Hypothesis Testing in Finite Mixture Models

by

Pengfei Li

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2007

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Mixture models provide a natural framework for unobserved heterogeneity in a population. They are widely applied in astronomy, biology, engineering, finance, genetics, medicine, social sciences, and other areas.

An important first step for using mixture models is the test of homogeneity. Before one tries to fit a mixture model, it might be of value to know whether the data arise from a homogeneous or heterogeneous population. If the data are homogeneous, it is not even necessary to go into mixture modeling. The rejection of the homogeneous model may also have scientific implications. For example, in classical statistical genetics, it is often suspected that only a subgroup of patients have a disease gene which is linked to the marker. Detecting the existence of this subgroup amounts to the rejection of a homogeneous null model in favour of a two-component mixture model. This problem has attracted intensive research recently. This thesis makes substantial contributions in this area of research.

Due to partial loss of identifiability, classic inference methods such as the likelihood ratio test (LRT) lose their usual elegant statistical properties. The limiting distribution of the LRT often involves complex Gaussian processes, which can be hard to implement in data analysis. The modified likelihood ratio test (MLRT) is found to be a nice alternative of the LRT. It restores the identifiability by introducing a penalty to the log-likelihood function. Under some mild conditions, the limiting distribution of the MLRT is $1/2\chi_0^2 + 1/2\chi_1^2$ where χ_0^2 is a point mass at 0. This limiting distribution is convenient to use in real data analysis. The choice of the penalty functions in the MLRT is very flexible. A good choice of the penalty enhances the power of the MLRT. In this thesis, we first introduce a new class of penalty functions, with which the MLRT enjoys a significantly improved power for testing homogeneity.

The main contribution of this thesis is to propose a new class of methods for testing homogeneity. Most existing methods in the literature for testing of homogeneity, explicitly or implicitly, are derived under the condition of finite Fisher information and a compactness assumption on the space of the mixing parameters. The finite Fisher information condition can prevent their usage to many important mixture models, such as the mixture of geometric distributions, the mixture of exponential distributions and more generally mixture models in scale distribution families. The compactness assumption often forces applicants to set artificial bounds for the parameters of interest and makes the resulting limiting distribution dependent on these bounds. Consequently, developing a method without such restrictions is a dream of many researchers. As it will be seen, the proposed EM-test in this thesis is free of these shortcomings.

The EM-test combines the merits of the classic LRT and score test. The properties of the EM-test are particularly easy to investigate under single parameter mixture models. It has a simple limiting distribution $0.5\chi_0^2 + 0.5\chi_1^2$, the same as the MLRT. This result is applicable to mixture models without requiring the restrictive regularity conditions described earlier.

The normal mixture model is a very popular model in applications. However it does not satisfy the strong identifiability condition, which imposes substantial technical difficulties in the study of the asymptotic properties. Most existing methods do not directly apply to the normal mixture models, so the asymptotic properties have to be developed separately. We investigate the use of the EM-test to normal mixture models and its limiting distributions are derived. For the homogeneity test in the presence of the structural parameter, the limiting distribution is a simple function of the $0.5\chi_0^2+0.5\chi_1^2$ and χ_1^2 distributions. The test with this limiting distribution is still very convenient to implement. For normal mixtures in both mean and variance parameters, the limiting distribution of the EM-test is found be to χ_2^2.

Mixture models are also widely used in the analysis of the directional data. The von Mises distribution is often regarded as the circular normal model. Interestingly, it satisfies the strong identifiability condition and the parameter space of the mean direction is compact. However the theoretical results in the single parameter mixture models can not directly apply to the von Mises mixture models. Because of this, we also study the application of the EM-test to von Mises mixture models in the presence of the structural parameter. The limiting distribution of the EM-test is also found to be $0.5\chi_0^2 + 0.5\chi_1^2$.

Extensive simulation results are obtained to examine the precision of the approximation of the limiting distributions to the finite sample distributions of the EM-test. The type I errors with the critical values determined by the limiting distributions are found to be close to nominal values. In particular, we also propose several precision enhancing methods, which are found to work well. Real data examples are used to illustrate the use of the EM-test.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor Professor Jiahua Chen for his insight, guidance, constant encouragement, and for his support in so many aspects. I also wish to thank Professor Paul Marriott for agreeing to be my co-supervisor. He has given me a lot of useful comments and suggestions, which help me to significantly improve my thesis. During my tenure as a doctoral student, I have learned a lot of things from them.

I am also very grateful for the assistance and valuable advice I received from my committee members: Professor Mu Zhu, Professor Shoja'eddin Chenouri, Professor Stéphanie Lluis and Professor Yongzhao Shao, and also for all helps they have given to me during the past couple of years. I also want to thank Yuejiao Fu from York University. The skills of research I learned from her during our collaboration works are priceless to me.

Many thanks to my friends, Baojiang Chen, Longyang Wu, Runhuan Feng, Yan Yuan, Yan Liu, Hui Zhao, Xu Wang and Fang Yang for their helps and encouragement during my time in Waterloo. I also want to take this opportunity to thank Professors Jerry Lawless, Peter Song, Changbao Wu and Grace Yi for their continual encouragement. Thanks also to the staff, Mary Lou Dufton, Gwen Sharp, Lucy Simpson, Anissa Anniss and Joan Hatton.

I am so indebted to my family, my parents and my sister, for their support and encouragement throughout my whole life. Last, but not least, I am deeply grateful for my wife, Weihong Hu, and my son, Daniel Li, without their understanding, love and patience, I would not have been able to finish this thesis.

To my parents

To my wife and my son

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Mixture Models

Since the work of Pearson (1894), finite mixture models have been widely used in many disciplines such as astronomy, biology, engineering, genetics, medicine, social sciences and so on. Mixture models can be easily applied to the data set in which two or more subpopulations are mixed together. Due to its flexibility in modeling, finite mixture model has enjoyed intensive attentions over the past years, from both a practical and a theoretical viewpoints.

Typically, there are two situations where mixture models are called for. The first situation is where there are some group labels or covariates that characterize the distributions of response variables, yet, this information is missing, not measured or unavailable when the data were collected. A simple albeit artificial example, is when the data consists of the heights of graduate students at the University of Waterloo. If at the time we collected the data, gender information was not recorded, the population is deemed non-homogeneous due to factors including the gender differences. Because the height distributions for the female and the male are unlikely the same, it is better to model the height data by a mixture of two parametric distributions, for example, a mixture of two normal distributions.

In some situations, the data are known to be a mixture of several sub-populations. The statistical problem might be to make inference on the membership of each sample in the data set. For example, given the height information of an individual student, the objective is to infer the gender for this student. One possible approach is to model the height data by a mixture of two normal distributions. Given the height information, one may compute the posterior probability of the individual being a female. In more complex applications, the number of homogeneous sub-populations may not be known. In this case, the statistical problem might be cluster analysis.

We finish this section by two examples to illustrate the wide spread applications of finite mixture models.

Example 1.1.1. *A data set presented in Newcomb* (1886) *and Pearson* (1894) *consists of measurements on the ratio of forehead to body length for 1000 crabs from the Bay of Naples. The histogram of these measurements exhibited obvious asymmetry and a single normal distribution could not capture this asymmetry very well* (McLachlan and Peel, 2000, P3. Fig. 1.1)*. Weldon* (1893) *guessed that the reason for the asymmetry might be that the population contained two subspecies of crabs. When the data were collected the labels for the subspecies were not available. This explanation motivated the fitting the data set by a mixture of two normal distributions* (Pearson, 1894)*. Obviously, this is a typical example for the first situation we discussed.*

Example 1.1.2. *Genetics studies offer good examples of the second situation. Recent circadian clock studies using gene expression times suggested that there exist some common circadian-related genes in two different tissues of mouse. These common genes may not be synchronized in phase or peak expression times. Instead, some circadian-related genes may* *be delayed by 4 to 8 hours in peak expression in one tissue relative to the other* (Panda et al., 2002). *The statistical and genetical question of interest is to distinguish the synchronized genes from genes that are systematically lagged in phase/peak expression times across two tissues. Liu et al.* (2006) *used a mixture of two von Mises distributions to cluster 48 common genes in heart and liver tissues into two groups.*

1.2 Basic Definitions

As discussed in the last section, mixture models are typically used to model data that arise from a heterogeneous population. Suppose the whole population can be divided into m sub-populations and for each subpopulation, the data can be modeled by a parametric distribution. The marginal distribution for the whole population is then a mixture model. The general definition for a mixture model is given as follows.

Definition 1.2.1. *Let $f(x;\theta)$ be a parametric density function which comes from a known family of distributions $\{f(x;\theta),\ \theta \in \Theta \subset \mathbf{R}^d,\ d \geq 1\}$. Let Ψ be a distribution function defined on Θ. Then the distribution with the following density function is a mixture distribution:*

$$f(x;\Psi) = \int_\Theta f(x;\theta) d\Psi(\theta). \tag{1.1}$$

We call Ψ the mixing distribution and f the kernel function or component density. If Ψ has finitely many support points $\theta_j \in \Theta$, $j = 1, 2, \ldots, m$, with corresponding weights $\alpha_1, \ldots, \alpha_m$ ($\alpha_j > 0$, $j = 1, \ldots, m$ and $\sum_{j=1}^m \alpha_j = 1$), that is,

$$\Psi(\theta) = \sum_{j=1}^m \alpha_j I(\theta_j \leq \theta),$$

then (1.1) becomes

$$f(x;\Psi) = \sum_{j=1}^m \alpha_j f(x;\theta_j). \tag{1.2}$$

We call this model a finite mixture model with m components; m is called the number of components or the order of the mixture model; the weights $(\alpha_1, \ldots, \alpha_m)$ are called the mixing proportions, and the support points $(\theta_1, \ldots, \theta_m)$ are called the component parameters.

In the above formulation of the finite mixture model, the number of components m is considered fixed. But of course in many applications, the value of m is unknown and has to be inferred from the available data, along with the mixing proportions and the component parameters.

For general mixture models as in (1.1), the parameter space consists of mixing distributions. For finite mixture models as in (1.2), when m is known, the parameter space is finite dimensional and we write it as

$$\Omega \;=\; \Big\{\Psi(\theta) = \sum_{j=1}^{m} \alpha_j I(\theta_j \leq \theta) : \sum_{j=1}^{m} \alpha_j = 1, \alpha_j \geq 0, \theta_j \in \Theta, \text{ for } j = 1, \ldots, m\Big\}. \quad (1.3)$$

In Example 1.1.1, Pearson (1894) used the following normal mixture model:

$$f(x; \Psi) = \frac{1-\alpha}{\sqrt{2\pi\sigma_1^2}} \phi\Big(\frac{x-\mu_1}{\sigma_1}\Big) + \frac{\alpha}{\sqrt{2\pi\sigma_2^2}} \phi\Big(\frac{x-\mu_2}{\sigma_2}\Big), \quad (1.4)$$

where ϕ is the density function for standard normal distribution. For this model, $m = 2$, $\alpha_1 = 1 - \alpha$, $\alpha_2 = \alpha$, $\theta_1 = (\mu_1, \sigma_1)$, $\theta_2 = (\mu_2, \sigma_2)$, and

$$\Psi(\mu, \sigma) = \sum_{j=1}^{2} \alpha_j I(\mu_j \leq \mu, \sigma_j \leq \sigma).$$

A very important concept associated with the mixture model is identifiability, which is the foundation for estimation problems. The estimation of Ψ will become meaningless if the parameters in Ψ are not identifiable. In general, a parametric distribution family is said to be identifiable if different parametric values give different members of the family. The identifiability for the finite mixture model is defined similarly. That is,

Definition 1.2.2. *Let $f(x;\Psi) = \sum_{j=1}^{m} \alpha_j f(x;\theta_j)$ be the member of a parametric family of finite mixture models, where $\Psi \in \Omega$ and Ω is given in* (1.3)*. This class of finite mixture models is said to be identifiable if for any two members $f(x;\Psi)$ and $f(x;\Psi^*)$,*

$$\sum_{j=1}^{m} \alpha_j f(x;\theta_j) = \sum_{j=1}^{m^*} \alpha_j^* f(x;\theta_j^*)$$

if and only if $m = m^*$, $(\alpha_1, \ldots, \alpha_m) = (\alpha_1^*, \ldots, \alpha_{m^*}^*)$ and $(\theta_1, \ldots, \theta_m) = (\theta_1^*, \ldots, \theta_{m^*}^*)$ after permuting the component labels.

Teicher (1963) showed that except for mixtures of uniform densities, many finite mixtures of continuous densities, especially univariate mixtures of normals, mixtures of exponentials and Gamma distributions, are generally identifiable. These results are extended to multivariate families such as multivariate mixtures of normals, see Yakowitz and Spragins (1968). Mixtures of discrete distributions need not to be identifiable, for example, mixtures of binomial distributions are not identifiable if the common size parameter for binomial distribution is smaller than $2m-1$. Whereas finite mixtures of Poisson distributions (Teicher, 1960) as well as finite mixtures of negative binomial distributions (Yakowitz and Spragins, 1968) are identifiable. See Titterington et al. (1985, Section 3.1) for a detailed account of the identifiability of finite mixture models. Many of the above mentioned mixture models are also strongly identifiable. Strong identifiability was first introduced by Chen (1995), which is required in developing some useful asymptotic results in hypothesis testing in finite mixture model, see Chen and Chen (2001), Chen et al. (2001, 2004).

We finish this section with the following comment on the identifiability for the finite mixture model. As we mentioned before, the mixture of normal distributions, the mixture of Poisson distributions and so on, are identifiable or even strongly identifiable. However, when m is quite large, some of the mixing proportions becomes close to 0, these mixture models are close to non-identifiable in the sense that the mixture density with m components might be empirically indistinguishable from one with fewer than m components. Because of this, in some applications, people use the mixture model with continuous mixing distribution instead of finite mixing distribution.

1.3 Literature Reviews

1.3.1 Estimation in Mixture Models

Over the past years, a variety of methods have been developed for estimating the parameters in finite mixture models. Four of them are widely used in practice and cited in the literature, they are method of moments, minimum-distance method, maximum likelihood method and Bayesian method. The method of moments is the earliest method for estimating the parameters in finite mixture models. In Pearson's classic paper, he used this method for estimating five parameters in normal mixture models (1.4). Moment estimators enjoyed a wide application until computers were fast enough to find the maximum of the log-likelihood function. Based on the method of moments, several useful diagnostic tools had been developed for mixture models, see Lindsay (1989a, 1989b) and Lindsay and Roeder (1992a). Even today, the moment estimators still serve as useful initial values for iterative numerical methods such as EM algorithm to compute maximum likelihood estimates; see Lindsay (1995). Some developments about moment estimators can also be found in Lindsay and Basak (1993), Furman and Lindsay (1994a, 1994b), Lindsay (1995), Withers (1996) and Craigmile and Titterington (1998).

Another general method for estimating the mixing distribution Ψ in finite mixture model is to minimize the distance between the empirical distribution and the mixture distribution or the distance between the kernel density estimation and the mixture density. Titterington et al. (1985) gave a detailed review of the minimum-distance estimators.

Maximum likelihood estimator can also be viewed as a special case of minimum-distance estimators, simply because it minimizes the Kullback-Leibler (1951) distance between the empirical distribution and the mixture distribution.

Due to the rapid improvement in computing power, finding numerical solutions of a likelihood equation becomes feasible. Likelihood-based inference has enjoyed fast development and plays an important role in the scope of finite mixture models. Let $X_1, \ldots, X_n$ be a random sample from the mixture model $f(x; \Psi)$ in (1.2). The log-likelihood function of Ψ is given by

$$l_n(\Psi) = \sum_{i=1}^{n} \log f(X_i; \Psi) = \sum_{i=1}^{n} \log \Big\{ \sum_{j=1}^{m} \alpha_j f(X_i; \theta_j) \Big\}.$$

The maximum likelihood estimator (MLE) of Ψ is defined to be

$$\hat{\Psi} = \arg \max_{\Psi \in \Omega} l_n(\Psi)$$

when this exists. For finite mixture models, the explicit expression for the MLEs are typically not available, and a number of numerical algorithms have been developed for maximizing the log-likelihood function. Expectation-maximization (EM) algorithm of Dempster et al. (1977) is the most popular method for finding the MLE. This is because a finite mixture model is a special case of the model for incomplete data, to which the EM algorithm can be easily applied. In the following, we give some details on the EM algorithm for the mixture of two components. The general idea of EM algorithm is similar and can be easily applied to mixture models with more than two components.

Suppose $X_1, \ldots, X_n$ is a random sample coming from the mixture model:

$$(1 - \alpha) f(x; \theta_1) + \alpha f(x; \theta_2).$$

In the EM framework, the data can be viewed as incomplete since their associated component-indicator variables, $I_1, \ldots, I_n$, are missing, where $I_i = 1$ if X_i comes from $f(x; \theta_2)$ and

$I_i = 0$ if X_i comes from $f(x;\theta_1)$, $i = 1, \ldots, n$. From this point of view, the complete data are $\{(X_i, I_i), i = 1, \ldots, n\}$. Note that $I_1, \ldots, I_n$ can be seen as a random sample from distribution $Bernoulli(\alpha)$. Hence the complete log-likelihood is given by

$$l_n^c(\Psi) = \sum_{i=1}^{n}[(1 - I_i)\{\log(1-\alpha) + \log f(X_i;\theta_1)\} + I_i\{\log(\alpha) + \log f(X_i;\theta_2)\}], \qquad (1.5)$$

where $\Psi(\theta) = (1-\alpha)I(\theta_1 \leq \theta) + \alpha I(\theta_2 \leq \theta)$.

Let $\Psi^{(0)}$ be the value initially specified for Ψ. On the first iteration of the EM algorithm, the E-step requires the computation of the conditional expectation of $l_n^c(\Psi)$ given the data $X_1, \ldots, X_n$ and the initial value $\Psi^{(0)}$ for Ψ. That is,

$$Q(\Psi, \Psi^{(0)}) = E\{l_n^c(\Psi)|X_1, \ldots, X_n, \Psi^{(0)}\}.$$

In general, on the $k+1$ iteration, the E-step requires the calculation of $Q(\Psi, \Psi^{(k)})$, where $\Psi^{(k)}$ is the value of Ψ after the kth EM iteration. Since the complete-data log likelihood, $l_n^c(\Psi)$, is linear in the unobservable component indicator variables $I_i's$, the E-step (on the $k+1$ iteration) simply requires the calculation of the conditional expectations of $I_i's$ given the observed data $X_1, \ldots, X_n$. The conditional expectations are the posterior probabilities of $X_i's$ belonging to the second component. As in McLachlan and Peel (2000, P20), for the E-step, we update the conditional expectations or the posterior probabilities ($w_i^{(k)}{'s}$) as follows:

$$w_i^{(k)} = E(I_i|X_i;\Psi^{(k)}) = \frac{\alpha^{(k)}f(X_i;\theta_2^{(k)})}{(1-\alpha^{(k)})f(X_i;\theta_1^{(k)}) + \alpha^{(k)}f(X_i;\theta_2^{(k)})}, \ i = 1, \ldots, n.$$

In the M-step, on the $k+1$ iteration, we update Ψ by maximizing $Q(\Psi, \Psi^{(k)})$ with respect to $\Psi \in \Omega$. Let $\Psi^{(k+1)}$ be the updated value for Ψ. That is,

$$\Psi^{(k+1)} = \arg\max_{\Psi \in \Omega} Q(\Psi, \Psi^{(k)}).$$

Due to the convenient structure of function $Q(\Psi, \Psi^{(k)})$, the maximization can be accomplished as follows:

$$\begin{aligned}
\alpha^{(k+1)} &= \sum_{i=1}^{n} w_i^{(k)}/n, \\
\theta_1^{(k+1)} &= \arg\max_{\theta_1 \in \Theta} \sum_{i=1}^{n} \{(1 - w_i^{(k)}) \log f(X_i; \theta_1)\}, \\
\theta_2^{(k+1)} &= \arg\max_{\theta_2 \in \Theta} \sum_{i=1}^{n} \{w_i^{(k)} \log f(X_i; \theta_2)\}.
\end{aligned}$$

In many situations, the above optimization problem has explicit solutions. One such example is the mixture of normal distributions. In some other situations, it maybe hard to find a close form for Ψ in the M-step. In the so-called generalized EM (GEM) algorithm (Dempster et al., 1977), in the M-step, $\Psi^{(k+1)}$ is not necessarily required to be a maximum of $Q(\Psi, \Psi^{(k)})$, but a value that makes

$$Q(\Psi^{(k+1)}, \Psi^{(k)}) \geq Q(\Psi^{(k)}, \Psi^{(k)})$$

for $k = 0, 1, 2, \ldots$. Several generalizations, particularizations, and accelerated versions of the EM algorithm have been proposed; see McLachlan and Krishnan (1997) and references therein.

The E and M steps are iterated repeatedly until the difference

$$|l_n(\Psi^{(k+1)}) - l_n(\Psi^{(k)})|$$

changes by an arbitrarily small value. One nice property of the EM (or GEM) algorithm is the so-called monotonicity property. That is,

$$l_n(\Psi^{(k+1)}) \geq l_n(\Psi^{(k)})$$

for $k = 0, 1, 2, \ldots$. This property is the fundamental reason behind the local convergence of the algorithm under some very general conditions. See Wu (1983). In some applications, we often modify the likelihood function by adding a penalty $p(\alpha)$ to $l_n(\Psi)$ and try to find the maximum of $pl_n(\Psi) = l_n(\Psi) + p(\alpha)$. In this case, the idea of the EM algorithm is still applicable. We only need to make a minor modification in the M-step for updating α. That is,

$$\alpha^{(k+1)} = \arg\max_{\alpha \in [0,1]} \Big\{ (n - \sum_{i=1}^{n} w_i^{(k)}) \log(1-\alpha) + \sum_{i=1}^{n} w_i^{(k)} \log(\alpha) + p(\alpha) \Big\}.$$

With this adjustment, the EM algorithm retains the property (Dempster et al. 1977)

$$pl_n(\Psi^{(k+1)}) \geq pl_n(\Psi^{(k)})$$

for $k = 0, 1, 2, \ldots$.

Due to its monotonicity property, the sequence $\{l_n(\Psi^{(k)}) : k = 0, 1, \ldots\}$ must converge to a local mode when the log-likelihood function is bounded above. Denote the limit of the log-likelihood value by l_n^*. Dempster et al. (1977) showed that under some weak conditions on the kernel function, l_n^* is a local maximum of $l_n(\Psi)$ if the sequence is not trapped at some saddle point. Wu (1983) gave a more rigorous treatment of the convergence properties of the EM algorithm in a general setup. Further details about the convergence of EM algorithm can be found in the monograph of McLachlan and Krishnan (1997).

The EM algorithm was initially criticized that it did not automatically provide an estimate of the covariance matrix of the MLEs. A number of methods have been suggested for estimating the covariance matrix of the MLE of the mixing distribution Ψ. Most suggestions are based on the observed information matrix $I(\Psi)$ given by

$$I(\Psi) = -\partial^2 l_n(\Psi)/\partial\Psi\partial\Psi^T.$$

There are two typical methods to estimate or approximate the observed information matrix. The first method replaces Ψ by the MLE $\hat{\Psi}$ in the above formula. This method involves the calculation of the second derivative of the log-likelihood function with respect to Ψ. The second method extracts the observed information matrix from the complete-data log-likelihood. Especially, if $X_1, \dots, X_n$ are independent and identically distributed (I.I.D.), the approximation has a very simple form, which only contains the first derivative of the complete log-likelihood function. The details can be seen in McLachlan and Krishnan (1997) and McLachlan and Peel (2000).

In spite of its popularity in the application of finite mixture models, the ordinary MLE are not well defined or not consistent under many classes of important mixture models, such as the normal mixture model given in (1.4). Under this model, the MLE is not well defined since $l_n(\Psi) \to \infty$ by letting $\mu_1 = X_1$ and $\sigma_1^2 \to 0$ with other parameters fixed. Several remedies have been suggested in the literature to account for this case. Hathaway (1985) and Tan et al. (2006) discussed the use of the constrained MLE, Chen et al. (2007) investigated the properties of the penalized MLE. The Penalized MLE and the constrained MLE are shown to be strongly consistent and asymptotically efficient under normal mixture models in both mean and variance parameters.

The fourth method for estimating Ψ is the Bayesian method. Let $L_n(X_1, \dots, X_n|\Psi)$ be the likelihood function of Ψ. In the framework of the Bayesian approach, one needs to assume that a prior distribution $p(\Psi)$ on Ψ is available. Using Bayes' theorem, we can obtain the posterior density $p(\Psi|X_1, \dots, X_n)$, which is given by

$$p(\Psi|X_1, \dots, X_n) \propto L_n(X_1, \dots, X_n|\Psi)p(\Psi).$$

As summarized in Frühwirth-Schnatter (2006), there are two main reasons why people may be interested in using the Bayesian method in finite mixture models. Firstly, including a suitable prior distribution for Ψ in the framework of the Bayesian approach may avoid spurious modes when maximizing the log-likelihood function. The idea for the penalized MLE in Chen et al. (2007) can be seen as putting a proper prior distribution on the variance parameters. Secondly, when the posterior distribution for the unknown parameters is available, the Bayesian method can yield valid inference without relying on the asymptotic normality. As warned by McLachlan and Peel (2000, p.68), the asymptotic theory of the MLE can apply only when the sample size n is very large. Hence the second advantage of the Bayesian method become obvious when the sample size n is small.

Unfortunately, for the likelihood function $L_n(X_1, \ldots, X_n|\Psi)$, it is impossible to find the conjugate prior for Ψ, which means whatever prior $p(\Psi)$ we choose, the posterior distribution $p(\Psi|X_1, \ldots, X_n)$ may not belong to any tractable distribution family. This problem no longer poses serious obstacle to the application of Bayesian method after the widespread use of Markov Chain Monte Carlo (MCMC) methods. The main idea of Bayesian estimation using the MCMC methods followed Dempster et al. (1977) by realizing a mixture model is a special case of incomplete data problem with the missing component indicator variables $I_1, \ldots, I_n$. After introducing $I_1, \ldots, I_n$ or the data augmentation, the idea of Bayesian estimation was to estimate the augmented parameter $(I_1, \ldots, I_n, \Psi)$ by sampling from the complete-data posterior distribution $p(I_1, \ldots, I_n, \Psi|X_1, \ldots, X_n)$, which consists of two main steps. In the first step, given the component indicator variables $I_1, \ldots, I_n$, we are back in the complete-data Bayesian estimation. In many situations, we can simulate the parameter Ψ by using Gibbs sampling. In the second step, given the simulated parameter Ψ, for the ith observation, we can sample I_i based on the posterior probability. More details can be seen in Tanner and Wong (1987), Gelfand and Smith (1990) and Frühwirth-Schnatter (2006).

1.3.2 Nonparametric Maximum Likelihood Estimate and Local Mixture Model

The review in the last subsection about the estimation methods in finite mixture model are under the assumption that the order of finite mixture model is known. When the order of the mixture model is unknown or the mixing distribution is not discrete, one may employ a non-parametric assumption on the mixing distribution. A nonparametric maximum likelihood estimate (NPMLE) of Ψ is a distribution function which maximizes the log-likelihood function over all possible mixing distributions (Laird, 1978). The identifiability problems for the NPMLE of a mixing distribution have been studied by Teicher (1963), Barndorff-Nielsen (1965), Chandra (1977), Jewell (1992) and Lindsay and Roeder (1992b). With the identifiability, the NPMLE has many interesting properties, such as the consistency of the NPMLE under very general conditions (Kiefer and Woldowitz, 1956 and Leroux, 1992). The most important result, so called "fundamental theorem of the nonparametric maximum likelihood esitmation", is summarized in Lindsay (1995).

Part I. *Existence, discreteness and uniqueness.* There exists an NPMLE which is discrete with no more than h distinct support points, where h is the number of the distinct points in the data set. Further, fitted log-likelihood values, namely,

$$(\log f(X_1; \tilde{\Psi}), \ldots, \log f(X_n; \tilde{\Psi}))$$

are unique, where $\tilde{\Psi}$ is the NPMLE of Ψ. That is, even if two distributions both maximize the log-likelihood, the log-likelihood vectors are equal. The mathematical tool for proving this part is the convex geometry, see Lindsay (1983) and Marriott (2002) for details.

The second part of the fundamental theorem is on how to determine whether a given distribution function Ψ_0 is the NPMLE or not. A useful tool for this investigation is called the directional derivative. Given two mixing distributions Ψ_0 and Ψ_1, the directional derivative of $l_n(\Psi)$ at Ψ_0 towards Ψ_1 is defined to be

$$\begin{aligned} D(\Psi_1; \Psi_0) &= \lim_{\epsilon \to 0^+} \frac{l_n\{(1-\epsilon)\Psi_0 + \epsilon\Psi_1\} - l_n(\Psi_0)}{\epsilon} \\ &= \sum_{i=1}^{n} \frac{f(X_i; \Psi_1) - f(X_i; \Psi_0)}{f(X_i; \Psi_0)}. \end{aligned}$$

If Ψ_1 is a point mass function at θ, the gradient function is defined to be

$$D(\theta; \Psi_0) = \sum_{i=1}^{n} \frac{f(X_i; \theta) - f(X_i; \Psi_0)}{f(X_i; \Psi_0)}.$$

Intuitively, if Ψ_0 is the NPMLE, then $l_n(\Psi)$ can not increase in any direction starting from Ψ_0, hence the gradient function $D(\theta; \Psi_0)$ should be non-positive.

Part II. *Gradient characterization and support point properties.* The distribution function $\tilde{\Psi}$ is the NPMLE of Ψ if and only if

$$D(\theta; \tilde{\Psi}) \leq 0 \quad \forall\theta.$$

Further, the supports of $\tilde{\Psi}$ are contained in the set of θ such that $D(\theta; \tilde{\Psi})$=0.

The fundamental theorem has important applications. It provides the basis for the algorithms for computing the NPMLE of the mixing distribution. For the details of the computational methods for finding the NPMLE, the reader is referred to the monographs by Lindsay (1995) and Böhning (1999).

An alternative and useful tool for understanding and making inference on mixture model, which allows for unknown number of discrete components, or continuous mixing distributions, is the use of the local mixture model (Marriott, 2002). The idea behind the local mixture model is to assume that Ψ is close to a point mass function at some fixed point θ_0 and then approximate $f(x; \Psi)$ by

$$f(x; \Psi) \approx f(x; \theta_0) + \sum_{k=2}^{r} \lambda_k f^{(k)}(x; \theta_0),$$

where

$$f^{(k)}(x;\theta_0) = \frac{\partial^k}{\partial\theta^k} f(x;\theta_0).$$

Here $f(x;\theta_0) + \sum_{k=2}^{r} \lambda_k f^{(k)}(x;\theta_0)$ is called the local mixture model of $f(x;\Psi)$ with order r. By this approximation, the integral in (1.1) has been changed to a function with r parameters. Some traditional methods, such as the Bayesian method can be easily applied. This idea has been proved to be powerful and efficient in measurement error modeling (Marriott, 2003), in Bayesian prediction (Marriott, 2002), in lifetime data analysis and influence analysis (Critchley and Marriott, 2004). Further, the idea of the local mixture model has been applied to exponential family (Anaya-Izquierdo and Marriott, 2007a) and the scale dispersion mixture (Anaya-Izquierdo and Marriott, 2007b). Marriott (2007) studied several ways in which the local assumption about Ψ can be relaxed.

1.3.3 Order Selection in Finite Mixture Models

Testing the number of components or the order m in a mixture model is an important problem when the prior information on m is unavailable. In some applications, m is the crucial parameter under consideration. For example, in cluster analysis, m is the number of clusters contained in the data; in latent structure analysis, m is the number of latent classes required to provide a reasonable model; in genetic analysis, if a quantitative trait is determined by a simple gene with two alleles, $m = 2$ means that the mode of inheritance is dominant, whereas $m = 3$ or more means the mode of inheritance is additive or more complex in nature. In other applications, m determines the model complexity. We may be interested in determining how large m needs to describe the data adequately; for parsimony, we prefer the less complex model.

Due to its importance, accessing the number of components or order selection in a finite mixture model has attracted the attention of many statisticians. In the literature, there are at least five types of order selection methods: information-based methods, for example, Akaike information criterion (AIC) and Bayes information criterion (BIC) (Leroux, 1992); penalized distance-based methods (Chen and Kalbfleisch, 1996, James et al., 2001 and Woo and Sriram, 2006); penalized log-likelihood based method (Chen and Khalili, 2006); Bayesian method (Carlin and Chib, 1995, Richardson and Green, 1997, Gruet, et al., 1999, Stephens, 2000, Ishwaran et al., 2001) and testing hypothesis-based method (Chen, 1998, Dacunha-Castelle and Gassiat, 1999, Chen et al., 2001, Chen and Chen, 2003, Liu and Shao, 2003, Chen et al., 2004, Charnigo and Sun, 2004, Chen and Kalbfleisch, 2005).

The AIC (Akaike, 1973) and the BIC (Schwarz, 1978) were first designed for the model selection problems. The AIC criterion aims to minimize the Kullback-Leibler distance between the true distribution and the distributions for the candidate models, while the BIC criterion tries to maximize the posterior probability in the space of all candidate models. For the order selection problem in finite mixture models, the AIC criterion selects the order to minimize

$$-2l_n(\hat{\Psi}) + 2d$$

and BIC criterion selects the order which minimizes

$$-2l_n(\hat{\Psi}) + d\log n,$$

where $\hat{\Psi}$ is the MLE of Ψ and d is the number of free parameters in Ψ under the given order, respectively. Under some regularity conditions, the order selected by the BIC is consistent (Keribin, 2000). The AIC and the BIC are asymptotically optimal under some criteria for the regular models. However, due to the non-regularity of the finite mixture model, these optimality properties do not hold in the framework of the mixture model. For this reason, several other information based criteria have been developed, for example, bootstrap-based information criterion (Ishiguro et al, 1997) and cross-validation-based information criterion (Smyth, 2000).

The penalized distance method is another popular method for the order selection in finite mixture models. There are several distance-based methods proposed in the literature. Let us reviewed three of them, other methods use the similar ideas. Chen and Kalbfleisch (1996) suggested choosing the order which minimizes the penalized distance between the empirical distribution function and the fitted cumulative distribution function with given order. They showed that under some conditions, the estimated order is strongly consistent for the true order. James et al. (2001) consider the order selection problem in normal mixture model in both mean and variance parameters. They proposed choosing the order by minimizing the penalized Kullback-Leibler distance between two density functions: the kernel density function of the true density function and the convolution of a normal density function and the density of finite normal mixtures with given order. They showed under certain conditions, the order estimator is consistent. Woo and Sriram (2006) considered choosing the order which minimizes the penalized Hellinger distance between the nonparametric kernel density function and the density function of finite mixture with given order. Their simulation results showed that the order estimation is robust to model assumptions.

The idea of the penalized log-likelihood based method (Chen and Khalili, 2006) is similar to the penalized distance method. The innovative part is the penalty functions added to the log-likelihood function. Chen and Khalili (2006) suggested putting two penalty functions to prevent the two types of over-fitting. The first penalty is to prevent any of the mixing proportions getting too close to 0. The second penalty function is to prevent fitting a model containing several sub-populations which only differ slightly. Under some regularity conditions, the estimated order is strongly consistent. The advantage of this method is that it does not need to compare all the candidate models. By choosing a suitable tuning parameter, the method can select the order automatically. How to determine the suitable tuning parameter in a computation-easy way is still under investigation.

In the literature, there are two types of Bayesian methods considering the order selection for finite mixture model. The first type of method is to apply trans-dimensional MCMC to sample from the joint posterior density $p(m, \Psi_m|X_1, \ldots, X_n)$, where Ψ_m the mixing distribution with m components. The major difficulty is that the number of parameters is not fixed when sampling from the posterior distribution. There are three popular methods in this category, which are product-space MCMC (Carlin and Chib, 1995), reversible jump MCMC (Richardson and Green, 1997) and birth and death MCMC (Stephens, 2000). The second type of method is to compute the marginal density $p(X_1, \ldots, X_n|m)$ for all possible orders and further to apply the Bayes' rule to quantify the posterior evidence for each order. The challenge is the computation of the marginal density. A lot of papers have been contributed to the approximation and the estimation of the marginal density, for example, sampling based approximation and density ratio based estimation, see Frühwirth-Schnatter (2004) for detailed reviews and comparisons. Based on the decomposition of the marginal density function, Ishwaran et al. (2001) proposed a weighted Bayes factor method, which can consistently estimate the order of the finite mixture model.

The fifth method for the order selection is the testing hypothesis-based method. In the past a few decades, statisticians devoted substantial effort to the hypothesis testing problem in finite mixture models. One popular method for testing the number of components is the $C(\alpha)$ test proposed by Neyman and Scott (1966). The $C(\alpha)$ test is a score test of homogeneity against heterogeneity. One advantage for this method is that the test statistics is not affected any specific mixture alternatives, and its limiting distribution is distribution free. When the kernel function $f(x;\theta)$ comes from the exponential family, the $C(\alpha)$ test simply measures the difference between the sample variance and the theoretical variance under the null model (Lindsay, 1995 and Anaya-Izquiordo and Marriott, 2007b). The $C(\alpha)$ method is also known to be asymptotically best in the sense of best power under local alternatives, but it is not efficient for detecting non-local alternatives (Chen, 1998). Here, the local alternative, intuitively, means that the distance between the null model and the alternative model is small. That is, some components have mixing proportions close to 0 or 1, and other components have component parameters close to each other.

The method of moments is not only used for estimating the unknown parameters in a finite mixture model, but it has also been applied for testing the number of components. Lindsay (1989b) suggested a statistic for testing the order of the finite mixture model based on the determinant of the moment matrix (Lindsay, 1989a) of the mixing distribution Ψ. For testing homogeneity in the mixture models, the suggested statistic is equivalent to $C(\alpha)$ test statistic for distributions in the exponential family (Lindsay, 1989b). The same idea was also used to test the order in the mixture of exponentials (Heckman et al., 1990). As Lindsay (1989b) pointed out this idea could not be directly used for testing the order of the finite mixture models in the presence of a structural parameter, for example, normal mixture models with same and unknown variance for each component.

The likelihood ratio test (LRT) is the most extensively used method in the parametric hypothesis test. The LRT statistic has a chi-squared null limiting distribution not depending on the true distribution for regular models and this property makes it easy to use. Due to the non-regularity of the finite mixture models, the large sample property of likelihood-based tests was an enigma until Ghosh and Sen (1985) and Hartigan (1985). For the test of homogeneity, the limiting distribution of the LRT statistic often involves a Guassian process, see Chen and Chen (2001), Dacunha-Castelle and Gassiat (1999), and Liu and Shao (2003). The percentiles of such a statistic is very hard to determine and hence the result on LRT is hard to implement in practice (Alder, 1990 and Sun, 1993).

The modified likelihood ratio test (MLRT) proposed in Chen (1998), Chen et al. (2001, 2004) and Chen and Kalbfleisch (2005), restores the simplicity of the likelihood based test by adding a penalty function on the mixing proportions. The limiting distribution of the MLRT statistic is chi-squared or a mixture of chi-squared distributions for a large variety of mixture models. The modified likelihood method has the advantage of giving a natural and general approach to testing problems in finite mixture models. The MLRT is asymptotically locally most powerful. Simulation indicates that the MLRT is more efficient compared to $C(\alpha)$ test when the Kullback-Leibler distance between the null model and the alternative model increases. Due to the penalty function, the MLRT may be inefficient when one of the mixing proportions is close to 0 or 1.

Charnigo and Sun (2004) proposed another class of D-test for testing the order of mixture models. The D-test statistic measures some L^2 distance between the best fitted uni-component model and the alternative model. This method enjoys the advantage that the statistic has a closed-form expression in terms of parameter estimators for a large class of kernel densities. In addition, a weighting function can be used to achieve high power of the D-test. The percentiles of the test statistic are usually obtained via computer simulation. Simulation studies in Charnigo and Sun (2004) suggested that the D-test and the MLRT are competitive.

1.4 Main Contributions and the Presentation of the Thesis

In some applications, prior information on the order of the finite mixture is known, and our task is reduced to select the order out of a few competitors. The most important albeit also most simple example is to choose between a homogeneous model vs a mixture model with two components. In these applications, it is often sensible to conduct a hypothesis test with the homogeneous model as the null model. For example, in classical statistical genetics, it is often suspected that only a subgroup of patients have a disease gene which is linked to the marker. Detecting the existence of this subgroup amounts to the rejection of a homogeneous null model in favour of a two component mixture model (Ott, 1999). In the literature testing for homogeneity, or specifically, testing for homogeneous model against an alternative of mixture of two components, has attracted substantial research recently. The objective of the thesis is to make substantial contributions in this area of research.

The classical LRT is a favored method in general, its application to test of homogeneity is not successful due to its inconvenient limiting distribution. What complicates the null limiting distribution of the LRT is the partial loss of identifiability of mixture models. The MLRT restores the identifiability by introducing a penalty function to the log-likelihood function. Under some conditions, the limiting distribution of the MLRT is $1/2\chi_0^2 + 1/2\chi_1^2$, which is convenient to implement in real data analysis. The choice of the penalty functions in the MLRT is very flexible. A good choice of the penalty enhances the power of the MLRT. Chen et al. (2001) suggested the use of penalty function $p(\alpha) = C \log\{4\alpha(1-\alpha)\}$ with some positive constant C, where $1-\alpha$ and α are the mixing proportions for the first and second components, respectively. A natural question is "can we find a new penalty function which can improve the approximation of the limiting distribution while retaining or improving the efficiency of the test under finite sample size". Chapter 2 contributes to the answer of this question. After detailed reviews of the LRT and the MLRT, the new penalty function for the MLRT is suggested. Extensive simulations based on Poisson mixtures, Binomial mixtures and Normal mixtures with known variance are conducted to compare the performance of the two penalty functions. We find that the MLRT with the new penalty function enjoys a significantly improved power for testing of homogeneity when α is close to 0 or 1.

