Biometrika (2008), **??**, ?, *pp*. 1–16

Advance Access publication on ?? 2008
Printed in Great Britain

Non-finite Fisher information and homogeneity: an EM approach

BY P. LI
Department of Mathematical & Statistical Sciences, University of Alberta, Edmonton, T6G 2G1, Canada
pli@stat.ubc.ca

J. CHEN
Department of Statistics, University of British Columbia, Vancouver, V6T 1Z2, Canada
jhchen@stat.ubc.ca

AND P. MARRIOTT
Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, N2L 3G1, Canada
pmarriot@math.uwaterloo.ca

SUMMARY

Even simple examples of finite mixture models can fail to fulfil the regularity conditions that are routinely assumed in standard parametric inference problems. Many methods have been investigated for testing for homogeneity in finite mixture models, for example, but all rely on regularity conditions including the finiteness of the Fisher information and the space of the mixing parameter being a compact subset of some Euclidean space. Very simple examples where such assumptions fail include mixtures of two geometric distributions and two exponential distributions, and, more generally, mixture models in scale distribution families. To overcome these difficulties, we propose and study an EM-test statistic, which has a simple limiting distribution for examples in this paper. Simulations show that the EM-test has accurate type I errors and is more efficient than existing methods when they are applicable. A real example is also included.

Some key words: Chi-squared limiting distribution; Compactness; Exponential mixture; Finite mixture model; Homogeneity; Likelihood ratio test; Score test.

1. INTRODUCTION

1·1. *Examples*

Many first-order asymptotic results for standard parametric models are based on the fact that the asymptotic distribution of the score vector is very tractable. However, even for very simple mixture models the behaviour of the score and the shape of the loglikelihood function can be very different from that expected from standard first-order results. Consider, for example, a mixture distribution with density function given by

$$f(x;\Psi) = \int f(x;\theta)d\Psi(\theta) = (1-\alpha)f(x;\theta_1) + \alpha f(x;\theta_2), \qquad (1)$$

where $f(x;\theta)$ is a density function belonging to some parametric family of distributions and $\Psi = (1-\alpha)I(\theta_1 \leqslant \theta) + \alpha I(\theta_2 \leqslant \theta)$ with $\theta_j \in \Theta$, $j = 1, 2$ and $0 \leqslant \alpha \leqslant 1$. We call $1-\alpha$ and α mixing proportions, θ_1 and θ_2 mixing parameters, and Θ the mixing parameter space. The score function with respect to α at $\alpha = 0$ is

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

If the variance of this score, i.e. the Fisher information at $\alpha = 0$, is not finite, then standard asymptotic results based on the finiteness of the Fisher information must be re-examined.

Example 1. Let $X_1, \ldots, X_n$ be a random sample from the mixture of exponentials $(1-\alpha)\mathrm{Ex}(1) + \alpha\mathrm{Ex}(\theta)$, where $\mathrm{Ex}(\theta)$ denotes the exponential distribution with mean θ. The score statistic for α at $\alpha = 0$ and given θ has the form

$$S(\theta) = \sum_{i=1}^{n} \left\{ \frac{\theta^{-1}\exp(-\theta^{-1}X_i)}{\exp(-X_i)} - 1 \right\},$$

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

$$E\{S^2(\theta)\} = \begin{cases} \{n(1-\theta)^2\}/\{\theta(2-\theta)\}, & \theta < 2, \\ \infty, & \theta \geqslant 2. \end{cases}$$

Hence the only way to ensure a finite Fisher information is to require $\Theta \subset (0, 2)$.

Standard first-order asymptotic theory leads one to expect that the shape of the loglikelihood function is mostly determined by the expected or observed Fisher information. However, this intuition can be misleading with models which do not satisfy the regularity conditions considered in this paper. The loglikelihood function for simple mixture models such as (1) in fact can be very far from quadratic; see Anaya-Izquierdo & Marriott (2007a, b) and Marriott (2007). Furthermore, this shape can be dominated by a few highly influential observations even when the model is correctly specified (Marriott, 2007).

Example 2. Consider a simple normal mixture model by $(1-\alpha)N(0,1) + \alpha N(\mu, 1)$ with $\mu \in \Theta \subset R$ where R is the set of real numbers. It is common to consider the likelihood ratio test for the hypothesis $H_0 : \alpha\mu = 0$ based on a random sample $X_1, \ldots, X_n$. Hartigan (1985) showed that the likelihood ratio statistic goes to infinity in probability as $n \to \infty$ when $\Theta = R$. That is, the classical chi-squared limiting distributional result of Wilks (1938) is not applicable.

These two examples illustrate how standard asymptotic results derived from many testing procedures are only applicable to models that satisfy Assumptions A1-A5 in Appendix, and

Assumption 1. the Fisher information $E[\{f(X;\theta)/f(X;\theta_0) - 1\}^2]$ is finite under the homogeneous model $f(x;\theta_0)$ for all $\theta \in \Theta$.

Assumption 2. Θ is a compact subset of some Euclidean space.

This paper looks at ways of developing testing procedures which have standard χ^2 behaviour even when Assumptions 1 and 2 fail.

1·2. *Testing for homogeneity*

For clarity we concentrate on the case of testing the hypothesis of homogeneity against the alternative of a two-component mixture. As pointed out in Anaya-Izquierdo & Marriott (2007a) this can be challenging since the mixture can be close to the unmixed model in two quite distinct ways, either that the two components θ_1 and θ_2 in (1) are both close to θ, or that the components are far from each other but the mixing parameter α is very close to 0 or 1. It is in the second case that the Fisher information in the α-parameter direction causes most problems. Furthermore if this mixing parameter is much smaller than the inverse of the sample size then it is effectively not estimable.

