Number of components M					
Maxima	1	2	3	4	5	6
(M)LE	4.54	0.422	0.423	*0.000500*	*0.000500*	*0.0175*
2	-	1.88	0.422	*0.0175*	*0.000500*	*0.000500*
3	-	0.560	0.644	0.0201	*0.0430*	*0.0201*
4	-	0.0203	0.0202	0.438	*0.0175*	*0.00200*
5	-		*0.000500*			
A	-	0.921		0.421	0.0201	
B	-				0.422	

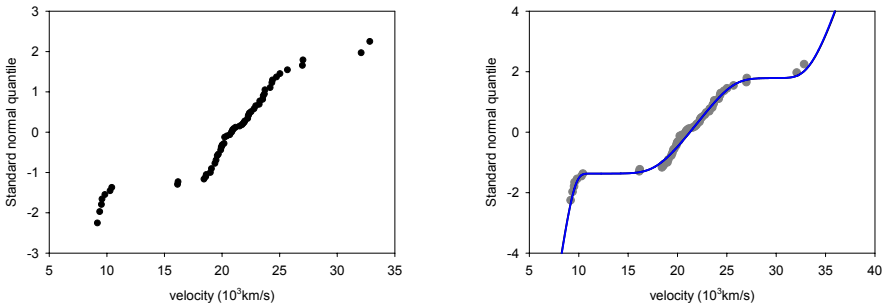
(a) Non-common scale parameter.

	Number of components M.					
Maxima	1	2	3	4	5	6
MLE	4.54	3.03	2.08	1.32	1.10	0.81
2	-	4.02	-	-	1.07	-

(b) Common scale parameter.

both situations, findings are the same. For M=2, the sample is unstable due to an unsuitable model, while for M=3 the sample is stable. Figure 6.9b illustrates that there is an negligible difference between the likelihood and the adapted likelihood estimate for the 3-component mixture model.

With more than 3 components, finding the adapted likelihood estimate for the M-component mixture is not obvious. Although some distinct spurious maxima present in the likelihood function are disappeared in the adapted likelihood function (for example, those maxima with one of the scale parameters equal to 0.0005), the surface of the latter is still highly unstable. It has a lot of maxima attained outside the parameter space and many other maxima with plausible parameter values. The introduction of δ does not



(a) QQ-plot of the interval-censored galaxy sample.

(b) LE and adapted LE fit of the 3-component mixture model.

Figure 6.9: Interval-censored galaxy sample.

lead to a reliable adapted likelihood estimate for any of the mixture models with more than 3 components. Moreover, for M=6 we did not identify a largest local maximum. Note that maximizing the adapted likelihood function with M larger than 3, is quite complicated compared to the classical case.

In summary, this adapted likelihood analysis confirms the results obtained with the likelihood analysis. The sample is too small to clearly recognize a fourth component.

Common scale parameter Practically the same results as with the likelihood analysis are obtained. Until at least 6 components, the surface of the adapted likelihood function looks like the surface of the classical likelihood function. To illustrate this, the differences in logarithm of likelihood value between the MLEs at M and M+1 components is tabulated in Table 6.7 for both analyses. As noted, differences between both situations are minor. Further, no more or other maxima are found and adapted MLEs are equal

Table 6.6: Value of the adapted log likelihood function for the largest maxima found in case of an M-component general normal mixture.

	Number of components M					
Maxima	1	2	3	4	5	6
(M)LE	**-486.070**	**-465.989**	**-449.313**	-	-	*-428.000*
2	-	-466.037	-455.624	*-443.337*	-	-
3	-	-474.738	-457.879	*-442.403*	*-435.159*	*-429.496*
4	-	*-476.859*	*-457.515*	-443.457	-	*-434.320*
5	-		*-457.882*			
A	-	-482.663		-445.157	*-437.290*	
					-440.477	
B	-				-442.016	

Note: Values in bold correspond to the adapted likelihood estimate. Values in italic are related to maxima attained outside the parameter space. The numbers in the column "maxima" refer to the maxima in Table 6.3

Table 6.7: Difference in log likelihood value between the (adapted) MLEs for the M-component and (M+1)-component mixture model with common scale parameter.

	Combinations of mixture components				
Method	1-2	2-3	3-4	4-5	5-6
Likelihood	9.917	17.817	4.434	2.903	8.051
Adapted likelihood	9.879	17.815	4.399	2.876	7.932

to the MLEs for at least 2 significant numbers. As a result, conclusions concerning the dominance of the global maximum are similar.

Also the results of the likelihood ratio test to assess the equality of the scale parameters (given a number of mixture components) are comparable up to 3 components. Namely, for the two-component mixture, the value of the likelihood ratio test statistic for the likelihood analysis is 20.61 as opposed to 20.34 for the adapted analysis. For the 3-component mixture this becomes 18.40 against 18.13. In case of more than 3 components results

are not comparable anymore. However, in this situation carrying out this kind of test is rather useless due to the nature of the surface of both the classical and adapted likelihood of the general finite mixture model.

6.2.3 Conclusions

Based on an analysis of the likelihood surface, it can be concluded that no more than 3 components are appropriate when considering a general normal mixture model. The sample contains not enough information to obtain a useful likelihood estimate for a 4-component general normal mixture model. Although derivative and cosine plots weakly suggest a fourth component, the sample size is too small to recognize it as not at random.

This is in contrast to results obtained in the literature. Both Aitkin (2001) and McLachlan and Peel (2000) state that at least four components could be used.

When the general normal mixture model is estimated, information concerning the likelihood surface should be given. If only likelihood estimates are given, they should be considered with caution. Likewise, results of bootstrap likelihood ratio tests should be handled carefully. We belief that a likelihood ratio test to assess M against M+1 components, should not be carried out if the sample is highly unstable with respect to an M-component mixture.