The main contribution of the thesis is to propose a new class of methods for testing of homogeneity. Most existing methods, explicitly or implicitly, are derived under the condition of finite Fisher information and a compactness assumption on the component parameter space (Ghosh and Sen, 1985, Chen and Chen, 2001, Chen et al., 2001, Dacunha-Castelle and Gassiat, 1999, Liu and Shao, 2003 and, Charnigo and Sun, 2004). The finite Fisher information condition can prevent their usage to many important mixture models. Two typical examples are the mixture of geometric distributions and the mixture of exponential distributions. The compactness assumption often forces applicants to set artificial bounds for the parameters of interest and makes the resulting limiting distribution dependent on these bounds. Many researchers in this area dream to develop a test method without such restrictions. In Chapter 3, a new class of methods, called EM-test, are proposed, which are shown to be free of all these shortcomings. In this chapter, we first give more details for our motivation by two simple and illustrative examples. Then the EM-test is described, which is found to inherit the advantages of the classic LRT and score test (Liang and Rathouz, 1999). The asymptotic properties of the EM-test are studied for single parameter mixture models. It has a simple limiting distribution $0.5\chi_0^2+0.5\chi_1^2$, the same as the MLRT. This result is applicable to mixture models without requiring the restrictive regularity conditions described earlier. The adjustment of the limiting distribution is suggested for the finite sample size. The Edgeworth expansion is used to approximate the non-zero proportion of the EM-test statistic. The performance of the new method is compared with the MLRT, the constrained LRT (Chen and Cheng, 1995 and Lemdani and Pons, 1995) and the D-test (Charnigo and Sun, 2004) by extensive simulation. The new method works well compared with these methods. We also illustrate the use of the EM-test by analyzing a number of real data sets.

The normal mixture model is probably the most popular model in applications. Chen and Chen (2003) noted that normal distribution does not satisfy the strong identifiability condition, which imposes substantial technical difficulties in the study of the asymptotic properties. For example, the convergence rate of the MLE of the mixing distribution is $O_p(n^{-1/8})$ instead of $O_p(n^{-1/4})$, the optimal rate for the single parameter mixture models (Chen, 1995). Most existing methods can not directly apply to normal mixture models. Their asymptotic properties have to be developed separately. In Chapter 4, after reviewing the developments of the recent researchers on the homogeneity test in normal mixtures, the use of the EM-test is investigated and its limiting distributions are derived. For the homogeneity test in the presence of the structural parameter, a penalty function is suggested to overcome the underestimation effect of the variance parameter. The limiting distribution is a simple function of $0.5\chi_0^2 + 0.5\chi_1^2$ and χ_1^2 distributions. The test with this limiting distribution is still very convenient to implement. For normal mixtures in both mean and variance parameters, a penalty on the variance parameters is added to avoid the unboundedness of the log-likelihood function. The limiting distribution of the EM-test is found to be χ_2^2. Simulations are conducted to check the precision of the approximation of the limiting distributions to the finite sample distribution of the EM-test statistics. A real example is used to illustrate the use of the EM-test.

Circular data arise in many disciplines, including astronomy, biology, ecology, geology and medicine. As a circular analog of the normal distribution on the real line, the von Mises distribution is called the circular normal distribution. It is the most commonly used distribution for circular data, see Mardia and Jupp (2000) for its general properties. Similar to the normal mixture model for the linear data, the mixture of von Mises distributions is often used to model the heterogeneity in the circular data. Interestingly, it satisfies the strong identifiability condition and its mean space is compact. However the theoretical results in the single parameter mixture model can not directly apply to the mixture of von Mises distributions. In Chapter 5, we study the application of the MLRT and the EM-test to von Mises mixture models in the presence of a structural parameter. At the beginning of this chapter, a general introduction about the von Mises distribution and von Mises mixture is given, which is followed by the circular moment properties for the mixture of two von Mises distributions. After that the asymptotic property of the LRT, the MLRT and the EM-test in the von Mises mixture model are studied. A penalty function is introduced to prevent the overestimation effect of the structural parameter, which significantly enhances the precision of the MLRT test and the EM-test. Two real data sets in Grimshaw et al. (2001) are used to illustrate the idea of these two tests.

Chapter 6 concludes the thesis and discuss additional problems for future research.

Chapter 2

More Effective Penalty Function for the Modified Likelihood Ratio Test

2.1 Introduction

The modified likelihood ratio test (MLRT) was first introduced in Chen (1998) for testing homogeneity in mixture of multinomial distributions. In Chen et al. (2001), the MLRT was generalized to the mixture of general kernel functions. Since then, the MLRT has been applied to a number of more general testing problems in finite mixture models, for example, Chen et al. (2004) used the MLRT to test the null hypothesis of $m = 2$ against $m \geq 3$; Chen and Kalbfleisch (2005) applied MLRT to test of homogeneity at the presence of the structural parameter; Fu et al. (2007) studied the use of the MLRT to mixture models for directional data.

For regular parametric models, the likelihood ratio test (LRT) enjoys a very simple limiting distribution and high efficiency. Theory for the likelihood ratio test in finite mixture models enjoyed a fast development in the past years since the work of Ghosh and Sen (1985) and Hartigan (1985). It is one of the most discussed method for testing homogeneity in the literature. However due to the non-regularity of the finite mixture models, the limiting distribution of the LRT often involves the supremum of Gaussian process and hence the test is hard to implement because the computation of the asymptotic quantiles for the LRT statistic is challenging (Adler, 1990 and Sun, 1993). To overcome this difficulty, Chen (1998) and others proposed adding a penalty function to the log-likelihood function. The resulting MLRT has a simple limiting distribution for a large number of mixture models and it is easy to use in applications. The result is valid for a large variety of penalty functions. The choice of a good penalty function is an important research problem. In this chapter, we propose the use of a new class of penalty functions. We show that the new penalty function significantly improves the power of the MLRT for testing of homogeneity when α is close to 0 or 1.

2.2 The LRT and the MLRT

To discuss the new penalty functions, let us consider the use of the LRT and the MLRT for testing homogeneity under the single parameter mixture models. Let $X_1, \ldots, X_n$ be a random sample from the mixture density:

$$f(x;\Psi) = (1-\alpha)f(x;\theta_1) + \alpha f(x;\theta_2), \tag{2.1}$$

where $\Psi(\theta) = (1-\alpha)I(\theta_1 \le \theta) + \alpha I(\theta_2 \le \theta)$, $\theta_i \in \Theta$, $i = 1, 2$ and $0 \le \alpha \le 1$. We wish to test

$$H_0 : \alpha(1-\alpha)(\theta_1 - \theta_2) = 0. \tag{2.2}$$

We assume Θ is a compact subset of real line. The log-likelihood function is given by

$$l_n(\alpha, \theta_1, \theta_2) = \sum_{i=1}^{n} \log\{(1-\alpha)f(X_i;\theta_1) + \alpha f(X_i;\theta_2)\}. \tag{2.3}$$

Let $(\hat{\alpha}, \hat{\theta}_1, \hat{\theta}_2)$ be the MLE of $(\alpha, \theta_1, \theta_2)$ under the full model and $\hat{\theta}_0$ maximize $l_n(1/2, \theta, \theta)$ under the null model. The likelihood ratio test (LRT) statistic is defined to be

$$R_n = 2\{l_n(\hat{\alpha}, \hat{\theta}_1, \hat{\theta}_2) - l_n(1/2, \hat{\theta}_0, \hat{\theta}_0)\}$$

and the LRT rejects the null hypothesis when R_n is large.

Chen and Chen (2001) studied the large sample behavior of the LRT using the so called sandwich method. Dacunha-Castelle and Gassiat (1999) and Liu and Shao (2003) derived the limiting distribution of the LRT using a parameter transformation technique. Under the regularity conditions listed in Section 2.5 at the end of this chapter, the limiting distribution is given by the following theorem.

Theorem 2.2.1. *If Conditions A1-A5 in Section 2.5 hold, the limiting distribution of R_n under null model $f(x; \theta_0)$ is that of*

$$\sup_{\theta \in \Theta} \{W^+(\theta)\}^2,$$

where $W(\theta)$ is a Gaussian process with mean 0, variance 1 and the autocorrelation function

$$\rho(\theta, \theta') = \frac{cov\{Z_i(\theta) - h(\theta)Y_i, Z_i(\theta') - h(\theta')Y_i\}}{\sqrt{var\{Z_i(\theta) - h(\theta)Y_i\} var\{Z_i(\theta') - h(\theta')Y_i\}}}.$$

Here

$$Y_i(\theta) = \frac{f(X_i; \theta) - f(X_i; \theta_0)}{(\theta - \theta_0) f(X_i; \theta_0)}, \ \theta \neq \theta_0; \ Y_i = Y_i(\theta_0) = \frac{f'(X_i; \theta_0)}{f(X_i; \theta_0)}, \tag{2.4}$$

$$Z_i(\theta) = \frac{Y_i(\theta) - Y_i(\theta_0)}{(\theta - \theta_0)}, \ \theta \neq \theta_0; \ Z_i = Z_i(\theta_0) = \frac{f''(X_i; \theta_0)}{2 f(X_i; \theta_0)}, \tag{2.5}$$

and $h(\theta) = E\{Y_i Z_i(\theta)\}/E(Y_i^2)$.

We emphasize here that the Conditions A1-A5 include the compact parameter space assumption and the finite Fisher information condition. If the parameter space is not compact, R_n may go to infinity in probability as $n \to \infty$ (Hartigan, 1985, Bickel and

Chernoff, 1993 and Liu and Shao, 2004). It is seen that the correlation function $\rho(\theta, \theta')$ in Theorem 2.2.1 is meaningful only when $E\{Z_i^2(\theta)\} < \infty$. Note that $E\{Z_i^2(\theta)\} < \infty$ implies that the Fisher information at $\alpha = 0$ and $\theta_1 = \theta_0$ or the second moment of the centered density ratio $f(x;\theta)/f(x;\theta_0) - 1$ is finite. More discussions will be given in Chapter 3.

When the kernel function $f(x;\theta) = (\theta^x/x!)\exp\{-\theta x\}$ is the probability mass function of the Poisson distribution, the correlation function is

$$\rho(\theta, \theta') = \frac{e^{vv'} - 1 - vv'}{\sqrt{(e^{v^2} - 1 - v^2)(e^{v'^2} - 1 - v'^2)}},$$

where $v = (\theta - \theta_0)/\sqrt{\theta_0}$ and $v' = (\theta' - \theta_0)/\sqrt{\theta_0}$. See Chen and Chen (2001) for additional examples of the correlation function. Typically, the limiting distribution is used to approximate the critical values of the corresponding test. In this example, we must compute the quantiles of the supremum of the above truncated Gaussian process to obtain the critical values, which is not an easy task (Adler, 1990 and Sun, 1993). Ghosh and Sen (1985) suggested a discretization method to approximate the p-value of the LRT statistic. It is to select an appropriate sequence of $\theta_i \in \Theta$, and approximate the limiting distribution by that of $\max_i\{W^+(\theta_i)\}^2$. This idea may have the following two disadvantages in applications (Chen et al., 2001). Firstly, the approximation may depend on the choice of θ_i's. Secondly, the asymptotic distribution of the LRT under the null model depends on the kernel function and the true value of Ψ. So for different kernel functions and different values of Ψ, we need to do different approximations.

To overcome the difficulty of the LRT, new tests with simple limiting distributions, and similar efficiency are sought by researchers in this area. The MLRT is one of most satisfying solutions.

As pointed out by Chen et al. (2001) and Anaya-Izquierdo and Marriott (2007a), there are two sources of non-regularity which complicate the asymptotic null distribution of the LRT:

(1) the null hypothesis lies on the boundary of parameter space $\alpha = 0$ or $\alpha = 1$;

(2) the mixture model is not identifiable under null model, that is, $\alpha = 0$, $\alpha = 1$ and $\theta_1 = \theta_2$ are equivalent.

The MLRT overcomes the boundary problem and non-identifiability by adding a regularity-restoring penalty function. Let

$$pl_n(\alpha, \theta_1, \theta_2) \quad = \quad l_n(\alpha, \theta_1, \theta_2) + p(\alpha)$$

such that $p(\alpha) \to -\infty$ as $\alpha \to 0$ or 1 and $p(\alpha)$ achieves its maximal value at $\alpha = 0.5$.

A main cause of non-regularity of mixture models is the possibility of $\alpha = 0$ or 1. Because of the penalty function $p(\alpha)$, the fitted value of α under the modified likelihood is bounded away from 0 and 1. Thus, the penalty has effectively placed a soft constraint on α. Let $\tilde{\alpha}$, $\tilde{\theta}_1$ and $\tilde{\theta}_2$ maximize $pl_n(\alpha, \theta_1, \theta_2)$ under the full model and $\tilde{\theta}_0$ maximize $pl_n(1/2, \theta, \theta)$. The modified likelihood ratio test statistic is then defined to be

$$M_n = 2\{pl_n(\tilde{\alpha}, \tilde{\theta}_1, \tilde{\theta}_2) - pl_n(1/2, \tilde{\theta}_0, \tilde{\theta}_0)\}.$$

The MLRT is asymptotically distribution-free and can be conveniently implemented.

Theorem 2.2.2. *If Conditions A1-A5 in Section 2.5 hold, then the asymptotic null distribution of the MLRT statistic M_n is the mixture of χ^2_1 and χ^2_0 with the same weights, i.e.*

$$0.5\chi^2_0 + 0.5\chi^2_1,$$

where χ^2_0 is a degenerate distribution with all its mass at 0.

By Theorem 2.2.1, we have $R_n = O_p(1)$. Since $0 \le M_n \le R_n$, we get

$$-R_n \le 2\{p(\tilde{\alpha}) - p(0.5)\} \le 0.$$

Therefore, $p(\tilde{\alpha}) = O_p(1)$, which implies that $\tilde{\alpha}$ is bounded away from 0 and 1 with probability approaching 1. Hence it follows that $\tilde{\theta}_1$ and $\tilde{\theta}_2$ must converge to the true value θ_0 under null model. The Taylor's expansion at $\theta_1 = \theta_0$ and $\theta_2 = \theta_0$ is then used to find the quadratic approximation of the modified likelihood ratio statistic and typical technique of quadratic approximation is further applied to yield the desired result. For more details, see Chen et al. (2000). The purpose of the penalty function $p(\alpha)$ is to bound the fitting of α away from 0 or 1, which is guaranteed under the finite Fisher information condition and the compact parameter space assumption. If these two conditions are not satisfied, the LRT statistic R_n may go to infinity in probability as $n \to \infty$, then the penalty function $p(\alpha)$ may not fully fulfill its purpose and the simple limiting distribution may not be applicable.

2.3 More Effective Penalty Function

According to Theorem 2.2.2, the limiting distribution of the MLRT does not depend on the specific form of $p(\alpha)$, but the precision of the approximation and its power do. Chen et al. (2001) suggested the use of penalty function

$$p(\alpha) = C \log\{4\alpha(1-\alpha)\} \tag{2.6}$$

with some positive constant C. For a number of mixture models such as Binomial, Poisson and Normal in mean, Chen et al. (2001) suggested using $C = \log(M)$ if the parameters θ_1 and θ_2 are restricted to $[-M, M]$. For example, for a Poisson mixture, if $\theta_i \in [0, 50]$, i =1, 2 we let $C = \log(50)$. In these cases, the influence of C on the type I error or the power is minor for C within some appropriate range.

Now we motivate the new penalty function by analyzing two simulated data sets from Poisson mixture models. We let the sample size be 200. The probability mass function of the first Poisson mixture model is

$$0.05\text{Pois}(0.127) + 0.95\text{Pois}(5.256)$$

and the second one is

$$0.1\text{Pois}(1.646) + 0.9\text{Pois}(5.373),$$

where Pois(θ) denotes the Poisson probability mass function with mean θ. These two data sets are given in Table 2.1.

Table 2.1: Simulated data sets from Poisson mixture models.

	Observed values											
	0	1	2	3	4	5	6	7	8	9	10	11
	Frequency											
Dataset I	7	9	10	27	32	40	30	20	11	6	8	0
Dataset II	4	11	16	22	28	28	33	33	14	5	3	3

We first compute the MLRT statistics with the penalty function in (2.6) with $C = \log(50)$ for the above two data sets. For the first data set, we find

$$(\tilde{\alpha}, \tilde{\theta}_1, \tilde{\theta}_2) = (0.919, 0.743, 5.185) \text{ and } M_n = 0.881$$

and for the second data set, we find

$$(\tilde{\alpha}, \tilde{\theta}_1, \tilde{\theta}_2) = (0.791, 2.751, 5.615) \text{ and } M_n = 0.960.$$

In view of the limiting distribution $0.5\chi_0^2+0.5\chi_1^2$, the asymptotic p-values for the above two MLRT statistics are 0.174 and 0.164, respectively. Thus, we do not have strong evidence to reject the uni-component Poisson distribution hypothesis in both cases. One reason is that the test statistic puts too much penalty on the mixing proportion when it is close to

0 or 1. For data set I, the penalty term contributes -9.479 to M_n and for data set II, the penalty term contributes -3.236 to M_n. It appears that the penalty in both cases played a heavy role, hence the power of the test is severely affected. Can we find a new penalty function, based on which the null critical values of MLRT is still close to the theoretical values and the power of MLRT increases even when the mixing proportion is close to 0 and 1?

For the validity of the asymptotic result, $p(\alpha)$ must decrease to negative infinity when $\alpha \to 0$ or 1. The only other considerations for the choice of $p(\alpha)$ are computational convenience and statistical efficiency. We suggest the following penalty function

$$p(\alpha) = C^* \log(1 - |1 - 2\alpha|) \tag{2.7}$$

for some positive C^*. Our arguments are as follows. Firstly, we find

$$\log(1 - |1 - 2\alpha|) \leq \log(1 - |1 - 2\alpha|^2) = \log\{4\alpha(1 - \alpha)\}.$$

Thus, when $C = C^*$, the penalty (2.7) is more severe than the penalty (2.6) which enhances the accuracy of the approximation. Secondly, when $\alpha \approx 0.5$, $\log(1 - |1 - 2\alpha|) \approx -|1 - 2\alpha|$. So the penalty (2.7) is a Lasso-type penalty (Tibshirani, 1996), that is, it is a continuous function for all α, but not smooth at $\alpha = 0.5$. Therefore the new penalty function has similar properties to the Lasso-type penalty for linear regression (Tibshirani, 1996); the probability of the fitted value of α being 0.5 is positive. In comparison, the penalty $\log\{4\alpha(1 - \alpha)\}$ is smooth at $\alpha = 0.5$ and does not have this property. In conclusion, the new penalty function can achieve the same precision under the null distribution at a lower level of modification C^*. Hence, its usage may lead to higher power.

In the simulation study to be presented, we found the MLRT statistics constructed from the penalty functions in (2.6) with $C = \log(50)$ and (2.7) with $C^* = 1$ have the similar null rejection rates for Poisson mixtures and Binomial mixtures. Figure 2.1 compares

$\log(1-|1-2\alpha|)$ and $\log(50)\log\{4\alpha(1-\alpha)\}$. These two penalty functions are almost the same when α is close to 0.5 and quite different when α is close to 0 and 1. Hence the MLRT constructed from the penalty function (2.7) is expected to have larger power when α is close to 0 and 1.

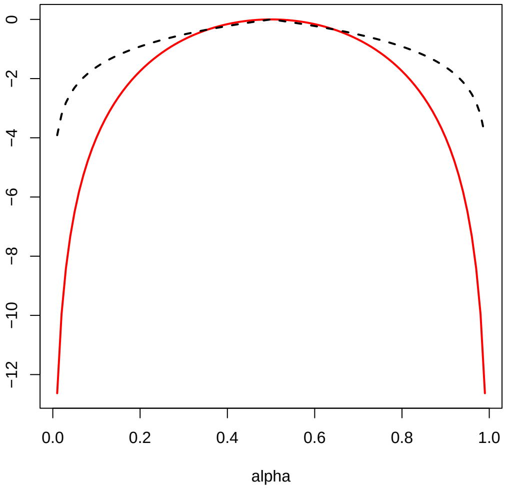

Figure 2.1: Comparisons of $\log(50)\log\{4\alpha(1-\alpha)\}$ (solid line) and $\log(1-|1-2\alpha|)$ (dashed line).

Remark 2.3.1. *The two penalty functions are special cases of* $C\log(1-|1-2\alpha|^h)$ *for some* $0<h\leq 2$*. A choice of* $0<h<1$ *may further improve the power of the MLRT. We still recommend the choice of* $h=1$ *due to the following reasons. Firstly, when* $h=1$*, for the EM algorithm introduced in Chapter 1,* α *values can be easily updated in M-step as follows:*

$$\alpha^{(k+1)} = \begin{cases} \min\left\{\frac{\sum_{i=1}^n w_i^{(k)}+C}{n+C}, 0.5\right\}, & \text{if } \frac{\sum_{i=1}^n w_i^{(k)}}{n} < 0.5 \\ 0.5, & \text{if } \frac{\sum_{i=1}^n w_i^{(k)}}{n} = 0.5 \\ \max\left\{\frac{\sum_{i=1}^n w_i^{(k)}}{n+C}, 0.5\right\}, & \text{if } \frac{\sum_{i=1}^n w_i^{(k)}}{n} > 0.5 \end{cases}$$

for $k = 0, 1, 2 \ldots$. *Secondly, there is a natural generalization of the current penalty function to the hypothesis testing problem with more than two components. Note that* $\log(1 - |1 - 2\alpha|) = \min[\log(2\alpha), \log\{2(1-\alpha)\}]$. *For the mixture model with* m *components, the penalty function can be set as* $\min\{\log(\alpha_1), \ldots, \log(\alpha_m)\}$. *However, when* $h < 1$, *the penalty function loses the above two properties.*

Remark 2.3.2. *Our final comment concerns the choice of the level of modification* C^*. *A natural choice is* $C^* = 1$. *Although a specific reason of choosing* $C^* = 1$ *is lacking, there is ample evidences that it is a very sensible choice in a wide range of applications. In simulation study,* $C^* = 1$ *works quite well for mixture of Poisson kernels, Binomial kernels and normal kernels with known variance. In each new application, we recommend a pilot simulation study to ensure that* C^* *is chosen such that the simulated type I errors are no more than 5.5% when the target is 5%.*

2.4 Simulation Study

The purpose of this simulation study is to compare the power of the MLRT statistics constructed using the penalty functions in (2.6) and (2.7). In the literature, there are many others methods for testing homogeneity, such as the $C(\alpha)$ test and the D-test. We do not include the comparison of the MLRT and the $C(\alpha)$ test, since this has been done in Chen (1998) and Chen et al. (2001). We also do not include the comparison of the MLRT and the D-test because this will be investigated in Chapter 3.

For convenience of presentation, let M_n and M_n^* denote the MLRT statistics with penalty functions in (2.6) and (2.7), respectively. The simulation experiment was conducted under the Poisson and Binomial kernels with size 10, and the Normal kernel with a known variance of 1. The mean values for the null distribution and alternative distributions for

Poisson kernel and Binomial kernel are 5, and for normal kernel are 0. Four alternative models are selected for each kernel as follows: we set $1 - \alpha = 0.5, 0.25, 0.1, 0.05$ and the variances for the alternative models are set to be 1.25 times the variance under null model. Under this setup, we have for the Poisson mixture,

$$\alpha(1-\alpha)(\theta_1 - \theta_2)^2 = 1.25,$$

for the Binomial mixture,

$$\alpha(1-\alpha)(\theta_1 - \theta_2)^2 = 1/144$$

and for the Normal mixture,

$$\alpha(1-\alpha)(\theta_1 - \theta_2)^2 = 1/4.$$

Together we have twelve alternative models and their parameter values are summarized in Table 2.2. We also provide the Kullback-Leibler information of these models with respect to the corresponding null models, namely,

$$KL(f, g) = E_f[\log\{f(X)/g(X)\}].$$

The M_n^* values were computed with $C^* = 1$ for all kernels, and M_n values were computed with $C = \log(50)$ for Poisson and Binomial kernels and with $C = \log(10)$ for the normal kernel, as suggested in Chen et al. (2001). In each simulation run, three levels; 10%, 5% and 1%, and two sample sizes 100 and 200 were considered. The null rejection rates were calculated based on 20,000 repetitions and the powers were computed based on 10,000 repetitions. Table 2.3 summarizes the rejection rates under the null models. Tables 2.4, 2.5 and 2.6 present the power comparisons under the alternative models. The simulation results show that when α is close to 0 or 1, for example, when $1 - \alpha = 0.05$, the new penalty function significantly improves the power of the MLRT. When M_n^* is applied to the two simulated data sets, for data set I,

$$(\tilde{\alpha}^*, \tilde{\theta}_1^*, \tilde{\theta}_2^*) = (0.947, 0.460, 5.128) \text{ and } M_n^* = 7.738$$

and for the second data set,

$$(\tilde{\alpha}^*, \tilde{\theta}_1^*, \tilde{\theta}_2^*) = (0.902, 1.653, 5.402) \text{ and } M_n^* = 4.176,$$

which gives strong evidence to reject the null hypothesis.

2.5 Appendix: Regularity Conditions

The following regularity conditions on the kernel density function are used in obtaining the asymptotic properties of the LRT and the MLRT.

A1. *Compact parameter space.* Θ is a compact subset of the real line.

A2. *Wald's integrability conditions.* (i) $E|\log f(X;\theta_0)| < \infty$, and (ii) for sufficiently small ρ and for sufficiently large r, the expected values $E\log\{1+f(X;\theta,\rho)\} < \infty$ for $\theta \in \Theta$ and $E\log\{1+\varphi(X,r)\} < \infty$, where

$$f(x;\theta,\rho) = \sup_{|\theta'-\theta|\le\rho} f(x;\theta')$$

and

$$\varphi(x;r) = \sup_{|\theta|\ge r} f(x;\theta).$$

A3. *Smoothness.* The kernel function $f(x;\theta)$ has common support and is three times continuously differentiable with respect to θ. The first two derivatives are denoted by $f'(x;\theta)$ and $f''(x;\theta)$.

A4. *Strong identifiability.* The kernel function $f(x;\theta)$ is strongly identifiable, ie.

(a) for any two mixing distribution functions Ψ_1 and Ψ_2 with two supporting points such that

$$\int f(x;\theta)d\Psi_1(\theta) = \int f(x;\theta)d\Psi_2(\theta), \text{ for all } x,$$

we must have $\Psi_1 = \Psi_2$;

(b) for any $\theta_1 \neq \theta_2$ in Θ,

$$\sum_{j=1}^{2}\{a_j f(x;\theta_j) + b_j f'(x;\theta_j) + c_j f''(x;\theta_j)\} = 0$$

implies that $a_j = b_j = c_j$, $j = 1, 2$.

A5. *Strong law of large numbers.* There exists a g with finite expectation such that

(a) $|Y_i(\theta)|^3 \leq g(X_i)$ and $|Z_i(\theta)|^3 \leq g(X_i)$ for all $\theta \in \Theta$;

(b) $|Z_i''(\theta)|^2 \leq g(X_i)$ for $\theta \in N(\theta_0)$, where $N(\theta_0)$ is some neighborhood of θ_0.

Remark 2.5.1. *Condition A5 is sufficient to ensure that the process*

$$n^{-1/2}\sum_{i=1}^{n}[\{Z_i(\theta) - Z_i(\theta_0)\}/(\theta - \theta_0)]$$

is tight in a small neighborhood of θ_0, see Billingsley (1968)*.*

Table 2.2: Parameters in Poisson, Binomial and Normal mixture models.

Model	$1-\alpha$	θ_1	θ_2	$100KL$
	Poisson mixtures:			
I	0.50	3.882	6.118	1.751
II	0.25	3.064	5.645	2.017
III	0.10	1.646	5.373	2.827
IV	0.05	0.127	5.256	5.081
	Binomial mixtures:			
I	0.50	0.417	0.583	1.989
II	0.25	0.356	0.548	2.059
III	0.10	0.250	0.528	2.358
IV	0.05	0.137	0.519	2.846
	Normal mixtures:			
I	0.50	-0.500	0.500	1.358
II	0.25	-0.866	0.289	1.444
III	0.10	-1.500	0.167	1.842
IV	0.05	-2.179	0.115	2.583

KL: Kullback-Leibler information.

Table 2.3: Null rejection rates (%) of the MLRT statistics.

	M_n			M_n^*		
Level	10%	5%	1%	10%	5%	1%
Poisson mixtures						
$n = 100$	9.7	5.0	1.0	9.9	5.2	1.1
$n = 200$	9.9	4.9	0.9	10.0	5.1	1.0
Binomial mixtures						
$n = 100$	9.5	4.9	1.0	9.8	5.1	1.1
$n = 200$	9.7	4.9	1.0	9.9	5.1	1.1
Normal mixtures						
$n = 100$	9.6	4.7	1.0	9.5	4.7	1.0
$n = 200$	9.8	4.8	1.1	9.7	4.7	1.1

Table 2.4: Power comparisons of the MLRT statistics under Poisson mixture models.

Model	10%		5%		1%	
	M_n	M_n^*	M_n	M_n^*	M_n	M_n^*
	$n = 100$					
I	62.9	62.5	49.4	48.8	25.4	25.0
II	65.2	65.5	51.9	51.8	27.7	27.8
III	65.8	68.8	53.8	57.5	31.3	35.6
IV	71.6	82.0	63.1	76.4	47.6	62.6
	$n = 200$					
I	83.2	82.8	74.2	73.7	50.8	49.5
II	84.8	84.9	76.3	76.3	54.3	53.8
III	85.9	88.4	78.1	82.2	60.1	65.8
IV	90.8	96.5	87.0	95.2	78.2	91.0

Table 2.5: Power comparisons of the MLRT statistics under Binomial mixture models.

Model	10%		5%		1%	
	M_n	M_n^*	M_n	M_n^*	M_n	M_n^*
	$n = 100$					
I	67.6	66.9	54.0	53.0	29.5	28.5
II	67.9	67.5	54.2	53.3	29.9	29.5
III	65.4	66.6	52.6	54.1	30.0	31.9
IV	64.3	69.5	52.4	59.7	32.9	40.5
	$n = 200$					
I	87.3	87.0	78.0	77.4	55.6	53.8
II	86.7	86.6	78.1	77.7	55.6	54.3
III	85.7	86.7	76.5	78.5	56.0	58.6
IV	83.5	88.0	74.9	82.2	57.4	68.5

Table 2.6: Power comparisons of the MLRT statistics under Normal mixture models.

Model	10%		5%		1%	
	M_n	M_n^*	M_n	M_n^*	M_n	M_n^*
	$n = 100$					
I	63.8	63.7	51.3	50.5	26.6	26.2
II	63.2	63.3	50.5	50.1	26.9	26.7
III	62.2	62.5	49.5	49.8	27.9	28.5
IV	59.9	61.6	49.7	51.5	30.6	33.2
	$n = 200$					
I	84.0	83.9	75.1	74.9	49.4	47.7
II	83.3	83.1	74.4	74.4	49.0	47.3
III	82.3	82.3	73.4	73.9	50.7	51.2
IV	79.3	81.5	71.9	75.0	53.9	57.0

Chapter 3

EM-test in Single Parameter Mixture Models

3.1 Motivation

In the last chapter, we reviewed the MLRT method and proposed a new class of penalty functions. In general, the MLRT is very convenient to use and widely applicable. At the same time, there are still many useful finite mixture models to which the MLRT can not be directly applied. To address this problem, we propose a new class of testing methods. We first provide some motivating insights.

Many first order asymptotic results in standard parametric models are based on the fact that the asymptotic distribution of the score vector is very tractable. However even for very simple mixture models the behaviour of the score and the shape of the log-likelihood function can be very different from the standard first order results. Consider, the mixture distribution in (2.1),

$$f(x;\Psi) = (1-\alpha)f(x;\theta_1) + \alpha f(x;\theta_2).$$

The score vector with respect to α at $\alpha = 0$ is based on

$$\frac{\partial}{\partial\alpha}\log f(x;\Psi)|_{\alpha=0} = \frac{f(x;\theta_2)}{f(x;\theta_1)} - 1.$$

It is immediately clear that if the covariance of this score is not finite then all standard asymptotic results based on the limiting normal distribution of the score will not hold.

Example 3.1.1. *Let $X_1, \ldots, X_n$ be a random sample from the following mixture of exponentials:*

$$(1-\alpha)Exp(1) + \alpha Exp(\theta),$$

where $Exp(\theta)$ denotes the exponential distribution with mean θ. The score statistic for α at $\alpha = 0$ and $\theta_1 = 1$ is given by

$$S(\theta) = \sum_{i=1}^{n}\Big[\frac{\theta^{-1}\exp\{-\theta^{-1}X_i\}}{\exp\{-X_i\}} - 1\Big]$$

which is a centered density ratio. Under the null model where $\alpha = 0$, however, we find

$$E\{S(\theta)^2\} = \begin{cases} \frac{n(1-\theta)^2}{\theta(2-\theta)} & \text{if } \theta < 2, \\ \infty & \text{if } \theta \geq 2. \end{cases}$$

Hence the only way to ensure a finite Fisher information is to restrict the range of θ to be less than 2.

Many inference procedures which are based on the shape of the log-likelihood function rely on this shape being approximately quadratic. The log-likelihood function for simple mixture models such as (2.1) in fact can be very far from a quadratic, see Anaya-Izquierdo and Marriott (2007a, 2007b) and Marriott (2007). Furthermore the shape can be dominated by a few highly influential observations even when the model is correctly specified, Marriott (2007).

Example 3.1.2. *Consider a simple normal mixture model given by* $(1-\alpha)N(0,1)+\alpha N(\mu,1)$ *with* $\mu \in \Theta \subset R$. *It is common to consider the likelihood ratio test for the hypothesis*

$$H_0 : \alpha\mu = 0$$

based on a random sample $X_1, X_2, \ldots, X_n$. *Hartigan* (1985) *showed that the likelihood ratio statistic goes to* ∞ *in probability as* $n \to \infty$ *when* $\Theta = R$. *That is, the classical chi-square limiting distribution result of Wilks* (1938) *is not applicable.*

In order to be able to use standard testing procedures many authors have been forced to make assumptions regarding the existence of the Fisher information and compactness of parameter spaces.

As pointed out in Anaya-Izquierdo and Marriott (2007a), the homogeneity testing problem in (2.2) can be challenging since the mixture can be close to the unmixed model in two quite distinct ways. One is that the two components θ_1 and θ_2 in (2.1) are both close to θ. Secondly the components might be very far from each other but the mixing parameter is very close to 0 or 1. It is in the second case that the Fisher information in the α-parameter direction causes most problems. Furthermore if this mixing parameter is much smaller than the inverse of the sample size then it is effectively not estimable. It should be noted that the MLRT employs a penalty function to prevent the manifestation of this ill effect. Yet the MLRT still relies on the compactness assumption and the finiteness of the Fisher information.

Finding an effective and convenient method for the test of homogeneity has challenged statisticians for a long time. Hartigan (1985) was the first to notice this challenge and provided the Example 3.1.2 above as a case where standard methods fail.

Although Bickel and Chernoff (1993) and Liu and Shao (2004) successfully derived the limiting distribution of the LRT under this specific model, the general problem under more useful models where Θ is not compact remains open. Recent advances are mostly obtained by confining the mixing parameter(s) into a compact space (Dacunha-Castelle and Gassiat, 1999; Chen and Chen, 2001; Liu and Shao, 2003, etc).

In addition, either explicitly or implicitly, these results are based on the assumption that all density ratios, $f(x;\theta)/f(x;\theta_0)$, has finite second or even higher moment under H_0 for all $\theta \in \Theta$, where $f(x;\theta_0)$ is the true distribution under the null hypothesis. Thus they are assuming the Fisher information is finite. To better explore the problem, we show what happens when a score test proposed by Davies (1977) is attempted. We present this case as follows.

Example 3.1.1. (Continued) *We wish to test the homogeneity null hypothesis*

$$H_0 : \alpha(\theta - 1) = 0.$$

According to Davies (1977)*, for each given θ, we first calculate a score statistic as the derivative of the log-likelihood function with respect to α at $\alpha = 0$.*

As a general rule, the test statistic is to be defined as

$$\sup_{\theta\in\Theta} n^{-1/2} S(\theta)/\sqrt{E\{S(\theta)^2\}}.$$

The score test is clearly not sensible because the supremum is effectively restricted to the range of $\theta < 2$ by the implicit assumption regarding the finiteness of the Fisher information.

Similar comments also apply to tests based on the log-likelihood. Investigation reveals that a finite Fisher information is explicitly or implicitly assumed in all the papers we are aware of (Chen and Chen, 2001; Chen et al. 2001, 2004; Dacunha-Castelle and Gassiat, 1999; Ghosh and Sen, 1985; Liu and Shao 2003, and Charnigo and Sun, 2004).

In this chapter, we propose an EM-test that is completely free from the two difficulties illustrated by the above examples. The EM-test statistic has a simple limiting distribution

$0.5\chi_0^2+0.5\chi_1^2$ for mixture models with single parameter kernel function. We further discuss a precision enhancing technique to improve the chi-square approximation to the finite sample distribution of the test statistics.

The remainder of this chapter is organized as follows. In Section 3.2, we first introduce the EM-test and then we examine the asymptotic properties of the EM-test subsequently. Finally a higher order adjustment for the non-zero proportion is investigated, which enhances the precision of the type I error calibrated by the limiting distribution. In Section 3.3, simulation studies are used to examine the performance of the EM-test, followed by the application of the EM-test to some real data examples in Section 3.4. For the convenience of presentation, the regularity conditions and proofs are deferred to Section 3.5.

3.2 The EM-test and Its Asymptotic Properties

3.2.1 Testing Procedure

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a two-component mixture model (2.1). We are interested in the homogeneity testing problem given in (2.2). The modified log-likelihood function is defined as follows.

$$\begin{aligned} pl_n(\alpha, \theta_1, \theta_2) &= \sum_{i=1}^{n} \log\{(1-\alpha)f(X_i;\theta_1) + \alpha f(X_i;\theta_2)\} + p(\alpha) \\ &= l_n(\alpha, \theta_1, \theta_2) + p(\alpha) \end{aligned}$$

where $l_n(\alpha, \theta_1, \theta_2)$ is the usual log-likelihood function defined in (2.3) and $p(\alpha)$ is a penalty function on α such that $p(\alpha)$ achieves the maximal value at $\alpha = 0.5$. The role of the penalty function $p(\alpha)$ has been discussed in Chapter 2.

We propose an EM-test procedure for homogeneity as follows. For each fixed $\alpha = \alpha_0 \in$

$(0, 0.5]$, we compute a penalized likelihood ratio test statistic of the form

$$M_n(\alpha_0) = 2\{pl_n(\alpha_0, \tilde{\theta}_{01}, \tilde{\theta}_{02}) - pl_n(0.5, \tilde{\theta}_0, \tilde{\theta}_0)\}$$

with $\tilde{\theta}_{01}$ and $\tilde{\theta}_{02}$ being the maximizers of $pl_n(\alpha_0, \theta_1, \theta_2)$ and $\tilde{\theta}_0$ being the maximizer of $pl_n(0.5, \theta, \theta)$.

Because $0 < \alpha_0 \leq 0.5$, the mixture model is fully identifiable. It is easy to verify that when $n \to \infty$, we have $\tilde{\theta}_{01} - \theta_0 = o_p(1)$ and $\tilde{\theta}_{02} - \theta_0 = o_p(1)$. Consequently, we can show that $M_n(\alpha_0)$ has a simple χ^2-type null limiting distribution without imposing any restrictive conditions. It is thus mathematically very convenient to conduct a test based on $M_n(\alpha_0)$. Note that if the data are from an alternative model with α far from α_0, this test is likely to be inefficient. To improve the power, we adopt an EM-like algorithm (Dempster et al., 1977) to iteratively update $M_n(\alpha)$, for a fixed and finite number of times, between α and (θ_1, θ_2). In addition, we choose a number of initial values of α_0 to accelerate this process so that only a few iterations are necessary to capture the true value of θ if the data are from the alternative model. We then use the maximum value of $M_n(\alpha_0)$'s as our test statistic.

The EM-test statistic is best explained by the following pseudo code.

Step 0. Choose a number of initial α values, say $\alpha_1, \alpha_2, \ldots, \alpha_J \in (0, 0.5]$. Compute

$$\tilde{\theta}_0 = \arg\max_{\theta} pl_n(0.5, \theta, \theta).$$

Let $j = 1, k = 0$.

Step 1. Let $\alpha_j^{(k)} = \alpha_j$.

Step 2. Compute

$$(\theta_{j1}^{(k)}, \theta_{j2}^{(k)}) = \arg\max_{\theta_1, \theta_2} pl_n(\alpha_j^{(k)}, \theta_1, \theta_2)$$

and

$$M_n^{(k)}(\alpha_j) = 2\{pl_n(\alpha_j^{(k)}, \theta_{j1}^{(k)}, \theta_{j2}^{(k)}) - pl_n(0.5, \tilde{\theta}_0, \tilde{\theta}_0)\}.$$

Step 3. For $i = 1, 2, \ldots, n$, compute the weights which are the conditional expectations in the E-step,

$$w_{ij}^{(k)} = \frac{\alpha_j^{(k)} f(X_i; \theta_{j2}^{(k)})}{(1-\alpha_j^{(k)}) f(X_i; \theta_{j1}^{(k)}) + \alpha_j^{(k)} f(X_i; \theta_{j2}^{(k)})}.$$

Now following the M-step, let

$$\alpha_j^{(k+1)} = \arg\max_{\alpha}\{(n - \sum_{i=1}^{n} w_{ij}^{(k)}) \log(1-\alpha) + \sum_{i=1}^{n} w_{ij}^{(k)} \log(\alpha) + p(\alpha)\},$$

$$\theta_{j1}^{(k+1)} = \arg\max_{\theta_1}\{\sum_{i=1}^{n} (1 - w_{ij}^{(k)}) \log f(X_i; \theta_1)\}$$

and

$$\theta_{j2}^{(k+1)} = \arg\max_{\theta_2}\{\sum_{i=1}^{n} w_{ij}^{(k)} \log f(X_i; \theta_2)\}.$$

Compute

$$M_n^{(k+1)}(\alpha_j) = 2\{pl_n(\alpha_j^{(k+1)}, \theta_{j1}^{(k+1)}, \theta_{j2}^{(k+1)}) - pl_n(0.5, \tilde{\theta}_0, \tilde{\theta}_0)\}.$$

Let $k = k + 1$ and repeat Step 3 for a fixed number of iterations in k.

Step 4. Let $j = j + 1$, $k = 0$ and go to Step 1, until $j = J$.

Step 5. For each k, calculate the test statistic as

$$EM_n^{(k)} = \max\{M_n^{(k)}(\alpha_j), j = 1, 2, \ldots, J\}.$$

In the above algorithm, fixing α to be one of $\alpha_j's$ and choosing a fixed finite number of iterations on k can be seen as two soft compactness conditions. Under these two conditions, when the data are from the null model $f(x; \theta_0)$, we can show that the value of α will be in a small neighborhood of the initial value of α and the values of θ_1 and θ_2 are in the small neighborhood of θ_0 after k iterations. So the limiting distribution of $EM^{(k)}$ will not rely on the finite Fisher information condition and the compact parameter space assumption. If the index k was allowed to grow unboundedly this nice property may disappear.