A test of homogeneity for models of the form (1) is a test of the null hypothesis

$$H_0: \ \alpha(1-\alpha)(\theta_1-\theta_2)=0$$

against the alternative where $\alpha(1-\alpha)(\theta_1-\theta_2)\neq 0$. As a result of symmetry, we may and will assume $0\leqslant\alpha\leqslant 1/2$ instead of $0\leqslant\alpha\leqslant 1$.

Finding an effective and convenient method for the test of homogeneity has challenged statisticians for a long time; see Titterington et al. (1985, Ch. 5) and McLachlan & Peel (2000, Ch. 6). Although Bickel & Chernoff (1993) and Liu & Shao (2004) successfully derived the limiting distribution of the likelihood ratio statistic under the specific model in Example 2, the general problem under more useful models where Θ is not a compact subset of some Euclidean space remains open. Recent advances are mostly obtained under Assumption 2 (Dacunha-Castelle & Gassiat, 1999; Chen & Chen, 2001; Liu & Shao, 2003). In addition, either explicitly or implicitly, these results rely on Assumption 1. To better explore the problem, we show what happens when a score test is attempted.

Example 1. (Continued) We wish to test the homogeneity null hypothesis $H_0:\alpha(\theta-1)=0$. According to Davies (1977), for each given θ, we first calculate a score statistic as the derivative of the loglikelihood function with respect to α at $\alpha=0$. As a general rule, the test statistic is to be defined as $\sup_{\theta\in\Theta} n^{-1/2}S(\theta)/\surd[E\{S^2(\theta)]\}$. A test based on this statistics is not sensible because the supremum is always attained in the range of $\theta<2$.

As one of the referees pointed out, a possible remedy when using the score test in Example 1 is self-normalization by the observed Fisher information. Giné et al. (1997) showed that the self-normalized score will have a standard normal limiting distribution when $S(\theta)$ lies in the domain of attraction of the normal law even if $E\{S^2(\theta)\}=\infty$. As far as we know, the infinite Fisher information problem has not been discussed before in the mixture model context. The self-normalization technique may be useful but investigation of this is beyond the scope of this paper.

2. THE EM-TEST AND ITS ASYMPTOTIC PROPERTIES

Let $X_1,\ldots,X_n$ be a random sample of size n from a two-component mixture model (1) and let

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

be the ordinary loglikelihood function. We define the penalized loglikelihood function

$$\mathrm{PL}_n(\alpha,\theta_1,\theta_2)=l_n(\alpha,\theta_1,\theta_2)+p(\alpha)$$

where $p(\alpha)$ is a penalty function on α. The exact form of the penalty function $p(\alpha)$ will be discussed later but the idea is to bound away from cases where α is very close to zero or one.

We propose a procedure for testing for homogeneity which has been motivated by the form of the EM-algorithm. For each fixed $\alpha=\alpha_0\in(0,0.5]$, for example 0.5, we compute a penalized likelihood ratio statistic

$$M_n(\alpha_0)=2\{\mathrm{PL}_n(\alpha_0,\tilde\theta_{01},\tilde\theta_{02})-\mathrm{PL}_n(0.5,\tilde\theta_0,\tilde\theta_0)\} \tag{2}$$

with $\tilde\theta_{01}$ and $\tilde\theta_{02}$ being the maximizers of $\mathrm{PL}_n(\alpha_0,\theta_1,\theta_2)$ and $\tilde\theta_0$ being the maximizer of $pl_n(0.5,\theta,\theta)$. It can be shown that under the null model $f(x;\theta_0)$ the statistic $M_n(\alpha_0)$ has a simple χ^2-type limiting distribution even when Assumptions 1 and 2 are not satisfied. Thus it is mathematically convenient to conduct a test based on $M_n(\alpha_0)$.

If the data are from an alternative model with α different from α_0, the test based on (2) is likely to be inefficient. We solve this problem by updating the α values via an EM-iteration. The mixture model can be regarded as a model for incomplete data, where the information on the membership of observations are unknown. The EM-algorithm (Dempster et al., 1977) can be used to update iteratively the fitted values of the mixing proportions α and the mixing parameters (θ_1, θ_2). The EM-test for homogeneity follows this strategy to update the value of α_0 to achieve a better efficiency than that of $M_n(\alpha_0)$. In addition, we choose a number of initial values of α_0 to accelerate this process so that only a few iterations are necessary in order to capture the true value of θ if the data are from the alternative model. We then use the maximum value of the $M_n(\alpha_0)$-values as our test statistic.

The EM-test statistic is best explained by the procedure, initialized by choosing a number of α values, $\alpha_1, \ldots, \alpha_J$, say, computing $\tilde{\theta}_0 = \arg\max_\theta \mathrm{PL}_n(0.5, \theta, \theta)$, and letting $j = 1$ and $k = 0$.

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

Step 2. Compute $(\theta_{j1}^{(k)}, \theta_{j2}^{(k)}) = \arg\max_{\theta_1,\theta_2} \mathrm{PL}_n(\alpha_j^{(k)}, \theta_1, \theta_2)$ and

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

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

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

Now following the M-step, let

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

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

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

Compute

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

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

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

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

$$\mathrm{EM}_n^{(k)} = \max\{M_n^{(k)}(\alpha_j), j = 1, \ldots, J\}.$$

When the number of EM-iterations tends to infinity under the assumption that the EM-algorithm converges to a global maximum, the EM-test statistic becomes the modified likelihood ratio test, see Chen (1998), Chen et al. (2001, 2004). The modified likelihood ratio test enjoys a simple limiting distribution only under Assumptions 1 and 2. Therefore, although letting $k = \infty$ may further increase the value of the EM-test statistic, its nice asymptotic properties become inapplicable for providing a critical value of the test. Without the penalty term $p(\alpha)$, the EM-test reduces to the ordinary likelihood ratio test when $k = \infty$. The likelihood ratio test has a complicated limiting distribution which is available only under more restrictive conditions.