The results of the adapted likelihood analysis confirm the conclusions of the likelihood analysis. Findings could be derived directly through considering the magnitude of the scale parameters of the likelihood estimates only.

Chapter 7

Conclusions

The basic aim of this project was to arrive at a sensible and workable approach for the maximum likelihood estimation of general finite mixtures. At the end, this work can be looked at as a kind of "guidebook" for the (maximum) likelihood estimation of (general) finite mixtures with a (log)location-scale distribution as component density. We believe that it provides the theoretical background as well as the practical tools to carry out a principled likelihood based analysis of "multimodal" samples. Two of its main applications are illustrated by the case studies in Chapter 6.

While for many basic distributions the application of the ML method is straightforward with the commonly accepted global maximum as MLE, the use of the classical ML method for the general finite mixture distribution is generally assumed to be problematic. In Chapter 4, we have given one possible framework, which allows to handle this estimation problem. We investigated and compared some existing techniques and solutions dealing both with the nonexistent classical MLE and with the choice of one proper maximum out of the many of the classical likelihood function. We have not only explained that most standard methods which try to solve the problem of the nonexistent MLE, either do not solve the problem of the non-existence of a consistent global maximum or are not an option, but also that the al-

ternative likelihood method can be considered as a natural extension of the classical ML method.

Moreover, we believe to have shown that there is no reason to not use the likelihood method for the estimation of general finite mixtures as long as one keeps in mind some basic rules. First, instead of concentrating only on one specific maximum, the nature of the likelihood surface has to be considered too. Specifically, if the largest local maximum is not dominating the likelihood function, inference results based on *any* maximum should be questioned or even not used at all. On the contrary, if the largest local maximum dominates the likelihood function, principled inference results can be obtained from the likelihood estimate in the same way as from a classical MLE. Second, if distinct spurious maxima are at the top of the likelihood function, there are problems with the numerical identifiability of the estimation problem. Reliable or meaningful estimates cannot be obtained from maxima of the likelihood function. Nevertheless, these spurious maxima make it possible to recognize too small sample sizes or a mixture model with too many components. Their appearance at the top of the likelihood reveal that the asymptotic properties of the LE cannot be guaranteed at this specific sample size.

The incorporation of a measurement error will not solve the problems encountered with the classical ML method. Moreover, not only these adapted likelihood methods deal with the same kind of problems, but also the same kind of strategy as for a likelihood analysis has to be followed for an adapted likelihood analysis. Based on the likelihood estimates of the scale parameters, in case of a location-scale distribution as component density, and the possible values for the measurement error, it can be derived whether an adapted likelihood analysis could lead to different conclusions.

We proposed some tools that allow an exploration of the samples with regard to the possible number of mixture components as well as the automatic fitting of general finite mixtures. The tangent-rico method has

an excellent performance as a starting value method, in particular in combination with the EM-algorithm. For two-component general finite mixtures, it has been shown that this method works at least as good as using the true values as starting values. In addition, it can be supplemented with a feasible method that, regardless of the sample size, give almost always rise to the largest local maximum. For mixtures with a small number of components, it has been made possible to carry out reliable simulations and bootstrap procedures. But most importantly, irrespective of the number of mixture components, the maxima identified with the starting values of the tangent-rico method, give in general a good indication of the stability of the sample.

Unless the mixture components are very poorly separated or censoring is involved, the EM-algorithm in combination with the tangent-rico method only required a limited amount of time. Although, its performance was slower for smallest extreme value or Weibull mixtures compared to (log)normal mixtures, certainly for two or three-component mixtures it is still workable in practice. However, in case of moderate or heavily censored samples, the slowness of the EM-algorithm is a potential problem. When using the NR-method, often the same maxima as with the EM-algorithm were obtained. Still more research is required to find out whether it would be a good alternative. As a final note, it should be realized that poorly separated mixtures, whether the scale parameter is common or not, will always require a huge sample size to obtain reliable (maximum) likelihood estimates. Their estimation will always be problematic, no matter what iterative procedure is used.

Apart from the NR-method, the methods used in this work are implemented by the authors. Quite likely their performance can be improved by a better implementation. However, the methods, as proposed here, are a good basis for the set up of a complete software package, which easily and in sound way, handles the likelihood estimation of general finite mixtures.

This work should not be regarded as an end-point. Many questions remain open and many extensions are still worth the effort to be investigated. One topic which deserves more attention is the relation between the stability of the sample and the need for a formal testing procedure for the specific number of mixture components. In particular, whether such a test is still useful when distinct spurious maxima are found at the top of the likelihood function. Further, there is the possible extension of this work, and in particular the tangent-rico method, to the second important type of heterogeneous failure time distribution, namely the minimum-type model. Also, there remains open the question whether the tangent-rico method can be extended to the case of multivariate mixtures. Still, the biggest challenge is to incorporate all of this in a performant software package, which is accessible to all kinds of domains, including the reliability of electronic components.

