Statistical Theory and Related Fields

ISSN: 2475-4269 (Print) 2475-4277 (Online) Journal homepage: https://www.tandfonline.com/loi/tstf20

On finite mixture models

Jiahua Chen

To cite this article: Jiahua Chen (2017) On finite mixture models, Statistical Theory and Related Fields, 1:1, 15-27, DOI: 10.1080/24754269.2017.1321883

To link to this article: https://doi.org/10.1080/24754269.2017.1321883

STATISTICAL THEORY AND RELATED FIELDS, 2017
VOL. 1, NO. 1, 15–27
https://doi.org/10.1080/24754269.2017.1321883

Taylor & Francis
Taylor & Francis Group

On finite mixture models

Jiahua Chen[a,b]

[a]Research Institute of Big Data, Yunnan University, Yunnan, China; [b]Department of Statistics, University of British Columbia, Vancouver, Canada

ABSTRACT

Finite mixture models are widely used in scientific investigations. Due to their non-regularity, there are many technical challenges concerning inference problems on various aspects of the finite mixture models. After decades of effort by statisticians, substantial progresses are recorded recently in characterising large sample properties of some classical inference methods when applied to finite mixture models, providing effective numerical solutions for mixture model-based data analysis, and the invention of novel inference approaches. This paper aims to provide a comprehensive summary on large sample properties of some classical statistical methods and recently developed modified likelihood ratio test and EM-test for the order of the finite mixture model. The presentation de-emphasises the rigour in order to gain some insights behind some complex technical issues. The paper wishes to recommend the EM-test as the most promising approach to data analysis problems from all models with mixture structures.

ARTICLE HISTORY
Received 15 March 2017
Accepted 19 April 2017

KEYWORDS
C(α)-test; EM-test; hidden Markov model; homogeneity; modified likelihood ratio test; structural parameter

1. Introduction

Let $f(x;\theta)$ for each θ in a parameter space Θ be a density function with respect to some σ-finite measure. A parametric distribution family is formed as the collection of distributions $\{f(x;\theta)\colon \theta \in \Theta\}$. Naturally, a family is not formed by an arbitrary collection of distributions. Classical distribution families contain distributions connected through some common scientific background. For instance, a binomial distribution family is made of distributions modelling the number of successes in a fixed number of independent trials repeated under identical conditions.

A finite mixture model builds on a classical distribution family so that its density functions are finite convex combinations of the densities in some parametric family $\{f(x;\theta)\colon \theta \in \Theta\}$:

$$f(x;G) = \sum_{j=1}^{m} \pi_j f(x;\theta_j) = \int_{\Theta} f(x;\theta) dG(\theta). \quad (1)$$

Let $\mathbb{1}(\cdot)$ be an indicator function. The mixing distribution G in the above definition refers to its cumulative distribution function (c.d.f.) or its probability masses on some support points:

$$G(\theta) = \sum_{j=1}^{m} \pi_j \mathbb{1}(\theta_j \le \theta) = \sum_{j=1}^{m} \pi_j \{\theta_j\}. \quad (2)$$

Clearly, even if G has continuous support rather than being a finite discrete distribution, $f(x;G)$ remains a well-defined density function. We focus on inference problems when G has form (2).

A distribution is also a mathematical way to characterise a population. Imagine taking some measurements of interest on a random unit from a population. The distribution of these measurements represents one aspect of this population. Suppose a population is made of m subpopulations, each equating a distribution in a parametric family $\{f(x;\theta)\colon \theta \in \Theta\}$. A random unit from such a population is also a unit from one of these subpopulations: with probability π_j for the jth subpopulation, $j = 1, 2, \ldots, m$. Without knowing the subpopulation, we work with the marginal distribution of this unit which is (1).

Finite mixture model finds its applications in a wide range of disciplines and goes back deep into history. When a biological population made of a single species has reached equilibrium, the random variations between individuals are then completely attributed to cumulative effect of numerous minor factors. The resulting uncertainty is therefore well approximated by a normal distribution, according to the classical central limit theory. Normal model is hence broadly assumed in biometrics. If a data set displays non-normality, a finite mixture model is a natural alternative and extension.

Figure 1 contains, among others, the histogram of a data set containing measurements of 1000 crabs sampled from Bay of Naples provided to Pearson by a biologist (Pearson, 1894). The histogram displays an apparent departure from normality. The departure can be sensibly explained by the possibility that the crab population contains two subpopulations (species). If so, a two-component finite normal mixture model should fit the data well. Indeed, Pearson (1894) found a

CONTACT Jiahua Chen jhchen@stat.ubc.ca

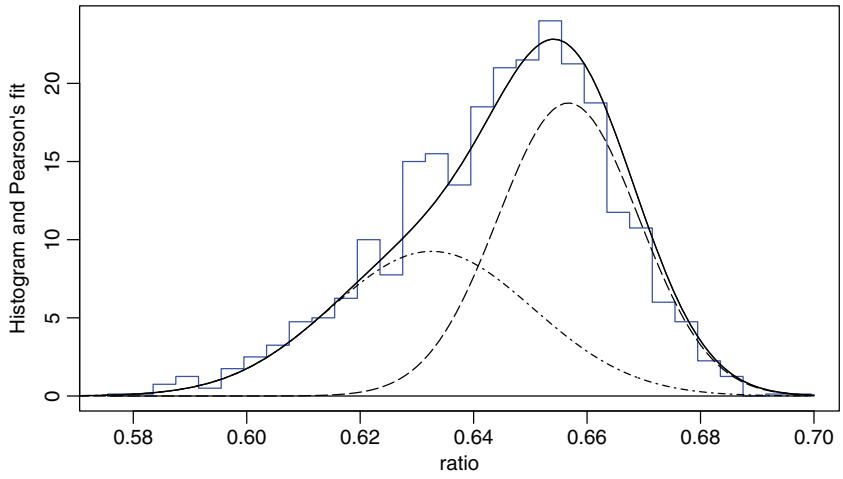

Figure 1. Histogram of the crab data and the fitted density of two-component normal mixture.

satisfactory fit via the method of moments to the data set by a finite normal mixture model of order $m = 2$. We remark that the data can also be well fitted by a Gamma distribution. A finite normal mixture is preferred not for its fit, but for its natural justification.

Based on the same interpretation as in Pearson (1894), finite mixture models are routinely used to accommodate the genetic heterogeneity thought to underlie many human diseases. See Chernoff and Lander (1995), Friedlander and Leitersdorf (1995), Schork, Allison, and Thiel (1996) and Ott (1999) for many examples. There are also abundant application examples in Titterington, Smith, and Makov (1985), McLachlan and Peel (2004), and Frühwirth-Schnatter (2006) in other disciplines. We further recommend Lindsay (1995) for some insightful discussions on mixture models.

In addition to their importance in applications, mixture models form a fertile and challenging field of statistical research. The most urgent tasks include point estimation of the mixing distribution G and the related numerical calculation. Interestingly, under mild conditions on $\{f(x;\theta)\colon \Theta\}$, even the nonparametric maximum likelihood estimator (MLE) of G is consistent, given a set of independent and identically distributed (i.i.d.) observations (Chen, 2016; Kiefer & Wolfowitz, 1956). Geometric properties of the nonparametric MLE were nicely illustrated in Lindsay (1995). These discussions form the solid foundation for numerical computations (Böhning, 2000).

This paper focuses on the hypothesis test problem for order m of the finite mixture model. Due to partial loss of identifiability and its induced non-regularity, the likelihood ratio test (LRT) losses its well-behaved large sample property represented by the famous Wilks theorem (Wilks, 1938). The research activities on this topic may be divided into several stages. Early researchers such as Hartigan (1985) and Ghosh and Sen (1985) revealed the ill effect of the non-regularity. Subsequent research contains rigorous mathematical answers to the asymptotic properties. Some representative results are Chernoff and Lander (1995), Dacunha-Castelle and Gassiat (1999), Liu and Shao (2003) and Chen and Chen (2003). These results are interesting but also conclude that the straightforward LRT is generally impractical. For instance, there are currently no effective numerical methods to evaluate the p-value according to these limiting distributions.

One way to circumvent this obstacle is the famous C(α) test proposed by Neyman and Scott (1965) which is locally optimal and easy to use. However, it is limited to test for homogeneity: namely for H_0: $m = 1$ when the population contains only one subpopulation. Another line of approach was proposed by Chen (1998) and further developed in Chen, Chen, and Kalbfleisch (2001) and Chen, Chen, and Kalbfleisch (2004). This approach partially restores the regularity through placing a soft bound on mixing proportions. The modified likelihood ratio test (mLRT), as it is now called, leads to Wilks-like asymptotic properties for many but still limited number of commonly used mixture models.