The mboxEM-test is partially motivated by the score test discussed in Liang & Rathouz (1999). Their score test can be directly used for the models in Examples 1 and 2. In both tests, a pre-chosen value of the mixing proportion is used. However, the EM-test iterates to find a more suitable mixing proportion which improves the power, while the score test has no such mechanism: it uses a single α value regardless of the actual fitting of the data.

Chen & Cheng (1995) and Lemdani & Pons (1995) proposed a constrained test based on

$$R_n(\epsilon_0) = 2\big\{ \sup_{\alpha\in[\epsilon_0,\, 0.5],\, \theta_1,\, \theta_2} l_n(\alpha, \theta_1, \theta_2) - l_n(0.5, \tilde{\theta}_0, \tilde{\theta}_0)\big\}$$

where $\epsilon_0 \in (0, 1/2]$ is a fixed positive constant. There are some similarities between this method and the EM-test, because the EM-test requires that the pre-chosen mixing proportions be larger than zero. However, the EM-iteration allows us to recoup the mixture models with smaller mixing proportions while the constrained method does not.

The computation of $\text{EM}_n^{(k)}$ is very simple. In practice, when the sample size is small, one might want to simulate the empirical critical values of the EM-test by a Monte Carlo or bootstrap method. The computational advantage of the EM-test will make such a simulation easy.

Under very general conditions, for fixed finite k and any finite set of pre-chosen α_j, the test statistic $\text{EM}_n^{(k)}$ has the limiting distribution $0.5\chi_0^2 + 0.5\chi_1^2$. This is shown in the following theorems, whose proofs are given in the Appendix.

THEOREM 1. *Suppose that $f(x;\theta)$ satisfies Assumptions* A1-A5 *given in the Appendix, and $p(\alpha)$ is a continuous function such that $p(\alpha) \to -\infty$ as $\alpha \to 0$ and which attains its maximal value at $\alpha = 0.5$. Under the null distribution $f(x;\theta_0)$, we have, for $j = 1, \ldots, J$ and any fixed finite k,*

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

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

Based on the above result, we can easily derive the null distribution of $\text{EM}_n^{(k)}$.

THEOREM 2. *Assume the same conditions as in Theorem* 1, *and that one of the α_j's is equal to* 0.5. *Under the null distribution $f(x;\theta_0)$, and for any fixed finite k, as $n \to \infty$,*

$$\text{EM}_n^{(k)} \to 0.5\chi_0^2 + 0.5\chi_1^2,$$

in distribution.

Remark 1. For each given $\alpha \in (0, 0.5]$, $M_n(\alpha)$ in (2) can be written as the sum of two terms, one from the likelihood function and the other from the penalty term. Under the null model, the first term has the same quadratic approximation for all α-values. However, different α-values result in different sizes of the penalty function. Since the penalty $p(\alpha)$ attains its maximum at $\alpha = 0.5$, including $\alpha = 0.5$ implies that the limiting distribution is determined by the quadratic approximation only, and hence has the simplest form.

We emphasize here that Assumptions 1 and 2 are not required for the above results. Hence, the EM-test is both convenient in applications and widely applicable.

3. Two precision-enhancing measures

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

For the validity of the asymptotic result, $p(\alpha)$ must decrease to $-\infty$ when $\alpha \to 0$ and must be maximized at $\alpha = 1/2$. For a finite sample size, the choice of the penalty function $p(\alpha)$ may affect the accuracy of the null limiting distribution. It is important to choose a penalty which best balances the Type I error and the power. Other considerations include computational convenience. In the current paper, we find that the penalty function

$$p(\alpha) = C\log(1 - |1 - 2\alpha|) \tag{3}$$

for some positive C is a very good choice. Since

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

with the same value of constant C, this penalty is more severe than the penalty function $C\log\{4\alpha(1 - \alpha)\}$ introduced for the modified likelihood ratio test (Chen et al., 2001). The difference is relatively small when $\alpha - 0.5 \simeq 0$, and large when $\alpha - 0.5$ deviates from 0. As a result, the current choice helps to reduce the Type I error without limiting the power of the EM-test. In addition, when $\alpha \simeq 0.5$, $\log(1 - |1 - 2\alpha|) \simeq -|1 - 2\alpha|$. The penalty (3) is therefore a lasso-type penalty (Tibshirani, 1996); that is, it is a continuous function for all α, but not smooth at $\alpha = 0.5$. It has therefore similar properties to the lasso-type penalty for linear regression (Tibshirani, 1996), the probability of the fitted value of α being 0.5 is positive. In comparison, the penalty $\log\{4\alpha(1 - \alpha)\}$ is smooth at $\alpha = 0.5$ and does not have this property.

These two penalty functions are special cases of $C\log(1 - |1 - 2\alpha|^h)$ for some $0 < h \leqslant 2$. A choice of $0 < h < 1$ may further improve the power of the EM-test. We recommend the choice of $h = 1$ for the following reasons. First, when $h = 1$, in Step 3 of the algorithm the α values can be easily updated as follows:

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

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

The next precision enhancing measure is motivated by the following observation. It is suggestive that $(1 - p_n)\chi_0^2 + p_n\chi_1^2$ with $p_n = \mathrm{pr}(\mathrm{EM}_n^{(k)} > 0)$ may approximate better the finite-sample distribution than does the asymptotic limit given above. A good approximation for p_n might therefore be useful. Let $\mu(f)$ and $\sigma^2(f)$ be, respectively, the mean and variance under the homogeneous model. Furthermore, let $S = E[\{X_1 - \mu(f)\}^2] - \sigma^2(f)$ be an over-dispersion measure, where the mixture model would only be justified when $S > 0$. Note that $S_n = \sum_{i=1}^n (X_i - \bar{X})^2/n - \hat{\sigma}^2(f)$ provides consistent estimation of the over-dispersion measure S, where $\hat{\sigma}^2(f)$ is a consistent estimator of $\sigma^2(f)$. Intuitively, if $S_n \leqslant 0$, the homogeneous model should be not rejected and therefore we approximate p_n by $\mathrm{pr}\{S_n > 0\}$.