Appendix A

Conditions for (maximum) likelihood estimation

A.1 Conditions of Cramér and Wald

Appendix B

Additional tables

B.1 The influence on the starting values of the choice of the plotting positions

Table B.1: Comparison of candidates found for inflection points or nodes using different sorts of plotting positions for some worst-case scenarios.

n	m	$1-e^{-S_i}$	$\frac{i-0.5}{n}$	$\frac{i-0.3175}{n+0.365}$	$\frac{i}{n+1}$
20	3	?	?	?	?
	5				
50	3				
	5				
	11				
100	5				
	11				
	21				
500	25				
	51				
	101				
1000	51				
	101				
	201				

(a) Comparison of "best" inflection point found based on a SEV QQ-plot when samples originate from a 2-component normal mixture with true parameter values $\mu_1 = 0, \sigma_1 = 1, \mu_2 = 3, \sigma_2 = 0.5, p_1 = 0.2$.

		p_i			
n	m	$1-e^{-S_i}$	$\frac{i-0.5}{n}$	$\frac{i-0.3175}{n+0.365}$	$\frac{i}{n+1}$
20	3	91	96	95	94
	5	81 (84)	89 (92)	90 (92)	86 (88)
50	3				
	5				
	11				
100	5				
	11				
	21				
500	25				
	51				
	101				
1000	51				
	101				
	201				

(b) Comparison of "best" couple nodes found based on a normal QQ-plot when samples originates from a normal distribution with true parameter values $\mu = 0, \sigma = 4$.

B.2 A simulation study for two-component SEV mixtures

B.2.1 Search for the largest local maximum

Table B.2: Comparison between the largest maximum obtained with method II and the largest maximum obtained with method I or III.

		$\mu_2 = 1$		$\mu_2 = 2$		$\mu_2 = 3$		$\mu_2 = 4$	
n		*I*	*III*	*I*	*III*	*I*	*III*	*I*	*III*
20	=					53	100	64	100
	>					47	0	36	0
50	=					29	100	69	100
	>					71	0	31	0

(a) Group 1: separation in shape parameter.

		$\sigma_1 = 1$		$\sigma_1 = 0.5$		$\sigma_1 = 0.2$		$\sigma_1 = 0.1$	
n		*I*	*III*	*I*	*III*	*I*	*III*	*I*	*III*
20	=					71*	99*	86	100
	>					28	0	14	0
50	=					92	100	100	100
	>					8	0	0	0

*Only 99 simulations are carried out. Simulation 100 compared maxima with a singularity

(b) Group 2: separation in scale parameter. The number between brackets is the number of samples where no maximum at all was found.

		$\pi_1 = 0.2$		$\pi_1 = 0.4$		$\pi_1 = 0.6$		$\pi_1 = 0.8$	
n		*I*	*III*	*I*	*III*	*I*	*III*	*I*	*III*
20	=					73	100	64**	98**
	>					27	0	33	0
50	=					90	100	89	100
	>					10	0	11	0

*Only 98 simulations are carried out. Simulation 99, 100 compared maxima with a singularity

(c) Group 3: varying the proportion parameter.

Table B.3: The number of times (k) out of 100 that the starting value of a certain method leads to the LE for the sets of parameter values of group 1.

	$\mu_2 = 1$			$\mu_2 = 2$		
n	A	B	D	A	B	D
20				30	14 (4)	22
50				18	5 (4)	7
100				15	12 (2)	12
500						

	$\mu_2 = 3$			$\mu_2 = 4$		
n	A	B	D	A	B	D
20	52	28 (1)	43	52	34	46
50	46	31	32	79	75	75
100	56	50	53	91	91	91
500						

B.2.2 Consistency of the tangent-rico method

Table B.4: The number of times (k) out of 100 that the starting value of a certain method leads to the LE for the sets of parameter values of group 2.

	$\sigma_1 = 1$			$\sigma_1 = 0.5$		
n	A	B	D	A	B	D
20				39 (1)	17 (9)	28 (3)
50				56	36	41
100				71	67	67
500						

	$\sigma_1 = 0.2$			$\sigma_1 = 0.1$		
n	A	B	D	A	B	D
20	67	49	62	90	73	82
50	98	93	97	100	98	99
100	100	100	100	100	100	100
500						

B.3 Tables for Section 4.5

MLE versus adapted MLEs for large value of $\hat{\sigma}^{\mathbf{MLE}}$ The sample from Table B.6 has size 50 and is simulated from a two-component normal mixture with parameter values $\mu_1 = 0$, $\mu_2 = 7$, $\sigma = 3$ and $\pi_1 = 0.5$. Except for some boundary solutions, all likelihood functions had only one maximum. The MLE is given by $\hat{\mu}_1 = -0.327(1.11)$, $\hat{\mu}_2 = 5.86(1.03)$, $\hat{\sigma} = 2.79(0.435)$ and $\hat{\pi}_1 = 0.435(0.156)$.

LE versus adapted LEs

Table B.5: The number of times (k) out of 100 that the starting value of a certain method leads to the MLE for the sets of parameter values of group 3.

	$\pi_1 = 0.2$			$\pi_1 = 0.4$		
n	A	B	D	A	B	D
20	47	18 (1)	31 (1)	62	26 (2)	44 (1)
50	30	18	13	70	55	62
100	44	30	33	99	97	98
500						

	$\pi_1 = 0.6$			$\pi_1 = 0.8$		
n	A	B	D	A	B	D
20	61 (1)	45 (2)	51 (1)	66 (1)	57 (10)	45 (20)
50	95	84	92	94	90	91
100	100	99	100	100	99	99
500						

Table B.6: Maximum absolute difference between parameter estimates (first table) and estimated standard errors (second table) of the different methods. The parameter where the difference is largest, is indicated between brackets.

