University of Alberta

HOMOGENEITY TEST IN FINITE MIXTURE MODELS USING EM-TEST

by

Xiaoqing Niu

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Statistics

Department of Mathematical and Statistical Sciences

Spring 2014
Edmonton, Alberta

Abstract

The class of finite mixture models is widely used in many areas, including science, humanities, medicine, engineering, among many others. Testing homogeneity is one of the important and challenging problems in the application of finite mixture models. It has been investigated by many researchers and most of the existing works have focused on the univariate mixture models, normal mixture models on the mean parameters only, and normal mixture models on both mean and variance parameters. This thesis concentrates on testing homogeneity in multivariate mixture models, scale mixtures of normal distributions, and a class of contaminated normal models.

We first propose the use of the EM-test (Li, Chen, & Marriott, 2009) to test homogeneity in multivariate mixture models. We show that the EM-test statistic has asymptotically the same distribution as the likelihood ratio test for testing the restricted mean of a multivariate normal distribution given one observation. Based on this result, we suggest a resampling procedure to approximate the p-value of the EM-test.

Scale mixture of normal distributions, i.e., mixture of normal distributions on the variance parameters, has wide applications. However, an effective testing procedure specifically for testing homogeneity in this class of mixture models is not available. We retool the EM-test (Chen & Li, 2009) for testing homogeneity in the scale mixture of normal distributions. We show that the retooled EM-test has the simple limiting distribution $\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$.

Large-scale hypothesis testing problem appears in many areas such as microarray studies. We propose a new class of contaminated normal models,

which is a two-component normal mixture model with one component mean being zero and different component variances, and can be used in large-scale hypotheses. We further design a new EM-test for testing homogeneity in this class of mixture models. It is shown that the new EM-test statistic has a simple shifted $\frac{1}{2}\chi^2_1 + \frac{1}{2}\chi^2_2$ limiting distribution.

In all the three scenarios, extensive simulation studies are conducted to examine whether the limiting distributions approximate the finite sample distributions reasonably well and whether the EM-tests have appropriate power to detect heterogeneity in the alternative models. To demonstrate the application of the proposed methods, several real-data examples are analyzed.

Acknowledgements

I would never have been able to finish my dissertation without the guidance of my committee members, support from my friends, and encouragement from my family.

Foremost, I would like to express my sincere gratitude to my supervisors Dr. Pengfei Li and Dr. Ivan Mizera for their continuous support of my Ph.D study and research. Dr. Li's patience, motivation, and solid theoretical knowledge have greatly inspired me. Dr. Mizera's enthusiasm, immense knowledge and exquisite computational skills have deeply impressed me. They showed me the skills of research, the way of professional presenting and discussion, how to improve in the areas that I'm not good at and how to form good professional research habits. Their guidance helped me in all the time of transforming from an undergraduate student to a qualified Ph.D candidate.

My sincere thanks also goes to my committee member Dr. Rohana Karunamuni for his encouragement and insightful comments.

I would also like to thank our Department of Mathematical and Statistical Sciences and Dr. Rhonda Rosychuk, for offering me the teaching assistantship and research assistantship, which made my pursuit of Ph.D happen and continue smoothly. Thanks also goes to the staff in the department general office, Tara Schuetz, Patti Bobowsky, and Rick Mickalonis for their help and care.

I thank my fellow colleagues and friends: Dr. Dandan Luo, Manli Yan, Shuo Tong, Dr. Khurram Nadeem, Phoebe Ye, Yingming Or, Xin Zhang, Dr. Yidan Ding, Eason Shen, Dr. Xiaoxia Ye, Xuan Wu, and many others; from whom I learned new things, and with whom I gained new experiences, worked hard,

and had fun during the past five years. I also want to take this opportunity and thank Drs. Bo Li and Hongwei Liu in China Central Normal University for enlightening me the first glance of mathematics and research.

I wish to thank my parents, Hanmin Niu and Hui Yu. Without their constant encouragement and believing in me, I would not have completed this thesis with such courage and persistence. Finally, I would like to thank my husband, who is also a great colleague and friend, Yin Li. He is always there standing by me through the good times and bad.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Finite Mixture Modeling

In this thesis, we contribute to testing homogeneity in finite mixture models, i.e., testing if the data come from a homogeneous or heterogeneous population. Before moving to the details, we give a general introduction and provide definitions about finite mixture models in this section.

1.1.1 General introduction

The concept of finite mixture modeling can be dated back to more than one hundred years ago. In 1894, Professor Karl Pearson used a mixture of two normal distributions to model the heterogeneity in the crab data (Pearson, 1894). After that, finite mixture models have been widely used in many areas involving statistical modeling such as science, humanities, medicine, engineering, and so on. Comprehensive reviews can be found in Titterington, Smith, & Makov (1985), Lindsay (1995), McLachlan & Peel (2000), Schlattmann (2009), Zucchini & MacDonald (2009) and references therein.

In general, finite mixture models are called for if several subgroups of data are mixed together but the group label for each observation is unobserved, missing, or unknown due to some other reasons. In some cases, we may have prior knowledge on the the number of subgroups. In this situation, our interest is to recover the group labels and to estimate the proportion of each subgroup. For example, in microarray studies, we observe the gene expression levels of a large number of genes of both healthy individuals and patients with a certain disease. The purpose is to identify those differentially expressed genes in two samples. A t-test often serves this purpose. To account for the multiple hypothesis testing issue, geneticists favour the notion of controlling the false discovery rate (Benjamini & Hochberg, 1995). Among many recipes for controlling this rate, Efron (2004) used a finite normal mixture to classify the genes into the null and alternative subgroups based on the z-scores derived from the individual t-tests. A mixture of two normal distributions is often used to model the z-scores due to the existence of null and alternative subgroups (McLachlan, Bean, & Jones, 2006; Dai & Charnigo, 2010). Based on this model, we can estimate the proportion of the differentially expressed genes and calculate the probability that a given gene is differentially expressed in two samples.

More often in practice, the number of groups can not be given by externally existing information. Sometimes, we even do not know whether the mixture structure truly exists. In this scenario, a first and foremost step is to estimate the number of groups.

In the following, we give two examples which illustrate the wide applications of finite mixture models.

Example 1.1. (Crab data) *The crab data, collected by a biologist Professor*

Walter Weldon, consist of the forehead breadth to body length ratios for 1000 crabs sampled at Naples. Weldon (1892) noticed an obvious asymmetry pattern in the histogram of 1000 ratios (See also, McLachlan & Peel, 2000, pp. 3, Fig. 1.1). He suspected that there may exist evolutionary divergence of crab species, which well explains Pearson (1894)'s motivation of using a mixture of two normal distributions to model the crab data.

Example 1.2. (Emergency department visit data) *It is important to monitor and study the patient flow in a certain hospital for patients, medical care staff, and the hospital administrators. The patient flow in hospital emergency department (ED) is mainly represented by the patients' bed occupancy time. Patients visit the ED for different reasons; some come for acute care, some for rehabilitation, and some for long-term-care. Due to such a difference, the occupancy time in ED may not be reasonably modeled by a homogeneous distribution such as exponential distribution (Harrison, 2001). Instead, a mixture of two or three exponential distributions is thought to be more suitable, given the presence of acute care, rehabilitation, and possible long-term-care patients (Harrison & Millard, 1991; Li & Chen, 2010). A question of practical importance is the existence of the patient subgroups. If they exist, what is the number of patient subgroups? Obviously, this is a typical example for the second situation we discussed above.*

1.1.2 Definitions

In last subsection, we illustrate the use of finite mixture models in different areas. In this subsection, we give formal definitions of finite mixture models and the identifiability in the finite mixture model context. We further define

some notation that are used throughout the thesis.

Finite mixture models can be formulated with an incomplete-data structure. Denote $X_1, X_2, \ldots, X_n$ as a random sample of size n. For each random variable or vector X_i, we define an associate group-label vector $\mathbf{Z}_i$ of dimension m. The jth element Z_{ij} of $\mathbf{Z}_i$ is defined to be one or zero, depending on whether X_i is from the jth group or not. Let $f(x;\theta)$ be a density function from the parametric distribution family $\{f(x;\theta)|\theta \in \Theta \subseteq \mathbb{R}^d, d \geq 1\}$.

Suppose $\mathbf{Z}_i$ independently and identically follows a multinomial distribution with size 1 and m categories with corresponding probability vector being $(\alpha_1, \alpha_2, \ldots, \alpha_m)^\tau$ such that $\alpha_j \geq 0$ and $\sum_{j=1}^m \alpha_j = 1$. Precisely,

$$Pr\{\mathbf{Z}_i = \mathbf{z}_i\} = \alpha_1^{z_{i1}} \alpha_2^{z_{i2}} \cdots \alpha_m^{z_{im}},$$

where $\mathbf{z}_i = (z_{i1}, \ldots, z_{im})^\tau$. Further we assume that $X_1, X_2, \ldots, X_n$ are conditionally independent given $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_n$ with conditional density function:

$$f(x_i|z_{ij} = 1) = f(x_i;\theta_j).$$

Under this formulation, the marginal distribution of $X_1, X_2, \ldots, X_n$ is a m-component finite mixture model:

$$\sum_{j=1}^{m} \alpha_j f(x;\theta_j). \tag{1.1}$$

Let Ψ be the cumulative distribution function with m support points in Θ

such that

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

Then the finite mixture model in (1.1) can be further written as

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

In (1.2), $f(x; \theta)$ is called the kernel or the component density distribution; $\Psi(\theta)$ is the mixing distribution; m is the number of components in, or the order of, the finite mixture model; the weights α_j, $j = 1, 2, \ldots, m$ are the mixing proportions; and $\theta_j, j = 1, 2, \ldots, m$ are the mixing parameters or component parameters.

In Example 1.2, Li & Chen (2010) suggest that the following exponential mixture model is the most suitable one to model the ED data:

$$f(x; \Psi) = \alpha_1 \theta_1 e^{-\theta_1 x} + \alpha_2 \theta_2 e^{-\theta_2 x} + \alpha_3 \theta_3 e^{-\theta_3 x}, \quad x > 0.$$

In this mixture model, the component density distribution $f(x; \theta)$ is the probability density function of the exponential distribution with rate θ; the order $m = 3$; the mixing distribution $\Psi(\theta)$ is given as

$$\Psi(\theta) = \sum_{j=1}^{3} \alpha_j I(\theta_j \leq \theta).$$

We now define an important concept associated with finite mixture models – identifiability. Identifiability of parameters in general models means that in a parametric distribution family, the same distribution infers the same set of

parameters. In other words, different sets of parameters must define different distributions in that family. The identifiability in finite mixture models is similarly defined, but allows permutations among the component parameters.

Definition 1.1. (Identifiability) *Let* $f(x;\Psi) = \sum_{j=1}^{m} \alpha_j f(x;\theta_j)$ *be a member of the parametric family of finite mixture models. This family of mixture models is defined as identifiable if for any two distributions* $f(x;\Psi)$ *and* $f(x;\Psi^*)$,

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

infers that $m = m^*$, $(\alpha_1,\ldots,\alpha_m) = (\alpha_1^*,\ldots,\alpha_m^*)$, *and* $(\theta_1,\ldots,\theta_m) = (\theta_1^*,\ldots,\theta_m^*)$ *after permutations of the component labels.*

The identifiability of finite mixture models has been well studied in the literature. The aforementioned finite mixture of normal models in Example 1.1 and finite mixture of exponential models in Example 1.2 are both shown to be identifiable (Teicher, 1963). We refer to McLachlan & Peel (2000, pp. 26–28) and Charnigo & Pilla (2007).

We finish this section with some discussions. In applications, the order m in (1.2) may not be known in advance. Sometimes we even do not know if the data come from a homogeneous ($m = 1$) or heterogeneous ($m > 1$) population. This thesis mainly concentrates on testing homogeneity, i.e., testing $m = 1$ in some finite mixture models. In the next section, we first review the inference procedures on the order of a finite mixture model, which include testing $m = 1$ as a special case. In Section 1.3, we justify the importance of testing $m = 1$ and present our contributions and the organization of the thesis.

1.2 Inference on the Order of a Finite Mixture Model

In the application of finite mixture models, the order m is a crucial parameter. From the theoretical point of view, if the used order is larger than the true order, the optimal convergence rate of the mixing distribution can only be at most $n^{-1/4}$, instead of $n^{-1/2}$ (Chen, 1995). In real applications, the order often has important scientific implications, see Chapter 6 of McLachlan & Peel (2000), Li & Chen (2010), and Chen, Li, & Fu (2012). More specifically, in Example 1.1, the order of the normal mixture model represents the number of crab species at Naples. In Example 1.2, the order of the exponential mixture model is the number of patient subgroups in the ED. Another enlightening example in statistical genetics is presented below.

Example 1.3. (Genotype detection) *The phenotypes that display continuous or quantitative variation in human population are decided by genes and slightly affected by the environment. In the simplest case, the phenotype is decided by a gene with two alleles A and a. There are three genotypes AA, Aa (equivalent as aA), and aa that one individual can possess. Suppose the phenotypes associated with individuals possessing the three different genotypes are distributed as* $N(\mu_{AA}, \sigma^2_{AA})$, $N(\mu_{Aa}, \sigma^2_{Aa})$, *and* $N(\mu_{aa}, \sigma^2_{aa})$ *respectively. Here,* $N(\mu, \sigma^2)$ *denotes the normal distribution with mean* μ *and variance* σ^2. *Then the finite normal mixture model can be applied to model the population phenotype distribution (Schork, Allison, & Thiel, 1996; McLachlan & Peel, 2000). The order* $m = 1$ *represents the proposition that the hypothesized major gene does not exist; the order* $m = 2$ *implies that there are two different phenotypes and the major gene is dominant; the order* $m = 3$ *means that the major gene is*

additive; the order $m > 3$ tells that there may exist multiple influential genes.

Due to the importance of the order in the application of finite mixture models as illustrated in the above examples, determining the order (also known as the order selection problem) is a very important problem in the mixture model area. In recent decades, there have been considerable developments on the order selection problem. The major procedures include: (1) information-based approaches; (2) penalized distance based approaches; (3) penalized log-likelihood based approaches; (4) fully Bayesian approaches; (5) hypothesis testing based approaches. We briefly review these five types of procedures in the following five subsections.

1.2.1 Information-based approaches

The information-based approaches mainly consider Kullback-Leibler (KL; Kullback & Leibler, 1951) information (Akaike, 1973, 1974; Ishiguro, Sakamoto, & Kitagawa, 1997; Pan, 1999; Smyth, 2000; Miloslavsky & van der Laan, 2003; Windham & Cutler, 1992; Bozdogan, 1990, 1993), Bayesian information (Schwarz, 1978; Raftery, 1996; Roberts et al., 1998; Ishwaran, James, & Sun, 2001; Frühwirth-Schnatter, 2004), and classification-based information (Biernacki & Govaert, 1997; Celeux & Soromenho, 1996; Biernacki, Celeux, & Govaert, 1998).

The Kullback-Leibler information measures the distance between two density functions $f(x;\Psi)$ and $f(x;\hat{\Psi})$ in terms of

$$\begin{aligned} KL(f(x;\Psi), f(x;\hat{\Psi})) &= E_{f(x;\Psi)} \log\{f(X;\Psi)/f(X;\hat{\Psi})\} \\ &= E_{f(x;\Psi)} \log f(X;\Psi) - E_{f(x;\Psi)} \log f(X;\hat{\Psi}). \quad (1.3) \end{aligned}$$

Here $\hat{\Psi}$ is the estimator of Ψ for a given order. Minimizing the KL distance of the true model and the fitted model is a natural way of model selection. In (1.3), there are two terms, only the second term $-E_{f(x;\Psi)} \log f(X; \hat{\Psi})$ is affected by the fitted value $\hat{\Psi}$. Thus, minimizing $KL(f(x;\Psi), f(x;\hat{\Psi}))$ is equivalent to maximizing $E_{f(x;\Psi)} \log f(X; \hat{\Psi})$, which is the expected log fitted density. A simple estimator of the expected log fitted density is

$$\frac{1}{n} \sum_{i=1}^{n} \log f(x_i; \hat{\Psi}) = \frac{1}{n} l_n(\hat{\Psi}), \tag{1.4}$$

which uses the empirical distribution function to calculate the expectation. Here $l_n(\Psi)$ is the log-likelihood function of Ψ for the given data. However, this approximation often overestimates the expected log fitted density. Let $F(x;\Psi)$ be the cumulative distribution function of $f(x;\Psi)$. To deal with the bias $b(F)$ of (1.4) as an estimator of the expected log fitted density, several different approaches are proposed in literature. These different approaches lead to different model selection criteria.

The Akaike information criterion (AIC; Akaike, 1973, 1974), is based on the fact that the bias equals d asymptotically, where d is the total number of parameters in the model. In the context of order selection, the AIC criterion selects the order that minimizes

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

In this formula, $\hat{\Psi}$ is the maximum likelihood estimate (MLE) of Ψ, and d is the number of free parameters under the given order. In late 1990s, the bias term was proposed to be estimated using bootstrap (Efron, 1979) by Ishig-

uro, Sakamoto, & Kitagawa (1997) and Pan (1999). The Efron (bootstrap) information criterion (EIC) selects the order that minimizes

$$-2l_n(\hat{\Psi}) + 2b(\hat{F}_n).$$

Here, $b(\hat{F}_n)$ denotes the nonparametric bootstrap bias given by B bootstrap samples. Another bias correction approach proposed by Smyth (2000) is via cross-validation. This cross-validation-based information criterion (CVIC) utilizes v-fold cross-validation ($v \geq 1$) to estimate the expected log fitted density. The CVIC was further developed by Miloslavsky & van der Laan (2003). The minimum information ratio criterion (MIR; Windham & Cutler, 1992) and the informational complexity criterion (ICOMP; Bozdogan, 1990, 1993) also stemmed from the KL information minimization approach.

The Bayesian information criterion (BIC; Schwarz, 1978) is developed under the Bayesian framework. The objective function to be minimized is

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

which is equivalent to maximizing the leading term of the posterior probability of a given model. Although the BIC is derived under the certain regularity conditions, which are not satisfied by finite mixture models, it is still widely used to select the order of a finite mixture, see Fraley & Raftery (1998), Leroux (1992), Roeder & Wasserman (1997), Campbell et al. (1997), Dasgupta & Raftery (1998) and reference therein. On the basis of the Laplace's approximation to the posterior probability of a given model, Raftery (1996) proposed the Laplace-Metropolis Criterion (LMC). Unlike the technique used in the BIC,

the LMC used the simulated posterior replicates of Ψ to calculate the posterior probability. The Laplace-Empirical Criterion (LEC) is close to BIC and LMC. The only difference is that the posterior probability is approximated by utilizing the empirical information matrix, see Roberts et al. (1998). Other similar methods include weighted Bayes factor method with decomposition of the posterior probability for a given model (Ishwaran, James & Sun, 2001) and the Bridge sampling techniques (Frühwirth-Schnatter, 2004).

The classification-based criteria concentrates on the notion of classification likelihood, which is referred to as the complete-data likelihood in the EM framework. Suppose that $X_1, \ldots, X_n$ is a random sample from $f(x; \Psi)$. With the notation defined in Section 1.1.2, the classification log-likelihood (or the complete-data log-likelihood) is given as

$$l_c(\Psi; \mathbf{Z}_1, \ldots, \mathbf{Z}_n) = \sum_{i=1}^{n} \sum_{j=1}^{m} z_{ij} \{\log \alpha_j + \log f_j(X_i; \theta_j)\}. \tag{1.5}$$

The posterior probability of the observation X_i belonging to the j-th component is defined as

$$\tau_{ij} = \tau_j(X_i; \Psi) = \frac{\alpha_j f_j(X_i; \theta_j)}{\sum_{k=1}^{m} \alpha_k f(X_i; \theta_k)}.$$

With the MLE of Ψ, we estimate τ_{ij} as:

$$\hat{\tau}_{ij} = \tau_j(X_i; \hat{\Psi}), \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, m.$$

Replacing z_{ij} and Ψ in $l_c(\Psi; \mathbf{Z}_1, \ldots, \mathbf{Z}_n)$ in (1.5) by $\hat{\tau}_{ij}$ and $\hat{\Psi}$, respectively, we

have

$$l_c(\hat{\Psi}; \mathbf{Z}_1, \ldots, \mathbf{Z}_n) = l_n(\hat{\Psi}) + \sum_{i=1}^{n} \sum_{j=1}^{m} \hat{\tau}_{ij} \log \hat{\tau}_{ij}. \tag{1.6}$$

In the context of classification, negative of the second term on the right hand side in (1.6) is called the entropy $EN(\hat{\boldsymbol{\tau}})$ of the fuzzy classification matrix $\hat{\boldsymbol{\tau}} = (\hat{\tau}_{ij})_{n \times m}$. Based on (1.6), Biernacki & Govaert (1997) proposed the classification likelihood information criterion (CLC), which minimizes

$$-2l_n(\hat{\Psi}) + 2EN(\hat{\boldsymbol{\tau}}).$$

Other classification-based criteria include the normalized entropy criterion (NEC; Celeux & Soromenho, 1996) and the integrated classification likelihood (ICL; Biernacki, Celeux, & Govaert, 1998). The detailed description and discussion can be found in McLachlan & Peel (2000) and reference therein.

1.2.2 Penalized distance based approaches

Penalized distance based approaches use the distance between the fitted mixture model with a certain order and the nonparametric estimate of the population distribution. In literature, there are several penalized distance based methods. Different methods use different measures of distance or different penalty functions. We review three of them in details here. Other methods use similar ideas and can be referred to the reference therein.

Chen & Kalbfleisch (1996) proposed adding a penalty term $\sum_{j=1}^{m} \log \alpha_j$ to the distance measure and then choosing the model by minimizing the penalized distance. It is shown that the estimated order is strongly consistent under some regularity conditions. In James, Priebe, & Marchette (2001), the order

selection procedure in normal mixture model was suggested by minimizing the penalized KL distance. The fitted model is the convolution of a normal distribution and a finite normal mixture distribution with a given order. They showed that the estimated order has almost sure convergence towards the true order. Another approach was proposed by Woo & Sriram (2006, 2007), which minimizes the penalized Hellinger distance. Through simulations, this method is shown to be robust and perform well under model misspecification.

1.2.3 Penalized log-likelihood based approaches

The third important stream of the order selection procedure is based on the penalized log-likelihood function. The AIC and BIC can also be regarded as two approaches in this class. Under some regularity conditions, Leroux (1992) proved that the AIC and BIC do not underestimate the true order almost surely. Keribin (2000) further showed that BIC can consistently estimate the order. More recently, Chen & Khalili (2008) suggested adding two penalty terms to the log-likelihood function. One of the penalty is to prevent any of the mixing proportions from being too close to zero. The other penalty is to prevent fitting a mixture model whose sub-populations only differ slightly. Under some regularity conditions, the estimated order is strongly consistent. Two other penalized log-likelihood approaches focusing on the normal mixture model were proposed in Fujisawa & Eguchi (2006) and Huang, Peng, & Zhang (2013) respectively.

1.2.4 Fully Bayesian approaches

In the literature, there are two main types of Bayesian approaches for selecting the order a finite mixture model. The first type is to compute the posterior probability for all possible orders and then select the order with the largest posterior probability. The challenging part is the computation of the posterior probability. This type of Bayesian approach has been already reviewed before in Section 1.2.1.

The second type is to sample from the joint posterior density

$$p(m, \Psi_m | X_1, \ldots, X_n),$$

where Ψ_m is the mixing distribution with the order m. The major difficulty is that the dimension of the parameters varies with the order m. This violates the condition of convergence in usual Markov chain Monte Carlo (MCMC) methods. Carlin & Chib (1995) proposed the product-space MCMC to sample from mixtures with varying number of components. Richardson & Green (1997) used the reversible jump MCMC methods developed by Green (1995). By splitting or combing the mixture sub-populations, the model moves forward or backward with different order m. Stephens (2000) proposed the birth-death MCMC. The model parameters are viewed as a point process with each point representing one mixture component. Then a continuous time Markov birth-death process is constructed allowing the mixture components to be "born" and to "die". Recently, McGrory & Titterington (2007) showed that the variational approach gives automatic selection of the order in the normal mixture model. It is shown that the variational approach is a useful alternative to MCMC.

1.2.5 Hypothesis testing approaches

The last but not least procedure for selecting the order is the hypothesis testing based method. In the following, we review the results for testing $m = 1$, whose importance will be discussed in Section 1.3. We defer the discussion about testing the general order to Chapter 5.

The testing of homogeneity or testing $m = 1$ under finite mixture models has attracted much attention in the last two decades. The likelihood ratio test (LRT) is the most extensively studied method for this kind of hypothesis testing problem. Because of the nonregularity of mixture models, the limiting distribution of the LRT is no longer the χ^2–distribution but involves the supremum of a Gaussian process and is not convenient in practice (Chen & Chen, 2001; Dacunha-Castelle & Gassiat, 1999; Liu & Shao, 2003). Furthermore, the limiting distribution of the LRT is derived under two undesirable conditions: (i) the compactness of the parameter space; and (ii) the finiteness of the Fisher information in the mixing proportion direction, see Chen & Chen, 2001, Dacunha-Castelle & Gassiat, 1999, and Li, Chen, & Marriott, 2009. If the parameter space Θ for the mixing parameter is not bounded, the LRT diverges to ∞ as the sample size n goes to infinity (Hartigan, 1985; Liu, Pasarica, & Shao, 2003; Liu & Shao, 2003).

The modified likelihood ratio test (MLRT) proposed by Chen (1998) and Chen, Chen, & Kalbfleisch (2001) can be implemented easily. Under some mild conditions, the limiting distribution of the MLRT is a mixture of χ^2 distributions when the mixing parameter is univariate. The result of the MLRT for multivariate mixture models is however not available. Also, the asymptotic results of the MLRT depend on the compact parameter space and finite

Fisher information conditions mentioned above. Later on, Chen & Kalbfleisch (2005) applied the MLRT to test the homogeneity in normal mixture models with common and unknown component variance. They found that the asymptotic upper bound of the MLRT is the χ^2_2 with the exact limiting distribution remaining unknown.

Charnigo & Sun (2004, 2008) proposed a class of D-test for testing the homogeneity in finite mixture models. The D-test measures the L_2-distance between the homogeneous model and the alternative mixture model. They show that the D-test has many computational advantages and nice asymptotic properties under univariate mixture models and under normal mixture models with common and unknown component variance (Charnigo & Sun, 2010). Later on, the D-test has further been applied to the contaminated models with one component being fully specified, and contaminated normal mixture model with one component mean being 0 and the common and unknown component variance (Dai & Charnigo, 2007, 2008, 2010; Charnigo, Zhou, & Dai, 2013). It is worth noting that the asymptotic properties of the D-test also rely on the compact parameter space and finite Fisher information conditions.

The EM-test was proposed recently by Chen & Li (2009) and Li, Chen, & Marriott (2009). Some of the ideas of the EM-test can be traced back to the MLRT, but it has several additional advantages. Most notably, it is more widely applicable; for example, when used to test homogeneity under univariate mixture models, its limiting distribution does not rely on the two aforementioned undesirable conditions. Subsequently, Chen & Li (2011) proposed a computer experiment based approach to address the tuning parameter selection issue for the EM-test. The approximation precision of the EM-test is further improved. Note that all the published results for the EM-test are for

(i) univariate mixture models; (ii) normal mixture models with common and unknown component variance; (iii) normal mixture models on both means and variances.

1.3 Contributions and Outline of the Thesis

Testing homogeneity, that is, testing whether the data come from a homogeneous or a heterogeneous population, is one of the most important problems in the application of finite mixture models. For the sake of parsimony, if the data come from a homogeneous population, there is no need to apply a mixture model. In some applications, the order $m = 1$ is often proposed to represent some default proposition with scientific significance; the rejection of which usually leads to propositions of greater interest (Chen, Li, & Fu, 2012). For instance, in Example 1.1, $m = 1$ represents the default proposition that the evolutionary divergence of crab species does not exist; the rejection of this supports the existence of evolutionary divergence. In Example 1.3, $m = 1$ represents the default proposition that the suspected major gene does not exist; the rejection of this supports the existence of a major gene.