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

PROPOSITION 1. *Under the null hypothesis and if $E(X_1^6) < \infty$, then*

$$p_n = \mathrm{pr}\{S_n > 0\} = 0.5 + (2\pi n)^{-1/2}(a - b/6) + o_p(n^{-1/2}), \quad (4)$$

where

$$a = \lim_{n\to\infty} n^{1/2} E\Big\{\frac{S_n}{\surd\{\mathrm{var}(S_n)\}}\Big\},\ \ b = \lim_{n\to\infty} n^{1/2} E\Big\{\frac{S_n - E(S_n)}{\surd\{\mathrm{var}(S_n)\}}\Big\}^3.$$

Furthermore, if $E(X_1^{10}) < \infty$, then the remainders term $o_p(n^{-1/2})$ in (4) can be strengthened to $O_p(n^{-3/2})$.

In the above proposition, the Edgeworth approximation relies on the condition $E(X_1^6) < \infty$. There exists some distributions, such as the exponential distribution and the geometric distribution, which satisfy this condition or even the condition $E(X_1^{10}) < \infty$, but do not satisfy Assumption 1. The condition $E(X_1^6) < \infty$ is therefore not as restrictive as Assumption 1. The quantities a and b may depend on unknown parameters, in which case we replace them by their consistent estimates under the homogeneity model.

The penalty function $p(\alpha)$ clearly has effects on the probability of $\mathrm{EM}_n^{(k)} = 0$, but this is not reflected in the Edgeworth expansion. Simulations show that the expansion works well for the penalty in (3) and a range of C values. A refined approximation which depends on the choice of the penalty and the value of C is worth further investigation.

For many commonly used distributions, we can compute a and b analytically and the results are presented in the Table 1. In the Poisson and binomial examples, we can replace the unknown θ by its maximum likelihood estimate under the null model.

Table 1. *Edgeworth approximations of p_n for commonly used kernel functions.*

Kernel	Edgeworth approximation
$N(\mu, \sigma_0^2)$	$0.5 - 5/\{6\surd(\pi n)\} + O_p(n^{-3/2})$
Po(θ)	$0.5 - (5\theta + 1)/\{6\theta\surd(\pi n)\} + O_p(n^{-3/2})$
Bi(m, θ)	$0.5 - \{\theta(1-\theta)(5m - 11) + 1\}/[6\theta(1-\theta)\surd\{\pi n m(m-1)\}] + O_p(n^{-3/2})$
Ex(θ)	$0.5 - 8/\surd(18\pi n) + O_p(n^{-3/2})$

σ_0^2 in the normal kernel is assumed known

We recommend the use of penalty function (3) for the EM-test, together with its higher order adjustment. These two practical considerations enhance the performance of the new method.

4. SIMULATION STUDY

Our simulation study examines many aspects of the EM-test and related issues. First, we examine the precision of the Edgeworth expansion for $p_n = \mathrm{pr}(\mathrm{EM}_n^{(k)} > 0)$. We considered null models with kernels $N(0, 1)$, Po(5), Ex(5), and Bi(10,0.5). In each case, we generated random samples of sizes $n = 100$ and $n = 200$. We computed $\mathrm{EM}_n^{(k)}$ for $k = 0, 1, 2$ for each kernel using (3) with $C = 1$, and for two sets of initial values for α: $\{0.1, 0.2, 0.3, 0.4, 0.5\}$ and $\{0.1, 0.3, 0.5\}$. The nonzero proportions of $\mathrm{EM}_n^{(k)}$, $k = 0, 1, 2$ were calculated based on 20000 repetitions. Since the results for two sets of initial α-values are almost identical, we only report the results based

Table 2. *Simulated nonzero proportions for the* EM*-test statistics.*

Kernel	$\text{EM}_n^{(0)}$	$\text{EM}_n^{(1)}$	$\text{EM}_n^{(2)}$	Edgeworth approximation	Standard deviation
			$n = 100$		
$N(0, 1)$	0.449	0.449	0.449	0.453	0.0035
Po(5)	0.449	0.449	0.449	0.451	0.0035
Bi(10,0.5)	0.453	0.453	0.453	0.457	0.0035
Ex(5)	0.395	0.395	0.395	0.394	0.0035
			$n = 200$		
$N(0, 1)$	0.463	0.463	0.463	0.467	0.0035
Po(5)	0.465	0.465	0.465	0.465	0.0035
Bi(10,0.5)	0.467	0.467	0.467	0.470	0.0035
Ex(5)	0.423	0.423	0.423	0.425	0.0035

on the second set; see Table 2. Clearly, (4) provides a very good approximation to p_n in all cases considered.