The latest invention, EM-test, enjoys many advantages over the mLRT. When the mLRT works, the EM-test shares its asymptotic properties. The Wilks-like asymptotic properties of the EM-test hold much more broadly. The design of the EM-test comes with many build-in flexibilities. They allow users to utilise features of the specific mixture model to find one version of the EM-test with Wilks-like asymptotic properties. We anticipate advances of the EM-test for multi-parameter mixture models, location-scale mixture models, hidden Markov models (HMMs) and other models with non-i.i.d. data.

This paper is organised as follows. In Section 2, we explain the non-regularity and its implication to classical results on the LRT for homogeneity. Section 3 is devoted to a classical C(α) test for homogeneity. Section 4 contains results on restricted LRT which seem to have

limited usage but form a conceptual step-stone for further development. Section 5 presents mLRT and EM-test for homogeneity. Sections 6–8 show how the idea of EM-test is developed to obtain interesting results for the order of finite mixture of single parameter distribution, normal distribution, or data from HMM. We end the paper with some discussions on the future development of the EM-test.

2. Properties of finite mixture models

2.1. Partial identifiability

A parametric model is viable in applications only if it is identifiable. Being identifiable means that

$$f(x;\theta_1) = f(x;\theta_2)$$

for almost all x (with respect to the underlying σ-finite measure) must imply $\theta_1 = \theta_2$. A mixture model is identifiable if

$$f(x;G_1) = f(x;G_2)$$

for almost all x implies $G_1 = G_2$. We may regard G in (2) as a functional valued parameter for mixture model (1). On this platform, most commonly used finite mixture models are identifiable. If G is allowed to be any distribution on Θ, the mixture model (1) is still identifiable for many important families of $f(x;\theta)$.

In applications, our interest goes beyond generic G to include its compositions. Under a finite mixture model, we call m the order of the mixture, π_j the mixing proportions and θ_j the component parameter values of the component distribution $f(x;\theta)$. Clearly, we have $\Sigma_j \pi_j = 1$ and $\pi_j \geq 0$. When any $\pi_j = 0$, then the imaginary jth subpopulation does not show up in the mixture. When $\theta_{j_1} = \theta_{j_2}$ for some $1 \leq j_1 \neq j_2 \leq m$, then these two subpopulations are the same. In both cases, the number of distinct subpopulations is reduced at least by 1 and the model has lost identifiability in this respect. We generally refer them as two types of partial loss of identifiability. They are largely responsible for abnormal asymptotic properties.

2.2. Likelihood ratio test for homogeneity

Let $X_1, \ldots, X_n$ be an i.i.d. sample from a finite mixture model of order $m = 2$:

$$\pi_1 f(x,\theta_1) + \pi_2 f(x,\theta_2). \tag{3}$$

The problem of testing for homogeneity is specified by two opposing hypotheses

$$H_0 : \pi_1\pi_2(\theta_2 - \theta_1) = 0 \ \text{ against}$$
$$H_1 : \pi_1\pi_2(\theta_2 - \theta_1) \neq 0. \tag{4}$$

LRT for homogeneity is the first choice, but we need a solid statistical foundation.

Let us first consider a simpler and more specific homogeneity problem (Hartigan, 1985):

$$H_0 : \ N(0,1) \text{ against } \ H_1 : (1-\pi)N(0,1) + \pi N(\theta,1) \tag{5}$$

for $\theta \in \mathcal{R}$. Its non-regularity is rooted in the fact that when $\pi = 0$, the model becomes a null model for any θ value. In fact, both π and θ only present in H_1, a situation of particular interest (Davies, 1977, 1987).

Under the full model, the log likelihood function is given by

$$\ell_n(\pi;\theta) = \sum_{i=1}^{n} \log \phi(X_i;0,1) + \sum_{i=1}^{n} \log[1 + \pi\{\exp(\theta X_i - \theta^2/2) - 1\}],$$

where $\phi(x;\mu,\sigma)$ is the density function of $N(\mu,\sigma^2)$. Define the log likelihood ratio function:

$$R_n(\pi;\theta) = 2\sum_{i=1}^{n} \log[1 + \pi\{\exp(\theta X_i - \theta^2/2) - 1\}].$$

Let

$$Y_i(\theta) = \exp(\theta X_i - \theta^2/2) - 1, \quad \bar{Y}_n(\theta) = n^{-1}\sum_{i=1}^{n} Y_i(\theta),$$

and $\sigma^2(\theta) = \text{VAR}(Y_i(\theta))$. For fixed θ, $Y_i(\theta)$ are i.i.d. with zero mean and finite variance $\sigma^2(\theta)$ under H_0. Some asymptotic algebra shows that,

$$\sup_{\pi} R_n(\pi;\theta) = n\sigma^{-2}(\theta)\{[\bar{Y}_n]^+(\theta)\}^2 + o_p(1) \tag{6}$$

$$\to 0.5\chi_0^2 + 0.5\chi_1^2, \tag{7}$$

where $[\bar{Y}_n]^+$ is the positive part of $\bar{Y}_n$ for each $\theta \neq 0$. The leading term in (6) is a quadratic approximation to the log likelihood ratio function $R_n(\pi;\theta)$.

The authentic LRT statistic for homogeneity without having θ fixed is clearly

$$R_n = \sup_{\theta}\{\sup_{\pi} R_n(\pi;\theta)\}.$$

Over any finite interval, say $\theta \in [0, M]$ with $M < \infty$,

$$n^{1/2}\sigma^{-1}(\theta)\bar{Y}_n(\theta)$$

is easily seen to converge in distribution to a well-defined Gaussian process $\{Z(\theta): 0 \leq \theta \leq M\}$. Let $\theta_1, \theta_2, \ldots, \theta_K$ be K arbitrarily selected values in $\mathcal{R}$. Then,

$$R_n \geq \max_{1\leq j\leq K}\{n\sigma^{-2}(\theta_j)\{[\bar{Y}_n]^+(\theta_j)\}^2 + o_p(1)\}$$
$$\to \max_{1\leq j\leq K}\{Z^+(\theta_j)\}^2.$$

By choosing sufficiently dispersed θ_j over $\mathcal{R}$ and large K, we can make

$$\max_{1\leq j\leq K}\{Z^+(\theta_j)\}^2$$

arbitrarily large in probability. This leads to the conclusion that $R_n \to \infty$ in probability, a result that bans the use of usual LRT for homogeneity.

The above result does not totally diminish the enthusiasm on LRT for homogeneity. One aberration observed in Hartigan's example can be avoided by placing a compact condition on Θ: $|\theta| \le M < \infty$. Under this additional condition, it can be shown that

$$R_n \to \sup_{0\le\theta\le M} \{Z^+(\theta)\}^2 \tag{8}$$

in distribution, a potentially applicable result. At the same time, a rigorous proof of (8) is not obvious because the quadratic approximation (6) is valid for each $\theta \neq 0$, but not uniformly over $\theta \in \Theta$.

The non-uniformity leads to a technical challenging issue, especially for the original homogeneity problem (4). To avoid this difficulty, Ghosh and Sen (1985) introduced a separation condition $|\theta_1 - \theta_2| \ge \epsilon$ for some $\epsilon > 0$ in addition to compact constraints $|\theta_1| \le M$, $|\theta_2| \le M$ for (4). Over the space of (θ_1, θ_2) such that $|\theta_1 - \theta_2| \ge \epsilon$, a uniform quadratic approximation similar to (6) is obtained. The corresponding LRT statistic was then shown to have a limiting distribution in the familiar form $\sup_{\epsilon<|\theta|\le M}\{Z^+(\theta)\}^2$. The Gaussian process $Z(\theta)$ and its range are decided by the data-generating distribution $f(x;\theta_0)$ and the specific component parameter space Θ.

Subsequent attempts were made to establish the limiting distribution without the separation condition. Chernoff and Lander (1995) were the first to succeed when the component distribution is binomial. Their technique is to transform the parameter space so that the space of H_1 becomes a cone and H_0 becomes a single point on the tip of H_1. Because the component parameter space Θ is naturally compact, they obtained the limiting distribution without placing artificial constraints.