δ	$\frac{\delta}{\hat{\alpha}^{MLE}}$	MLE - MLEδ	MLE - MLEδ^*	MLEδ - MLEδ^*	MLE - MLEδ_s
1e-6	3.58e-7	1.67e-7 (μ_1)	1.24e-7 (μ_1)	2.91e-7 (μ_1)	1.37e-9 ($t_{0.001}$)
1e-5	3.58e-6	1.52e-6 (μ_2)	2.57e-6 ($t_{0.001}$)	2.15e-6 ($t_{0.001}$)	4.38e-10 (μ_1)
1e-4	3.58e-5	1.86e-5 ($t_{0.001}$)	4.37e-5 ($t_{0.001}$)	6.23e-5 ($t_{0.001}$)	4.58e-10 ($t_{0.001}$)
1e-3	3.58e-4	1.28e-4 (μ_2)	4.00e-4 (μ_2)	5.29e-4 (μ_2)	4.23e-8 ($t_{0.001}$)
1e-2	3.58e-3	2.31e-3 ($t_{0.001}$)	3.45e-3 ($t_{0.001}$)	5.75e-3 ($t_{0.001}$)	4.23e-6 ($t_{0.001}$)
0.1	3.58e-2	2.78e-2 ($t_{0.001}$)	3.43e-2 ($t_{0.001}$)	3.96e-2 (μ_1)	4.23e-4 ($t_{0.001}$)
1	0.358	0.152 ($t_{0.001}$)	0.274 (μ_1)	0.359 (μ_1)	4.26e-2 ($t_{0.001}$)
1.5	0.538	0.232 (μ_1)	0.224 ($t_{0.001}$)	0.398 (μ_1)	9.68e-2 ($t_{0.001}$)
2	0.717	0.294 (μ_2)	0.675 (μ_1)	0.864 (μ_2)	0.174 ($t_{0.001}$)

δ	# I	MLE - MLEδ	MLE - MLEδ^*	MLEδ - MLEδ^*	MLE - MLEδ_s
1e-6	50	7.39e-8 (μ_1)	9.02e-8 (μ_1)	1.64e-7 (μ_1)	2.19e-9 (μ_2)
1e-5	50	6.42e-7 (μ_2)	5.27e-7 (μ_2)	1.40e-7 (μ_1)	6.60e-10 ($t_{0.001}$)
1e-4	50	7.90e-6 (μ_1)	1.78e-5 (μ_1)	2.57e-5 (μ_1)	8.67e-11 ($t_{0.001}$)
1e-3	50	6.29e-5 (μ_2)	2.17e-4 (μ_1)	2.69e-4 (μ_1)	6.03e-9 ($t_{0.001}$)
1e-2	50	1.77e-3 (μ_1)	1.98e-3 (μ_1)	3.75e-3 (μ_1)	6.03e-7 ($t_{0.001}$)
0.1	44	1.11e-2 (μ_1)	1.69e-2 (μ_1)	3.96e-2 (μ_2)	6.03e-5 ($t_{0.001}$)
1	17/18	0.118 (μ_2)	8.93e-2 (μ_2)	0.207 (μ_2)	6.04e-3 ($t_{0.001}$)
1.5	13/12	4.97e-2(μ_2)	0.234 (μ_1)	0.230 (μ_2)	1.36e-2 ($t_{0.001}$)
2	10/9	0.136 (μ_1)	0.256 (μ_2)	0.235 (μ_2)	2.42e-2 ($t_{0.001}$)

Note: # I is the the number of different intervals in the sample for the adapted ML methods. In case of an unequal number between the two methods, the first number refers to the sample corresponding to the estimator MLEδ.

Table B.7: Maximum absolute difference between parameter estimates and estimated standard errors of the different methods: d_1 refers to $\max|\hat{\theta} - \tilde{\theta}|$ and d_2 to $\max|\hat{se}(\hat{\theta}) - \tilde{se}(\tilde{\theta})|$, with $\hat{\theta}$ and $\tilde{\theta}$ one of the estimates LE, MLEδ or LE^*_δ. The parameter where the difference was largest is indicated between brackets.

δ	$\frac{\delta}{\min(\hat{\sigma}_1^{LE}, \hat{\sigma}_2^{LE})}$		LE - MLEδ	LE - LE^*_δ	MLEδ - LE^*_δ
$1e^{-6}$	$1.81e^{-3}$	d_1	$7.55e^{-11}$ (σ_1)	$7.08e^{-7}$ (p_1)	$7.08e^{-7}$ (p_1)
		d_2	$3.10e^{-11}$ (σ_1)	$1.44e^{-7}$ (μ_1)	$1.44e^{-7}$ (μ_1)
$1e^{-5}$	$1.81e^{-2}$	d_1	$7.55e^{-9}$ (σ_1)	$3.35e^{-6}$ (p_1)	$3.35e^{-6}$ (p_1)
		d_2	$3.10e^{-9}$ (σ_1)	$6.59e^{-7}$ (μ_1)	$6.59e^{-7}$ (μ_1)
$1e^{-4}$	0.181	d_1	$7.56e^{-7}$ (σ_1)	$7.89e^{-5}$ (p_1)	$7.89e^{-5}$ (p_1)
		d_2	$3.11e^{-7}$ (σ_1)	$1.60e^{-5}$ ($t_{0.001}$)	$1.60e^{-5}$ ($t_{0.001}$)
0.0005	0.905	d_1	$2.01e^{-5}$ (σ_1)	0.058 (p_1)	0.058 (p_1)
		d_2	$8.54e^{-6}$ (σ_1)	0.018 ($t_{0.001}$)	0.018 ($t_{0.001}$)
$1e^{-3}$	1.81	d_1	$1.04e^{-4}$ (σ_1)	$2.24e^{-4}$ (p_1)	$2.11e^{-4}$ (p_1)
		d_2	$5.08e^{-5}$ (σ_1)	$4.77e^{-5}$ (μ_1)	$8.43e^{-5}$ (σ_1)
0.005	9.05	d_1	0.058 (p_1)	0.058 (p_1)	0.0015 ($t_{0.001}$)
		d_2	0.018 ($t_{0.001}$)	0.018 ($t_{0.001}$)	0.00076 (σ_1)
0.01	18.1	d_1	$\hat{\sigma}_1^{MLE\delta} = 0$ (*)	0.34 (p_1)	-
		d_2	-	0.14 ($t_{0.001}$)	-
0.05	90.5	d_1	0.33 (p_1)	0.32 (p_1)	0.017 (μ_1)
		d_2	0.13 ($t_{0.001}$)	0.13 ($t_{0.001}$)	0.0032 (p_1)
0.1	181	d_1	$\hat{\sigma}_1^{MLE\delta} = 0$ (*)	0.61 (σ_1)	-
		d_2	-	0.086 ($t_{0.001}$)	-

Table B.8: Testing $\sigma_1 = \sigma_2$ with the likelihood ratio test. The LR statistic is assumed to have a χ^2 distribution with 1 df. The last column indicates the lowest α-level (of the commonly used) on which the H_0 would be rejected.