Next, we compare the EM-test and the modified likelihood ratio test for the Poisson mixture case. The mean values for the null distribution and the alternative distribution are chosen to be 5. Four alternative models are selected so that $1 - \alpha = 0.5, 0.25, 0.1, 0.05$ and the variances of the mixing distributions are set to be $\Delta = \alpha(1-\alpha)(\theta_1 - \theta_2)^2 = 1.25$; see Table 3 for details and for the corresponding Kullback–Leibler information with respect to the null model. For the modified likelihood ratio test, we used the penalty function $p(\alpha) = C \log\{4\alpha(1-\alpha)\}$ with $C = \log(50)$. The EM-test statistics were computed in the same way as before. The choice of $C = \log(50)$ for the modified likelihood ratio test was made in accordance to the recommendations in Chen et al. (2001); furthermore it worked well in our pre-trials, in that the two methods have close nominal Type I errors in all cases considered. The EM-test statistics were computed with penalty function (3) and $C = 1$. Although a specific reason for choosing $C = 1$ is lacking, we have seen ample evidence that it is a sensible choice in a wide range of applications. It would be ideal if a data-driven procedure with some theoretical justification could be found to justify this choice, but, lacking that, we recommend a pilot simulation study or a literature search before each application to ensure that C is chosen so that the Type I errors are no more than 5.5% when the target is 5%.

We computed the null rejection rates based on 20000 repetitions and the powers based on 10000 repetitions. The results are reported in Table 4. We find that the null rejection rates of both the modified likelihood ratio test and the EM-test in all cases are close to the nominal values. The EM-tests are generally more efficient particularly when $|\alpha - 0.5|$ is relatively large. Using the EM-tests with five initial α values does not noticeably improve its power. Also, $\text{EM}_n^{(1)}$ and $\text{EM}_n^{(2)}$ have better powers compared to $\text{EM}_n^{(0)}$, but they do not differ much. Thus, we come to the general recommendation of $\text{EM}_n^{(1)}$ with $\alpha \in \{0.1, 0.3, 0.5\}$ paired with the penalty function (3) and $C = 1$.

We also investigated the modified likelihood ratio test with (3) and $C = 1$, and found that the modified likelihood ratio test and the EM-test have similar Type I and Type II errors in all cases. This is not unexpected because the modified likelihood ratio test is the EM-test statistic with $k = \infty$. A crucial difference is that the asymptotic result of the EM-test is much more widely applicable. The power comparisons between the EM-test and the modified likelihood ratio test

Table 3. *Parameters in the alternative models.*

	$1-\alpha$	θ_1	θ_2	Δ	100KL
Poisson mixtures:					
Model I	0.50	3.882	6.118	1.25	1.751
Model II	0.25	3.064	5.645	1.25	2.017
Model III	0.10	1.646	5.373	1.25	2.827
Model IV	0.05	0.127	5.256	1.25	5.081
Exponential mixtures:					
Model I	0.50	3.129	6.871	3.50	1.008
Model II	0.25	2.128	5.957	2.75	0.956
Model III	0.10	0.757	5.471	2.00	1.252
Model IV	0.05	0.127	5.256	1.25	1.996

Δ, variance of the mixing distribution;
KŁ, Kullback–Leibler information.

Table 4. *Rejection rates of the* EM*-test and the modified likelihood ratio test under Poisson mixtures at the 5% level.*

Model	MLRT	$\text{EM}_n^{(0)}$	$\text{EM}_n^{(1)}$	$\text{EM}_n^{(2)}$	$\text{EM}_n^{(0)}$	$\text{EM}_n^{(1)}$	$\text{EM}_n^{(2)}$
				$n = 100$			
H_0	5.0	5.1	5.2	5.2	5.1	5.1	5.1
I	49.4	49.0	49.0	49.0	49.0	49.0	49.0
II	51.9	51.8	51.8	51.9	51.8	51.8	51.8
III	53.8	57.1	57.3	57.4	56.8	57.1	57.2
IV	63.1	72.0	74.3	74.5	72.0	74.3	74.5
				$n = 200$			
H_0	4.9	4.9	5.0	5.0	4.9	4.9	4.9
I	74.2	73.9	73.9	73.9	73.9	73.9	73.9
II	76.3	76.5	76.5	76.5	76.4	76.4	76.4
III	78.1	81.6	81.7	81.8	81.5	81.7	81.7
IV	87.0	91.5	92.2	92.4	91.5	92.2	92.4

Results in columns (3, 4, 5) used $\alpha = (0.1, 0.2, 0.3, 0.4, 0.5)$.
Results in columns (6, 7, 8) used $\alpha = (0.1, 0.3, 0.5)$.
MLRT: Modified likelihood ratio test.

under the binomial kernel and normal kernel are similar to what we have seen under the Poisson kernel. For the sake of brevity, we do not present details here.

We now study the EM-test in cases where the asymptotic results of the modified likelihood ratio test or the likelihood ratio test are not applicable. The exponential kernel is used in this simulation. We set the mean of the mixture model to be 5 in all cases and the same parameter values for alternative models as in Table 3. Although the limiting distributions of the likelihood ratio statistic, denoted by R_n, and D-test (Charnigo & Sun, 2004) are not available, these tests can be calibrated using simulated quantiles under the null models. Hence, they are included in the simulation to serve as efficiency barometers. We use the notation $d(2, n)$, $d_1(2, n)$ and $d_2(2, n)$ for the D-test with weighting functions 1, x and x^2, respectively. The modified likelihood ratio test can also be calibrated by simulated quantiles, but it is bounded by the EM-test and the likelihood ratio test and therefore is not included. The constrained likelihood ratio test $R_n(\epsilon_0)$ is applicable under the same conditions as the EM-test. We find that it has large Type I errors unless the lower bound for α, ϵ_0, is relatively large. Through some pilot simulation studies, we found that its Type I errors match those of EM-tests when $\epsilon_0 = 0.45$. We thus included the constrained likelihood ratio test $R_n(0.45)$ in our simulation.