Due to the importance of testing homogeneity in the application of finite mixture models, this thesis makes further contributions to this area of research. We devote to developing effective hypothesis testing procedures for three classes of finite mixture models: multivariate mixture models, scale mixtures of normal models, and a class of contaminated normal mixture models.

Multivariate mixture models, or mixture models with multivariate mixing parameters, have a lot of applications. For example, the mixture of multinomial distributions has been used to model the heterogeneity in the transformed

repeated reaction times of 197 children in a developmental psychology study (Cruz-Medina, Hettmansperger, & Thomas, 2004). Detecting the heterogeneity in the reaction times is an important problem with scientific interest. As we have reviewed in Section 1.2.5, the asymptotic results for most existing methods concentrate on the univariate mixture models. Motivated by this example, we propose the use of the EM-test for testing homogeneity in multivariate mixture models in Chapter 2. We show that the EM-test statistic has asymptotically the same distribution as the likelihood ratio test for testing the restricted mean of a multivariate normal distribution based on one observation. On the basis of this result, we suggest a resampling procedure to approximate the p-value of the EM-test. Simulation studies show that the EM-test has accurate type-I error and adequate power, and is more powerful and computationally efficient than the bootstrap likelihood ratio test. Two real-data sets are analyzed to illustrate the application of our theoretical results.

Scale mixtures of normal distributions, i.e., mixtures of normal distributions on the variance parameters, are widely used to model the heavy-tailed data. They have a lot of applications in clinical chemistry, image data, finance, and economics. The testing of homogeneity is one of the fundamental problem in the application of scale mixtures of normal distributions. Chen & Li (2009) proposed a class of EM-tests for testing homogeneity in mixtures of normal distributions on the mean parameters only and in mixtures of normal distributions on both the mean and variance parameters. An effective testing procedure specifically designed for the scale mixtures of normal distributions is lacking in the literature. In Chapter 3, we retool the EM-test proposed in Chen & Li (2009) for testing homogeneity in scale mixtures of normal distributions. We show that the retooled EM-test has the simple lim-

iting distribution $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$. We also use a computational method to provide an empirical value for the tuning parameter selection. Simulation studies show that the retooled EM-test has an accurate size and is more powerful than existing methods such as the likelihood ratio test and the method in Chen & Li (2009). To demonstrate the application of the proposed method, we analyze two real-data examples.

This work of Chapter 4 is mainly motivated by modelling the z-scores derived from the two-sample t-tests (see pp. 2 in Section 1.1.1) in large-scale hypothesis testing problem. Dai & Charnigo (2010) suggested modelling the z-scores by a contaminated normal model, which is a two-component normal mixture with one component mean being zero and the common and unknown component variances. They further proposed testing the existence of differentially expressed genes by testing homogeneity in the proposed contaminated normal model. We observe that in many applications, the component variances may not be the same. One particular example will be given in Section 4.4. More examples can be found in Efron (2004) and McLachlan, Bean, & Jones (2006). In Chapter 4, we first suggest modelling the z-scores by a new class of contaminated normal model. The new model is also a two-component normal mixture model, in which one component mean is still zero, but two component variances can be different. We further propose an EM-test to detect the existence of the differentially expressed genes by testing the homogeneity in the new class of contaminated normal models. We show that the EM-test statistic asymptotically has a simple shifted $\frac{1}{2}\chi_1^2 + \frac{1}{2}\chi_2^2$ distribution. Extensive simulation studies show that, the proposed testing procedure has accurate type-I error and adequate powers for detecting the heterogeneity. A real-data example is analyzed to illustrated the proposed method.

Chapter 5 includes a brief summary of the thesis and provides some future works.

Chapter 2

EM-test in Multivariate Mixture Models[1]

2.1 Introduction and Motivating Example

The class of multivariate mixture models has attracted growing attention in the literature and has applications in genetics, biology, epidemiology, psychology, and many other fields. A mixture of multivariate normals is the most popular statistical model for model-based clustering analysis (McLachlan & Peel, 2000). A mixture of multivariate Poisson distributions is often used for multivariate count data, where the incidences of several related events are observed (Karlis & Meligkotsidou, 2007). For example, a mixture of bivariate Poisson distributions has been used to model the number of two dependent kinds of claims in automobile insurance (Partrat, 1994), the number of surface and interior faults of lenses (Karlis & Meligkotsidou, 2007), and the number of

[1]A version of this chapter has been published. Niu, X., Li, P., & Zhang, P. (2011). *Canadian Journal of Statistics*, 39, 218–238.

parasites in two species of pheasants (Kopocinski, 1999). A mixture of multinomial distributions is useful for categorical data exhibiting overdispersion or heterogeneity. See Cruz-Medina, Hettmansperger, & Thomas (2004), Morel & Nagaraj (1993), and Cadez, Smyth, & Mannila (2001) for its applications to a psychological study, a housing satisfaction study, and marketing, respectively. More examples involving multivariate mixture models can be found in Johnson, Kotz, & Balakrishnan (1997), McLachlan & Peel (2000), Roos (2003), Karlis & Meligkotsidou (2007), and the references therein.

In the following, we present a motivating example showing a possible application of multivariate mixture models.

Example 2.1. (Reaction-time data) *It is well accepted that there are large differences among individuals, especially children, with regards to their visual, hearing, neurological, and mental abilities. In developmental psychology, this different performance is characterized by different probability distributions corresponding to a task. The reaction time captures the main characteristic of the response process. The data come from a reaction-time experiment that studied normally developing nine-year-old children. Each received two visual stimuli and indicated by pressing the appropriate key whether the image on the right was an exact copy or a mirror image of the one on the left. Six repeated measurements were recorded for each child. The data set is composed of the reaction time (in milliseconds) for every trial of the 197 children who responded to all six trials correctly (Cruz-Medina, Hettmansperger, & Thomas, 2004).*

As stated in Cruz-Medina, Hettmansperger, & Thomas (2004), the question of interest is whether the reaction-time data provide strong evidence for the hypothesis that there are subgroups of respondents, each following a different reaction-time distribution. Instead of imposing parametric assumptions on the

reaction-time distribution, Cruz-Medina, Hettmansperger, & Thomas (2004) suggested dividing the real line into $d+1$ nonoverlapping intervals. For each child, a $(d+1)$-dimensional vector is then obtained by counting the number of measurements from the six trials that fall into each of the $d+1$ intervals. If there are subgroups of respondents with different reaction-time distributions, a mixture of multinomial distributions is appropriate for modeling the $(d+1)$-dimensional count vectors; otherwise a homogeneous multinomial distribution is more suitable. Therefore, a testing procedure for homogeneity under the mixture multinomial distribution will provide the necessary justification.

As we have discussed in Chapter 1, most results of the existing methods for testing homogeneity can not be directly applied to multivariate mixture models. In this chapter, we propose the use of the EM-test for testing homogeneity under multivariate mixture models. We explore its asymptotic properties and develop software implementing the test for some commonly used multivariate kernels. The software is written in the R language (R Development Core Team, 2011).

The rest of this chapter is organized as follows. In Section 2.2, we set up the problem and present the asymptotic results of the EM-test. A resampling procedure is proposed to approximate the p-values of the EM-test based on its asymptotic results. In Section 2.3, simulation studies are performed to explore the type-I error and the power of the test, and to compare the EM-test with the bootstrap LRT (McLachlan, 1987). Two real-data examples are included in Section 2.4. For ease of presentation, all the proofs are in the last Section 2.6.

2.2 Main Results

Suppose $X_1, \ldots, X_n$ are a random sample of size n from the two-component multivariate mixture model

$$(1-\alpha)f(x;\boldsymbol{\theta}_1) + \alpha f(x;\boldsymbol{\theta}_2),$$

where $f(x;\boldsymbol{\theta})$ belongs to a parametric family of probability density functions, the mixing parameters $\boldsymbol{\theta}_j = (\theta_{j1}, \ldots, \theta_{jd})^\tau \in \Theta$ for $j = 1, 2$, and the mixing proportion $\alpha \in [0, 1]$. We aim to test

$$H_0 : \alpha(1-\alpha)(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) = 0,$$

that is, to test whether the data are from a homogeneous model $f(x, \boldsymbol{\theta})$. Without loss of generality, we will assume that $\alpha \in [0, 0.5]$. Throughout this chapter, we use $\boldsymbol{\theta}$ instead of θ to denote the mixing parameter to emphasize that the mixing parameter is a vector. In this chapter, the random variables $X_1, \ldots, X_n$ can be of dimension one or higher. For simplicity of presentation, we still use the notation $X_1, \ldots, X_n$.

In this section, we first present the EM-test and then investigate its asymptotic properties.

2.2.1 The EM-test

We denote the log-likelihood function as

$$l_n(\alpha, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \sum_{i=1}^{n} \log\{(1-\alpha)f(X_i;\boldsymbol{\theta}_1) + \alpha f(X_i;\boldsymbol{\theta}_2)\}$$

and define the modified log-likehood function to be

$$pl_n(\alpha, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = l_n(\alpha, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) + p(\alpha).$$

The penalty function $p(\alpha)$ is a continuous function of α, and it is used to prevent α from taking a value close to zero. A concrete recommendation for $p(\alpha)$ will be given in Section 2.3.

Before giving the details of the procedure, we want to highlight the basic idea of the EM-test. The primary motivation of the EM-test is to gain back the efficiency lost by the constrained LRT in which the mixing proportion α is fixed to be $\alpha_0 \in (0, 0.5]$. The goal of using the EM-algorithm (Dempster, Laird, & Rubin, 1977) to update the mixing proportion and other component parameters is to improve the power of the test. The resulting EM-test statistic is defined as the maximum of the LRT statistics from several different starting values of α, and usually a few iterations suffice.

The EM-test procedure is initialized by choosing a finite set of $\{\alpha_1, \ldots, \alpha_J\} \subset (0, 0.5]$ for α and a positive integer K. An example of $\{\alpha_1, \ldots, \alpha_J\}$ is $\{0.1, 0.3, 0.5\}$. Here K is the number of EM-iterations at the time the test is applied.

For each $j = 1, 2, \ldots, J$, compute

$$(\boldsymbol{\theta}_{j,1}^{(1)}, \boldsymbol{\theta}_{j,2}^{(1)}) = \arg\max_{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2} pl_n(\alpha_j, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2).$$

The EM-algorithm or the downhill simplex method (optim function in R) can be used to search for $\boldsymbol{\theta}_{j,1}^{(1)}$ and $\boldsymbol{\theta}_{j,2}^{(1)}$. Let $\alpha_j^{(1)} = \alpha_j$. The EM-iteration starts from here.

Suppose $\alpha_j^{(k)}$, $\boldsymbol{\theta}_{j,1}^{(k)}$, and $\boldsymbol{\theta}_{j,2}^{(k)}$ are available. The calculation for $k = 1$ has

been illustrated above. For $i = 1, 2, \ldots, n$, and the current k, we use an E-step to compute the posterior probabilities

$$w_{ij}^{(k)} = \frac{\alpha_j^{(k)} f(X_i; \boldsymbol{\theta}_{j,2}^{(k)})}{(1-\alpha_j^{(k)}) f(X_i; \boldsymbol{\theta}_{j,1}^{(k)}) + \alpha_j^{(k)} f(X_i; \boldsymbol{\theta}_{j,2}^{(k)})}$$

and then update the mixing proportion α and the two mixing parameters $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$ by an M-step such that

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

and

$$\left(\boldsymbol{\theta}_{j,1}^{(k+1)}, \boldsymbol{\theta}_{j,2}^{(k+1)}\right) = \arg\max_{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2} \Big\{ \sum_{i=1}^{n} (1 - w_{ij}^{(k)}) \log f(X_i; \boldsymbol{\theta}_1) + \sum_{i=1}^{n} w_{ij}^{(k)} \log f(X_i; \boldsymbol{\theta}_2) \Big\}.$$

The E-step and the M-step are iterated $K - 1$ times.

For each k and j, we define

$$M_n^{(k)}(\alpha_j) = 2\{pl_n(\alpha_j^{(k)}, \boldsymbol{\theta}_{j,1}^{(k)}, \boldsymbol{\theta}_{j,2}^{(k)}) - pl_n(0.5, \hat{\boldsymbol{\theta}}_0, \hat{\boldsymbol{\theta}}_0)\}$$

where $\hat{\boldsymbol{\theta}}_0 = \arg\max_{\boldsymbol{\theta}} pl_n(0.5, \boldsymbol{\theta}, \boldsymbol{\theta})$. The EM-test statistic is then defined as

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

We reject the null hypothesis when $EM_n^{(K)}$ exceeds some critical value of the limiting distribution presented in the next subsection.

2.2.2 Asymptotic properties

We first investigate the asymptotic behavior of the EM-test procedure by presenting a result that sheds light on the iteration process.

Theorem 2.1: *Suppose that* $f(x;\boldsymbol{\theta})$ *and* $p(\alpha)$ *satisfy the regularity conditions in Section 2.6.1. Under the null distribution* $f(x;\boldsymbol{\theta}_0)$*, for each given* $\alpha_j \in (0, 0.5]$ *and* $k \leq K$*, we have*

$$\alpha_j^{(k)} - \alpha_j = o_p(1),\ \boldsymbol{\theta}_{j,1}^{(k)} - \boldsymbol{\theta}_0 = O_p(n^{-1/4}),\ \boldsymbol{\theta}_{j,2}^{(k)} - \boldsymbol{\theta}_0 = O_p(n^{-1/4}),$$

and $(1-\alpha_j^{(k)})(\boldsymbol{\theta}_{j,1}^{(k)} - \boldsymbol{\theta}_0) + \alpha_j^{(k)}(\boldsymbol{\theta}_{j,2}^{(k)} - \boldsymbol{\theta}_0) = O_p(n^{-1/2})$.

Note that the iteration changes the value of α by only an $o_p(1)$ quantity. This is the crucial property that simplifies the asymptotic theory of the EM-test.

To present the limiting distribution, we need some additional notation. For $i = 1, \ldots, n$ and $1 \leq h < l \leq d$, let

$$Y_{ih} = \frac{\partial f(X_i;\boldsymbol{\theta}_0)/\partial\theta_h}{f(X_i;\boldsymbol{\theta}_0)},\quad Z_{ih} = \frac{\partial^2 f(X_i;\boldsymbol{\theta}_0)/\partial\theta_h^2}{2f(X_i;\boldsymbol{\theta}_0)},\quad U_{ihl} = \frac{\partial^2 f(X_i;\boldsymbol{\theta}_0)/(\partial\theta_h\partial\theta_l)}{f(X_i;\boldsymbol{\theta}_0)},$$

and

$$\mathbf{b}_{1i} = \Big(Y_{i1}, \ldots, Y_{id}\Big)^\tau,$$
$$\mathbf{b}_{2i} = \Big(Z_{i1}, \ldots, Z_{id}, U_{i12}, \ldots, U_{i1d}, U_{i23}, \ldots, U_{i2d}, \ldots, U_{i\,d-1\,d}\Big)^\tau.$$

For $j, k = 1, 2$, set $\mathbf{B}_{jk} = E\left[\{\mathbf{b}_{ji} - E(\mathbf{b}_{ji})\}\{\mathbf{b}_{ki} - E(\mathbf{b}_{ki})\}^\tau\right]$. Here the expectation is taken with respect to the true model $f(x;\boldsymbol{\theta}_0)$ under the null hypothesis.

Furthermore, we orthogonalize $\mathbf{b}_{1i}$ and $\mathbf{b}_{2i}$ by introducing $\tilde{\mathbf{b}}_{2i} = \mathbf{b}_{2i} -$

$\mathbf{B}_{21}\mathbf{B}_{11}^{-1}\mathbf{b}_{1i}$, which has variance-covariance matrix $\tilde{\mathbf{B}}_{22} = \mathbf{B}_{22} - \mathbf{B}_{21}\mathbf{B}_{11}^{-1}\mathbf{B}_{12}$. The following theorem presents the limiting distribution of the EM-test.

Theorem 2.2: *Assume the same conditions as in Theorem 2.1, and that* $\alpha_1 = 0.5$. *Let* $\mathbf{w} = (w_1, \ldots, w_{d(d+1)/2})^\tau$ *be a zero-mean multivariate normal random vector with variance-covariance matrix* $\tilde{\mathbf{B}}_{22}^{-1}$. *Under the null distribution* $f(x; \boldsymbol{\theta}_0)$ *and for any fixed finite* K, *as* $n \to \infty$,

$$EM_n^{(K)} \to \mathbf{w}^\tau \tilde{\mathbf{B}}_{22}\mathbf{w} - \inf_{\mathbf{v}}(\mathbf{w}-\mathbf{v})^\tau \tilde{\mathbf{B}}_{22}(\mathbf{w}-\mathbf{v}) \tag{2.1}$$

in distribution. Here

$$\mathbf{v} = (v_1^2, \ldots, v_d^2, v_1v_2, \ldots, v_1v_d, v_2v_3, \ldots, v_2v_d, \ldots, v_{d-1}v_d)^\tau$$

consists of the lower triangular elements of the symmetric matrix

$$(v_1, \ldots, v_d)^\tau (v_1, \ldots, v_d).$$

All the possible values of $\mathbf{v}$ *form a cone in* d*-dimensional space.*

It should be noted here that the right-hand side of (2.1) is the distribution of the LRT for testing $\mathbf{v} = 0$ against $\mathbf{v} \neq 0$, based on one observation from the multivariate normal distribution with mean vector $\mathbf{v}$ and variance-covariance matrix $\tilde{\mathbf{B}}_{22}^{-1}$. When $d = 1$, $\mathbf{v} = (v_1^2)$ can take only nonnegative values and therefore the distribution in (2.1) is $0.5\chi_0^2 + 0.5\chi_1^2$, which is the same as the result in Li, Chen, & Marriott (2009). Theorem 2.2 extends the result to $d \geq 2$.

The limiting distribution depends on the matrix $\tilde{\mathbf{B}}_{22}$. For some kernels, it

has a simple form. For example, for a multivariate normal kernel $MN(\boldsymbol{\theta}, I_d)$ with mean vector $\boldsymbol{\theta}$ and known variance-covariance matrix equal to the identity matrix I_d, we have

$$\tilde{\mathbf{B}}_{22} = \text{diag}\left\{\underbrace{0.5, \ldots, 0.5}_{d}, \underbrace{1, \ldots, 1}_{d(d-1)/2}\right\}.$$

For a d-dimensional multivariate product Poisson distribution $MultiPois(\theta_1, \theta_2, \ldots, \theta_d)$, with the density

$$\prod_{i=1}^{d} \frac{\theta_i^{x_i} e^{-\theta_i}}{x_i!},$$

we have

$$\tilde{\mathbf{B}}_{22} = \text{diag}\left\{\frac{1}{2\theta_1^2}, \ldots, \frac{1}{2\theta_d^2}, \frac{1}{\theta_1\theta_2}, \ldots, \frac{1}{\theta_1\theta_d}, \frac{1}{\theta_2\theta_3}, \ldots, \frac{1}{\theta_{d-1}\theta_d}\right\},$$

where diag$\{\cdot\}$ denotes a diagonal matrix constructed from its argument. As for the multivariate product Poisson kernel, the calculation of $\tilde{\mathbf{B}}_{22}$ may depend on the value of $\boldsymbol{\theta}_0$, which can be estimated by $\hat{\boldsymbol{\theta}}_0$.

2.2.3 A resampling procedure

In applications, the limiting distribution is often used to calculate the approximate p-value of the test. However, the limiting distribution in Theorem 2.2 may not have an explicit form when $d \geq 2$. To overcome this difficulty, we propose the following resampling procedure to approximate the p-values of the EM-test on the basis of its limiting distribution.

Step 0. Calculate $\tilde{\mathbf{B}}_{22}$. If necessary, estimate $\boldsymbol{\theta}_0$ by $\hat{\boldsymbol{\theta}}_0$.

Step 1. Generate M random vectors, $\left\{\mathbf{w}^{(m)}, m = 1, \ldots, M\right\}$, from the

multivariate normal distribution with mean vector zero and variance-covariance matrix $\tilde{\mathbf{B}}_{22}^{-1}$. The appropriate choice of M depends on the desired precision level.

Step 2. For each vector $\mathbf{w}^{(m)}$, calculate the LRT statistic (denoted by Q_m) for testing $\mathbf{v} = 0$ against $\mathbf{v} \neq 0$. That is,

$$Q_m = (\mathbf{w}^{(m)})^\tau \tilde{\mathbf{B}}_{22}(\mathbf{w}^{(m)}) - \inf_{\mathbf{v}}(\mathbf{w}^{(m)} - \mathbf{v})^\tau \tilde{\mathbf{B}}_{22}(\mathbf{w}^{(m)} - \mathbf{v}).$$

Then, $\{Q_m, m = 1, \ldots, M\}$ can be viewed as M random observations from the limiting distribution.

Step 3. Approximate the p-values of the EM-test statistics by

$$\frac{\#\{m : Q_m > EM_n^{(K)}\}}{M}.$$

We have written an R function (R Development Core Team, 2011) to calculate the approximate p-values from the limiting distribution.

2.3 Simulation Studies

2.3.1 Simulation results for the EM-test

We first conduct extensive simulation to check whether the limiting distribution approximates the finite sample distribution of the EM-test reasonably well and whether the EM-test has appropriate power to detect heterogeneity in the alternative models. We use the penalty function $p(\alpha) = \log(1 - |1 - 2\alpha|)$, with $K = 2$ or 3 iterations, and three initial values of $\alpha \in \{0.1, 0.3, 0.5\}$ to calculate

the EM-test statistics.

We consider the dimension $d = 2$ or 3 and two kernels: (i) multinomial distribution $Multinom(m; \theta_1, \ldots, \theta_{d+1})$ and (ii) multivariate product Poisson distribution $MultiPois(\theta_1, \ldots, \theta_d)$. For each of the four combinations of the kernel and the dimension, we choose two null models and four alternative models. The four alternative models are formed by mixing the two null models with $1-\alpha = 0.5, 0.25, 0.1, 0.05$. In total, there are eight null models and sixteen alternative models. The details of the null models are given in Table 2.1. The details of the alternative models together with their corresponding Kullback-Leibler (KL) information with respect to the null model are given in Table 2.2.

Table 2.1: Eight null multivariate mixture models.

No.	Model	No.	Model
N1	$Multinom(12; 1/3, 1/3, 1/3)$	N2	$Multinom(12; 1/4, 1/2, 1/4)$
N3	$Multinom(12; 1/4, 1/4, 1/4, 1/4)$	N4	$Multinom(12; 1/6, 1/3, 1/3, 1/6)$
N5	$MultiPois(5, 5)$	N6	$MultiPois(3, 5)$
N7	$MultiPois(5, 5, 5)$	N8	$MultiPois(1, 3, 5)$

Table 2.2: Sixteen alternative multivariate mixture models.

No.	Model	100KL	No.	Model	100KL
A1	0.5N1+0.5N2	2.03	A2	0.25N1+0.75N2	1.22
A3	0.1N1+0.9N2	0.34	A4	0.05N1+0.95N2	0.11
A5	0.5N3+0.5N4	2.03	A6	0.25N3+0.75N4	1.29
A7	0.1N3+0.9N4	0.38	A8	0.05N3+0.95N4	0.13
A9	0.5N5+0.5N6	1.22	A10	0.25N5+0.75N6	0.85
A11	0.1N5+0.9N6	0.27	A12	0.05N5+0.95N6	0.09
A13	0.5N7+0.5N8	29.76	A14	0.25N7+0.75N8	23.32
A15	0.1N7+0.9N8	8.11	A16	0.05N7+0.95N8	3.23

In each simulation, we consider three significance levels, 10%, 5%, and 1%, and two sample sizes, $n = 100$ and 200. We compute the null rejection rates under each of the eight null models based on 5000 repetitions and the powers under each of the sixteen alternative models based on 1000 repetitions. The resampling procedure presented in Section 2.3 with $M = 10000$ is used to calculate the approximate p-values of the EM-test statistics.

The null rejection rates of the EM-test are given in Table 2.3. We find that the simulated null rejection rates are quite close to the nominal levels in all eight selected null models. In Fig. 2.1, we present the histograms of the EM-test statistic $EM_n^{(2)}$ based on the 5000 repetitions together with the kernel density estimates of the limiting distribution based on $M = 10000$ random observations from the limiting distribution. We consider four null models N1, N3, N5, and N7 with $n = 200$. It is shown that the density estimate of the limiting distribution matches the finite sample distribution well in all four null models. Therefore, we conclude that the limiting distribution provides a good approximation to the finite sample distribution of the EM-test.

The powers of the EM-test are given in Table 2.4. To save space, only powers at the 5% significance level are reported. In all sixteen alternative models, the powers of the EM-test are larger than the nominal level and increase as the sample size increases. We also observe that as the mixing proportion $1-\alpha$ increases from 0.05 to 0.5, for instance, from Model A4 to Model A1, the power of the EM-test increases. This is because there is cumulating information about the heterogeneity. We notice that for some alternative models, such as A3, A4, A7, A8, A11, and A12, the corresponding Kullback-Leibler information with respect to the null model is quite small. This explains why the power of the EM-test is quite low for those models. The large Kullback-Leibler information

Figure 2.1: Histograms of $EM_n^{(2)}$ and kernel density estimates of limiting distribution under N1, N3, N5, and N7 with $n = 200$ in the multivariate mixture models.

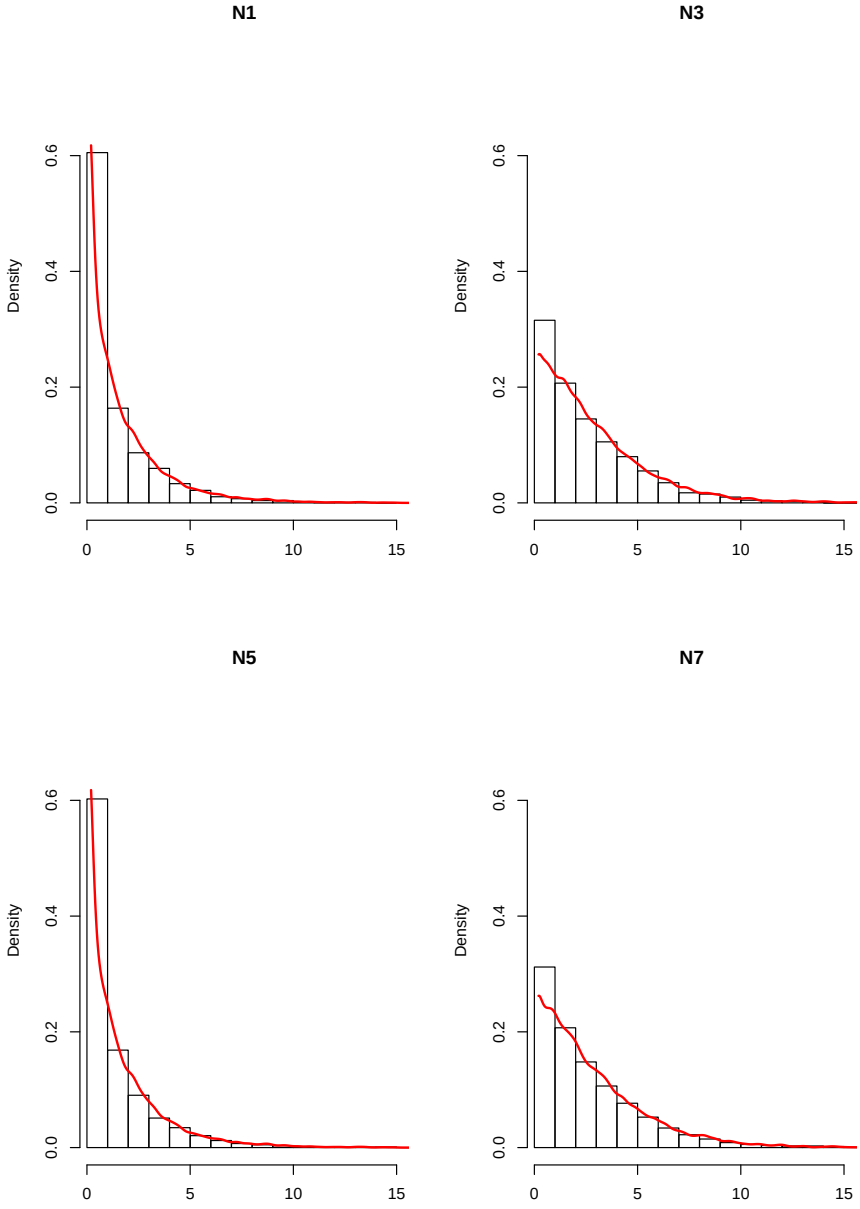


Table 2.3: Simulated type-I error rates (%) of the EM-test under null multivariate mixture models.