The EM-test is partially motivated by the score test proposed by Liang and Rathouz (1999). It is originally designed for the case when $\theta_1 = \theta_0$ is known and $\theta_2 = \theta$ is unknown, and the testing problem is

$$H_0 : \alpha(\theta - \theta_0) = 0.$$

For this problem, similar to Example 3.1.1, the score statistic for α at $\alpha = 0$ and $\theta_1 = \theta_0$ is given by

$$S(\theta) = \sum_{i=1}^{n} \Big[\frac{f(X_i; \theta)}{f(X_i; \theta_0)} - 1 \Big].$$

Liang and Rathouz (1999) proposed to first choose an fixed α value from $(0, 1]$. Given this α, a maximum likelihood estimator of θ is then obtained. Denote this estimator by $\hat{\theta}_\alpha$. The score test statistic is defined to be $T_\alpha = \alpha S(\hat{\theta}_\alpha)$. Under some conditions this statistic enjoys a χ^2-type limiting distribution under the null hypothesis. Simulation shows that for a number of mixture models, this method has very good power properties too.

The score test of Liang and Rathouz (1999) can be directly used for the models in Examples 3.1.1 and 3.1.2. In both tests, a pre-chosen value of the mixing proportion is utilized. However, the EM-test employs the likelihood ratio statistic which is more efficient than the score statistic by general consensus. In addition, the EM-test iterates to find a more suitable mixing proportion which improves efficiency, while the score test has no such mechanism. Namely, it uses a single α value regardless of the actual fitting of the data.

The MLRT can be regarded as the limiting case of the EM-test. When the number of iterations $k \to \infty$, and under the assumption that the EM-algorithm converges to a global maximum, then the EM-test statistic becomes the modified likelihood ratio test.

It is of interest to point out that the EM-test does not have any global/local maximum problems which can occur in other existing methods. In Figure 3.1, we plot the $M_n^{(k)}(\alpha)$ values of $k = 0, 5, 10$ based on two simulated data sets, one from a null model and another from an alternative model. Under the null model, the iteration does not increase the value of $M_n^{(k)}(\alpha)$ much and its value at $\alpha = 0.5$ dominates. In comparison, under the alternative, the value of $M_n^{(k)}(\alpha)$ increases at $\alpha = 0.1$ with k, and it dominates. Hence, the EM-test retains the most relevant value unlike the MLRT which searches exhaustively.

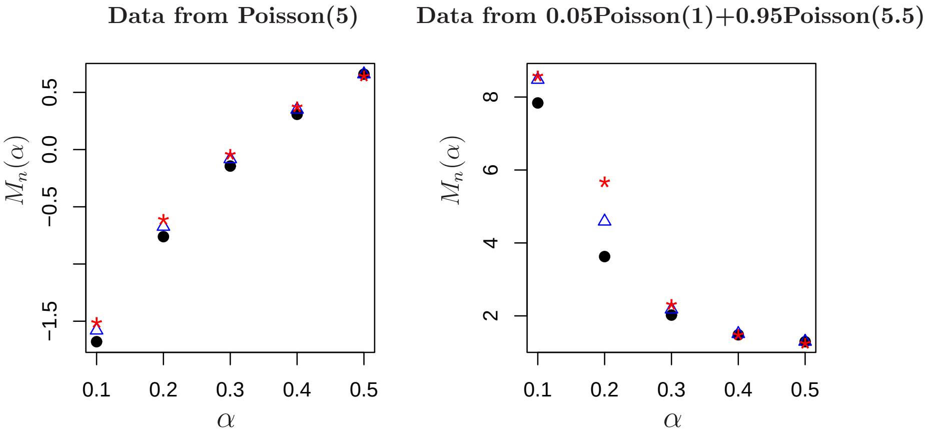

Figure 3.1: The $M_n^{(k)}(\alpha)$ at $\alpha = 0.1, 0.2, \ldots, 0.5$, $k = 0, 5, 10$ ($\bullet$: 0 iteration; $\triangle$: 5 iterations; $\star$: 10 iterations).

By removing the penalty term $p(\alpha)$, the EM-test reduces to the ordinary likelihood ratio test when the number of iterations $k = \infty$. However note that the likelihood ratio test has a complex limiting distribution available only under more restrictive conditions.

Chen and Cheng (1995) and Lemdani and Pons (1995) proposed a constrained test by requiring $\epsilon_0 \le \alpha$ for some fixed positive constant $\epsilon_0 \in (0, 1/2]$. There are some similarities between that method and the EM-test, because the EM-test requires pre-chosen mixing proportions be larger than 0. However, the EM-iteration allows us to recoup the mixture models with smaller mixing proportions while the other method does not.

3.2.2 Asymptotic Results for the EM-test

Under very general conditions, for fixed finite k and any finite set of pre-chosen α_j, we show that the test statistic $EM_n^{(k)}$ has simple limiting distribution $0.5\chi_0^2 + 0.5\chi_1^2$. The details are in the follow theorems with proofs given in Section 3.5.

Theorem 3.2.1. *Suppose that $f(x;\theta)$ satisfies Conditions B1-B5 given in Section 3.5, and $p(\alpha)$ is a continuous function such that $p(\alpha) \to -\infty$ as $\alpha \to 0$ or 1 and it attains its maximal value at $\alpha = 0.5$. Under the null distribution $f(x;\theta_0)$, we have, for $j = 1, \ldots, J$ and any fixed finite k,*

$$\begin{aligned} \alpha_j^{(k)} - \alpha_j &= o_p(1), \\ \theta_{j1}^{(k)} - \theta_0 &= O_p(n^{-1/4}), \\ \theta_{j2}^{(k)} - \theta_0 &= O_p(n^{-1/4}) \end{aligned}$$

and

$$m_{j1}^{(k)} = (1-\alpha_j^{(k)})(\theta_{j1}^{(k)} - \theta_0) + \alpha_j^{(k)}(\theta_{j2}^{(k)} - \theta_0) = O_p(n^{-1/2}).$$

Based on the above results, we can easily derive the null distribution of $EM_n^{(k)}$.

Theorem 3.2.2. *Assume the same conditions as in Theorem 3.2.1, and that one of α_j's is equal to 0.5. Under the null distribution $f(x;\theta_0)$, and for any fixed finite k, as $n \to \infty$,*

$$EM_n^{(k)} \xrightarrow{d} 0.5\chi_0^2 + 0.5\chi_1^2.$$

Remark 3.2.1. (*Understanding the limiting distribution*). *Let $\eta = \alpha(1-\alpha)(\theta_1 - \theta_2)$. The test of homogeneity is to test $\eta = 0$ against the alternative $\eta \neq 0$. Intuitively, the limiting distribution of the EM-test statistic should be χ_1^2. Note that the following two groups of parameters: $(\alpha, \theta_1, \theta_2)$ and $(1-\alpha, \theta_2, \theta_1)$ give the same density function in* (2.1)*. Due to symmetry, we can assume $\theta_1 \leq \theta_2$ without loss of any information. Therefore the parameter* *space for η is restricted to the non-positive part of the real line and the null hypothesis is on the boundary of the parameters space. Chernoff* (1954)*, Self and Liang* (1987) *and Lindsay* (1995) *all provide discussion on the limiting distributions when the null hypothesis is on the boundary. The limiting distributions in this case are often the mixture of χ^2 distributions instead of χ_1^2.*

Another way for understanding the limiting distribution is from the point view of moment. Let us use the two-component normal mixture model with same and known variance of 1 as the example. Without loss of generality, we further assume that the mean of the two-component normal mixture model is 0. Under this setup, the test of homogeneity is to select one model between $N(0, 1)$ and $(1 - \alpha)N(\mu_1, 1) + \alpha N(\mu_2, 1)$ with

$$(1 - \alpha)\mu_1 + \alpha\mu_2 = 0.$$

For any given $\alpha \in (0, 0.5]$, the second moment for the mixture model is given by

$$E(X_1^2) = 1 + \alpha\mu_1^2 + (1 - \alpha)\mu_2^2,$$

which is greater than or equal to 1, with the equality holding when the mixture model is the homogeneous model. So the test of homogeneity is equivalent to testing

$$H_0 : E(X_1^2) = 1 \textit{ versus } H_a : E(X_1^2) > 1.$$

The null hypothesis is on the boundary of the parameter space, and so the limiting distribution in this case is the mixture of χ^2 distributions.

Remark 3.2.2. *For each given $\alpha \in (0, 0.5]$, $M_n(\alpha)$ can be written as the summation of two terms: one is from the likelihood function and the other from the penalty. Under the null model, the first term has the same quadratic approximation for all α values. However, different α values result in different sizes of penalty. Since the penalty $p(\alpha)$ attains the* *maximum value of 0 at* $\alpha = 0.5$*, including* $\alpha = 0.5$ *implies that the limiting distribution is determined by the quadratic approximation only, and hence has the simplest form.*

We emphasize here that Conditions $B1 - B5$ do not include the condition that the mixing parameter is confined in a compact space, nor conditions on finiteness of the Fisher information. Hence, the EM-test is both more convenient in applications and more widely applicable.

Based on the above result, the EM-test rejects the null hypothesis when $EM_n^{(k)}$, for a prechosen k, is larger than some quantile of the limiting distribution. In theory, we can choose a dense set of α in the interval $(0, 0.5]$ as our initial values and iterate the EM-like algorithm many times. Simulation studies suggest that three initial values (0.1,0.3,0.5) for α, and one iteration are enough to arrive at an efficient EM-test statistic.

3.2.3 Precision Enhancing Methods

Before the EM-test is fully implemented, we suggest two precision enhancing measures to further improve its utility. In applications, the limiting distribution of the test statistic is usually used to provide a critical value for rejecting the null hypothesis. However when the sample size is not large, the calibration via the limiting distribution might not be precise enough. One way to improve the calibration precision is to choose a good penalty function.

For the validity of the asymptotic result, $p(\alpha)$ must decrease to negative infinity when $\alpha \to 0$ or 1. Other considerations include computational convenience and statistical efficiency. Based on our discussions in Chapter 2, the penalty function in (2.7) is recommended for computing the EM-test statistics.

The next precision enhancing measure is motivated from the following observation. By a quick inspection of the limiting distribution, it is suggestive that $(1-p_n)\chi_0^2+p_n\chi_1^2$ with $p_n = \Pr(EM_n^{(k)} > 0)$ may better approximate the finite sample distribution. Because of this, a good approximation for p_n can be useful. Let $\mu(f)$ and $\sigma^2(f)$ be the mean and variance under the homogeneous model, respectively. Further let $S = E[\{X_1 - \mu(f)\}^2] - \sigma^2(f)$ being an over-dispersion measure. The mixture model is not justified unless possibly when $S > 0$. Note that $S_n = \sum_{i=1}^{n}(X_i - \bar{X})^2/n - \hat{\sigma}^2(f)$ provides consistent estimation of the over-dispersion measure S, where $\hat{\sigma}^2(f)$ is the consistent estimate of $\sigma^2(f)$. Intuitively, if $S_n \leq 0$, the homogeneous model should be not rejected. Therefore we approximate p_n by $\Pr\{S_n > 0\}$.

In the following proposition, we use the Edgeworth expansion to find the leading term of this probability. We omit the proof because it is a routine application of the techniques in Hall (1992, p. 56).

Proposition 3.2.1. *Under null hypothesis, if $E(X_1^6) < \infty$, then*

$$p_n \approx Pr\{S_n > 0\} = 0.5 + \frac{1}{\sqrt{2\pi n}}\Big(a - \frac{b}{6}\Big) + o(n^{-1/2}), \tag{3.1}$$

where

$$a = \lim_{n\to\infty} n^{1/2} E\Big\{\frac{S_n}{\sqrt{\mathrm{Var}(S_n)}}\Big\} \text{ and } b = \lim_{n\to\infty} n^{1/2} E\Big\{\frac{S_n - E(S_n)}{\sqrt{\mathrm{Var}(S_n)}}\Big\}^3.$$

Further, if $E(X_1^{10}) < \infty$, then the remaining term $o(n^{-1/2})$ in (3.1) *can be strengthened to $O(n^{-3/2})$.*

In the above proposition, the Edgeworth approximation relies on the condition $E(X_1^6) < \infty$. Note that there exists some distributions, such as the Exponential distribution and the Geometric distribution, which satisfy this condition or even the condition $E(X_1^{10}) < \infty$, but do not satisfy the finite Fisher information condition. So the condition $E(X_1^6) < \infty$ is not restrictive compared with the finite Fisher information condition. The quantities a and b may depend on unknown parameters, in which case we replace them by their consistent estimates under the homogeneity model.

For many commonly used distributions, we can compute a and b analytically and the results are presented in the Table 3.1. In the Poisson and Binomial examples, we can replace the unknown θ by its maximum likelihood estimate under the null model.

Table 3.1: Edgeworth approximations of p_n for commonly used kernel functions.

Kernel	Edgeworth approximation
$N(\mu, \sigma_0^2)$	$0.5 - \frac{5}{6\sqrt{\pi n}} + O_p(n^{-3/2})$
$Pois(\theta)$	$0.5 - \frac{5\theta+1}{6\theta\sqrt{\pi n}} + O_p(n^{-3/2})$
$Binom(N, \theta)$	$0.5 - \frac{1}{\sqrt{\pi n}} \frac{(5N-11)\theta(1-\theta)+1}{6\theta(1-\theta)\sqrt{N(N-1)}} + O_p(n^{-3/2})$
$Exp(\theta)$	$0.5 - \frac{8}{3\sqrt{2\pi n}} + O_p(n^{-3/2})$

σ_0^2 in the normal kernel is assumed known

Needless to say, we recommend the use of penalty function (2.7) for the EM-test together with the higher order adjustment. These two practical considerations enhance the performance of the new method.

3.3 Simulation Study

Our simulation study examines many aspects of the EM-test and related issues.

First, we examine the precision of the Edgeworth expansion for $p_n = \Pr(EM_n^{(k)} > 0)$. We consider null models with kernels $N(0,1)$, $Pois(5)$, $Pois(3)$, $Exp(5)$ (mean=5), $Exp(1)$ (mean=1), $Binom(10, 0.3)$ and $Binom(10, 0.5)$. In each case, we generated random samples of sizes $n = 100$ and 200. The non-zero proportions of the EM-test statistics are calculated based on 20,000 repetitions for each kernel. The penalty function in (2.7) with $C^* = 1$ and two sets of initial values for α, (0.1,0.2,0.3,0.4,0.5) and (0.1,0.3,0.5), are used to compute the EM-test statistics. Note the non-zero proportions of the EM-tests are almost the same. Hence only the non-zero proportions of the EM-test statistics based on the second set of α are presented. The simulation results are summarized in Table 3.2. Clearly, (3.1) gives a very good approximation to p_n in all cases considered. Another observation from the simulation is that the non-zero proportions are not affected much by the value of C^*. For example, if the data are generated from $Exp(5)$, when C^* changed from 1 to 1.5, the non-zero proportion changed from 0.395 to 0.393 for $n = 100$, and from 0.423 to 0.422 for $n = 200$; if the data are generated from $Exp(1)$, when C^* changed from 1 to 1.5, the non-zero proportion changed from 0.396 to 0.394 for $n = 100$, and from 0.425 to 0.422 for $n = 200$.

Next, the Poisson mixture model is used to compare the performance of the MLRT and the EM-test when limiting distribution for both statistics are available. The setup for Poisson mixture model in Section 2.4 is used in this simulation. For the MLRT and the EM-test, we use the penalty function in (2.7) with $C^* = 1$. We computed the null rejection rates based on 20,000 repetitions and the powers based on 10,000 repetitions. The outcomes are summarized in Tables 3.3 and 3.4. We find that the null rejection rates of both the MLRT and the EM-test in all cases are close to the nominal values. The power of the EM-test and the MLRT are almost same for all four models especially when $|\alpha - 0.5|$ is not too large. As we pointed out earlier, the MLRT statistic is the EM-test statistic with $k = \infty$. The crucial point is: the asymptotic property of the EM-test is applicable to more general mixture models. We also find that there is no need of using more than three initials values of α or more than 1 iteration in EM-like algorithm. Additional initial values or iterations do not meaningfully improve the power.

We now study the EM-test under the models where the asymptotic results of the MLRT or the LRT are not applicable. Exponential kernel is used in this simulation. We set the mean of the mixture model 5 in all cases and the same parameter values for alternative

Table 3.2: Simulated non-zero proportions of the EM-test statistics.

				Edgeworth	
Kernel	$EM_n^{(0)}$	$EM_n^{(1)}$	$EM_n^{(2)}$	Approximation	Std
		$n = 100$			
$N(0,1)$	0.449	0.449	0.449	0.453	0.0035
$Pois(5)$	0.449	0.449	0.449	0.451	0.0035
$Pois(3)$	0.448	0.448	0.448	0.450	0.0035
$Binom(10, 0.5)$	0.453	0.453	0.453	0.457	0.0035
$Binom(10, 0.3)$	0.453	0.453	0.453	0.457	0.0035
$Exp(5)$	0.395	0.395	0.395	0.394	0.0035
$Exp(1)$	0.396	0.396	0.396	0.394	0.0035
		$n = 200$			
$N(0,1)$	0.463	0.463	0.463	0.467	0.0035
$Pois(5)$	0.465	0.465	0.465	0.465	0.0035
$Pois(3)$	0.465	0.465	0.465	0.465	0.0035
$Binom(10, 0.5)$	0.467	0.467	0.467	0.470	0.0035
$Binom(10, 0.3)$	0.468	0.468	0.468	0.469	0.0035
$Exp(5)$	0.423	0.423	0.423	0.425	0.0035
$Exp(1)$	0.425	0.425	0.425	0.425	0.0035

Table 3.3: Null rejection rates (%) under Poisson kernel.

Level	MLRT	$EM_n^{(0)}$	$EM_n^{(1)}$	$EM_n^{(2)}$	$EM_n^{(0)}$	$EM_n^{(1)}$	$EM_n^{(2)}$
				$n = 100$			
10%	9.9	9.8	9.8	9.8	9.8	9.8	9.8
5%	5.2	5.1	5.2	5.2	5.1	5.1	5.1
1%	1.1	1.1	1.1	1.1	1.1	1.1	1.1
				$n = 200$			
10%	10.0	9.9	9.9	9.9	9.8	9.8	9.9
5%	5.1	4.9	5.0	5.0	4.9	4.9	4.9
1%	1.0	1.0	1.0	1.0	0.9	1.0	1.0

Results in columns (3, 4, 5) used $\alpha = (0.1, 0.2, 0.3, 0.4, 0.5)$.

Results in columns (6, 7, 8) used $\alpha = (0.1, 0.3, 0.5)$.

models as in Table 3.5. Although the limiting distributions of the LRT (denoted by R_n) and the D-test (Charnigo and Sun, 2004) are not available, these tests be done by using simulated quantiles under some null models. They are included in the simulation to serve as efficiency barometers.

The definition of the D-test (Charnigo and Sun, 2004) is given as follows. Let $(\hat{\alpha}, \hat{\theta}_1, \hat{\theta}_2)$ and $(\hat{\theta}_0, \hat{\theta}_0)$ maximize $l_n(\alpha, \theta_1, \theta_2)$ and $l_n(1/2, \theta, \theta)$ respectively. The D-test statistic is defined as

$$D = d(2, n) = \int \left\{ (1 - \hat{\alpha}) f(x; \hat{\theta}_1) + \hat{\alpha} f(x; \hat{\theta}_2) - f(x; \hat{\theta}_0) \right\}^2 dx.$$

A weight function can also be incorporated in this definition to enhance the efficiency of the D-test. Charnigo and Sun (2004) considered weighting functions x and x^2 for testing homogeneity in mixtures in exponential family distributions. We call these two weighted

Table 3.4: Simulated powers (%) under Poisson mixture alternatives at the 5% level.

Model	MLRT	$EM_n^{(0)}$	$EM_n^{(1)}$	$EM_n^{(2)}$	$EM_n^{(0)}$	$EM_n^{(1)}$	$EM_n^{(2)}$
				$n = 100$			
I	48.8	49.0	49.0	49.0	49.0	49.0	49.0
II	51.8	51.8	51.8	51.9	51.8	51.8	51.8
III	57.5	57.1	57.3	57.4	56.8	57.1	57.2
IV	76.4	72.0	74.3	74.5	72.0	74.3	74.5
				$n = 200$			
I	73.7	73.9	73.9	73.9	73.9	73.9	73.9
II	76.3	76.5	76.5	76.5	76.4	76.4	76.4
III	82.2	81.6	81.7	81.8	81.5	81.7	81.7
IV	95.2	91.5	92.2	92.4	91.5	92.2	92.4

Results in columns (3, 4, 5) used $\alpha = (0.1, 0.2, 0.3, 0.4, 0.5)$.

Results in columns (6, 7, 8) used $\alpha = (0.1, 0.3, 0.5)$.

versions of the D-test statistics as $d_1(2, n)$ and $d_2(2, n)$.

The MLRT can also be calibrated by simulated quantiles, but it is bounded by the EM-test and the LRT and therefore is not included. The constrained LRT (Chen and Cheng, 1995; Lemdani and Pons, 1995) is applicable under the same conditions as the EM-test. Its test statistic is defined as

$$R_n(\epsilon_0) = 2\left\{ \sup_{\alpha\in[\epsilon_0,\ 1-\epsilon_0],\ \theta_1,\ \theta_2} l_n(\alpha, \theta_1, \theta_2) - l_n(0.5, \hat{\theta}_0, \hat{\theta}_0) \right\},$$

for some user chosen positive constant $\epsilon_0 \in (0, 0.5]$. We find that it has large type I errors unless we choose a large value of ϵ_0. Through some pilot simulation studies, we found that we have to select ϵ_0 as large as 0.45 to have its type I errors comparable with those of

Table 3.5: Parameters in four exponential mixture models.

	$1-\alpha$	θ_1	θ_2	Δ	$100KL$
Model I	0.50	3.129	6.871	3.50	1.008
Model II	0.25	2.128	5.957	2.75	0.956
Model III	0.10	0.757	5.471	2.00	1.252
Model IV	0.05	0.127	5.256	1.25	1.996

Δ: variance of mixing distribution.

KL: Kullback-Leibler information.

EM-tests. We thus included the constrained LRT with $\epsilon_0 = 0.45$ in our simulation.

Software for calculating the critical values of the D-test for the $Exp(1)$ distribution can be found at *http://stat.cwru.edu/˜rjc12*. For other null distributions, some transformations as suggested in Charnigo and Sun (2004) must be used. We computed $EM_n^{(k)}$ for $k = 0, 1, 2$ with $C^* = 1$ and $\alpha \in \{0.1, 0.3, 0.5\}$ first. Their type I errors are also somewhat larger than the nominal values, we hence also computed the EM-tests with $C^* = 1.5$. The null rejection rates of the EM-tests, D-tests and constrained LRT calibrated by limiting distributions or by critical values obtained from references are in Table 3.6, The EM-tests and constrained LRT have reasonably accurate type I errors. The D-test statistics may not sufficiently invariant to allow transformation of critical values between the $Exp(1)$ and the $Exp(5)$ null distributions. We note that the EM-tests are slightly over-sized but not too severely with both choices of $C^* = 1$ and $C^* = 1.5$. Both meet the recommendation criterion set in Remark 2.3.2, that is, choose a C^* value large enough such that the simulated null rejection rates is no more than 5.5% at 5% significance level.

The power calculations of all methods were done using simulated quantiles to ensure objective comparisons. In general, the efficiency of the EM-test is much better than other

Table 3.6: Null rejection rates (%) of the EM-test, the constrained LRT and the D-test under exponential mixtures.

				$C^* = 1$		$C^* = 1.5$		
Level	$d(2, n)$	$d_1(2, n)$	$d_2(2, n)$	$EM_n^{(0)}$	$EM_n^{(1)}$	$EM_n^{(0)}$	$EM_n^{(1)}$	$R_n(0.45)$
				$n = 100$				
10%	18.8	12.4	8.3	11.0	11.1	10.5	10.5	10.8
5%	12.2	5.1	4.0	5.8	6.0	5.3	5.4	5.5
1%	4.0	0.8	0.8	1.2	1.3	1.1	1.1	1.1
				$n = 200$				
10%	19.7	14.9	10.4	10.5	10.6	10.2	10.2	10.5
5%	13.5	7.3	4.3	5.5	5.5	5.2	5.2	5.3
1%	5.0	1.0	0.7	1.2	1.3	1.1	1.2	1.1

methods. The D-test based on $d(2, n)$ is less efficient than the EM-test when α is close to 0.5, but more efficient for alternatives when α is close to 1. This result may not be very useful because the type I error of the $d(2, n)$ based D-test is hard to control. An interesting result is that the EM-test is much more efficient than the LRT when α is close to 0.5. Due to the penalty function, the EM-test is expected to lose power when α is close 0.

3.4 Applications and Real Examples

In this section, we analyze a number of well-known real data sets to further demonstrate the use of the EM-test.

Example 3.4.3. First, we apply the EM-test to the data studied in Proschan (1963). The data consist of the times of successive failures for the air conditioning system of

Table 3.7: Simulated powers (%) of the D-test, the EM-test, the constrained LRT and the LRT under exponential mixture alternatives at the 5% level.

				$C^* = 1$		$C^* = 1.5$			
Model	$d(2,n)$	$d_1(2,n)$	$d_2(2,n)$	$EM_n^{(0)}$	$EM_n^{(1)}$	$EM_n^{(0)}$	$EM_n^{(1)}$	$R_n(0.45)$	R_n
$n = 100$									
I	17.6	30.1	32.6	33.6	33.4	34.1	34.0	34.6	29.8
II	22.1	30.7	30.3	31.0	30.8	31.3	31.3	31.2	27.9
III	35.4	31.9	24.3	29.2	29.6	27.6	27.9	24.9	32.6
IV	49.5	21.9	10.1	32.2	32.9	28.4	29.0	17.6	42.3
$n = 200$									
I	26.3	48.1	51.2	53.4	53.3	53.6	53.6	54.1	47.6
II	34.5	49.3	47.7	48.0	48.0	47.5	47.6	47.4	44.6
III	54.6	51.7	39.8	45.8	46.0	42.1	42.5	37.5	52.1
IV	66.4	34.9	12.4	46.8	48.6	42.2	43.6	22.5	61.5

each member in a fleet of 13 Boeing 720 jet airplanes. Proschan (1963) performed the Kolmogorov-Smirnov test to the pooled data, a total of 213 observations, to determine whether the exponential distribution offered a good fit of the pooled failure times. At the level of 0.05, the Kolmogorov-Smirnov test failed to reject the null hypothesis of exponential fit. However, the exponential distribution did not fit the pooled failure times very well. Figure 3.2 gives the plot of the log empirical survival curve for the pooled data and the log theoretical survival curve under the exponential model. Proschan (1963) observed that the log empirical survival curve lies consistently below the theoretical curve when the failure time is less than 150 and lies consistently above the theoretical curve when the failure time is larger than 150.

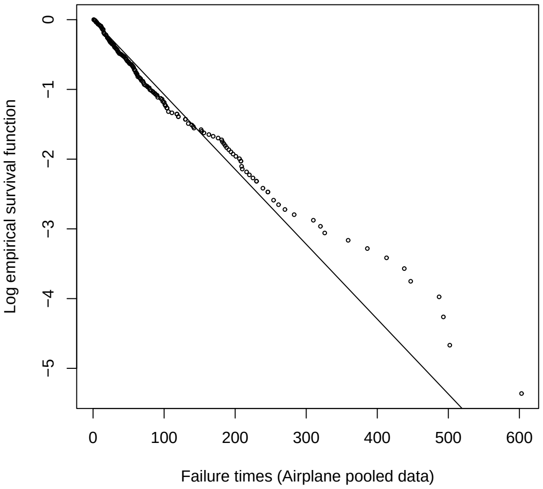

Figure 3.2: The log empirical (points) and null theoretical (solid line) survival functions of the airplane pooled data.

Proschan (1963) further used a more refined analysis to show that the failure distribution for each airplane separately was exponential, but for some airplanes the rates were different.

So it will be reasonable to assume the pooled failure times follows a mixture of exponential distributions. Now we conduct the test of homogeneity for the pooled data. The MLEs under the mixture model for $(\alpha, \theta_1, \theta_2)$ is $(0.430, 128.286, 46.506)$. Since $\hat{\theta}_2/\hat{\theta}_1 = 2.758 > 2$, most existing methods of testing the homogeneity are strictly not applicable because the density ratio may have infinite second moment, or infinite Fisher information. In contrast, a rigorous EM-test can be conducted. Accord to our simulations, $C^* = 1.5$ is a good choice for the level of modification for the pooled failure times. We computed the EM-statistics with $C^* = 1.5$ and three initial values (0.1,0.3,0.5) of α, and found $EM_n^{(0)} = EM_n^{(1)} = 6.221$. With the sample size of 213, according to Table 3.1, p_n will be well approximated by 0.427.

In view of the adjusted limiting distribution $0.573\chi_0^2 + 0.427\chi_1^2$, the asymptotic p-value for EM-test is 0.005. For the constrained LRT, we have $R_n(0.45) = 6.30$ with the asymptotic p-value 0.005. We also calculate the LRT statistic, $R_n = 6.31$. We simulated the quantiles of the LRT statistic with 10,000 repetitions and found the simulated p-value for the LRT is 0.019. So for the pooled failure data, the EM-test and the constrained LRT give stronger evidence than the LRT to reject the homogeneous exponential fit. We should note that the above analysis only tells us the two-component exponential mixture model provides a more suitable fitting for the pooled data than the homogeneous exponential model. If we want to know what is the order of the mixture if the finite mixture model is used or which model, the finite mixture model or the mixture model with continuous mixing distribution, is more suitable for the pooled data, further analysis needs to be conducted.

Example 3.4.4. The second example considers the failure times of a computer in 257 unspecified units, which can be found in Cox and Lewis (1968) Table 1.3. Since the units are unspecified, there may exist the heterogeneity in the failure times. Our interest is to test whether this heterogeneity can be easily detected or not. We conducted the test of homogeneity and found that the asymptotic p-value of the EM-test and the constrained LRT are all less than 0.00001. Figure 3.3 shows the plot of the log empirical survival curve and the log theoretical survival curve under the homogeneous exponential model. Clearly, the homogeneous exponential model does not fit the data. Yet the pool fit might be purely caused by the two largest observations. Hence we remove these two observations and reanalyzed the remaining data. The homogeneous model is rejected by all methods with very small p-values. To better show the difference between the EM-test and other methods, we delete the largest 7 observations and reanalyze the remaining data. The EM-test statistics with $C^* = 1.5$ and three initial values (0.1, 0.3, 0.5) of α, are $EM_n^{(0)} = 10.484$ and $EM_n^{(1)} = 10.582$, with the asymptotic p-values 0.0005. The constrained LRT is found to be $R_n(0.45) = 0.632$, with the asymptotic p-value 0.185. We also calculate the asymptotic p-value for the $C(\alpha)$ test, which is found to be 0.442. In this case, the EM-test has much stronger evidence than the constrained LRT and the $C(\alpha)$ test to reject the homogeneous exponential model.

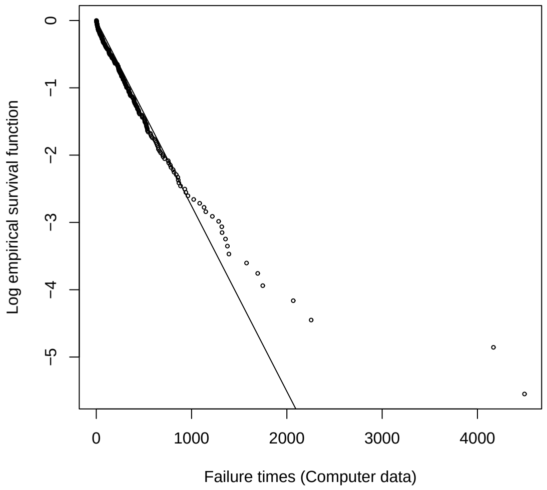

Figure 3.3: The log empirical (points) and null theoretical (solid line) survival functions of the computer data.

3.5 Appendix: Regularity Conditions and Technical Proofs

Regularity Conditions for the EM-test

The proofs are based on the following regularity conditions on the kernel density function.

B1. *Wald's integrability conditions.* (i) $E|\log f(X;\theta_0)| < \infty$, and (ii) for sufficiently small ρ and for sufficiently large r, the expected values $E\log\{1+f(X;\theta,\rho)\} < \infty$ for $\theta \in \Theta$ and $E\log\{1+\varphi(X,r)\} < \infty$, where

$$f(x;\theta,\rho) = \sup_{|\theta'-\theta|\leq\rho} f(x;\theta')$$

and

$$\varphi(x;r) = \sup_{|\theta|\geq r} f(x;\theta).$$

(iii) $\lim_{|\theta|\to\infty} f(x;\theta) = 0$ for all x except on a set with probability zero.

B2. *Smoothness.* The kernel function $f(x;\theta)$ has common support and is three times continuously differentiable with respect to θ. The first two derivatives are denoted by $f'(x;\theta)$ and $f''(x;\theta)$.

B3. *Identifiability.* For any two mixing distribution functions Ψ_1 and Ψ_2 with two supporting points such that

$$\int f(x;\theta)d\Psi_1(\theta) = \int f(x;\theta)d\Psi_2(\theta), \text{ for all } x,$$

we must have $\Psi_1 = \Psi_2$.

B4. *Strong law of large numbers.* For some neighborhood $N(\theta_0)$ of θ_0, there exists a g with finite expectation such that

$$|Y_i(\theta)|^3 \leq g(X_i),\ |Z_i(\theta)|^3 \leq g(X_i) \text{ and } |Z_i''(\theta)|^2 \leq g(X_i).$$

B5. *Positive definite.* The covariance matrix of (Y_i, Z_i) is positive definite.

Remark 3.5.1. *The conditions imposed for the EM-test are markedly weakened when compared to the MLRT or other likelihood based methods. Firstly, the parameter space of* θ

need not be bounded any more. Secondly, we only require $E\{Y_i(\theta)^3\} < \infty$ for θ in a small neighborhood of θ_0.

Techinical Proofs

We put two preliminary results as two lemmas here first. The first is seen as the extension of the results in Wald (1949).

Lemma 3.5.1. *Suppose that Condition B1 holds. Let $(\bar{\alpha}, \bar{\theta}_1, \bar{\theta}_2)$ be some estimators of $(\alpha, \theta_1, \theta_2)$ such that $\delta \leq \bar{\alpha} \leq 1 - \delta$ for some $\delta \in (0, 0.5]$. Assume that*

$$l_n(\bar{\alpha}, \bar{\theta}_1, \bar{\theta}_2) - l_n(0.5, \theta_0, \theta_0) \geq c > -\infty.$$

Then under null distribution $f(x; \theta_0)$, $\bar{\theta}_1 - \theta_0 = o_p(1)$ and $\bar{\theta}_2 - \theta_0 = o_p(1)$.

Proof. The parameter space under the full model (2.1) with the restriction becomes

$$\Lambda = [\delta, 1 - \delta] \times \Theta \times \Theta.$$

When the null model is true, the "true" parameter values form the set $\{(\alpha, \theta_0, \theta_0) : \delta \leq \alpha \leq 1 - \delta\}$.

Our proof in principle is similar to that of Wald (1949), but has some important differences.

First, for some positive constants ϵ and r, let

$$A(\alpha; \epsilon, r) = \{(\alpha', \theta_1, \theta_2) \in \Lambda; |\alpha' - \alpha| \leq \epsilon, |\theta_1| > r, |\theta_2| > r\}$$

and define

$$\psi(x; \alpha, \epsilon, r) = \sup\{\alpha' f(x; \theta_1') + (1 - \alpha') f(x; \theta_2,) : (\alpha', \theta_1', \theta_2') \in A(\alpha; \epsilon, r)\}.$$

By Condition B1, it is obvious that for all small enough ϵ and large enough r,

$$E\{\log \psi(X;\epsilon, r)\} < E\{\log f(X;\theta_0)\}$$

under the null distribution $f(X;\theta_0)$. Hence, by the law of large numbers,

$$\Pr[\sup\{l_n(\alpha', \theta_1', \theta_2') : A(\alpha;\epsilon, r)\} - l_n(\alpha, \theta_0, \theta_0) > c] \to 0$$

almost surely for any $c > -\infty$. By the classical arguments on the compact set $[\delta, 1-\delta]$ for α, the above conclusion is easily extended to

$$\Pr[\sup\{l_n(\alpha', \theta_1', \theta_2') : (\alpha', \theta_1', \theta_2') \in A\} - l_n(\alpha, \theta_0, \theta_0) > c] \to 0$$

where $A = \cup_{\delta \leq \alpha \leq 1-\delta} A(\alpha;\epsilon, r)$.

Next, the same conclusion and the proof are applicable to

$$B(\alpha, \theta_1;\epsilon, r) = \{(\alpha, \theta_1', \theta_2) \in \Lambda; |\alpha' - \alpha| \leq \epsilon, |\theta_1' - \theta_1| < \epsilon, |\theta_2| > r\}$$

and hence also to

$$B = \cup\{B(\alpha, \theta_1;\epsilon, r) : \delta \leq \alpha \leq 1-\delta, |\theta_1| \leq r\}.$$

In plain words, the log-likelihood at any parameter point with either one of θ_1 and θ_2 very large trails the log-likelihood at the true parameter point by an infinite amount.

What left is to prove the same conclusion for parameter points in the compact complement of $A \cup B$ but outside any small neighborhood of $(\alpha, \theta_0, \theta_0)$. However, this is the same as the classical consistent result by Wald (1949).

We hence conclude the proof. □

Lemma 3.5.2. *Suppose the same conditions of Theorem 3.2.1 on $f(x;\theta)$ and $p(\alpha)$ hold. Let $(\bar{\alpha}, \bar{\theta}_1, \bar{\theta}_2)$ be some estimators of $(\alpha, \theta_1, \theta_2)$ such that under the null hypothesis,*

$$\bar{\theta}_1 - \theta_0 = o_p(1) \text{ and } \bar{\theta}_2 - \theta_0 = o_p(1)$$

and $\bar{\alpha} - \alpha_0 = o_p(1)$ *for some* $\alpha_0 \in (0, 0.5]$. *If for all* n *and* $X_1, \ldots, X_n$,

$$pl_n(\bar{\alpha}, \bar{\theta}_1, \bar{\theta}_2) - pl_n(0.5, \theta_0, \theta_0) \geq c > -\infty,$$

then under the null distribution $f(x; \theta_0)$,

$$\bar{\theta}_1 - \theta_0 = O_p(n^{-1/4}),\ \bar{\theta}_2 - \theta_0 = O_p(n^{-1/4})$$

and

$$\bar{m}_1 = (1 - \bar{\alpha})(\bar{\theta}_1 - \theta_0) + \bar{\alpha}(\bar{\theta}_2 - \theta_0) = O_p(n^{-1/2}).$$

Proof. For $i = 1, \ldots, n$, let $W_i = Z_i - \beta Y_i$ with $\beta = E(Y_1 Z_1)/E(Y_1^2)$. Note that Y_i and Z_i are defined in (2.4) and (2.5), respectively. Further, let $\bar{m} = \bar{m}_1 + \beta \bar{m}_2$ with $\bar{m}_2 = (1 - \bar{\alpha})(\bar{\theta}_1 - \theta_0)^2 + \bar{\alpha}(\bar{\theta}_2 - \theta_0)^2$.

By the condition of the lemma, $\bar{\theta}_1$ and $\bar{\theta}_2$ are in a small neighborhood of θ_0 in probability. Therefore, by the Taylor's expansion at θ_0, we get:

$$\begin{aligned} 2\{pl_n(\bar{\alpha}, \bar{\theta}_1, \bar{\theta}_2) - pl_n(0.5, \theta_0, \theta_0)\} &\leq 2\{l_n(\bar{\alpha}, \bar{\theta}_1, \bar{\theta}_2) - l_n(0.5, \theta_0, \theta_0)\} \\ &\leq 2\sum_{i=1}^{n}\{\bar{m} Y_i + \bar{m}_2 W_i\} - \{\bar{m}^2 \sum_{i=1}^{n} Y_i^2 + \bar{m}_2^2 \sum_{i=1}^{n} W_i^2\}(1 + o_p(1)) + o_p(1) \\ &\leq \frac{\{(\sum_{i=1}^{n} W_i)^+\}^2}{\sum_{i=1}^{n} W_i^2} + \frac{(\sum_{i=1}^{n} Y_i)^2}{\sum_{i=1}^{n} Y_i^2} + o_p(1). \end{aligned} \tag{3.2}$$

We do not have cross terms in the second line because Y_i and W_i are uncorrelated. The last inequality is simply the property of the quadratic function and the non-negativeness of $\bar{m}_2$.

Together with the condition that $pl_n(\bar{\alpha}, \bar{\theta}_1, \bar{\theta}_2) - pl_n(0.5, \theta_0, \theta_0) \geq c$, the above inequality implies that

$$2\bar{m}_2 \sum_{i=1}^{n} W_i - \bar{m}_2^2 \{\sum_{i=1}^{n} W_i^2\}\{1 + o_p(1)\} = O_p(1).$$

Because $\sum_{i=1}^n W_i = O_p(n^{1/2})$ and $\sum_{i=1}^n W_i^2 = O_p(n)$, we get $\bar{m}_2 = O_p(n^{-1/2})$. Due to the condition that $\bar{\alpha} = \alpha_0 + o_p(1)$ and $0 < \alpha_0 \le 0.5$, we further conclude that

$$\bar{\theta}_1 - \theta_0 = O_p(n^{-1/4}) \text{ and } \bar{\theta}_2 - \theta_0 = O_p(n^{-1/4}).$$

Similarly, we have $\bar{m} = O_p(n^{-1/2})$ and therefore $\bar{m}_1 = O_p(n^{-1/2})$. These conclude the proof. □

Let $(\bar{\alpha}, \bar{\theta}_1, \bar{\theta}_2)$ be some estimators of $(\alpha, \theta_1, \theta_2)$ with the asymptotic properties as before. Define

$$\begin{aligned} Q_n(\alpha) &= (n - \sum_{i=1}^n \bar{w}_i)\log(1-\alpha) + \sum_{i=1}^n \bar{w}_i \log(\alpha) + p(\alpha) \\ &= R_n(\alpha) + p(\alpha) \end{aligned}$$

with

$$\bar{w}_i = \frac{\bar{\alpha} f(X_i;\bar{\theta}_2)}{(1-\bar{\alpha})f(X_i;\bar{\theta}_1) + \bar{\alpha} f(X_i;\bar{\theta}_2)}.$$

Let $\bar{\alpha}^* = \arg\max_\alpha Q_n(\alpha)$. The following lemma considers some asymptotic properties regarding $\bar{\alpha}^*$.