Software for calculating the critical values of the D-test for the Ex(1) distribution can be found at http://stat.cwru.edu/˜rjc12. For other null distributions, the transformations suggested in Charnigo & Sun (2004) were used. First we computed $\mathrm{EM}_n^{(k)}$ for $k = 0, 1, 2$ with $C = 1$ and $\alpha \in \{0.1, 0.3, 0.5\}$ first. Their Type I errors are also somewhat larger than the nominal values, and we therefore also computed the EM-tests with $C = 1.5$. The null rejection rates of the EM-tests, D-tests and constrained likelihood ratio test are calibrated by either limiting distributions or by critical values obtained from references and are shown in Table 5. The EM-tests and the constrained likelihood ratio test have reasonably accurate Type I errors. The D-test statistics may not be sufficiently invariant to allow transformation of critical values between the Ex(1) and the Ex(5) null distributions. Although the sizes of the EM-tests are slightly large, this effect is not too severe with both $C = 1$ and $C = 1.5$. Both meet the recommendation criterion set earlier.

The power calculations of all methods were done using simulated quantiles to ensure objective comparisons. In general, the efficiency of the EM-test is much better than that of other methods. The D-test based on $d(2, n)$ is less efficient than the EM-test when α is close to 0.5, but is more efficient for alternatives when α is close to 1. This result may not be very useful because the Type I error of the $d(2, n)$ based D-test is hard to control. An interesting result is that the EM-test is much more efficient than the likelihood ratio test when α is close to 0.5. As a result of the penalty function, the EM-test is expected to lose power when α is close 0.

Table 5. *Rejection rates of the* D*-test, the* EM*-test, the constrained likelihood ratio test and the likelihood ratio test under exponential mixture alternatives at the 5% level.*

				$C = 1$		$C = 1.5$			
Model	$d(2,n)$	$d_1(2,n)$	$d_2(2,n)$	$em_n^{(0)}$	$em_n^{(1)}$	$em_n^{(0)}$	$em_n^{(1)}$	$R_n(0.45)$	R_n
			$n = 100$						
H_0	12.2	5.1	4.0	5.8	6.0	5.3	5.4	5.5	–
I	17.6	30.1	32.6	33.6	33.4	34.1	34.0	34.6	29.8
II	22.1	30.7	30.3	31.0	30.8	31.3	31.3	31.2	27.9
III	35.4	31.9	24.3	29.2	29.6	27.6	27.9	24.9	32.6
IV	49.5	21.9	10.1	32.2	32.9	28.4	29.0	17.6	42.3
			$n = 200$						
H_0	13.5	7.3	4.3	5.5	5.5	5.2	5.2	5.3	–
I	26.3	48.1	51.2	53.4	53.3	53.6	53.6	54.1	47.6
II	34.5	49.3	47.7	48.0	48.0	47.5	47.6	47.4	44.6
III	54.6	51.7	39.8	45.8	46.0	42.1	42.5	37.5	52.1
IV	66.4	34.9	12.4	46.8	48.6	42.2	43.6	22.5	61.5

5. REAL DATA EXAMPLES

Example 3. First, we apply the EM-test to the data studied in Proschan (1963). The data consist of the times of successive failures for the air conditioning system of each member in a fleet of 13 Boeing 720 jet aircrafts. Proschan (1963) applied the Kolmogorov-Smirnov test to the pooled data, a total of 213 observations, to determine whether or not the exponential distribution offered a good fit to the pooled failure times. At the level of 0.05, the Kolmogorov-Smirnov test failed to reject the null hypothesis of exponential fit. However, the exponential distribution did not fit the pooled failure times very well. Proschan (1963) plotted the log empirical survival curve for the pooled data and the log theoretical survival curve under the exponential model and observed that the empirical curve lies consistently below the theoretical curve when the failure time is less than 150 and above the theoretical curve when the failure time is larger than 150.

Proschan (1963) further used a more refined analysis to show that the failure distribution for each aircraft separately was exponential, but for some aircrafts the rates were different. It is therefore reasonable to assume the pooled failure times follow a mixture of exponential distributions. Now we conduct a test of homogeneity for the pooled data. The maximum likelihood estimates for $(\alpha, \theta_1, \theta_2)$ under the mixture model are $(0.430, 128.286, 46.506)$. Since $\hat{\theta}_2/\hat{\theta}_1 = 2.758 > 2$, most existing methods of testing the homogeneity are strictly not applicable because the density ratio may have infinite second moment, and hence infinite Fisher information. In contrast, a rigorous EM-test can be conducted. According to our simulations, $C = 1.5$ is a good choice for the level of modification for the pooled failure times. We computed the EM-statistics with $C = 1.5$ and three initial values (0.1,0.3,0.5) of α, and found $em_n^{(0)} = em_n^{(1)} = 6.221$. With a sample size of 213, according to Table 1, p_n will be well approximated by 0.427. In view of the adjusted limiting distribution $0.573\chi_0^2 + 0.427\chi_1^2$, the asymptotic p-value for the EM-test is 0.005. For the constrained likelihood ratio test, we have $R_n(0.45) = 6.30$ with the asymptotic p-value 0.005. We also calculate the likelihood ratio statistic, $R_n = 6.31$. We simulated the quantiles of the likelihood ratio statistic with 10,000 repetitions and found the simulated p-value to be 0.019. For the pooled failure data, therefore, the EM-test and the constrained likelihood ratio test give stronger evidence than the likelihood ratio test for rejecting the homogeneous exponential fit.

Acknowledgement

This research was partially supported by the Natural Science and Engineering Research Council of Canada and the Mathematics of Information Technology and Complex Systems.