For more general component distribution $f(x;\theta)$, Chen and Chen (2001) used a different approach. Similar to Ghosh and Sen (1985), they also obtained a uniform quadratic approximation to the likelihood function over $|\theta_1 - \theta_2| > \epsilon$. At the same time, they obtained both lower and upper bounds for the likelihood ratio function over $|\theta_1 - \theta_2| \le \epsilon$. For finite mixture models of binomial, Poisson, normal with known variance, the difference between the upper and lower bounds reduces to zero as $\epsilon \to 0$. The limit perfectly matches the quadratic approximation obtained earlier when applied to $\theta_1 = \theta_2$. Therefore, the LRT statistics indeed has a limiting distribution in the form of $\sup_{|\theta|\le M}\{Z^+(\theta)\}^2$. Since the limit at $\theta_1 = \theta_2$ is obtained by squeezing two bounds, the technique is named as sandwich method.

Interestingly, for normal mixture model of order $m = 2$ with structural variance parameter:

$$(1-\pi)N(\theta_1, \sigma^2) + \pi N(\theta_2, \sigma^2),$$

the limit obtained by sandwich method is stochastically a sum of two independent standard normally distributed random variables. Hence, the limiting distribution of the LRT statistic for homogeneity turns out to have a generic form

$$Z_0^2 + \sup_{|\theta|\le M} \{Z^+(\theta)\}^2.$$

See Chen and Chen (2003) for details.

Two key intermediate results established in Chen and Chen (2001) are: a uniform quadratic approximation to the likelihood ratio function over $|\theta_1 - \theta_2| > \epsilon$, and its smooth limit when $\epsilon \to 0$ coincides with the quadratic approximation within $|\theta_1 - \theta_2| \le \epsilon$. To some degree, the most general results on homogeneity test and beyond given by Dacunha-Castelle and Gassiat (1999) are obtained by streamline conditions that lead to validity of these two intermediate results.

3. C(α) test for homogeneity

In the past few decades, statisticians succeeded at determining the large sample properties of the LRT statistics. In most cases, these results do not lead to practical homogeneity tests due to numerical difficulty of computing p-values. A C(α) test developed for composite hypotheses by Neyman and Scott (1965) is an attractive alternative.

Consider the situation where an i.i.d. sample from a one-parameter distribution family $\{f(x;\theta) : \theta \in \mathcal{R}\}$ is available for testing a simple null hypothesis $\theta = \theta_0$. When the distribution family is regular, the score function of θ is given by

$$S_n(\theta) = \sum_{i=1}^{n} \frac{f'(X_i;\theta)}{f(X_i;\theta)},$$

where $f'(x;\theta)$ is the derivative of f with respect to θ. It is well known that

$$\mathrm{E}_\theta\{S_n(\theta)\} = 0$$

for any θ. Hence, substantial deviation of $S_n(\theta_0)$ from value 0 is an evidence against H_0: $\theta = \theta_0$. The degree of deviation is decided by comparing $T_n = (n\mathbb{I})^{-1/2}S_n(\theta_0)$ to the reference standard normal distribution where $\mathbb{I}$ is the Fisher information. Score test as such is known to be locally optimal against one-sided alternative.

Consider now a null hypothesis H_0: $\theta = \theta_0$ under a model with multi-parameter θ and ξ. In the presence of nuisance parameter ξ, H_0 is *composite* such that it contains a set of distributions. We have available two zero-expectation functions:

$$S_{n1}(\theta,\xi) = \sum_{i=1}^{n} \frac{f_1'(X_i;\theta,\xi)}{f(X_i;\theta,\xi)},$$

$$S_{n2}(\theta,\xi) = \sum_{i=1}^{n} \frac{f_2'(X_i;\theta,\xi)}{f(X_i;\theta,\xi)},$$

where f_1' and f_2' are derivatives with respect to θ and ξ. Since ξ value is not specified under H_0, a root-n consistent estimator $\hat{\xi}$ will be utilised. The $C(\alpha)$ test seeks a statistic from a linear combination:

$$T_n = aS_{n1}(\theta_0, \hat{\xi}) + bS_{n2}(\theta_0, \hat{\xi}), \tag{9}$$

with a and b chosen by some optimality consideration.

We now discuss how a specific $C(\alpha)$ test is derived for homogeneity in the context of mixture model. Consider the mixture model with density function given by

$$f(x; G) = \int_\Theta f(x; \theta) dG(\theta)$$

with its parameter space $\mathbb{G}$ containing all distributions on Θ. We narrow the space $\mathbb{G}$ down slightly so that all its members have finite second moment. When $\Theta = \mathcal{R}$, we may rewrite the mixture density function as

$$\varphi(x; \theta, \sigma, G) = \int_\Theta f(x; \theta + \sqrt{\sigma}\xi) dG(\xi) \tag{10}$$

such that the standardised mixing distribution $G(\cdot)$ has mean 0 and variance 1. Under new parameterisation, the null hypothesis becomes H_0: $\sigma = 0$. Both θ and the mixing distribution G are nuisance parameters.

The partial derivative of $\log \varphi(x; \theta, \sigma, G)$ with respect to σ is given by

$$\frac{\partial \varphi(x; \theta, \sigma, G)}{\partial \sigma} = \frac{\int_\Theta \xi f'(x; \theta + \sqrt{\sigma}\xi) dG(\xi)}{2\sqrt{\sigma} \int_\Theta f(x; \theta + \sqrt{\sigma}\xi) dG(\xi)}.$$

At $\sigma = 0$ or let $\sigma \downarrow 0$, we find (and define it to be)

$$\left.\frac{\partial \varphi(x; \theta, \sigma, G)}{\partial \sigma}\right|_{\sigma \downarrow 0} = \frac{f''(x; \theta)}{2f(x; \theta)}.$$

This is the score function for σ based on a single observation. The choice of $\sqrt{\sigma}$ is to get a non-degenerate score function.

The partial derivative of $\log \varphi(x; \theta, \sigma, G)$ with respect to θ is given by

$$\frac{\partial \varphi(x; \theta, \sigma, G)}{\partial \theta} = \frac{\int_\Theta f'(x; \theta + \sqrt{\sigma}\xi) dG(\xi)}{\int_\Theta f(x; \theta + \sqrt{\sigma}\xi) dG(\xi)}$$

which leads to score function for θ based on a single observation as

$$\frac{\partial \varphi(x; \theta, 0, G)}{\partial \theta} = \frac{f'(x; \theta)}{f(x; \theta)}.$$

We have now identified two zero-expectation functions (under H_0):

$$y_i(\theta) = \frac{f'(x_i; \theta)}{f(x_i; \theta)}, \quad z_i(\theta) = \frac{f''(x_i; \theta)}{2f(x_i; \theta)}. \tag{11}$$

where x_i's are i.i.d. observations from the mixture model. The score functions based on the entire sample are $\sum_{i=1}^n Z_i(\theta)$ and $\sum_{i=1}^n Y_i(\theta)$ for the mean and variance of G. The optimal combination is given by $w_i(\theta) = z_i(\theta) - \beta(\theta) y_i(\theta)$, with β being the regression coefficient

$$\beta(\theta) = \frac{\mathrm{E}\{Y_1(\theta) Z_1(\theta)\}}{\mathrm{E}\{Y_1^2(\theta)\}}.$$

We have capitalised Y and Z to indicate their status as random variables in the above operation. The expectation is with respect to a homogeneous null distribution $f(x; \theta)$.

Replacing the unspecified θ by its MLE under homogeneous model $f(x, \theta)$, the C(α) statistic arrives at a simple form:

$$W_n = \frac{\sum_{i=1}^n W_i(\hat{\theta})}{\sqrt{n\nu(\hat{\theta})}} = \frac{\sum_{i=1}^n Z_i(\hat{\theta})}{\sqrt{n\nu(\hat{\theta})}} \tag{12}$$

with $\nu(\theta) = \mathrm{E}\{W_1^2(\theta)\}$. Clearly, W_n has standard normal limiting distribution under some general moment conditions on Y and Z. At a given significance level α, we reject the homogeneity hypothesis H_0 when $W_n > z_\alpha$. This is the C(α) test for homogeneity.

In deriving the C(α) statistic, we assumed that the parameter space $\Theta = \mathcal{R}$. Note that when $G(\theta)$ is a mixing distribution on Θ, so is $G((\theta - \theta^*)/\sigma^*)$ for any θ^* and $\sigma^* \geq 0$. If instead $\Theta = \mathcal{R}^+$ as in the Poisson mixture model in which component distribution has mean $\theta \geq 0$, $G((\theta - \theta^*)/\sigma^*)$ is not a legitimate mixing distribution for some θ^* and σ^*. One may re-parameterise the model through $\xi = \log \theta$ in this case. However, there is no unified approach in general.