δ	$\frac{\delta}{\min(\hat{\sigma}_1^{LE}, \hat{\sigma}_2^{LE})}$	Method	LRT-value	p-value	α-level A
		LE	6.639	0.00998	1%
$1e^{-6}$	$1.81e^{-3}$	LE^*_δ	6.636	0.00999	1%
		MLEδ	6.639	0.00998	1%
$1e^{-5}$	$1.81e^{-2}$	LE^*_δ	6.651	0.00991	1%
		MLEδ	6.639	0.00998	1%
$1e^{-4}$	0.181	LE^*_δ	6.361	0.0117	5%
		MLEδ	6.639	0.00997	1%
0.0005	0.905	LE^*_δ	5.944	0.0148	5%
		MLEδ	6.641	0.00996	1%
$1e^{-3}$	1.81	LE^*_δ	7.586	0.00588	1%
		MLEδ	6.691	0.00969	1%
0.005	9.05	LE^*_δ	5.855	0.0155	5%
		MLEδ	6.262	0.0123	5%
0.01	18.1	LE^*_δ	4.563	0.0327	5%
		MLEδ	5.032	0.0249 ?	5%
0.05	90.5	LE^*_δ	6.538	0.0106	5%
		MLEδ	4.237	0.0396	5%
0.1	181	LE^*_δ	2.509	0.113	12%
		MLEδ	5.162	0.0231	5%

Table B.9: Maximum absolute difference between parameter estimates and estimated standard errors of the different methods: d_1 refers to $\max|\hat{\theta} - \tilde{\theta}|$ and d_2 to $\max|\hat{se}(\hat{\theta}) - \tilde{se}(\tilde{\theta})|$, with $\hat{\theta}$ and $\tilde{\theta}$ one of the estimates LE, MLEδ or LE^*_δ. The parameter where the difference was largest is indicated between brackets.

δ	$\frac{\delta}{\min(\hat{\sigma}_1^{LE}, \hat{\sigma}_2^{LE})}$		LE - MLEδ	LE - LE^*_δ	MLEδ - LE^*_δ
$1e^{-6}$	$5.97e^{-6}$	d_1	$4.78e^{-11}$ $(t_{0.001})$	$3.37e^{-7}$ $(t_{0.001})$	$3.37e^{-7}$ $(t_{0.001})$
		d_2	$8.37e^{-11}$ $(t_{0.001})$	$8.34e^{-8}$ $(t_{0.001})$	$8.34e^{-8}$ $(t_{0.001})$
$1e^{-5}$	$5.97e^{-5}$	d_1	$4.53e^{-11}$ $(t_{0.001})$	$6.88e^{-6}$ $(t_{0.001})$	$6.88e^{-6}$ $(t_{0.001})$
		d_2	$1.12e^{-11}$ $(t_{0.001})$	$1.68e^{-6}$ $(t_{0.001})$	$1.68e^{-6}$ $(t_{0.001})$
$1e^{-4}$	$5.97e^{-4}$	d_1	$4.11e^{-9}$ $(t_{0.001})$	$6.17e^{-5}$ $(t_{0.001})$	$6.17e^{-5}$ $(t_{0.001})$
		d_2	$1.03e^{-9}$ $(t_{0.001})$	$2.23e^{-5}$ $(t_{0.001})$	$2.23e^{-5}$ $(t_{0.001})$
$1e^{-3}$	$5.97e^{-3}$	d_1	$4.11e^{-7}$ $(t_{0.001})$	$1.63e^{-4}$ (σ_2)	$1.63e^{-4}$ (σ_2)
		d_2	$1.03e^{-7}$ $(t_{0.001})$	$5.14e^{-5}$ (μ_2)	$5.14e^{-5}$ (μ_2)
$1e^{-2}$	$5.97e^{-2}$	d_1	$4.11e^{-5}$ $(t_{0.001})$	$4.33e^{-3}$ $(t_{0.001})$	$4.37e^{-4}$ $(t_{0.001})$
		d_2	$1.03e^{-5}$ $(t_{0.001})$	$1.73e^{-3}$ $(t_{0.001})$	$1.72e^{-3}$ $(t_{0.001})$
0.1	0.597	d_1	0.0041 $(t_{0.001})$	0.047 $(t_{0.001})$	0.052 $(t_{0.001})$
		d_2	0.0010 $(t_{0.001})$	0.016 $(t_{0.001})$	0.015 $(t_{0.001})$
0.2	1.19	d_1	0.017 $(t_{0.001})$	0.021 (σ_2)	0.032 (σ_2)
		d_2	0.0040 $(t_{0.001})$	0.0067 (μ_2)	0.0068 (μ_2)
0.3	1.79	d_1	0.038 $(t_{0.001})$	0.066 (μ_2)	0.067 (μ_2)
		d_2	0.0087 $(t_{0.001})$	0.0040 (μ_2)	0.0084 $(t_{0.001})$
0.5	2.99	d_1	0.11 $(t_{0.001})$	0.056 (μ_2)	0.12 $(t_{0.001})$
		d_2	0.020 $(t_{0.001})$	0.0055 $(t_{0.001})$	0.014 $(t_{0.001})$
0.7	4.18	d_1	$\hat{\sigma}_2^{MLE\delta} \to 0$	0.36 $(t_{0.001})$	-
		d_2	-	0.39 (μ_2)	-
1	5.97	d_1	$\hat{\sigma}_1^{MLE\delta} \to 0$ and $\hat{\sigma}_2^{MLE\delta} \to 0$	0.48 $(t_{0.001})$	-
		d_2	-	0.21 (μ_2)	-