Lemma 3.5.3. *Suppose the same conditions of Lemma 3.5.2 hold. Under the null distribution $f(x;\theta_0)$, we have $|\bar{\alpha}^* - \alpha_0| = o_p(1)$.*

Proof. For $i = 1, 2, \ldots, n$, let

$$\begin{aligned} \bar{\delta}_i &= (1-\bar{\alpha})\Big\{\frac{f(X_i;\bar{\theta}_1)}{f(X_i;\theta_0)} - 1\Big\} + \bar{\alpha}\Big\{\frac{f(X_i;\bar{\theta}_2)}{f(X_i;\theta_0)} - 1\Big\} \\ &= \bar{m}_1 Y_i + (1-\bar{\alpha})(\bar{\theta}_1 - \theta_0)^2 Z_i(\bar{\theta}_1) + \bar{\alpha}(\bar{\theta}_2 - \theta_0)^2 Z_i(\bar{\theta}_2), \end{aligned}$$

where Y_i and Z_i are defined in (2.4) and (2.5). Thus,

$$\max_{1\le i\le n} |\bar{\delta}_i| \le |\bar{m}_1| \max_{1\le i\le n} |Y_i| + \bar{m}_2 \max_{1\le i\le n}\{\sup_{\theta\in N(\theta_0)} |Z_i(\theta)|\}.$$

By Condition B4 and a result on order statistic in Serfling (1980, page 90), we have

$$\max_{1\le i\le n}\{\sup_{\theta\in N(\theta_0)}|Z_i(\theta)|\} = o_p(n^{1/2}) \text{ and } \max_{1\le i\le n}|Y_i| = o_p(n^{1/2}).$$

Consequently, we have $\max_i|\delta_i| = o_p(1)$.

Expanding $f(X_i;\bar\theta_j)$ at $\bar\theta_j=\theta_0$, $j=1,2$, we get

$$\begin{aligned}\bar w_i-\bar\alpha &= \bar\alpha(1-\bar\alpha)\frac{f(X_i;\bar\theta_2)-f(X_i;\bar\theta_1)}{(1-\bar\alpha)f(X_i;\bar\theta_1)+\bar\alpha f(X_i;\bar\theta_2)}\\ &= \frac{\bar\alpha(1-\bar\alpha)}{1+\delta_i}\{(\bar\theta_2-\bar\theta_1)Y_i+(\bar\theta_2-\theta_0)^2Z_i(\bar\theta_2)-(\bar\theta_1-\theta_0)^2Z_i(\bar\theta_1)\}.\end{aligned}$$

Hence, putting $\tilde\alpha = n^{-1}\sum_{i=1}^n\bar w_i$, we have

$$\begin{aligned}&|\tilde\alpha-\bar\alpha|\\ =\ &\Big\{(\bar\theta_2-\bar\theta_1)\sum_{i=1}^nY_i+(\bar\theta_2-\theta_0)^2\sum_{i=1}^nZ_i(\bar\theta_2)-(\bar\theta_1-\theta_0)^2\sum_{i=1}^nZ_i(\bar\theta_1)\Big\}O_p(n^{-1})\\ =\ &o_p(1).\end{aligned}$$

By this result and the assumption that $\bar\alpha-\alpha_0=o_p(1)$, we have

$$\tilde\alpha-\alpha_0=o_p(1)$$

and hence it suffices to prove that $\bar\alpha^*-\tilde\alpha=o_p(1)$.

Note that $R_n(\alpha)$ is a binomial log-likelihood. It attains its maximum at and decreases from $\tilde\alpha$ in both directions. For any $\epsilon>0$ and $\alpha\ge\tilde\alpha+2\epsilon$, by the mean value theorem,

$$R_n(\alpha)-R_n(\tilde\alpha)\le R_n(\tilde\alpha+2\epsilon)-R_n(\tilde\alpha+\epsilon)=\epsilon R_n'(\xi)$$

for some $\xi\in[\tilde\alpha+\epsilon,\tilde\alpha+2\epsilon]$. It is easy to verify that $R_n'(\xi)\to-\infty$ in probability as $n\to\infty$ uniformly for ξ in this range. On the other hand, we have

$$p(\alpha)-p(\tilde\alpha)=p(\alpha)-p(\alpha_0)+o_p(1)=O_p(1).$$

Hence, with probability approaching 1,

$$Q_n(\alpha) - Q_n(\tilde{\alpha}) = R_n(\alpha) - R_n(\tilde{\alpha}) + \{p(\alpha) - p(\tilde{\alpha})\} \to -\infty$$

uniformly for any $\alpha > \tilde{\alpha} + 2\epsilon$. Hence, we must have $\bar{\alpha}^* < \tilde{\alpha} + 2\epsilon$ in probability. Similarly, we can show that $\bar{\alpha}^* > \tilde{\alpha} - 2\epsilon$ in probability. Therefore, we have $\bar{\alpha}^* = \tilde{\alpha} + o_p(1)$ as claimed. □

Proof of Theorem 3.2.1

By the property of the EM algorithm (Dempster et al. 1977), the definition of $\alpha_j^{(k)}$ and others, for any finite k, we have

$$pl_n(\alpha_j^{(k)}, \theta_{j1}^{(k)}, \theta_{j2}^{(k)}) \geq pl_n(\alpha_j^{(0)}, \theta_{j1}^{(0)}, \theta_{j2}^{(0)}) \geq pl_n(\alpha_j, \theta_0, \theta_0).$$

Therefore

$$\begin{aligned} l_n(\alpha_j^{(0)}, \theta_{j1}^{(0)}, \theta_{j2}^{0)}) - l_n(\alpha_j^{(0)}, \theta_0, \theta_0) &\geq p(\alpha_j^{(0)}) - p(\alpha_j^{(k)}) \\ &\geq p(\alpha_j) - p(0.5) > -\infty. \end{aligned}$$

By Lemma 3.5.1, we have shown that $\theta_{j1}^{(0)}$ and $\theta_{j2}^{0)}$ are consistent for θ_0. Because of this, the conclusions of Lemmas 3.5.2 and 3.5.3 apply. Hence, we find

$$\alpha_j^{(1)} - \alpha_j = o_p(1);$$

and both

$$\theta_{j1}^{(1)} - \theta_0 = O_p(n^{-1/4}); \quad \theta_{j2}^{(1)} - \theta_0 = O_p(n^{-1/4}).$$

The above conclusions can then be used to show the same conclusions are true for $\alpha_j^{(2)}$, $\theta_{j1}^{(2)}$ and $\theta_{j2}^{(2)}$. By mathematical induction, the conclusion of the theorem is true for all finite k. □

Proof of Theorem 3.2.2

Due the properties proved in Theorem 3.2.1, the inequality (3.2) is applicable. Hence for any (j, k), we have

$$2\{pl_n(\alpha_j^{(k)}, \theta_{j1}^{(k)}, \theta_{j2}^{(k)}) - pl_n(0.5, \theta_0, \theta_0)\} \leq \frac{\{(\sum_{i=1}^n W_i)^+\}^2}{\sum_{i=1}^n W_i^2} + \frac{(\sum_{i=1}^n Y_i)^2}{\sum_{i=1}^n Y_i^2} + o_p(1).$$

It is obvious that

$$2\{\sup_{\theta\in\Theta} pl_n(0.5, \theta, \theta) - pl_n(0.5, \theta_0, \theta_0)\} = \frac{(\sum_{i=1}^n Y_i)^2}{\sum_{i=1}^n Y_i^2} + o_p(1).$$

Hence, we have

$$2\{pl_n(\alpha_j^{(k)}, \theta_{j1}^{(k)}, \theta_{j2}^{(k)}) - \sup_{\theta\in\Theta} pl_n(0.5, \theta, \theta)\} \leq \frac{\{(\sum_{i=1}^n W_i)^+\}^2}{\sum_{i=1}^n W_i^2} + o_p(1).$$

At the same time, it is simple to show that

$$2\{pl_n(\alpha_j^{(k)}, \theta_{j1}^{(k)}, \theta_{j2}^{(k)}) - \sup_{\theta\in\Theta} pl_n(0.5, \theta, \theta)\} \geq \frac{\{(\sum_{i=1}^n W_i)^+\}^2}{\sum_{i=1}^n W_i^2} + o_p(1)$$

when $\alpha_j = 0.5$. Thus,

$$EM_n^{(k)} = \frac{\{(\sum W_i)^+\}^2}{\sum W_i^2} + o_p(1).$$

Consequently, the limiting distribution is given by $0.5\chi_0^2 + 0.5\chi_1^2$. □

Chapter 4

EM-test in Normal Mixture Models

4.1 Introduction

After its first application in Pearson (1894), the normal mixture model has become one of the most popular model in real data analysis, see Reoder (1994), McLachlan and Peel (200), Chen and Chen (2003), Chen and Kalbfleisch (2005), Tadesse et al. (2005) and Früwirth-Schnatter (2006).

An example of great interest arises in genetics, in which one wants to know if there exists a major gene corresponding to a trait of interest. Assume that this major gene (if exists) has two possible alleles (say A and a) and further assume that A is dominant over a. Suppose the phenotypes or the trait values for the individuals with and without allele A are distributed by $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ distributions, respectively. Then if the major gene exits, the phenotype of a randomly selected individual from the population follows a two-component normal mixture model; if not, it follows uni-component normal distribution. The test for the presence of a major gene in principle is the test of homogeneity under the normal mixture model.

The test of homogeneity under normal mixture models is an important and challenging problem. Due to the lack of strong identifiability, the asymptotic properties of the LRT and the MLRT under normal mixture models are substantially different from those under regular one-parameter mixture models. The asymptotic properties of the LRT and MLRT have attracted the attention of many statisticians. Chen and Chen (2003) is the first paper addressing the asymptotic property of the ordinary likelihood test statistic under the univariate normal mixture model in the presence of a structure parameter ($\sigma_1^2 = \sigma_2^2 = \sigma^2$). Chen and Kalbfleish (2005) investigated the use of the MLRT under the same model, and found the MLRT does not have a simple limiting distribution, but has a very useful stochastic upper bound which has χ_2^2 distribution. Qin and Smith (2004) investigated the same problem and claimed that the limiting distribution for the MLRT is $0.5\chi_1^2 + 0.5\chi_2^2$. Yet this limiting distribution does not fit well with the finite sample distribution of the MLRT. Our simulations on normal mixture models reveal that the zero-proportions for the MLRT statistic under the null hypothesis are around 20% when the sample sizes are between 100 and 200. Clearly, this limiting distribution is not very useful to determine the critical value of the MLRT. In comparison, the χ_2^2 distribution is found to provide a good fit for the tail distribution of the MLRT (Chen and Kalbfleisch, 2005).

For normal mixture models in both mean and variance parameters, we could not find many theoretical results on the limiting distribution of the LRT or the MLRT. The general results in Dacunha-Castelle and Gassiat (1999) and others are not applicable to normal mixture models. Most publications focus on simulating the quantiles of the LRT statistics. Wolfe (1971) suggested that the χ_4^2 distribution provides a good approximation for the null limiting distribution of the LRT. However, McLachlan (1987) presented simulation results and suggested that the χ_6^2 distribution provides a better fit to the null limiting distribution of the LRT than the χ_4^2. Feng and McCulloch (1994) found that the null distribution of the LRT depends on a lower bound placed on the component variances.

In this chapter, we investigate the use of the EM-test to normal mixture models and derived its limiting distributions under two situations. For the test of homogeneity in the presence of the structural parameter ($\sigma_1^2 = \sigma_2^2 = \sigma^2$), the limiting distribution is a simple function of $0.5\chi_0^2 + 0.5\chi_1^2$ and χ_1^2 distributions. The test with this limiting distribution is still very convenient to implement. Its accuracy is examined with extensive simulations, and is very satisfactory. The power of the test is comparable to the MLRT coupled with χ_2^2 distribution. For normal mixture models in both mean and variance parameters, the limiting distribution of the EM-test is found be χ_2^2. Simulations are also conducted to confirm that the quantiles of limiting distribution provide accurate critical values of the EM-test. We include a real data example to illustrate the use of the EM-test.

4.2 Normal Mixture Models in the Presence of the Structural Parameter

4.2.1 The EM-test Procedure

Suppose $X_1, \ldots, X_n$ is a random sample from

$$(1-\alpha)N(\mu_1, \sigma^2) + \alpha N(\mu_2, \sigma^2).$$

We are interested in testing

$$H_0 : \alpha(1-\alpha)(\mu_1 - \mu_2) = 0. \tag{4.1}$$

For the above testing problem, the log-likelihood function is given by

$$l_n(\alpha, \mu_1, \mu_2, \sigma) = \sum_{i=1}^{n} \log\{(1-\alpha)f(X_i; \mu_1, \sigma) + \alpha f(X_i; \mu_2, \sigma)\},$$

where $f(x;\mu,\sigma)$ is the probability density function of normal distribution with mean μ and variance σ^2.

For the testing problem (4.1), Chen and Chen (2003) first noted that normal density function is not strongly identifiable, which is a consequence of

$$\frac{\partial f^2(x;\mu,\sigma)}{\partial \mu^2}|_{(\mu,\sigma^2)=(0,1)} = 2\frac{\partial f(x;\mu,\sigma)}{\partial(\sigma^2)}|_{(\mu,\sigma^2)=(0,1)}.$$

A direct effect of the above equality is that μ_1^2, μ_2^2 and σ^2 are fully confounded together apart from the confounding between μ_1 and μ_2. So the first moment and the second moment of the mixture model can not uniquely determine the values of μ_1^2, μ_2^2 and σ^2. To detach them, we need to go for the third and forth moments. Because of this, Chen and Chen (2003) found that the convergence rates for the MLEs of the mixing distribution and σ^2 under the null model are $O_p(n^{-1/8})$ and $O_p(n^{-1/4})$, respectively, which imposes substantial technical difficulties in the study of asymptotic properties. Note that the variance of the mixture distribution is the sum of the component variance σ^2 and the variance of the mixing distribution. Fitting a mixture model to data arising from a uni-component normal model tends to result in a smaller fitted component variance. This bias effect tends to make the LRT or the MLRT somewhat liberal, which is one possible reason that the upper bound χ_2^2 provides a good fit for the tail distribution of the MLRT. If we adjust the bias of the estimation of σ^2 under the null model, it will make the estimation of σ^2 biased under the alternative model.

As discussed in Chapter 3, under single parameter mixture models, the limiting distribution of the EM-test does not require the compactness of the parameter space and has comparable power to the MLRT. With the additional parameter, the asymptotic property of the EM-test obtained in Chapter 3 can not be directly used. It is of great interest to study the application of the EM-test to the testing problem (4.1).

To overcome the under-estimation effect, we define the modified log-likelihood function for the EM-test as follows:

$$pl_n(\alpha, \mu_1, \mu_2, \sigma) = l_n(\alpha, \mu_1, \mu_2, \sigma) + p_n(\sigma) + p(\alpha),$$

where $p_n(\sigma)$ is a penalty function on σ^2, which will be allowed to depend on the data, and $p(\alpha)$ is a penalty function on α such that $p(\alpha)$ achieves the maximal value at $\alpha = 0.5$. The role of the penalty function $p(\alpha)$ has been discussed in Chapter 2. The penalty function $p_n(\sigma)$ will be selected to prevent the underestimation of σ^2 under the null model.

The idea of the EM-test in Chapter 3 can be applied to normal mixture models in exactly the same way. For given $\alpha = \alpha_0$, we compute a modified likelihood ratio test statistic as follows

$$M_n(\alpha_0) = 2\{pl_n(\alpha_0, \mu_{01}^{(0)}, \mu_{02}^{(0)}, \sigma^{(0)}) - pl_n(0.5, \hat{\mu}_0, \hat{\mu}_0, \hat{\sigma}_0)\},$$

with $(\mu_{01}^{(0)}, \mu_{02}^{(0)}, \sigma^{(0)})$ being the maximizer of $pl_n(\alpha_0, \mu_1, \mu_2, \sigma)$ and $(\hat{\mu}_0, \hat{\sigma}_0)$ being the maximizer of $pl_n(0.5, \mu, \mu, \sigma)$. Several EM iterations will be applied between α and (μ_1, μ_2, σ) to capture more relevant parameter values if the data are from an alternative model. The value of $M_n(\alpha_0)$ will be updated accordingly. Several α_0 will be used simultaneously to accelerate the process. The EM-test statistic will be defined by the maximum value of these outcomes with a given number of iterations.

The analytic form of the EM-test statistic is not convenient to present. Instead, we provide the following pseudo code.

Step 0. Choose a number of initial α values, say $\alpha_1, \alpha_2, \ldots, \alpha_J \in (0, 0.5]$. Compute

$$(\hat{\mu}_0, \hat{\sigma}_0) = \arg\max_{\mu,\, \sigma^2} pl_n(1/2, \mu, \mu, \sigma).$$

Let $j = 1, k = 0$.

Step 1. Let $\alpha_j^{(k)} = \alpha_j$.

Step 2. Compute

$$(\mu_{j1}^{(k)}, \mu_{j2}^{(k)}, \sigma_j^{(k)}) = \arg\max_{\mu_1,\ \mu_2,\ \sigma} pl_n(\alpha_j^{(k)}, \mu_1,\ \mu_2,\ \sigma)$$

and

$$M_n^{(k)}(\alpha_j) = 2\{pl_n(\alpha_j^{(k)}, \mu_{j1}^{(k)}, \mu_{j2}^{(k)}, \sigma_j^{(k)}) - pl_n(1/2, \hat{\mu}_0, \hat{\mu}_0, \hat{\sigma}_0)\}.$$

Step 3. For $i = 1, 2, \ldots, n$, compute the weights which are the conditional expectations in the E-step.

$$w_{ij}^{(k)} = \frac{\alpha_j^{(k)} f(X_i; \mu_{j2}^{(k)}, \sigma_j^{(k)})}{(1-\alpha_j^{(k)}) f(X_i; \mu_{j1}^{(k)}, \sigma_j^{(k)}) + \alpha_j^{(k)} f(X_i; \mu_{j2}^{(k)}, \sigma_j^{(k)})}.$$

Now following the M-step, let

$$\begin{aligned}
\alpha_j^{(k+1)} &= \arg\max_{\alpha}\{(n - \sum_{i=1}^{n} w_{ij}^{(k)}) \log(1-\alpha) + \sum_{i=1}^{n} w_{ij}^{(k)} \log(\alpha) + p(\alpha)\}, \\
\mu_{j1}^{(k+1)} &= \sum_{i=1}^{n}(1 - w_{ij}^{(k)}) X_i \Big/ \sum_{i=1}^{n}(1 - w_{ij}^{(k)}), \\
\mu_{j2}^{(k+1)} &= \sum_{i=1}^{n} w_{ij}^{(k)} X_i \Big/ \sum_{i=1}^{n} w_{ij}^{(k)}, \\
\sigma_j^{(k+1)} &= \arg\max_{\sigma}\Big\{ -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(1 - w_{ij}^{(k)})(X_i - \mu_{j1}^{(k+1)})^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n} w_{ij}^{(k)}(X_i - \mu_{j2}^{(k+1)})^2 \\
&\quad -\frac{n}{2}\log\sigma^2 + p_n(\sigma)\Big\}.
\end{aligned}$$

Compute

$$M_n^{(k+1)}(\alpha_j) = 2\{pl_n(\alpha_j^{(k+1)}, \mu_{j1}^{(k+1)}, \mu_{j2}^{(k+1)}, \sigma_j^{(k+1)}) - pl_n(1/2, \hat{\mu}_0, \hat{\mu}_0, \hat{\sigma}_0)\}.$$

Let $k = k + 1$ and repeat Step 3 for a fixed number of iterations in k.

Step 4. Let $j = j + 1$, $k = 0$ and go to Step 1, until $j = J$.

Step 5. For each k, calculate the test statistic as

$$EM_n^{(k)} = \max\{M_n^{(k)}(\alpha_j), j = 1, 2, \ldots, J\}.$$

The homogeneous model is rejected when the EM-test statistic is larger than some critical value. The critical value will be determined by the limiting distribution of the EM-test statistic, which will be studied in the next subsection.

4.2.2 Asymptotic Behavior of the EM-test

We study the asymptotic properties of the EM-test under the following conditions on the penalty function $p_n(\sigma)$.

C1. $p_n(a\sigma; aX_1 + b, \ldots, aX_n + b) = p_n(\sigma; X_1, \ldots, X_n)$.

C2. $\sup_{\sigma>0} \max\{0, p_n(\sigma)\} = o(n)$ and $p_n(1) = o(n)$.

C3. $p_n'(\sigma) = o_p(n^{1/4})$ for $\sigma \in N(1)$ with $p_n'(\sigma)$ being the first derivative of $p_n(\sigma)$ and $N(1)$ being a small neighborhood of 1 .

Remark 4.2.1. *Under Condition C1, the EM-test has the invariance property, which is a desirable property of statistical inference for location-scale models. Condition C2 controls the influence of $p_n(\sigma)$ on the log-likelihood function.*

The following theorem assesses the asymptotic orders of $(\alpha_j^{(k)}, \mu_{j1}^{(k)}, \mu_{j2}^{(k)}, \sigma_j^{(k)})$ under the null hypothesis. The proofs will be given in Section 4.2.4.

Theorem 4.2.1. *Suppose Conditions C1-C3 hold, and $p(\alpha)$ is a continuous function such that $p(\alpha) \to -\infty$ as $\alpha \to 0$ and it attains its maximal value at $\alpha = 0.5$. Under the null distribution $N(\mu_0, \sigma_0^2)$, we have, for $j = 1, \ldots, J$ and any fixed finite k,*

(a) *if* $\alpha_j = 0.5$, *then*

$$\begin{aligned} \alpha_j^{(k)} - \alpha_j &= O_p(n^{-1/4}), \\ \mu_{j1}^{(k)} - \mu_0 &= O_p(n^{-1/8}), \\ \mu_{j2}^{(k)} - \mu_0 &= O_p(n^{-1/8}), \\ \sigma_j^{(k)} - \sigma_0 &= O_p(n^{-1/4}); \end{aligned}$$

(b) *if* $\alpha_j \neq 0.5$, *then*

$$\begin{aligned} \alpha_j^{(k)} - \alpha_j &= O_p(n^{-1/4}), \\ \mu_{j1}^{(k)} - \mu_0 &= O_p(n^{-1/6}), \\ \mu_{j2}^{(k)} - \mu_0 &= O_p(n^{-1/6}), \\ \sigma_j^{(k)} - \sigma_0 &= O_p(n^{-1/3}). \end{aligned}$$

It is of interest to note that the convergence rates of $(\mu_{j1}^{(k)}, \mu_{j2}^{(k)}, \sigma_j^{(k)})$ depend on the choice of initial α value. The reason behind this strange phenomenon is the loss of the strong identifiability of normal mixture models. More discussions will be given after the next theorem. When $\alpha_j = 0.5$, after several iterations, $\alpha_j^{(k)}$ is no longer 0.5. However the problem does not reduce to Case (b), because $\alpha_j^{(k)}$ is still in a very small neighborhood of 0.5, while in Case (b), α_j is outside a fixed small neighborhood of 0.5.

Based on the results in Theorem 4.2.1, we derive the limiting distribution of $EM_n^{(k)}$ and the proof will also be deferred to Section 4.2.4.

Theorem 4.2.2. *Assume the same conditions as in Theorem 4.2.1, and that one of* α_j*'s is equal to 0.5. Then under null distribution* $N(\mu_0, \sigma_0^2)$*, and for any fixed finite* k*, as* $n \to \infty$*,*

$$\Pr(EM_n^{(k)} \leq x) \to F(x - \Delta)\{0.5 + 0.5F(x)\},$$

where $F(x)$ is the cumulative density function (cdf) for χ_1^2 and

$$\Delta = 2 \max_{\alpha_j \neq 0.5} \{p(\alpha_j) - p(0.5)\}.$$

The results in Theorems 4.2.1 and 4.2.2 needs some interpretations. We can understand them from the point view of moments. Without loss of generality, we assume that the mean and variance of the normal mixture model $(1-\alpha)N(\mu_1, \sigma^2) + \alpha N(\mu_2, \sigma^2)$ are 0 and 1, respectively. The test of homogeneity is to choose one model between $N(0, 1)$ and $(1-\alpha)N(\mu_1, \sigma^2) + \alpha N(\mu_2, \sigma^2)$ with

$$(1-\alpha)\mu_1 + \alpha\mu_2 = 0 \text{ and } (1-\alpha)\mu_1^2 + \alpha\mu_2^2 + \sigma^2 = 1.$$

First, let us consider the case when α is fixed to be 0.5. Since the first moment of the mixture model is equal to 0, the mixture density is symmetric and the third moment of the mixture model is 0. Therefore the third moment can not tell the difference between the homogeneous model and the mixture model. After some calculations, the forth moment of the mixture model is found to be

$$E(X_1^4) = 3 - (\mu_1^4 + \mu_2^4), \tag{4.2}$$

which is smaller than 3 or equal to 3 with the equality holding when the mixture model reduces to the homogeneous model. Hence the test of homogeneity is equivalent to testing

$$H_0 : E(X_1^4) = 3 \text{ versus } H_a : E(X_1^4) < 3.$$

The null hypothesis is on the boundary of the parameter space, so the limiting distribution of $M_n(0.5)$ is $0.5\chi_0^2 + 0.5\chi_1^2$. When α is fixed to be $\alpha_0 \in (0, 0.5)$, we can tell the difference between the homogeneous model and the mixture model from the third moment. Note that

$$E(X_1^3) = (1-\alpha_0)\mu_1^3 + \alpha_0\mu_2^3,$$

which can be greater than 0 or less than 0 and becomes 0 when the mixture model is the homogeneous model. Therefore, the test of homogeneity is equivalent to testing

$$H_0 : E(X_1^3) = 0 \text{ versus } H_a : E(X_1^3) \neq 0.$$

In this case, the null hypothesis is the interior point of the parameter space, so $M_n(\alpha_0)$ has the asymptotic distribution $\chi_1^2 + 2\{p(\alpha_0) - p(0.5)\}$. The term $2\{p(\alpha_0) - p(0.5)\}$ is due to the penalty function. Since the third moment and the forth moment are asymptotically orthogonal, the limiting distribution of the EM-test involves the maximum of two independent distributions: the χ_1^2 and the $0.5\chi_0^2 + 0.5\chi_1^2$. The presence of Δ in Theorem 4.2.2 is because of the penalty function.

We can also understand the order assessment results in Theorem 4.2.1 from the point view of moments. If α is fixed to be 0.5, the MLE of the forth moment has the asymptotic order $n^{-1/2}$ and so from (4.2), the asymptotic orders for the MLEs of μ_1^4 and μ_2^4 are $n^{-1/2}$. Therefore the MLEs of μ_1 and μ_2 have the asymptotic orders $n^{-1/8}$. If α is fixed to be $\alpha_0 \in (0, 0.5)$, the MLE of the third moment has the asymptotic order $n^{-1/2}$. Similar arguments used before can lead to the result that the asymptotic orders for the MLEs of μ_1 and μ_2 are $n^{-1/6}$.

4.2.3 Simulation Study

In this section, we use simulation to study the accuracy of the limiting distribution of the EM-test statistics for the test of homogeneity in normal mixture models in the presence of the structural parameter. We also compare the power of the EM-test and the MLRT proposed in Chen and Kalbfleisch (2005). For this purpose, we introduce the definition of the MLRT in Chen and Kalbfleisch (2005). Let $(\hat{\alpha}, \hat{\mu}_1, \hat{\mu}_2, \hat{\sigma})$ and $(\hat{\mu}_0, \hat{\sigma}_0)$ be the maximizers of $pl_n(\alpha, \mu_1, \mu_2, \sigma)$ and $pl_n(0.5, \mu, \mu, \sigma)$ with $p_n(\sigma) = 0$ and $p(\alpha) = \log\{4\alpha(1-\alpha)\}$, respectively. The MLRT statistic is defined to be

$$M_n = 2\{l_n(\hat{\alpha}, \hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}) - l_n(0.5, \hat{\mu}_0, \hat{\mu}_0, \hat{\sigma}_0)\}.$$

We use the upper bound χ^2_2 distribution to compute the critical values as in Chen and Kalbfleisch (2005) to calculate the simulated null rejection rates. To ensure the comparability, we use the same penalty function $p(\alpha)$ as in Chen and Kalbflesich (2005).

For the EM-test statistics, we choose the penalty function

$$p_n(\sigma) = -\Big\{s_n/\sigma^2 + \log(\sigma^2/s_n)\Big\}, \tag{4.3}$$

where $s_n = \sum_{i=1}^n (X_i - \bar{X})^2/n$. The penalty function $p_n(\sigma)$ is equivalent to placing an inverse gamma prior on σ^2. It is easy to check that $p_n(\sigma)$ satisfies all Conditions C1-C3. It is also seen that with this $p_n(\sigma)$, $\sigma_j^{(k)}$ has a closed form expression in Step 3 of the EM-iteration. Note that the asymptotic property of the EM-test is valid even when $p_n(\sigma) = 0$. The purpose of using the penalty function $p_n(\sigma)$ is to prevent the underestimation of σ^2 under the null model. The penalty function $p_n(\sigma)$ is maximized at s_n, the MLE of the σ^2 under the homogeneous model. This penalty will push the estimation of the σ^2 towards s_n, which will improve the approximation of the limiting distribution of the EM-test to its finite sample distribution. From this viewpoint, the penalty function $p_n(\sigma)$ plays the role of the higher order adjustment.

For the penalty function $p(\alpha)$, we use

$$p(\alpha) = \log(1 - |1 - 2\alpha|).$$

The combination of $p_n(\sigma)$ and $p(\alpha)$ results in accurate type I errors for the EM-test statistics as we will see.

Similar to single parameter mixture models, we conduct the simulation with two groups of initial values for α: (0.1, 0.2, 0.3, 0.4, 0.5) and (0.1, 0.3, 0.5). We generate 20,000 random samples from $N(0,1)$ with sample size n (n=100, 200, 500). The simulated null rejection rates are summarized in Table 4.1. The EM-test and the MLRT both have very accurate type I errors, especially $EM_n^{(1)}$ with three initial values (0.1, 0.3, 0.5) of α.

Table 4.1: Simulated type I errors (%) of the EM-test and the MLRT under normal mixture models in the presence of the structural parameter.

Level	$EM_n^{(0)}$	$EM_n^{(1)}$	$EM_n^{(2)}$	$EM_n^{(0)}$	$EM_n^{(1)}$	$EM_n^{(2)}$	MLRT
	$n = 100$						
10%	8.9	9.1	9.2	9.2	9.9	10.2	10.9
5%	4.6	4.8	4.8	4.6	5.1	5.3	5.7
1%	0.9	1.0	1.0	0.9	1.0	1.1	1.2
	$n = 200$						
10%	9.3	9.4	9.5	9.7	10.0	10.3	9.8
5%	4.6	4.8	4.8	4.7	5.0	5.1	5.0
1%	1.0	1.1	1.1	0.9	1.1	1.1	1.1
	$n = 500$						
10%	9.6	9.6	9.7	9.9	10.1	10.2	8.7
5%	5.0	5.0	5.1	4.9	5.1	5.2	4.5
1%	1.1	1.1	1.1	1.0	1.1	1.1	0.9

Results in columns (2, 3, 4) used $\alpha = (0.1, 0.2, 0.3, 0.4, 0.5)$.

Results in columns (5, 6, 7) used $\alpha = (0.1, 0.3, 0.5)$.

For the power comparison, we select four alternative models. The parameters are shown in Table 4.2. The powers of the EM-test statistics and the MLRT statistics are calculated based on 10,000 repetitions and they are presented in Table 4.3. We use the simulated critical values to ensure fairness. The results tell us that the EM-test statistics based on three initial values have almost the same power as those from five initial values. Combing the type I error results and the power comparison results, we recommend the use of $EM_n^{(1)}$ with three initial values (0.1, 0.3, 0.5) of α in applications. We also observe that the EM-test and the MLRT have comparable power. The EM-test has higher power when the mixing proportion α is close to 0.5, while the MLRT statistic performs better when the mixing proportion α is close to 0. However, as in the previous chapter the limiting distribution of the EM-test does not require the compactness of the parameter space, while the upper bound result of the MLRT does.

Table 4.2: Parameters in alternative normal mixture models in the presence of the structural parameter.

	$1-\alpha$	θ_1	θ_2	σ	$100KL$
Model I	0.50	1	-1.25	1	2.978
Model II	0.25	1	-1.25	1	3.872
Model III	0.10	1.25	-1.25	1	4.202
Model IV	0.05	1.25	-1.5	1	3.108

KL: Kullback-Leibler information.

4.2.4 Technical Proofs

The null limiting distribution of the EM-test statistic does not depend on the true values of μ and σ^2 under the null model. Therefore, without loss of generality, we may assume that the null distribution is $N(0,1)$ in our asymptotic investigation.

Let us first study the consistency of the estimator of (μ_1, μ_2, σ). We assume that a random sample $X_1, \ldots, X_n$ is drawn from the null distribution $N(0,1)$. All the setups are the same as in Theorems 4.2.1 and 4.2.2.

Table 4.3: Simulated powers (%) of the EM-test and the MLRT under normal mixture models in the presence of the structural parameter at the 5% level .

Model	$EM_n^{(0)}$	$EM_n^{(1)}$	$EM_n^{(2)}$	$EM_n^{(0)}$	$EM_n^{(1)}$	$EM_n^{(2)}$	MLRT
	$n = 100$						
I	45.5	45.2	44.9	46.0	45.4	45.3	37.9
II	48.0	47.8	47.6	47.0	47.1	47.2	46.5
III	44.7	44.8	44.8	43.9	44.1	44.3	51.6
IV	31.7	32.6	32.9	32.4	32.9	33.2	44.4
	$n = 200$						
I	76.3	76.1	76.2	76.6	76.6	76.6	69.2
II	81.1	81.0	81.0	80.6	80.7	80.7	80.6
III	79.0	79.1	79.2	78.5	78.6	78.7	84.7
IV	61.3	61.7	61.9	62.2	62.5	62.7	74.6

Results in columns (2, 3, 4) used $\alpha = (0.1, 0.2, 0.3, 0.4, 0.5)$.

Results in columns (5, 6, 7) used $\alpha = (0.1, 0.3, 0.5)$.

Lemma 4.2.1. *Let* $(\bar{\alpha}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma})$ *be any estimators of* $(\alpha, \mu_1, \mu_2, \sigma)$*. If*

$$pl_n(\bar{\alpha}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}) - pl_n(0.5, 0, 0, 1) \geq c > -\infty$$

and $\bar{\alpha} \in [\delta, 1-\delta]$ *for some* $\delta \in (0, 0.5]$*, then under the null model* $N(0,1)$*,* $\bar{\mu}_1 = o_p(1)$*,* $\bar{\mu}_2 = o_p(1)$ *and* $\bar{\sigma} - 1 = o_p(1)$*.*

Proof. The idea of the proof is to first show that $\bar{\sigma}$ is bounded below with probability approaching 1 and then apply the result in Kiefer and Wolfowitz (1956) to show the consistency of $(\bar{\mu}_1, \bar{\mu}_2, \bar{\sigma})$.

Let $A = \{i : |X_i - \mu_1| < |\sigma \log \sigma| \text{ or } |X_i - \mu_2| < |\sigma \log \sigma|\}$. For any index set, say S, we define

$$l_n(\alpha, \mu_1, \mu_2, \sigma; S) = \sum_{i \in S} \log\{(1-\alpha) f(X_i; \mu_1, \sigma) + \alpha f(X_i; \mu_2, \sigma)\},$$

hence $l_n(\alpha, \mu_1, \mu_2, \sigma) = l_n(\alpha, \mu_1, \mu_2, \sigma; A) + l_n(\alpha, \mu_1, \mu_2, \sigma; A^c)$.

Let $n(A)$ be the number of observations in set A. Note that when $i \in A$, the mixture density is no larger than $1/\sqrt{2\pi\sigma^2}$, therefore

$$l_n(\alpha, \mu_1, \mu_2, \sigma; A) \leq -n(A) \log \sqrt{2\pi\sigma^2} = -\frac{1}{2} n(A) \log(2\pi\sigma^2). \tag{4.4}$$

When $i \in A^c$, the mixture density is no larger than $\exp\{-\log^2 \sigma/2\}/\sqrt{2\pi\sigma^2}$, so

$$l_n(\alpha, \mu_1, \mu_2, \sigma; A^c) \leq -\frac{1}{2} n(A^c)\{\log(2\pi\sigma^2) + \log^2 \sigma\}. \tag{4.5}$$

Combining (4.4) and (4.5), we have

$$l_n(\alpha, \mu_1, \mu_2, \sigma) \leq -\frac{1}{2} n \log(2\pi\sigma^2) - \frac{1}{2}\{n - n(A)\} \log^2 \sigma.$$

After some simple calculations, we get

$$l_n(\alpha, \mu_1, \mu_2, \sigma) - l_n(0.5, 0, 0, 1) \leq \frac{1}{2} \sum_{i=1}^{n} X_i^2 - n \log(\sigma) - \frac{1}{2}\{n - n(A)\} \log^2 \sigma.$$

By the definition of set A, its size decreases when σ goes to 0. Therefore there exits a positive ϵ such that when $\sigma < \epsilon$,

$$n(A) \leq n/2$$

almost surely and uniformly in μ_1, μ_2 and σ. For brevity, we do not include all the details, see Chen et al. (2007) for the proof of a similar result. Note that when ϵ is small enough, for any $\sigma < \epsilon$,

$$\frac{1}{4} \log^2 \sigma + \log \sigma \geq 1,$$

which will be used later.

When $\sigma < \epsilon$, under the null model $N(0,1)$, as $n \to \infty$,

$$\begin{aligned} l_n(\alpha, \mu_1, \mu_2, \sigma) - l_n(0.5, 0, 0, 1) &\leq \frac{1}{2}\sum_{i=1}^{n} X_i^2 - n\log(\sigma) - \frac{n}{4}\log^2 \sigma \\ &\leq \frac{1}{2}\sum_{i=1}^{n} X_i^2 - n = -\frac{n}{2} + o(n), \end{aligned}$$

almost surely. In the last step, we use $\sum_{i=1}^{n} X_i^2 = n + o(n)$, which is a direct result from the strong law of large numbers. Since $\sup_{\sigma>0} \max\{0, p_n(\sigma)\} = o(n)$ and $p_n(1) = o(n)$, and $p(\alpha) \leq p(0.5)$, when $\sigma < \epsilon$, under the null model $N(0,1)$, for large enough n,

$$pl_n(\alpha, \mu_1, \mu_2, \sigma) - pl_n(0.5, 0, 0, 1) \leq -\frac{n}{2} + o(n).$$

Thus for any estimator $(\bar{\alpha}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma})$ such that $pl_n(\bar{\alpha}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}) - pl_n(0.5, 0, 0, 1) \geq c > -\infty$, it is clear that under the null model $N(0,1)$,

$$\lim_{n\to\infty} P(\epsilon \leq \bar{\sigma}) = 1.$$

This result is equivalent to placing a positive constant lower bound for the variance parameter for searching the maximal value of $pl_n(\alpha, \mu_1, \mu_2, \sigma)$. Thus, the consistency of $(\bar{\mu}_1, \bar{\mu}_2, \bar{\sigma})$ is covered by the result in Kiefer and Wolfowitz (1956). Note that their proof can be modified to accommodate a penalty of size $o(n)$. □

Lemma 4.2.2. *Let $(\bar{\alpha}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma})$ be any estimators of $(\alpha, \mu_1, \mu_2, \sigma)$ such that*

$$pl_n(\bar{\alpha}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}) - pl_n(0.5, 0, 0, 1) \geq c > -\infty.$$

If $\bar{\alpha} - \alpha_0 = o_p(1)$ for some $\alpha_0 \in (0, 0.5]$, then $\bar{\mu}_1 = O_p(n^{-1/8})$, $\bar{\mu}_2 = O_p(n^{-1/8})$ and $\bar{\sigma}^2 - 1 = O_p(n^{-1/4})$.

Proof. For $i = 1, \dots, n$, let

$$Y_i = X_i, Z_i = (X_i^2 - 1)/2, U_i = (X_i^3 - 3X_i)/6 \text{ and } V_i = (X_i^4 - 6X_i^2 + 3)/24. \tag{4.6}$$

Further, let

$$\bar{s}_1 = \bar{m}_1,\ \bar{s}_2 = \bar{m}_2 + \bar{\sigma}^2 - 1,\ \bar{s}_3 = \bar{m}_3 \text{ and } \bar{s}_4 = \bar{m}_4 - 3\bar{m}_2^2 \tag{4.7}$$

with $\bar{m}_j = (1 - \bar{\alpha})\bar{\mu}_1^j + (1 - \bar{\alpha})\bar{\mu}_2^j$, $j = 1, 2, 3, 4$.

Following the result in Lemma 4.2.1, $(\bar{\mu}_1, \bar{\mu}_2, \bar{\sigma})$ are in a small neighborhood of $(0, 0, 1)$ in probability. Therefore the result in (28), page 363 in Chen and Chen (2003) is easily modified for $2\{l_n(\bar{\alpha}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}) - l_n(0.5, 0, 0, 1)\}$. Hence

$$\begin{aligned}
&2\{pl_n(\bar{\alpha}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}) - pl_n(0.5, 0, 0, 1)\} \\
\le\ & 2\{l_n(\bar{\alpha}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}) - l_n(0.5, 0, 0, 1)\} + 2\{p_n(\bar{\sigma}) - p_n(1)\} + 2\{p(\bar{\alpha}) - p(0.5)\} \\
\le\ & 2\bar{s}_1\sum_{i=1}^n Y_i - \bar{s}_1^2\sum_{i=1}^n Y_i^2\{1 + o_p(1)\} \\
& +2\bar{s}_2\sum_{i=1}^n Z_i - \bar{s}_2^2\sum_{i=1}^n Z_i^2\{1 + o_p(1)\} \\
& +2\bar{s}_3\sum_{i=1}^n U_i - \bar{s}_3^2\sum_{i=1}^n U_i^2\{1 + o_p(1)\} \\
& +2\bar{s}_4\sum_{i=1}^n V_i - \bar{s}_4^2\sum_{i=1}^n V_i^2\{1 + o_p(1)\} \\
& +o_p(n^{1/4})(\bar{\sigma}^2 - 1) + 2\{p(\alpha_0) - p(0.5)\} + o_p(1).
\end{aligned} \tag{4.8}$$

The term $o_p(n^{1/4})(\bar{\sigma}^2 - 1)$ in the last step, is obtained by using the mean value theorem on $p_n(\bar{\sigma}) - p_n(1)$ and by Condition C3. The term $2\{p(\alpha_0) - p(0.5)\} + o_p(1)$ in the last step comes from that $\bar{\alpha} - \alpha_0 = o_p(1)$ and $p(\alpha)$ is a continuous function of α.