	Level=10%			Level=5%			Level=1%		
Model	$EM_n^{(1)}$	$EM_n^{(2)}$	$EM_n^{(3)}$	$EM_n^{(1)}$	$EM_n^{(2)}$	$EM_n^{(3)}$	$EM_n^{(1)}$	$EM_n^{(2)}$	$EM_n^{(3)}$
				$n = 100$					
N1	8.9	9.1	9.1	4.4	4.5	4.6	0.9	0.9	0.9
N2	9.4	9.6	9.6	4.7	4.8	4.8	0.8	0.8	0.8
N3	9.3	9.5	9.5	4.8	4.9	5.0	1.0	1.1	1.1
N4	9.4	9.4	9.5	5.0	5.2	5.3	1.1	1.1	1.1
N5	9.0	9.1	9.1	4.4	4.5	4.5	0.8	0.8	0.8
N6	9.6	9.6	9.6	5.2	5.3	5.3	1.0	1.0	1.1
N7	9.3	9.5	9.6	5.4	5.6	5.6	1.0	1.1	1.1
N8	9.5	9.7	9.8	5.4	5.5	5.6	0.9	0.9	0.9
				$n = 200$					
N1	9.0	9.1	9.1	4.4	4.4	4.4	0.9	0.9	0.9
N2	9.6	9.6	9.6	4.9	4.9	4.9	0.8	0.8	0.8
N3	8.5	8.6	8.6	4.3	4.3	4.4	0.8	0.9	0.9
N4	9.4	9.5	9.5	4.8	4.9	4.9	0.9	0.9	0.9
N5	9.1	9.1	9.1	4.4	4.4	4.4	0.9	1.0	1.0
N6	9.1	9.1	9.1	4.9	5.0	5.0	1.0	1.0	1.0
N7	9.3	9.4	9.5	4.7	4.7	4.8	1.1	1.1	1.1
N8	9.6	9.6	9.7	5.1	5.1	5.1	0.7	0.8	0.8

of model A16 explains the large power of the EM-test, although the mixing proportion for the first component is small.

2.3.2 Comparison with the bootstrap LRT

In this subsection, we compare the EM-test with the bootstrap LRT, measuring the computational time to obtain the corresponding p-value, the type-I error, and the power. To compare the computational time, we first generate two random samples of size 100 and 200, respectively, from each of four models,

Table 2.4: Powers (%) of the EM-test at the 5% significance level under alternative multivariate mixture models.

	$n = 100$			$n = 200$		
Model	$EM_n^{(1)}$	$EM_n^{(2)}$	$EM_n^{(3)}$	$EM_n^{(1)}$	$EM_n^{(2)}$	$EM_n^{(3)}$
A1	55.2	55.2	55.2	79.6	79.7	79.7
A2	33.6	33.6	33.6	59.8	59.9	59.9
A3	12.5	12.8	12.8	20.8	20.8	20.8
A4	8.0	8.2	8.2	10.4	10.4	10.4
A5	42.3	42.4	42.4	71.4	71.4	71.4
A6	26.7	27.1	27.1	49.5	49.6	49.6
A7	13.5	13.6	13.7	17.0	17.2	17.2
A8	8.2	8.4	8.4	11.8	11.9	11.9
A9	34.1	34.2	34.2	56.9	56.9	56.9
A10	25.0	25.1	25.1	47.5	47.5	47.5
A11	13.7	13.9	13.9	18.4	18.5	18.6
A12	7.2	7.4	7.4	10.8	10.8	10.9
A13	100	100	100	100	100	100
A14	100	100	100	100	100	100
A15	95.6	95.6	95.6	99.9	99.9	99.9
A16	69.0	69.4	69.4	90.3	90.4	90.5

N1, N3, N5, and N7. For each random sample, we calculate the p-values of the EM-test statistic using the procedure in Section 2.3 with $M = 10000$, and the p-values of the LRT statistic using the parametric bootstrap procedure (McLachlan, 1987) with a bootstrap size of 500. The computational times are given in Table 2.5. All the results are obtained using the same computer. For the EM-test, it takes less than two minutes to obtain the corresponding p-value for $d = 2$ and less than three minutes for $d = 3$. The sample size does not greatly affect the computational time. For the bootstrap LRT, it is computationally intensive to obtain the p-value, especially for the multinomial mixture.

Table 2.5: Computational time (in minutes) of the EM-test and the bootstrap LRT in the multivariate mixture models.

	$n = 100$		$n = 200$	
Model	EM-test	Bootstrap LRT	EM-test	Bootstrap LRT
N1	1.0	226.8	1.4	571.0
N3	2.2	258.9	2.5	666.4
N5	1.1	9.1	1.1	16.0
N7	2.6	25.3	2.6	46.2

Next we compare the type-I error and the power of the EM-test with the bootstrap LRT. We consider two null models, N5 and N6, and four alternative models, A9, A10, A11, and A12. A bootstrap size of 500 is used for the bootstrap LRT. We compute the type-I errors and the powers of the bootstrap LRT based on 500 repetitions. The results at the 5% level are reported in Table 2.6. Comparing Table 2.6 with Tables 2.3 and 2.4, we see that the EM-test has larger power than the bootstrap LRT under A9 and A10 and similar power to the bootstrap LRT under A11 and A12.

Table 2.6: Simulated type-I error rates (%) of the bootstrap LRT at the 5% significance level under two null and four alternative multivariate mixture models.

	Bootstrap LRT	
Model	n=100	n=200
N5	5.0	5.4
N6	6.4	3.8
A9	31.8	48.8
A10	22.4	39.2
A11	12.4	16.6
A12	7.8	10.6

2.4 Real-data Examples

Example 2.1. (Continued) We now apply the EM-test to the reaction-time data. Following Cruz-Medina, Hettmansperger, & Thomas (2004), ten cut points are chosen: 500, 1000, 1200, 1400, 1600, 2000, 2500, 3000, 4000, and 5000. This leads to eleven intervals that start with $[448, 500]$ and end with $[5000, 7919]$. For each child, the six measurements are then transformed to an eleven-dimensional vector by counting the number of measurements in each of the eleven intervals. The EM-test is then applied to the resulting 197 eleven-dimensional vectors. Under the null hypothesis, the maximum likelihood estimate is

$$\hat{\boldsymbol{\theta}}_0 = (0.0008, 0.0440, 0.0753, 0.1328, 0.1294, 0.2217, 0.1633, 0.0990, 0.0880, 0.0305, 0.0152)^\tau.$$

The EM-test statistics are found to be $EM_n^{(1)} = 237.1917$, $EM_n^{(2)} = 238.1338$, and $EM_n^{(3)} = 238.3934$. To generate the random samples from the limiting distribution, we need to minimize a function in a ten-dimensional space. To save computational time, the approximate p-values for the EM-test statistics are calculated based on the resampling procedure with $M = 1000$ repeated samples, and they are all found to be zero. Thus, the data provide overwhelming evidence for the existence of subgroups of children, each of which follows a different reaction-time distribution. The total computational time for obtaining the EM-test statistics and their p-values is 4.7 minutes.

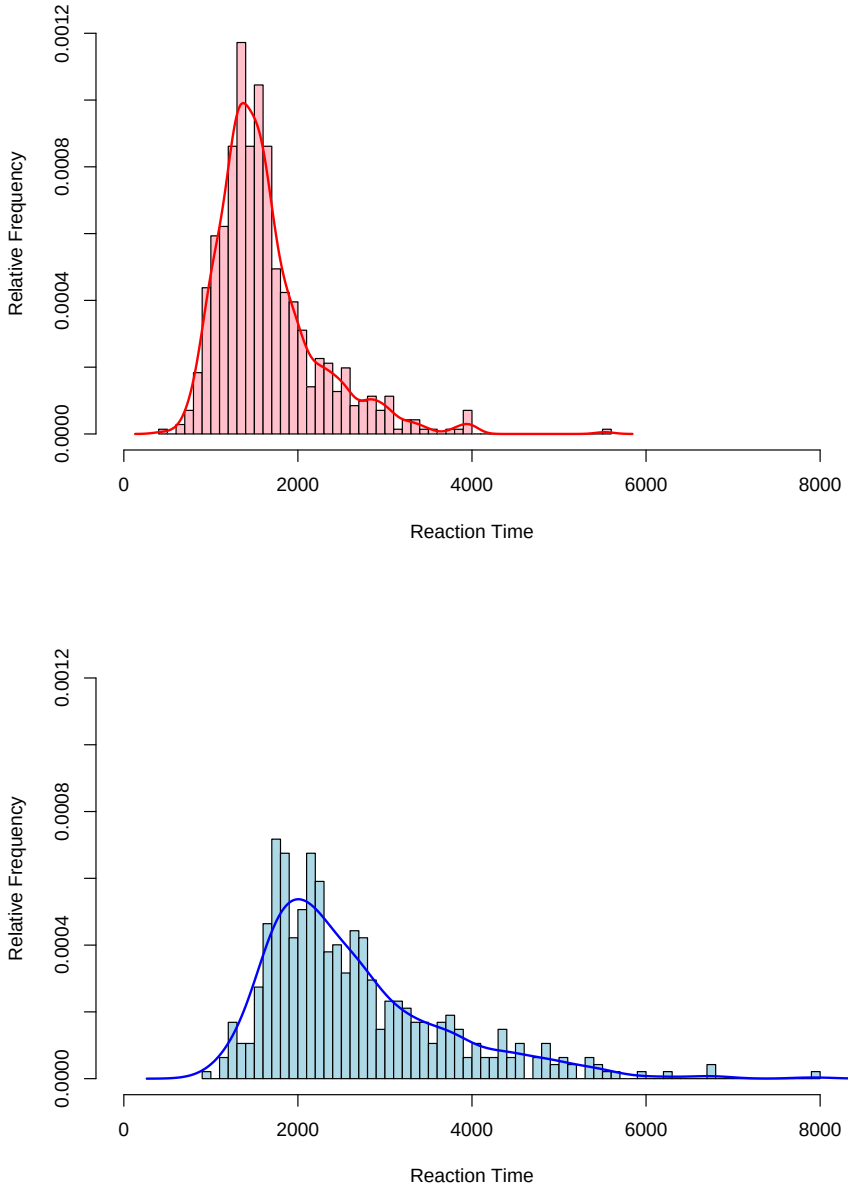
We next fit a two-component multinomial mixture model and then cluster

the 197 children into two subgroups according to the posterior probabilities. Figure 2.2 gives the histograms and kernel density estimates of the reaction time of the children in the two subgroups. Clearly, the density functions of the reaction time in these two subgroups are fairly different, which is in accordance with the EM-test result.

Example 2.2. (Lens-fault data) This example considers the counts of surface and interior faults in 100 lenses. These data were analyzed with a multivariate Poisson-log normal distribution (Aitchison & Ho, 1989), which is a continuous mixture of bivariate product Poisson distributions, and a finite mixture of bivariate product Poisson distributions (Karlis & Meligkotsidou, 2007).

Karlis & Meligkotsidou (2007) used the Akaike information criterion (Akaike, 1974) to choose the order of the finite mixture and found that a mixture model is more suitable than a homogeneous model. However, they did not derive a homogeneity test procedure. For illustration, we applied the EM-test to the lens-fault data. Under the null hypothesis, the maximum likelihood estimate $\hat{\boldsymbol{\theta}}_0 = (3.25, 2.93)^\tau$. The EM-test statistics are found to be $EM_n^{(1)} = 31.786$, $EM_n^{(2)} = 31.786$, and $EM_n^{(3)} = 31.786$. The resampling procedure in Section 2.3 with $M = 10000$ gives the approximate p-values as zero for all three statistics. Therefore, we have strong evidence in favor of the alternative hypothesis, that is, a mixture of bivariate product Poisson distributions is more suitable for the lens data. The total computational time for obtaining the EM-test statistics and their p-values is 1.1 minutes.

Figure 2.2: Histograms and kernel density estimates of the reaction time of the children in first subgroup (upper panel) and in second subgroup (lower panel).



2.5 Discussion

Comparing the EM-test with the bootstrap LRT

Both the EM-test in multivariate case and the bootstrap LRT involve resampling procedures to calculate the approximate p-values for the corresponding test statistics. The difference is that for the EM-test, the random observations are generated from its limiting distribution; while for the bootstrap LRT, the random observations are generated from the corresponding finite sample distribution. The EM-test was shown to be more computationally efficient than the bootstrap LRT for obtaining the approximate p-values. Moreover, the EM-test is more powerful than the bootstrap LRT when one of the two mixing proportions in the alternative model is close to 0.5, and it has comparable power to the bootstrap LRT when one of the two mixing proportions in the alternative model is close to 0.

Choice of the penalty $p(\alpha)$

To make the asymptotic results in Theorems 2.1 and 2.2 valid, the penalty $p(\alpha)$ must satisfy Condition A0 in Section 2.6.1. The penalty $p(\alpha) = \log(1 - |1 - 2\alpha|)$ and the penalty proposed in Chen, Chen, & Kalbfleisch (2001), $p(\alpha) = C\log\{4\alpha(1-\alpha)\}$, both meet the requirement. Our empirical experience shows that both penalty functions can tightly control the type-I error of the EM-test. However, the EM-test based on the penalty $p(\alpha) = \log(1 - |1 - 2\alpha|)$ has larger power than that based on $p(\alpha) = C\log\{4\alpha(1-\alpha)\}$ when one of the mixing proportions in the alternative model is close to 0 or 1. Furthermore, the penalty $p(\alpha) = \log(1 - |1 - 2\alpha|)$ does not complicate the implementation

of the EM-test. In the M-step of the EM-iteration, the value of α can be easily updated as follows:

$$\alpha_j^{(k+1)} = \begin{cases} \min\{(n+1)^{-1}(\sum_{i=1}^n w_{ij}^{(k)} + 1), 0.5\}, & n^{-1}\sum_{i=1}^n w_{ij}^{(k)} \le 0.5, \\ \max\{(n+1)^{-1}\sum_{i=1}^n w_{ij}^{(k)}, 0.5\}, & n^{-1}\sum_{i=1}^n w_{ij}^{(k)} > 0.5. \end{cases}$$

More details can be found in Li, Chen, & Marriott (2009).

Choice of the set $\{\alpha_1, \ldots, \alpha_J\}$

In general, the specific choice of the set $\{\alpha_1, \ldots, \alpha_J\}$ is not crucial. Our recommendation is $\{0.1, 0.3, 0.5\}$. The updated α-values from either $\alpha = 0.3$ or $\alpha = 0.4$ are likely to be close after two iterations. Further increasing J may not significantly improve the power of the EM-test, which is verified through the simulation.

Optimization issue in (2.1)

To generate the random sample from the limiting distribution, we need to find the infimum of a quartic polynomial of $v_1, \ldots, v_d$, that is,

$$\inf_{\mathbf{v}} (\mathbf{w} - \mathbf{v})^\tau \tilde{\mathbf{B}}_{22} (\mathbf{w} - \mathbf{v})$$

with $\mathbf{v} = (v_1^2, \ldots, v_d^2, v_1v_2, \ldots, v_1v_d, v_2v_3, \ldots, v_2v_d, \ldots, v_{d-1}v_d)^\tau$. For any given $\mathbf{w}$, the infimum is greater than or equal to 0 and bounded above by $\mathbf{w}^\tau \tilde{\mathbf{B}}_{22} \mathbf{w}$. The downhill simplex method (optim function in R) is used to search for the infimum. Since the quartic polynomial may have multiple local minimizers, we suggest using multiple initial values for $v_1, \ldots, v_d$. In our simulation and

real examples, we use five randomly generated initial values for $v_1, \ldots, v_d$. Further increasing the number of initial values to ten gives the same results. The global minimizer may not be unique, but this does not complicate the computation. Our experience indicates that the same infimum value for the quartic polynomial function can always be found.

Application scope of the asymptotic results

The validity of the asymptotic results in Theorems 2.1 and 2.2 requires that the kernel function $f(x; \boldsymbol{\theta})$ satisfies Conditions A1–A5 in Section 2.6.1. The kernel functions satisfying these conditions include the multinomial kernel, multivariate product Poisson kernel, and multivariate normal kernel with a known and same-component covariance matrix. Unfortunately, the multivariate normal kernel with unknown covariance matrix and the multivariate Poisson kernel introduced in Karlis & Meligkotsidou (2007) do not satisfy Condition A5, a weaker version of the strong identifiability condition introduced in Chen (1995). If Condition A5 is not satisfied, the mixture model is not strongly identifiable. Therefore, the best convergence rate $n^{-1/4}$ for the mixing distribution may not be achieved (Chen, 1995). Further study is needed for this special class of mixture models.

2.6 Proof

2.6.1 Regularity conditions

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

A0 The penalty function $p(\alpha)$ is continuous, maximized at $\alpha = 0.5$, and approaches negative infinity as α approaches zero.

A1 (*Wald's integrability conditions*) (i) $E(|\log f(X;\boldsymbol{\theta}_0)|) < \infty$; (ii) for sufficiently small ρ and for sufficient large r, $E[\log(1 + f(X;\boldsymbol{\theta},\rho)] < \infty$ for $\boldsymbol{\theta} \in \Theta$ and $E[\log(1+\varphi(X,r)] < \infty$, where $f(x;\boldsymbol{\theta},\rho) = \sup_{\|\boldsymbol{\theta}'-\boldsymbol{\theta}\|<\rho} f(x;\boldsymbol{\theta}')$ and $\varphi(x,r) = \sup_{\|\boldsymbol{\theta}\|\geq r} f(x;\boldsymbol{\theta})$; (iii) $f(x;\boldsymbol{\theta}) \to 0$ in probability as $\|\boldsymbol{\theta}\| \to \infty$. Here $\|\boldsymbol{\theta}\| = \sqrt{\sum_{i=1}^{d} \theta_i^2}$ is the norm of $\boldsymbol{\theta}$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^\tau$.

A2 (*Smoothness*) $f(x;\boldsymbol{\theta})$ has common support and continuous 5th order partial derivatives with respect to $\boldsymbol{\theta}$.

A3 (*Identifiability*) For any distribution functions Ψ_1 and Ψ_2 with two support points such that $\int_\Theta f(x;\boldsymbol{\theta})\, d\Psi_1(\boldsymbol{\theta}) = \int_\Theta f(x;\boldsymbol{\theta})\, d\Psi_2(\boldsymbol{\theta})$ for all x, we must have $\Psi_1 = \Psi_2$.

A4 (*Uniform boundedness*) For all $h \leq 5$ and $\theta_1, \ldots, \theta_h$, there exists a function g with finite expectation such that

$$\left|\frac{\partial^h f(X_i;\boldsymbol{\theta}_0)/\partial\theta_1 \cdots \partial\theta_h}{f(X_i;\boldsymbol{\theta}_0)}\right|^3 \leq g(X_i).$$

Moreover, there exists a positive ϵ such that

$$\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\leq\epsilon} \left|\frac{\partial^5 f(X_i;\boldsymbol{\theta})/\partial\theta_1 \cdots \partial\theta_h}{f(X_i;\boldsymbol{\theta}_0)}\right|^3 \leq g(X_i).$$

A5 (*Positive definiteness*) The variance-covariance matrix $\mathbf{B}$ of $(\mathbf{b}_{1i}^\tau, \mathbf{b}_{2i}^\tau)^\tau$ is positive definite.

A6 (*Interior point*) $\boldsymbol{\theta}_0$ is an interior point of the parameter space Θ.

2.6.2 Technical lemmas

To prove Theorems 2.1 and 2.2, we need the following three technical lemmas. Lemma 2.1 shows that any estimator with a large likelihood and α not close to zero is consistent for $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ under the null model. Lemma 2.2 strengthens the result of Lemma 1 by providing the exact convergence rate. Lemma 2.3 shows that an EM-iteration will only change the value of α by an $o_p(1)$ quantity. The theorems then follow easily.

Lemma 2.1: *Suppose that Conditions A0–A6 hold. Let* $(\bar{\alpha}, \bar{\boldsymbol{\theta}}_1, \bar{\boldsymbol{\theta}}_2)$ *be the estimators of* $(\alpha, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ *in* $\Lambda = [\delta, 0.5] \times \Theta \times \Theta$ *for some* $\delta \in (0, 0.5]$. *Assume that*

$$pl_n(\bar{\alpha}, \bar{\boldsymbol{\theta}}_1, \bar{\boldsymbol{\theta}}_2) - pl_n(0.5, \boldsymbol{\theta}_0, \boldsymbol{\theta}_0) \geq c > -\infty. \tag{2.2}$$

Under the null distribution $f(x; \boldsymbol{\theta}_0)$, *we have* $\bar{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_0 = o_p(1)$ *and* $\bar{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_0 = o_p(1)$.

Proof. The assumption in (2.2) together with Condition A0 implies that

$$l_n(\bar{\alpha}, \bar{\boldsymbol{\theta}}_1, \bar{\boldsymbol{\theta}}_2) - l_n(0.5, \boldsymbol{\theta}_0, \boldsymbol{\theta}_0) \geq c > -\infty. \tag{2.3}$$

The classic consistency result by Wald (1949) is that if $\hat{\Psi}_n = (1 - \hat{\alpha}) I(\hat{\boldsymbol{\theta}}_1 \leq \boldsymbol{\theta}) + \hat{\alpha} I(\hat{\boldsymbol{\theta}}_2 \leq \boldsymbol{\theta})$ satisfies

$$l_n(\hat{\alpha}, \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2) - l_n(0.5, \boldsymbol{\theta}_0, \boldsymbol{\theta}_0) \geq c > -\infty,$$

for all n, where $I(\cdot)$ is the indicator function, then $\hat{\Psi}_n$ is consistent for $\Psi_0 = I(\boldsymbol{\theta}_0 \leq \boldsymbol{\theta})$, the mixing distribution under the null hypothesis. This result

and (2.3) lead to the consistency of $\bar{\Psi} = (1-\bar{\alpha})I(\bar{\boldsymbol{\theta}}_1 \le \boldsymbol{\theta}) + \bar{\alpha} I(\bar{\boldsymbol{\theta}}_2 \le \boldsymbol{\theta})$ for Ψ_0, which is possible only if $\bar{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_0 = o_p(1)$ and $\bar{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_0 = o_p(1)$ under the assumption that $\bar{\alpha} \in [\delta, 0.5]$ for some $\delta > 0$. □

Lemma 2.2: *Suppose that the conditions of Lemma 2.1 hold. Then under the null distribution $f(x; \boldsymbol{\theta}_0)$,*

$$\bar{\boldsymbol{\theta}}_1 \ -\boldsymbol{\theta}_0 = O_p(n^{-1/4}), \quad \bar{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_0 = O_p(n^{-1/4}),$$
$$\bar{\mathbf{m}}_1 = (1-\bar{\alpha})(\bar{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_0) + \bar{\alpha}(\bar{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_0) = O_p(n^{-1/2}).$$

Proof. Let $R_{1n}(\bar{\alpha}, \bar{\boldsymbol{\theta}}_1, \bar{\boldsymbol{\theta}}_2) = 2[pl_n(\bar{\alpha}, \bar{\boldsymbol{\theta}}_1, \bar{\boldsymbol{\theta}}_2) - pl_n(0.5, \boldsymbol{\theta}_0, \boldsymbol{\theta}_0)]$. It has a natural lower bound

$$R_{1n}(\bar{\alpha}, \bar{\boldsymbol{\theta}}_1, \bar{\boldsymbol{\theta}}_2) \ge 2c.$$

The rest of the proof is devoted to finding an upper bound for R_{1n}, which leads to the order assessment results.

Since the penalty function $p(\alpha)$ is nonpositive, we have

$$R_{1n}(\bar{\alpha}, \bar{\boldsymbol{\theta}}_1, \bar{\boldsymbol{\theta}}_2) \le 2[l_n(\bar{\alpha}, \bar{\boldsymbol{\theta}}_1, \bar{\boldsymbol{\theta}}_2) - l_n(0.5, \boldsymbol{\theta}_0, \boldsymbol{\theta}_0)] = 2\sum_{i=1}^{n} \log(1+\delta_i)$$

with $\delta_i = (1-\bar{\alpha})\{\frac{f(X_i;\bar{\boldsymbol{\theta}}_1)}{f(X_i;\boldsymbol{\theta}_0)} - 1\} + \bar{\alpha}\{\frac{f(X_i;\bar{\boldsymbol{\theta}}_2)}{f(X_i;\boldsymbol{\theta}_0)} - 1\}$. By the inequality $\log(1+x) \le x - x^2/2 + x^3/3$, we then get

$$R_{1n}(\bar{\alpha}, \bar{\boldsymbol{\theta}}_1, \bar{\boldsymbol{\theta}}_2) \le 2\sum_{i=1}^{n} \log(1+\delta_i) \le 2\sum_{i=1}^{n} \delta_i - \sum_{i=1}^{n} \delta_i^2 + 2/3 \sum_{i=1}^{n} \delta_i^3. \qquad (2.4)$$

Next we need to find the asymptotic expansion of each term in (2.4). We start with the linear term. By Lemma 2.1, both $\bar{\boldsymbol{\theta}}_1$ and $\bar{\boldsymbol{\theta}}_2$ are in a small neighborhood of $\boldsymbol{\theta}_0$. Using the 2nd order Taylor's Expansion on $f(X_i; \boldsymbol{\theta}_1)$ and

$f(X_i;\boldsymbol{\theta}_2)$ around $\boldsymbol{\theta}_0$, we obtain

$$\begin{aligned}
\delta_i &= (1-\bar{\alpha})\{\frac{f(X_i;\bar{\boldsymbol{\theta}}_1)}{f(X_i;\boldsymbol{\theta}_0)}-1\}+\bar{\alpha}\{\frac{f(X_i;\bar{\boldsymbol{\theta}}_2)}{f(X_i;\boldsymbol{\theta}_0)}-1\} \\
&= \sum_{h=1}^{d}\{(1-\bar{\alpha})(\bar{\theta}_{1h}-\theta_{0h})+\bar{\alpha}(\bar{\theta}_{2h}-\theta_{0h})\}\frac{\partial f(X_i;\boldsymbol{\theta}_0)/\partial\theta_h}{f(X_i;\boldsymbol{\theta}_0)} \\
&\quad+\sum_{h=1}^{d}\{(1-\bar{\alpha})(\bar{\theta}_{1h}-\theta_{0h})^2+\bar{\alpha}(\bar{\theta}_{2h}-\theta_{0h})^2\}\frac{\partial^2 f(X_i;\boldsymbol{\theta}_0)/\partial\theta_h^2}{2f(X_i;\boldsymbol{\theta}_0)} \\
&\quad+\sum_{h<l}(1-\bar{\alpha})(\bar{\theta}_{1h}-\theta_{0h})(\bar{\theta}_{1l}-\theta_{0l})\frac{\partial^2 f(X_i;\boldsymbol{\theta}_0)/\partial\theta_h\partial\theta_l}{f(X_i;\boldsymbol{\theta}_0)} \\
&\quad+\sum_{h<l}\bar{\alpha}(\bar{\theta}_{2h}-\theta_{0h})(\bar{\theta}_{2l}-\theta_{0l})\frac{\partial^2 f(X_i;\boldsymbol{\theta}_0)/\partial\theta_h\partial\theta_l}{f(X_i;\boldsymbol{\theta}_0)}+\epsilon_{in},
\end{aligned}$$

where ϵ_{in} is the remainder term. The above equation and the notational substitution of Y_{ih}, Z_{ih}, and U_{ihl} lead to

$$\delta_i = \sum_{h=1}^{d}\bar{m}_{1h}Y_{ih}+\sum_{h=1}^{d}\bar{m}_{2h}Z_{ih}+\sum_{1\le h<l\le d}\bar{s}_{hl}U_{ihl}+\epsilon_{in},$$

with

$$\begin{aligned}
\bar{m}_{1h} &= (1-\bar{\alpha})(\bar{\theta}_{1h}-\theta_{0h})+\bar{\alpha}(\bar{\theta}_{2h}-\theta_{0h}), \\
\bar{m}_{2h} &= (1-\bar{\alpha})(\bar{\theta}_{1h}-\theta_{0h})^2+\bar{\alpha}(\bar{\theta}_{2h}-\theta_{0h})^2, \\
\bar{s}_{hl} &= (1-\bar{\alpha})(\bar{\theta}_{1h}-\theta_{0h})(\bar{\theta}_{1l}-\theta_{0l})+\bar{\alpha}(\bar{\theta}_{2h}-\theta_{0h})(\bar{\theta}_{2l}-\theta_{0l}),
\end{aligned}$$

for $1 \le h < l \le d$. Therefore, the linear term has the following asymptotic expansion

$$\sum_{i=1}^{n}\delta_i = \sum_{i=1}^{n}\left(\sum_{h=1}^{d}\bar{m}_{1h}Y_{ih}+\sum_{h=1}^{d}\bar{m}_{2h}Z_{ih}+\sum_{1\le h<l\le d}\bar{s}_{hl}U_{ihl}\right)+\sum_{i=1}^{n}\epsilon_{in}. \quad (2.5)$$

Let

$$\bar{\mathbf{m}} = (\bar{m}_{11}, \ldots, \bar{m}_{1d}, \bar{m}_{21}, \ldots, \bar{m}_{2d}, \bar{s}_{12}, \ldots, \bar{s}_{1d}, \bar{s}_{23}, \ldots, \bar{s}_{2d}, \ldots, \bar{s}_{d-1\,d})^{\tau} \quad (2.6)$$

and

$$\mathbf{b}_i = (Y_{i1}, \ldots, Y_{id}, Z_{i1}, \ldots, Z_{id}, U_{i12}, \ldots, U_{i1d}, U_{i23}, \ldots, U_{i2d}, \ldots, U_{i\,d-1\,d})^{\tau}.$$

Then, (2.5) becomes

$$\sum_{i=1}^{n} \delta_i = \sum_{i=1}^{n} \bar{\mathbf{m}}^{\tau} \mathbf{b}_i + \sum_{i=1}^{n} \epsilon_{in}. \quad (2.7)$$

For clarity, we first assume that the remainder term

$$\sum_{i=1}^{n} \epsilon_{in} = o_p(1) + o_p(n)\{||\bar{\mathbf{m}}||^2\}, \quad (2.8)$$

and we defer its proof to the end.