Table B.10: Testing $\sigma_1 = \sigma_2$ with the likelihood ratio test. The LR statistic is assumed to have a χ^2 distribution with 1 df. The last column indicates the lowest α-level (of the commonly used) on which the H_0 would be rejected.

δ	$\frac{\delta}{\min(\hat{\sigma}_1^{LE}, \hat{\sigma}_2^{LE})}$	Method	LRT-value	p-value	α-level A
		LE	2.149	0.143	15%
$1e^{-6}$	$5.97e^{-6}$	LE^*_δ	2.149	0.143	15%
		MLEδ	2.149	0.143	15%
$1e^{-5}$	$5.97e^{-5}$	LE^*_δ	2.149	0.143	15%
		MLEδ	2.149	0.143	15%
$1e^{-4}$	$5.97e^{-4}$	LE^*_δ	2.150	0.143	15%
		MLEδ	2.149	0.143	15%
$1e^{-3}$	$5.97e^{-3}$	LE^*_δ	2.140	0.143	15%
		MLEδ	2.149	0.143	15%
$1e^{-2}$	$5.97e^{-2}$	LE^*_δ	2.288	0.130	15%
		MLEδ	2.149	0.143	15%
0.1	0.597	LE^*_δ	3.223	0.0726	10%
		MLEδ	2.150	0.143	15%
0.2	1.19	LE^*_δ	1.281	0.258	30%
		MLEδ	2.160	0.142	15%
0.3	1.79	LE^*_δ	1.599	0.206	25%
		MLEδ	2.211	0.137	15%
0.5	2.99	LE^*_δ	3.345	0.0674	10%
		MLEδ	2.460	0.117	15%
0.7	4.18	LE^*_δ	0.00916	0.942	95%
		MLEδ	2.766	0.0963	10%
1	5.97	LE^*_δ	0.0372	0.847	85%
		MLEδ	0	1	non

References

Acheson and McElwee (1951). Concerning the reliability of electron tubes. *Sylvania Technologist 4*.

Aitchison, J. and S. Silvey (1958). Maximum likelihood estimation of parameters subject to restraints. *Annals of Mathematical Statistics 29*, 813–828.

Aitkin, M. (2001). Likelihood and bayesian analysis of mixtures. *Statistical modelling 1*, 287–304.

Aitkin, M. and D. Rubin (1985). Estimation and hypothesis testing in finite mixture models. *Journal of the Royal Statistical Society : Series B 47*, 67–75.

Anderson, I. and J. Bezdek (1984). Curvature and tangential deflection of discrete arcs: a theory based on the commutator of scatter matrix pairs and its application to vertex detection in planar shape data. *IEEE transactions on pattern analysis and machine intelligence 6*(1), 27–40.

Atakov, E., J. Clement, and B. Miner (1994). Two electromigration failure modes in polycrystalline aluminum interconnects. *IEEE/IRPS*, 213–223.

Barnett, V. (1966). Evaluation of the maximum-likelihood estimator where the likelihood equation has multiple roots. *Biometrika 53*, 151–165.

Battiti, R. and G. Tecchiolli (1996). The continuous reactive tabu search : Blending combinatorial optimization and stochastic search for global optimization. *Annals of 0peration Research 63*, 153–188.

Behboodian, J. (1970). On the modes of a mixture of two normal distributions. *Technometrics 12*, 131–139.

Behboodian, J. (1972). Information matrix for a mixture of two normal distributions. *Journal of Statistical Computation and Simulation 1*, 919–923.

Bezdek, J. and J. Dunn (1975). Optimal fuzzy partitions: a heuristic for estimating the parameters in a mixture of normal distributions. *IEEE transactions on computers ?*, 835–838.

Bhattacharya, C. (1967). A simple method for resolution of a distribution into its gaussian components. *Biometrics*.

Birolini, A. (1994). *Quality and Reliability of Technical Systems*. Springer - Verlag: Berlin.

Böhning, D. (1995). A review of reliable maximum likelihood algorithms for semiparametric mixture models. *Journal of Statistical Planning and Inference 47*, 5–28.

Böhning, D. (2000). *Computer-Assisted analysis fo mixtures and applications. Meta-analyis, disease mapping and others*. Chapman & Hall/CRC: Boca Raton.

Bowman, K. and L. Shenton (1973). Space solutions for a normal mixture. *Biometrika 60*(3), 629–636.

Carlin, B. and T. Louis (1996). *Bayes and Empirical Bayes Methods for Data Analysis*. London: Chapman & Hall.

Celeux, G., D. Chauveau, and J. Diebolt (1996). Stochastic versions of the em algorithm: an experimental study in the mixture case. *Journal of Statistical Computation and Simulation 55*, 287–314.

Celeux, G., S. Chrtien, F. Forbes, and A. Mkhadri (2001). A component-wise em algorithm for mixtures. *Journal of Computational and Graphical Statistics 10*, 697–712.

Cetin, B., J. Barhen, and J. Burdick (1993). Teriminal repeller unconstrained subenergy tunneling (trust) for fast global optimisation. *Journal of Optimization Theory and Applications 77*, 97–126.