Note that using the fact $|x| \le 1 + x^4$,

$$|o_p(n^{1/4})(\bar{\sigma}^2 - 1)| \le o_p(1)\{1 + n(\bar{\sigma}^2 - 1)^4\}.$$

By Lemma 4.2.3 to be shown,

$$(\bar{\sigma}^2 - 1)^4 = O_p\Big\{\sum_{j=1}^{4} \bar{s}_j^2\Big\}.$$

Therefore $o_p(n^{1/4})(\bar{\sigma}^2 - 1)$ is a higher order term of the quadratic terms in (4.8) and can be omitted from expansion in (4.8). Note that

$$2\bar{s}_1\sum_{i=1}^{n}Y_i - \bar{s}_1^2\sum_{i=1}^{n}Y_i^2\{1 + o_p(1)\} \le \frac{(\sum_{i=1}^{n} Y_i)^2}{\sum_{i=1}^{n} Y_i^2}\{1 + o_p(1)\} = O_p(1).$$

The same conclusion is applicable to other three terms. Hence

$$\begin{aligned} c &\le 2\{pl_n(\bar{\alpha}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}) - pl_n(0.5, 0, 0, 1)\} \\ &\le 2\bar{s}_1\sum_{i=1}^{n}Y_i - \bar{s}_1^2\sum_{i=1}^{n}Y_i^2\{1 + o_p(1)\} + O_p(1) \\ &\le O_p(1). \end{aligned}$$

Therefore

$$2\bar{s}_1\sum_{i=1}^{n}Y_i - \bar{s}_1^2\{\sum_{i=1}^{n}Y_i^2\}\{1 + o_p(1)\} = O_p(1).$$

Because $\sum_{i=1}^{n} Y_i = O_p(n^{1/2})$ and $\sum_{i=1}^{n} Y_i^2 = n + o_p(n)$, we get

$$\bar{s}_1 = O_p(n^{-1/2}). \tag{4.9}$$

Similarly, we have

$$\bar{s}_j = O_p(n^{-1/2}),\ j = 2, 3, 4. \tag{4.10}$$

Due to the condition that $\bar{\alpha} = \alpha_0 + o_p(1)$ and $0 < \alpha_0 \le 0.5$, we further conclude that

$$\bar{\mu}_1 = O_p(n^{-1/8}),\ \bar{\mu}_2 = O_p(n^{-1/8}) \text{ and } \bar{\sigma}^2 - 1 = O_p(n^{-1/4})$$

by using the results in Lemma 4.2.3. □

We skipped two technical details in the above proof and they will be covered in the following Lemma.

Lemma 4.2.3. *Let $(\bar{\alpha}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma})$ be any estimators of $(\alpha, \mu_1, \mu_2, \sigma)$ such at*

$$\bar{\mu}_1 = o_p(1),\ \bar{\mu}_2 = o_p(1),\ \bar{\sigma}^2 - 1 = o_p(1)$$

and $\bar{\alpha} \in [\delta, 1-\delta]$ for some $\delta \in (0, 0.5]$. Then

$$\bar{\mu}_1^4 = O_p\Big(\sum_{j=1}^4 |\bar{s}_j|\Big),\ \bar{\mu}_2^4 = O_p\Big(\sum_{j=1}^4 |\bar{s}_j|\Big),\ \textit{and}\ (\bar{\sigma}^2-1)^2 = O_p\Big(\sum_{j=1}^4 |\bar{s}_j|\Big).$$

Proof. Since $\bar{\mu}_1 = o_p(1)$, $\bar{\mu}_2 = o_p(1)$, $\bar{\sigma}^2 - 1 = o_p(1)$, then $\bar{s}_j = o_p(1)$, where $\bar{s}_j$ is defined in (4.7), $j = 1, 2, 3, 4$.

According to the definition of $\bar{s}_1$ in (4.7), we have

$$\bar{\mu}_1 = \frac{1}{1-\bar{\alpha}}\bar{s}_1 - \frac{\bar{\alpha}}{1-\bar{\alpha}}\bar{\mu}_2. \tag{4.11}$$

Plugging (4.11) in the definitions of $\bar{s}_3$ and $\bar{s}_4$ in (4.7) and using the condition $\bar{\alpha} \in [\delta, 1-\delta]$ for some $\delta \in (0, 0.5]$, we obtain

$$\begin{aligned}
\bar{s}_3 &= \frac{\bar{\alpha}(1-2\bar{\alpha})}{(1-\bar{\alpha})^2}\bar{\mu}_2^3 + o_p(\bar{s}_1) && (4.12)\\
\bar{s}_4 &= \frac{\bar{\alpha}(1-6\bar{\alpha}+6\bar{\alpha}^2)}{(1-\bar{\alpha})^3}\bar{\mu}_2^4 + o_p(\bar{s}_1)\\
&= \frac{\bar{\alpha}(1-6\bar{\alpha}+6\bar{\alpha}^2)}{(1-\bar{\alpha})^3}\bar{\mu}_2^4 - 3(1-2\bar{\alpha})\bar{\mu}_2\bar{s}_3/\{2(1-\bar{\alpha})\}\\
&\quad +3(1-2\bar{\alpha})\bar{\mu}_2\bar{s}_3/\{2(1-\bar{\alpha})\} + o_p(\bar{s}_3)\\
&= -\frac{\bar{\alpha}}{2(1-\bar{\alpha})^3}\bar{\mu}_2^4 + o_p(\bar{s}_1) + o_p(\bar{s}_3). && (4.13)
\end{aligned}$$

Hence

$$\bar{\mu}_2^4 = O_p\Big(\sum_{j=1}^4 |\bar{s}_j|\Big)$$

and consequently by (4.11)

$$\bar{\mu}_1^4 = O_p\Big(\sum_{j=1}^{4} |\bar{s}_j|\Big).$$

From the definition of $\bar{s}_2$ in (4.7), we further conclude

$$(\bar{\sigma}^2 - 1)^2 = O_p\Big(\sum_{j=1}^{4} |\bar{s}_j|\Big).$$

□

Now we show that after finite number of iterations, the fitted value of α remain in an infinite small neighborhood of the initial value under the null model. Let $(\bar{\alpha}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma})$ be some estimators of $(\alpha, \mu_1, \mu_2, \sigma)$ as before. Define

$$\begin{aligned} H_n(\alpha) &= (n - \sum_{i=1}^{n} \bar{w}_i) \log(1-\alpha) + \sum_{i=1}^{n} \bar{w}_i \log(\alpha) + p(\alpha) \\ &= R_n(\alpha) + p(\alpha) \end{aligned}$$

with

$$\bar{w}_i = \frac{\bar{\alpha} f(X_i; \bar{\mu}_2, \bar{\sigma})}{(1-\bar{\alpha}) f(X_i; \bar{\mu}_1, \bar{\sigma}) + \bar{\alpha} f(X_i; \bar{\mu}_2, \bar{\sigma})}.$$

Let $\bar{\alpha}^* = \arg\max_\alpha H_n(\alpha)$. The following lemma considers some asymptotic properties regarding $\bar{\alpha}^*$.

Lemma 4.2.4. *Let $(\bar{\alpha}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma})$ be any estimators of $(\alpha, \mu_1, \mu_2, \sigma)$. Suppose*

$$pl_n(\bar{\alpha}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}) - pl_n(0.5, 0, 0, 1) \geq c > -\infty.$$

If $\bar{\alpha} - \alpha_0 = O_p(n^{-1/4})$ for some $\alpha_0 \in (0, 1)$, then under the null distribution $N(0, 1)$, we have $\bar{\alpha}^* - \alpha_0 = O_p(n^{-1/4})$.

Proof. Putting $\hat{\alpha} = n^{-1}\sum_{i=1}^{n}\bar{w}_i$, we have

$$|\hat{\alpha} - \bar{\alpha}| \;=\; \frac{(1-\bar{\alpha})\bar{\alpha}}{n}\Big|\sum_{i=1}^{n}\frac{f(X_i;\bar{\mu}_2,\bar{\sigma}) - f(X_i;\bar{\mu}_1,\bar{\sigma})}{(1-\bar{\alpha})f(X_i;\bar{\mu}_1,\bar{\sigma}) + \bar{\alpha}f(X_i;\bar{\mu}_2,\bar{\sigma})}\Big|. \tag{4.14}$$

Following the results in Lemma 4.2.1, $(\bar{\mu}_1, \bar{\mu}_2, \bar{\sigma})$ are in a small neighborhood of $(0, 0, 1)$ in probability. Therefore, by the Taylor's expansion at $(0, 0, 1)$ of the function on the right hand side of (4.14) to order 1, we get

$$\begin{aligned} |\hat{\alpha} - \bar{\alpha}| &= \frac{(1-\bar{\alpha})\bar{\alpha}}{n}\Big|(\bar{\mu}_2 - \bar{\mu}_1)\sum_{i=1}^{n}Y_i + O_p(n)\{\bar{\mu}_1^2 + \bar{\mu}_2^2 + (\bar{\sigma}^2-1)^2\}\Big| \\ &= O_p(n^{-1/4}). \end{aligned}$$

Here in the last step, we use the order assessment results in Lemma 4.2.2. Due to the assumption that $\bar{\alpha} - \alpha_0 = O_p(n^{-1/4})$, we have $\hat{\alpha} - \alpha_0 = O_p(n^{-1/4})$ and therefore it suffices to prove that $\bar{\alpha}^* - \hat{\alpha} = O_p(n^{-1/4})$.

First, using the similar arguments in Lemma 3.5.3, we have

$$\bar{\alpha}^* - \hat{\alpha} = o_p(1).$$

Next, note that

$$H_n(\hat{\alpha}) = R_n(\hat{\alpha}) + p(\hat{\alpha}) \;\;\le\;\; H_n(\bar{\alpha}^*) = R_n(\bar{\alpha}^*) + p(\bar{\alpha}^*)$$

and $R_n'(\hat{\alpha}) = 0$. By applying the first order Taylor expansion at $\hat{\alpha}$ for $R_n(\bar{\alpha}^*)$, the above inequality becomes

$$R_n(\hat{\alpha}) + p(\hat{\alpha}) \;\;\le\;\; R_n(\hat{\alpha}) + R_n''(\eta)(\bar{\alpha}^* - \hat{\alpha})^2 + p(\bar{\alpha}^*),$$

for some η between $\hat{\alpha}$ and $\bar{\alpha}^*$. Hence

$$-R_n''(\eta)(\bar{\alpha}^* - \hat{\alpha})^2 \le p(\bar{\alpha}^*) - p(\hat{\alpha}) = o_p(1).$$

Note that $\eta = \alpha_0 + o_p(1)$ and hence

$$-R_n''(\eta) = \frac{n}{(1-\eta)^2}(1-\hat{\alpha}) + \frac{n}{\eta^2}\hat{\alpha} = \frac{n}{\alpha_0(1-\alpha_0)}\{1 + o_p(1)\}.$$

So it is easily seen that $\bar{\alpha}^* - \hat{\alpha} = o_p(n^{-1/2}) = O_p(n^{-1/4})$ as claimed. □

Lemma 4.2.5. *Let* $(\bar{\alpha}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma})$ *be any estimators of* $(\alpha, \mu_1, \mu_2, \sigma)$*. Suppose*

$$pl_n(\bar{\alpha}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}) - pl_n(0.5, 0, 0, 1) \geq c > -\infty.$$

(a) If $\bar{\alpha} - 0.5 = O_p(n^{-1/4})$*, then*

$$\begin{aligned} &2\{pl_n(\bar{\alpha}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}) - pl_n(0.5, 0, 0, 1)\} \\ &\leq \frac{(\sum_{i=1}^n Y_i)^2}{\sum_{i=1}^n Y_i^2} + \frac{(\sum_{i=1}^n Z_i)^2}{\sum_{i=1}^n Z_i^2} + \frac{\{(\sum_{i=1}^n V_i)^-\}^2}{\sum_{i=1}^n V_i^2} + o_p(1). \end{aligned}$$

(b) If $\bar{\alpha} - \alpha_0 = o_p(1)$ *for* $\alpha_0 \in (0, 0.5)$*, then*

$$\bar{\mu}_1 = O_p(n^{-1/6}),\ \bar{\mu}_2 = O_p(n^{-1/6}), \bar{\sigma}^2 - 1 = O_p(n^{-1/3})$$

and

$$\begin{aligned} &2\{pl_n(\bar{\alpha}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}) - pl_n(0.5, 0, 0, 1)\} \\ &\leq \frac{(\sum_{i=1}^n Y_i)^2}{\sum_{i=1}^n Y_i^2} + \frac{(\sum_{i=1}^n Z_i)^2}{\sum_{i=1}^n Z_i^2} + \frac{(\sum_{i=1}^n U_i)^2}{\sum_{i=1}^n U_i^2} + 2\{p(\alpha_0) - p(0.5)\} + o_p(1). \end{aligned}$$

Proof. (a) When $\bar{\alpha} - 0.5 = O_p(n^{-1/4})$, from (4.9), (4.12) and Lemma 4.2.2, we get

$$\bar{s}_3 = O_p(n^{-1/4}) \cdot O_p(n^{-3/8}) + o_p(\bar{s}_1) = o_p(n^{-1/2}).$$

So the third quadratic function in (4.8) becomes $o_p(1)$. Further from (4.9), (4.10) and (4.13), we obtain $\bar{s}_4 = -2\bar{\mu}_2^4 + o_p(n^{-1/2})$, which is always non-positive in probability.

Using the property of a quadratic function and Lemma 4.2.2, the upper bound (4.8) can be strengthened to

$$\begin{aligned}
&2\{pl_n(\bar{\alpha},\bar{\mu}_1,\bar{\mu}_2,\bar{\sigma}) - pl_n(0.5,0,0,1)\}\\
\leq\ & 2\{\bar{s}_1\sum_{i=1}^{n}Y_i + \bar{s}_2\sum_{i=1}^{n}Z_i + \bar{s}_4\sum_{i=1}^{n}V_i\}\\
&-\{\bar{s}_1^2\sum_{i=1}^{n}Y_i^2 + \bar{s}_2^2\sum_{i=1}^{n}Z_i^2 + \bar{s}_4^2\sum_{i=1}^{n}V_i^2\}\{1+o_p(1)\} + o_p(1)\\
\leq\ & \frac{(\sum_{i=1}^n Y_i)^2}{\sum_{i=1}^n Y_i^2} + \frac{(\sum_{i=1}^n Z_i)^2}{\sum_{i=1}^n Z_i^2} + \frac{\{(\sum_{i=1}^n V_i)^-\}^2}{\sum_{i=1}^n V_i^2} + o_p(1).
\end{aligned}$$

(b) If $\bar{\alpha} - \alpha_0 = o_p(1)$ for $\alpha_0 \in (0, 0.5)$, then from (4.9), (4.10) and (4.12), we get $\bar{\mu}_2 = O_p(n^{-1/6})$. Together with (4.11), we further have $\bar{\mu}_1 = O_p(n^{-1/6})$. From the definition of $\bar{s}_2$ in (4.7), we conclude $\bar{\sigma}^2 - 1 = O_p(n^{-1/3})$.

Next, from (4.13) and the above results, it is seen that $\bar{s}_4 = o_p(n^{-1/2})$. So the forth quadratic function in (4.8) becomes $o_p(1)$. Using the property of the quadratic function and Lemma 4.2.2, the upper bound in (4.8) can be strengthened to

$$\begin{aligned}
&2\{pl_n(\bar{\alpha},\bar{\mu}_1,\bar{\mu}_2,\bar{\sigma}) - pl_n(0.5,0,0,1)\}\\
\leq\ & 2\{\bar{s}_1\sum_{i=1}^{n}Y_i + \bar{s}_2\sum_{i=1}^{n}Z_i + \bar{s}_3\sum_{i=1}^{n}U_i\}\\
&-\{\bar{s}_1^2\sum_{i=1}^{n}Y_i^2 + \bar{s}_2^2\sum_{i=1}^{n}Z_i^2 + \bar{s}_3^2\sum_{i=1}^{n}U_i^2\}\{1+o_p(1)\} + 2\{p(\alpha_0) - p(0.5)\} + o_p(1)\\
\leq\ & \frac{(\sum_{i=1}^n Y_i)^2}{\sum_{i=1}^n Y_i^2} + \frac{(\sum_{i=1}^n Z_i)^2}{\sum_{i=1}^n Z_i^2} + \frac{(\sum_{i=1}^n U_i)^2}{\sum_{i=1}^n U_i^2} + 2\{p(\alpha_0) - p(0.5)\} + o_p(1).
\end{aligned}$$

□

Proof of Theorem 4.2.1

By the property of the EM algorithm (Dempster et al., 1977), the definition of $\alpha_j^{(k)}$ and others, for any finite k, we have

$$pl_n(\alpha_j^{(k)}, \mu_{j1}^{(k)}, \mu_{j2}^{(k)}, \sigma_j^{(k)}) \geq pl_n(\alpha_j^{(0)}, \mu_{j1}^{(0)}, \mu_{j2}^{(0)}, \sigma_j^{(0)}) \geq pl_n(\alpha_j, 0, 0, 1).$$

Therefore

$$pl_n(\alpha_j^{(k)}, \mu_{j1}^{(k)}, \mu_{j2}^{(k)}, \sigma_j^{(k)}) - pl_n(0.5, 0, 0, 1) \geq p(\alpha_j) - p(0.5) > -\infty.$$

Following Lemma 4.2.4, $\alpha_j^{(1)} - \alpha_j = O_p(n^{-1/4})$. By mathematical induction, we can further have

$$\alpha_j^{(k)} - \alpha_j = O_p(n^{-1/4}).$$

Hence the conclusions in Lemmas 4.2.2 and 4.2.5 can apply to prove the results in Theorem 4.2.1 is true. □

Proof of Theorem 4.2.2

According the classic results for the regular models, we have that

$$2\{\sup_{\mu,\sigma} pl_n(0.5, \mu, \mu, \sigma) - pl_n(0.5, 0, 0, 1)\} = \frac{(\sum_{i=1}^n Y_i)^2}{\sum_{i=1}^n Y_i^2} + \frac{(\sum_{i=1}^n Z_i)^2}{\sum_{i=1}^n Z_i^2} + o_p(1).$$

Due the properties proved in Theorem 4.2.1, the conclusions in Lemma 4.2.5 are applicable. So immediately, we have

$$M_n^{(k)}(0.5) \leq \frac{\{(\sum_{i=1}^n V_i)^-\}^2}{\sum_{i=1}^n V_i^2} + o_p(1)$$

and for $\alpha_j \neq 0.5$,

$$M_n^{(k)}(\alpha_j) \leq \frac{(\sum_{i=1}^n U_i)^2}{\sum_{i=1}^n U_i^2} + \Delta + o_p(1),$$

where Δ is defined in Theorem 4.2.2. Hence

$$EM_n^{(k)} \leq \max\left[\frac{(\sum_{i=1}^n U_i)^2}{\sum_{i=1}^n U_i^2} + \Delta, \frac{\{(\sum_{i=1}^n V_i)^-\}^2}{\sum_{i=1}^n V_i^2}\right] + o_p(1).$$

It is seen that the upper bound is also achievable. Therefore,

$$EM_n^{(k)} = \max\left[\frac{(\sum_{i=1}^n U_i)^2}{\sum_{i=1}^n U_i^2} + \Delta, \frac{\{(\sum_{i=1}^n V_i)^-\}^2}{\sum_{i=1}^n V_i^2}\right] + o_p(1).$$

It is easy to verify that U_i and V_i are uncorrelated. Further, $\sum_{i=1}^n U_i/\sqrt{n}$ and $\sum_{i=1}^n V_i/\sqrt{n}$ are jointly asymptotical bivariate normal and therefore asymptotically independent. Consequently, the limiting distribution is given by $F(x-\Delta)\{0.5 + 0.5F(x)\}$ with $F(x)$ being the cdf of χ_1^2 distribution. □

4.3 Normal Mixture Models in Both Mean and Variance Parameters

4.3.1 The EM-test Procedure

In the last section, we applied the EM-test to normal mixture models in the presence of the structural parameter and studied its asymptotic properties. In this section, we apply the EM-test to the test of homogeneity in the normal mixture model when both mean and variance parameters are unconstrained. Suppose $X_1, \ldots, X_n$ is a random sample from

$$(1-\alpha)N(\mu_1, \sigma_1^2) + \alpha N(\mu_2, \sigma_2^2).$$

Our interest is to test

$$H_0 : \alpha = 0 \text{ or } \alpha = 1 \text{ or } (\mu_1, \sigma_1^2) = (\mu_2, \sigma_2^2). \tag{4.15}$$

The log-likelihood function for the above testing problem is given by

$$l_n(\alpha, \mu_1, \mu_2, \sigma_1, \sigma_2) = \sum_{i=1}^n \log\{(1-\alpha)f(X_i; \mu_1, \sigma_1) + \alpha f(X_i; \mu_2, \sigma_2)\}.$$

For the testing problem (4.15), apart from the non-strong identifiability of normal mixture model, the asymptotic properties of the LRT or the MLRT are further complicated due to the following two special technical difficulties.

(1) The log-likelihood function is unbounded because for any given n, $l_n(\alpha, \mu_1, \mu_2, \sigma_1, \sigma_2) \to \infty$ when $\mu_1 = X_1$ and $\sigma_1 \to 0$ with other parameters fixed and hence the MLEs are not well defined;

(2) Normal kernel does not satisfy the finite Fisher information condition, i.e., under the null model $N(0,1)$,
$$E\Big\{\frac{f(X_i;\mu,\sigma)}{f(X_i;0,1)} - 1\Big\}^2$$
can be infinity for some μ and $\sigma > 0$.

Because of these two difficulties, especially the second one, the limiting distributions of the LRT and the MLRT are still under investigation. As we discussed in Chapter 3, the EM-test is a likelihood based method, whose asymptotic results are free from the finite Fisher information condition. It could be a useful method for the current testing problem. As with the EM-test in the last section, we first introduce a modified log-likelihood function as follows:

$$pl_n(\alpha, \mu_1, \mu_2, \sigma_1, \sigma_2) = l_n(\alpha, \mu_1, \mu_2, \sigma_1, \sigma_2) + p_n(\sigma_1) + p_n(\sigma_2) + p(\alpha), \tag{4.16}$$

where $p_n(\sigma)$ is used to prevent the fitted value of σ^2 to be close 0 and $p(\alpha)$ is added for the same purpose as in the last section. Note that it is not necessary to choose an additive penalty function for σ_1^2 and σ_2^2. This choice is to enable efficient numerical computation in the calculation of the EM-test statistic to be defined. The EM-test statistic is defined in exactly the same way as in last section except that the modified log-likelihood function in the definition is replaced by the function in (4.16) and the maximization must be performed with respect to five parameters $(\alpha, \mu_1, \mu_2, \sigma_1, \sigma_2)$ under the alternative model. The analytic form of the EM-test statistic is not convenient to present. Again, we provide the following pseudo code instead.

Step 0. Choose a number of initial α values, say $\alpha_1, \alpha_2, \ldots, \alpha_J$. Compute

$$(\hat{\mu}_0, \hat{\sigma}_0) = \arg\max_{\mu,\, \sigma} pl_n(1/2, \mu, \mu, \sigma, \sigma).$$

Let $j = 1, k = 0$.

Step 1. Let $\alpha_j^{(k)} = \alpha_j$.

Step 2. Compute

$$(\mu_{j1}^{(k)}, \mu_{j2}^{(k)}, \sigma_{j1}^{(k)}, \sigma_{j2}^{(k)}) = \arg \max_{\mu_1,\, \mu_2,\, \sigma_1,\, \sigma_2} pl_n(\alpha_j^{(k)}, \mu_1,\ \mu_2,\ \sigma_1,\ \sigma_2)$$

and

$$M_n^{(k)}(\alpha_j) = 2\{pl_n(\alpha_j^{(k)}, \mu_{j1}^{(k)}, \mu_{j2}^{(k)}, \sigma_{j1}^{(k)}, \sigma_{j2}^{(k)}) - pl_n(1/2, \hat{\mu}_0, \hat{\mu}_0, \hat{\sigma}_0, \hat{\sigma}_0)\}.$$

Step 3. For $i = 1, 2, \ldots, n$, compute the weights, which are the conditional expectations in the E-step,

$$w_{ij}^{(k)} = \frac{\alpha_j^{(k)} f(X_i; \mu_{j2}^{(k)}, \sigma_{j2}^{(k)})}{(1 - \alpha_j^{(k)}) f(X_i; \mu_{j1}^{(k)}, \sigma_{j1}^{(k)}) + \alpha_j^{(k)} f(X_i; \mu_{j2}^{(k)}, \sigma_{j2}^{(k)})}.$$

Now following the M-step, let

$$\begin{aligned}
\alpha_j^{(k+1)} &= \arg\max_{\alpha}\{(n - \sum_{i=1}^n w_{ij}^{(k)}) \log(1-\alpha) + \sum_{i=1}^n w_{ij}^{(k)} \log(\alpha) + p(\alpha)\},\\
\mu_{j1}^{(k+1)} &= \sum_{i=1}^n (1 - w_{ij}^{(k)}) X_i \Big/ \sum_{i=1}^n (1 - w_{ij}^{(k)}),\\
\mu_{j2}^{(k+1)} &= \sum_{i=1}^n w_{ij}^{(k)} X_i \Big/ \sum_{i=1}^n w_{ij}^{(k)},\\
\sigma_{j1}^{(k+1)} &= \arg\max_{\sigma_1} \Big\{ - \frac{1}{2\sigma_1^2} \sum_{i=1}^n (1 - w_{ij}^{(k)})(X_i - \mu_{j1}^{(k+1)})^2 - \frac{1}{2} \sum_{i=1}^n (1 - w_{ij}^{(k)}) \log \sigma_1^2 + p_n(\sigma_1) \Big\},\\
\sigma_{j2}^{(k+1)} &= \arg\max_{\sigma_2} \Big\{ - \frac{1}{2\sigma_2^2} \sum_{i=1}^n w_{ij}^{(k)} (X_i - \mu_{j2}^{(k+1)})^2 - \frac{1}{2} \sum_{i=1}^n w_{ij}^{(k)} \log \sigma_2^2 + p_n(\sigma_2) \Big\}.
\end{aligned}$$

Compute

$$M_n^{(k+1)}(\alpha_j) = 2\{pl_n(\alpha_j^{(k+1)}, \mu_{j1}^{(k+1)}, \mu_{j2}^{(k+1)}, \sigma_{j1}^{(k+1)}, \sigma_{j2}^{(k+1)}) - pl_n(1/2, \hat{\mu}_0, \hat{\mu}_0, \hat{\sigma}_0, \hat{\sigma}_0)\}.$$

Let $k = k+1$ and repeat Step 3 for a fixed number of iterations in k.

Step 4. Let $j = j+1$, $k = 0$ and go to Step 1, until $j = J$.

Step 5. For each k, calculate the test statistic as

$$EM_n^{(k)} = \max\{M_n^{(k)}(\alpha_j), j = 1, 2, \ldots, J\}.$$

The EM-test rejects the homogeneous model when $EM_n^{(k)}$, for a prechosen k, is larger than some critical value. The critical value will be determined by the limiting distribution of the EM-test statistic, which will be studied in the next subsection.

4.3.2 Asymptotic Behavior of the EM-test

We study the asymptotic properties of the EM-test under the following conditions on the penalty function $p_n(\sigma)$ in addition to the Conditions C1 and C2:

C4. $p_n'(\sigma) = o_p(n^{1/6})$ for $\sigma \in N(1)$ with $p_n'(\sigma)$ is the first derivative of $p_n(\sigma)$ and $N(1)$ being a small neighborhood of 1 .

C5. $p_n(\sigma) \leq 4(\log n)^2 \log(\sigma)$, when $\sigma \leq 1/n$ as n is large enough.

The first theorem considers the consistency of $(\alpha_j^{(k)}, \mu_{j1}^{(k)}, \mu_{j2}^{(k)}, \sigma_{j1}^{(k)}, \sigma_{j2}^{(k)})$. The proof will be given in Section 4.3.4.

Theorem 4.3.1. *Suppose Conditions C1-C2 and C4-C5 hold, and $p(\alpha)$ is a continuous function such that $p(\alpha) \to -\infty$ as $\alpha \to 0$ and it attains its maximal value at $\alpha = 0.5$. Under the null distribution $N(\mu_0, \sigma_0^2)$, we have, for $j = 1, \ldots, J$ and any fixed finite k,*

$$\alpha_j^{(k)} - \alpha_j = o_p(1), \text{ and } \mu_{jh}^{(k)} - \mu_0 = o_p(1) \text{ and } \sigma_{jh}^{(k)} - \sigma_0 = o_p(1), \; h = 1, 2.$$

Based on the above consistency result, we can get the null distribution of $EM_n^{(k)}$ for any given $\alpha_j, j = 1, 2, \ldots, J$, and finite k. The proof will be deferred to Section 4.3.4.

Theorem 4.3.2. *Assume the same conditions as in Theorem 4.3.1, and that one of α_j's is equal to 0.5. Under null distribution $N(\mu_0, \sigma_0^2)$, for any fixed finite k, as $n \to \infty$,*

$$EM_n^{(k)} \xrightarrow{d} \chi_2^2.$$

It is surprise to see that the additional scale parameter in the model makes the limiting distribution much simpler. This is in sharp comparison to the case for the likelihood ratio test. We interpret this phenomenon from the point view of moments. Without loss of generality, we assume the mean and variance of the normal mixture model $(1-\alpha)N(\mu_1, \sigma_1^2) + \alpha N(\mu_2, \sigma_2^2)$ are 0 and 1, respectively. The test of homogeneity is to choose one model between $N(0, 1)$ and $(1-\alpha)N(\mu_1, \sigma_1^2) + \alpha N(\mu_2, \sigma_2^2)$ with

$$(1-\alpha)\mu_1 + \alpha\mu_2 = 0 \text{ and } (1-\alpha)(\mu_1^2 + \sigma_1^2) + \alpha(\mu_2^2 + \sigma_2^2) = 1.$$

Let $\beta_1 = \mu_1^2 + \sigma_1^2 - 1$. When α is fixed to be $\alpha_0 \in (0, 0.5]$ (for simplicity, let $\alpha_0 = 0.5$), the third moment and the forth moment of the mixture model are found to be

$$\begin{aligned} E(X_1^3) &= 3\mu_1\beta_1, \\ E(X_1^4) &= 3\beta_1^2 - 2\mu_1^3 + 3. \end{aligned}$$

It is easy to verify that $\{E(X_1^3), E(X_1^4)\} = \{0, 3\}$ if and only if the mixture model is the homogeneous model. So the test of homogeneity is equivalent to testing

$$H_0 : \{E(X_1^3), E(X_1^4)\} = \{0, 3\} \text{ versus } H_a : \{E(X_1^3), E(X_1^4)\} \neq \{0, 3\}.$$

As seen in Figure 4.1, $\{0, 3\}$ is the interior point of the parameter space of $\{E(X_1^3), E(X_1^4)\}$. Therefore the limiting distribution of the EM-test in this case is χ_2^2. Clearly, this limiting distribution is convenient for the calculation of the critical values for the EM-test statistics.

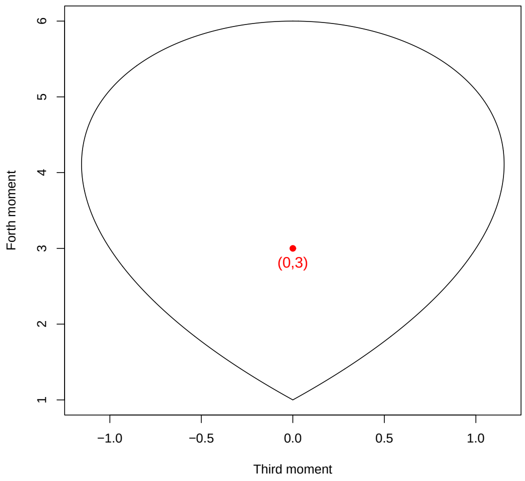

Figure 4.1: The parameter space (area inside the solid line) for $\{E(X_1^3), E(X_1^4)\}$.

4.3.3 Simulation Studies

Now we use the simulation studies to examine the finite sample performance of the limiting distribution for the EM-test and the choice of the penalty function $p_n(\sigma)$. We also compare the power of the EM-test and the MLRT. Here the MLRT statistic is defined as

$$M_n = 2\left\{\sup_{\alpha,\mu_1,\mu_2,\sigma_1,\sigma_2} pl_n(\alpha,\mu_1,\mu_2,\sigma_1,\sigma_2) - pl_n(0.5,\hat{\mu}_0,\hat{\mu}_0,\hat{\sigma}_0,\hat{\sigma}_0)\right\}.$$

Although the limiting distribution for the MLRT is not available. It is included in the simulation to serve as an efficiency barometer.

The penalty function $p_n(\sigma)$ used in Section 4.2.3 is found to satisfy all Conditions C1-C2 and C4-C5. As we mentioned before, this penalty function is equivalent to placing an inverse gamma prior on σ^2. When we have one isolated point in the data, this penalty function or the inverse gamma prior provides extra information for estimating the variance parameters, which will prevent the degenerate situation. Therefore, we choose the same type of penalty function as in Section 4.2.3 for $p_n(\sigma)$. Based on our simulation results, we suggest the use of the following two penalty functions, $p_n(\sigma)$ and $p(\alpha)$,

$$p_n(\sigma) = -0.25\Big\{s_n/\sigma^2 + \log(\sigma^2/s_n)\Big\} \text{ and } p(\alpha) = \log(1 - |1 - 2\alpha|).$$

The combination of $p_n(\sigma)$ and $p(\alpha)$ results in the EM-test statistics with accurate type I errors.

Table 4.4: Simulated type I errors (%) of the EM-test under normal mixture model in both mean and variance parameters.

Level	$EM_n^{(0)}$	$EM_n^{(1)}$	$EM_n^{(2)}$	$EM_n^{(0)}$	$EM_n^{(1)}$	$EM_n^{(2)}$
	$n = 100$					
10%	10.8	10.9	10.9	10.5	10.6	10.6
5%	5.5	5.5	5.6	5.3	5.4	5.4
1%	1.2	1.2	1.2	1.1	1.2	1.2
	$n = 200$					
10%	10.7	10.7	10.7	10.4	10.5	10.5
5%	5.4	5.4	5.4	5.1	5.2	5.2
1%	1.1	1.1	1.1	1.0	1.0	1.0
	$n = 500$					
10%	10.3	10.4	10.4	10.1	10.2	10.2
5%	5.3	5.3	5.3	5.2	5.2	5.2
1%	1.0	1.0	1.0	1.0	1.0	1.0

Results in columns (2, 3, 4) used $\alpha = (0.1, 0.2, 0.3, 0.4, 0.5)$

Results in columns (5, 6, 7) used $\alpha = (0.1, 0.3, 0.5)$

In simulations, the type I errors are calculated based on 20,000 samples from N(0,1).

In a similar way to to Section 4.2.3, we use two groups of initial values (0.1, 0.2, 0.3, 0.4, 0.5) and (0.1, 0.3, 0.5) to calculate $EM_n^{(k)}$. The simulation results are summarized in Table 4.4. The EM-test statistics based on (0.1, 0.3, 0.5) give accurate type I errors.

For power comparison, we select five alternative models. The parameters are shown in Table 4.5. The powers of the EM-test and the MLRT are calculated based on 10,000 repetitions and they are presented in Table 4.6. Since the limiting distribution of the MLRT is unavailable in this case, the critical values have to be simulated if we forcefully implement it. In this sense, the MLRT is not truly a viable method. Because we need a yardstick for the EM-test, we decide to carry out the MLRT nevertheless using the simulated critical values. The comparison should provide us information on whether the EM-test happen to be poor in the current case. As will be seen, the EM-test passed this test and works well. The simulation results tell us, $EM_n^{(0)}$ and $EM_n^{(1)}$ based on three initial values (0.1,0.3,0.5) of α almost have the same power as the MLRT. Further increasing the number of the iterations or the number of initial values on α may not increase the power of the EM-test statistics. So in applications, we suggest the use of $EM_n^{(0)}$ or $EM_n^{(1)}$ based on three initial values (0.1, 0.3, 0.5) of α. In Table 4.5, the last alternative model is also used to check whether the penalty function $p_n(\sigma)$ can handle the case when the data have some isolated points. In the 10,000 repetitions, the estimations of σ_1^2 and σ_2^2 are all far away from 0. That is, the penalty function $p_n(\sigma)$ efficiently prevents the degenerate situation.

4.3.4 Technical Proofs

In this subsection, we first prove some general results and then apply these results to show Theorems 4.3.1 and 4.3.2. Without loss of generality, the null distribution is assumed to be $N(0, 1)$.

Lemma 4.3.1. *Suppose that Conditions C1-C2 and C5 hold. Let* $(\bar{\alpha}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}_1, \bar{\sigma}_2)$ *be any*

Table 4.5: Parameters in alternative normal mixture models in both mean and variance parameters.

	$1-\alpha$	θ_1	θ_2	σ_1	σ_2	$100KL$
Model I	0.50	0.75	-0.75	1.20	0.80	4.258
Model II	0.25	0.65	-0.65	1.20	0.80	4.647
Model III	0.10	0.85	-0.85	1.20	0.80	4.611
Model IV	0.05	1.15	-1.15	1.20	0.80	4.974
Model V	$2/n$	1.50	-1.50	0.75	0.25	–

KL: Kullback-Leibler information.

estimators of $(\alpha, \mu_1, \mu_2, \sigma_1, \sigma_2)$ *such that* $\delta \leq \bar{\alpha} \leq 1-\delta$ *for some* $\delta \in (0, 0.5]$. *Assume that*

$$pl_n(\bar{\alpha}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}_1, \bar{\sigma}_2) - pl_n(0.5, 0, 0, 1, 1) \geq c > -\infty.$$

Then under null distribution $N(0,1)$, $\bar{\mu}_h = o_p(1)$ *and* $\bar{\sigma}_h - 1 = o_p(1)$ *for* $h = 1, 2$.

Under Conditions C1-C2 and C5, Chen et al. (2007) proved that if $pl_n(\bar{\alpha}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}_1, \bar{\sigma}_2) - pl_n(0.5, 0, 0, 1, 1) \geq c > -\infty$, then

$$\bar{\Psi}(\mu, \sigma) = \sum_{j=1}^{2} \bar{\alpha}_j I(\bar{\mu}_j \leq \mu, \bar{\sigma}_j \leq \sigma)$$

is a consistent estimator of

$$\Psi_0(\mu, \sigma) = I(0 \leq \mu, 1 \leq \sigma),$$

the mixing distribution under the null model. Using the condition that $\alpha \in [\delta, 1-\delta]$ for some $\delta \in (0, 0.5]$, Lemma 4.3.1 follows directly.

Let $(\bar{\alpha}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}_1, \bar{\sigma}_2)$ be some estimators of $(\alpha, \mu_1, \mu_2, \sigma_1, \sigma_2)$ as before. Define

$$H_n(\alpha) \;\; = \;\; (n - \sum_{i=1}^{n} \bar{w}_i) \log(1-\alpha) + \sum_{i=1}^{n} \bar{w}_i \log(\alpha) + p(\alpha)$$

Table 4.6: Simulated powers (%) of the EM-test and the MLRT under normal mixture models in both mean and variance parameters at the 5% level.

Model	$EM_n^{(0)}$	$EM_n^{(1)}$	$EM_n^{(2)}$	$EM_n^{(0)}$	$EM_n^{(1)}$	$EM_n^{(2)}$	MLRT
	$n = 100$						
I	47.8	47.8	47.7	47.9	47.7	47.7	47.6
II	55.1	55.1	54.9	55.4	55.2	55.0	54.7
III	55.9	55.9	55.8	56.2	55.9	55.9	55.6
IV	58.4	58.4	58.4	58.7	58.4	58.4	58.4
V	48.7	48.7	48.6	48.9	48.6	48.6	48.8
	$n = 200$						
I	81.9	81.8	81.8	82.1	81.9	81.9	81.6
II	87.5	87.4	87.4	87.8	87.6	87.6	87.3
III	86.0	85.9	85.9	86.2	86.1	86.1	85.9
IV	87.3	87.2	87.2	87.5	87.4	87.4	87.1
V	46.7	46.5	46.5	46.9	46.7	46.7	46.6

Results in columns (2, 3, 4) used $\alpha = (0.1, 0.2, 0.3, 0.4, 0.5)$.

Results in columns (5, 6, 7) used $\alpha = (0.1, 0.3, 0.5)$.

with

$$\bar{w}_i = \frac{\bar{\alpha} f(X_i; \bar{\mu}_2, \sigma_2)}{(1-\bar{\alpha}) f(X_i; \bar{\mu}_1, \sigma_1) + \bar{\alpha} f(X_i; \bar{\mu}_2, \sigma_2)}.$$

Let $\bar{\alpha}^* = \arg\max_\alpha H_n(\alpha)$. Using the same techniques in Lemma 3.5.3 and Lemma 4.2.4, we have the following lemma considering the asymptotic properties of $\bar{\alpha}^*$.

Lemma 4.3.2. *Let* $(\bar{\alpha}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}_1, \bar{\sigma}_2)$ *be any estimators of* $(\alpha, \mu_1, \mu_2, \sigma_1, \sigma_2)$ *such that under* *null hypothesis,*

$$\bar{\mu}_h = o_p(1) \ \textit{and} \ \bar{\sigma}_h - 1 = o_p(1), \ \textit{for} \ h = 1, 2.$$

If $|\bar{\alpha} - \alpha_0| = o_p(1)$ *for some* $\alpha_0 \in (0, 1)$*. Then under the null distribution* $N(0, 1)$*, we have* $|\bar{\alpha}^* - \alpha_0| = o_p(1)$*.*

The results in Lemmas 4.3.1 and 4.3.2 will be used to prove Theorem 4.3.1 as follows.