Since the remainder terms resulting from the square and cubic sums in (2.4) have at least the order of the remainder term from the linear sum, we obtain the following asymptotic expansions for the quadratic and cubic terms in (2.4):

$$\sum_{i=1}^{n} \delta_i^2 = \sum_{i=1}^{n} \left(\bar{\mathbf{m}}^{\tau} \mathbf{b}_i\right)^2 + O_p\left(\sum_{i=1}^{n} \epsilon_{in}\right), \quad (2.9)$$

$$\sum_{i=1}^{n} \delta_i^3 = \sum_{i=1}^{n} \left(\bar{\mathbf{m}}^{\tau} \mathbf{b}_i\right)^3 + O_p\left(\sum_{i=1}^{n} \epsilon_{in}\right). \quad (2.10)$$

By the law of large numbers and the positive definiteness of $\mathbf{B}$ as assumed in

Condition A5,

$$\sum_{i=1}^{n} (\bar{\mathbf{m}}^{\tau}\mathbf{b}_i)^2 = n\bar{\mathbf{m}}^{\tau}\mathbf{B}\bar{\mathbf{m}}\{1 + o_p(1)\}. \tag{2.11}$$

Similarly,

$$\sum_{i=1}^{n} (\bar{\mathbf{m}}^{\tau}\mathbf{b}_i)^3 = o_p(n)||\bar{\mathbf{m}}||^2. \tag{2.12}$$

Combining (2.4) and order assessments (2.7)–(2.12), we arrive at the upper bound

$$R_{1n}(\bar{\alpha}, \bar{\boldsymbol{\theta}}_1, \bar{\boldsymbol{\theta}}_2) \leq 2\bar{\mathbf{m}}^{\tau} \sum_{i=1}^{n} \mathbf{b}_i - n\bar{\mathbf{m}}^{\tau}\mathbf{B}\bar{\mathbf{m}}\{1 + o_p(1)\} + o_p(1). \tag{2.13}$$

With the lower bound $R_{1n}(\bar{\alpha}, \bar{\boldsymbol{\theta}}_1, \bar{\boldsymbol{\theta}}_2) \geq 2c$, the above upper bound implies that $\bar{\mathbf{m}} = O_p(n^{-1/2})$. Let $\bar{\alpha} \in [\delta, 0.5]$ for some $\delta \in (0, 0.5]$, then

$$\bar{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_0 = O_p(n^{-1/4}),\ \bar{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_0 = O_p(n^{-1/4}),$$

and $\bar{\mathbf{m}}_1 = O_p(n^{-1/2})$.

We now proceed to prove (2.8). We expand the remainder term to order 5

and get

$$
\begin{aligned}
&\sum_{i=1}^{n}\epsilon_{in}\\
=&\sum_{i=1}^{n}\sum_{j_1,j_2,j_3=1}^{d}(1-\bar{\alpha})\prod_{s=1}^{3}(\bar{\theta}_{1j_s}-\theta_{0j_s})\frac{\partial^3 f(X_i;\boldsymbol{\theta}_0)/\partial\theta_{j_1}\partial\theta_{j_2}\partial\theta_{j_3}}{3!\,f(X_i;\boldsymbol{\theta}_0)}\\
&+\sum_{i=1}^{n}\sum_{j_1,j_2,j_3=1}^{d}\bar{\alpha}\prod_{s=1}^{3}(\bar{\theta}_{2j_s}-\theta_{0j_s})\frac{\partial^3 f(X_i;\boldsymbol{\theta}_0)/\partial\theta_{j_1}\partial\theta_{j_2}\partial\theta_{j_3}}{3!\,f(X_i;\boldsymbol{\theta}_0)}\\
&+\sum_{i=1}^{n}\sum_{j_1,\ldots,j_4=1}^{d}(1-\bar{\alpha})\prod_{s=1}^{4}(\bar{\theta}_{1j_s}-\theta_{0j_s})\frac{\partial^4 f(X_i;\boldsymbol{\theta}_0)/\partial\theta_{j_1}\cdots\partial\theta_{j_4}}{4!\,f(X_i;\boldsymbol{\theta}_0)}\\
&+\sum_{i=1}^{n}\sum_{j_1,\ldots,j_4=1}^{d}\bar{\alpha}\prod_{s=1}^{4}(\bar{\theta}_{2j_s}-\theta_{0j_s})\frac{\partial^4 f(X_i;\boldsymbol{\theta}_0)/\partial\theta_{j_1}\cdots\partial\theta_{j_4}}{4!\,f(X_i;\boldsymbol{\theta}_0)}\\
&+\sum_{i=1}^{n}\sum_{j_1,\ldots,j_5=1}^{d}(1-\bar{\alpha})\prod_{s=1}^{5}(\bar{\theta}_{1j_s}-\theta_{0j_s})\frac{\partial^5 f(X_i;\boldsymbol{\xi}_1)/\partial\theta_{j_1}\cdots\partial\theta_{j_5}}{5!\,f(X_i;\boldsymbol{\theta}_0)}\\
&+\sum_{i=1}^{n}\sum_{j_1,\ldots,j_5=1}^{d}\bar{\alpha}\prod_{s=1}^{5}(\bar{\theta}_{2j_s}-\theta_{0j_s})\frac{\partial^5 f(X_i;\boldsymbol{\xi}_2)/\partial\theta_{j_1}\cdots\partial\theta_{j_5}}{5!\,f(X_i;\boldsymbol{\theta}_0)},
\end{aligned}
$$

where the $\boldsymbol{\xi}_j$ are between $\bar{\boldsymbol{\theta}}_j$ and $\boldsymbol{\theta}_0$, $j=1,2$. By Condition A4 and the consistency of $\bar{\boldsymbol{\theta}}_1$ and $\bar{\boldsymbol{\theta}}_2$, we further have

$$
\begin{aligned}
&\left|\sum_{i=1}^{n}\epsilon_{in}\right|\\
&=O_p(n^{1/2})(\|\bar{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_0\|^3+\|\bar{\boldsymbol{\theta}}_2-\boldsymbol{\theta}_0\|^3)+O_p(n^{1/2})(\|\bar{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_0\|^4+\|\bar{\boldsymbol{\theta}}_2-\boldsymbol{\theta}_0\|^4)\\
&\quad+O_p(n)(\|\bar{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_0\|^5+\|\bar{\boldsymbol{\theta}}_2-\boldsymbol{\theta}_0\|^5)\\
&=o_p(n^{1/2})(\|\bar{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_0\|^2+\|\bar{\boldsymbol{\theta}}_2-\boldsymbol{\theta}_0\|^2)+o_p(n)(\|\bar{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_0\|^4+\|\bar{\boldsymbol{\theta}}_2-\boldsymbol{\theta}_0\|^4) \quad (2.14)
\end{aligned}
$$

Since $\bar{\alpha} \in [\delta, 0.5]$ for some $0 < \delta \leq 0.5$, the first term in (2.14)

$$\begin{aligned} & o_p(n^{1/2})(\|\bar{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_0\|^2 + \|\bar{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_0\|^2) \\ = & o_p(n^{1/2})(\sum_{s=1}^{d} \bar{m}_{2s}) = o_p(n^{1/2})(||\bar{\mathbf{m}}||) \\ \leq & o_p(1) + o_p(n)||\bar{\mathbf{m}}||^2 \end{aligned}$$

and the second term in (2.14)

$$\begin{aligned} & o_p(n)(\|\bar{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_0\|^4 + \|\bar{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_0\|^4) \\ = & o_p(n)(\sum_{s=1}^{d} \bar{m}_{2s}^2) \\ = & o_p(n)||\bar{\mathbf{m}}||^2. \end{aligned}$$

Therefore,

$$\left| \sum_{i=1}^{n} \epsilon_{in} \right| = o_p(1) + o_p(n)||\bar{\mathbf{m}}||^2.$$

□

Now we show that under the null distribution, the EM-iteration changes the fitted value of α by only $o_p(1)$. Let $(\bar{\alpha}, \bar{\boldsymbol{\theta}}_1, \bar{\boldsymbol{\theta}}_2)$ be estimators of $(\alpha, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. Define

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

$R_n(\alpha) = \sum_{i=1}^{n} (1 - \bar{w}_i) \log(1-\alpha) + (\sum_{i=1}^{n} \bar{w}_i) \log \alpha$, and $Q_n(\alpha) = R_n(\alpha) + p(\alpha)$. The EM-algorithm updates α by searching for $\bar{\alpha}^* = \arg\max Q_n(\alpha)$.

Lemma 2.3: *Suppose the conditions in Lemma 2.1 hold and* $\bar{\alpha} - \alpha_0 = o_p(1)$ *for some* $\alpha_0 \in (0, 0.5]$. *Then under the null distribution* $f(x; \boldsymbol{\theta}_0)$, *we have*

$|\bar{\alpha}^* - \alpha_0| = o_p(1)$.

Proof. The proof is similar to that of Lemma A3 in Li, Chen, & Marriott (2009) and is therefore omitted. □

2.6.3 Proof of Theorem 2.1

With the above three technical lemmas, the proof is the same as that of Theorem 2.1 in Li, Chen, & Marriott (2009).

2.6.4 Proof of Theorem 2.2

Since the null model $f(x;\boldsymbol{\theta})$ is regular, the following classical expansion is applicable

$$\begin{aligned} R_{0n} &= 2\{pl_n(0.5, \hat{\boldsymbol{\theta}}_0, \hat{\boldsymbol{\theta}}_0) - pl_n(0.5, \boldsymbol{\theta}_0, \boldsymbol{\theta}_0)\} \\ &= \left(\sum_{i=1}^{n} \mathbf{b}_{1i}\right)^{\tau} (n\mathbf{B}_{11})^{-1} \left(\sum_{i=1}^{n} \mathbf{b}_{1i}\right) + o_p(1). \end{aligned}$$

Let $R_{1n}(\alpha_j^{(k)}, \boldsymbol{\theta}_{j,1}^{(k)}, \boldsymbol{\theta}_{j,2}^{(k)}) = 2\{pl_n(\alpha_j^{(k)}, \boldsymbol{\theta}_{j,1}^{(k)}, \boldsymbol{\theta}_{j,2}^{(k)}) - pl_n(0.5, \boldsymbol{\theta}_0, \boldsymbol{\theta}_0)\}$. With the results in Theorem 2.1, the upper bound in (2.13) is applicable. Hence,

$$R_{1n}(\alpha_j^{(k)}, \boldsymbol{\theta}_{j,1}^{(k)}, \boldsymbol{\theta}_{j,2}^{(k)}) \leq 2(\mathbf{m}^{(k)})^{\tau} \sum_{i=1}^{n} \mathbf{b}_i - n(\mathbf{m}^{(k)})^{\tau} \mathbf{B}(\mathbf{m}^{(k)})\{1 + o_p(1)\} + o_p(1),$$

where $\mathbf{m}^{(k)}$ is defined similarly to (2.6) with $(\bar{\alpha}, \bar{\boldsymbol{\theta}}_1, \bar{\boldsymbol{\theta}}_2)$ replaced by $(\alpha_j^{(k)}, \boldsymbol{\theta}_{j,1}^{(k)}, \boldsymbol{\theta}_{j,2}^{(k)})$. With the order assessment in Theorem 2.1, we further have

$$R_{1n}(\alpha_j^{(k)}, \boldsymbol{\theta}_{j,1}^{(k)}, \boldsymbol{\theta}_{j,2}^{(k)}) \leq 2(\mathbf{m}^{(k)})^{\tau} \sum_{i=1}^{n} \mathbf{b}_i - n(\mathbf{m}^{(k)})^{\tau} \mathbf{B}(\mathbf{m}^{(k)}) + o_p(1). \qquad (2.15)$$

To further explore the upper bound in (2.15), we re-parametrize as follows. By the definition of $m_{1h}^{(k)}$, we have

$$\theta_{j,2h}^{(k)} - \theta_{0h} = \frac{m_{1h}^{(k)} - (1-\alpha_j^{(k)})(\theta_{j,1h}^{(k)} - \theta_{0h})}{\alpha_j^{(k)}},\ h = 1, 2, \cdots, d.$$

Substituting the above equation into the definitions of $m_{2h}^{(k)}$ and $s_{hl}^{(k)}$, we obtain

$$\begin{aligned} m_{2h}^{(k)} &= \frac{1-\alpha_j^{(k)}}{\alpha_j^{(k)}} \left(\theta_{j,1h}^{(k)} - \theta_{0h}\right)^2 + o_p(m_{1h}^{(k)}), \\ s_{hl}^{(k)} &= \frac{1-\alpha_j^{(k)}}{\alpha_j^{(k)}} (\theta_{j,1h}^{(k)} - \theta_{0h})(\theta_{j,1l}^{(k)} - \theta_{0l}) + o_p(m_{1h}^{(k)}) + o_p(m_{1l}^{(k)}). \end{aligned}$$

Let

$$\begin{aligned} & v_h^{(k)} = \sqrt{\frac{1-\alpha_j^{(k)}}{\alpha_j^{(k)}}} \left(\theta_{j,1h}^{(k)} - \theta_{0h}\right), \\ & \mathbf{v}^{(k)} \\ &= \left((v_1^{(k)})^2 \ldots (v_d^{(k)})^2, v_1^{(k)} v_2^{(k)}, \ldots, v_1^{(k)} v_d^{(k)}, v_2^{(k)} v_3^{(k)}, \ldots, v_2^{(k)} v_d^{(k)}, \ldots, v_{d-1}^{(k)} v_d^{(k)}\right)^\tau, \end{aligned}$$

and $\mathbf{t}^{(k)} = ((\mathbf{m}_1^{(k)})^\tau, (\mathbf{v}^{(k)})^\tau)^\tau$ with $\mathbf{m}_1^{(k)}$ consisting of the first d elements of $\mathbf{m}^{(k)}$. Using the order assessment in Theorem 2.1, $\mathbf{m}^{(k)} = \mathbf{t}^{(k)} + o_p(n^{-1/2})$. Therefore,

$$R_{1n}(\alpha_j^{(k)}, \boldsymbol{\theta}_{j,1}^{(k)}, \boldsymbol{\theta}_{j,2}^{(k)}) \leq 2(\mathbf{t}^{(k)})^\tau \sum_{i=1}^{n} \mathbf{b}_i - n(\mathbf{t}^{(k)})^\tau \mathbf{B}(\mathbf{t}^{(k)}) + o_p(1).$$

Letting $\tilde{\mathbf{m}}_1^{(k)} = \mathbf{m}_1^{(k)} + \mathbf{B}_{11}^{-1}\mathbf{B}_{12}\mathbf{v}^{(k)}$, we get

$$2(\mathbf{t}^{(k)})^\tau \sum_{i=1}^{n} \mathbf{b}_i - n(\mathbf{t}^{(k)})^\tau \mathbf{B}(\mathbf{t}^{(k)}) = 2(\tilde{\mathbf{m}}_1^{(k)})^\tau \sum_{i=1}^{n} \mathbf{b}_{1i} - n(\tilde{\mathbf{m}}_1^{(k)})^\tau \mathbf{B}_{11}(\tilde{\mathbf{m}}_1^{(k)}) + 2(\mathbf{v}^{(k)})^\tau \sum_{i=1}^{n} \tilde{\mathbf{b}}_{2i} - n(\mathbf{v}^{(k)})^\tau \tilde{\mathbf{B}}_{22}(\mathbf{v}^{(k)}).$$

Hence,

$$R_{1n}(\alpha_j^{(k)}, \boldsymbol{\theta}_{j,1}^{(k)}, \boldsymbol{\theta}_{j,2}^{(k)}) \leq 2(\tilde{\mathbf{m}}_1^{(k)})^\tau \sum_{i=1}^{n} \mathbf{b}_{1i} - n(\tilde{\mathbf{m}}_1^{(k)})^\tau \mathbf{B}_{11}(\tilde{\mathbf{m}}_1^{(k)}) + 2(\mathbf{v}^{(k)})^\tau \sum_{i=1}^{n} \tilde{\mathbf{b}}_{2i} - n(\mathbf{v}^{(k)})^\tau \tilde{\mathbf{B}}_{22}(\mathbf{v}^{(k)}) + o_p(1).$$

For $M_n^{(k)}(\alpha_j)$, we have

$$\begin{aligned} M_n^{(k)}(\alpha_j) &= R_{1n}(\alpha_j^{(k)}, \boldsymbol{\theta}_{j,1}^{(k)}, \boldsymbol{\theta}_{j,2}^{(k)}) - R_{0n} \\ &\leq \sup_{\mathbf{m}_1}\{2\mathbf{m}_1{}^\tau \sum_{i=1}^{n} \mathbf{b}_{1i} - n\mathbf{m}_1^\tau \mathbf{B}_{11}\mathbf{m}_1\} + \sup_{\mathbf{v}}\{2\mathbf{v}^\tau \sum_{i=1}^{n} \tilde{\mathbf{b}}_{2i} - n\mathbf{v}^\tau \tilde{\mathbf{B}}_{22}\mathbf{v}\} \\ &\quad - \left(\sum_{i=1}^{n} \mathbf{b}_{1i}\right)^\tau (n\mathbf{B}_{11})^{-1} \left(\sum_{i=1}^{n} \mathbf{b}_{1i}\right) + o_p(1) \\ &\leq \sup_{\mathbf{v}}\{2\mathbf{v}^\tau \sum_{i=1}^{n} \tilde{\mathbf{b}}_{2i} - n\mathbf{v}^\tau \tilde{\mathbf{B}}_{22}\mathbf{v}\} + o_p(1). \end{aligned}$$

Here $\mathbf{v} = (v_1^2, \ldots, v_d^2, v_1v_2, \ldots, v_1v_d, v_2v_3, \ldots, v_2v_d, \ldots, v_{d-1}v_d)^\tau$. The leading term in the above equation does not depend on α and therefore it also serves as the upper bound of $EM_n^{(K)}$.

We now show that the upper bound is achievable. Since the EM-iteration increases the modified likelihood (Dempster, Laird, & Rubin, 1977), we only need to show that this is the case when $k = 1$. To prove this for $k = 1$, we only need to find a set of parameter values $\hat{\alpha}, \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2$ at which the upper bound

is attained. We first calculate

$$\hat{\mathbf{v}} = \arg\sup_{\mathbf{v}} \{2\mathbf{v}^\tau \sum_{i=1}^{n} \tilde{\mathbf{b}}_{2i} - n\mathbf{v}^\tau \tilde{\mathbf{B}}_{22}\mathbf{v}\}$$

and

$$\hat{\mathbf{m}}_1 = \left(\sum_{i=1}^{n} \mathbf{b}_{1i}\right)^\tau (n\mathbf{B}_{11})^{-1} \left(\sum_{i=1}^{n} \mathbf{b}_{1i}\right) + \mathbf{B}_{11}^{-1}\mathbf{B}_{12}\hat{\mathbf{v}}.$$

Next we choose $\hat{\alpha} = 0.5$, then determine $\hat{\boldsymbol{\theta}}_1$ by the equations

$$\hat{v}_h = \sqrt{\frac{1-\hat{\alpha}}{\hat{\alpha}}} \left(\hat{\theta}_{1h} - \theta_{0h}\right), \ h = 1, \ldots, d$$

and $\hat{\boldsymbol{\theta}}_2$ by the equation

$$\hat{\mathbf{m}}_1 = (1-\hat{\alpha})(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_0) + \hat{\alpha}(\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_0).$$

The existence of $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ is obvious. It can be shown that

$$\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_0 = O_p(n^{-1/4}), \ \hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_0 = O_p(n^{-1/4}).$$

This order assessment can be used to show that

$$\begin{aligned} EM_n^{(K)} \geq M_n^{(1)}(0.5) &\geq R_{1n}(0.5, \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2) - R_{0n} \\ &= \sup_{\mathbf{v}} \{2\mathbf{v}^\tau \sum_{i=1}^{n} \tilde{\mathbf{b}}_{2i} - n\mathbf{v}^\tau \tilde{\mathbf{B}}_{22}\mathbf{v}\} + o_p(1). \end{aligned}$$

Combining the lower and upper bounds of $EM_n^{(K)}$, we have

$$\begin{aligned} EM_n^{(K)} &= \sup_{\mathbf{v}}\{2\mathbf{v}^\tau \sum_{i=1}^{n} \tilde{\mathbf{b}}_{2i} - n\mathbf{v}^\tau \tilde{\mathbf{B}}_{22}\mathbf{v}\} + o_p(1) \\ &= \sup_{\mathbf{v}}\{2\mathbf{v}^\tau \tilde{\mathbf{B}}_{22}\tilde{\mathbf{B}}_{22}^{-1} \sum_{i=1}^{n} \tilde{\mathbf{b}}_{2i} - n\mathbf{v}^\tau \tilde{\mathbf{B}}_{22}\mathbf{v}\} + o_p(1). \end{aligned}$$

Note that $n^{-1/2}\tilde{\mathbf{B}}_{22}^{-1} \sum_{i=1}^{n} \tilde{\mathbf{b}}_{2i}$ asymptotically follows the zero-mean multivariate normal distribution with variance-covariance matrix $\tilde{\mathbf{B}}_{22}^{-1}$. Then

$$EM_n^{(K)} \to \sup_{\mathbf{v}}(2\mathbf{v}^\tau \tilde{\mathbf{B}}_{22}\mathbf{w} - \mathbf{v}^\tau \tilde{\mathbf{B}}_{22}\mathbf{v})$$

in distribution, where $\mathbf{w}$ is a random vector from the zero-mean multivariate normal distribution with variance-covariance matrix $\tilde{\mathbf{B}}_{22}^{-1}$ and $\mathbf{v}$ still takes the form $(v_1^2, \ldots, v_d^2, v_1v_2, \ldots, v_1v_d, v_2v_3, \ldots, v_2v_d, \ldots, v_{d-1}v_d)^\tau$. After some simple algebraic work, we have

$$EM_n^{(K)} \to \sup_{\mathbf{v}}(2\mathbf{v}^\tau \tilde{\mathbf{B}}_{22}\mathbf{w} - \mathbf{v}^\tau \tilde{\mathbf{B}}_{22}\mathbf{v}) = \mathbf{w}^\tau \tilde{\mathbf{B}}_{22}\mathbf{w} - \inf_{\mathbf{v}}(\mathbf{w}-\mathbf{v})^\tau \tilde{\mathbf{B}}_{22}(\mathbf{w}-\mathbf{v}).$$

The right-hand side of the above equation is just the distribution of the LRT for testing $\mathbf{v} = 0$ based on one observation $\mathbf{w}$ from the multivariate normal distribution with mean $\mathbf{v}$ and variance-covariance matrix $\tilde{\mathbf{B}}_{22}^{-1}$. This finishes the proof for Theorem 2.2. □

Chapter 3

EM-test in a Scale Mixture of Normal Models[1]

3.1 Introduction and Motivating Example

The class of scale mixtures of normal distributions, i.e., mixtures of normal distributions on the variance parameters, contains many continuous unimodal and symmetric distributions on the real line, such as the Student t family, the logistic distribution, the Laplace distribution, and the stable family; see Andrews & Mallows (1974), West (1984), and West (1987). Since it has heavier tails than the normal family, this class serves as a natural alternative to the normal family when heavy tails are observed in the data. Naylor & Smith (1983) applied a two-component scale mixtures of normal distributions to model the heavy tail and heterogeneity in some biochemical measurements in clinical chemistry. Wainwright & Simoncelli (2000) and Doulgeris & Eltoft (2010) used scale mixtures of normal distributions in the analysis of image data. The

[1]A version of this chapter is ready for submission.

model has also played an important role in finance and economics. It has been used to explain the heavy tail in future price movements (Hall, Brorsen, & Irwin, 1989) and daily changes in the logarithm of exchange rates (Boothe & Glassman, 1987; Rocha, Pestana, & Menezes, 2012). It has also been widely used in the analysis of data with outliers (West, 1984 and references therein).

In practice, before one tries to apply a scale mixtures of normal distributions, it is helpful to know whether the data arise from a homogeneous or heterogeneous population. Applying the model to data from a homogeneous population may result in an efficiency loss when estimating the unknown parameters. We present below a motivating example showing the application of a scale mixtures of normal distributions.

Example 3.1. (Blood chloride data) *In clinical chemistry, the clinical assessment of biochemical measurements is typically carried out by reference to a "normal range," which is the 95% confidence interval of the mean measurement for a "healthy" population (Naylor & Smith, 1983). One way of obtaining such a normal range is to first collect a large sample of biochemical measurements from a healthy population. However, in practice, it may be difficult to collect measurements only from healthy individuals. Instead, measurements from a contaminated sample, containing both healthy and unhealthy individuals, are obtained. Because of the potential existence of heterogeneity in the contaminated sample, mixtures of normal distributions are widely used in such analyses.*

Naylor & Smith (1983) used a scale mixture of two normal distributions to model a contaminated sample of 542 blood-chloride measurements collected during routine analysis by the Department of Chemical Pathology at the Derbyshire Royal Infirmary. Based on this mixture model, they derived the normal

range for the healthy population. A statistical problem of interest here is to test whether unhealthy individuals really exist. If not, the normal range derived from a scale mixtures of normal distributions may not be as accurate as that derived from a homogeneous normal distribution.

The design of an effective method for testing homogeneity is a challenging problem for scale mixtures of normal distributions. The likelihood function is unbounded (Hathaway, 1985; Chen, Tan, & Zhang, 2008), and the Fisher information on the mixing proportion direction can be infinity (Chen & Li, 2009). Because of these two irregularities, many elegant asymptotic results for existing methods such as the LRT, the MLRT, and the D-test (Charnigo & Sun, 2004) cannot be directly applied unless the parameter spaces for the mean and variance are constrained.