Chan, V. and W. Meeker (1999). A failure-time model for infant-mortality and wearout failure modes. *IEEE Transactions on Reliability 48*, 377–387.

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

Chelouah, R. (2000). A continuous genetic algorithm designed for the global optimization of mulitmodal functions. *Journal of Heuristics 6*, 191–213.

Cheng, R. and T. Iles (1987). Corrected maximum likelihood in non-regular problems. *Journal of the Royal Statistical Society: Series B 49*, 95–101.

Cox, D. and D. Hinkley (1974). *Theoretical statistics*. London: Chapman and Hall.

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton: Princeton University Press.

Crowder, M., A. Kimber, R. Smith, and T. Sweeting (1991). *Statistical analysis of reliability data*. London, Chapman and Hall.

D'agostino, R. and M. Stephens (1986). *Goodness-of-fit techniques*. New York: Marcel Dekker, Inc.

Davidian, M. and D. Giltinan (1995). *Nonlinear Models for Repeated Measurement Data*. London: Chapman & Hall.

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

Degraeve, R. (1998). *Time dependent dielectric breakdown in thin oxides: mechanisms, statistics and oxide reliability prediction*. Ph. D. thesis, K.U.Leuven.

Degraeve, R., J. Ogier, R. Bellens, P. Roussel, G. Groeseneken, and H. Maes (1998). A new model for the field dependence of intrinsic and extrinsic time-dependent dielectric breakdown. *IEEE Transactions on Electron Devices 45*, 472–481.

Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the em-algorithm. *Journal of the royal statistical society: series B 39*, 1–38.

Duda, R. and P. Hart (1973). *Pattern Classification and Scene Analysis*. New York: Wiley.

Engelman, L. and J. Hartigan (1969). Percentage points of a test for clusters. *Journal of the American Statistical Association ?*, 1647–1648.

Everitt, B. (1984). Maximum likelihood estimation of the parameters in a mixture of two univariate normal distributions: a comparison of different algorithms. *The statistician 33*, 205–215.

Everitt, B. and D. Hand (1981). *Finite mixture distributions*. Monographs on applied probability and statistics. Chapman & Hall: London.

Failure. Software package for reliabilty data distributed by xact, weg naar as 322, 3600 genk, belgium.

Finch, S., N. Mendell, and H. jr. Thode (1989). Probabilistic measures of adequacy of a numerical search for a global maximum. *Journal of the American Statistical Association 84*, 1020–1023.

Fisher, A., A. Abel, M. Lepper, A. Zitzelsberger, and A. von Glasgow (2000). Experimental data and statistical models for bimodal em failures. In *IEEE international reliability physics symposium proceedings.*

Fowlkes, E. (1979). Some methods for studying the mixture of two normal (lognormal) distributions. *Journal of the American Statistical Association 74*(367), 561–575.

Fraley, C. and A. Raftery (2000). Model-based clustering, discriminant analysis and density estimtation. Technical Report 380, University of Washington.

Furman, W. and B. Lindsay (1994). Measuring the relative effectiveness of moment estimators as starting values in maximizing likelihoods. *Computational Statistics & Data Analysis 17*, 493–507.

Gumbel, E. (1958). *Statistics of Extremes.* Columbia University Press.

Harding, J. (1948). The use of probability paper for the graphical analysis of polymodal frequency distributions. *Journal of the Marine Biological Association of the United Kingdom 28*, 141–153.

Hasselblad, V. (1969). Estimation of finite mixtures of distributions from the exponential family. *American Statistical Association Journal ?*, 1459–1471.

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The elements of statistical learning: data mining, inference and prediction.* New York: Springer.

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

Hathaway, R. and J. Bezdek (1986). On the asymptotic properties of fuzzy c-means cluster prototypes as estimators of mixture subpopulation centers. *Communication in statistics: theory and methods 15*, 505–513.

Haughton, D. (1997). Packages for estimating finite mixtures: a review. *American Statistician 51*, 194–.

Hosmer, D. (1973). On mle of the parameters of a mixture of two normal distributions when the sample size is small. *Communications in Statistics 1*, 217–227.

Huzurbazar, V. (1948). The likelihood equations, consistency, and the maximum of the likelihood function. *Annals of Eugenetics 14*, 185.

JEDEC. Solid state technology association, www.jedec.org.

Jewell, N. (1982). Mixtures of exponential distributions. *The Annals of Statistics 10*, 479–484.

Jiang, R. and D. Murthy (1995). Modeling failure-data by mixture of 2 weibull distributions: A graphical approach. *IEEE transactions on reliability 44*(3), 477–488.

Johnson, R. and D. Wichern (1998). *Applied Multivariate Statistical Analysis* (4^{th} ed.). New Jersey: Prentice Hall.

Joyce, W., R. Dixon, and R. Hartman (1976). Statistical characterization of the lifetimes of continuously operated (al,ga)as double-heterostructure lasers. *Applied physics letters 28*(11), 684–686.

Kao, J. (1959). A graphical estimation of mixed weibull parameters in life-testing of electron tubes. *Technometrics 4*.

Kiefer, J. and J. Wolfowitz (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics*, 887–906.

Kiefer, N. (1978). Discrete parameter variation: efficient estimation of a switching regression model. *Econometrica 46*, 427–433.

Lawless, J. (1982). *Statistical Models and Methods for Lifetime Data*. New York: John Wiley & Sons.

Lehmann, E. (1980). Efficient likelihood estimators. *The American Statistician 34*, 233–235.

Lehmann, E. (1983). *Theory of point estimation*. Wiley: New York.

Lindsay, B. (1983a). The geometry of mixture likelihoods : a general theory. *The Annals of Statistics 11*, 86–94.