Proof of Theorem 4.3.1

By the property of the EM algorithm (Dempster et al., 1977), the definition of $\alpha_j^{(k)}$ and others, for any finite k, we have

$$pl_n(\alpha_j^{(k)}, \mu_{j1}^{(k)}, \mu_{j2}^{(k)}, \sigma_{j1}^{(k)}, \sigma_{j2}^{(k)}) \geq pl_n(\alpha_j^{(0)}, \mu_{j1}^{(0)}, \mu_{j2}^{(0)}, \sigma_{j1}^{(0)}, \sigma_{j2}^{(0)}) \geq pl_n(\alpha_j, 0, 0, 1, 1).$$

Therefore

$$pl_n(\alpha_j^{(k)}, \mu_{j1}^{(k)}, \mu_{j2}^{(k)}, \sigma_{j1}^{(k)}, \sigma_{j2}^{(k)}) - pl_n(0.5, 0, 0, 1, 1) \quad \geq \quad p(\alpha_j) - p(0.5) > -\infty.$$

By Lemma 4.3.2, we find

$$\alpha_j^{(1)} - \alpha_j = o_p(1).$$

Then applying the results in Lemma 4.3.1, we get

$$\mu_{jh}^{(1)} - 0 = o_p(1) \text{ and } \sigma_{jh}^{(1)} - 1 = o_p(1), \ h = 1, 2.$$

Applying this conclusions repeatedly, it is seen that

$$\alpha^{(k)} - \alpha_j = o_p(1), \, \mu_{jh}^{(k)} - 0 = o_p(1) \text{ and } \sigma_{jh}^{(k)} - 1 = o_p(1), \ h = 1, 2.$$

is true for all finite k. □

We now study the asymptotic distribution of the EM-test. We employ here the "sandwich method" to derive the limiting distribution of the EM-test statistic $EM_n^{(k)}$ under the null hypothesis.

Suppose $(\bar{\alpha}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}_1, \bar{\sigma}_2)$ are the estimators of $(\alpha, \mu_1, \mu_2, \sigma_1, \sigma_2)$ such that $\bar{\mu}_h = o_p(1)$ and $\bar{\sigma}_h - 1 = o_p(1)$, $h = 1, 2$, and $\delta \leq \bar{\alpha} \leq 1 - \delta$ for some $\delta \in (0, 0.5]$. In the first stage, we derive an upper bound for $2\{pl_n(\bar{\alpha}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}_1, \bar{\sigma}_2) - pl_n(0.5, \hat{\mu}_0, \hat{\mu}_0, \hat{\sigma}_0, \hat{\sigma}_0)\}$ and then apply the result to $EM_n^{(k)}$.

Note that

$$\begin{aligned}
& 2\{pl_n(\bar{\alpha}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}_1, \bar{\sigma}_2) - pl_n(0.5, \hat{\mu}_0, \hat{\mu}_0, \hat{\sigma}_0, \hat{\sigma}_0)\} \\
= \; & 2\{l_n(\bar{\alpha}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}_1, \bar{\sigma}_2) - l_n(0.5, \hat{\mu}_0, \hat{\mu}_0, \hat{\sigma}_0, \hat{\sigma}_0) \\
& + p_n(\bar{\sigma}_1) - p_n(1) + p_n(\bar{\sigma}_2) - p_n(1) + p(\bar{\alpha}) - p(0.5)\} \\
\leq \; & 2\{l_n(\bar{\alpha}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}_1, \bar{\sigma}_2) - l_n(0.5, \hat{\mu}_0, \hat{\mu}_0, \hat{\sigma}_0, \hat{\sigma}_0) \\
& + p_n(\bar{\sigma}_1) - p_n(1) + p_n(\bar{\sigma}_2) - p_n(1)\} \\
= \; & r_{1n}(\bar{\alpha}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}_1, \bar{\sigma}_2) + r_{2n} + r_{3n}(\bar{\sigma}_1, \bar{\sigma}_2),
\end{aligned}$$

where

$$r_{1n}(\bar{\alpha}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}_1, \bar{\sigma}_2) = 2\{l_n(\bar{\alpha}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}_1, \bar{\sigma}_2) - l_n(0.5, 0, 0, 1, 1)\},$$

$$r_{2n} = 2\{l_n(0.5, 0, 0, 1, 1) - l_n(0.5, \hat{\mu}_0, \hat{\mu}_0, \hat{\sigma}_0, \hat{\sigma}_0)\}$$

and

$$r_{3n}(\bar{\sigma}_1, \bar{\sigma}_2) = 2\{p_n(\bar{\sigma}_1) - p_n(1) + p_n(\bar{\sigma}_2) - p_n(1)\}.$$

To analyze r_{1n}, express $r_{1n} = 2\sum_{i=1}^n \log(1 + \bar{\delta}_i)$, where

$$\bar{\delta}_i = (1 - \bar{\alpha})\Big\{\frac{f(X_i, \bar{\mu}_1, \bar{\sigma}_1)}{f(X_i; 0, 1)} - 1\Big\} + \bar{\alpha}\Big\{\frac{f(X_i, \bar{\mu}_2, \bar{\sigma}_2)}{f(X_i; 0, 1)} - 1\Big\}.$$

By the inequality $2\log(1 + x) \leq 2x - x^2 + 2/3x^3$, we have

$$r_{1n} \leq 2\sum_{i=1}^n \delta_i - \sum_{i=1}^n \delta_i^2 + 2/3\sum_{i=1}^n \delta_i^3.$$

For $l = 1, 2, 3, 4$ and $s = 1, 2, 3, 4$, we define

$$\bar{m}_{l,s} = (1 - \bar{\alpha})\bar{\mu}_1^l(\bar{\sigma}_1^2 - 1)^s + \bar{\alpha}\bar{\mu}_2^l(\bar{\sigma}_2^2 - 1)^s.$$

Using the Taylor's expansions for $f(X_i; \bar{\mu}_h, \bar{\sigma}_h)$ up to order 4, $h = 1, 2$, we have

$$\delta_i = \sum_{l+s=1}^{4} \binom{l+s}{s} \bar{m}_{l,s} \frac{f^{(l,s)}(X_i; 0, 1)}{(l+s)! f(X_i; 0, 1)} + \epsilon_{in}^{(1)}$$

and the remainder term $\epsilon_n^{(1)} = \sum_{i=1}^n \epsilon_{in}^{(1)}$ satisfies

$$\epsilon_n^{(1)} = O_p(n^{1/2}) \Big\{ \sum_{h=1}^{2} \sum_{k=0}^{5} |\bar{\mu}_h|^k |\bar{\sigma}_h^2 - 1|^{5-k} \Big\}. \tag{4.17}$$

Here $f^{(l,s)}(X_i; \mu, \sigma)$ is defined to be

$$f^{(l,s)}(x; \mu, \sigma) = \partial^{l+s} f(x; \mu, \sigma) / \{\partial \mu^l \partial (\sigma^2)^s\}, \; l, s \geq 0.$$

Note that when $k = 0, 1, 2$,

$$O_p(n^{1/2}) \sum_{h=1}^{2} |\bar{\mu}_h|^k |\bar{\sigma}_h^2 - 1|^{5-k} \leq O_p(n^{1/2}) \sum_{h=1}^{2} |\bar{\sigma}_h^2 - 1|^3$$

and when $k = 3, 4$,

$$O_p(n^{1/2}) \sum_{h=1}^{2} |\bar{\mu}_h|^k |\bar{\sigma}_h^2 - 1|^{5-k} \leq O_p(n^{1/2}) \sum_{h=1}^{2} |\bar{\mu}_h|^3 |\bar{\sigma}_h^2 - 1|.$$

Therefore, (4.17) is simplified to

$$\epsilon_n^{(1)} = O_p(n^{1/2}) \{ |\bar{\mu}_h|^5 + |\bar{\mu}_h|^3 |\bar{\sigma}_h^2 - 1| + |\bar{\sigma}_h^2 - 1|^3 \}.$$

Absorbing the term $\bar{m}_{l,s}$ such that $l + 2s \geq 5$ into the remainder term, we have

$$\delta_i \;\; = \;\; \sum_{l+2s=1}^{4} \binom{l+s}{s} \bar{m}_{l,s} \frac{f^{(l,s)}(X_i; 0, 1)}{(l+s)! f(X_i; 0, 1)} + \epsilon_{in} \tag{4.18}$$

with

$$\epsilon_n = \sum_{i=1}^{n} \epsilon_{in} = O_p(n^{1/2}) \sum_{h=1}^{2} \{|\bar{\mu}_h|^5 + |\bar{\mu}_h|^3 |\bar{\sigma}_h^2 - 1| + |\bar{\mu}_h| (\bar{\sigma}_h^2 - 1)^2 + |\bar{\sigma}_h^2 - 1|^3\}. \quad (4.19)$$

Note that all the other terms in δ_i have been considered in the above absorption. For example,

$$O_p(n^{1/2}) \bar{m}_{2,2} = O_p(n^{1/2}) \sum_{h=1}^{2} \bar{\mu}_h^2 (\sigma_h^2 - 1)^2 \leq O_p(n^{1/2}) \sum_{h=1}^{2} |\bar{\mu}_h| (\sigma_h^2 - 1)^2$$

and

$$O_p(n^{1/2}) \bar{m}_{0,4} = O_p(n^{1/2}) \sum_{h=1}^{2} (\sigma_h^2 - 1)^4 \leq O_p(n^{1/2}) \sum_{h=1}^{2} |\sigma_h^2 - 1|^3.$$

Using the fact that $|2x| \leq 1 + x^2$, we have

$$O_p(n^{1/2}) |\bar{\mu}_h|^3 |\bar{\sigma}_h^2 - 1| \leq O_p(n^{1/2}) \{|\bar{\mu}_h|^5 + |\bar{\mu}_h| (\bar{\sigma}_h^2 - 1)^2\}.$$

So (4.19) reduces to

$$\epsilon_n = \sum_{i=1}^{n} \epsilon_{in} = O_p(n^{1/2}) \sum_{h=1}^{2} \{|\bar{\mu}_h|^5 + |\bar{\mu}_h| (\bar{\sigma}_h^2 - 1)^2 + |\bar{\sigma}_h^2 - 1|^3\}. \quad (4.20)$$

Next we come to simplify the dominant term of δ_i. By calculation, (4.18) further reduces to

$$\delta_i = \bar{t}_1 Y_i + \bar{t}_2 Z_i + \bar{t}_3 U_i + \bar{t}_4 V_i + \epsilon_{in},$$

where Y_i, Z_i, U_i and V_i are the same as those defined in (4.6) and

$$\bar{t}_1 = \bar{m}_{1,0},\ \bar{t}_2 = \bar{m}_{2,0} + \bar{m}_{0,1},\ \bar{t}_3 = \bar{m}_{3,0} + 3\bar{m}_{1,1} \text{ and } \bar{t}_4 = \bar{m}_{4,0} + 6\bar{m}_{2,1} + 3\bar{m}_{0,2}. \quad (4.21)$$

Note that the remainder term for square term and cubic term will be as high as ϵ_n. So

$$\begin{aligned} r_{1n}(\bar{\alpha}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}_1, \bar{\sigma}_2) \ \leq \ & 2\{\bar{t}_1 \sum_{i=1}^{n} Y_i + \bar{t}_2 \sum_{i=1}^{n} Z_i + \bar{t}_3 \sum_{i=1}^{n} U_i + \bar{t}_4 \sum_{i=1}^{n} V_i\} \\ & -\{\bar{t}_1^2 \sum_{i=1}^{n} Y_i^2 + \bar{t}_2^2 \sum_{i=1}^{n} Z_i^2 + \bar{t}_3^2 \sum_{i=1}^{n} U_i^2 + \bar{t}_4^2 \sum_{i=1}^{n} V_i^2\}\{1 + o_p(1)\} \\ & +2/3\{\bar{t}_1 \sum_{i=1}^{n} Y_i + \bar{t}_2 \sum_{i=1}^{n} Z_i + \bar{t}_3 \sum_{i=1}^{n} U_i + \bar{t}_4 \sum_{i=1}^{n} V_i\}^3 + O_p(\epsilon_n). \end{aligned}$$

It is easy to verify that (Y_i, Z_i, U_i, V_i) are mutually orthogonal, therefore we do not have the cross term in the above square term. Further using the inequality

$$(a + b)^3 \leq 4(a^3 + b^3), \ a, b \geq 0$$

repeatedly, we have that

$$\begin{aligned} & \Big| \{\bar{t}_1 \sum_{i=1}^{n} Y_i + \bar{t}_2 \sum_{i=1}^{n} Z_i + \bar{t}_3 \sum_{i=1}^{n} U_i + \bar{t}_4 \sum_{i=1}^{n} V_i\}^3 \Big| \\ \leq \ & 16\Big\{ |\bar{t}_1|^3 \sum_{i=1}^{n} |Y_i|^3 + |\bar{t}_2|^3 \sum_{i=1}^{n} |Z_i|^3 + |\bar{t}_3|^3 \sum_{i=1}^{n} |U_i|^3 + |\bar{t}_4|^3 \sum_{i=1}^{n} |V_i|^3 \Big\} \\ = \ & O_p(n) \Big\{ \sum_{l=1}^{4} |\bar{t}_l|^3 \Big\} \\ = \ & o_p(n) \Big\{ \sum_{l=1}^{4} \bar{t}_l^2 \Big\}, \end{aligned}$$

which means the cubic term is dominated by the square term. Hence,

$$\begin{aligned} r_{1n}(\bar{\alpha}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}_1, \bar{\sigma}_2) \ \leq \ & 2\{\bar{t}_1 \sum_{i=1}^{n} Y_i + \bar{t}_2 \sum_{i=1}^{n} Z_i + \bar{t}_3 \sum_{i=1}^{n} U_i + \bar{t}_4 \sum_{i=1}^{n} V_i\} \\ & -\{\bar{t}_1^2 \sum_{i=1}^{n} Y_i^2 + \bar{t}_2^2 \sum_{i=1}^{n} Z_i^2 + \bar{t}_3^2 \sum_{i=1}^{n} U_i^2 + \bar{t}_4^2 \sum_{i=1}^{n} V_i^2\}\{1 + o_p(1)\} \\ & +O_p(\epsilon_n). \end{aligned} \tag{4.22}$$

Our next step is to show ϵ_n is also a high order term than the square term. From (4.20), the key point is to show that

$$\epsilon_n = o_p(n)\Big\{\sum_{l=1}^{4}\bar{t}_l^2\Big\}, \tag{4.23}$$

which is an immediate consequence of the following lemma.

Lemma 4.3.3. *Suppose $(\bar{\alpha},\bar{\mu}_1,\bar{\mu}_2,\bar{\sigma}_1,\bar{\sigma}_2)$ are the estimators of $(\alpha,\mu_1,\mu_2,\sigma_1,\sigma_2)$ such that under null hypothesis, $\bar{\mu}_h = o_p(1)$ and $\bar{\sigma}_h - 1 = o_p(1)$, $h = 1,2$ and $\delta \le \bar{\alpha} \le 1-\delta$ for some $\delta \in (0, 0.5]$. Then under null distribution $N(0,1)$,*

$$\bar{\mu}_h^5 = o_p\Big\{\sum_{l=1}^{4}|\bar{t}_l|\Big\},\ \bar{\mu}_h(\bar{\sigma}_h^2-1)^2 = o_p\Big\{\sum_{l=1}^{4}|\bar{t}_l|\Big\}\ \textit{and}\ (\bar{\sigma}_h^2-1)^3 = o_p\Big\{\sum_{l=1}^{4}|\bar{t}_l|\Big\},\ h=1,2.$$

Proof. Because $\bar{\mu}_h = o_p(1)$ and $\bar{\sigma}_h - 1 = o_p(1)$, $h = 1,2$, we have $\bar{t}_l = o_p(1)$, $l = 1,2,3,4$. Let $\bar{\beta}_h = \bar{\mu}_h^2 + \bar{\sigma}_h^2 - 1$, $h = 1,2$. According to the definitions of $\bar{t}_1$ and $\bar{t}_2$ in (4.21), we can obtain the following relationships:

$$\bar{\mu}_2 = \bar{t}_1/\bar{\alpha} - (1-\bar{\alpha})\bar{\mu}_1/\bar{\alpha}, \tag{4.24}$$

$$\bar{\beta}_2 = \bar{t}_2/\bar{\alpha} - (1-\bar{\alpha})\bar{\beta}_1/\bar{\alpha}. \tag{4.25}$$

Plugging (4.24) and (4.25) into the definitions of $\bar{t}_3$ and $\bar{t}_4$ in (4.21) and using the condition that $\delta \le \bar{\alpha} \le 1-\delta$ for some $\delta \in (0, 0.5]$, we can easily show

$$\bar{t}_3 = 3\frac{1-\bar{\alpha}}{\bar{\alpha}}\Big\{\bar{\mu}_1\bar{\beta}_1 - \frac{2(2\bar{\alpha}-1)}{3\bar{\alpha}}\bar{\mu}_1^3\Big\} + o_p(\bar{t}_1) + o_p(\bar{t}_2), \tag{4.26}$$

$$\bar{t}_4 = 3\frac{1-\bar{\alpha}}{\bar{\alpha}}\Big\{\bar{\beta}_1^2 - \frac{2(1-3\bar{\alpha}+3\bar{\alpha}^2)}{3\bar{\alpha}^2}\bar{\mu}_1^4\Big\} + o_p(\bar{t}_1) + o_p(\bar{t}_2). \tag{4.27}$$

From (4.26) × $\Big\{\bar{\beta}_1 + \frac{2(2\bar{\alpha}-1)}{3\bar{\alpha}}\bar{\mu}_1^2\Big\}$ − (4.27) × $\bar{\mu}_1$, we get

$$\frac{2(1-\bar{\alpha})(1-\bar{\alpha}+\bar{\alpha}^2)}{3\bar{\alpha}^3}\bar{\mu}_1^5 = o_p(\bar{t}_1) + o_p(\bar{t}_2) + o_p(\bar{t}_3) + o_p(\bar{t}_4).$$

Using the fact $\bar{\alpha} \in [\delta, 1-\delta]$, for some $\delta \in (0, 0.5]$, we conclude that

$$\bar{\mu}_1^5 = o_p\Big(\sum_{l=1}^{4} |\bar{t}_l|\Big). \tag{4.28}$$

From (4.27) × $\bar{\mu}_1$ and the above result, we have

$$\bar{\mu}_1 \bar{\beta}_1^2 = o_p\Big(\sum_{l=1}^{4} |\bar{t}_l|\Big).$$

From the definition of $\bar{\beta}_1$, it is seen that

$$|\bar{\mu}_1(\bar{\sigma}_1^2 - 1)^2| \leq 2|\bar{\mu}_1|(\bar{\beta}_1^2 + \bar{\mu}_1^4) = o_p\Big(\sum_{l=1}^{4} |\bar{t}_l|\Big).$$

From (4.27) × $\bar{\beta}_1$ + (4.26) × $\frac{2(1-3\bar{\alpha}+3\bar{\alpha}^2)}{3\bar{\alpha}^2}\bar{\mu}_1^3$ and (4.28), we get

$$|\bar{\beta}_1|^3 = o_p\Big(\sum_{l=1}^{4} |\bar{t}_l|\Big).$$

Using the inequality $(a+b)^3 \leq 4(a^3+b^3),\ a, b \geq 0$ and the definition of $\bar{\beta}_1$, we obtain

$$|(\bar{\sigma}_1^2 - 1)^3| \leq 4(|\bar{\beta}_1|^3 + |\bar{\mu}_1|^6) = o_p\Big(\sum_{l=1}^{4} |\bar{t}_l|\Big).$$

Other parts can be done similarly. □

Combining (4.22) and (4.23), the upper bound in (4.22) can further reduce to

$$\begin{aligned} r_{1n}(\bar{\alpha}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}_1, \bar{\sigma}_2) \quad \leq \quad & 2\{\bar{t}_1 \sum_{i=1}^{n} Y_i + \bar{t}_2 \sum_{i=1}^{n} Z_i + \bar{t}_3 \sum_{i=1}^{n} U_i + \bar{t}_4 \sum_{i=1}^{n} V_i\} \\ & -\{\bar{t}_1^2 \sum_{i=1}^{n} Y_i^2 + \bar{t}_2^2 \sum_{i=1}^{n} Z_i^2 + \bar{t}_3^2 \sum_{i=1}^{n} U_i^2 + \bar{t}_4^2 \sum_{i=1}^{n} V_i^2\}\{1 + o_p(1)\}. \end{aligned} \tag{4.29}$$

According to the classic results about regular models, we have

$$r_{2n} = -\frac{(\sum_{i=1}^{n} Y_i)^2}{\sum_{i=1}^{n} Y_i^2} - \frac{(\sum_{i=1}^{n} Z_i)^2}{\sum_{i=1}^{n} Z_i^2} + o_p(1). \tag{4.30}$$

Now we come to analyze $r_{3n}(\bar{\sigma}_1, \bar{\sigma}_2)$. Using mean theorem on $p_n(\bar{\sigma}_h) - p_n(1)$, $h = 1, 2$, and Condition C4, we can get

$$\begin{aligned} r_{3n}(\bar{\sigma}_1, \bar{\sigma}_2) &= o_p(n^{1/6})(|\bar{\sigma}_1^2 - 1| + |\bar{\sigma}_2^2 - 1|) \\ &\leq o_p(1) + o_p(n^{1/2})\{(\bar{\sigma}_1^2 - 1)^3 + (\bar{\sigma}_2^2 - 1)^3\} \\ &= o_p(1) + o_p(n^{1/2})\Big\{\sum_{l=1}^{4} |\bar{t}_l|\Big\} \\ &\leq o_p(1) + o_p(n)\Big\{\sum_{l=1}^{4} \bar{t}_l^2\Big\}. \end{aligned} \tag{4.31}$$

In the above second step, we use the fact that $|x| \leq 1 + |x|^3$, and in the third step, we apply the result in Lemma 4.3.3. Combining (4.29) and (4.31), we have

$$\begin{aligned} & r_{1n}(\bar{\alpha}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}_1, \bar{\sigma}_2) + r_{3n}(\bar{\sigma}_1, \bar{\sigma}_2) \\ \leq\ & 2\{\bar{t}_1\sum_{i=1}^{n} Y_i + \bar{t}_2\sum_{i=1}^{n} Z_i + \bar{t}_3\sum_{i=1}^{n} U_i + \bar{t}_4\sum_{i=1}^{n} V_i\} \\ & -\{\bar{t}_1^2\sum_{i=1}^{n} Y_i^2 + \bar{t}_2^2\sum_{i=1}^{n} Z_i^2 + \bar{t}_3^2\sum_{i=1}^{n} U_i^2 + \bar{t}_4^2\sum_{i=1}^{n} V_i^2\}\{1 + o_p(1)\} + o_p(1). \end{aligned} \tag{4.32}$$

Inequality (4.32) implies $r_{1n}(\bar{\alpha}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}_1, \bar{\sigma}_2) + r_{3n}(\bar{\sigma}_1, \bar{\sigma}_2)$ is stochastically bounded by the maximum of the following quadratic function:

$$\begin{aligned} Q(t_1, t_2, t_3, t_4) &= 2\{t_1\sum_{i=1}^{n} Y_i + t_2\sum_{i=1}^{n} Z_i + t_3\sum_{i=1}^{n} U_i + t_4\sum_{i=1}^{n} V_i\} \\ &\quad -\{t_1^2\sum_{i=1}^{n} Y_i^2 + t_2^2\sum_{i=1}^{n} Z_i^2 + t_3^2\sum_{i=1}^{n} U_i^2 + t_4^2\sum_{i=1}^{n} V_i^2\}. \end{aligned}$$

We see that $Q(t_1, t_2, t_3, t_4)$ is maximized at $t_l = \hat{t}_l$, $l = 1, 2, 3, 4$, with

$$\hat{t}_1 = \frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n Y_i^2},\ \hat{t}_2 = \frac{\sum_{i=1}^n Z_i}{\sum_{i=1}^n Z_i^2},\ \hat{t}_3 = \frac{\sum_{i=1}^n U_i}{\sum_{i=1}^n U_i^2} \text{ and } \hat{t}_4 = \frac{\sum_{i=1}^n V_i}{\sum_{i=1}^n V_i^2} \tag{4.33}$$

and

$$Q(\hat{t}_1, \hat{t}_2, \hat{t}_3, \hat{t}_4) = \frac{(\sum_{i=1}^n Y_i)^2}{\sum_{i=1}^n Y_i^2} + \frac{(\sum_{i=1}^n Z_i)^2}{\sum_{i=1}^n Z_i^2} + \frac{(\sum_{i=1}^n U_i)^2}{\sum_{i=1}^n U_i^2} + \frac{(\sum_{i=1}^n V_i)^2}{\sum_{i=1}^n V_i^2} + o_p(1).$$

This implies that

$$\begin{aligned} & r_{1n}(\bar{\alpha}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}_1, \bar{\sigma}_2) + r_{3n}(\bar{\sigma}_1, \bar{\sigma}_2) \\ \leq\ & \frac{(\sum_{i=1}^n Y_i)^2}{\sum_{i=1}^n Y_i^2} + \frac{(\sum_{i=1}^n Z_i)^2}{\sum_{i=1}^n Z_i^2} + \frac{(\sum_{i=1}^n U_i)^2}{\sum_{i=1}^n U_i^2} + \frac{(\sum_{i=1}^n V_i)^2}{\sum_{i=1}^n V_i^2} + o_p(1). \end{aligned} \tag{4.34}$$

Combining (4.30) and (4.34), we can have the following lemma.

Lemma 4.3.4. *Suppose Conditions C1-C2 and C4-C5 hold. Let* $(\bar{\alpha}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}_1, \bar{\sigma}_2)$ *be any estimators of* $(\alpha, \mu_1, \mu_2, \sigma_1, \sigma_2)$ *such that under null hypothesis,* $\bar{\mu}_h = o_p(1)$ *and* $\bar{\sigma}_h - 1 = o_p(1)$, $h = 1, 2$, *and* $\delta \leq \bar{\alpha} \leq 1 - \delta$ *for some* $\delta \in (0, 0.5]$. *Then under null distribution* $N(0, 1)$,

$$\begin{aligned} & 2\{pl_n(\bar{\alpha}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}_1, \bar{\sigma}_2) - pl_n(0.5, \hat{\mu}_0, \hat{\mu}_0, \hat{\sigma}_0, \hat{\sigma}_0)\} \\ \leq\ & \frac{(\sum_{i=1}^n U_i)^2}{\sum_{i=1}^n U_i^2} + \frac{(\sum_{i=1}^n V_i)^2}{\sum_{i=1}^n V_i^2} + o_p(1). \end{aligned} \tag{4.35}$$

This result is going to be used to establish Theorem 4.3.2 as follows.

Proof of Theorem 4.3.2

According to Theorem 4.3.1, we can get that (4.35) also serves as an upper bound for $EM_n^{(k)}$. Now we come to show that the upper bound is also achievable.

Let $\tilde{\alpha} = 0.5$ and $\tilde{\mu}_h$ and $\tilde{\sigma}_h^2$, $h = 1, 2$, be the solutions for the following four equations:

$$\left\{ \begin{array}{rcl} 1/2\mu_1 + 1/2\mu_2 & = & \hat{t}_1 \\ 1/2\beta_1 + 1/2\beta_2 & = & \hat{t}_2 \\ 3\mu_1\beta_1 & = & \hat{t}_3 \\ 3\beta_1^2 - 2\mu_1^4 & = & \hat{t}_4 \end{array} \right. , \tag{4.36}$$

where $\beta_h = \sigma_h^2 - 1 + \mu_h^2$, $h = 1, 2$. Note that from (4.26) and (4.27), $3\mu_1\beta_1$ and $3\beta_1^2 - 2\mu_1^4$ are the dominant term of t_3 and t_4, respectively. By setting the equations in the way of (4.36), it is easy to show the solutions exist as follows

From the last two equations of (4.36), we need to solve the following equation for μ_1:

$$g(\mu_1^2) = 6\mu_1^6 + 3\hat{t}_4\mu_1^2 - \hat{t}_3^2 = 0. \tag{4.37}$$

Note that $g(0) < 0$ and $g(\mu_1^2) \to \infty$ as $\mu_1 \to \infty$, therefore there exists a positive solution for μ_1^2. Let $\tilde{\mu}_1$ be the smallest positive solution. With $\tilde{\mu}_1$, we find $\tilde{\beta}_1$ from the third equation of (4.36), and then from the first two equations of (4.36), we can get $\tilde{\mu}_2$ and $\tilde{\beta}_2$. Therefore, the solutions for the four equations in (4.36) exist.

Note that directly solving the equation in (4.37), we can further get $\tilde{\mu}_1 = o_p(1)$. From (4.36), we conclude

$$\tilde{\mu}_2 = o_p(1),\ \tilde{\sigma}_1^2 - 1 = o_p(1) \text{ and } \tilde{\sigma}_2^2 - 1 = o_p(1).$$

Using the relationships in (4.26) and (4.27), we have

$$\tilde{t}_l = \hat{t}_l + o_p(n^{-1/2}),\ l = 1, 2, 3, 4, \tag{4.38}$$

where $\tilde{t}_l$ is similarly defined in (4.21) with $(\bar{\alpha}, \bar{\mu}_1, \bar{\mu}_2, \bar{\sigma}_1, \bar{\sigma}_2)$ being replaced by $(\tilde{\alpha}, \tilde{\mu}_1, \tilde{\mu}_2, \tilde{\sigma}_1, \tilde{\sigma}_2)$.

By (4.33) and (4.38), $(\tilde{\alpha}, \tilde{\mu}_1, \tilde{\mu}_2, \tilde{\sigma}_1, \tilde{\sigma}_2)$ are such that

$$\begin{aligned} & r_{1n}(\tilde{\alpha}, \tilde{\mu}_1, \tilde{\mu}_2, \tilde{\sigma}_1, \tilde{\sigma}_2) \\ = \ & \frac{(\sum_{i=1}^n Y_i)^2}{\sum_{i=1}^n Y_i^2} + \frac{(\sum_{i=1}^n Z_i)^2}{\sum_{i=1}^n Z_i^2} + \frac{(\sum_{i=1}^n U_i)^2}{\sum_{i=1}^n U_i^2} + \frac{(\sum_{i=1}^n V_i)^2}{\sum_{i=1}^n V_i^2} + o_p(1). \end{aligned} \tag{4.39}$$

Note that using the similar techniques in Lemma 4.3.3, we can show

$$(\tilde{\sigma}_h^2 - 1)^3 = o_p\Big\{\sum_{l=1}^4 |\tilde{t}_l|\Big\},\ h = 1, 2$$

and hence

$$(\tilde{\sigma}_h^2 - 1) = o_p(n^{-1/6}),\ h = 1, 2.$$

Applying the mean value theorem,

$$r_{3n}(\tilde{\sigma}_1, \tilde{\sigma}_2) = o_p(1). \tag{4.40}$$

Combining (4.30), (4.39) and (4.40), we have

$$\begin{aligned}
&2\{pl_n(\tilde{\alpha}, \tilde{\mu}_1, \tilde{\mu}_2, \tilde{\sigma}_1, \tilde{\sigma}_2) - pl_n(0.5, \hat{\mu}_0, \hat{\mu}_0, \hat{\sigma}_0, \hat{\sigma}_0)\} \\
&= \frac{(\sum_{i=1}^n U_i)^2}{\sum_{i=1}^n U_i^2} + \frac{(\sum_{i=1}^n V_i)^2}{\sum_{i=1}^n V_i^2} + o_p(1).
\end{aligned}$$

So if one of α_j's is equal to 0.5, then

$$\begin{aligned}
EM_n^{(k)} &\geq 2\{\sup_{(\mu_1,\mu_2,\sigma_1,\sigma_2)} pl_n(0.5, \mu_1, \mu_2, \sigma_1, \sigma_2) - pl_n(0.5, \hat{\mu}_0, \hat{\mu}_0, \hat{\sigma}_0, \hat{\sigma}_0)\} \\
&\geq 2\{pl_n(\tilde{\alpha}, \tilde{\mu}_1, \tilde{\mu}_2, \tilde{\sigma}_1, \tilde{\sigma}_2) - pl_n(0.5, \hat{\mu}_0, \hat{\mu}_0, \hat{\sigma}_0, \hat{\sigma}_0)\} \\
&= \frac{(\sum_{i=1}^n U_i)^2}{\sum_{i=1}^n U_i^2} + \frac{(\sum_{i=1}^n V_i)^2}{\sum_{i=1}^n V_i^2} + o_p(1).
\end{aligned}$$

That is, the upper bound for $EM_n^{(k)}$ is also an lower bound. Hence

$$EM_n^{(k)} = \frac{(\sum_{i=1}^n U_i)^2}{\sum_{i=1}^n U_i^2} + \frac{(\sum_{i=1}^n V_i)^2}{\sum_{i=1}^n V_i^2} + o_p(1).$$

Consequently, the limiting distribution of $EM_n^{(k)}$ is given by χ_2^2. □

4.4 Application and Real Example

Mixtures of multivariate normal distributions have been widely used in cluster analysis in multi-dimensional data sets, see McLachlan et al. (2002) and Tadesse et al. (2005). In many applications, only a small subset of variables is useful for cluster analysis. Including the unnecessary variables in cluster analysis could complicate or even mask the recovery of the clusters (see Tadesse et al. 2005 and the references therein). Variable selection has become a very important step before using the mixture of multivariate normal distributions in cluster analysis. McLachlan et al. (2002) and Charnigo and Sun (2004) suggested conducting the test of homogeneity for each variable to examine whether or not this variable is important for clustering.

To see how the EM-test might work in this situation, we apply it to Fisher's Iris data. This data set consists of 150 four-dimensional variables for three species of iris (Iris setosa, Iris versicolour, Iris virginica). Four measurements, sepal length, sepal width, petal length and petal width, are taken for each plant. The Iris data has been analyzed by several authors (see, e.g., Tadesse et al, 2005) in the framework of classification and clustering.

For illustration purposes, we take the first measurement, sepal length, for the first two species of iris: Iris setosa and Iris versicolour, which results in 100 observations in total. For the sepal length measurement, we test whether or not this variable is useful for cluster the two species of iris. Before conducting the formal test, we do some preliminary analysis for the 100 observations. Figure 4.2 is the Q-Q plot of this measurement, which suggests some deviance from the uni-component normal model. A rigorous EM-test can also be conducted. The analysis results are presented in Table 4.7. The results show strong evidence to reject the homogeneity in sepal length measurements. The Q-Q plot and the rigorous EM-test both favor the two-component normal mixture model. So the sepal length measurement is potentially important for clustering. Note that from our analysis we can only conclude that the two-component normal mixture model provides a more suitable description about the data than the homogeneous normal model. If we want to know whether or not the data is from a uni-component nonnormal model, we have to rely on some other methods or the scientific background behind this data.

Table 4.7: Homogeneity testing results for the sepal length observations under two normal mixture models.

	Normal mixture model $(\sigma_1^2 = \sigma_2^2 = \sigma^2)$		Normal mixture model $(\sigma_1^2 \neq \sigma_2^2)$
Measurement	MLRT	$IM_n^{(1)}$	$IM_n^{(1)}$
Sepal length	7.693	5.847	7.548
Asymptotic p-value	0.021	0.017	0.023
Simulated p-value	0.026	0.017	0.025

Results for the EM-test used $\alpha = (0.1, 0.3, 0.5)$

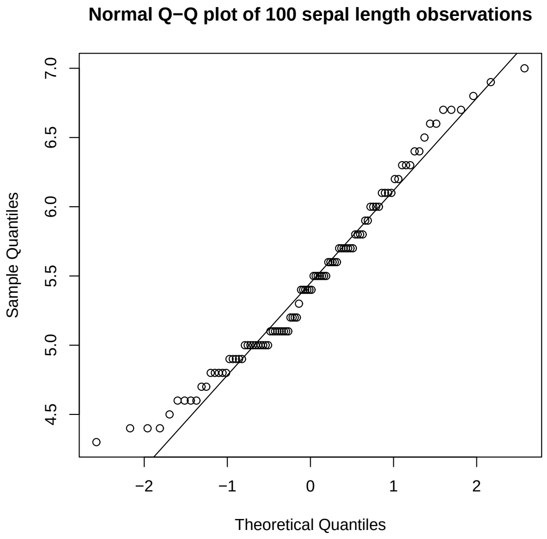

Figure 4.2: The Q-Q plot of the sepal length observations.

Chapter 5

Homogeneity Test in Mixture of Circular Distributions

5.1 Introduction

Circular data, which are measured in the form of angles or two dimensional orientations, arise commonly in many disciplines, including astronomy, biology, ecology, geology, physics and medicine. Examples of such data include the direction of flight of birds or the orientation of the movement of animals, wind and ocean current directions, circadian and other biorhythms. Describing and analyzing such data statistically poses a lot of interesting and challenging problems. For example, the sample mean defined for linear data is no longer appropriate to measure the center of circular data. Suppose two turtles moved at 10° and 350° measured clockwise from north. Their arithmetic mean is 180°, due south, while the two movements point toward north.

[1] The paper Chen, et al. (2007), based on this chapter, has been accepted for publication in *Canadian Journal of Statistics*.

The followings are several key monographs regarding the theory of circular statistics. Batschelet (1981) contains the descriptive and inferential tools for circular observations with a wealth of excellent examples. Fisher (1993) provides a nice introduction to statistical methods for analyzing circular data including an interesting historical overview of the subject. Mardia and Jupp (2000) cover a wide variety of topics in statistics of directional data. Jammalamadaka and Sengupta (2001) present the latest developments in circular statistics. Statistical inferences in circular or directional mixture models have been discussed by many authors, such as Stephens (1969), Fraser, et al. (1981), Hsu, et al. (1986), Kim and Koo (2000), and Holzmann, et al. (2004).

As a circular analog of the normal distribution on the real line, the von Mises distribution is the most commonly used distribution for circular data. The von Mises distribution also has two parameters, one represents the mean direction and the other describes the variation around the mean direction, called the concentration parameter. When the data are drawn from a heterogeneous population, mixtures of von Mises distributions with same concentration parameter are often used, see Grimshaw et al. (2001) for an example in geology. At the same time, it might be of value to know whether the data arise from a homogeneous or heterogeneous population. If the data are homogeneous, it is not even necessary to go into mixture modeling. Similar to the linear case, unless the components are fairly distinct, testing the order or the number of components in a circular mixture is a challenging problem.

The following is an illustrating example involving the orientation directions of elongate bones due to Grimshaw et al. (2001).

Example 5.1.1. *Information about the flow directions of ancient rivers (paleoflow direction) helps scientists better understand how certain rock units are oriented, which in turn leads to more efficient exploration of natural resources and better understanding of land-*

scape development and climate change. Primary bedforms are usually used to interpret the paleoflow direction, since the orientation of the foreset lamination of the bedforms parallels to the current direction. However, bedforms are often masked or destroyed by various physical and chemical processes, scientists then have to analyze other available data to obtain information on the paleoflow direction. Morris et al. (1996) *proposed the use of the orientation of elongate bones to identify the paleoflow direction. The Dinosaur National Monument and Dry Mesa Dinosaur Quarry are two ideal quarries for comparison of directions of elongate bone and paleoflow, since both dinosaur bone and well-preserved bedforms exist.*

The measurements on elongate dinosaur bones are axial data with period π*, since there is no reason to make a distinction of two ends of the fossil bone. In order to use the vectorial probability models, one can double the angles modulo 2*π*. The values of the transformed axial data then range from 0 to* 2π*. Dinosaur National Monument and Dry Mesa Dinosaur Quarry have 444 and 555 dinosaur bones direction measurements, respectively.*

As mentioned in Grimshaw et al. (2001)*, elongate bones can be classified into two categories: symmetrical and asymmetrical. Symmetrical bones tend to orient themselves vertical to the paleoflow direction, while asymmetrical bones, which display additional bone mass on only one end, tend to orient themselves parallel to the paleoflow direction. The primary interest of Grimshaw et al.* (2001) *is to make a rigorous comparison of dinosaur bone direction with primary bedforms. If these two directions are same, the scientists will have some confidence for the use of dinosaur bone orientations to estimate paleoflow direction when the bedforms are not visible. Before doing that, the statistical problem of interest is to test whether there exist two types of elongate bones in the two quarries.*

In the above example, since we do not have prior information on the orientation directions of the two types of bones, a two-component von Mises mixture in mean direction with the same but unknown concentration parameter is suitable. In this case, we have a straightforward test of homogeneity problem. Interestingly, unlike the normal mixture model at the presence of the structural parameter, von Mises mixture model satisfies the strong identifiability condition and the parameter space of the mean direction is compact. Therefore the theoretical results for normal mixture model do not apply to the mixture of von Mises distributions. In this chapter, we study the application of the MLRT and the EM-test to von Mises mixture models at the presence of the structural parameter. The remainder of this chapter is organized as follows. In Section 5.2, we study the circular moment property for a mixture of two von Mises distributions. The results suggest that the structural parameter tends to be overestimated when a two-component von Mises mixture is used to fit the data arising from a homogeneous von Mises distribution. This phenomenon motivates an additional penalty on the large values of the structural parameter. The asymptotic results of the MLRT and the EM-test are developed in Section 5.3. In Section 5.4, we present some simulation results and analyze the data sets in Example 5.1.1. Some additional discussions are given in Section 5.5. For the convenience of presentation, all the proofs are deferred to Section 5.6.

5.2 The von Mises Distribution and Circular Moments

The von Mises distribution was first introduced as a statistical model for directional data by von Mises (1918). It plays a key role in statistical inference for circular data. Because of its importance and similarities to the normal distribution on the line, it is also called the circular normal distribution. The von Mises distribution $M(\mu, \kappa)$ has probability density function (pdf)

$$f(x; \mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} \exp\{\kappa \cos(x - \mu)\}, \quad |x| \leq \pi,$$

where $|\mu| \leq \pi$ and $\kappa \geq 0$. The function $I_0(\kappa)$ is the normalizing constant and is known as the modified Bessel function of the first kind and order zero. In general, the modified Bessel function I_p of the first kind and order p (sometimes also called Bessel function of purely imaginary argument) is defined by

$$I_p(\kappa) = \frac{1}{2\pi} \int_0^{2\pi} \cos(px) \exp(\kappa \cos x) dx.$$

Also, we define

$$A(\kappa) = \frac{I_1(\kappa)}{I_0(\kappa)}.$$

Properties of these functions can be found in Abramowitz and Stegun (1965).

Figures 5.1 and 5.2 give plots of three density functions of $M(0,1)$, $M(0,2)$ and $M(0,3)$ on the interval $[-\pi, \pi]$ and on the circle, respectively. In Figure 5.2, the unit circle serves the role of the x axis in the linear plot in Figure 5.1. The distance between the plotted line and the unit circle is the value of the probability density function. If we cut at the point π and stretch the plot such that the unit circle becomes the real line, we will get the linear plot in Figure 5.1. It is seen that these probability density functions are unimodal and symmetric about μ, and as κ increases, the density function becomes more peaked at μ.