In Chen & Li (2009), a class of EM-tests were proposed for testing homogeneity in normal mixture models on the mean parameters and in normal mixture models on both the mean and variance parameters. In these two cases, the EM-tests have a χ^2-type limiting distribution without any constraints on the parameter spaces for the mean and variance. In principle, the EM-test for testing homogeneity in normal mixtures on both the mean and variance parameters can be applied to scale mixtures of normal distributions. However, a tailor-made testing procedure specifically for the scale mixtures of normal distributions is expected to be more powerful.

In this chapter, we retool the EM-test for testing homogeneity in scale mixtures of normal distributions. The retooled method first applies a penalty function on the component variances to obtain a bounded penalized log-likelihood. Based on this penalized log-likelihood, we define the EM-test statistics and show that the limiting distribution of the retooled EM-test is $0.5\chi^2_0 + 0.5\chi^2_1$.

The penalty function contains a tuning parameter that affects the precision of the test. We use a computational method to provide an easy-to-use empirical value for the tuning parameter. Simulation studies show that the retooled EM-test has an accurate size. When the data is generated from the scale mixtures of normal distributions, the EM-test is more powerful than the likelihood ratio test and the method in Chen & Li (2009). When the data is generated from a normal mixture on both means and variances, the EM-test is again more powerful than the likelihood ratio test in most situations and is more powerful than or comparable to the method in Chen & Li (2009) when the sample size is small or the mixing proportion in one component is small. Software implementing the test has also been developed in the R language (R Development Core Team, 2011).

The rest of this chapter is organized as follows. In Section 3.2, we present the EM-test procedure, its asymptotic properties, and the empirical value for the tuning parameter. In Section 3.3, we present simulation studies, and in Section 3.4, we apply the retooled EM-test to two real-data examples. For convenience of presentation, all the proofs are given in Section 3.5.

3.2 Main Results

Suppose $X_1, \ldots, X_n$ is a random sample of size n from a scale mixture of two normal distributions

$$(1-\alpha)f(x;\mu,\sigma_1) + \alpha f(x;\mu,\sigma_2).$$

Here $f(x;\mu,\sigma)$ denotes the probability density function of the normal distribution $N(\mu,\sigma^2)$ with mean μ and variance σ^2. We aim to test

$$H_0 : \alpha(1-\alpha)(\sigma_1-\sigma_2) = 0.$$

Without loss of generality, we assume $\alpha \in [0, 0.5]$.

3.2.1 The EM-test statistic

We denote the log-likelihood function as

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

and define the penalized log-likelihood function as

$$pl_n(\alpha,\mu,\sigma_1,\sigma_2) = l_n(\alpha,\mu,\sigma_1,\sigma_2) + p_n(\sigma_1) + p_n(\sigma_2) + p(\alpha).$$

The penalty function $p(\alpha)$ is the same as the one used in Chapter 2. We still use $p(\alpha) = \log(1-|1-2\alpha|)$. More discussions can be found in Section 2.5. The smooth-penalty function $p_n(\sigma)$ goes to negative infinity when σ goes to either 0 or infinity. We recommend

$$p_n(\sigma) = -a_n\{s_n^2/\sigma^2 + \log(\sigma^2/s_n^2)\}$$

with $s_n^2 = n^{-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$ and $\overline{X} = n^{-1}\sum_{i=1}^{n} X_i$. The choice of the tuning parameter a_n will be discussed in Section 3.2.3. The penalty $p_n(\sigma)$ prevents the fittings of σ_1 and σ_2 from being close to zero, which results in the bounded penalized log-likelihood.

Similar to Chapter 2, the EM-test statistics is constructed in the following iterative way. We first choose a finite set of $\{\alpha_1, \ldots, \alpha_J\} \subset (0, 0.5]$ and a positive integer K. As suggested in Chapter 2, we use $\{\alpha_1, \ldots, \alpha_J\} = \{0.1, 0.3, 0.5\}$ and $K = 2$ or 3.

For each $j = 1, 2, \ldots, J$, we proceed as follows. Let

$$(\mu_j^{(1)}, \sigma_{j,1}^{(1)}, \sigma_{j,2}^{(1)}) = \arg\max_{\mu,\sigma_1,\sigma_2} pl_n(\alpha_j, \mu, \sigma_1, \sigma_2).$$

Further let $k = 1$ and $\alpha_j^{(1)} = \alpha_j$.

Then we update $(\alpha, \mu, \sigma_1, \sigma_2)$ by using the EM-iteration $K - 1$ times. In each iteration, for $i = 1, \ldots, n$, and the current k, we first use an E-step to calculate the posterior probabilities,

$$w_{ij}^{(k)} = \frac{\alpha_j^{(k)} f(X_i; \mu_j^{(k)}, \sigma_{j,2}^{(k)})}{(1 - \alpha_j^{(k)}) f(X_i; \mu_j^{(k)}, \sigma_{j,1}^{(k)}) + \alpha_j^{(k)} f(X_i; \mu_j^{(k)}, \sigma_{j,2}^{(k)})}.$$

Then we use an M-step to update α via

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

and $(\mu, \sigma_1, \sigma_2)$ via

$$\Big(\mu_j^{(k+1)}, \sigma_{j,1}^{(k+1)}, \sigma_{j,2}^{(k+1)}\Big) = \arg\max_{\mu,\sigma_1,\sigma_2} \Big\{ \sum_{i=1}^n (1 - w_{ij}^{(k)}) \log f(X_i; \mu, \sigma_1) + \sum_{i=1}^n w_{ij}^{(k)} \log f(X_i; \mu, \sigma_2) + p_n(\sigma_1) + p_n(\sigma_2) \Big\}.$$

Let $k = k + 1$. The EM-iteration continues until $k = K$.

Define the test statistics

$$M_n^{(K)}(\alpha_j) = 2\{pl_n(\alpha_j^{(K)}, \mu_j^{(K)}, \sigma_{j,1}^{(K)}, \sigma_{j,2}^{(K)}) - pl_n(0.5, \hat{\mu}_0, \hat{\sigma}_0, \hat{\sigma}_0)\}$$

where $(\hat{\mu}_0, \hat{\sigma}_0) = \arg\max_{\mu,\sigma} pl_n(0.5, \mu, \sigma, \sigma)$. The EM-test statistic is then defined as

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

We reject the null hypothesis when $EM_n^{(K)}$ exceeds some critical value of the limiting distribution presented in Section 3.2.2.

3.2.2 Asymptotic distribution

The asymptotic distribution of $EM_n^{(K)}$ is obtained via a careful choice of the two penalty functions $p(\alpha)$ and $p_n(\sigma)$.

C1 The penalty function $p(\alpha)$ is continuous, maximized at $\alpha = 0.5$, and approaches negative infinity as α approaches 0 or 1.

C2 $\sup_{\sigma>0}\{|p_n(\sigma)|\} = o(n)$.

C3 $p_n'(\sigma) = o_p(n^{1/6})$ at any $\sigma > 0$.

C4 $p_n(\sigma) \leq 4(\log n)^2 \log(\sigma)$, when $\sigma \leq n^{-1}$ and n is large.

The penalty functions recommended in Section 3.2.1 satisfy Conditions C1–C4. Since the user has the freedom to choose the penalty functions, these conditions are not restrictive as long as such functions exist. The following theorem presents the limiting distribution of $EM_n^{(K)}$; the proof is given in Section 3.5.

Theorem 3.1: *Suppose that the penalty functions $p(\alpha)$ and $p_n(\sigma)$ satisfy Conditions C1–C4 and that $\alpha_1 = 0.5$. Under the null hypothesis and for any fixed finite K, as $n \to \infty$,*

$$EM_n^{(K)} \to \frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$$

in distribution.

The limiting distribution of the EM-test is quite simple and can be conveniently used to calculate the asymptotic p-values of the EM-test statistic. Its approximation to the finite-sample distribution of the EM-test statistic will be examined in Section 3.3 through a simulation study.

3.2.3 Choice of tuning parameters

The definition of $EM_n^{(K)}$ involves a few tuning parameters: the finite set $\{\alpha_1, \ldots, \alpha_J\}$ for α, the iteration number K, and the penalty functions $p(\alpha)$ and $p_n(\sigma)$. As suggested in Chapter 2, we recommend using the initial values $\alpha \in \{0.1, 0.3, 0.5\}$, $K = 2$ or 3 iterations, and $p(\alpha) = \log(1 - |1 - 2\alpha|)$. For the penalty function $p_n(\sigma)$, we recommend

$$p_n(\sigma) = -a_n\{s_n^2/\sigma^2 + \log(\sigma^2/s_n^2)\}.$$

The penalty function $p_n(\sigma)$ satisfies Conditions C2–C4 with $a_n = o(n^{1/6})$. That is, the specific choice of a_n will not affect the limiting distribution. Carefully tuning the value of a_n will improve the precision of the approximation of the limiting distribution to the finite-sample distribution of $EM_n^{(K)}$.

Ideally, we wish to choose the value of a_n such that the limiting distribution

is the same as the finite-sample distribution for all $x > 0$, that is,

$$\Pr(EM_n^{(K)} \geq x) = 0.5\Pr(\chi_1^2 \geq x).$$

Unfortunately, such an a_n may not exist. Let q, usually 5%, be a given significance level and x_{2q} be the $1-2q$ upper quantile of the χ_1^2 distribution, which is also the $1-q$ upper quantile of the limiting distribution $0.5\chi_0^2 + 0.5\chi_1^2$. We then consider choosing the value of a_n such that

$$\Pr(EM_n^{(K)} \geq x_{2q}) = 0.5\Pr(\chi_1^2 \geq x_{2q}).$$

That is, we choose the value of a_n such that the $1-q$ quantiles of the finite-sample distribution and the limiting distribution of $EM_n^{(K)}$ are the same. Theoretically, this can be challenging even though q is given. Instead, we developed an empirical value for a_n through a computational method (Chen & Li, 2011).

The idea is as follows. First, we choose several representative sample sizes n and a_n values. Next, for each combination of n and a_n, we generate M random samples of size n from the null distribution $N(0,1)$ and record the percentage of times that the EM-test statistic $EM_n^{(K)}$ is greater than or equal to x_{2q}. We denote this observed percentage, also called the simulated type-I error rate, by $\hat{q}$. Last, we define a discrepancy between the simulated type-I error and the target significance level q as

$$y = \log\{\hat{q}/(1-\hat{q})\} - \log\{q/(1-q)\}.$$

We then fit a regression model between y and (n, a_n). Setting the fitted value $\hat{y}$ equal to 0, we then obtain an empirical value for a_n.

In our computer experiment, we set $q = 0.05$ because this is the most commonly used significance level. In the simulation study, we also checked the precision of the EM-test at the 1% level. We consider three sample sizes, $n \in \{100, 300, 500\}$, and six a_n values, $a_n \in \{1/128, 1/64, 1/32, 1/16, 1/8, 1/4\}$. Hence, a 3×6 full factorial design is used in our computer experiment. For each combination of n and a_n we generate 2500 random samples from $N(0, 1)$, and for each random sample we use $K = 2$ to calculate the EM-test statistic $EM_n^{(K)}$. The discrepancies between $\hat{q}$ and q in terms of $y = \log\{\hat{q}/(1-\hat{q})\} - \log\{q/(1-q)\}$ with $q = 0.05$ are reported in Table 3.1.

Table 3.1: Discrepancy between $\hat{q}$ and q in terms of y under the scale mixtures of normal distributions.

	a_n=1/128	a_n=1/64	a_n=1/32	a_n=1/16	a_n=1/8	a_n=1/4
$n = 100$	-0.021	0.021	0.041	-0.065	-0.134	-0.402
$n = 300$	0.101	-0.021	-0.134	0.101	-0.088	-0.287
$n = 500$	0	0.081	0.041	-0.043	-0.158	-0.315

By analysis of variance, we observe that only a_n has a significant effect on the response y. After some exploratory analysis, the covariate in the form of a_n itself gives the most satisfactory outcome. We next fit a linear regression between y and a_n. Based on the 18 observations, the fitted model is

$$\hat{y} = 0.05194 - 1.5019 a_n$$

with $R^2 = 81.1\%$. Setting $\hat{y} = 0$ gives the empirical value $a_n = 0.035$.

Note that since our method is invariant to the scale transformation, the value $a_n = 0.035$ is applicable to the general null distribution $N(\mu, \sigma^2)$. The performance of the EM-test with a_n=0.035 will be examined in the next section

through a simulation study.

3.3 Simulation Study

The purpose of the simulation study is twofold: (i) to check the approximation of the limiting distribution to the finite sample distribution; and (ii) to examine the power of the retooled EM-test. The EM-test is calculated based on the recommendations for $\{\alpha_1, \ldots, \alpha_J\}$, K, and the two penalty functions $p(\alpha)$ and $p_n(\sigma)$ in Section 3.2.3.

We consider the null model $N(0,1)$ since $EM_n^{(K)}$ is invariant to the scale transformation. For the null model $N(0,1)$ and the 11 sample sizes $50, 100, \ldots,$ 1000, we calculate the simulated type-I error rates based on 10000 repetitions. The simulation results for two significance levels, 5% and 1%, are summarized in Table 3.2. For comparison, we further conducted the simulation for the EM-test in Chen & Li (2009), designed for homogeneity under a normal mixture on the mean and variance parameters. We use $\widetilde{EM}_n^{(K)}$ to denote this EM-test in Chen & Li (2009). The same null model and sample sizes are considered. The results for 5% and 1% significance levels of $\widetilde{EM}_n^{(K)}$ are also tabulated in Table 3.2.

From the results in Table 3.2, we observe that our proposed approach has comparable accuracy with that in Chen & Li (2009). The simulated type-I error rates of both methods are quite close to the significance level for all the considered sample sizes. Hence, we conclude that the limiting distribution provides an accurate approximation of the finite-sample distribution for both tests.

To examine the power of the retooled EM-test, we choose eight alternative

Table 3.2: Simulated type-I error rates (%) of the retooled EM-test and the EM-test in Chen & Li (2009) under the scale mixtures of normal distributions.

	Retooled approach			Method in Chen & Li (2009)		
n	$EM_n^{(1)}$	$EM_n^{(2)}$	$EM_n^{(3)}$	$\widetilde{EM}_n^{(1)}$	$\widetilde{EM}_n^{(2)}$	$\widetilde{EM}_n^{(3)}$
Level=5%						
50	5.3	5.4	5.4	4.6	4.8	4.8
100	5.2	5.2	5.3	5.2	5.2	5.3
200	4.8	4.8	4.9	5.5	5.5	5.6
300	4.7	4.7	4.8	5.4	5.4	5.4
400	4.9	4.9	4.9	5.5	5.5	5.5
500	4.8	4.8	4.8	5.5	5.5	5.5
600	4.8	4.9	4.9	5.5	5.5	5.5
700	4.6	4.6	4.6	5.4	5.4	5.4
800	4.7	4.8	4.8	5.6	5.6	5.6
900	4.8	4.8	4.8	5.4	5.4	5.5
1000	4.7	4.7	4.7	5.3	5.3	5.3
Level=1%						
50	1.2	1.3	1.3	0.8	0.9	0.9
100	1.1	1.1	1.1	1.0	1.0	1.0
200	1.0	1.0	1.1	1.1	1.1	1.1
300	0.9	1.0	1.0	1.1	1.1	1.1
400	1.1	1.2	1.2	1.2	1.2	1.2
500	1.1	1.1	1.1	1.1	1.1	1.2
600	1.2	1.2	1.2	1.2	1.2	1.2
700	1.1	1.1	1.1	1.2	1.2	1.2
800	1.1	1.1	1.1	1.1	1.1	1.1
900	1.0	1.0	1.0	1.2	1.2	1.2
1000	1.0	1.0	1.0	1.1	1.1	1.1

models in the form of $(1-\alpha)N(\mu,\sigma_1^2)+\alpha N(\mu,\sigma_2^2)$, as well as another eight models in which the means in two components are different. The parameter settings and their Kullback–Leibler (KL) information with respect to the null model are listed in Table 3.3 and Table 3.4. We compare the power of the retooled EM-test $EM_n^{(K)}$ with the likelihood ratio test (LRT) and the EM-test $\widetilde{EM}_n^{(K)}$ in Chen & Li (2009). Since the limiting distribution of the LRT is unknown when the parameter spaces of the mean and variance are unconstrained, the quantiles of the LRT are obtained from 10000 repetitions under the null model. Further, to avoid an unbounded log-likelihood, we add a penalty on σ_1 and σ_2 to the log-likelihood to obtain the maximum likelihood estimator of the unknown parameters.

Table 3.3: First set of eight alternative models $(1-\alpha)N(\mu,\sigma_1^2)+\alpha N(\mu,\sigma_2^2)$.

Model	α	μ	σ_1	σ_2	100KL
A1	0.5	0	0.5	1.1	2.502
A2	0.25	0	0.5	1.05	2.392
A3	0.1	0	0.5	1.25	2.490
A4	0.05	0	0.5	1.55	2.573
A5	0.5	0	0.5	1.2	3.418
A6	0.25	0	0.5	1.15	3.571
A7	0.1	0	0.5	1.4	3.935
A8	0.05	0	0.5	1.85	4.891

In the comparison, we consider three sample sizes: $n = 50, 100, 200$. For each model and sample size, we calculate the power of each test at the 5% level based on 5000 repetitions. The simulation results are summarized in Table 3.5 and Table 3.6. As expected, in the first set of eight models, the retooled EM-test is more powerful than that in Chen & Li (2009) for detecting heterogeneity in scale mixtures of normal distributions. The retooled EM-test

Table 3.4: Second set of eight alternative models $(1-\alpha)N(\mu_1, \sigma_1^2)+\alpha N(\mu_2, \sigma_2^2)$.

Model	α	μ_1	μ_2	σ_1	σ_2	100KL
A9	0.5	0	0.5	0.5	1.1	2.875
A10	0.25	0	0.5	0.5	1.05	3.095
A11	0.1	0	0.5	0.5	1.25	3.045
A12	0.05	0	0.5	0.5	1.55	2.950
A13	0.5	0	0.5	0.5	1.2	3.761
A14	0.25	0	0.5	0.5	1.15	4.331
A15	0.1	0	0.5	0.5	1.4	4.604
A16	0.05	0	0.5	0.5	1.85	5.392

is comparable to the LRT and is sometimes more powerful than the LRT for detecting heterogeneity in scale mixtures of normal distributions. Compared with the LRT, another advantage of the retooled EM-test is that we do not need to use the bootstrap method to calculate the quantiles.

In the second set of eight models, there is a 0.5-difference in two component means. The results in Table 3.6 show that our proposed method is quite robust against the existence of mean difference. In all the eight models, the retooled EM-test has adequate powers. With sample sizes $n = 50$ and 100, the retooled EM-test has even higher power than that in Chen & Li (2009). With $n = 200$, the retooled EM-test is slightly less powerful than that in Chen & Li (2009) under models A9 and A13, where two mixing proportions are equal to 0.5. In other models, the retooled EM-test is slightly more powerful. Again, the retooled EM-test has comparable power to the LRT.

Table 3.5: Comparison of powers (%) of the retooled EM-test, the EM-test in Chen & Li (2009), and the LRT at the 5% level in the first set of eight alternative scale mixtures of normal distributions.

	Retooled approach				Method in Chen & Li (2009)		
Model	$EM_n^{(1)}$	$EM_n^{(2)}$	$EM_n^{(3)}$	LRT	$\widetilde{EM}_n^{(1)}$	$\widetilde{EM}_n^{(2)}$	$\widetilde{EM}_n^{(3)}$
				$n = 50$			
A1	42.8	42.7	42.5	37.9	25.2	25.1	25.2
A2	41.4	41.3	41.2	40.3	27.7	27.4	27.2
A3	42.4	42.3	42.3	39.7	31.2	31.0	31.0
A4	40.2	40.1	40.0	40.8	31.1	31.0	31.0
A5	53.7	53.7	53.5	48.0	31.2	31.0	30.9
A6	52.9	52.7	52.6	50.2	35.7	35.7	35.6
A7	52.2	52.2	52.0	51.6	41.9	41.9	41.8
A8	49.5	49.6	49.5	51.1	43.4	43.2	43.2
				$n = 100$			
A1	69.9	69.9	69.8	63.5	47.6	47.6	47.6
A2	67.3	67.3	67.1	63.2	47.6	47.6	47.6
A3	66.6	66.6	66.6	64.3	51.9	51.9	51.9
A4	63.6	63.6	63.5	62.2	53.0	53.1	53.1
A5	79.7	79.7	79.6	74.3	58.8	58.8	58.7
A6	81.4	81.4	81.3	77.7	63.7	63.6	63.6
A7	77.5	77.4	77.4	76.8	66.9	66.8	66.8
A8	75.1	75.1	75.1	75.1	67.4	67.2	67.1
				$n = 200$			
A1	92.1	92.2	92.1	89.4	78.5	78.5	78.5
A2	91.3	91.3	91.3	88.6	78.7	78.7	78.6
A3	88.3	88.3	88.2	88.4	79.0	79.0	79.0
A4	85.0	85.0	85.0	85.4	77.8	77.8	77.8
A5	96.9	97.0	96.9	95.9	90.1	90.0	90.0
A6	97.4	97.4	97.4	95.7	91.7	91.6	91.6
A7	95.3	95.3	95.2	94.7	91.5	91.4	91.4
A8	93.2	93.2	93.2	93.0	89.6	89.5	89.5

Table 3.6: Comparison of powers (%) of the retooled EM-test, the EM-test in Chen & Li (2009), and the LRT at the 5% level in the second set of eight alternative scale mixtures of normal distributions.

	Retooled approach				Method in Chen & Li (2009)		
Model	$EM_n^{(1)}$	$EM_n^{(2)}$	$EM_n^{(3)}$	LRT	$\widetilde{EM}_n^{(1)}$	$\widetilde{EM}_n^{(2)}$	$\widetilde{EM}_n^{(3)}$
				$n = 50$			
A9	40.5	40.6	40.3	36.9	36.4	36.4	36.3
A10	48.8	48.9	48.8	45.1	39.3	39.0	39.0
A11	44.3	44.2	44.2	49.7	36.3	36.2	36.2
A12	42.6	42.4	42.4	42.8	36.8	36.8	36.8
A13	48.6	48.5	48.5	41.5	42.8	42.6	42.6
A14	59.3	59.3	59.0	54.8	49.3	49.0	49.0
A15	58.6	58.4	58.3	56.9	47.7	47.6	47.6
A16	53.4	53.4	53.4	54.6	46.7	46.4	46.3
				$n = 100$			
A9	62.8	62.8	62.7	56.1	62.2	62.4	62.5
A10	74.2	74.0	74.0	69.5	68.0	68.1	68.1
A11	72.6	72.6	72.5	70.7	62.6	62.6	62.6
A12	65.8	65.7	65.6	66.9	58.0	58.0	58.0
A13	74.3	74.4	74.1	67.9	73.2	73.3	73.3
A14	84.8	84.8	84.7	81.8	78.5	78.5	78.5
A15	82.2	82.2	82.0	81.8	73.1	73.1	73.1
A16	77.6	77.6	77.6	76.3	70.8	70.9	70.9
				$n = 200$			
A9	85.5	85.6	85.4	81.0	92.5	92.5	92.5
A10	94.9	94.9	94.9	93.4	94.1	94.1	94.1
A11	93.9	93.9	93.8	93.6	88.1	88.1	88.1
A12	88.3	88.3	88.3	87.9	83.0	83.0	83.0
A13	93.1	93.2	93.0	91.5	95.8	95.8	95.8
A14	98.7	98.7	98.7	97.8	97.1	97.0	97.0
A15	97.2	97.2	97.2	96.9	94.9	94.9	94.9
A16	94.1	94.1	94.1	94.4	91.6	91.6	91.5

3.4 Real-data Examples

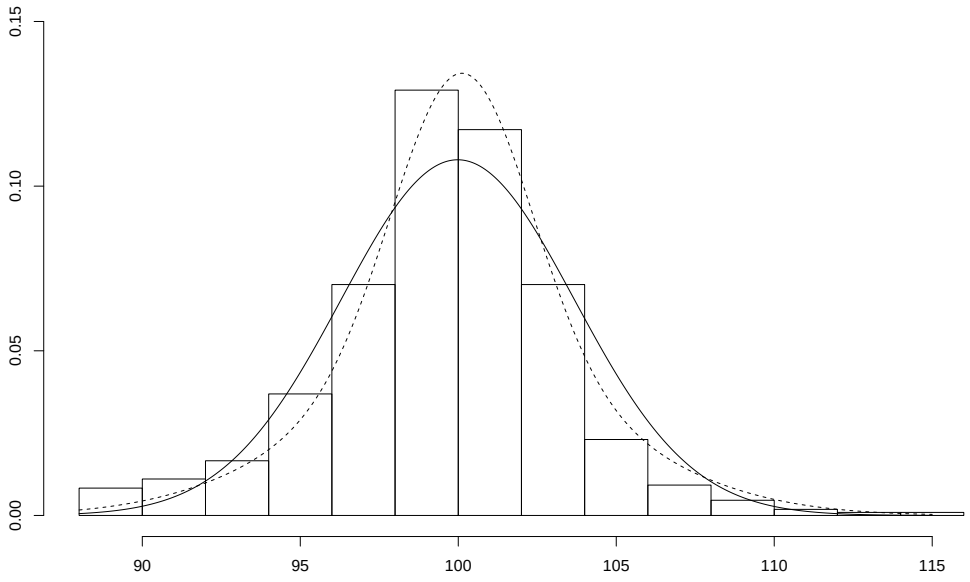
Example 3.1. (Continued) We now apply the retooled EM-test to the 542 blood-chloride measurements. The retooled EM-test statistics are $EM_n^{(1)} = 33.56$, $EM_n^{(2)} = 33.58$, and $EM_n^{(3)} = 33.59$. Calibrated by the limiting distribution $0.5\chi_0^2 + 0.5\chi_1^2$, the corresponding p-values are all around $3e-9$. The p-values of the EM-test statistics $\widetilde{EM}_n^{(1)}$, $\widetilde{EM}_n^{(2)}$, and $\widetilde{EM}_n^{(3)}$ in Chen & Li (2009) are all around $1e-8$. Both methods suggest overwhelming evidence against the homogeneous model. This leads to the conclusion that the data support the existence of unhealthy individuals in the collected sample. Since the retooled EM-test is specifically designed for testing homogeneity in scale mixtures of normal distributions, it provides stronger evidence than the test in Chen & Li (2009), which is designed to be more general.

Figure 3.1 compares the fittings of the scale mixture of two normal distributions (dashed line) and the homogeneous normal distribution (solid line). Clearly, the former provides a better fit, which further supports the results for the retooled EM-test.

Example 3.2. (Age of onset of schizophrenia) This example considers the age of onset of schizophrenia for 152 male schizophrenics from a schizophrenia study reported in Lewine (1981). As suggested by Lewine (1981), there may be two types of schizophrenia. The first type is characterized by early-onset, typical symptoms and poor premorbid competence; the second is associated with late-onset, atypical symptoms and good premorbid competence. An interesting question is whether the two types of schizophrenia really exist.

Everitt, Landau, & Leese (2001) analyzed the data by fitting a two-component normal mixture on the mean and variance parameters to the logarithm of the

Figure 3.1: Histogram and two fitted densities of 542 blood-chloride measurements: the density from the homogeneous normal distribution (solid line) and the density from the two-component scale mixture of normal distributions (dashed line).