Regardless of the hidden validity issue of the mathematical derivation, W_n remains a useful metric on homogeneity hypothesis. Hence, it remains an effective test statistic.

Many commonly used distributions in statistics belong to a group of natural exponential families with quadratic variance function (NEF-QVF; Morris (1982)). The examples include normal, Poisson, binomial, and exponential. The density function in one-parameter natural exponential family has a unified analytical form

$$f(x; \theta) = h(x) \exp\{x\phi - A(\phi)\},$$

with respect to some σ-finite measure, where $\theta = A'(\phi)$ is the mean parameter. Let $\sigma^2 = A''(\phi)$ be the variance under $f(x; \theta)$. To be a member of NEF-QVF, the variance must be a quadratic function of the mean:

$$\begin{aligned} \sigma^2 = A''(\phi) &= a\{A'(\phi)\}^2 + bA'(\phi) + c \\ &= a\theta^2 + b\theta + c. \end{aligned} \tag{13}$$

Take Poisson distribution as an example; it satisfies (13) with $a = 0$, $b = 1$ and $c = 0$.

C(α) statistic has a particularly simple analytical form under NEF-QVF,

$$W_n = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 - n\hat{\sigma}^2}{\sqrt{2n(a+1)}\hat{\sigma}^2},$$

where $\bar{x} = n^{-1} \sum_{i=1}^n x_i$ and $\hat{\sigma}^2 = a\bar{x}^2 + b\bar{x} + c$.

Let X be a random variable with a NEF-QVF mixture distribution and θ be a random variable with distribution G. Denote $\mu = \mathrm{E}\{\mathrm{E}(X|\theta)\} = \mathrm{E}(\theta)$. When G degenerates, the variance of X is given by $\sigma^2(\mu) = a\mu^2 + b\mu + c$. When G does not degenerate,

$$\begin{aligned}\mathrm{VAR}(X) &= \mathrm{E}\{\mathrm{VAR}(X|\theta)\} + \mathrm{VAR}\{\mathrm{E}(X|\theta)\} \\ &= \mathrm{E}(a\theta^2 + b\theta + c) + \mathrm{VAR}(\theta).\end{aligned}$$

Combining these calculations, we find

$$\mathrm{VAR}(X) - \sigma^2(\mu) = (a+1)\mathrm{VAR}(\theta) > 0.$$

This inequality shows that the presence of mixture inflates the variance of X by a quantity of size $(a+1)\mathrm{VAR}(\theta)$. The test statistic W_n is an over-dispersion measure. Because of this, C(α) test coincides with the detection of over-dispersion in the case of exponential family mixtures.

4. Restricted likelihood ratio test

$C(\alpha)$ test is simple and easy to implement. Yet statisticians are not ready to completely give up the likelihood approach. Many slightly altered likelihood approaches are developed and are effective to some degree. The likelihood method also has potential to test for general order of the mixture model.

As mentioned earlier, there are two types of partial loss of identifiability in the context of homogeneity test. One is when a subpopulation has a very small mixing proportion so that it is practically absent from any sample, the other is when two subpopulations are nearly identical. They are the culprits of the non-standard large sample properties of the LRT. The artificial separation condition is to partly restore regularity to attain some useful results.

Placing a similar condition on mixing proportion is also possible to partly restore the regularity. Chen and Cheng (2000) and Lemdani and Pons (1999) investigated the asymptotic distribution of the LRT under the restriction of $\min\{\pi_1, \pi_2\} \geq \epsilon$ for some $\epsilon > 0$. Under this restriction, the model defined by (3) becomes "regular" in some sense.

Assume that we have an i.i.d. sample from (3). Under mild regularity conditions on component distribution $f(x;\theta)$, the MLE $\hat{G}_n$ of G is consistent (Chen, 2016). The consistency in the functional space $\mathbb{G}$ is best interpreted as

$$\int |\hat{G}_n(\theta) - G(\theta)| \exp(-|\theta|) d\theta \to 0$$

almost surely as the sample size n goes to infinite. This conclusion is not affected when $\min\{\pi_1, \pi_2\} \geq \epsilon$ is applied as long as the true mixing distribution G^* satisfies this restriction. When $G^*(\theta) = \mathbb{1}(\theta^* \leq \theta)$, the consistency leads to

$$\hat{\theta}_1 - \theta^* = o(1); \quad \text{and} \quad \hat{\theta}_2 - \theta^* = o(1). \tag{14}$$

This leads to a useful expansion of the log-likelihood function. Let m_1 and m_2 be moments of $\hat{G}_n$ centred at θ^*:

$$\begin{aligned} m_1 &= \hat{\pi}_1(\hat{\theta}_1 - \theta^*) + \hat{\pi}_2(\hat{\theta}_2 - \theta^*); \\ m_2 &= \hat{\pi}_1(\hat{\theta}_1 - \theta^*)^2 + \hat{\pi}_2(\hat{\theta}_2 - \theta^*)^2. \end{aligned}$$

By Taylor's expansion, one gets

$$\begin{aligned} &\log f(x_i; \hat{G}_n) - \log f(x_i; G^*) \\ &\approx m_1 y_i(\theta^*) + m_2 z_i(\theta^*) \\ &\quad -(1/2)\{m_1 y_i(\theta^*) + m_2 z_i(\theta^*)\}^2 \end{aligned}$$

after the high-order terms are ignored.

The further development is algebraically simplest when the random versions of y_i, z_i are uncorrelated. When this is the case, the above expansion leads to

$$\begin{aligned} \ell_n(\hat{G}_n) - \ell_n(G^*) \approx\ & m_1 \sum_{i=1}^n y_i(\theta^*) + m_2 \sum_{i=1}^n z_i(\theta^*) \\ & - (1/2) m_1^2 \sum y_i^2(\theta^*) \\ & - (1/2) m_2 \sum_{i=1}^n z_i^2(\theta^*). \end{aligned}$$

Because $\hat{G}_n$ by definition maximises $\ell_n(G)$ among all two-point mixing distributions satisfying $\min\{\pi_1, \pi_2\} \geq \epsilon$, and because of the above expansion, we infer that the moments of $\hat{G}_n$ must satisfy

$$m_1 \approx \frac{\sum_{i=1}^n y_i(\theta^*)}{\sum_{i=1}^n y_i^2(\theta^*)}; \quad m_2 \approx \frac{[\sum_{i=1}^n z_i(\theta^*)]^+}{\sum_{i=1}^n z_i^2(\theta^*)}$$

taking note that m_2 is nonnegative.

Under the null hypothesis that the true mixing distribution G^* degenerates, the MLE is searched after under a regular model $f(x;\theta)$. We also have

$$\begin{aligned} &\ell_n(\hat{\theta}) - \ell(\theta^*) \\ &\approx (\hat{\theta} - \theta^*) \sum_{i=1}^n y_i(\theta^*) - (1/2)(\hat{\theta} - \theta^*)^2 \sum y_i^2(\theta^*) \end{aligned}$$

with the MLE $\hat{\theta}$ under the null hypothesis satisfying

$$\hat{\theta} - \theta^* \approx \frac{\sum_{i=1}^n y_i(\theta^*)}{\sum_{i=1}^n y_i^2(\theta^*)}.$$

These informal derivations point to an approximation to the likelihood ratio statistics (under the restriction $\min\{\pi, 1-\pi\} > \epsilon > 0$)

$$\begin{aligned} R_n = 2\{\ell_n(\hat{G}_n) - \ell_n(\hat{\theta})\} &\approx \frac{\{[\sum_{i=1}^n z_i(\theta^*)]^+\}^2}{\sum_{i=1}^n z_i^2(\theta^*)} \\ &\approx [Z^+(\theta^*)]^2 \end{aligned}$$

which clearly has a limiting distribution $0.5\chi_0^2 + 0.5\chi_1^2$, the same as in (7).

This result is mathematically neat because the limiting distribution does not depend on the true distribution θ^* nor the choice of ϵ. It is also easy for numerical computation of the p-value in applications. Nevertheless, the choice of ϵ is problematic in applications. Its choice affects how well the limiting distribution approximates the finite sample distribution. Our experience indicates that the approximation has poor precision unless ϵ is somewhat large. Hence, this result is more for insight than of practical value.