The measures of location and dispersion of circular data are defined differently from those for linear data. Let X be a circular random variable. The circular mean direction and the circular variance can be defined as

$$\mathrm{CE}(X) = \mathrm{argmin}_{\mu \in [-\pi,\pi]} \mathrm{E}\{2 - 2\cos(X - \mu)\}$$

and

$$\mathrm{CVar}(X) = \mathrm{E}[2 - 2\cos\{X - \mathrm{CE}(X)\}], \tag{5.1}$$

respectively. We may note that $2 - 2\cos(X - \mu) = 4\sin^2\{(X - \mu)/2\}$. Hence, the circular mean is the minimum point of a trigonometry distance, and the circular variance is the

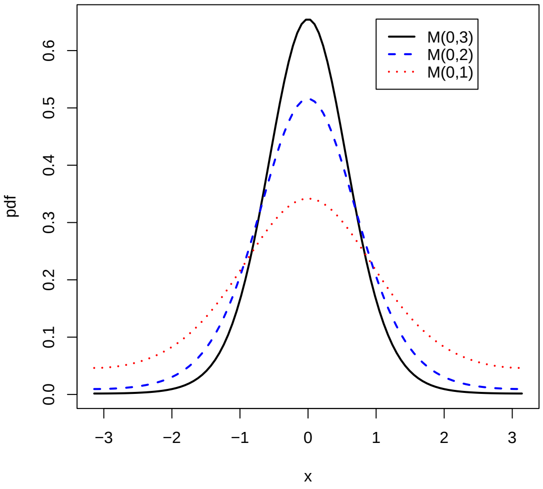

Figure 5.1: Three densities of von Mises distribution on $[-\pi, \pi]$.

resulting minimum value. Replacing the sine function in the definition by the identity function leads to the usual mean and variance for linear data.

For the von Mises distribution $M(\mu, \kappa)$, the mean direction is μ and the circular variance is $2 - 2A(\kappa)$. As κ increases, the circular variance decreases and the distribution places more mass close to the mean direction. Hence κ is also called the concentration parameter.

We now give the circular mean and variance for a mixture of two von Mises distributions. Let X be a circular random variable with distribution $(1-\alpha)M(\mu_1, \kappa) + \alpha M(\mu_2, \kappa)$ for some $\kappa > 0$. Then, for any μ, we have

$$E\{\cos(X - \mu)\} \quad = \quad A(\kappa)\cos(\eta - \mu)\sqrt{1 - 4\alpha(1-\alpha)\sin^2(\frac{\mu_1 - \mu_2}{2})},$$

where $\eta \in [-\pi, \pi]$ is an angle such that

$$\cos\eta = \{(1-\alpha)\cos\mu_1 + \alpha\cos\mu_2\}/\sqrt{1 - 4\alpha(1-\alpha)\sin^2(\frac{\mu_1 - \mu_2}{2})}$$

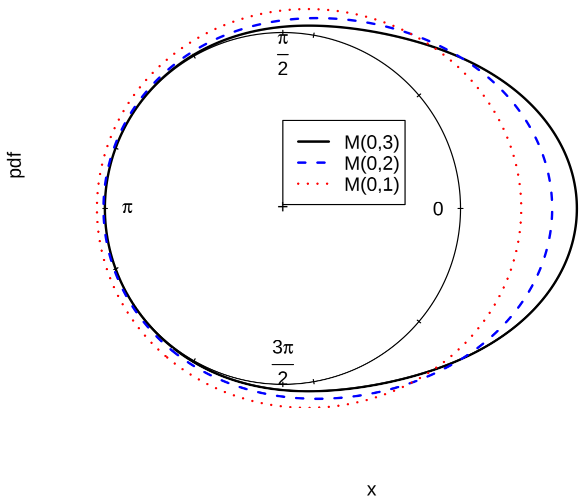

Figure 5.2: Three densities of von Mises distribution on the circle.

and

$$\sin\eta = \{(1-\alpha)\sin\mu_1 + \alpha\sin\mu_2\}/\sqrt{1-4\alpha(1-\alpha)\sin^2(\frac{\mu_1-\mu_2}{2})}.$$

Hence, we have

$$\mathrm{CE(X)} = \arg\max_{\mu} \mathrm{E}\{\cos(\mathrm{X}-\mu)\} = \eta.$$

Consequently, by (5.1)

$$\mathrm{CVar}(X) = 2 - 2A(\kappa)\sqrt{1-4\alpha(1-\alpha)\sin^2(\frac{\mu_1-\mu_2}{2})}.$$

It is seen that the variance of a heterogeneous model, where $\alpha(1-\alpha) \neq 0$ and $\mu_1 \neq \mu_2$, is larger than that of a homogeneous model with the same κ. Thus, because $A(\kappa)$ is an increasing function of κ, fitting a heterogeneous model to data arising from a homogeneous model tends to result in a larger fitted concentration parameter.

5.3 Asymptotic Properties of Likelihood-based Tests

Assume that a circular random sample $X_1, \ldots, X_n$ is drawn from the von Mises mixture distribution $(1-\alpha)M(\mu_1, \kappa) + \alpha M(\mu_2, \kappa)$, where $0 \le \alpha \le 1$, $-\pi \le \mu_1, \mu_2 \le \pi$ and $\kappa \ge 0$. We are interested in testing

$$H_0 : \alpha(1-\alpha)(\mu_1 - \mu_2) = 0 \tag{5.2}$$

versus the full model. This section focuses on the asymptotic properties of likelihood-based testing procedures.

5.3.1 The Likelihood Ratio Test

The log-likelihood function can be expressed as

$$l_n(\alpha, \mu_1, \mu_2, \kappa) = -n \log I_0(\kappa) + \sum \log[(1-\alpha)\exp\{\kappa \cos(X_i - \mu_1)\} + \alpha \exp\{\kappa \cos(X_i - \mu_2)\}].$$

Let $\hat{\mu}_0$ and $\hat{\kappa}_0$ be the MLEs under the null hypothesis and let $\hat{\alpha}$, $\hat{\mu}_1$, $\hat{\mu}_2$ and $\hat{\kappa}$ be the MLEs under the full model. The likelihood ratio test (LRT) statistic is defined as

$$R_n = 2\{l_n(\hat{\alpha}, \hat{\mu}_1, \hat{\mu}_2, \hat{\kappa}) - l_n(0.5, \hat{\mu}_0, \hat{\mu}_0, \hat{\kappa}_0)\}.$$

Without loss of generality, let $M(0, \kappa_0)$ be the null distribution.

The MLEs have some interesting properties as stated in the following two lemmas.

Lemma 5.3.1. *Assume that the distribution of the random sample $X_1, \ldots, X_n$ is given by $M(0, \kappa_0)$ for some $\kappa_0 > 0$. Let $\hat{\kappa}$ be the MLE of κ under the full model $(1-\alpha)M(\mu_1, \kappa) + \alpha M(\mu_2, \kappa)$. Then there exists a constant $0 < \Delta < \infty$ such that*

$$\lim_{n \to \infty} P(\hat{\kappa} \le \Delta) = 1.$$

As a consequence of Lemma 5.3.1, the parameter space under consideration can be reduced to a compact one for theoretical derivations. With identifiability (Fraser, et al. 1981, Holzmann, et al., 2004), Lemma 5.3.1 implies the consistency of the MLEs, which is a direct application of the result in Kiefer and Wolfowitz (1956). The proof is omitted.

Lemma 5.3.2. *Assume that the distribution of the random sample* $X_1, \dots, X_n$ *is given by* $M(0, \kappa_0)$. *Let* $\hat{\alpha}$, $\hat{\mu}_1$, $\hat{\mu}_2$, *and* $\hat{\kappa}$ *be the MLEs of* α, μ_1, μ_2, *and* κ *under the full model* $(1-\alpha)M(\mu_1, \kappa) + \alpha M(\mu_2, \kappa)$. *Then* $(1-\hat{\alpha})\hat{\mu}_1 + \hat{\alpha}\hat{\mu}_2 \to 0$, $(1-\hat{\alpha})\hat{\mu}_1^2 + \hat{\alpha}\hat{\mu}_2^2 \to 0$ *and* $\hat{\kappa} \to \kappa_0$ *in probability, as* $n \to \infty$.

The asymptotic distribution of the LRT statistic is given in the following theorem.

Theorem 5.3.1. *Let* $X_1, \dots, X_n$ *be a random sample from the mixture distribution* $(1-\alpha)M(\mu_1, \kappa) + \alpha M(\mu_2, \kappa)$, *where* $0 \le \alpha \le 1$, $-\pi \le \mu_1, \mu_2 \le \pi$ *and* $\kappa \ge 0$. *Let* R_n *be the LRT statistic for testing* $H_0 : \alpha(1-\alpha)(\mu_1 - \mu_2) = 0$. *Then under the null distribution* $M(0, \kappa_0)$, *as* $n \to \infty$,

$$R_n \xrightarrow{d} \sup_{|\mu| \le \pi} \{\zeta^+(\mu)\}^2,$$

where $\zeta(\mu)$, $|\mu| \le \pi$, *is a Gaussian process with mean 0, variance 1 and autocorrelation* $\rho(s,t)$ *which is given by*

$$\rho(s,t) = \frac{g(s,t)}{\{g(s,s)g(t,t)\}^{\frac{1}{2}}}, \qquad \text{for } s, t \ne 0,$$

where

$$\begin{aligned} g(s,t) \quad = \quad & \frac{1}{st}\Big[\frac{I_0[\kappa_0\{(\cos s + \cos t - 1)^2 + (\sin s + \sin t)^2\}^{\frac{1}{2}}]}{I_0(\kappa_0)} - 1 \\ & - \frac{A^2(\kappa_0)(\cos s - 1)(\cos t - 1)}{1 - A(\kappa_0)/\kappa_0 - A^2(\kappa_0)} - \kappa_0 A(\kappa_0) \sin s \sin t\Big]. \end{aligned} \tag{5.3}$$

Chen and Chen (2003) gives the asymptotic distribution of the LRT statistic in a two-component normal mixture with a structural parameter. In comparison with the result above, their Gaussian process contains a spike at 0. More discussions on the subtle difference between normal and von Mises mixtures will be given in Section 5.5.

5.3.2 The Modified Likelihood Approaches

Similar to the asymptotic results obtained in other finite mixture models, this limiting distribution of the LRT is not convenient to use in practice. An obvious alternative approach is the MLRT, which has been found easy to apply in the literature. Let the modified log-likelihood function be

$$pl_n(\alpha, \mu_1, \mu_2, \kappa) = l_n(\alpha, \mu_1, \mu_2, \kappa) + p(\alpha). \tag{5.4}$$

The MLRT statistic is defined as

$$M_n^* = 2\{pl_n(\hat{\alpha}^*, \hat{\mu}_1^*, \hat{\mu}_2^*, \hat{\kappa}^*) - pl_n(1/2, \hat{\mu}_0^*, \hat{\mu}_0^*, \hat{\kappa}_0^*)\},$$

where $(\hat{\alpha}^*, \hat{\mu}_1^*, \hat{\mu}_2^*, \hat{\kappa}^*)$ maximizes $pl_n(\alpha, \mu_1, \mu_2, \kappa)$ over the region $0 < \alpha < 1,\ -\pi \le \mu_1, \mu_2 \le \pi,\ \kappa \ge 0$, and $(\hat{\mu}_0^*, \hat{\mu}_0^*, \hat{\kappa}_0^*)$ maximizes $pl_n(1/2, \mu, \mu, \kappa)$ which is the modified log-likelihood function under the null hypothesis.

The result in Theorem 5.3.1 is instrumental in analyzing the asymptotic properties of the MLRT. The limiting distribution of the MLRT can be shown to be a mixture of chi-squared distributions, which is very convenient to implement in practice. However simulation studies indicate that the finite sample distribution of the above MLRT statistic under the null model is not well approximated by this null limiting distribution unless the sample size is very large. Two accuracy enhancing measures are hence proposed in the next subsection.

5.3.3 Accuracy Enhancing Methods

Due to the moment properties discussed in Section 5.2, κ tends to be overestimated by the MLE or the modified MLE under the heterogeneous model. As a consequence, the finite sample distribution of the MLRT statistic is stochastically larger than the limiting distribution which inflates the type I error rate. To overcome this problem, we propose penalizing the fit with larger values of κ. More specifically, we suggest adding $-\log(\kappa+1)$ to pl_n in (5.4) and the resulting modified likelihood function becomes

$$pl_n(\alpha, \mu_1, \mu_2, \kappa) = l_n(\alpha, \mu_1, \mu_2, \kappa) - \log(\kappa + 1) + p(\alpha). \tag{5.5}$$

The corresponding MLRT statistic M_n is defined in the same fashion as M_n^*. The limiting distribution of the new MLRT is not affected by this additional penalty. The result will be presented in next subsection. The penalty function $-\log(\kappa+1)$ is a decreasing function on κ with the upper bound 0. It prevents the overestimation of κ under the null model and improve the type-I error of the MLRT statistic. The second enhancing method is to select a more effective $p(\alpha)$. As we discussed in Chapter 2, we choose $p(\alpha) = C^* \log(1-|1-2\alpha|)$. Both enhancing methods are found to be effective as will be demonstrated by simulation studies.

5.3.4 The EM-test

In this subsection, the EM-test is adapted to the testing problem in (5.2). Under the current model, the EM-test procedure is carried out as follows.

Step 0: Choose a number of initial α values, say $0 < \alpha_1, \alpha_2, \ldots, \alpha_J \leq 0.5$. Compute

$$(\hat{\mu}_0^*, \hat{\kappa}_0^*) = \arg\max_{\mu,\, \kappa} pl_n(1/2, \mu, \mu, \kappa)$$

with pl_n given by (5.5).

Let $j = 1$ and $k = 0$.

Step 1: Let $\alpha_j^{(k)} = \alpha_j$.

Step 2: Compute

$$(\mu_{j1}^{(k)}, \mu_{j2}^{(k)}, \kappa_j^{(k)}) = \arg \max_{\mu_1,\, \mu_2,\, \kappa} pl_n(\alpha_j^{(k)}, \mu_1, \mu_2, \kappa)$$

and

$$M_n^{(k)}(\alpha_j) = 2\{pl_n(\alpha_j^{(k)}, \mu_{j1}^{(k)}, \mu_{j2}^{(k)}, \kappa_j^{(k)}) - pl_n(1/2, \hat{\mu}_0^*, \hat{\mu}_0^*, \hat{\kappa}_0^*)\}.$$

Step 3: For $i = 1, 2, \ldots, n$, compute the weights, which are conditional expectations in the E-step,

$$w_{ij}^{(k)} = \frac{\alpha_j^{(k)} f(X_i; \mu_{j2}^{(k)}, \kappa_j^{(k)})}{(1-\alpha_j^{(k)}) f(X_i; \mu_{j1}^{(k)}, \kappa_j^{(k)}) + \alpha_j^{(k)} f(X_i; \mu_{j2}^{(k)}, \kappa_j^{(k)})}.$$

Now following the M-step, let

$$\begin{aligned}
\alpha_j^{(k+1)} &= \arg\max_{\alpha} \{(n - \sum_{i=1}^{n} w_{ij}^{(k)}) \log(1-\alpha) + \sum_{i=1}^{n} w_{ij}^{(k)} \log(\alpha) + p(\alpha)\}, \\
\mu_{j1}^{(k+1)} &= \arg\max_{\mu_1} \Big\{ \sum_{i=1}^{n} (1 - w_{ij}^{(k)}) \cos(X_i - \mu_1) \Big\}, \\
\mu_{j2}^{(k+1)} &= \arg\max_{\mu_2} \Big\{ \sum_{i=1}^{n} w_{ij}^{(k)} \cos(X_i - \mu_2) \Big\}, \\
\kappa_j^{(k+1)} &= \arg\max_{\kappa} \Big[\kappa \{ \sum_{i=1}^{n} (1 - w_{ij}^{(k)}) \cos(X_i - \mu_{j1}^{(k+1)}) + \sum_{i=1}^{n} w_{ij}^{(k)} \cos(X_i - \mu_{j2}^{(k+1)}) \} \\
&\quad -n \log\{I_0(\kappa)\} - \log(\kappa + 1) \Big].
\end{aligned}$$

Compute

$$M_n^{(k+1)}(\alpha_j) = 2\{pl_n(\alpha_j^{(k+1)}, \mu_{j1}^{(k+1)}, \mu_{j2}^{(k+1)}, \kappa_j^{(k+1)}) - pl_n(1/2, \hat{\mu}_0^*, \hat{\mu}_0^*, \hat{\kappa}_0^*)\}.$$

Let $k = k + 1$ and repeat Step 3 until fixed number of iterations in k .

Step 4: Let $j = j + 1$, $k = 0$ and go to Step 1, until $j = J$.

Step 5: Calculate the EM-test statistic

$$EM_n^{(k)} = \max\{M_n^{(k)}(\alpha_j), \quad j = 1, 2, \ldots, J\}.$$

With these preparations, we present the asymptotic results of the MLRT and the EM-test together in the following theorem.

Theorem 5.3.2. *Let $X_1, \ldots, X_n$ be a random sample from the von Mises mixture distribution $(1-\alpha)M(\mu_1, \kappa) + \alpha M(\mu_2, \kappa)$, where $0 < \alpha < 1$, $-\pi \le \mu_1, \mu_2 \le \pi$, $\kappa \ge 0$. Suppose that $p(\alpha)$ is a continuous function such that $p(\alpha) \to -\infty$ as $\alpha \to 0$ or 1 and it attains its maximal value at $\alpha = 0.5$. Then under the null distribution $M(0, \kappa_0)$, as $n \to \infty$,*

(a) $M_n \xrightarrow{d} \frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$;

(b) if one of the initial α values is equal to 0.5, then for any fixed finite k, $EM_n^{(k)} \xrightarrow{d} \frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$.

Intuitively, we should choose large values of J and k to ensure the efficiency of the test. However, our simulation suggests that $EM_n^{(1)}$ with $(\alpha_1, \alpha_2, \alpha_3) = (0.1, 0.3, 0.5)$ captures most power of M_n, which was also observed in Chapter 3. The reason for the necessity of including $\alpha = 0.5$ in the above theorem is the same in Theorem 3.2.2. See comments in Remark 3.2.2.

5.4 Simulations and Applications

Simulation studies were conducted to assess the performance of the proposed testing procedures. Let

$$R_n^* = 2\{\sup_{\mu_1,\mu_2,\kappa} l_n(1/2, \mu_1, \mu_2, \kappa) - l_n(1/2, \hat{\mu}_0, \hat{\mu}_0, \hat{\kappa}_0)\}.$$

Let $EM_n^{(k)}$, M_n and M_n^* denote the EM-test, the MLRT with pl_n defined by (5.5) and the MLRT with pl_n defined by (5.4), respectively. The penalty $p(\alpha)$ for all those tests is chosen to be $p(\alpha) = C^* \log(1 - |1 - 2\alpha|)$. The choice of C^* has some influences on the type I error of the test. Simulation studies are often used to find a suitable range of C^*. In the current testing problem, $C^* = 3$ has been found satisfactory.

Table 5.1: Simulated null rejection rates (%) of the MLRT and the EM-test statistics.

		$\kappa = 2$					$\kappa = 3$				
n	Level	$EM_n^{(0)}$	$EM_n^{(1)}$	M_n	M_n^*	R_n^*	$EM_n^{(0)}$	$EM_n^{(1)}$	M_n	M_n^*	R_n^*
100	10%	10.3	10.4	10.7	13.3	12.0	10.1	10.3	10.7	15.0	13.6
100	5%	5.1	5.3	5.4	7.2	6.3	5.0	5.1	5.5	8.6	7.0
100	1%	1.0	1.2	1.2	1.8	1.4	1.1	1.2	1.3	2.2	1.6
200	10%	9.8	9.8	10.0	11.9	11.2	10.0	10.0	10.4	13.5	12.6
200	5%	5.1	5.1	5.3	6.7	5.9	5.1	5.1	5.4	7.6	6.5
200	1%	1.1	1.2	1.3	1.5	1.2	1.1	1.2	1.2	2.0	1.4

The empirical null distributions of the test statistics were calculated based on 10,000 repetitions for various combination of $n (= 100,\ 200)$ and $\kappa (= 2,\ 3)$. The simulated null rejection rates of the above test statistics are reported in Table 5.1. Clearly, the simulated type I error rates of M_n^* are much larger than nominal values and the problem gets worse with larger κ. Note that increasing C^* should lower the simulated levels. However, even if $C^* = \infty$, at which M_n^* reduces to R_n^*, the simulated type I errors are still larger than the nominal levels. On the other hand, the simulated null rejection rates of M_n and $EM_n^{(k)}$ are very close to the nominal levels and they do not depend much on the sample size nor on the size of the concentration parameter. Thus the penalty function $-\log(\kappa + 1)$ is very important.

Table 5.2: Simulated powers (%) of the MLRT and the EM-test at the 5% level.

		$\kappa = 2$			$\kappa = 3$		
n	$1-\alpha$	$EM_n^{(0)}$	$EM_n^{(1)}$	M_n	$EM_n^{(0)}$	$EM_n^{(1)}$	M_n
100	0.05	5.8	6.1	6.2	24.0	24.7	24.4
100	0.10	11.2	19.7	20.7	70.6	73.0	72.7
100	0.25	80.1	84.0	84.1	99.6	99.9	99.9
200	0.05	7.5	8.0	8.1	47.2	54.8	54.8
200	0.10	37.9	44.8	45.8	97.0	97.5	97.5
200	0.25	96.4	98.6	99.0	100.0	100.0	100.0

To compare the power of the tests, we considered the following alternative models

$$(1-\alpha)M(\pi/2,\kappa) + \alpha M(-\pi/2,\kappa),$$

with $(1-\alpha)(= 0.05, 0.10, 0.25)$ and $\kappa = 2, 3$. Simulated critical values were used for power calculation and the power was calculated based on 10,000 repetitions. The results are in Table 5.2. It is seen that with one iteration and three initial values for α, the test based on $EM_n^{(1)}$ captures most power of M_n.

Now we turn to the analysis of two real data sets in Example 5.1.1. For the Dinosaur National Monument data set, we find $M_n = 26.81$, $EM_n^{(0)} = 26.68$ and $EM_n^{(1)} = 26.77$, which suggest that there is strong evidence to reject uni-component von Mises distribution. For the Dry Mesa Dinosaur Quarry data set, we find $EM_n^{(0)} = EM_n^{(1)} = M_n = 0$, all suggest lack of evidence to reject the uni-component von Mises distribution. In Figure 5.3, we plot the fitted uni-component von Mises density functions and the kernel density functions of the two dinosaur data sets. For the Dinosaur National Monument data set, two functions differ substantially which is consistent to the conclusion of our formal test. For the Dry

Mesa Dinosaur Quarry data set, it appears the uni-component von Mises distribution fits the data very well which explains the insignificant outcome of the formal test.

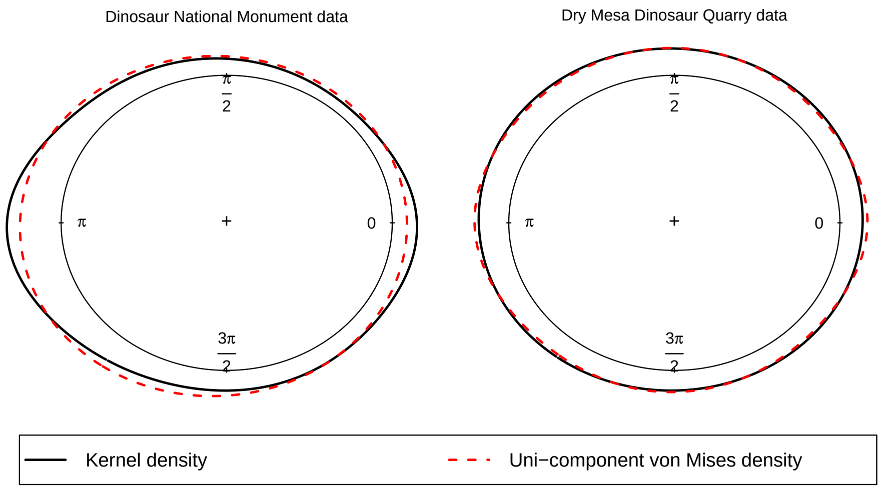

Figure 5.3: Kernel and uni-component von Mises densities.

5.5 Discussion

The von Mises distribution is often regarded as normal distribution for circular data. However, the von Mises mixtures have quite different properties from normal mixtures. Most notably, the von Mises mixture is strongly identifiable. Because of this, the MLRT has quite different asymptotic properties when applied to normal mixtures and von Mises mixture models. For von Mises mixtures, the convergence rates of the MLEs of the mixing distribution and κ are $O_p(n^{-1/4})$ and $O_p(n^{-1/2})$ respectively, and the limiting distribution of the MLRT has a simple form, which are different from the conclusions for normal mixtures (Chen and Chen 2003, Chen and Kalbfleisch 2005). At the same time, when κ goes to infinity, the von Mises distribution converges to a normal distribution and the strong identifiability of the von Mises distribution weakens, which is reflected in (5.13) in the proof of Theorem 5.3.1. Although for each given κ, the ratio in (5.13) remains $o_p(1)$, the asymptotic approximation requires larger n for larger κ.

5.6 Appendix: Technical Proofs

Preliminaries and Notation.

Define

$$\begin{aligned}
U_i(\kappa) &= \frac{1}{\kappa-\kappa_0}\Big\{\frac{f(X_i;0,\kappa)}{f(X_i;0,\kappa_0)}-1\Big\} = \frac{1}{\kappa-\kappa_0}\Big[\frac{I_0(\kappa_0)}{I_0(\kappa)}\exp\{(\kappa-\kappa_0)\cos X_i\}-1\Big],\\
Y_i(\mu,\kappa) &= \frac{1}{\mu}\Big\{\frac{f(X_i;\mu,\kappa)}{f(X_i;0,\kappa_0)}-\frac{f(X_i;0,\kappa)}{f(X_i;0,\kappa_0)}\Big\}\\
&= \frac{I_0(\kappa_0)}{\mu I_0(\kappa)}[\exp\{\kappa\cos(X_i-\mu)-\kappa_0\cos X_i\}-\exp\{(\kappa-\kappa_0)\cos X_i\}],\\
Z_i(\mu) &= \frac{Y_i(\mu,\kappa_0)-Y_i(0,\kappa_0)}{\mu},
\end{aligned}$$

and let $Y_i(0,\kappa)$, $U_i(\kappa_0)$ and $Z_i(0)$ be their continuity limits. For convenience, we put $Y_i(\mu)=Y_i(\mu,\kappa_0)$, $Y_i=Y_i(0)$, $U_i=U_i(\kappa_0)$ and $Z_i=Z_i(0)$. The following proposition assesses the stochastic orders of some relevant stochastic processes.

Proposition 5.6.1. *Indexed by the parameters $\kappa\in[\kappa_0-\delta,\kappa_0+\delta]$ for some $\delta>0$, and $|\mu|\le\pi$, the following processes are tight*

$$\begin{aligned}
U_n^*(\kappa) &= n^{-1/2}\sum\{U_i(\kappa)-U_i\}/(\kappa-\kappa_0),\\
Y_n^*(\mu) &= n^{-1/2}\sum\{Y_i(\mu)-Y_i\}/\mu,\\
Y_n^*(\mu,\kappa) &= n^{-1/2}\sum\{Y_i(\mu,\kappa)-Y_i(\mu)\}/(\kappa-\kappa_0),\\
Z_n^*(\mu) &= n^{-1/2}\sum\{Y_i(\mu)-Y_i-\mu Z_i\}/\mu^2.
\end{aligned}$$

Proof. According to Billingsley (1968, p.95), it suffices to verify the following Lipschitz conditions are satisfied:

$$E\{U_n^*(\kappa_1) - U_n^*(\kappa_2)\}^2 \leq B(\kappa_1 - \kappa_2)^2,$$

$$E\{Y_n^*(\mu_1) - Y_n^*(\mu_2)\}^2 \leq B(\mu_1 - \mu_2)^2,$$

$$E\{Y_n^*(\mu_1, \kappa_1) - Y_n^*(\mu_2, \kappa_2)\}^2 \leq B[(\mu_1 - \mu_2)^2 + (\kappa_1 - \kappa_2)^2],$$

$$E\{Z_n^*(\mu_1) - Z_n^*(\mu_2)\}^2 \leq B(\mu_1 - \mu_2)^2$$

for some constant B. Consider the following functions

$$\frac{U_i(\kappa) - U_i}{\kappa - \kappa_0}, \quad \frac{Y_i(\mu) - Y_i}{\mu}, \quad \frac{Y_i(\mu, \kappa) - Y_i(\mu)}{\kappa - \kappa_0} \text{ and } \frac{Y_i(\mu) - Y_i - \mu Z_i}{\mu^2}.$$

The Lipschitz condition is satisfied if the derivatives of the above functions have bounded second moments uniformly in μ and κ. This is obvious since their second moments are continuous in μ and κ inside a compact parameter space. □

Proof of Lemma 5.3.1 Note that

$$(1-\alpha)\exp\{\kappa\cos(X_i-\mu_1)\}+\alpha\exp\{\kappa\cos(X_i-\mu_2)\} \leq \exp[\kappa\{\max(\cos(X_i-\mu_1), \cos(X_i-\mu_2))\}].$$

Thus we have

$$l_n(\alpha, \mu_1, \mu_2, \kappa) \leq -n\log I_0(\kappa) + \kappa\sum[\max\{\cos(X_i - \mu_1), \cos(X_i - \mu_2)\}].$$

Using (A.4) in Mardia and Jupp (2000, p. 349) we have, as $\kappa \to \infty$,

$$I_0(\kappa) = \frac{e^\kappa}{(2\pi\kappa)^{\frac{1}{2}}}\{1 + o(1)\},$$

hence

$$\begin{aligned} l_n(\alpha, \mu_1, \mu_2, \kappa) &\leq -n\kappa + \frac{n}{2}\log(2\pi\kappa) + \kappa \sum [\max\{\cos(X_i - \mu_1), \cos(X_i - \mu_2)\}] + o_p(1) \\ &= -\kappa \sum [1 - \max\{\cos(X_i - \mu_1), \cos(X_i - \mu_2)\}] + \frac{n}{2}\log(2\pi\kappa) + o_p(1). \end{aligned}$$

Let $S(\mu_1, \mu_2) = E\left[1 - \max\{\cos(X - \mu_1), \cos(X - \mu_2)\}\right]$. Note that $1 - \max\{\cos(X_i - \mu_1), \cos(X_i - \mu_2)\}$ is a uniformly continuous function in $|X_i| \leq \pi$, $|\mu_1| \leq \pi$ and $|\mu_2| \leq \pi$. So the uniform strong law of large numbers implies

$$n^{-1} \sum [1 - \max\{\cos(X_i - \mu_1), \cos(X_i - \mu_2)\}] \to S(\mu_1, \mu_2),$$

almost surely and uniformly in $|\mu_1| \leq \pi$ and $|\mu_2| \leq \pi$ (see Rubin, 1956). For any $|X| \leq \pi$, we have

$$1 - \max\{\cos(X - \mu_1), \cos(X - \mu_2)\} \geq 0,$$

where the equality holds only if $X = \mu_1$ or μ_2, which has zero probability to occur when $0 < \kappa_0 < \infty$. Therefore, under the null distribution $M(0, \kappa_0)$ with $\kappa_0 > 0$, $S(\mu_1, \mu_2)$ is continuous and positive, for all the values of μ_1 and μ_2. Thus,

$$q = \min_{\mu_1, \mu_2} S(\mu_1, \mu_2) > 0.$$

Then with probability approaching one uniformly in α, μ_1, μ_2, and κ,

$$l_n(\alpha, \mu_1, \mu_2, \kappa) \leq -n\{q\kappa - \log(2\pi\kappa)/2\} + o_p(1).$$

Clearly, there exists a $\Delta > 0$ such that when $\kappa > \Delta$, we have $q\kappa - \log(2\pi\kappa)/2 > 0$. Note that $l_n(0, 0, 0, 0) = 0$. The function $l_n(\alpha, \mu_1, \mu_2, \kappa) - l_n(0, 0, 0, 0) < 0$ in probability when $\kappa > \Delta$. This shows that $\lim P(\hat{\kappa} > \Delta) = 0$ for some constant Δ. □

Proof of Theorem 5.3.1 By symmetry, we assume that $0 \leq \alpha \leq 1/2$ instead of $0 \leq \alpha \leq 1$. Let

$$r_n(\alpha, \mu_1, \mu_2, \kappa) = 2\{l_n(\alpha, \mu_1, \mu_2, \kappa) - l_n(0, \hat{\mu}_0, \hat{\mu}_0, \hat{\kappa}_0)\}.$$

Also, let $r_{1n}(\alpha, \mu_1, \mu_2, \kappa) = 2\{l_n(\alpha, \mu_1, \mu_2, \kappa) - l_n(0, 0, 0, \kappa_0)\}$ and $r_{2n} = 2\{l_n(0, 0, 0, \kappa_0) - l_n(0, \hat{\mu}_0, \hat{\mu}_0, \hat{\kappa}_0)\}$. Then the LRT statistic $R_n = r_n(\hat{\alpha}, \hat{\mu}_1, \hat{\mu}_2, \hat{\kappa})$, and $r_n(\alpha, \mu_1, \mu_2, \kappa) = r_{1n}(\alpha, \mu_1, \mu_2, \kappa) + r_{2n}$.

We study R_n through quadratic expansions of r_{1n} and r_{2n}. We work on r_{1n} first. Express

$$r_{1n}(\alpha, \mu_1, \mu_2, \kappa) = 2\sum_{i=1}^{n} \log(1 + \delta_i),$$

where

$$\delta_i = (1-\alpha)\Big\{\frac{f(X_i; \mu_1, \kappa)}{f(X_i; 0, \kappa_0)} - 1\Big\} + \alpha\Big\{\frac{f(X_i; \mu_2, \kappa)}{f(X_i; 0, \kappa_0)} - 1\Big\}.$$

We can also write δ_i as

$$\delta_i = (\kappa - \kappa_0)U_i(\kappa) + (1-\alpha)\mu_1 Y_i(\mu_1, \kappa) + \alpha\mu_2 Y_i(\mu_2, \kappa). \tag{5.6}$$

By Lemma 5.3.2 and the assumption that $0 \leq \alpha \leq 1/2$, under the null distribution, $\hat{\mu}_1 = o_p(1)$ and $\hat{\alpha}\hat{\mu}_2 = o_p(1)$. Hence, for asymptotic consideration, we only need to expand r_{1n} at μ_1 values in an arbitrarily small neighborhood of 0. Expansion of r_{1n} with respect to μ_2 will be done in

$$\Omega_1(\epsilon) = \{|\mu_2| > \epsilon\} \text{ and } \Omega_2(\epsilon) = \{|\mu_2| \leq \epsilon\}$$

for arbitrarily small $\epsilon > 0$, respectively. Let $R_n(\epsilon, I)$ denote the supremum of r_n over $\Omega_1(\epsilon)$ and $R_n(\epsilon, II)$ denote the supremum of r_n over $\Omega_2(\epsilon)$. Then $R_n = \max\{R_n(\epsilon, I), R_n(\epsilon, II)\}$. Since $\hat{\kappa}$ is a consistent estimator of κ_0 as shown in Lemma 5.3.2, we need only expand r_{1n} with respect to κ in $[\kappa_0 - \delta, \kappa_0 + \delta]$ for some arbitrarily small $\delta > 0$.

We first analyze $R_n(\epsilon, I)$. In the region of Ω_1, we expand δ_i as follows

$$\begin{aligned}\delta_i &= (\kappa - \kappa_0)U_i(\kappa_0) + (1-\alpha)\mu_1 Y_i(0, \kappa_0) + \alpha\mu_2 Y_i(\mu_2, \kappa_0) + \epsilon_{in} \\ &= (\kappa - \kappa_0)U_i + (1-\alpha)\mu_1 Y_i + \alpha\mu_2 Y_i(\mu_2) + \epsilon_{in},\end{aligned}$$

where ϵ_{in} is the remainder term. Let $\epsilon_n = \sum_{i=1}^n \epsilon_{in}$. By Proposition 5.6.1, we can show

$$|\epsilon_n| \;\; \leq \;\; n^{1/2}\{(\kappa-\kappa_0)^2+\alpha^2+\mu_1^2\}O_p(1).$$

Since the remainder resulting from the square and cubic sums has at least the order of the remainder from the linear sum, we have

$$\begin{aligned} r_{1n}(\alpha,\mu_1,\mu_2,\kappa) &\leq 2\sum_{i=1}^n\delta_i - \sum_{i=1}^n\delta_i^2 + \frac{2}{3}\sum_{i=1}^n\delta_i^3 \\ = \;\; & 2\sum_{i=1}^n\{(\kappa-\kappa_0)U_i+(1-\alpha)\mu_1Y_i+\alpha\mu_2Y_i(\mu_2)\} \\ & -\sum_{i=1}^n\{(\kappa-\kappa_0)U_i+(1-\alpha)\mu_1Y_i+\alpha\mu_2Y_i(\mu_2)\}^2 \\ & +n^{1/2}\{(\kappa-\kappa_0)^2+\alpha^2+\mu_1^2\}O_p(1)+n\{(\kappa-\kappa_0)^3+\alpha^3+\mu_1^3\}O_p(1). \qquad (5.7) \end{aligned}$$

Note that, under the null distribution $M(0,\kappa_0)$,

$$\begin{aligned} E(U_i^2) &= 1-A(\kappa_0)/\kappa_0-A^2(\kappa_0), \\ E(Y_i^2) &= \kappa_0A(\kappa_0), \\ E\{Y_i(\mu_2)U_i\} &= A(\kappa_0)(\cos\mu_2-1)/\mu_2, \\ E\{Y_i(\mu_2)Y_i\} &= \kappa_0A(\kappa_0)\sin\mu_2/\mu_2. \end{aligned}$$

Let

$$\begin{aligned} V_i(\mu_2) &= \frac{1}{\mu_2}\Big[Y_i(\mu_2)-\frac{E\{Y_i(\mu_2)U_i\}}{E(U_i^2)}U_i-\frac{E\{Y_i(\mu_2)Y_i\}}{E(Y_i^2)}Y_i\Big] \\ &= \frac{1}{\mu_2}\Big[Y_i(\mu_2)-\frac{A(\kappa_0)(\cos\mu_2-1)}{\mu_2\{1-A(\kappa_0)/\kappa_0-A^2(\kappa_0)\}}U_i-\frac{\sin\mu_2}{\mu_2}Y_i\Big] \end{aligned}$$

and $V_i = V_i(0)$ be the continuity limit of $V_i(\mu_2)$. Then

$$(\kappa-\kappa_0)U_i+(1-\alpha)\mu_1Y_i+\alpha\mu_2Y_i(\mu_2) = t_1U_i+t_2Y_i+t_3V_i(\mu_2),$$

where $t_3 = \alpha\mu_2^2$ and

$$\begin{aligned} t_1 &= \kappa - \kappa_0 + \frac{A(\kappa_0)(\cos\mu_2 - 1)}{\mu_2^2\{1 - A(\kappa_0)/\kappa_0 - A^2(\kappa_0)\}} t_3, \\ t_2 &= (1-\alpha)\mu_1 + \frac{\sin\mu_2}{\mu_2^2} t_3. \end{aligned}$$

It is easy to verify that U_i, Y_i and $V_i(\mu_2)$ are mutually orthogonal for all μ_2. We restrict our attention to a small neighborhood of $(t_1, t_2, t_3) = (0, 0, 0)$ as suggested by the consistency results of the MLEs in Lemma 5.3.2. Consequently, we may regard t_1, t_2 and t_3 as $o_p(1)$. We have

$$\begin{aligned} r_{1n}(\alpha, \mu_1, \mu_2, \kappa) &\le 2\sum_{i=1}^n \{t_1 U_i + t_2 Y_i + t_3 V_i(\mu_2)\} \\ &\quad - \sum_{i=1}^n \{t_1^2 U_i^2 + t_2^2 Y_i^2 + t_3^2 V_i^2(\mu_2)\}\{1 + o_p(1)\}. \end{aligned} \tag{5.8}$$

The remainder terms in (5.7) are summarized in the $o_p(1)$ in (5.8). Furthermore, the right-hand side of (5.8) is asymptotically less than or equal to the maximum of the following quadratic function

$$Q(t_1, t_2, t_3) = 2\sum_{i=1}^n \{t_1 U_i + t_2 Y_i + t_3 V_i(\mu_2)\} - \sum_{i=1}^n \{t_1^2 U_i^2 + t_2^2 Y_i^2 + t_3^2 V_i^2(\mu_2)\}.$$

Note that for any fixed $\epsilon < |\mu_2| \le \pi$, $t_3 \ge 0$ and $Q(t_1, t_2, t_3)$ is maximized at $(t_1, t_2, t_3) = (\tilde{t}_1, \tilde{t}_2, \tilde{t}_3)$, where

$$\tilde{t}_1 = \frac{\sum U_i}{\sum U_i^2}, \quad \tilde{t}_2 = \frac{\sum Y_i}{\sum Y_i^2} \quad \text{and} \quad \tilde{t}_3 = \frac{\{\sum V_i(\mu_2)\}^+}{\sum V_i^2(\mu_2)}. \tag{5.9}$$

Thus

$$r_{1n}(\hat{\alpha}, \hat{\mu}_1, \hat{\mu}_2, \hat{\kappa}) \le \frac{\{\sum U_i\}^2}{\sum U_i^2} + \frac{\{\sum Y_i\}^2}{\sum Y_i^2} + \sup_{\epsilon < |\mu_2| \le \pi} \frac{[\{\sum V_i(\mu_2)\}^+]^2}{\sum V_i^2(\mu_2)} + o_p(1). \tag{5.10}$$

On the other hand, the classic analysis gives

$$r_{2n} = 2\{l_n(0,0,0,\kappa_0) - l_n(0,\hat{\mu}_0,\hat{\mu}_0,\kappa_0)\} = -\frac{\{\sum U_i\}^2}{\sum U_i^2} - \frac{\{\sum Y_i\}^2}{\sum Y_i^2} + o_p(1). \tag{5.11}$$

Combining (5.10) and (5.11) yields

$$R_n(\epsilon, I) \leq \sup_{\epsilon<|\mu_2|\leq\pi} \frac{[\{\sum V_i(\mu_2)\}^+]^2}{\sum V_i^2(\mu_2)} + o_p(1).$$

We thus obtained a upper bound for $R_n(\epsilon, I)$. We next show that this upper bound is achievable.