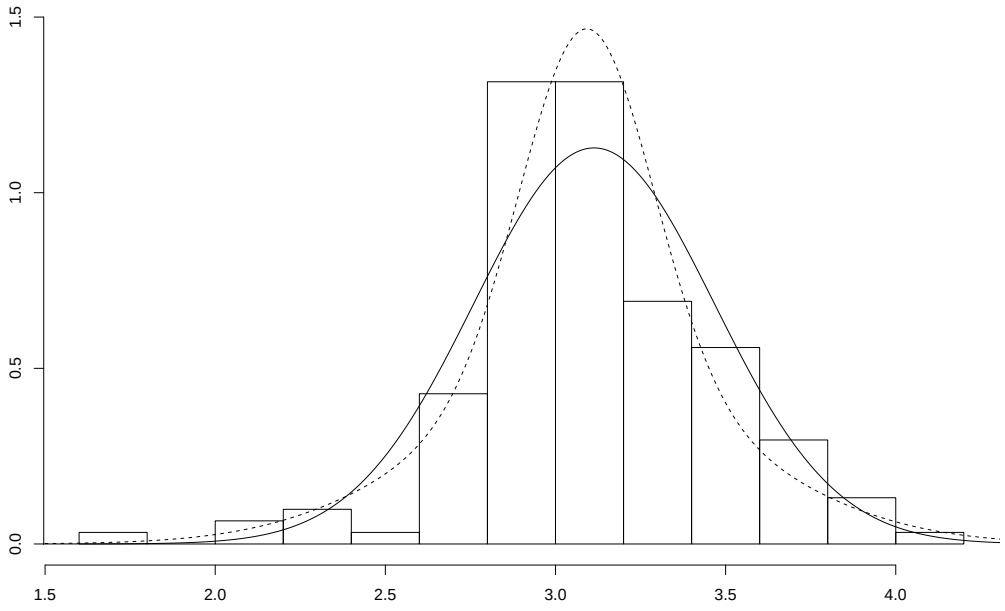


152 observations. They used the LRT to test homogeneity in this model. Chen & Li (2009) applied the EM-test for homogeneity; they found that the p-values of $\widetilde{EM}_n^{(1)}$, $\widetilde{EM}_n^{(2)}$, and $\widetilde{EM}_n^{(3)}$ are all around $1e-3$. Therefore, the data strongly support the heterogenous model for the 152 log-transformed observations. They further noticed that the means for the two components are almost the same, and the variances for the two components are quite different. For the purposes of illustration, we fit the 152 log-transformed observations by a two-component scale mixtures of normal distributions and apply the retooled EM-test. The retooled EM-test statistics are $EM_n^{(1)} = 11.3380$, $EM_n^{(2)} = 11.3380$, and $EM_n^{(3)} = 11.3380$ with corresponding p-values around

$4e-4$. Therefore, the retooled EM-test also strongly supports the existence of two types of schizophrenia, and it is more powerful than that in Chen & Li (2009).

Figure 3.2 compares the fittings of the homogenous normal model and the scale mixture of two normal distributions. Clearly, the mixture model provides a more reasonable fit, which again supports the results for the retooled EM-test.

Figure 3.2: Histogram and two fitted densities of 152 log-transformed ages of onset of schizophrenia for males: the density from the homogeneous normal distribution (solid line) and the density from the two-component scale mixture of normal distributions (dashed line).



3.5 Proof

Since the retooled EM-test is invariant to the scale transformation, without loss of generality, we assume that the true distribution under the null hypothesis is $N(0,1)$.

3.5.1 Two useful lemmas

We first present two useful lemmas. The first is about the consistency of $(\alpha_j^{(K)}, \mu_j^{(K)}, \sigma_{j,1}^{(K)}, \sigma_{j,2}^{(K)})$ under the null hypothesis.

Lemma 3.1: *Suppose that Conditions C1–C4 are satisfied. Then under the null distribution $N(0,1)$, we have, for $j = 1, 2, \ldots, J$ and any finite K,*

$$\alpha_j^{(K)} - \alpha_j = o_p(1),\ \mu_j^{(K)} = o_p(1),\ \sigma_{j,1}^{(K)} - 1 = o_p(1),\ \text{and } \sigma_{j,2}^{(K)} - 1 = o_p(1).$$

Proof. The proof is similar to that of Theorem 3 in Chen & Li (2009) and is therefore omitted. □

The next lemma concerns the expansion of the modified log-likelihood function when $(\mu, \sigma_1, \sigma_2)$ are in small neighborhoods of the true values. For $i = 1, 2, \ldots, n$, we define

$$Z_i = \frac{X_i^2 - 1}{2},\ U_i = \frac{X_i^3 - 3X_i}{6},\ \text{and } V_i = \frac{X_i^4 - 6X_i^2 + 3}{24}.$$

Lemma 3.2: *Assume that the conditions of Lemma 3.1 hold. Suppose $(\bar{\alpha}, \bar{\mu}, \bar{\sigma}_1, \bar{\sigma}_2)$ are estimators of $(\alpha, \mu, \sigma_1, \sigma_2)$ such that $(\bar{\mu}, \bar{\sigma}_1, \bar{\sigma}_2) = (0, 1, 1) + o_p(1)$ and $\bar{\alpha} \in (\delta, 0.5]$ for some $\delta > 0$. Under the null distribution $N(0,1)$, we*

have

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

Here, $(x)^+$ *denotes* $\max\{x,0\}$*, the positive part of* x*.*

Proof. We first write $2\{pl_n(\bar{\alpha},\bar{\mu},\bar{\sigma}_1,\bar{\sigma}_2) - pl_n(0.5,0,1,1)\}$ as the sum of two terms:

$$\begin{aligned} & 2\{pl_n(\bar{\alpha},\bar{\mu},\bar{\sigma}_1,\bar{\sigma}_2) - pl_n(0.5,0,1,1)\} \\ = & \, 2\{l_n(\bar{\alpha},\bar{\mu},\bar{\sigma}_1,\bar{\sigma}_2) - l_n(0.5,0,1,1)\} \\ & +2\{p_n(\bar{\sigma}_1) + p_n(\bar{\sigma}_2) + p(\bar{\alpha}) - 2p_n(1) - p(0.5)\}. \end{aligned}$$

We next find separate upper bounds for the two terms on the right-hand side of the above equation.

From (A.20) in Chen & Li (2008), we directly have

$$\begin{aligned} & 2\{l_n(\bar{\alpha},\bar{\mu},\bar{\sigma}_1,\bar{\sigma}_2) - l_n(0.5,0,1,1)\} \\ \leq & \, 2\Big\{\bar{t}_1\sum_{i=1}^n X_i + \bar{t}_2\sum_{i=1}^n Z_i + \bar{t}_3\sum_{i=1}^n U_i + \bar{t}_4\sum_{i=1}^n V_i\Big\} \\ & -\Big\{\bar{t}_1^2\sum_{i=1}^n X_i^2 + \bar{t}_2^2\sum_{i=1}^n Z_i^2 + \bar{t}_3^2\sum_{i=1}^n U_i^2 + \bar{t}_4^2\sum_{i=1}^n V_i^2\Big\}\{1+o_p(1)\} + o_p(1). \end{aligned}$$

Here

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

with

$$\bar{m}_{l,s} = (1-\bar{\alpha})\bar{\mu}^l(\bar{\sigma}_1^2-1)^s + \bar{\alpha}\bar{\mu}^l(\bar{\sigma}_2^2-1)^s, \quad l=1,2,3,4, \ s=1,2.$$

We notice that $\bar{m}_{l,s}$ can be simplified to

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

Given the conditions on $(\bar{\alpha}, \bar{\mu}, \bar{\sigma}_1, \bar{\sigma}_2)$, the $\bar{t}_l$'s are as follows:

$$\bar{t}_1 = \bar{\mu}, \quad \bar{t}_2 = \tilde{t}_2 + o_p(\bar{t}_1), \quad \bar{t}_3 = o_p(\bar{t}_1), \quad \bar{t}_4 = \tilde{t}_4 + o_p(\bar{t}_1)$$

where $\tilde{t}_2 = (1-\bar{\alpha})(\bar{\sigma}_1^2-1)+\bar{\alpha}(\bar{\sigma}_2^2-1)$ and $\tilde{t}_4 = 3\{(1-\bar{\alpha})(\bar{\sigma}_1^2-1)^2+\bar{\alpha}(\bar{\sigma}_2^2-1)^2\}$. Hence, the upper bound of $2\{l_n(\bar{\alpha}, \bar{\mu}, \bar{\sigma}_1, \bar{\sigma}_2) - l_n(0.5, 0, 1, 1)\}$ can be refined to

$$\begin{aligned}
&2\{l_n(\bar{\alpha}, \bar{\mu}, \bar{\sigma}_1, \bar{\sigma}_2) - l_n(0.5, 0, 1, 1)\} \\
&\leq 2\Big\{\bar{\mu}\sum_{i=1}^n X_i + \tilde{t}_2\sum_{i=1}^n Z_i + \tilde{t}_4\sum_{i=1}^n V_i\Big\} \\
&\quad -\Big\{\bar{\mu}^2\sum_{i=1}^n X_i^2 + \tilde{t}_2^2\sum_{i=1}^n Z_i^2 + \tilde{t}_4^2\sum_{i=1}^n V_i^2\Big\}\{1+o_p(1)\} + o_p(1). \qquad (3.1)
\end{aligned}$$

We now consider the upper bound for $2\{p_n(\bar{\sigma}_1) + p_n(\bar{\sigma}_2) + p(\bar{\alpha}) - 2p_n(1) - p(0.5)\}$. Using Conditions C1 and C3, we get

$$\begin{aligned}
&2\{p_n(\bar{\sigma}_1) + p_n(\bar{\sigma}_2) + p(\bar{\alpha}) - 2p_n(1) - p(0.5)\} \\
&\leq 2\{p_n(\bar{\sigma}_1) + p_n(\bar{\sigma}_2) - 2p_n(1)\} = o_p(n^{1/6})\{|\bar{\sigma}_1^2-1| + |\bar{\sigma}_2^2-1|\} \\
&\leq o_p(n)\Big\{\tilde{t}_2^2 + \tilde{t}_4^2\Big\} + o_p(1). \qquad (3.2)
\end{aligned}$$

Combining (3.1) and (3.2) gives

$$
\begin{aligned}
& 2\{pl_n(\bar{\alpha}, \bar{\mu}, \bar{\sigma}_1, \bar{\sigma}_2) - pl_n(0.5, 0, 1, 1)\} \\
\leq\ & 2\Big\{\bar{\mu}\sum_{i=1}^{n} X_i + \tilde{t}_2\sum_{i=1}^{n} Z_i + \tilde{t}_4\sum_{i=1}^{n} V_i\Big\} \\
& -\Big\{\bar{\mu}^2\sum_{i=1}^{n} X_i^2 + \tilde{t}_2^2\sum_{i=1}^{n} Z_i^2 + \tilde{t}_4^2\sum_{i=1}^{n} V_i^2\Big\}\{1 + o_p(1)\} + o_p(1) \\
\leq\ & \sup_{\mu,\ t_2,\ t_4}\Bigg[2\Big\{\mu\sum_{i=1}^{n} X_i + t_2\sum_{i=1}^{n} Z_i + t_4\sum_{i=1}^{n} V_i\Big\} \\
& -\Big\{\mu^2\sum_{i=1}^{n} X_i^2 + t_2^2\sum_{i=1}^{n} Z_i^2 + t_4^2\sum_{i=1}^{n} V_i^2\Big\}\Bigg] + o_p(1).
\end{aligned}
$$

Here $t_2 = (1-\alpha)(\sigma_1^2 - 1) + \alpha(\sigma_2^2 - 1)$ and $t_4 = 3\{(1-\alpha)(\sigma_1^2-1)^2 + \alpha(\sigma_2^2-1)^2\}$. Since $t_4 \geq 0$, we further have

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

This finishes the proof. □

3.5.2 Proof of Theorem 3.1

Let

$$
r_{1n}(\alpha_j^{(K)}, \mu_j^{(K)}, \sigma_{j,1}^{(K)}, \sigma_{j,2}^{(K)}) = 2\{pl_n(\alpha_j^{(K)}, \mu_j^{(K)}, \sigma_{j,1}^{(K)}, \sigma_{j,2}^{(K)}) - pl_n(0.5, 0, 1, 1)\}
$$

and

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

Then $M_n^{(K)}(\alpha_j) = r_{1n}(\alpha_j^{(K)}, \mu_j^{(K)}, \sigma_{j,1}^{(K)}, \sigma_{j,2}^{(K)}) + r_{2n}$.

The conclusion of Lemma 3.1 implies that the upper bound in Lemma 3.2 is also applicable to $r_{1n}(\alpha_j^{(K)}, \mu_j^{(K)}, \sigma_{j,1}^{(K)}, \sigma_{j,2}^{(K)})$. That is,

$$\begin{aligned} & r_{1n}(\alpha_j^{(K)}, \mu_j^{(K)}, \sigma_{j,1}^{(K)}, \sigma_{j,2}^{(K)}) \\ &= 2\{pl_n(\alpha_j^{(K)}, \mu_j^{(K)}, \sigma_{j,1}^{(K)}, \sigma_{j,2}^{(K)}) - pl_n(0.5, 0, 1, 1)\} \\ &\leq \frac{(\sum_{i=1}^n X_i)^2}{\sum_{i=1}^n X_i^2} + \frac{(\sum_{i=1}^n Z_i)^2}{\sum_{i=1}^n Z_i^2} + \frac{\{(\sum_{i=1}^n V_i)^+\}^2}{\sum_{i=1}^n V_i^2} + o_p(1). \end{aligned} \tag{3.3}$$

Applying some of the classic results about regular models, we have

$$r_{2n} = -\frac{(\sum_{i=1}^n X_i)^2}{\sum_{i=1}^n X_i^2} - \frac{(\sum_{i=1}^n Z_i)^2}{\sum_{i=1}^n Z_i^2} + o_p(1). \tag{3.4}$$

Combining (3.3) and (3.4), we get

$$M_n^{(K)}(\alpha_j) = r_{1n}(\alpha_j^{(K)}, \mu_j^{(K)}, \sigma_{j,1}^{(K)}, \sigma_{j,2}^{(K)}) + r_{2n} \leq \frac{\{(\sum_{i=1}^n V_i)^+\}^2}{\sum_{i=1}^n V_i^2} + o_p(1).$$

The leading term in the above equation does not depend on α, and therefore it also serves as an upper bound of $EM_n^{(K)}$, i.e.,

$$EM_n^{(K)} \leq \frac{\{(\sum_{i=1}^n V_i)^+\}^2}{\sum_{i=1}^n V_i^2} + o_p(1).$$

We now show that the upper bound is achievable. Since the EM-iteration increases the penalized log-likelihood (Dempster, Laird, & Rubin, 1977), we only need to show that this is the case when $K = 1$. It suffices to find a set of parameter values $\hat{\alpha}$, $\hat{\mu}$, $\hat{\sigma}_1$, $\hat{\sigma}_2$ at which the upper bound is attained. Let

$\hat{\alpha} = 0.5$ and $\hat{\mu} = \sum_{i=1}^n X_i / \sum_{i=1}^n X_i^2$. We choose $\hat{\sigma}_1$ and $\hat{\sigma}_2$ such that

$$0.5(\hat{\sigma}_1^2 - 1) + 0.5(\hat{\sigma}_2^2 - 1) = \frac{\sum_{i=1}^n Z_i}{\sum_{i=1}^n Z_i^2},$$
$$3\{0.5(\hat{\sigma}_1^2 - 1)^2 + 0.5(\hat{\sigma}_2^2 - 1)^2\} = \frac{\{(\sum_{i=1}^n V_i)^+\}}{\sum_{i=1}^n V_i^2}.$$

It can be checked that $\hat{\sigma}_1$ and $\hat{\sigma}_2$ exist. Further, $\hat{\mu} = O_p(n^{-1/2})$, $\hat{\sigma}_1^2 - 1 = O_p(n^{-1/4})$, and $\hat{\sigma}_2^2 - 1 = O_p(n^{-1/4})$. With this order information, we obtain

$$2\{pl_n(\hat{\alpha}, \hat{\mu}, \hat{\sigma}_1, \hat{\sigma}_2) - pl_n(0.5, \hat{\mu}_0, \hat{\sigma}_0, \hat{\sigma}_0)\} = \frac{\{(\sum_{i=1}^n V_i)^+\}^2}{\sum_{i=1}^n V_i^2} + o_p(1).$$

Because the EM-iteration always increases the penalized log-likelihood, we must have

$$\begin{aligned} EM_n^{(K)} \geq M_n^{(1)}(0.5) &\geq 2\{pl_n(\hat{\alpha}, \hat{\mu}, \hat{\sigma}_1, \hat{\sigma}_2) - pl_n(0.5, \hat{\mu}_0, \hat{\sigma}_0, \hat{\sigma}_0)\} \\ &= \frac{\{(\sum_{i=1}^n V_i)^+\}^2}{\sum_{i=1}^n V_i^2} + o_p(1). \end{aligned}$$

This shows that the asymptotic upper bound of $EM_n^{(K)}$ is identical to its lower bound, which implies that

$$EM_n^{(K)} = \frac{\{(\sum_{i=1}^n V_i)^+\}^2}{\sum_{i=1}^n V_i^2} + o_p(1).$$

By central limit theorem,

$$\frac{\sum_{i=1}^n V_i}{\sqrt{\sum_{i=1}^n V_i^2}} \to N(0, 1)$$

in distribution. Therefore, $EM_n^{(K)}$ asymptotically follows the distribution $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$. □

Chapter 4

Testing Homogeneity in a Contaminated Normal Model

4.1 Introduction

This chapter is mainly motivated by detecting the alternative hypotheses in large-scale hypothesis testing problem, in which we need to conduct thousands and sometimes millions of hypothesis tests on parallel data sets. One particular area involving large-scale hypothesis testing is the microarray study. The following is an illustrating example from this area. More applications can be found in Efron (2010).

Example 4.1. (Prostate data) *The prostate data consist of the gene expression levels of* $n = 6033$ *genes for 102 male individuals: 52 prostate cancer patients and 50 normal control subjects. The principal goal of the study is to find the genes that are differentially expressed between the prostate cancer patients and the normal control subjects. Such genes will be used for further investigation.*

For the ith gene, we test

H_{0i} : *gene i is not differentially expressed in two samples or gene i is "null".*

Hence we need to deal with over 6000 hypotheses at the same time. The commonly used test statistic for H_{0i} *is the traditional two-sample test statistic* t_i*. Under* H_{0i}*,* t_i *follows or approximately follows a t-distribution with 100 degrees of freedom. Efron (2010) suggested transforming* t_i *to* z_i *by*

$$z_i = \Phi^{-1}(F_{100}(t_i)),$$

where Φ *and* F_{100} *are the cumulative distribution functions for* $N(0,1)$ *and t-distribution with 100 degrees of freedom. Under* H_{0i}*,* z_i *follows or approximately follows* $N(0,1)$*, known as the theoretical null distribution. Efron (2010) then used* z_i *to test* H_{0i}*.*

In large-scale hypothesis testing problem, scientists are not interested in controlling the type-I error individually. Instead they prefer the notion of controlling the false discovery rate (Benjamini & Hochberg, 1995). Among many methods for controlling this rate, Efron (2004) proposed the use of a finite normal mixture to model the z-scores z_i's. See also McLachlan, Bean, & Jones (2006) and Dai & Charnigo (2010). An appropriate candidate model is

$$(1-\alpha)f(x;\mu_1,\sigma_1) + \alpha f(x;\mu_2,\sigma_2)$$

with the 1st component corresponding to the null genes and the 2nd component corresponding to the differentially expressed genes. Here $f(x;\mu,\sigma)$ denotes the probability density function of the normal distribution $N(\mu,\sigma^2)$.

In theory, $f(x;\mu_1,\sigma_1)$ should be the probability density function of $N(0,1)$, the theoretical null distribution. In practice, the theoretical null distribution may fail to work due to several reasons. For example, the null distribution of t_i may not be the exact t-distribution, see Efron (2010, pp. 105–109). Hence Efron (2010) suggested the notion of empirical null distribution, which in many examples can be well approximated by $N(0,\sigma_1^2)$. Hence we suggest modelling z-scores by a contaminated normal model:

$$(1-\alpha)f(x;0,\sigma_1)+\alpha f(x;\mu,\sigma_2). \tag{4.1}$$

Before identifying the genes that are differentially expressed in two samples, we first detect the existence of these genes by testing the homogeneity under the contaminated normal model (4.1). That is, we aim to test

$$H_0 : \alpha = 0 \text{ or } (0,\sigma_1) = (\mu,\sigma_2). \tag{4.2}$$

Developing an effective testing procedure for (4.2) is again a challenging problem. Similar to the scale mixtures of normal distributions, the log-likelihood function is unbounded and the Fisher information in the mixing proportion direction may be infinity. The asymptotic results for all the existing methods can not be directly applied. In this chapter, we design a new class of EM-tests specifically for (4.2) under the contaminated model (4.1).

The rest of this chapter is organized as follows. In Section 4.2, we describe the new EM-test procedure, and present the asymptotic properties. In Section 4.3, simulation studies are used to check the type-I errors and powers of the new EM-test. We further compare the performance of the new EM-test with

other methods in the literature. In Section 4.4, a real-data example is analyzed to illustrate the proposed EM-test. All the proofs are given in Section 4.5.

We finish this section with two remarks.

Remark 4.1. Dai & Charnigo (2010) suggested modelling the z-score by another contaminated normal model:

$$(1-\alpha)N(0,\sigma^2)+\alpha N(\mu,\sigma^2). \tag{4.3}$$

They further proposed two methods: the MLRT and the D-test for testing homogeneity in (4.3). The motivation of our new model in (4.1) is twofold. First, in some examples, the variances in two components are observed to be different. One particular example is given in Section 4.4. Second, as shown in Section 4.3, the proposed EM-test based on model (4.1) is comparable to the two methods in Dai & Charnigo (2010) even when z-scores are generated from model (4.3). The proposed EM-test becomes more powerful when z-scores are not generated from model (4.3).

Remark 4.2. Another stream for detecting the existence of differentially expressed genes is based on p-values: $p_i = 2\{1-\Phi(|z_i|)\}$, $i = 1,\ldots,n$. A commonly used model for modelling p-values is the contaminated Beta model:

$$(1-\alpha)B(1,1)+\alpha B(a,b), \tag{4.4}$$

where $B(a,b)$ denotes the Beta distribution with two parameters a and b. Note that $B(1,1)$ is also the uniform distribution over (0,1). See Allison et al., (2002), Peng (2003), Dai & Charnigo (2008), and the reference therein. Dai & Charnigo (2008) further proposed the use of MLRT and D-test for testing

homogeneity in (4.4). In Section 4.3, our new EM-test is further compared with these two tests for detecting the existence of differentially expressed genes.

4.2 Main Results

4.2.1 The new EM-test procedure

Suppose $X_1, \ldots, X_n$ are a random sample of size n from the contaminated normal model (4.1). We are interested in the homogeneity test problem in (4.2).

We denote the log-likelihood function as

$$l_n(\alpha, \mu, \sigma_1, \sigma_2) = \sum_{i=1}^{n} \log\{(1-\alpha) f(X_i; 0, \sigma_1) + \alpha f(X_i; \mu, \sigma_2)\}$$

and define the modified log-likelihood function as

$$pl_n(\alpha, \mu, \sigma_1, \sigma_2) = l_n(\alpha, \mu, \sigma_1, \sigma_2) + p(\alpha) + p_n(\sigma_1) + p_n(\sigma_2).$$

The penalty function $p(\alpha)$ is used to prevent the fitting of α being close to 0. Here we do not penalize the fitting of α being close to 1 since in large-scale hypothesis testing problem, the true value of α is in general quite small, for example, smaller than 0.25 (Efron, 2010). One example for $p(\alpha)$ is $p(\alpha) = \log(\alpha)$, which has been used in Fu, Chen, & Li (2008) for testing homogeneity in a class of contaminated von Mises model. The penalty $p_n(\sigma)$ prevents the fitting of σ_1^2 and σ_2^2 being close to 0, which help avoid the unbounded likelihood

(Chen, Tan, & Zhang, 2008). An example of $p_n(\sigma)$ is

$$p_n(\sigma) = -a_n \cdot \left(\frac{\hat{\sigma}_0^2}{\sigma^2} + \log \frac{\sigma^2}{\hat{\sigma}_0^2} \right),$$

where $\hat{\sigma}_0^2 = \sum_{i=1}^n X_i^2/n$ is the maximum likelihood estimator of the variance parameter under the null model. Using the penalty function, $\hat{\sigma}_0^2$ also maximizes the modified log-likelihood function under the null hypothesis. The choice of a_n is discussed in Section 4.2.3.

Similar to Chapters 2 and 3, the EM-test statistics are constructed in the following iterative way. We first choose a finite set of $\{\alpha_1, \ldots, \alpha_J\}$ for α and a positive integer K. The specific choices of $\{\alpha_1, \ldots, \alpha_J\}$ and K will be suggested in Section 4.2.3.

For each $j = 1, 2, \ldots, J$, compute

$$(\mu_j^{(1)}, \sigma_{j,1}^{(1)}, \sigma_{j,2}^{(1)}) = \arg\max_{\mu,\sigma_1,\sigma_2} pl_n(\alpha_j, \mu, \sigma_1, \sigma_2).$$

Further let $k = 1$ and $\alpha_j^{(1)} = \alpha_j$. The EM-iteration starts from here.

In the E-step of the k-th iteration ($k = 1, \ldots, K$), we compute the posterior probabilities

$$w_{ij}^{(k)} = \frac{\alpha_j^{(k)} f(X_i; \mu_j^{(k)}, \sigma_{j,2}^{(k)})}{(1 - \alpha_j^{(k)}) f(X_i; 0, \sigma_{j,1}^{(k)}) + \alpha_j^{(k)} f(X_i; \mu_j^{(k)}, \sigma_{j,2}^{(k)})}.$$

In the M-step of the k-th EM-iteration, we then update $(\alpha, \mu, \sigma_1, \sigma_2)$ by

$$\begin{aligned}
\alpha_j^{(k+1)} &= \arg\max_\alpha \Big\{ \sum_{i=1}^{n} (1 - w_{ij}^{(k)}) \log(1-\alpha) + \sum_{i=1}^{n} w_{ij}^{(k)} \log(\alpha) + p(\alpha) \Big\}, \\
\mu_j^{(k+1)} &= \arg\max_\mu \Big\{ \sum_{i=1}^{n} w_{ij}^{(k)} \log f(X_i; \mu, \sigma_2) \Big\}, \\
\Big(\sigma_{j,1}^{(k+1)}, \sigma_{j,2}^{(k+1)}\Big) &= \arg\max_{\sigma_1,\sigma_2} \Big\{ \sum_{i=1}^{n} (1 - w_{ij}^{(k)}) \log f(X_i; 0, \sigma_1) + p_n(\sigma_1) + p_n(\sigma_2) \\
&\qquad + \sum_{i=1}^{n} w_{ij}^{(k)} \log f(X_i; \mu_j^{(k)}, \sigma_2) \Big\}.
\end{aligned}$$

For each k and j, we define

$$M_n^{(k)}(\alpha_j) = 2\{pl_n(\alpha_j^{(k)}, \mu_j^{(k)}, \sigma_{j,1}^{(k)}, \sigma_{j,2}^{(k)}) - pl_n(1, 0, \hat{\sigma}_0, \hat{\sigma}_0)\}.$$

The EM-test statistic $EM_n^{(K)}$ is defined as

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

The null hypothesis is rejected when $EM_n^{(K)}$ exceeds the critical value of the limiting distribution given in section 4.2.2.

4.2.2 Limiting distribution of the EM-test

We derive the limiting distribution of the EM-test statistics $EM_n^{(K)}$ under the following conditions on the penalty functions $p(\alpha)$ and $p_n(\sigma)$.

D1 The penalty function $p(\alpha)$ is continuous, approaches negative infinity as α approaches zero. Further $p(1) = 0$.

D2 $\sup\{|p_n(\sigma)|\} = o(n)$.

D3 $p_n'(\sigma) = o_p(n^{1/6})$ at any $\sigma > 0$.

D4 $p_n(\sigma) \leq 4(\log n)^2 \log(\sigma)$, when $\sigma \leq n^{-1}$ and n is large.

Conditions D1-D4 are very similar to Conditions C1-C4 in Chapter 3. These conditions guarantee that the new EM-test has a simple limiting distribution. They are satisfied by the suggested penalty functions in Section 4.2.1.

Theorem 4.1: *Suppose that the penalty functions $p(\alpha)$, $p_n(\sigma)$ satisfy Conditions D1-D4 and the initial set $\{\alpha_1, \ldots, \alpha_J\} \in (0, 1)$. Under the null hypothesis and for any fixed finite K, as $n \to \infty$,*

$$EM_n^{(K)} \to \frac{1}{2}\chi_1^2 + \frac{1}{2}\chi_2^2 + 2\max_j p(\alpha_j)$$

in distribution.