5. Modified likelihood ratio test and EM-test for homogeneity

Placing a hard restriction $\min\{\pi_1, \pi_2\} \geq \epsilon$ leads to difficulties at specifying ϵ in applications. An attractive alternative is developed by placing a soft restriction of similar nature in Chen (1998) and further in Chen et al. (2001), Chen et al. (2004) and others. For test of homogeneity, a modified likelihood as follows is introduced

$$\tilde{\ell}_n(G) = \ell_n(G) + C\log\{4\pi_1\pi_2\}$$

for some positive constant C when G is a mixing distribution with mixing proportions π_1 and π_2. When G degenerates, we use $\pi_1 = \pi_2 = 0.5$ in this definition. When Θ is compact, $\ell_n(G) - \ell_n(G^*) = O_p(1)$ under some model assumptions. Note that $C\log\{4\pi_1\pi_2\} \to -\infty$ when $\min\{\pi_1, \pi_2\} \to 0$. Hence, the maximum of $\tilde{\ell}_n(G)$ is attained at some G with $\min\{\pi_1, \pi_2\} \geq c_n > 0$. That is, the modified likelihood is an implicitly restricted likelihood. Does it work like a restricted likelihood?

Let $\tilde{G}$ be the two support point mixing distribution that maximises the modified likelihood function $\tilde{\ell}_n(G)$. Let us define the mLRT statistic to be

$$\tilde{R}_n = 2\{\ell_n(\tilde{G}) - \ell_n(\hat{\theta})\}.$$

As shown in Chen et al. (2001), under mild conditions on $f(x;\theta)$ and with a compact assumption on Θ, the modified MLE $\tilde{G}$ satisfies

$$\tilde{\theta}_1 - \theta^* = o_p(1); \quad \text{and} \quad \tilde{\theta}_2 - \theta^* = o_p(1). \tag{15}$$

This mimics (14) and validates informal expansions that followed. Not surprisingly,

$$\tilde{R}_n \to 0.5\chi_0^2 + 0.5\chi_1^2$$

in distribution under the null hypothesis. Unlike the restricted LRT, the fitted values of π_1, π_2 in G under modified likelihood are allowed to be arbitrarily close to zero. Hence, the mLRT largely avoids the difficulty to choose an ϵ.

We seem to merely replace one challenge with another: choosing a properly sized C instead of a properly sized ϵ. The choice of C is a less touchy issue because it does not lead to direct restrictions on mixing distribution G. The influence of C is mild and smooth on mixing proportions. Chen et al. (2001) found that in most cases, putting $C = 1$ leads to acceptable precision of the approximation based on limiting distribution $0.5\chi_0^2 + 0.5\chi_1^2$.

In spite of its usefulness as illustrated above, the mLRT exhibits two soft spots. One is that the component parameter space Θ is compact and the other is that Fisher information with respect to mixing proportion for any $\theta \in \Theta$ must be finite. The first requirement is largely ignored in applications and simulation studies such as in Chen et al. (2001). In other words, even though the asymptotic conclusion is possible only if Θ is compact, it matters little in applications.

The second requirement translates into the condition that for any $\theta \in \Theta$,

$$\mathrm{E}\left\{\frac{f(X;\theta)}{f(X;G^*)}\right\}^2 < \infty. \tag{16}$$

Consider the finite mixture of exponential distribution where the component density function

$$f(x;\theta) = \frac{1}{\theta}\exp\{-\frac{x}{\theta}\}.$$

At homogeneous distribution when $\theta^* = 1$,

$$\mathrm{E}\left\{\frac{f(X;\theta)}{f(X;G^*)}\right\}^2 = \begin{cases} \frac{(1-\theta)^2}{\theta(2-\theta)} & \theta < 2; \\ \infty & \theta \geq 2. \end{cases}$$

That is, the finite Fisher information condition is violated at $\theta = 2$. Hence, the asymptotic conclusion of the mLRT is not applicable to finite exponential mixture in general. The finite Fisher information condition is also violated for finite mixture of Gamma distributions, of distribution in a scale family, and many more. Unlike the compact condition on Θ, the violation of finite Fisher information condition noticeably changes the asymptotic properties of the mLRT. See Chen and Li (2011) for evidences.

In summary, although the mLRT advances markedly from the original LRT, further development is needed which leads to EM-test. Even though the EM-test and mLRT differ a lot on surface, they are closely connected.

A common thread of the restricted LRT and the mLRT is to apply some constraints to the mixing proportion. Taking the restriction to extreme, let us allow only a few pre-chosen mixing proportions in the space of alternative distributions. This sounds unreasonable, but it is how the EM-test is obtained.

Let $X_1, \ldots, X_n$ be an i.i.d. sample from a two-component finite mixture model

$$f(x;G) = f(x;\pi,\theta_1,\theta_2) = \pi f(x,\theta_1) + (1-\pi)f(x,\theta_2), \tag{17}$$

the same as before. The log likelihood function under this model assumption is given by

$$\ell_n(\pi;\theta_1,\theta_2)=\sum_{i=1}^{n}\log\{\pi f(x_i,\theta_1)+(1-\pi)f(x_i,\theta_2)\}. \quad (18)$$

Consider the following pair of opposing hypotheses:

$$H_0: f(x;\theta) \text{ against}$$
$$H_1: (0.3)f(x,\theta_1)+(0.7)f(x,\theta_2). \quad (19)$$

When H_0 is true and the model $f(x;\theta)$ is regular, the statistic

$$R_n(0.3)=2\Big\{\sup_{\theta_1,\theta_2}\ell_n(\pi=0.3;\theta_1,\theta_2)-\sup_{\theta}\ell_n(\pi;\theta,\theta)\Big\}$$

has now the typical null limiting distribution $(0.5)\chi_0^2+(0.5)\chi_1^2$. This result does not depend on the specific choice of $\pi=0.3$. We have therefore found a shortcut for homogeneity test: rejecting H_0 when $R_n(0.3)$ is large according to this reference distribution.

Yet, $R_n(0.3)$ has an obvious shortcoming. Suppose the true distribution has two subpopulations with mixing proportions $\pi=0.16$ and $1-\pi=0.84$. The most effective test should use $R_n(0.16)$ instead of $R_n(0.3)$. In applications, we do not know the true mixing proportion. To guard against all possibilities, we end up using $\sup_\pi R_n(\pi)$ as the test statistics. However, this statistic is now the LRT statistic, whose distribution was found too difficult to handle.

EM-test breaks this cycle (Li, Chen, & Marriott, 2009) but the strategy and the motivation are somewhat involved. A simplistic and not rigorous version goes as follows. We first specify a mixing proportion for the potential alternative mixture, say $\pi=0.3$,after which, we numerically locate the maximum point of $\ell_n(\pi;\theta_1,\theta_2)$ and denote them as $\theta_1^{(0)},\theta_2^{(0)}$. Given $\theta_1^{(0)},\theta_2^{(0)}$, we locate $\pi^{(0)}$ that maximises

$$\tilde{\ell}_n(G)=\ell_n(\pi;\theta_1^{(0)},\theta_2^{(0)})+C\log(1-|1-2\pi|)$$

with some $C>0$. Note a regularisation term $C\log(1-|1-2\pi|)$ slightly different from the one for mLRT is added with a similar purpose. It can be ignored without harming the main line of thinking.

Let $G_{0.3}^{(0)}$ be the mixing distribution formed by $\pi^{(0)},\theta_1^{(0)},\theta_2^{(0)}$. Applying EM-iteration (Dempster, Laird, & Rubin, 1977) with $G^{(0)}$ as the initial to obtain $G_{0.3}^{(K)}$ after K iterations. If $K\to\infty$, $G_{0.3}^{(K)}$ converges to at least a local maximum of the log-likelihood function $\tilde{\ell}_n(G)$ under model (17). See Wu (1983) for the convergence issue related to EM-algorithm.

To avoid falling back to the LRT statistics, the EM-test constructs a statistic before the iteration converges. For a finite K and given π, define

$$M_n^{(K)}(\pi)=2\{\ell_n(G_\pi^{(K)})-\ell_n(\pi;\hat{\theta};\hat{\theta})\}$$

with $\hat{\theta}$ being the MLE of θ under the null model. Li et al. (2009) found that for each fixed K and J choices of π value, as $n\to\infty$,

$$\text{EM}_n=\max\{M_n^{(K)}(\pi_1),M_n^{(K)}(\pi_2),\ldots,M_n^{(K)}(\pi_J)\}\to(0.5)\chi_0^2+(0.5)\chi_1^2 \quad (20)$$

in distribution. Hence, EM_n is a suitable test statistic for homogeneity.