Lindsay, B. (1983b). The geometry of mixture likelihoods, part ii: the exponential family. *The Annals of Statistics 11*, 783–792.

Lindsay, B. (1995). *Mixture Models: Theory, Geometry and Applications*, Volume 5 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Hayward: Institute of Statistical Mathematics.

Ling, J. and J. Pan (1998). A new method for selection of population distribution and parameter estimation. *Reliability Engineering and System Safety 60*, 247–255.

Lloyd, J. (1979). On the lognormal distribution of electromigration lifetimes. *Journal of Applied Physics 50*, 5062–5064.

Lloyd, J. and J. Kitchin (1990). The electromigration failure distribution: the fine-line case. *Journal of Applied Physics 69*, 2117–2127.

Maritz, J. and T. Lwin (1989). *Empirical Bayes Estimation* (2nd ed.), Volume 35 of *Monographs on Statistics and Applied Probability*. London: Chapman and Hall.

Markatou, M. (2000). Mixture models, robustness and the weighted likelihood methodology. *Biometrics 56*, 483–486.

Markatou, M., A. Basu, and B. Lindsay (1998). Weighted likelihood equations with bootstrap root search. *Journal of the American Statistical Association 93*, 740–750.

Martin, A., P. O'Sullivan, and A. Mathewson (1998). Dielectric reliability measurement methods: a review. *Microelectronics and Reliability 38*, 37–72.

McLachlan, G. (1988). On the choice of starting values for the em algorithm in fitting mixture models. *The Statistician 37*, 417–425.

McLachlan, G. and D. Peel (1998). Mixfit: an algorithm for the automatic fitting and testing of normal mixture models. In *IEEE??*, pp. 553–556.

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

Meeker, W. and L. Escobar (1998). *Statistical methods for reliability data*. New York: Wiley.

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

Møltoft, J. (1983). Behind the bathtub curve: a new model and its conseqences. *Microelectronics and reliability 23*, 489–500.

Mu, F., C. Tan, and M. Xu (2000). Proportional difference estimate method of determining the characteristic parameters of monomodal and multi-

modal weibull distributions of time-dependent dielectric breakdown. *Solid State Electronics 44*, 1419–1424.

Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics 42*, 12–25.

Nelson, W. (1982). *Applied Life Data Analysis*. New York: John Wiley & Sons.

Ogawa, E., K.-D. Lee, H. Matsuhashi, K.-S. Ko, P.-R. Justison, A. Ramamurthi, and A. B. P. Ho (2001). Statistics of electromigration early failures in cu/oxide dual-damascene interconnects. In *IEEE Annual Proceedings Reliability Physcics*, Volume 39, pp. 341–349.

Ohring, M. (1998). *Reliability and Failure of Electronic Materials and Devices*. London : Academic Press.

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

Perlman, M. (1983). The limiting behavior of multiple roots of the likelihood equations. In M. Rizvi, J. Rustagi, and D. Siegmund (Eds.), *Recent Advances in Statistics: Papers in honor of Herman Chernoff on his Sixtieth Birthday*, pp. 339–370. New York: Academic Press.

Preston, E. (1952). A graphical method for the analysis of statistical distributions into two normal components. *Biometrika 40*, 460–464.

Quandt, R. (1978). A new approach to estimating switching regressions. *Journal of the American statistical association 76*, 306–310.

Quandt, R. and J. Ramsey (1978). Estimating mixtures of normal distributions and switching regressions. *Journal of the American statistical association 73*(364), 730–751.

Rao, C. (1948). The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society 10*, 159–203.

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

Redner, R., R. Hathaway, and J. Bezdek (1987). Estimating the parameters of mixture models with modal estimators. *Communications in Statistics: Theory and Methods 16*, 2639–2660.

Redner, R. and H. Walker (1984). Mixture densities, maximum likelihood and the em-algorithm. *SIAM review 29*, 195–239.

Reeds, J. (1985). Asymptotic number of roots of cauchy location likelihood equations. *The Annals of Statistics 13*, 775–784.

Rider, P. (1962). Estimating the parameters of mixed poisson, binomial, and weibull distributions (by the method of moments). *Bulletin of the International Statistical Institute 39*, 225–232.

Sichart, K. and R. Vollertsen (1991). Bimodal lifetime distributions of dielectrics for integrated circuits. *Quality and Reliability Engineering International 7*, 299–305.

Small, G., J. Wang, and Z. Yang (2000). Eliminating multiple root problems in estimation. *Statistical Science 15*, 313–341.

Sundberg, R. (1974). Maximum likelihood theory for incomplete data from an exponential family. *Scandinavian Journal of Statistics: Theory and Applications 1*, 49–58.

Tarone, R. and Gruenhage (1975). A note on the uniqueness of roots of the likelihood equations for vector-valued parameters. *Journal of the American Statistical Association 70*, 903–904.

Teicher, H. (1961). Identifiability of mixtures. *Annals of Mathematical Statistics 34*, 244–248.

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

Titterington, D., A. Smith, and U. Makov (1985). *Statistical Analysis of Finite Mixtures*. New York: Wiley.

Ueda, N., R. Nakano, Z. Ghahramani, and G. Hinton (2000). Smem algorithm for mixture models. *Neural Computation 12*, 2109–2128.

Verbeke, G. and G. Molenberghs (1997). *Linear Mixed Models in Practice*. New-York: Springer-Verlag.

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

Weibull. Software package for reliability data distributed by the reliasoft corporation, www.weibull.com.

Wolfe, J. (1970). Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research ?*, 329–350.

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

Zhang, T. and Y. Ren (2002). Failure data analysis by models involving 3 weibull distributions. In *2002 Proceedings: Annual Reliability and Maintainability Symposium*, pp. 45–50.