The result in Theorem 4.1 needs some interpretations. In the EM-test procedure, α is bounded away from 0. With $\alpha \neq 0$, testing homogeneity in (4.1) is to test $\mu = 0$ and the homogeneity of variances in two components. For any given α_j, $M_n^{(K)}(\alpha_j)$ contains two terms: a term due to log-likelihood difference and a term due to penalty functions. Testing $\mu = 0$ contributes a χ_1^2 to the limiting distribution of the term due to log-likelihood difference; testing the homogeneity of variances contributes to another $0.5\chi_0^2 + 0.5\chi_1^2$ to the limiting distribution of the term due to log-likelihood difference, as shown in Chapter 3. Roughly speaking, adding two parts together, the term due to log-likelihood difference in $M_n^{(K)}(\alpha_j)$ has a $\frac{1}{2}\chi_1^2 + \frac{1}{2}\chi_2^2$ limiting distribution. Hence $EM_n^{(K)}$ has the limiting distribution given in Theorem 4.1 with $2\max_j p(\alpha_j)$ coming from the penalty functions.

4.2.3 Choice of the penalty functions

Before the EM-test is implemented in applications, we need to specify several tuning parameters: the finite set $\{\alpha_1, \ldots, \alpha_J\}$, the iteration number K, the penalty functions $p(\alpha)$ and $p_n(\sigma)$.

For $\{\alpha_1, \ldots, \alpha_J\}$, we suggest using $\{0.05, 0.15, 0.25\}$. As discussed in Efron (2010), in most situations, the value of α is quite small, for example less than 0.25. Hence we choose three initial values which are less than or equal to 0.25 for α. Empirical experience suggests that further increasing the number of initial values may not significantly improve the power of the EM-test. Similar to Chapters 2 and 3, we use $K = 2$ or 3 as the iteration number. For $p(\alpha)$, we use the penalty $p(\alpha) = \log(\alpha)$ as suggested in Fu, Chen, & Li (2008). This penalty function satisfies Condition D1 for the theoretical development. Further, in the M-step of the EM-iteration, α can be updated via an explicit form. For the penalty $p_n(\sigma)$, we recommend

$$p_n(\sigma) = -a_n \cdot \left(\frac{\hat{\sigma}_0^2}{\sigma^2} + \log \frac{\sigma^2}{\hat{\sigma}_0^2} \right).$$

As long as $a_n = o(n^{1/6})$, $p_n(\sigma)$ satisfies Conditions D2-D4. Carefully tuning the value of a_n in the penalty functions will further improve the precision of the approximation of the limiting distribution to the finite-sample distribution of $EM_n^{(K)}$. We adapt the computer experiment approach used in Chapter 3 and Chen & Li (2011) to obtain an empirical formula for a_n.

The idea of computer experiment is similar to that in Section 3.2.3. We omit it here. At the beginning of the computer experiment, we carry out pilot experiment for many choices of sample sizes n. We find that when $a_n \leq 1.4$ the simulated type-I errors of the EM-test are larger than the nominal levels.

Hence, in designed experiment, a_n is chosen to be from 1.6 to 4.0, with the step length 0.2. In total, 13 values of a_n are considered. We consider three quite large sample sizes: 500, 1000, and 1500 since in large-scale hypothesis testing problem, the number of parallel hypotheses are around thousands. Then a 13×3 full factorial design is used in our computer experiment. For each combination of a_n and n, 5000 random samples of size n from the $N(0,1)$ are used to calculate the simulated type-I errors $\hat{q}$ of $EM_n^{(2)}$ at the target significance level q. Similar to Section 3.2.3, the discrepancy between $\hat{q}$ and q is calculated as

$$y = \log\{\hat{q}/(1-\hat{q})\} - \log\{q/(1-q)\}.$$

Table 4.1 presents the discrepancy between $\hat{q}$ and q when $q = 0.05$, the same level used in Section 3.2.3.

Table 4.1: Discrepancy between $\hat{q}$ and q in term of y under the contaminated normal models.

a_n	n=500	n=1000	n=1500
1.6	0.157	0.175	0.278
1.8	0.061	0.101	0.210
2.0	0.081	0.081	0.157
2.2	0.061	0.101	0.101
2.4	0.061	0.041	0.138
2.6	0.000	0.101	0.101
2.8	-0.043	0.041	0.061
3.0	-0.043	0.061	-0.021
3.2	-0.111	0.041	-0.021
3.4	-0.065	-0.021	0.138
3.6	-0.111	-0.065	0.081
3.8	-0.021	0.041	0.041
4.0	-0.234	-0.065	0.061

Analysis of variance suggests both n and a_n have significant effects on y. After some brainstorming and exploratory analysis, the covariates in the form of $1/n$ and $\log\{a_n - 1.4\}$ gives the most satisfactory outcomes in terms of both the goodness of fit and the simplicity of the resulting formula for a_n. The covariate $\log\{a_n - 1.4\}$ effectively confines the value of a_n in $(1.4, \infty)$, as suggested by our pilot study.

We next regress y in $1/n$ and $\log\{a_n - 1.4\}$. Based on 39 observations, the fitted regression model is

$$\hat{y} = 0.158 - 82.899/n - 0.094 \log(a_n - 1.4)$$

with $R^2 = 73.9\%$. Setting $\hat{y} = 0$ gives the following empirical formula for a_n:

$$a_n = \exp(1.681 - 881.904/n) + 1.4.$$

Since our EM-test procedure is invariant to the scale transformation, the empirical formula of a_n is applicable to the general null distribution $N(0, \sigma^2)$. In the next section, we examine the performance of the derived empirical formula of a_n and other suggested tuning parameters.

4.3 Simulation Study

The purpose of simulation study is twofold: (i) check if the limiting distribution of the EM-test provides accurate approximation to the finite sample distribution; (ii) compare the power of the EM-test with the MLRT ($\lambda_{n,N}$) and the D-test ($d_{n,N}$) proposed in Dai & Charnigo (2010) under the contaminated normal model in (4.3) and the MLRT ($\lambda_{n,B}$) and the D-test ($d_{n,B}$) proposed in

Dai & Charnigo (2008) under the contaminated Beta model in (4.4). Note that all the five methods can be used to detect the existence of the gene that are differentially expressed in two samples. The EM-test statistics are calculated based on the recommended tuning parameters in Section 4.2.3.

To check the approximation of the limiting distribution of $EM_n^{(K)}$, we generate 10000 replications respectively for $n = 100, 200, \ldots, 1000, 1500, 2000, \ldots, 4500, 5000, 10000$ from the null model $N(0, 1)$. The limiting distribution in Theorem 4.1 is used to calculate the critical values. The simulated type-I error rates of $EM_n^{(K)}$ at the nominal levels $10\%, 5\%$, and 1% are summarized in Table 4.2. As we can see, for all considered sample sizes, the simulated type-I errors are very close to the nominal levels, which indicates that the limiting distribution approximates the finite sample distribution reasonably well.

Next, we conduct the simulation under the alternative models. We choose twelve alternative models, which are listed in Table 4.3. The first six are the alternative models used in Dai & Charnigo (2010). In these models, the component variances are the same for all components. In the second six models, the component variances are chosen to be different; other settings are the same as the first six models. In all these twelve models, σ_1 is set to be 1.

We consider sample sizes $n = 100, 200, 500, 1000, 1500$. For each combination of model and sample size, 5000 replications are used to compute the power of the five tests: $EM_n^{(K)}$, $\lambda_{n,N}$, $d_{n,N}$, $\lambda_{n,B}$, and $d_{n,B}$. We only present the power comparison with sample size n=500 at the 5% level in Table 4.4. The simulation results for other sample sizes and significance levels show the similar trend and therefore we omit them.

From Table 4.4, we have the following two observations:

Table 4.2: Simulated type-I error rates (%) of the EM-test under the contaminated normal model.

	Level=10%			Level=5%			Level=1%		
n	$EM_n^{(1)}$	$EM_n^{(2)}$	$EM_n^{(3)}$	$EM_n^{(1)}$	$EM_n^{(2)}$	$EM_n^{(3)}$	$EM_n^{(1)}$	$EM_n^{(2)}$	$EM_n^{(3)}$
100	9.4	9.9	10.3	4.9	5.2	5.4	0.9	1.0	1.1
200	10.3	10.5	10.7	5.1	5.3	5.4	1.1	1.1	1.1
300	10.1	10.3	10.5	5.0	5.1	5.2	0.9	0.9	1.0
400	10.2	10.3	10.5	5.1	5.1	5.2	0.9	0.9	0.9
500	10.1	10.2	10.3	5.0	5.1	5.1	1.0	1.0	1.1
600	10.1	10.2	10.2	5.0	5.0	5.0	0.9	1.0	1.0
700	10.0	10.1	10.2	4.9	4.9	4.9	1.0	1.0	1
800	9.5	9.6	9.7	4.7	4.7	4.8	1.0	1.0	1.0
900	9.6	9.6	9.7	4.6	4.6	4.7	0.8	0.8	0.9
1000	9.7	9.7	9.8	5.0	5.0	5.0	1.0	1.0	1.0
1500	9.5	9.5	9.6	4.7	4.7	4.8	1.0	1.0	1.0
2000	9.7	9.7	9.7	4.8	4.8	4.8	1.0	1.0	1.0
2500	9.6	9.6	9.6	4.8	4.8	4.8	0.9	0.9	0.9
3000	10.1	10.1	10.1	5.1	5.1	5.1	1.0	1.0	1.0
3500	9.7	9.7	9.7	4.9	4.9	4.9	0.9	0.9	0.9
4000	10.2	10.2	10.2	5.3	5.3	5.3	1.0	1.0	1.0
4500	10.3	10.3	10.3	5.0	5.0	5.0	1.0	1.0	1.0
5000	10.0	10.0	10.0	4.9	4.9	4.9	1.1	1.1	1.1
10000	10.0	10.0	10.0	5.0	5.0	5.0	1.0	1.0	1.0

Table 4.3: Twelve alternative contaminated normal models.

No.	Model
A1	$0.95N(0,1)+0.05N(1,1)$
A2	$0.95N(0,1)+0.05N(2,1)$
A3	$0.9N(0,1)+0.1N(1,1)$
A4	$0.9N(0,1)+0.1N(2,1)$
A5	$0.9N(0,1)+0.05N(1,1)+0.05N(-1,1)$
A6	$0.9N(0,1)+0.05N(2,1)+0.05N(-2,1)$
A7	$0.95N(0,1)+0.05N(1,2)$
A8	$0.95N(0,1)+0.05N(1,0.5)$
A9	$0.9N(0,1)+0.1N(2,2)$
A10	$0.9N(0,1)+0.1N(2,0.5)$
A11	$0.9N(0,1)+0.05N(1,2)+0.05N(-1,0.5)$
A12	$0.9N(0,1)+0.05N(2,2)+0.05N(-2,0.5)$

(i) $EM_n^{(K)}$ is more powerful than $\lambda_{n,N}$ and $d_{n,N}$ under A2 and A9, in which we may observe quite large z-scores. $EM_n^{(K)}$ is also more powerful than $\lambda_{n,N}$ and $d_{n,N}$ under A5-A6 and A11-A12, in which the estimator of μ is very close to 0.

(ii) $EM_n^{(K)}$ is more powerful than or comparable to $\lambda_{n,B}$ and $d_{n,B}$ under A1-A4 and A7-A10. However, $EM_n^{(K)}$ has less power than $\lambda_{n,B}$ and $d_{n,B}$ under A5-A6 and A11-A12.

Based on these two observations, we conclude that (i) the new EM-test based on model (4.1) is as powerful as or more powerful than than the MLRT and D-test based on model (4.3); (ii) the new EM-test based on model (4.1) is as powerful as or more powerful than the MLRT and D-test based on model (4.4) when z-scores are generated from model (4.1); (iii) the new EM-test becomes less powerful than the MLRT and D-test based on model (4.4) when there is symmetry between overexpression and underexpression.

Table 4.4: Powers (%) of the EM-test under twelve alternative contaminated normal models at the 5% significance level, n=500.

Model No.	$\lambda_{n,N}$	$d_{n,N}$	$\lambda_{n,B}$	$d_{n,B}$	$EM_n^{(1)}$	$EM_n^{(2)}$	$EM_n^{(3)}$
A1	19.2	19.2	10.0	10.9	17.5	17.6	17.7
A2	54.8	54.8	74.8	79.3	79.0	79.0	79.1
A3	57.2	57.2	26.5	29.9	54.6	54.7	54.8
A4	97.1	97.0	99.8	99.8	99.5	99.5	99.5
A5	5.3	5.3	25.6	28.5	6.4	6.4	6.5
A6	5.0	5.0	99.7	99.8	52.5	53.0	53.3
A7	54.4	39.8	76.8	82.7	81.5	81.5	81.6
A8	19.2	19.2	7.1	8.2	17.1	17.2	17.3
A9	100	99.9	97.4	98.9	100	100	100
A10	97.5	97.6	98.2	99	96.9	97	97.1
A11	45.4	31	78.2	83.3	74.7	74.7	74.8
A12	74.8	65	99.1	99.8	92.2	92.3	92.3

4.4 Real-data Example

Example 4.2. (Police data) This example is taken from Efron (2010, pp. 95–97). In 2006 at New York City, a study was conducted to investigate whether there are some police officers that have racial bias with pedestrian stops. The preliminary data included $\mathbf{x}_{ij}$, the vector of covariates for police officer i, stop j; y_{ij}= 0 or 1, the indicator of whether the stopped person belonged to a certain minority group or not. A logistic regression model

$$\log \frac{\Pr(y_{ij}=1)}{1-\Pr(y_{ij}=1)} = \beta_i + \boldsymbol{\gamma}^\tau \mathbf{x}_{ij}$$

was used to estimate the "officer effect" β_i (Efron, 2010). The z-score of the i-th officer is defined as

$$z_i = \hat{\beta}_i / \mathrm{se}(\hat{\beta}_i).$$

In total, $n = 2749$ z-scores are obtained. Large positive z_i's are considered as signs of possible racial bias.

Figure 4.1: Histogram and two fitted densities of the police data: the density from the homogeneous normal distribution (solid line) and the density from the two-component contaminated normal distribution (dashed line).

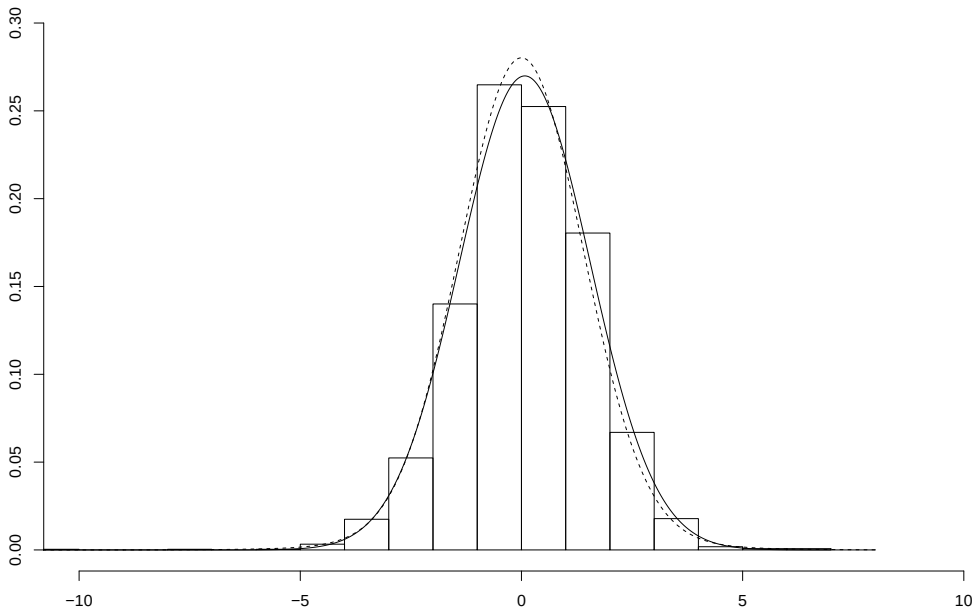


Figure 4.1 shows the histogram of the 2749 z-scores. The solid line and the dashed line are the fitted density curves for the homogeneous normal distribution and the contaminated normal distribution in (4.1), respectively. Merely based on this figure, it is hard to tell which model provides a better fitting. Hence a formal test is required here. The EM-test statistics in this example are found to be $EM_n^{(2)}$ =41.026 with the p-value being around $1.7e-10$ calibrated by its limiting distribution. Based on the p-value, the null hypothesis is soundly rejected. We also apply the MLRTs and D-tests under both

the contaminated normal model in (4.3) and the contaminated Beta model in (4.4) to the police data example. The p-values for the test statistics $\lambda_{n,N}$, $d_{n,N}$, $\lambda_{n,B}$, and $d_{n,B}$ calibrated by their respect limiting distributions are respectively $7.2e-06$, $3.9e-1$, 0, and 0. Clearly, the proposed EM-test provides stronger evidence than $\lambda_{n,N}$ and $d_{n,N}$. It looks that $\lambda_{n,B}$ and $d_{n,B}$ provide even stronger evidence.

The fitted contaminated normal model for the 2749 z-scores is

$$0.951N(0, 1.391^2) + 0.049N(0.021, 2.610^2).$$

The two component variances are quite different, which explains why the proposed EM-test is more powerful than $\lambda_{n,N}$ and $d_{n,N}$. Further note that first component distribution is quite different from the theoretical null distribution $N(0, 1)$. Efron (2010) derived the empirical null and found its variance is 1.40^2, which is quite close to 1.391^2 but far away from 1. Both suggest that the theoretical null distribution may not work here. To see the effect of the failure of theoretical null distribution on $\lambda_{n,B}$ and $d_{n,B}$, we get 10000 random samples of sample size $n = 2749$ from $N(0, 1.391^2)$ and find that the simulated type-I error rates of $\lambda_{n,B}$ and $d_{n,B}$ at the 5% level are around 100%. Hence the limiting distributions of $\lambda_{n,B}$ and $d_{n,B}$ do not provide reasonable approximations if the theoretical null distribution fails to work. Therefore the results based on $\lambda_{n,B}$ and $d_{n,B}$ are questionable. However, the EM-test, $\lambda_{n,N}$, and $d_{n,N}$ are invariant to the scale transformation and hence the conclusion based on these three tests are more trustable.

4.5 Proof

Since the EM-test $EM_n^{(K)}$ is invariant to the scale transformation, without loss of generality, we assume that under the null hypothesis the true distribution is $N(0,1)$. All the derivations are under this distribution.

The roadmap of the proof for Theorem 4.1 is similar to that of Theorem 3.1. We first prove two useful technical lemmas. Lemma 4.1 shows the consistency of $(\alpha_j^{(K)}, \mu_j^{(K)}, \sigma_{j,1}^{(K)}, \sigma_{j,2}^{(K)})$ and Lemma 4.2 derives an upper bound for the modified log-likelihood difference, which will be used to derive an upper bound for $EM_n^{(K)}$. Next, we show the upper bound of $EM_n^{(K)}$ is achievable and derive the limiting distribution of $EM_n^{(K)}$.

Lemma 4.1: *Suppose Conditions D1-D4 are satisfied. Then under the null distribution $N(0,1)$, we have, for $j = 1, 2, \ldots, J$ and any $k \leq K$,*

$$\alpha_j^{(k)} - \alpha_j = o_p(1),\ \mu_j^{(k)} = o_p(1),\ \sigma_{j,1}^{(k)} - 1 = o_p(1) \text{ and } \sigma_{j,2}^{(k)} - 1 = o_p(1).$$

Proof. The proof is similar to that of Lemma 6, Lemma 7, and Theorem 3 in Chen & Li (2009). Thus it is omitted. □

The next lemma concerns the upper bound of the modified log-likelihood difference when $(\mu, \sigma_1, \sigma_2)$ are in small neighbourhood of the true values. For $i = 1, 2, \ldots, n$, we define

$$Z_i = \frac{X_i^2 - 1}{2},\ U_i = \frac{X_i^3 - 3X_i}{6},\ \text{and } V_i = \frac{X_i^4 - 6X_i^2 + 3}{24}.$$

These notation are the same as those in Section 3.5.1.

Lemma 4.2: *Assume that the same conditions in Lemma 4.1 hold. Suppose*

$(\bar{\alpha}, \bar{\mu}, \bar{\sigma}_1, \bar{\sigma}_2)$ are estimators of $(\alpha, \mu, \sigma_1, \sigma_2)$ such that $(\bar{\mu}, \bar{\sigma}_1, \bar{\sigma}_2) = (0, 1, 1) + o_p(1)$ and $\bar{\alpha} \in (\delta, 1-\delta)$ for some $\delta > 0$. Under the null distribution $N(0,1)$, we have

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

Proof. Let

$$r_{1n}(\alpha, \mu, \sigma_1, \sigma_2) = 2\{l_n(\alpha, \mu, \sigma_1, \sigma_2) - l_n(1, 0, 1, 1)\}.$$

Then

$$\begin{aligned} & 2\{pl_n(\bar{\alpha}, \bar{\mu}, \bar{\sigma}_1, \bar{\sigma}_2) - pl_n(1, 0, 1, 1)\} \\ = & \, r_{1n}(\bar{\alpha}, \bar{\mu}, \bar{\sigma}_1, \bar{\sigma}_2) + 2\{p_n(\bar{\sigma}_1) + p_n(\bar{\sigma}_2) + p(\bar{\alpha}) - 2p_n(1) - p(1)\}. \quad (4.5) \end{aligned}$$

The upper bounds for the two terms in the above summation will be assessed separately.

We first consider $r_{1n}(\bar{\alpha}, \bar{\mu}, \bar{\sigma}_1, \bar{\sigma}_2)$. From (A.20) in Chen & Li (2008), we directly have

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

In this inequality, $\bar{t}_l$ are defined by

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

where $\bar{m}_{l,s}$ are the first four moments of the mixing distribution such that

$$\bar{m}_{l,s} = (1-\bar{\alpha})0^l(\bar{\sigma}_1^2 - 1)^s + \bar{\alpha}\bar{\mu}^l(\bar{\sigma}_2^2 - 1)^s.$$

After some simple algebra calculations, we have the following simpler forms of $\bar{t}_l$:

$$\bar{t}_1 = \bar{\alpha}\bar{\mu},\ \bar{t}_2 = \tilde{t}_2 + o_p(\bar{t}_1),\ \bar{t}_3 = o_p(\bar{t}_1),\ \bar{t}_4 = \tilde{t}_4 + o_p(\bar{t}_1),$$

where $\tilde{t}_2 = (1-\bar{\alpha})(\bar{\sigma}_1^2-1)+\bar{\alpha}(\bar{\sigma}_2^2-1)$ and $\tilde{t}_4 = 3\{(1-\bar{\alpha})(\bar{\sigma}_1^2-1)^2+\bar{\alpha}(\bar{\sigma}_2^2-1)^2\}$. Hence

$$\begin{aligned} r_{1n}(\bar{\alpha},\bar{\mu},\bar{\sigma}_1,\bar{\sigma}_2) \le\ & 2\Big\{\bar{t}_1\sum_{i=1}^n X_i + \tilde{t}_2\sum_{i=1}^n Z_i + \tilde{t}_4\sum_{i=1}^n V_i\Big\} \\ & -\Big\{\bar{t}_1^2\sum_{i=1}^n X_i^2 + \tilde{t}_2^2\sum_{i=1}^n Z_i^2 + \tilde{t}_4^2\sum_{i=1}^n V_i^2\Big\}\{1+o_p(1)\} \\ & +o_p(1). \end{aligned} \tag{4.6}$$

We now assess the upper bound for $2\{p_n(\bar{\sigma}_1)+p_n(\bar{\sigma}_2)+p(\bar{\alpha})-2p_n(1)-p(1)\}$. Similar to (3.2), using Conditions D1 and D3, we have

$$\begin{aligned} & 2\{p_n(\bar{\sigma}_1) + p_n(\bar{\sigma}_2) + p(\bar{\alpha}) - 2p_n(1) - p(1)\} \\ & \le 2\{p_n(\bar{\sigma}_1) + p_n(\bar{\sigma}_2) - 2p_n(1) + p(\bar{\alpha})\} = o_p(n^{1/6})\{|\bar{\sigma}_1^2 - 1| + |\bar{\sigma}_2^2 - 1|\} + 2p(\bar{\alpha}) \\ & \le 2p(\bar{\alpha}) + o_p(n)\Big\{\tilde{t}_2^2 + \tilde{t}_4^2\Big\} + o_p(1). \end{aligned} \tag{4.7}$$

Combining (4.5)-(4.7), we get

$$
\begin{aligned}
&2\{pl_n(\bar{\alpha},\bar{\mu},\bar{\sigma}_1,\bar{\sigma}_2) - pl_n(1,0,1,1)\} \\
&\le 2\Big\{\bar{t}_1\sum_{i=1}^n X_i + \tilde{t}_2\sum_{i=1}^n Z_i + \tilde{t}_4\sum_{i=1}^n V_i\Big\} \\
&\quad -\Big\{\bar{t}_1^2\sum_{i=1}^n X_i^2 + \tilde{t}_2^2\sum_{i=1}^n Z_i^2 + \tilde{t}_4^2\sum_{i=1}^n V_i^2\Big\}\{1+o_p(1)\} + 2p(\bar{\alpha}) + o_p(1). \qquad (4.8)
\end{aligned}
$$

Define

$$
\begin{aligned}
&Q(t_1,t_2,t_4) \\
&= 2\Big\{t_1\sum_{i=1}^n X_i + t_2\sum_{i=1}^n Z_i + t_4\sum_{i=1}^n V_i\Big\} - \Big\{t_1^2\sum_{i=1}^n X_i^2 + t_2^2\sum_{i=1}^n Z_i^2 + t_4^2\sum_{i=1}^n V_i^2\Big\}
\end{aligned}
$$

as a function of (t_1,t_2,t_4) with $t_4 \ge 0$. With

$$
\hat{t}_1 = \frac{\sum_{i=1}^n X_i}{\sum_{i=1}^n X_i^2},\ \hat{t}_2 = \frac{\sum_{i=1}^n Z_i}{\sum_{i=1}^n Z_i^2},\ \text{and } \hat{t}_4 = \frac{(\sum_{i=1}^n V_i)^+}{\sum_{i=1}^n V_i^2}, \qquad (4.9)
$$

this quadratic function is maximized. Further the maximized value of $Q(t_1,t_2,t_4)$ is

$$
Q(\hat{t}_1,\hat{t}_2,\hat{t}_4) = \frac{(\sum_{i=1}^n X_i)^2}{\sum_{i=1}^n X_i^2} + \frac{(\sum_{i=1}^n Z_i)^2}{\sum_{i=1}^n Z_i^2} + \frac{\{(\sum_{i=1}^n V_i)^+\}^2}{\sum_{i=1}^n V_i^2}.
$$

Together with (4.8), we get

$$
\begin{aligned}
&2\{pl_n(\bar{\alpha},\bar{\mu},\bar{\sigma}_1,\bar{\sigma}_2) - pl_n(1,0,1,1)\} \\
&\le Q(\hat{t}_1,\hat{t}_2,\hat{t}_4) + 2p(\bar{\alpha}) + o_p(1) \\
&\le \frac{(\sum_{i=1}^n X_i)^2}{\sum_{i=1}^n X_i^2} + \frac{(\sum_{i=1}^n Z_i)^2}{\sum_{i=1}^n Z_i^2} + \frac{\{(\sum_{i=1}^n V_i)^+\}^2}{\sum_{i=1}^n V_i^2} + 2p(\bar{\alpha}) + o_p(1).
\end{aligned}
$$

This finishes the proof. □

We now move to the proof of Theorem 4.1. The consistency results in Lemma 4.1 enable us to apply Lemma 4.2 to $2\{pl_n(\alpha_j^{(K)}, \mu_j^{(K)}, \sigma_{j,1}^{(K)}, \sigma_{j,2}^{(K)}) - pl_n(1, 0, 1, 1)\}$. That is,

$$\begin{aligned}
&2\{pl_n(\alpha_j^{(K)}, \mu_j^{(K)}, \sigma_{j,1}^{(K)}, \sigma_{j,2}^{(K)}) - pl_n(1, 0, 1, 1)\} \\
&\leq \frac{(\sum_{i=1}^n X_i)^2}{\sum_{i=1}^n X_i^2} + \frac{(\sum_{i=1}^n Z_i)^2}{\sum_{i=1}^n Z_i^2} + \frac{\{(\sum_{i=1}^n V_i)^+\}^2}{\sum_{i=1}^n V_i^2} + 2p(\alpha_j) + o_p(1).
\end{aligned}$$

Note that classic theory for regular models implies

$$2\{pl_n(1, 0, \hat{\sigma}_0, \hat{\sigma}_0) - pl_n(1, 0, 1, 1)\} = \frac{(\sum_{i=1}^n Z_i)^2}{\sum_{i=1}^n Z_i^2} + o_p(1).$$

Hence

$$\begin{aligned}
M_n^{(K)}(\alpha_j) &= 2\{pl_n(\alpha_j^{(K)}, \mu_j^{(K)}, \sigma_{j,1}^{(K)}, \sigma_{j,2}^{(K)}) - pl_n(1, 0, \hat{\sigma}_0, \hat{\sigma}_0)\} \\
&= 2\{pl_n(\alpha_j^{(K)}, \mu_j^{(K)}, \sigma_{j,1}^{(K)}, \sigma_{j,2}^{(K)}) - pl_n(1, 0, 1, 1)\} \\
&\quad -2\{pl_n(1, 0, \hat{\sigma}_0, \hat{\sigma}_0) - pl_n(1, 0, 1, 1)\} \\
&\leq \frac{(\sum_{i=1}^n X_i)^2}{\sum_{i=1}^n X_i^2} + \frac{\{(\sum_{i=1}^n V_i)^+\}^2}{\sum_{i=1}^n V_i^2} + 2p(\alpha_j) + o_p(1).
\end{aligned}$$

The upper bound of $EM_n^{(K)}$ is then given by

$$EM_n^{(K)} \leq \frac{(\sum_{i=1}^n X_i)^2}{\sum_{i=1}^n X_i^2} + \frac{\{(\sum_{i=1}^n V_i)^+\}^2}{\sum_{i=1}^n V_i^2} + 2\max_j p(\alpha_j) + o_p(1). \quad (4.10)$$

Next, we show that the upper bound in (4.10) is achievable. Since the EM-iteration increases the modified likelihood (Dempster, Laird, & Rubin, 1977), we only need to show that this is the case when $K = 1$. If suffices to find a set of parameter values $\hat{\alpha}$ and $(\hat{\mu}, \hat{\sigma}_1^2, \hat{\sigma}_2^2)$ at which the upper bound (4.10) is

attained.