Appendix

Some notation and regularity conditions

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

Assumption A1: *Wald's integrability conditions.* (i) $E|\log f(X;\theta_0)| < \infty$; (ii) for sufficiently small ρ and for sufficiently large r, the expected values $E\log\{1 + f(X;\theta,\rho)\} < \infty$ for $\theta \in \Theta$ and $E\log\{1+\varphi(X,r)\} < \infty$, where $f(x;\theta,\rho) = \sup_{|\theta'-\theta|\leqslant\rho} f(x;\theta')$ and $\varphi(x;r) = \sup_{|\theta|\geqslant r} f(x;\theta)$; (iii) $\lim_{|\theta|\to\infty} f(x;\theta) = 0$ for all x except on a set with probability zero.

Assumption A2: *Smoothness.* The kernel function $f(x;\theta)$ has common support and is three times continuously differentiable with respect to θ. The first two derivatives are denoted by $f'(x;\theta)$ and $f''(x;\theta)$.

Assumption A3: *Identifiability*. For any two mixing distribution functions Ψ_1 and Ψ_2 with two supporting points such that $\int f(x;\theta)d\Psi_1(\theta) = \int f(x;\theta)d\Psi_2(\theta)$, for all x, we must have $\Psi_1 = \Psi_2$.

Assumption A4: *Uniform boundedness.* Let

$$Y_i(\theta) = \frac{f(X_i;\theta) - f(X_i;\theta_0)}{(\theta-\theta_0)f(X_i;\theta_0)}, \;\; \theta \neq \theta_0; \;\; Y_i = Y_i(\theta_0) = \frac{f'(X_i;\theta_0)}{f(X_i;\theta_0)} \tag{A1}$$

$$Z_i(\theta) = \frac{Y_i(\theta) - Y_i(\theta_0)}{(\theta-\theta_0)}, \; \theta \neq \theta_0; \; Z_i = Z_i(\theta_0) = \frac{f''(X_i;\theta_0)}{2f(X_i;\theta_0)}. \tag{A2}$$

For some neighbourhood $N(\theta_0)$ of θ_0, there exists a g with finite expectation such that $|Y_i(\theta)|^3 \leqslant g(X_i)$, $|Z_i(\theta)|^3 \leqslant g(X_i)$ and $|Z_i''(\theta)|^2 \leqslant g(X_i)$.

Assumption A5: *Positive definiteness.* The covariance matrix of (Y_i, Z_i) is positive definite.

Proofs of Theorems 1 and 2

A brief roadmap for the proofs is as follows. Lemma A1 shows that any estimator with α bounded away from 0 or 1, and with a large likelihood value, is consistent for θ_1 and θ_2 under the null model, which can be seen as the extension of the results in Wald (1949). Lemma A2 strengthens Lemma A1 by providing specific convergence rates. Lemma A3 makes Lemmas A1 and A2 applicable to $(\alpha_j^{(k)}, \theta_{j1}^{(k)}, \theta_{j2}^{(k)})$, by showing that the EM-iteration keeps $\alpha_j^{(k)}$ in a small neighbourhood of α_j and therefore away from 0 or 1. Theorems 1 and 2 then follow easily.

LEMMA A1. *Suppose that Assumptions A1-A3 hold. Let $(\bar\alpha, \bar\theta_1, \bar\theta_2)$ be estimators of $(\alpha, \theta_1, \theta_2)$ such that $\delta \leqslant \bar\alpha \leqslant 0.5$ for some $\delta \in (0, 0.5]$. Assume that*

$$l_n(\bar\alpha, \bar\theta_1, \bar\theta_2) - l_n(0.5, \theta_0, \theta_0) \geqslant c > -\infty.$$

Then, under the null distribution $f(x;\theta_0)$, $\bar\theta_1 - \theta_0 = o_p(1)$ and $\bar\theta_2 - \theta_0 = o_p(1)$.

Proof. The parameter space under the full model (1) with the restriction $\delta \leqslant \bar\alpha \leqslant 0.5$ becomes $\Lambda = [\delta, 0.5] \times \Theta \times \Theta$. The parameter value of a null model belongs to $\{(\alpha, \theta_0, \theta_0) : \delta \leqslant \alpha \leqslant 0.5\}$.

First, for some positive constants ϵ and r, let

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

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

By Assumptions A1 and A2, it is obvious that, for all small enough ϵ and large enough r,

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

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

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

almost surely for any $c > -\infty$. By compactness, there exist α_j, $j = 1, \ldots, J$, such that $[\delta, 0.5] \subset A = \cup_{j=1}^J A(\alpha_j; \epsilon, r)$ and each $A(\alpha_j; \epsilon, r)$ has the above property. Therefore

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

The same conclusion and proof are applicable to

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

and hence also to $B = \cup\{B(\alpha, \theta_1; \epsilon, r) : \delta \leqslant \alpha \leqslant 1 - \delta, |\theta_1| \leqslant r\}$. In words, the loglikelihood at any parameter point with either θ_1 or θ_2 very large trails the loglikelihood at the true parameter point by an infinite amount.

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

LEMMA A2. *Suppose the conditions of Theorem* 1 *on $f(x;\theta)$ and $p(\alpha)$ hold. Let $(\bar\alpha, \bar\theta_1, \bar\theta_2)$ be estimators of $(\alpha, \theta_1, \theta_2)$ such that, under the null hypothesis, $\bar\theta_1 - \theta_0 = o_p(1)$, $\bar\theta_2 - \theta_0 = o_p(1)$, $\delta \leqslant \bar\alpha \leqslant 0.5$, for some $\delta \in (0, 0.5]$. If, for all n and $X_1, \ldots, X_n$,*

$$\mathrm{PL}_n(\bar\alpha, \bar\theta_1, \bar\theta_2) - pl_n(0.5, \theta_0, \theta_0) \geqslant c > -\infty,$$

then, under the null distribution $f(x;\theta_0)$, $\bar\theta_1 - \theta_0 = O_p(n^{-1/4})$, $\bar\theta_2 - \theta_0 = O_p(n^{-1/4})$, $\bar m_1 = (1 - \bar\alpha)(\bar\theta_1 - \theta_0) + \bar\alpha(\bar\theta_2 - \theta_0) = O_p(n^{-1/2})$.