Li et al. (2009) generally recommended to have $K=3$ and $J=3$ with π taken from $\{0.1, 0.3, 0.5\}$ in applications. Although EM-iteration for the purpose of computing maximum likelihood estimate is slow, the increment of $\ell_n(G_\pi^{(k)})$ from $k=0, 1, 2, 3$ is mostly in the first two iterations. The change from $k=2$ to $k=3$ is usually small and the effect on p-value is beyond third decimal place. Further iteration makes little difference. In addition, by iterating from three initial π in $\{0.1, 0.3, 0.5\}$, the value of the mixing proportion in $G_\pi^{(k)}$ effectively covers the range $[0, 0.5]$. By symmetry, the range extends to $[0, 1]$. Hence, increasing the number of initial mixing proportions is not needed.

One may quickly realise that the value of the EM-test statistic is close to a usual LRT statistic, given the same data set. However, EM-test statistic gauges how fast the likelihood increases initially, and the LRT statistic measures how much the likelihood ultimately increases. The advantages of the EM-test include: (a) Θ need not be compact, (b) the second moment as in (16) need not be finite, (c) much simpler asymptotics, (d) elegant extension to finite normal mixture models (Chen & Li, 2016; Chen, Li, & Fu, 2012).

6. Test for $m=m_0\geq 2$

So far, the discussions are limited to homogeneity test ($m=1$) with one-dimensional θ. We now move to H_0: $m=m_0\geq 2$ for one-dimensional θ and for finite normal mixture in both mean and variance.

Consider the one-dimensional θ but general m_0 first. The generic modified likelihood is defined as

$$\tilde{\ell}_n(G)=\ell_n(G)+C\sum_{j=1}^{m}\log\pi_j$$

for G with m support points and mixing proportions π_j. That is, when a model of order $m>2$ is fitted to an i.i.d. sample, the likelihood is penalised by $C\log\pi_j$ for a subpopulation with proportion π_j. Under mild conditions on $f(x;\theta)$ and a compact condition on Θ, $\ell_n(G)-\ell_n(G^*)$ is stochastically bounded as $n\to\infty$, where G^* stands for the true distribution. This shows that if $\tilde{G}_n$ maximises $\tilde{\ell}_n(G)$ among G with m support points, we must have

$$\sum_{j=1}^{m}\log\tilde{\pi}_j=O_p(1).$$

At the same time, the modified MLE $\tilde{G}_n$ is known to be consistent for G^* under some conditions, i.e., $\|\tilde{G}_n-G^*\|\to 0$.

In the context of homogeneity test, the null $G^* = \{\theta^*\}$ is fitted with a mixing distribution $\tilde{G}_n$ under H_1 and $\tilde{\theta}$ under H_0. Moving from $G = \{\tilde{\theta}\}$ to $G = \tilde{G}_n$, $2\ell_n(G)$ is increased by $\{Z^+(\theta^*)\}^2$ for some standard normal Z. When the null is H_0: $m = 2$ with

$$G^* = \pi_1^*\{\theta_1^*\} + \pi_2^*\{\theta_2^*\},$$

we compare the size of $\ell_n(G)$ maximised over

$$G = \pi_1\{\theta_1\} + \pi_2\{\theta_2\}$$

or over

$$G = \pi_1 G_1 + \pi_2 G_2$$

directly. From the experience of homogeneity test, we have reason to believe that $2\ell_n(G)$ will increase by

$$\{Z^+(\theta_1^*)\}^2 + \{Z^+(\theta_2^*)\}^2.$$

If $Z^+(\theta_1^*)$ and $Z^+(\theta_2^*)$ are independent, the limiting distribution would be

$$(0.25)\chi_0^2 + (0.5)\chi_1^2 + (0.25)\chi_2^2.$$

The truth is more complex but related.

Let $\mathbf{Z} = (Z_1, Z_2)^\tau$ and denote its covariance matrix $\mathbf{B}$. Chen et al. (2004) proved that the mLRT statistic for H_0: $m = 2$ against H_1: $m > 2$ after some manipulation asymptotically equals

$$\sup_{\mathbf{t}>0}\{2\mathbf{Z}^\tau\mathbf{t} - \mathbf{t}^\tau\mathbf{B}\mathbf{t}\},$$

where $\mathbf{t} = (t_1, t_2)^\tau$ and $\mathbf{t} > 0$ is interpreted componentwise. The distribution of this random variable is a mixture of chi-squares

$$\left(\frac{1}{2} - \frac{\arccos(\rho)}{2\pi}\right)\chi_0^2 + \frac{1}{2}\chi_1^2 + \frac{\arccos(\rho)}{2\pi}\chi_2^2,$$

where ρ is the correlation coefficient and here $\pi = 3.14159\cdots$.

EM-test for H_0: $m = 2$ against H_1: $m > 2$ has the same limiting distribution and this is not coincident. Unlikely mLRT, EM-test directly specifies a rigid form of mixing distributions permitted in H_1 when maximising $\tilde{\ell}_n(G)$ in the initial step. Suppose $\tilde{G}_0 = \tilde{\pi}_1\{\tilde{\theta}_1\} + \tilde{\pi}_2\{\tilde{\theta}_2\}$ maximises $\tilde{\ell}_n(G)$ under H_0. Li and Chen (2010) first defined a class of mixing distributions

$$G = \tilde{\pi}_1 G_1 + \tilde{\pi}_2 G_2 \tag{21}$$

such that for $k = 1, 2$, each G_k is a mixing distribution of exactly two support points in vicinity of $\tilde{\theta}_k$ with specific mixing proportions. Among mixing distributions of this form, obtain $G^{(0)}$ that maximises $\tilde{\ell}_n(G)$.

After $G^{(0)}$ is obtained, EM-iteration is applied K times to get $G^{(K)}$ as the usual EM-algorithm without restrictions. The meticulously specified $G^{(0)}$ makes the outcome of the EM-iteration $G^{(K)}$ retain the form of $\tilde{G}_n$ when H_0 is true. The resultant EM-test statistic EM_n therefore has the same limiting distribution as the mLRT.

The advantage is that the EM-test is multiple. It works well with generic m_0. To test H_0: $m = m_0$ against H_1: $m > m_0$, the form (21) is replaced by

$$G = \pi_1 G_1 + \pi_2 G_2 + \cdots + \pi_{m_0} G_{m_0}.$$

The EM-test statistic EM_n was found to have the same asymptotic expansion:

$$\sup_{\mathbf{t}>0}\{2\mathbf{Z}^\tau\mathbf{t} - \mathbf{t}^\tau\mathbf{B}\mathbf{t}\}.$$

The only difference is the dimensions of $\mathbf{t}$, $\mathbf{Z}$ and $\mathbf{B}$. The limiting distribution is a mixture of chi-squares with degrees of freedom ranging from 0 to m_0.

The expansion of the mLRT statistic is established after examining all potential maximisers of the modified likelihood. In comparison, that of the EM-test is established after examining a much reduced class. This key difference leads to simpler technical deliberation and broader applicability of the conclusion. The EM-test does not require compact Θ nor finite Fisher information. The research on the mLRT stopped at H_0: $m = 2$.

7. EM-test for finite normal mixture model

Two key innovations in developing the EM-test are (a) tactical selection of a special structured class of mixing distributions from H_1 based on the fitted mixing distribution under H_0; (b) use of EM-iteration to measure the improvement in likelihood from H_1 over H_0. These two strategies are broadly applicable. The outcomes are particularly interesting for finite normal mixture models. Let $\phi(x)$ be the density function of the standard normal. The density function of a finite normal mixture distribution is given by

$$f(x; G) = \frac{\pi_1}{\sigma_1}\phi\left(\frac{x-\mu_1}{\sigma_1}\right) + \frac{\pi_2}{\sigma_2}\phi\left(\frac{x-\mu_2}{\sigma_2}\right) + \cdots + \frac{\pi_m}{\sigma_m}\phi\left(\frac{x-\mu_m}{\sigma_m}\right). \tag{22}$$

Normal mixture distributions are the most important mixture models but their inference is also technically most challenging. The likelihood function of the normal mixture model is unbounded based on a set of random samples, unless an artificial bound is placed on its component variance parameter (Hathaway, 1985). Moreover, the model is not strongly identifiable (Chen, 1995) so it is hard to differentiate between overdispersion caused by mixture or by a large variance. It has infinite Fisher information with respect to mixing proportions. One must regularise the likelihood with some well-designed penalty function to achieve consistent point estimation (Chen, 2016), or settle for restricted MLE (Hathaway, 1985).