We first choose $\hat{\alpha}$ such that $p(\hat{\alpha}) = \max_j p(\alpha_j)$. Without loss of generality, we assume that $\hat{\alpha} = \alpha_1$. We next choose $(\hat{\mu}, \hat{\sigma}_1^2, \hat{\sigma}_2^2)$ such that

$$\begin{cases} \hat{\alpha}\hat{\mu} = \hat{t}_1, \\ (1-\hat{\alpha})(\hat{\sigma}_1^2 - 1) + \hat{\alpha}(\hat{\sigma}_2^2 - 1) = \hat{t}_2, \\ 3\{(1-\hat{\alpha})(\hat{\sigma}_1^2 - 1)^2 + \hat{\alpha}(\hat{\sigma}_2^2 - 1)^2\} = \hat{t_4}, \end{cases}$$

where the expressions of $\hat{t}_l$ are given in (4.9). It can be checked that $(\hat{\mu}, \hat{\sigma}_1^2, \hat{\sigma}_2^2)$ exists and

$$\hat{\mu} = O_p(n^{-1/2}),\ \hat{\sigma}_1^2 - 1 = O_p(n^{-1/4}),\ \hat{\sigma}_2^2 - 1 = O_p(n^{-1/4}).$$

With these order information, we obtain

$$\begin{aligned} & 2\{pl_n(\hat{\alpha}, \hat{\mu}, \hat{\sigma}_1^2, \hat{\sigma}_2^2) - pl_n(1, 0, \hat{\sigma}_0, \hat{\sigma}_0)\} \\ = & \frac{(\sum_{i=1}^n X_i)^2}{\sum_{i=1}^n X_i^2} + \frac{\{(\sum_{i=1}^n V_i)^+\}^2}{\sum_{i=1}^n V_i^2} + 2p(\hat{\alpha}) + o_p(1). \end{aligned}$$

Note that $\hat{\alpha} = \alpha_1$ and $p(\hat{\alpha}) = p(\alpha_1) = \max_j p(\alpha_j)$. Thus for $EM_n^{(K)}$, we have

$$\begin{aligned} EM_n^{(K)} \geq M_n^{(1)}(\alpha_1) & \geq 2\{\sup_{\mu,\sigma_1,\sigma_2} pl_n(\alpha_1, \mu, \sigma_1, \sigma_2) - pl_n(1, 0, \hat{\sigma}_0, \hat{\sigma}_0)\} \\ & \geq 2\{pl_n(\hat{\alpha}, \hat{\mu}, \hat{\sigma}_1^2, \hat{\sigma}_2^2) - pl_n(1, 0, \hat{\sigma}_0, \hat{\sigma}_0)\} \\ & = \frac{(\sum_{i=1}^n X_i)^2}{\sum_{i=1}^n X_i^2} + \frac{\{(\sum_{i=1}^n V_i)^+\}^2}{\sum_{i=1}^n V_i^2} + 2\max_j p(\alpha_j) + o_p(1). \end{aligned}$$

This shows the asymptotic upper bound of $EM_n^{(K)}$ is also the asymptotic lower

bound, which implies

$$EM_n^{(K)} = \frac{(\sum_{i=1}^n X_i)^2}{\sum_{i=1}^n X_i^2} + \frac{\{(\sum_{i=1}^n V_i)^+\}^2}{\sum_{i=1}^n V_i^2} + 2\max_j p(\alpha_j) + o_p(1).$$

By central limit theorem, both $\sum_{i=1}^n X_i \Big/ \sqrt{\sum_{i=1}^n X_i^2}$ and $\sum_{i=1}^n V_i \Big/ \sqrt{\sum_{i=1}^n V_i^2}$ converge in distribution to $N(0,1)$. Further X_i and V_i are uncorrelated. Therefore, $EM_n^{(K)}$ asymptotically follows the distribution

$$\frac{1}{2}\chi_1^2 + \frac{1}{2}\chi_2^2 + 2\max_j p(\alpha_j).$$

This finishes the proof. □

Chapter 5

Summary and Future Work

In this chapter, we first summarize the main contributions of the thesis and then outline some future research problems.

5.1 Summary of the Thesis

In this thesis, we have considered testing homogeneity in three classes of finite mixture models: multivariate mixture models, the scale mixtures of normal distributions, and a new class of contaminated normal models in (4.1).

In Chapter 2, we propose the use of the EM-test for testing homogeneity in multivariate mixture models. We derived the limiting distribution of the EM-test statistic. Based on that, a resampling procedure is designed to approximate the p-values of the EM-test. Simulation studies show that the EM-test has accurate type-I error and adequate power, and is more powerful and computationally efficient than the bootstrap likelihood ratio test.

In Chapter 3, we retool the EM-test proposed in Chen & Li (2009) for testing homogeneity in scale mixtures of normal distributions. We show that

the retooled EM-test has the simple limiting distribution $\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$. We also use a computational method to provide an empirical formula for the tuning parameter selection. Simulation studies show that the retooled EM-test has an accurate size and is more powerful than existing methods when the data is generated from the scale mixtures of normal distributions. Further the retooled EM-test has adequate power and sometimes is more powerful than other methods even when the two component means in normal mixture models are slightly different.

In Chapter 4, we propose a class of contaminated normal models for modelling the z-scores in large-scale hypothesis testing problem and develop a new EM-test for homogeneity in this model. We show that the EM-test statistic asymptotically has simple shifted $\frac{1}{2}\chi^2_1+\frac{1}{2}\chi^2_2$ distribution. Extensive simulation studies show that, the proposed testing procedure has accurate type-I error and prominent powers for detecting heterogeneity. Further the new EM-test also compares favourably to other existing methods.

5.2 Future Work

In this section, we outline some future research problems.

Strong identifiability of a multivariate mixture model

As we discussed in Section 2.5, the developed theory for the EM-test in multivariate mixture models can be applied only if the kernel density $f(x;\theta)$ satisfies Condition A5, a weaker version of the strong identifiability condition (Chen, 1995). Verifying this condition in practice is a tedious work. Chen (1995) and Holzmann, Munk, & Stratmann (2004) used the characteristic function

technique to show that many univariate mixture models and von Mises mixture model are strongly identifiable. In the future, we would like to develop some general theory about the strong identifiability of multivariate mixture models. Based on that, we plan to find some easy-to-use criterion for checking the strong identifiability of multivariate mixture models. Further, developing the asymptotic distribution of the EM-test when the model is not strongly identifiable is another interesting topic.

Testing the order of a multivariate mixture model

Testing the order of a multivariate mixture model is an interesting, important, and more general hypothesis testing problem. In the literature, Dacunha-Castelle & Gassiant (1999) and Liu & Shao (2003) have derived the limiting distribution of the LRT for testing

$$H_0 : \ m = p \text{ versus } H_A : \ m = q$$

with $q > p$. In general, the limiting distribution involves the supremum of a Gaussian process, which may not be easy to use in practice. Chen, Chen, & Kalbfleisch (2004) proposed a MLRT for testing

$$H_0 : \ m = 2 \text{ versus } H_A : \ m > 2$$

in univariate mixture models. The limiting distribution of the MLRT is shown to be a mixture of χ^2-distributions. Recently Li & Chen (2010) and Chen, Li,

& Fu (2012) proposed two classes of EM-tests for testing

$$H_0 : \ m = m_0 \text{ versus } H_A : \ m > m_0$$

in univariate mixture models and in normal mixture models with unknown component variances. They showed that for any positive integer m_0, the limiting distributions of the EM-tests are the χ^2-type. We expect that the EM-test idea will be effective in developing a convenient statistical procedure for testing the order of a multivariate mixture model. We plan to investigate the possibility in details in the future.

Homogeneity test in general contaminated models

The contaminated models have many applications. The contaminated exponential distributions and more generally the contaminated Gamma distributions have been used in software reliability analysis, see Slud (1997) and Liu, Pasarica, & Shao (2003). The contaminated von Mises distributions have been used in paleoflow direction study, see Grimshaw, Whiting, & Morris (2001) and Fu, Chen, & Li (2008).

A contaminated model takes the following model:

$$(1 - \alpha) f(x; \beta_0, \gamma_1) + \alpha f(x; \beta, \gamma_2) \tag{5.1}$$

with β_0 being known and $(\beta, \gamma_1, \gamma_2)$ being unknown. The contaminated normal model is a special case of the above formulation. Testing homogeneity in the contaminated model is an important problem, see the discussion in Chapter 4, Liu, Pasarica, & Shao (2003), and Fu, Chen, & Li (2008). Due to the success

of the EM-test in Chapter 4, we plan to further apply the idea of EM-test to test homogeneity in the general contaminated model in (5.1).

EM-test in finite mixture of regression models

Finite mixture of regression (FMR) model , which is a natural extension of finite mixture model to incorporate the covariate information, has been widely used in many areas such as machine learning (Jacobs et al., 1991; Jiang & Tanner, 1999a, 1999b), finance and social science (Kamakura et al., 2003; Skrondal & Rabe-Hesketh, 2004), clinical, medical and psychological studies (Schlattmann, 2009), and so on.

In the literature, the identifiability issue in FMR models was addressed by Hennig (2000). The variable selection problems in FMR models were explored through AIC and BIC in 1990s. Recently, Khalili & Chen (2007) proposed the new variable selection approaches using the least absolute shrinkage and selection operator (LASSO; Tibshirani, 1996) and the smoothly clipped absolute deviation (SCAD; Fan & Li, 2001). The study on testing the order of a FMR model is quite limited. Zhu & Zhang (2004) investigated the asymptotic distributions of the LRT and the MLRT for testing homogeneity in FRM models. Dai & Charnigo (2007) considered testing homogeneity in a contaminated regression model using the MLRT and the D-test. Applying the idea of EM-test for testing homogeneity or more generally testing the order of a FMR model is an interesting research problem. Continuing effort on this problem is part of my future research plan.

Bibliography

[1] Aitchison, J., & Ho, C. H. (1989). The multivariate Poisson-log normal distribution. *Biometrika*, 76, 643–653.

[2] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second international symposium on information theory* (pp. 267–281). Akademinai Kiado.

[3] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716–723.

[4] Allison, D. B., Gadbury, G. L., Heo, M., Fernández, J. R., Lee, C. K., Prolla, T. A., & Weindruch, R. (2002). A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics & Data Analysis*, 39, 1–20.

[5] Andrews, D. F., & Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36, 99–102.

[6] Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57, 289–300.

[7] Biernacki, C., Celeux, G., & Govaert, G. (1998). Assessing a mixture model for clustering with the integrated classification likelihood. *Technical Report No. 3521.*

[8] Biernacki, C., & Govaert, G. (1997). Using the classification likelihood to choose the number of clusters. *Computing Science and Statistics*, 451–457.

[9] Boothe, P., & Glassman, D. (1987). The statistical distribution of exchange rates. *Journal of International Economics*, 22, 297–319.

[10] Bozdogan, H. (1990). On the information-based measure of covariance complexity and its application to the evaluation of multivariate linear models. *Communications in Statistics-Theory and Methods*, 19, 221–278.

[11] Bozdogan, H. (1993). Choosing the number of component clusters in the mixture-model using a new informational complexity criterion of the inverse-Fisher information matrix. In *Information and classification* (pp. 40–54). Springer Berlin Heidelberg.

[12] Cadez, I. V., Smyth, P., & Mannila, H. (2001). Probabilistic modeling of transaction data with applications to profiling, visualization and prediction. In: F. Provost and R. Srikant, *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco (CA), USA, 37–46.

[13] Campbell, J. G., Fraley, C., Murtagh, F., & Raftery, A. E. (1997). Linear flaw detection in woven textiles using model-based clustering. *Pattern Recognition Letters*, 18, 1539–1548.

[14] Carlin, B. P., & Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, 473–484.

[15] Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of classification*, 13, 195–212.

[16] Charnigo, R., & Pilla, R. S. (2007). Semiparametric mixtures of generalized exponential families. *Scandinavian Journal of Statistics*, 34, 535–551.

[17] Charnigo, R., & 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.

[18] Charnigo, R., & Sun, J. (2008). Testing homogeneity in discrete mixtures. *Journal of Statistical Planning and Inference*, 138(5), 1368–1388.

[19] Charnigo, R., & Sun, J. (2010). Asymptotic relationships between the D-test and likelihood ratio-type tests for homogeneity. *Statistica Sinica*, 20(2), 497.

[20] Charnigo, R., Zhou, F., & Dai, H. (2013). Contaminated Chi-Square Modeling and Large-Scale ANOVA Testing. *Journal of Biometrics & Biostatistics*.

[21] Chen, H., & Chen, J. (2001). The likelihood ratio test for homogeneity in the finite mixture models. *The Canadian Journal of Statistics*, 29, 201–215.

[22] Chen, H., Chen, J., & Kalbfleisch, J. D.(2001). A modified likelihood ratio test for homogeneity in finite mixture models. *Journal of the Royal Statistical Society, Series B*, 63, 19–29.

[23] Chen, H., Chen, J., & Kalbfleisch, J. D. (2004). Testing for a finite mixture model with two components. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66, 95–115.

[24] Chen, J. (1995). Optimal rate of convergence for finite mixture models. *The Annals of Statistics*, 23, 221–233.

[25] Chen, J. (1998). Penalized likelihood ratio test for finite mixture models with multinomial observations. *The Canadian Journal of Statistics*, 26, 583–599.

[26] Chen, J., & Kalbfleisch, J. D. (1996). Penalized minimum-distance estimates in finite mixture models. *Canadian Journal of Statistics*, 24, 167–175.

[27] Chen, J., & 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.

[28] Chen, J., & Khalili, A. (2008). Order selection in finite mixture models with a nonsmooth penalty. *Journal of the American Statistical Association*, 103, 1674–1683.

[29] Chen, J., & Li, P. (2008). Hypothesis test for normal mixture models: The EM approach. *Technical report*, University of British Columbia.

[30] Chen, J., & Li, P. (2009). Hypothesis test for normal mixture models: The EM approach. *The Annals of Statistics*, 37, 2523–2542.

[31] Chen, J., & Li, P. (2011). Tuning the EM-test for finite mixture models. *Canadian Journal of Statistics*, 39, 389–404.

[32] Chen, J., Li, P., & Fu, Y. (2012). Inference on the order of a normal mixture. *Journal of American Statistical Association*, 107, 1096–1105.

[33] Chen, J., Tan, X., & Zhang, R. (2008). Inference for normal mixtures in mean and variance. *Statistica Sinica*, 18, 443–465.

[34] Cruz-Medina, I. R., Hettmansperger, T. P., & Thomas, H. (2004). Semi-parametric mixture models and repeated measures: The multinomial cut point model. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 53, Part 3, 463–474.

[35] Dacunha-Castelle, D., & 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.

[36] Dai, H., & Charnigo, R. (2007). Inferences in contaminated regression and density models. *Sankhyā: The Indian Journal of Statistics*, 842–869.

[37] Dai, H., & Charnigo, R. (2008). Omnibus testing and gene filtration in microarray data analysis. *Journal of Applied Statistics*, 35, 31–47.

[38] Dai, H., & Charnigo, R. (2010). Contaminated normal modeling with application to microarray data analysis. *Canadian Journal of Statistics*, 38, 315–332.

[39] Dasgupta, A., & Raftery, A. E. (1998). Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association*, 93, 294–302.

[40] Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)*, 39, 1–38.

[41] Doulgeris, A. P., & Eltoft, T. (2010). Scale mixture of Gaussian modelling of polarimetric SAR data. *Hindawi Publishing Corporation, EURASIP Journal on Advances in Signal Processing*, Article ID 874592.

[42] Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The annals of Statistics*, 7, 1–26.

[43] Efron, B. (2004). Large-scale simultaneous hypothesis testing. *Journal of the American Statistical Association*, 99, 96–104.

[44] Efron, B. (2010). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction* (Vol. 1). Cambridge University Press.

[45] Everitt, B. S., Landau, S., & Leese, M. (2001). *Cluster Analysis*, 4th ed. Arnold, London.

[46] Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96, 1348–1360.

[47] Fraley, C., & Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The computer journal*, 41, 578–588.

[48] Frühwirth-Schnatter, S. (2004). Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques. *The Econometrics Journal*, 7, 143–167.

[49] Fu, Y., Chen, J., & Li, P. (2008). Modified likelihood ratio test for homogeneity in a mixture of von Mises distributions. *Journal of Statistical Planning and Inference*, 138, 667–681.

[50] Fujisawa, H., & Eguchi, S. (2006). Robust estimation in the normal mixture model. *Journal of Statistical Planning and Inference*, 136, 3989–4011.

[51] Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82, 711–732.

[52] Grimshaw, S. D., Whiting, D. G., & Morris, T. H. (2001). Likelihood ratio tests for a mixture of two von Mises distributions. *Biometrics*, 57, 260–265.

[53] Hall, J. A., Brorsen, W., & Irwin, S. H. (1989). The distribution of futures prices: A test of the stable Paretian and mixture of normals hypotheses. *Journal of Financial and Quantitative Analysis*, 24, 105–116.

[54] Harrison, G. W. (2001). Implications of mixed exponential occupancy distributions and patient flow models for health care planning. *Health Care Management Science*, 4, 37–45.

[55] Harrison, G. W., & Millard, P. H. (1991). Balancing acute and long-term care: the mathematics of throughput in departments of geriatric medicine. *Methods of information in medicine*, 30, 221–228.

[56] Hartigan, J. A. (1985). A failure of likelihood asymptotics for normal mixtures. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer* (Vol. 2, pp. 807–810). Wadsworth, Belmont, CA.

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

[58] Hennig, C. (2000). Identifiablity of models for clusterwise linear regression. *Journal of Classification*, 17, 273–296.

[59] Holzmann, H., Munk, A., & Stratmann, B. (2004). Identifiability of finite mixtures-with applications to circular distributions. *Sankhyā: The Indian Journal of Statistics*, 440–449.

[60] Huang, T., Peng, H., & Zhang, K. (2013). Model Selection for Gaussian Mixture Models. *arXiv preprint arXiv:1301.3558*.

[61] Ishiguro, M., Sakamoto, Y., & Kitagawa, G. (1997). Bootstrapping log-likelihood and EIC, an extension of AIC. *Annals of the Institute of Statistical Mathematics*, 49, 411–434.

[62] Ishwaran, H., James, L. F., & Sun, J. (2001). Bayesian model selection in finite mixtures by marginal density decompositions. *Journal of the American Statistical Association*, 96, 1316–1332.

[63] Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural computation*, 3, 79–87.

[64] James, L. F., Priebe, C. E., & Marchette, D. J. (2001). Consistent estimation of mixture complexity. *The Annals of Statistics*, 29, 1281–1296.

[65] Jiang, W., & Tanner, M. A. (1999a). Hierarchical mixtures-of-experts for exponential family regression models: approximation and maximum likelihood estimation. *Annals of Statistics*, 987–1011.

[66] Jiang, W., & Tanner, M. A. (1999b). On the approximation rate of hierarchical mixtures-of-experts for generalized linear models. *Neural computation*, 11, 1183–1198.

[67] Johnson, N., Kotz, S., & Balakrishnan, N. (1997). *Discrete Multivariate Distributions*. Wiley, New York.

[68] Kamakura, W. A., Wedel, M., De Rosa, F., & Mazzon, J. A. (2003). Cross-selling through database marketing: a mixed data factor analyzer for data augmentation and prediction. *International Journal of Research in marketing*, 20, 45–65.

[69] Karlis, D., & Meligkotsidou, L. (2007). Finite mixtures of multivariate Poisson distributions with application. *Journal of Statistical Planning and Inference*, 137, 1942–1960.

[70] Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā: The Indian Journal of Statistics, Series A*, 62, 49–66.

[71] Khalili, A., & Chen, J. (2007). Variable selection in finite mixture of regression models. *Journal of the american Statistical association*, 102, 1025–1038.

[72] Kopocinski, B. (1999). Multivariate negative binomial distributions generated by multivariate exponential distributions. *Applicationes mathematicae*, 25, 463–472.

[73] Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22, 79–86.

[74] Leroux, B. G. (1992). Consistent estimation of a mixing distribution. *The Annals of Statistics*, 20, 1350–1360.

[75] Lewine, R. R. J. (1981). Sex differences in schizophrenia: Timing or subtypes? *Psychological Bulletin*, 90, 432–444.

[76] Li, P., & Chen, J. (2010). Testing the order of a finite mixture. *Journal of the American Statistical Association*, 105, 1084–1092.

[77] Li, P., Chen, J., & Marriott, P. (2009). Non-finite Fisher information and homogeneity: The EM approach. *Biometrika*, 96, 411–426.

[78] Lindsay, B. G. (1995). Mixture models: theory, geometry and applications. In *NSF-CBMS regional conference series in probability and statistics* (pp. i–163). Institute of Mathematical Statistics and the American Statistical Association.

[79] Liu, X., Pasarica, C., & Shao, Y. (2003). Testing homogeneity in gamma mixture models. *Scandinavian Journal of Statistics*, 30, 227–239.

[80] Liu, X., & Shao, Y. Z. (2003). Asymptotics for likelihood ratio tests under loss of identifiability. *The Annals of Statistics*, 31, 807–832.

[81] McGrory, C. A., & Titterington, D. M. (2007). Variational approximations in Bayesian model selection for finite mixture distributions. *Computational Statistics & Data Analysis*, 51(11), 5352–5367.

[82] McLachlan, G. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 36, 318–324.

[83] McLachlan, G. J., Bean, R. W., & Jones, L. B. T. (2006). A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics*, 22(13), 1608–1615.

[84] McLachlan, G., & Peel, D. (2000). *Finite Mixture Models*. John Wiley and Sons, New York.

[85] Miloslavsky, M., & van der Laan, M. J. (2003). Fitting of mixtures with unspecified number of components using cross validation distance estimate. *Computational statistics & data analysis*, 41, 413–428.

[86] Morel, J. G., & Nagaraj, N. K. (1993). A finite mixture distribution for modelling multinomial extra variation. *Biometrika*, 80, 363–371.

[87] Naylor, J. C., & Smith, A. F. M. (1983). A contamination model in clinical chemistry: An illustration of a method for the efficient computation of posterior distributions. *Journal of the Royal Statistical Society: Series D*, 32, 82–87.

[88] Niu, X., Li, P., & Zhang, P. (2011). Testing homogeneity in a multivariate mixture model. *Canadian Journal of Statistics*, 39, 218–238.

[89] Pan, W. (1999). Bootstrapping likelihood for model selection with small samples. *Journal of Computational and Graphical Statistics*, 8, 687–698.

[90] Partrat, C. (1994). Compound model for two dependent kinds of claim. *Insurance: Mathematics and Economics*, 15, 219–231.

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

[92] Peng, X. (2003). Simultaneous inference and sample size considerations in microarray data analysis. *Doctoral dissertation, University of Kentucky.*

[93] R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN: 3-900051-07-0, URL http://www.R-project.org.

[94] Raftery, A. E. (1996). Hypothesis testing and model selection via posterior simulation. *Markov chain Monte Carlo in practice*, 163–188.

[95] Richardson, S., & Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (statistical methodology)*, 59, 731–792.

[96] Roberts, S. J., Husmeier, D., Rezek, I., & Penny, W. (1998). Bayesian approaches to Gaussian mixture modeling. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20, 1133–1142.

[97] Rocha, M. L., Pestana, D., & Menezes, A. G. (2012). Heavy tails and mixtures of normal random variables. *CEEAplA Working Paper No. 6.*

[98] Roeder, K., & Wasserman, L. (1997). Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, 92, 894–902.

[99] Roos, B. (2003). Poisson approximation of multivariate Poisson mixtures. *Journal of Applied Probability*, 40, 376–390.

[100] Schlattmann, P. (2009). *Medical applications of finite mixture models*. Springer.

[101] Schork, N. J., Allison, D. B., & Thiel, B. (1996). Mixture distributions in human genetics research. *Statistical Methods in Medical Research*, 5, 155–178.

[102] Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 6, 461–464.

[103] Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. CRC Press.

[104] Slud, E. (1997). Testing for imperfect debugging in software reliability. *Scandinavian journal of statistics*, 24, 555–572.

[105] Smyth, P. (2000). Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and Computing*, 10, 63–72.

[106] Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components-an alternative to reversible jump methods. *Annals of Statistics*, 28, 40–74.

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

[108] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

[109] Titterington, D. M., Smith, A. F., & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions* (Vol. 7). New York: Wiley.

[110] Wainwright, M. J., & Simoncelli, E. P. (2000). Scale mixtures of Gaussians and the statistics of natural images. *Advances in Neural Information Processing Systems*, 12, 855–861.

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

[112] Weldon, W. F. R. (1892). Certain correlated variations in Crangon vulgaris. *Proceedings of the Royal Society of London*, 51(308-314), 1–21.

[113] West, M. (1984). Outlier models and prior distributions in Bayesian linear regression. *Journal of the Royal Statistical Society: Series B*, 46, 431–439.

[114] West, M. (1987). On scale mixtures of normal distributions. *Biometrika*, 74, 646–648.

[115] Windham, M. P., & Cutler, A. (1992). Information ratios for validating mixture analyses. *Journal of the American Statistical Association*, 87, 1188–1192.

[116] Woo, M. J., & Sriram, T. N. (2006). Robust estimation of mixture complexity. *Journal of the American Statistical Association*, 101, 1475–1486.

[117] Woo, M. J., & Sriram, T. N. (2007). Robust estimation of mixture complexity for count data. *Computational statistics & data analysis*, 51(9), 4379–4392.

[118] Zhu, H. T., & Zhang, H. (2004). Hypothesis testing in mixture regression models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66, 3–16.

[119] Zucchini, W., & MacDonald, I. L. (2009). *Hidden Markov models for time series: an introduction using R*. CRC Press.