For $\epsilon < |\mu_2| \leq \pi$ fixed, let $\tilde{\alpha}$, $\tilde{\mu}_1$ and $\tilde{\kappa}$ be the solutions for α, μ_1 and κ of (5.9). Then $\tilde{\alpha} = O_p(n^{-1/2})$, $\tilde{\mu}_1 = O_p(n^{-1/2})$ and $\tilde{\kappa} - \kappa_0 = O_p(n^{-1/2})$ uniformly in μ_2. Note that

$$r_{1n}(\tilde{\alpha}, \tilde{\mu}_1, \mu_2, \tilde{\kappa}) = 2\sum_{i=1}^{n}\tilde{\delta}_i - \sum_{i=1}^{n}\tilde{\delta}_i^2(1+\tilde{\eta}_i)^{-2},$$

where $|\tilde{\eta}_i| < |\tilde{\delta}_i|$ and $\tilde{\delta}_i$ is equal to δ_i in (5.6) with $\alpha = \tilde{\alpha}$, $\mu_1 = \tilde{\mu}_1$ and $\kappa = \tilde{\kappa}$. Since $U_i(\kappa)$ and $Y_i(\mu, \kappa)$ are bounded functions for $|X_i| \leq \pi$, $|\mu| \leq \pi$, and $\kappa \in [\kappa_0 - \delta, \kappa_0 + \delta]$, we have $\max_{1\leq i\leq n} |\tilde{\delta}_i| = O_p(n^{-1/2}) = o_p(1)$. It follows that uniformly in $\epsilon < |\mu_2| \leq \pi$,

$$\max_{1\leq i\leq n} |\tilde{\eta}_i| = o_p(1).$$

Then we can easily get

$$r_{1n}(\tilde{\alpha}, \tilde{\mu}_1, \mu_2, \tilde{\kappa}) \;\; = \;\; 2\sum_{i=1}^{n}\tilde{\delta}_i - \{1+o_p(1)\}\sum_{i=1}^{n}\tilde{\delta}_i^2.$$

By (5.9), $\tilde{\alpha}$, $\tilde{\mu}_1$ and $\tilde{\kappa}$ are such that

$$\sup_{\epsilon<|\mu_2|\leq\pi} r_n(\tilde{\alpha}, \tilde{\mu}_1, \mu_2, \tilde{\kappa}) = \sup_{\epsilon<|\mu_2|\leq\pi} \frac{\{[\sum V_i(\mu_2)]^+\}^2}{\sum V_i^2(\mu_2)} + o_p(1).$$

That is,

$$R_n(\epsilon, I) = \sup_{\epsilon<|\mu_2|\leq\pi} \frac{\{[\sum V_i(\mu_2)]^+\}^2}{\sum V_i^2(\mu_2)} + o_p(1). \tag{5.12}$$

This concludes the analysis of $R_n(\epsilon, I)$.

Next, we try to expand $R_n(\epsilon, II)$. Since the MLEs of μ_1 and κ are consistent, in addition to $|\mu_2| \leq \epsilon$, we can restrict μ_1 and κ in the following analysis to the region of $|\mu_1| \leq \epsilon$ and $|\kappa - \kappa_0| \leq \epsilon$, respectively.

In the sequel, $\hat{\alpha}$, $\hat{\mu}_1$, $\hat{\mu}_2$ and $\hat{\kappa}$ denote the MLEs of α, μ_1, μ_2 and κ within the region defined by $0 \leq \alpha \leq 1/2$, $|\mu_1| \leq \epsilon$, $|\mu_2| \leq \epsilon$ and $|\kappa - \kappa_0| \leq \epsilon$. We write

$$\delta_i \quad = \quad (\kappa - \kappa_0)U_i + m_1 Y_i + m_2 Z_i + \epsilon_{in},$$

where ϵ_{in} is the remainder term, $m_1 = (1-\alpha)\mu_1 + \alpha\mu_2$ and $m_2 = (1-\alpha)\mu_1^2 + \alpha\mu_2^2$. Let $\epsilon_n = \sum_{i=1}^n \epsilon_{in}$. By Proposition 5.6.1, we find

$$\begin{aligned}\epsilon_n \quad = \quad & n^{1/2}(\kappa - \kappa_0)^2 O_p(1) + n^{1/2} m_1(\kappa - \kappa_0) O_p(1) \\ & + n^{1/2}(1-\alpha)\mu_1^3 O_p(1) + n^{1/2}\alpha\mu_2^3 O_p(1).\end{aligned}$$

Using the facts $|2x| \leq 1 + x^2$ for any x and $|\mu_2| \leq \epsilon$ and $|\kappa - \kappa_0| \leq \epsilon$, we have

$$|\epsilon_n| \quad \leq \quad n\epsilon O_p(1)\{(\kappa - \kappa_0)^2 + m_1^2 + m_2^2\} + \epsilon O_p(1).$$

Note that $n^{-1}\sum_{i=1}^n \{(\kappa - \kappa_0)U_i + m_1 Y_i + m_2 Z_i\}^2$ converges to a positive definite quadratic form in $\kappa - \kappa_0$, m_1 and m_2. Thus

$$\frac{\sum_{i=1}^n |(\kappa - \kappa_0)U_i + m_1 Y_i + m_2 Z_i|^3}{\sum_{i=1}^n \{(\kappa - \kappa_0)U_i + m_1 Y_i + m_2 Z_i\}^2} \leq (|\kappa - \kappa_0| + |m_1| + |m_2|) O_p(1) \leq \epsilon O_p(1). \quad (5.13)$$

Using a few similar techniques employed for $R_n(\epsilon, I)$, we have

$$\begin{aligned} r_{1n}(\alpha, \mu_1, \mu_2, \kappa) \quad \leq \quad & 2\sum_{i=1}^n \{(\kappa - \kappa_0)U_i + m_1 Y_i + m_2 Z_i\} \\ & - \sum_{i=1}^n \{(\kappa - \kappa_0)U_i + m_1 Y_i + m_2 Z_i\}^2 \{1 + \epsilon O_p(1)\} + \epsilon O_p(1).(5.14)\end{aligned}$$

We now conduct the orthogonal transformation as follows

$$(\kappa-\kappa_0)U_i+m_1Y_i+m_2Z_i=s_1U_i+m_1Y_i+m_2V_i,$$

where

$$s_1=\kappa-\kappa_0-\frac{A(\kappa_0)}{2\{1-A(\kappa_0)/\kappa_0-A^2(\kappa_0)\}}m_2.$$

Thus, (5.14) becomes

$$\begin{aligned} r_{1n}(\alpha,\mu_1,\mu_2,\kappa) &\leq 2\sum_{i=1}^{n}\{s_1U_i+m_1Y_i+m_2V_i\} \\ &\quad -\sum_{i=1}^{n}\{s_1U_i+m_1Y_i+m_2V_i\}^2\{1+\epsilon O_p(1)\}+\epsilon O_p(1) \\ &= 2\sum_{i=1}^{n}\{s_1U_i+m_1Y_i+m_2V_i\} \\ &\quad -\sum_{i=1}^{n}\{s_1^2U_i^2+m_1^2Y_i^2+m_2^2V_i^2\}\{1+\epsilon O_p(1)\}+\epsilon O_p(1). \end{aligned}$$

According to the same technique leading to (5.10), we get

$$r_{1n}(\hat{\alpha},\hat{\mu}_1,\hat{\mu}_2,\hat{\kappa})\leq\{1+\epsilon O_p(1)\}^{-1}\left[\frac{\{\sum U_i\}^2}{\sum U_i^2}+\frac{\{\sum Y_i\}^2}{\sum Y_i^2}+\frac{[\{\sum V_i\}^+]^2}{\sum V_i^2}\right]+\epsilon O_p(1).$$

Recall

$$r_{2n}=2\{l_n(0,0,0,\kappa_0)-l_n(0,\hat{\mu}_0,\hat{\mu}_0,\kappa_0)\}=-\frac{\{\sum U_i\}^2}{\sum U_i^2}-\frac{\{\sum Y_i\}^2}{\sum Y_i^2}+o_p(1).$$

Then,

$$r_n(\hat{\alpha},\hat{\mu}_1,\hat{\mu}_2,\hat{\kappa})\leq\frac{\epsilon O_p(1)}{1+\epsilon O_p(1)}\left[\frac{\{\sum U_i\}^2}{\sum U_i^2}+\frac{\{\sum Y_i\}^2}{\sum Y_i^2}\right]+\frac{[\{\sum V_i\}^+]^2}{\{1+\epsilon O_p(1)\}\sum V_i^2}+\epsilon O_p(1).$$

Therefore

$$R_n(\epsilon,II)\leq\frac{[\{\sum V_i\}^+]^2}{\sum V_i^2}+\epsilon O_p(1). \tag{5.15}$$

Next, let $\tilde{\alpha} = 1/2$, and $\tilde{\mu}_1$, $\tilde{\mu}_2$ and $\tilde{\kappa}$ be determined by

$$\tilde{s}_1 = \frac{\sum U_i}{\sum U_i^2}, \quad \tilde{m}_1 = \frac{\sum Y_i}{\sum Y_i^2} \quad \text{and} \quad \tilde{m}_2 = \frac{\{\sum V_i\}^+}{\sum V_i^2}. \tag{5.16}$$

It is easy to see that

$$r_n(\tilde{\alpha}, \tilde{\mu}_1, \tilde{\mu}_2, \tilde{\kappa}) = \frac{[\{\sum V_i\}^+]^2}{\sum V_i^2} + o_p(1) \tag{5.17}$$

and hence

$$R_n(\epsilon, II) \geq \frac{[\{\sum V_i\}^+]^2}{\sum V_i^2} + o_p(1). \tag{5.18}$$

For any $\epsilon > 0$, $R_n = \max\{R_n(\epsilon, I), R_n(\epsilon, II)\}$. Combining (5.12), (5.15) and (5.18), we have

$$R_n \leq \max\Big\{\frac{[\{\sum V_i\}^+]^2}{\sum V_i^2}, \sup_{\epsilon < |\mu_2| \leq \pi} \frac{\{[\sum V_i(\mu_2)]^+\}^2}{\sum V_i^2(\mu_2)} + \epsilon O_p(1)\Big\} + o_p(1)$$

and

$$R_n \geq \max\Big\{\frac{[\{\sum V_i\}^+]^2}{\sum V_i^2}, \sup_{\epsilon < |\mu_2| \leq \pi} \frac{\{[\sum V_i(\mu_2)]^+\}^2}{\sum V_i^2(\mu_2)}\Big\} + o_p(1).$$

By the uniform strong law of large numbers and the tightness of the process of $Y_n^*(\mu)$, the process

$$\{\sum V_i^2(\mu)\}^{-1/2} \sum V_i(\mu), \ |\mu| \leq \pi$$

converges weakly to a Gaussian process $\zeta(\mu)$ with mean 0, standard deviation 1 and the autocorrelation function $\rho(s,t)$ which is given by

$$\rho(s,t) = \frac{g(s,t)}{\{g(s,s)g(t,t)\}^{\frac{1}{2}}}, \qquad \text{for } s, t \neq 0,$$

where $g(s,t) = E\{V_1(s)V_1(t)\}$. By letting $n \to \infty$ and then $\epsilon \to 0$, we conclude that R_n converges in probability to $\sup_{|\mu| \leq \pi}\{\zeta^+(\mu)\}^2$. The only thing left is to calculate the function $g(s,t)$. The result in (5.3) follows by some tedious but simple calculations. □

In order to prove Theorem 5.3.2, we need the following lemma which states the consistency property of the modified MLEs.

Lemma 5.6.1. *Let* $X_1, \ldots, X_n$ *be a random sample from the mixture population* $(1-\alpha)M(\mu_1,\kappa)+\alpha M(\mu_2,\kappa)$. *Under the null distribution* $M(0,\kappa_0)$,

(a) $\hat{\mu}_0^* = o_p(1)$, $\hat{\kappa}_0^* - \kappa_0 = o_p(1)$ *and*

$$pl_n(1/2,\hat{\mu}_0^*,\hat{\mu}_0^*,\hat{\kappa}_0^*) - pl_n(1/2,0,0,\kappa_0) \quad = \quad \frac{\{\sum U_i\}^2}{\sum U_i^2} + \frac{\{\sum Y_i\}^2}{\sum Y_i^2} + o_p(1); \quad (5.19)$$

(b) $\hat{\mu}_1^* = o_p(1)$, $\hat{\mu}_2^* = o_p(1)$ *and* $\hat{\kappa}^* - \kappa_0 = o_p(1)$.

Proof. (a) Note that $\hat{\mu}_0^*$ and $\hat{\kappa}_0^*$ are the modified MLEs of μ and κ under the null model, hence the consistency of $\hat{\mu}_0^*$ and $\hat{\kappa}_0^*$ follows from the classic theory. Using this consistency result, we have

$$pl_n(1/2,\hat{\mu}_0^*,\hat{\mu}_0^*,\hat{\kappa}_0^*) - pl_n(1/2,0,0,\kappa_0) = 2\{\sum_{i=1}^n \log f(X_i;\hat{\mu}_0^*,\hat{\kappa}_0^*) - \sum_{i=1}^n \log f(X_i;0,\kappa_0)\} + o_p(1).$$

Thus, the proof of (5.19) for the modified likelihood reduces to the proof of the classical result for the usual LRT. Then, the result follows.

(b) Firstly, we prove that the modified MLE of α is bounded away from 0 or 1 in probability. Note that

$$\begin{aligned}
0 \le M_n &= 2\{pl_n(\hat{\alpha}^*,\hat{\mu}_1^*,\hat{\mu}_2^*,\hat{\kappa}^*) - pl_n(1/2,\hat{\mu}_0^*,\hat{\mu}_0^*,\hat{\kappa}_0^*)\}\\
&\le 2\{pl_n(\hat{\alpha}^*,\hat{\mu}_1^*,\hat{\mu}_2^*,\hat{\kappa}^*) - pl_n(1/2,\hat{\mu}_0,\hat{\mu}_0,\hat{\kappa}_0)\}\\
&= 2\{l_n(\hat{\alpha}^*,\hat{\mu}_1^*,\hat{\mu}_2^*,\hat{\kappa}^*) - \log(\hat{\kappa}^*+1) + p(\hat{\alpha}^*)\\
&\quad -l_n(1/2,\hat{\mu}_0,\hat{\mu}_0,\hat{\kappa}_0) + \log(\hat{\kappa}_0+1) - p(0.5)\}\\
&\le R_n + 2\{p(\hat{\alpha}^*) - p(0.5)\} + 2\log(\hat{\kappa}_0+1)\\
&= R_n + 2\{p(\hat{\alpha}^*) - p(0.5)\} + 2\log(\kappa_0+1) + o_p(1).
\end{aligned}$$

The last step uses the consistency of $\hat{\kappa}_0$. By Theorem 5.3.1, $R_n = O_p(1)$, which implies $p(\hat{\alpha}^*) - p(0.5) = O_p(1)$. Hence there exists $\epsilon_0 > 0$, such that $P(\epsilon_0 \le \hat{\alpha}^* \le 1-\epsilon_0) \to 1$

as $n \to \infty$. Hence, the problem reduces to the consistency of the modified MLEs in a compact and identifiable parameter space. Consequently, the consistency of the modified MLEs follows. □

Proof of Theorem 5.3.2 (a) By (5.11), (5.19) and the consistency of $\hat{\kappa}^*$, we have

$$\begin{aligned} M_n &= r_{1n}(\hat{\alpha}^*, \hat{\mu}_1^*, \hat{\mu}_2^*, \hat{\kappa}^*) + r_{2n} + 2\{p(\hat{\alpha}^*) - p(0.5)\} + o_p(1) \\ &= r_n(\hat{\alpha}^*, \hat{\mu}_1^*, \hat{\mu}_2^*, \hat{\kappa}^*) + 2\{p(\hat{\alpha}^*) - p(0.5)\} + o_p(1) \\ &\leq r_n(\hat{\alpha}^*, \hat{\mu}_1^*, \hat{\mu}_2^*, \hat{\kappa}^*) + o_p(1). \end{aligned} \tag{5.20}$$

Since $(\hat{\mu}_1^*, \hat{\mu}_2^*, \hat{\kappa}^*)$ are consistent estimators of $(0, 0, \kappa_0)$, $R_n(\epsilon, II)$ can serve as an upper bound for $r_n(\hat{\alpha}^*, \hat{\mu}_1^*, \hat{\mu}_2^*, \hat{\kappa}^*)$. Combining (5.18) and (5.20), we have

$$M_n \leq \frac{\{[\sum V_i]^+\}^2}{\sum V_i^2} + o_p(1).$$

We take $\tilde{\mu}_1$, $\tilde{\mu}_2$ and $\tilde{\kappa}$ as determined by (5.16) when $\tilde{\alpha} = 1/2$. Then $\tilde{\kappa} - \kappa_0 = o_p(1)$ and so

$$\begin{aligned} M_n &\geq 2\{pl_n(\tilde{\alpha}, \tilde{\mu}_1, \tilde{\mu}_2, \tilde{\kappa}) - pl_n(1/2, \hat{\mu}_0^*, \hat{\mu}_0^*, \hat{\kappa}_0^*)\} \\ &= r_n(\tilde{\alpha}, \tilde{\mu}_1, \tilde{\mu}_2, \tilde{\kappa}) + o_p(1) \\ &= \frac{\{[\sum V_i]^+\}^2}{\sum V_i^2} + o_p(1). \end{aligned}$$

The result in the last step follows from (5.17). Combining the above results, we have

$$M_n = \frac{\{[\sum V_i]^+\}^2}{\sum V_i^2} + o_p(1).$$

Consequently, the limiting distribution of M_n is given by $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$.

(b) Obviously,

$$EM_n^{(k)} \leq M_n \leq \frac{\{[\sum V_i]^+\}^2}{\sum V_i^2} + o_p(1).$$

If one of α_j's is equal to 0.5,

$$EM_n^{(k)} \geq 2\{pl_n(0.5, \tilde{\mu}_1, \tilde{\mu}_2, \tilde{\kappa}) - pl_n(1/2, \hat{\mu}_0^*, \hat{\mu}_0^*, \hat{\kappa}_0^*)\} = \frac{\{[\sum V_i]^+\}^2}{\sum V_i^2} + o_p(1).$$

Hence

$$EM_n^{(k)} = \frac{\{[\sum V_i]^+\}^2}{\sum V_i^2} + o_p(1).$$

Consequently, the limiting distribution of $EM_n^{(k)}$ is given by $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$. □

Chapter 6

Future Work

In this chapter, we will first provide a summary of what has achieved in the thesis. The techniques employed in this thesis and the new method proposed are seen to be applicable to much more general mixture models. We also outline a number of future research problems.

6.1 Summary of the Current Achievements

In this thesis, we have considered a number of hypothesis testing problems in finite mixture models. More specifically, we study the homogeneity test in two-component mixture models.

In Chapter 2, we discussed the problem of choosing a more effective penalty function for the MLRT. The MLRT with the new penalty enjoys a significant improvement on the power for testing homogeneity when the true mixing proportion is close to 0 and 1. The simulation studies suggested that the simulated null rejection rates are very close to the theoretical values by setting the level of modification to be 1.

In Chapter 3, motivated from designing a homogeneity test procedure for geometric mixture models, exponential mixture models and mixture models on the scale parameter, we proposed a new class of methods (EM-test) for testing homogeneity in mixture models with two components. The EM-test combines and exceeds the advantages of the score test (Liang and Rathouz, 1999) and the MLRT (Chen, 1998, Chen et al. 2001, 2004). We find that the EM-test has a simple null distribution, is more efficient, and has broader applications than the MLRT and other methods. The EM-test also compares favorably to the D-test and the constrained LRT.

In Chapter 4, we considered the use of the EM-test to the test of homogeneity in normal mixture models. For the test of homogeneity at the presence of the structural parameter, a penalty function on the variance parameter is suggested to overcome the under-estimation effect. The limiting distribution is a simple function of $0.5\chi_0^2 + 0.5\chi_1^2$ and χ_1^2 distributions. The test with this limiting distribution is still very convenient to implement in applications. For normal mixture models in both mean and variance parameters, the penalty functions on the component variance parameters are added to the log-likelihood function to avoid the unboundedness of the log-likelihood function. The limiting distribution of the EM-test is shown to be χ_2^2. The simulation results review the good fitting of the limiting distributions to the finite sample distributions of the EM-test statistics.

In Chapter 5, we applied both the MLRT and the EM-test to test of homogeneity in the mixture of circular distributions, especially, the mixture of von Mises distributions with unknown but equal concentration parameters. A new penalty function on κ is suggested to overcome the over-estimation effect of κ. The MLRT and the EM-test are applied to test of homogeneity. Simulation studies suggested that the EM-test based three initial values of α, (0.1, 0.3, 0.5), and one iteration, and the MLRT have comparable power. The EM-test has null rejection rates very close to nominal values. Two real data sets in Grimshaw et al. (2001) are used to illustrate the idea of two tests.

In the next a few sections, we present some research problems which are the natural extension to the development of what have been achieved so far.

6.2 Homogeneity Test in Mixture Models with Multi-dimensional Parameters

In Section 4.3, we discussed the application of the EM-test to test of homogeneity under normal mixture models in both mean and variance parameters. The EM-test has been proved to have a simple limiting distribution in this case. The normal mixture model in both mean and variance parameters is a special case of mixture models with multi-dimensional parameters, namely, the density function is given by

$$(1-\alpha)f(x;\theta_1)+\alpha f(x;\theta_2), \tag{6.1}$$

where $\theta_1, \theta_2 \in \Theta \subset \mathbf{R}^d$, $d \geq 2$. Dacunha-Castelle and Gassiat (1999) and Liu and Shao (2003) found that the limiting distribution of the LRT for testing of homogeneity in mixture models with multi-dimensional parameters is much more complicated than the univariate case even if the mixture density in (6.1) is strongly identifiable and the kernel function $f(x;\mu,\kappa)$ satisfies the finite Fisher information condition. Due to the nice properties of the EM-test discussed before, one of our future research directions is to extend and apply the idea of the EM-test to test of homogeneity in mixture models with multi-dimensional parameters.

One possible way for extension is to follow the pseudo code described in Section 3.2.1 step by step to calculate the EM-test statistics. In Section 4.3, we followed this idea and applied the EM-test to normal mixture model in both mean and variance parameters. The application was proved to be successful. The EM-test enjoys a simple limiting distribution.

When this idea is applied to mixture of general distributions, such as mixture of gamma distributions and mixture of multivariate normal distributions, the limiting distribution of the EM-test may still be tractable, but not as neat as in normal mixture models. The main reason is that we may have several cross terms in the asymptotic expansion of the expression of the EM-test statistic.

Another way for extension is called “one dimension at one time”. Note that for calculating the EM-test statistics, the crucial step is to first find an initial value for $(\alpha, \theta_1, \theta_2)$ for the EM-iterations, which is the Step 2 of the pseudo code described in Section 3.2.1. The idea of “one dimension at one time” differs from the original one on how to choose the initial value for $(\alpha, \theta_1, \theta_2)$. We suggest only allowing one dimension can have different values on two components when we find the initial value for $(\alpha, \theta_1, \theta_2)$. By allowing the first dimension can have different values on two components in Step 2 of the pseudo code in Section 3.2.1, we can calculate an EM-test statistic $EM_{n1}^{(k)}$. Similarly, we have $EM_{n2}^{(k)}, \ldots, EM_{nd}^{(k)}$. The final test statistic is defined to be

$$EM_n^{(k)} = \max\{EM_{nh}^{(k)}, h = 1, \ldots, d\}.$$

Since we only allow one dimension can have different parameters on two components at the first stage, other parameters can be treated as the structural parameters. Our application experiences in the normal mixture model and the von Mises mixture models at the presence of structural parameter suggest that $EM_{nh}^{(0)}$ may have a simple χ^2-type limiting distribution. Further, intuitively, finite number of iterations will not change the values of $(\alpha, \theta_1, \theta_2)$ too much under null hypothesis, we expect $EM_{nh}^{(k)}$, $h = 1, \ldots, d$, will still enjoys a simple χ^2-type limiting distribution. Then the second type of EM-test is expected to have a simple limiting distribution, the maximum of some χ^2-type distributions.

In our future research, we will give a comprehensive comparison of these two ideas from the theoretical and practical issues.

6.3 Testing the Order in Finite Mixture Models

Apart from the homogeneity test, testing the order of the finite mixture model is also an interesting, challenging and more general problem.

In Chen et al (2004), the MLRT is applied to test

$$H_0 : m = 2 \text{ versus } H_a : m > 2.$$

The asymptotic null distribution for the MLRT test statistic was shown to be a mixture of χ^2_0, χ^2_1 and χ^2_2. A more general problem is to test

$$H_0 : m = q_0 \text{ versus } H_a : m = p > q_0,$$

or more specifically,

$$H_0 : m = q_0 \text{ versus } H_a : m = q_0 + 1, \tag{6.2}$$

Dacunha-Castelle and Gassiat (1999) and Liu and Shao (2003) investigated the use of the LRT. They found that the limiting distributions for the LRT are in general related to the supremum of some the Gaussian processes. Not only it is hard to determining the quantiles of the supremum of a general Gaussian process, but the structure of the Gaussian process in these results is also very complex. The application of the MLRT has so far only met with limited success such as the one in Chen et al. (2004). We envisage that the EM-test idea could be more effective in developing convenient statistical procedures. We have started working on this problem, yet due to the nature of the problem, it is likely a long term effort.

6.4 Data Adaptive Level of Penalty of the MLRT and the EM-test

The choice of penalty function, particularly the choice of the level of modification, is a crucial step in the implementation of the MLRT or the EM-test. A new class of penalty functions was introduced in (2.7) and it was found to have some superior properties. At this stage, the level of the modification was mostly determined by simulation studies. It appears that the effectiveness of the MLRT and the EM-test are not greatly affected by the level of modification. Yet it would be ideal if a data-driven procedure with some theoretical justification can be found. Continued effort on this research problem is part of my future research plan.

Bibliography

[1] Abramowitz, M. and Stegun, I. A. (1965). *Handbook of Mathematical Functions.* Dover, New York.

[2] Adler, R. J. (1990). *An Introduction to Continuity, Extrema, and Related Topics for General Gaussian Processes. IMS Lecture Notes - Monograph Series*, **12**. Institute of Mathematical Statistics, Hayward.

[3] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *In Second International Symposium on Information Theory*, B. N. Petrov and F. Csaki (Eds). Budapest: Akadémiai Kiadó, pp. 261-281.

[4] Anaya-Izquierdo K. A. and Marriott P. (2007a). Local mixture models of exponential families. *Bernoulli*, **13**, 623-640.

[5] Anaya-Izquierdo K. A. and Marriott P. (2007b). Local mixtures of the exponential distribution. *Annals of the Institute of Statistical Mathematics*, **59**, 111-134.

[6] Barndorff-Nielsen, O. (1965). Identifiability of mixtures of exponential families. *Journal of Mathematical Analysis and Applications*, **12**, 115-121.

[7] Batschelet, E. (1981). *Circular Statistics in Biology*. London: Academic Press.

[8] Bickel, P. and Chernoff, H. (1993). Asymptotic distribution of the likelihood ratio statistic in a prototypical non regular problem, in *Ghosh, J.K. (Ed.), Statistics and Probability*, Wiley Eastern Limited. pp. 83-96.

[9] Billingsley, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

[10] Böhning, D. (1999). *Computer-Assisted Analysis of Mixtures and Applications: Meta Analysis, Disease Mapping and Others*. New York: Chapman and Hall/CRC.

[11] Carlin, B. P. and Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B*, **57**, 473-484.

[12] Chandra, S. (1977). On the mixtures of probability distributions. *Scandinavian Journal of Statistics: Theory and Applications*, **39**, 105-112.

[13] Charnigo, R. and Sun J. (2004). Testing homogeneity in a mixture distribution via the L^2-distance between competing models. *Journal of the American Statistical Association*, **99**, 488-498.

[14] Chen, H. and Chen, J. (2001). The likelihood ratio test for homogeneity in the finite mixture models. *Canadian Journal of Statistics*, **29**, 201-215.

[15] Chen, H. and Chen, J. (2003). Tests for homogeneity in normal mixtures with presence of a structural parameter. *Statistica Sinica*, **13**, 351-365.

[16] Chen, H., Chen, J. and Kalbfleisch, J. D. (2000). A modified likelihood ratio test for homogeneity in the finite mixture models. Working Paper 2000-01, Department of Statistics and Actuarial Science, University of Waterloo.

[17] Chen, H., Chen, J. and Kalbfleisch, J. D. (2001). A modified likelihood ratio test for homogeneity in finite mixture models. *Journal of the Royal Statistical Society, Series B*, **63**, no. 1, 19-29.

[18] Chen, H., Chen, J. and Kalbfleisch, J. D. (2004). Testing for a finite mixture model with two components. *Journal of the Royal Statistical Society, Series B*, **66**, no. 1, 95-115.

[19] Chen, J. (1995). Optimal rate of convergence in finite mixture models. *The Annals of Statistics*, **23**, 221-234.

[20] Chen, J. (1998). Penalized likelihood ratio test for finite mixture models with multinomial observations. *Canadian Journal of Statistics*, **26**, 583-599.

[21] Chen, J. and Cheng, P. (1995). The limit distribution of the restricted likelihood ratio statistic for finite mixture models. *Northeastern Mathematical Journal* **11**, 365-374.

[22] Chen, J. and Kalbfleisch, J. D. (1996). Penalized minimum-distance estimates in finite mixture models. *Canadian Journal of Statistics*, **24**,167-175.

[23] Chen, J. and Kalbfleisch, J. D. (2005). Modified likelihood ratio test in finite mixture models with a structural parameter. *Journal of Statistical Planning and Inference*, **129**, 93-107.

[24] Chen, J. and Khalili, A. (2006). Order selection in finite mixture models. Working Paper 2006-03, Department of Statistics and Actuarial Science, University of Waterloo.

[25] Chen, J., Li, P. and Fu, Y. (2007). Testing homogeneity in a mixture of von Mises distributions with a structural parameter. *Canaidian Journal of Statistics. To appear.*

[26] Chen, J., Tan, X. and Zhang, R. (2007). Inference for normal mixtures in mean and variance. *Statistica Sincia*. To appear.

[27] Chernoff, H. (1954). On the distribution of the likelihood ratio. *Annals of Mathematical Statistics*, **25**, 573-578.

[28] Cox, D. R. and Lewis, P. A. W. (1968). *The Statistical Analysis of Series of Events*. Chapman & Hall, London.

[29] Craigmile, P. F. and Titterington, D. M. (1998). Parameter estimation for finite mixtures of uniform distributions. *Communications in Statistics–Theory and Methods*, **26**, 1981-1995.

[30] Critchley, F. and Marriott, P. (2004). Data-informed influence analysis. *Biometrika*, **91**, 125-140.

[31] Dacunha-Castelle, D. and Gassiat, E. (1999). Testing the order of a model using locally conic parametrization: population mixtures and stationary ARMA processes. *The Annals of Statistics*, **27**, 1178-1209.

[32] Davies R. B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, **64**, 247-254.

[33] Dempster, A. P., Laird, N. M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, **39**, 1-38.

[34] Feng, Z. D. anc McCulloch, C. E. (1994). On the likelihood ratio test statistic for the number of components in a normal mixture with unequal variances. *Biometrics*, **50**, 1158-1162.

[35] Fraser, M. D., Hsu, Y. S. and Walker, J. J. (1981). Identifiability of finite mixtures of von Mises distributions. *The Annals of Statistics*, **9**, 1130-1131.

[36] Fisher, N. I. (1993). *Statistical Analysis of Circular Data*. Cambridge: Cambridge University Press.

[37] Fu, Y., Chen, J. and Li, P. (2007). Modified likelihood ratio test for homogeneity in a mixture of von Mises distributions. *Journal of Statistical Planning and Inference. To appear.*

[38] Furman, W. D. and Lindsay, B. G. (1994a). Testing for the number of components in a mixture of normal distributions using moment estimators. *Computational Statistics and Data Analysis*, **17**, 473-492.

[39] Furman, W. D. and Lindsay, B. G. (1994b). Measuring the effectiveness of moment estimators as starting values in maximizing mixture likelihoods. *Computational Statistics and Data Analysis*, **17**, 493-507.

[40] Frühwirth-Schnatter, S. (2004). Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techiniques. *The Econometrics Journal*, **7**, 143-167.

[41] Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer Series in Statistics. New York/Berlin/Heidelberg: Springer.

[42] Gelfand, A. E. and Smith A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398-409.

[43] Ghosh, J. K. and Sen, P. K. (1985). On the asymptotic performance of the log-likelihood ratio statistic for the mixture model and related results, in *Proc. Berkely*

Conf. in Honor of J. Neyman and Kiefer, Volume 2, eds L. LeCam and R. A. Olshen, 789-806.

[44] Gruet, M.-A., Philippe, A. and Robert, C. P. (1999). MCMC control spreadsheets for exponential mixture estimation. *Journal of Computational and Graphical Statistics*, **8**, 298-317.

[45] Grimshaw, S. D., Whiting, D. G. and Morris, T. H. (2001). Likelihood ratio tests for a mixture of two von Mises distributions. *Biometrics*, **57**, 260-265.

[46] Hall, P. (1992). *The Bootstrap and Edgeworth Expansion.* Springer Series in Statistics. New York/Berlin/Heidelberg: Springer.

[47] Hartigan, J. A. (1985). A failure of likelihood asymptotics for normal mixtures, in *Proc. Berkely Conf. in Honor of J. Neyman and Kiefer, Volume 2*, eds L. LeCam and R. A. Olshen, 807-810.

[48] Hathaway, R. J. (1985). A constrained formulation of maximun-likelihood estimation for normal mixture distributions. *The Annals of Statistics*, **13**, 795-800.

[49] Heckman, J. J, Robb, R. and Walker, J. R. (1990). Testing the mixture of exponentials hypothesis and estimating the mixing distribution by the methods of moments. *Journal of the American Statistical Association*, **85**, 582-589

[50] Holzmann, H., Munk, A. and Stratmann, B. (2004). Identifiability of finite mixtures-with applications to circular distributions. *Sankhya*, **66**, 440-450.

[51] Hsu, Y. S., Walker, J. J. and Ogren, D. E. (1986). A step-wise method for determining the number of component distribution in a mixture. *Mathematical Geology*, **18**, 153-160.

[52] Ishiguro, M., Sakamoto, Y. and Kitagawa, G. (1997). Bootstrapping log-likelihood and EIC: an extension of AIC. *Annals of the Institute of Statistical Mathematics*. **49**, 411-434.

[53] Ishwaran, H., James, L. F. and Sun, J. (2001). Bayesian model selection in finite mixtures by marginal density decompositions. *Journal of the American Statistical Association*, **96**, 1316-1332.

[54] James, L. F., Priebe, C. E. and Marchette, D. J. (2001). Consistent estimation of mixture complexity. *The Annals of Statistics*, **29**, 1281-1296.

[55] Jammalamadaka, S. R. and SenGupta, A. (2001). *Topics in Circular Statistics*. World Scientific Publishing, Co.

[56] Jewell, N. P. (1982). Mixtures of exponential dsitributions. *Applied Statistics*, **10**, 479-484.

[57] Keribin, C. K. (2000). Consistent estimation of the order of mixture models. *Sankhya*, **62**, 49C62.

[58] Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimates in the presence of infinitely many incidental paramters. *Annals of Mathematical Statistics*, **27**, 887-906.

[59] Kim, P. T. and Koo, J. Y. (2000). Directional mixture models and optimal estimation of the mixing density. *Canadian Journal of Statistics*, **28**, 383-398.

[60] Kullback, S. and Leibler, R. A. (1951). On the information and sufficiency. *Annals of Mathematical Statistics*, **22**, **22**, 79-86.

[61] Laird, N. M. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, **73**, 805-811.

[62] Lemdani, M. and Pons, O. (1995). Tests for genetic linkage and homogeneity. *Biometrics* **51**, 1033-1041.

[63] Leroux, B. (1992). Consistent estimation of a mixture distribution. *The Annals of Statistics*, **20**, 1350-1360.

[64] Liang, K. Y. and Rathouz, P. J. (1999). Hypothesis testing under mixture models: application to genetic linkage analysis. *Biometics*, **55**, 65-74.

[65] Lindsay, B. G. (1983). The geometry of likelihoods: a general theory. *The Annals of Statistics*, **11**, 86-94.

[66] Lindsay, B. G. (1989a). On the determinants of moment matrices. *The Annals of Statistics*, **17**, 711-721.

[67] Lindsay, B. G. (1989b). Moment matrices: applications in mixtures. *The Annals of Statistics*, **17**, 722-740.

[68] Lindsay, B. G. (1995). *Mixture Models: Theory, Geometry and Applications.* Hayward: Institute for mathematical Statistics.

[69] Lindsay, B. G. and Basak, P. (1993). Multivariate normal mixtures: a fast, consistent method of moments. *Journal of the American Statistical Association*, **86**, 96-107.

[70] Lindsay, B. G. and Roeder, K. (1992a). Residual diagnostics i n the mixture model. *Journal of the American Statistical Association*, **87**, 785-795.

[71] Lindsay, B. G. and Roeder, K. (1992b). Uniqueness of estimation and identifiability in mixture models. *Canadian Journal of Statistics*, **21**, 139-147.

[72] Liu, D., Peddada, S. D., Li, L. and Weinberg, C. R. (2006). Phase analysis of circadian-related genes in two tissues. *BMC Bioinformatics*, **7**:87.

[73] Liu, X. and Shao, Y. Z. (2003). Asymptotics for likelihood ratio tests under loss of identifiability. *The Annals of Statistics*, **31**, 807-832.

[74] Liu, X. and Shao, Y. Z. (2004). Asymptotics for the likelihood ratio test in a two-component normal mixture model. *Journal of Statistical Planning and Inference*, **123**, 61-81.

[75] Mardia, K. V. and Jupp, P. E. (2000). *Directional Statistics.* John Wiley and Sons.

[76] Marriott, P. (2002). On the local geometry of mixture models. *Biometrika*, **89**, 77-93.

[77] Marriott, P. (2003). On the geometry of measurement error models. *Biometrika*, **90**, 567-576.

[78] Marriott, P. (2007). Extending local mixture models. *Annals of the Institute of Statistical Mathematics*, **59**, 95-110.

[79] McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test statistics for the number of components in a normal mixture. *Applied Statistics*, **36**, 318-324.

[80] McLachlan, G. J., Bean, R. W., and Peel, D. (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*,**18**, 413-422.

[81] McLachlan, G. J. and Krishnan, T. (1997). *The EM algorithm and Extensions*. New York: Wiley.

[82] McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. New York: Wiley.

[83] Morris, T. H., Richmond, D. R., and Grimshaw, S. D. (1996). Orientation of dinosaur bones in riverine environments: insights into sedimentary dynamics and taphonomy. *The Continental Jurassic*, Morales, M., ed., 521-530. Museum of Northern Arizona, Flagstaff, Arizona.

[84] Newcomb, S. (1886). A generalized theory of the combination of the observations so as to obtain the best result. *American Journal of Mathematics*, **8**, 343-366.

[85] Neyman, J. and Scott, E. L. (1966). On the use of $C(\alpha)$ optimal test of composite hypotheses. *Bulletin de l'Institut International de Statistique* **41**, 447-497.

[86] Ott, J. (1999). *Analysis of Human Genetic Linkage, Third Edition*. Baltimore: The Johns Hopkins Unversity Press.

[87] Panda, S, Antoch, M. P., Miller, B. H., Su, A. I., Schook, A. B., Straume, M., Schultz, P. G., Kay, S. A., Takahashi, J. S. and Hogenesch, J. B. (2002). Coordinated transcription of key pathways in the mouse by the circadian clock. *Cell*, **109**, 307-320.

[88] Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society Of London A*, **185**, 71-110.

[89] Proschan, F. (1963). Theoretical Explanation of Observed Decreasing Failure Rate. *Technometrics*, **5**, 375-383.

[90] Qin, Y. and Smith, B. (2004). Likelihood ratio test for homogeneity in normal mixtures in the presence of a structural parameter. *Statistica Sinica*, **14**, 1165-1177.

[91] Reoder, K. (1994). A graphical technique for determining the number of components in a mixture of normals. *Journal of the American Statistical Association*, **89**, 487-500.

[92] Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, Series B*, **59**, 731-792.

[93] Rubin, H. (1956). Uniform convergence of random functions with applications to statistics. *Annals of Mathematical Statistics*, **27**, 200-203.

[94] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461-464.

[95] Self, S. G. and Lang, K. Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, **82**, 605-610.

[96] Serfling, R. J. (1980). *Approximation Theorem of Mathematical Statistics.* New York: Wiley.

[97] Smyth, P. (2000). Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and Computing*, **10**, 63-72.

[98] Stephens, M. (1969). Techniques for directional data. Technical Report 150, Department of Statistics, Stanford University.

[99] Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components - An alternative to reversible jump methods. *The Annalys of Statistics*, **28**, 40-74.

[100] Sun, J. (1993). Tail probabilities of the maxima of Guassian random fields. *Annals of Probability* **21**, 34-71.

[101] Tadesse, M., Sha, N., and Vannucci, M. (2005). Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association*, **100**, 602-617.

[102] Tan X., Chen J. and Zhang R. (2006). Consistency of the constrained maximum likelihood estimator in finite normal mixture models. *Submitted*.

[103] Tanner, M. Y. and Wong W. H. (1987) The calculation of poserior distribution by data augmentation. *Journal of the American Statistical Association*. **67**, 702-708.

[104] Teicher, H. (1960). On the mixture of distributions. *Annals of Mathematical Statistics*, **31**, 55-73.

[105] Teicher, H. (1963). Identifiability of finite mixtures. *Annals of Mathematical Statistics*, **34**, 1265-1269.

[106] Tibshirani, R. J. (1996). Regression shrinkage and selection via the LASSO, *Journal of the Royal Statistical Society, Series B*, **58**, 267-288.

[107] Titterington, D. M., Smith, A. F. M. and Markov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley.

[108] von Mises, R. (1918). *Über* die "ganzzahligkeit" der atomgewichte und verwandte fragen. Phys. Z., **19**, 490-500.

[109] Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics*, **20**, 595-601.

[110] Weldon, W. F. R. (1893). On certain correlated variations in Crangon vulgaris. *Proceedings of the Royal Society of London*, **54**, 318-329.

[111] Wilks S. S. (1938). The large sample distribution of the likelihood ratio for testing composition hypotheses. *Annals of Mathematical Statistics* **9**, 60-62.

[112] Withers, C. S. (1996). Moment estimates for mixtures of several distributions its different means of scales. *Communications in Statistics–Theory and Methods*, **25**, 1799-1824.

[113] Wolfe, J. H. (1971). A Monte Carlo study of the sampling distribution of the likelihood ratio for mixtures of multinormal distributions. *Technical Bulletin* **STB 72-2**. San Diego: U.S. Naval Personnel and Training Research Laboratory.

[114] Woo, M. J and Sriram, T. N. (2006). Robust estimation of mixture complexity. *Journal of the American Statistical Association*, **101**, 1475-1486

[115] Wu, C. F. J. (1983). On the convergence properties of the EM alorighm. *The Annals of Statistics*, **11**, 95-103.

[116] Yakowitz, S. J. and Spragins, J. D. (1968). On the identifiability of finite mixtures. *Annals of Mathematical Statistics*, **39**, 209-214.