Let $\ell_n(G) = \Sigma \log f(x_i; G)$ still be the log likelihood function given a set of i.i.d. observations from a

finite normal mixture distribution and s_n^2 be the sample variance of x_i, and σ_j^2 be component variances. One approach to achieve consistent estimation of G is through penalised likelihood defined as

$$p\ell_n(G) = \ell_n(G) + \sum_{j=1}^{m} p_n(\sigma_h^2; s_n^2)$$

for some $\hat{\sigma}^2$-dependent smooth penalty function $p_n(\sigma^2; \hat{\sigma}^2)$ (see Chen, Tan, and Zhang (2008) or Chen (2016)). The key requirement on $p_n(\sigma^2; \hat{\sigma}^2)$ is that its value goes to negative infinity when σ goes to 0 or infinity at an appropriate rate. One such a choice is

$$p_n(\sigma^2; \hat{\sigma}^2) = -a_n\left\{\left(\frac{\hat{\sigma}^2}{\sigma^2}\right) + \log\left(\frac{\hat{\sigma}^2}{\sigma^2}\right) - 1\right\}.$$

The factor a_n determines the level of penalty, and the other factor places a Gamma distribution prior on σ^{-2}. A recommended choice for a_n is $a_n = n^{-1/2}$. As long as $a_n > 0$, $p\ell_n(G)$ becomes a bounded function of G for each finite m.

The likelihood principle that underlies the hypothesis test is to look for $\hat{G}_0$ and $\hat{G}_1$ that maximise the likelihood over spaces of H_0: $m = m_0$ and H_1: $m > m_0$, respectively, to form a test statistic. The size $\ell_n(\hat{G}_1) - \ell_n(\hat{G}_0)$ reflects the improvement of the fit on H_1 over H_0, and its effectiveness is supported by the classical Neyman–Pearson lemma. At the same time, the test is useful only if asymptotic property of $2\{\ell_n(\hat{G}_1) - \ell_n(\hat{G}_0)\}$ is manageable. In Section 6, we have illustrated that by strategic choice of a subspace of H_1, an EM-test is effective for order of the finite mixture of one-parameter component distribution.

With normal component distribution, Chen et al. (2012) identified a subspace of H_1, made of mixing distributions of order $2m_0$ for each carefully selected vector β as follows:

$$\Omega_{2m_0}(\beta) = \left\{ \sum_{j=1}^{m_0} \pi_j[\beta_j\{\theta_{1j}, \sigma_{1j}\} + (1-\beta_j)\{\theta_{2j}, \sigma_{2j}\}] : \theta_{1j}, \theta_{2j} \in I_j, \right\} \quad (23)$$

where I_j's are non-overlapping intervals containing $\hat{\theta}_{0j}$ induced by $\hat{G}_0$. For each β, one computes

$$G_\beta^{(0)} = \arg\max\{p\ell_n(G) : G \in \Omega_{2m_0}(\beta)\}.$$

EM-iteration is then applied to $p\ell_n(G)$ with $G_\beta^{(0)}$ being the initial value to get $G_\beta^{(K)}$, for a pre-specified K. Define

$$M_n^{(K)}(\beta) = 2\{p\ell_n(G_\beta^{(K)}) - p\ell_n(\hat{G}_0)\}.$$

The EM-test statistic is defined as the maximum $M_n^{(K)}(\beta)$ over a number of carefully selected values of β.

For the ease of presentation, I purposely erred at some details with the general idea preserved. Rigorous readers should consult Chen et al. (2012) for more accurate accounts. Almost like a gift, for testing H_0: $m = m_0$,

$$\mathrm{EM}_n^{(K)} \to \chi^2_{2m_0}$$

in distribution as $n \to \infty$. When the EM-test is applied to Pearson's crab data discussed in the beginning for H_0 of $m = 1, 2$, the p-values are found to be 8×10^{-11} and 0.53, respectively. Hence, an order-2 finite normal mixture model is well supported by the data.

Chen and Li (2016) further studied the order test problem for finite normal mixture model with common component variance:

$$f(x; G) = \frac{\pi_1}{\sigma}\phi\left(\frac{x-\mu_1}{\sigma}\right) + \frac{\pi_2}{\sigma}\phi\left(\frac{x-\mu_2}{\sigma}\right) + \cdots + \frac{\pi_m}{\sigma}\phi\left(\frac{x-\mu_m}{\sigma}\right).$$

Using the same strategy, an EM-test was constructed and found to have limiting distribution $\chi^2_{m_0-1}$ for H_0: $m = m_0$ when $m_0 \geq 2$. Their simulation, however, showed the limiting distribution is not sufficiently accurate for the finite sample distribution. Some additional research is needed.

8. EM-test for hidden Markov model

Under finite mixture models, data $X_1, X_2, \ldots, X_n$ may be regarded as generated in two steps. A hidden variable S_i selects one of m subpopulations with probability π_i in the first step. Given S_i, X_i is a sample from this subpopulation in the second step. The hidden states $S_1, \ldots, S_n$ are themselves i.i.d.

When $S_{1:T} = \{S_1, S_2, \ldots, S_T\}$ form a Markov chain instead, the time series $X_{1:T}$ are no longer i.i.d. However, at equilibrium, the marginal distributions of X_t are identical and have a finite mixture distribution. When $S_{1:T}$ are not observed, the model for $X_{1:T}$ under this formulation is called hidden Markov model (HMM).

In finance applications, market indexes often exhibit distinct stochastic properties over different periods. Such behaviour prompts the suggestion that the stochastic property of related process X_t is determined by some hidden state S_t. A two-state Markov chain for $S_{1:T}$ works well in many specific instances. For example, two states may represent periods of expanding and shrinking economy. Such models are investigated in statistical finance such as Hamilton (2010), Engel (1994) and Chen, Huang, and Wang (2016). At the same time, statistical evidence for the use of multi-state HMM is indispensable.

Consider a simple two-state HMM with state space $\mathcal{S} = \{1, 2\}$. At equilibrium, the marginal distribution of X_t has density function

$$f_h(x; G) = \pi_1 f(x; \theta_1) + \pi_2 f(x; \theta_2), \quad (24)$$

where $\pi_1 = P(S_t = 1)$ and $\pi_2 = P(S_t = 2)$. The stochastic property of the hidden states $S_{1:T}$ is determined by

transition probabilities

$$p_{ij} = P(S_t = j | S_{t-1} = i).$$

Because $S_{1:T}$ are generally dependent, so are $X_{1:T}$. To permit proper equilibrium distribution, the transition probabilities must satisfy certain conditions. When $\mathcal{S} = \{1, 2\}$, the HMM has four free parameters: $\{\theta_1, \theta_2, p_{12}, p_{21}\}$. The HMM reduces to a homogenous model when $\theta_1 = \theta_2$, or when $p_{12} = 0$ or $p_{21} = 0$.

The likelihood function of $\{\theta_1, \theta_2, p_{12}, p_{21}\}$ may be written as

$$L_n(\theta_1, \theta_2, p_{12}, p_{21}, p_1, p_2) = \sum_{s_{1:T}} \left\{ \prod_{t=1}^{T} f(x; \theta_{s_t}) \right\} P(S_{1:t} = s_{1:T}),$$

where the summation is over all possible state sequence $s_{1:T}$,

$$P(S_{1:t} = s_{1:T}) = p_{s_1} \prod_{t=2}^{T} p_{s_{t-1}, s_t}$$

and two extra parameters for the distribution of the initial state S_1: $p_1 = 1 - p_2 = P(S_1 = 1)$.

Interestingly, an easy-to-use forward–backward algorithm is available to compute the MLE of the parameters (Baum, Petrie, Soules, & Weiss, 1970) but evaluating L_n or $\log L_n$ is not as straightforward. Full likelihood-based data analysis under HMM is hence challenging in general. One strategy is to replace the likelihood function with a function with similar properties but easier to handle. One such candidate is composite likelihood (CL).

A simplistic introduction of CL is as follows. Additional references can be found in Lindsay (1988), Varin (2008) and Varin, Reid, and Firth (2011). A user may first identify a class of subsets of the observations. A likelihood function can be formed based on each subset of the data under the full model assumption, after which, a log CL is formed as a weighted sum of these log likelihood functions. By a selective choice of these subsets, the CL can be made robust against some degree of model mis-specification, simpler for asymptotic analysis, and easier for numerical computation.

A simple composite log likelihood as the sum of log likelihood based on single x_t is then

$$\ell_c(G) = \sum_{t=1}^{T} \log f_h(x_t; G). \tag{25}$$

Note this likelihood has identical algebraic expression when $x_{1:T}$ are T i.i.d. observations. Hence, all tests discussed under i.i.d. assumptions can be mathematically carried out. Interestingly, the large sample properties of the mLRT and EM-test given in previous sections remain valid. See Dannemann and Holzmann (2008a, 2008b), , Holzmann and Schwaiger (2016) for details. We are saved from adding more details.

9. Future directions

EM-test is currently the most successful approach for testing the order of the finite mixture models. EM-test is a class of tests created based on a conceptually simple principle. They have pushed the boundary of the order test. At the same time, many problems remain unsolved.

There is still lack of effective tests for the order of the finite mixture models with multi-parametric $f(x; \theta)$. Within this category, finite mixture of one-dimensional normal distribution is special. Two parameters here specify two distinct aspects of the distribution. EM-test has properly addressed the order test problem in its plain form. At the same time, the EM-test-like solutions often depend on individual features as in Li et al. (2015) and Shen and He (2015). The former deals with the feature where multiple samples are available and the latter works on feature where the mixing proportions are structured.

Finite normal mixture is a special mixture of distributions in a location-scale family. These models are obtained when $\phi(\cdot)$ in (22) is logistic, Cauchy, Student and so on. Unfortunately, the neat conclusions on the EM-test for normal mixtures are not true. One must start all over again to determine the asymptotic properties of similarly formulated EM-tests. Both interestingly and unfortunately, the large sample properties depend on the specific location-scale family. Some preliminary results obtained by my collaborators indicate that one version of the EM-test for homogeneity of location-scale mixture has limiting distribution assembling

$$\sup_{\mathbf{t}} \{2\mathbf{Z}^\tau \mathbf{t} - \mathbf{t}^\tau \mathbf{B} \mathbf{t}\} \tag{26}$$

with the range on the surface of some three-dimensional cone.

The idea of the EM-test is equally applicable to finite mixture of other generic multi-parameter component distribution (Niu, Li, & Zhang, 2011). The challenge is that the range in the generic (26) is dependent on the specific model as well as the specific choice of the subspace of H_1. Searching for specific choice that leads to simple limiting distribution is the major task for the future research.

Finally, I wish to turn the attention to non-i.i.d. data such as data from HMM. The EM-test has been found to work well based on CL constructed from marginal distributions as demonstrated by references given earlier. However, this line of approach ignores the time series nature of the HMM. There should be a way to have the transition information accommodated in the EM-test. Finite mixture of regressions forms another rich source of non-i.i.d. data. At this moment, there have been little discussion on tests for the order of regression mixtures.

Acknowledgments

The author likes to thank research fundings from the National Natural Science Foundation of China (Grant No 11690011) and the Natural Science and Engineering Research Council (RGPIN-2014- 03743).

Disclosure statement

No potential conflict of interest was reported by the author.

References

Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics,* 41(1), 164–171.

Böhning, D. (2000). *Computer-assisted analysis of mixtures and applications: Meta-analysis, disease mapping and others*. London, UK: Chapman and Hall.

Chen, H., & Chen, J. (2001). The likelihood ratio test for homogeneity in finite mixture models. *Canadian Journal of Statistics,* 29(2), 201–215.

Chen, H., & Chen, J. (2003). Tests for homogeneity in normal mixtures in the presence of a structural parameter. *Statistica Sinica,* 13, 351–365.

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(1), 19–29.

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,* 66, 95–115.

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

Chen, J. (1998). Penalized likelihood-ratio test for finite mixture models with multinomial observations. *Canadian Journal of Statistics,* 26(4), 583–599.

Chen, J. (2016). Consistency of the mle under mixture models. *Statistical Science*. arXiv: 1607.01251.

Chen, J., & Cheng, P. (2000). The limiting distribution of the restricted likelihood ratio statistic for finite mixture models. *Chinese Journal of Applied Probability and Statistics,* 2, 159–167.

Chen, J., Huang, Y., & Wang, P. (2016). Composite likelihood under hidden Markov model. *Statistica Sinica,* 26(4), 1569–1586.

Chen, J., & Li, P. (2011). Tuning the em-test for finite mixture models. *Canadian Journal of Statistics,* 39(3), 389–404.

Chen, J., & Li, P. (2016). Testing the order of a normal mixture in mean. *Communications in Mathematics and Statistics,* 4(1), 21–38.

Chen, J., Li, P., & Fu, Y. (2012). Inference on the order of a normal mixture. *Journal of the American Statistical Association,* 107(499), 1096–1105.

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

Chernoff, H., & Lander, E. (1995). Asymptotic distribution of the likelihood ratio test that a mixture of two binomials is a single binomial. *Journal of Statistical Planning and Inference,* 43(1), 19–40.

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(4), 1178–1209.

Dannemann, J., & Holzmann, H. (2008a). Likelihood ratio testing for hidden Markov models under non-standard conditions. *Scandinavian Journal of Statistics,* 35(2), 309–321.

Dannemann, J., & Holzmann, H. (2008b). Testing for two states in a hidden Markov model. *Canadian Journal of Statistics,* 36(4), 505–520.

Davies, R. B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika,* 64(1), 247–254.

Davies, R. B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika,* 74(1), 33–43.

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,* 39, 1–38.

Engel, C. (1994). Can the Markov switching model forecast exchange rates? *Journal of International Economics, 36*(1–2), 151–165.

Friedlander, Y., & Leitersdorf, E. (1995). Segregation analysis of plasma lipoprotein (a) levels in pedigrees with molecularly defined familial hypercholesterolemia. *Genetic Epidemiology,* 12(2), 129–143.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. New York, NY: Springer Science & Business Media.

Ghosh, J. K., & Sen, P. K. (1985). On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results. In L. LeCam & R. A. Olshen' (Eds.), *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer* (Vol. 2, pp. 789–806). Belmont, CA: Wadsworth.

Hamilton, J. D. (2010). Regime switching models. In *Macroeconometrics and time series analysis* (pp. 202–209). London, UK: Palgrave Macmillan.

Hartigan, J. A. (1985). A failure of likelihood asymptotics for normal mixtures. In L. LeCam & R. A. Olshen' (Eds.), *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer* (Vol. 2, pp. 807–810). Belmont, CA: Wadsworth.

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

Holzmann, H., & Schwaiger, F. (2016). Testing for the number of states in hidden Markov models. *Computational Statistics and Data Analysis,* 100, 318–330.

Kiefer, J., & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics,* 27, 887–906.

Lemdani, M., & Pons, O. (1999). Likelihood ratio tests in contamination models. *Bernoulli,* 5(4), 705–719.

Li, P., & Chen, J. (2010). Testing the order of a finite mixture. *Journal of the American Statistical Association,* 105(491), 1084–1092.

Li, P., Chen, J., & Marriott, P. (2009). Non-finite fisher information and homogeneity: An EM approach. *Biometrika,* 96, 411–426.

Li, S., Chen, J., Guo, J., Jing, B.-Y., Tsang, S.-Y., & Xue, H. (2015). Likelihood ratio test for multi-sample mixture model and its application to genetic imprinting. *Journal of the American Statistical Association,* 110(510), 867–877.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary Mathematics,* 80(1), 221–239.

Lindsay, B. G. (1995). Mixture models: Theory, geometry and applications. In *NSF-CBMS regional conference series in probability and statistics*. Hayward, CA: Institute of Mathematical Statistics.

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

McLachlan, G., & Peel, D. (2004). *Finite mixture models*. New York, NY: John Wiley & Sons.

Morris, C. N. (1982). Natural exponential families with quadratic variance functions. *The Annals of Statistics,* 10, 65–80.

Neyman, J., & Scott, E. (1965). On the use of c (alpha) optimal tests of composite hypotheses. *Bulletin of the International Statistical Institute,* 41(1), 477–497.

Niu, X., Li, P., & Zhang, P. (2011). Testing homogeneity in a multivariate mixture model. *Canadian Journal of Statistics,* 39(2), 218–238.

Ott, J. (1999). *Analysis of human genetic linkage*. Baltimore, MD: JHU Press.

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

Schork, N. J., Allison, D. B., & Thiel, B. (1996). Mixture distributions in human genetics research. *Statistical Methods in Medical Research,* 5(2), 155–178.

Shen, J., & He, X. (2015). Inference for subgroup analysis with a structured logistic-normal mixture model. *Journal of the American Statistical Association,* 110(509), 303–312.

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

Varin, C. (2008). On composite marginal likelihoods. *AStA – Advances in Statistical Analysis,* 92(1), 1–28.

Varin, C., Reid, N., & Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica,* 21, 5–42.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics,* 9(1), 60–62.

Wu, C. J. (1983). On the convergence properties of the em algorithm. *The Annals of statistics,* 11, 95–103.