Proof. For $i = 1, \ldots, n$, let $W_i = Z_i - \beta Y_i$ with $\beta = E(Y_1 Z_1)/E(Y_1^2)$. Furthermore, let $\bar{m} = \bar{m}_1 + \beta \bar{m}_2$ with $\bar{m}_2 = (1-\bar{\alpha})(\bar{\theta}_1 - \theta_0)^2 + \bar{\alpha}(\bar{\theta}_2 - \theta_0)^2$.

Since $\bar{\theta}_1$ and $\bar{\theta}_2$ are in a small neighbourhood of θ_0, in probability, by Taylor expansion, we obtain

$$\begin{aligned}
&2\{\mathrm{PL}_n(\bar{\alpha}, \bar{\theta}_1, \bar{\theta}_2) - pl_n(0.5, \theta_0, \theta_0)\} \\
&\leqslant 2\{l_n(\bar{\alpha}, \bar{\theta}_1, \bar{\theta}_2) - l_n(0.5, \theta_0, \theta_0)\} \\
&\leqslant 2\sum_{i=1}^n (\bar{m} Y_i + \bar{m}_2 W_i) - \left(\bar{m}^2 \sum_{i=1}^n Y_i^2 + \bar{m}_2^2 \sum_{i=1}^n W_i^2\right)\{1 + o_p(1)\} + o_p(1) \\
&\leqslant \frac{\{(\sum_{i=1}^n W_i)^+\}^2}{\sum_{i=1}^n W_i^2} + \frac{(\sum_{i=1}^n Y_i)^2}{\sum_{i=1}^n Y_i^2} + o_p(1). \qquad (A3)
\end{aligned}$$

We do not have cross terms in the second line because Y_i and W_i are uncorrelated. The last inequality follows from the property of the quadratic function and the nonnegativeness of $\bar{m}_2$.

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

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

Since $\sum_{i=1}^n W_i = O_p(n^{1/2})$ and $\sum_{i=1}^n W_i^2 = O_p(n)$, we obtain $\bar{m}_2 = O_p(n^{-1/2})$. Since $\delta \leqslant \bar{\alpha} \leqslant 0.5$ for some $\delta \in (0, 0.5]$, we further conclude that $\bar{\theta}_1 - \theta_0 = O_p(n^{-1/4})$, $\bar{\theta}_2 - \theta_0 = O_p(n^{-1/4})$. Similarly, we have $\bar{m}_1 = O_p(n^{-1/2})$ and therefore $\bar{m}_1 = O_p(n^{-1/2})$. □

Now we show that, under the null model, the EM-iteration changes the fitted value of α by $o_p(1)$. Let $(\bar{\alpha}, \bar{\theta}_1, \bar{\theta}_2)$ be some estimators of $(\alpha, \theta_1, \theta_2)$ with the same asymptotic properties as before, and let

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

We further define

$$R_n(\alpha) = \left(n - \sum_{i=1}^n \bar{w}_i\right) \log(1-\alpha) + \sum_{i=1}^n \bar{w}_i \log(\alpha)$$

and $H_n(\alpha) = R_n(\alpha) + p(\alpha)$. The EM-test updates α by searching for $\bar{\alpha}^* = \arg\max_\alpha Q_n(\alpha)$.

LEMMA A3. *Suppose that the conditions of Lemma* A2 *hold and* $\bar{\alpha} - \alpha_0 = o_p(1)$ *for some* $\alpha_0 \in (0, 0.5]$. *Under the null distribution* $f(x; \theta_0)$, *we have* $|\bar{\alpha}^* - \alpha_0| = o_p(1)$.

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

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

where Y_i and Z_i are defined in (A1) and (A2). Thus,

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

By Assumption A4 and a result on order statistics in Serfling(1980, p. 90), we have

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

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

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

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

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

$$|\tilde\alpha - \bar\alpha| = \Big\{(\bar\theta_2 - \bar\theta_1)\sum_{i=1}^n Y_i + (\bar\theta_2-\theta_0)^2\sum_{i=1}^n Z_i(\bar\theta_2) - (\bar\theta_1-\theta_0)^2\sum_{i=1}^n Z_i(\bar\theta_1)\Big\}O_p(n^{-1}) = o_p(1).$$

By this result and the assumption that $\bar\alpha - \alpha_0 = o_p(1)$, we have $\tilde\alpha - \alpha_0 = o_p(1)$ and hence it suffices to prove that $\bar\alpha^* - \tilde\alpha = o_p(1)$.

As $R_n(\alpha)$ is a binomial loglikelihood, it attains its maximum at $\tilde\alpha$ and decreases on both sides. For any $\epsilon > 0$ and $\alpha \geqslant \tilde\alpha + 2\epsilon$, by the mean value theorem,

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

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

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

Hence, with probability approaching 1,

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

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

We now prove Theorems 1 and 2 by showing that the slightly more general results in previous lemmas are applicable.

Proof of Theorem 1. By the property of EM algorithm (Dempster et al., 1977), the definition of $\alpha_j^{(k)}$, for any finite k, we have

$$\mathrm{PL}_n(\alpha_j^{(k)}, \theta_{j1}^{(k)}, \theta_{j2}^{(k)}) \geqslant \mathrm{PL}_n(\alpha_j, \theta_{j1}^{(0)}, \theta_{j2}^{(0)}) \geqslant \mathrm{PL}_n(\alpha_j, \theta_0, \theta_0).$$

Therefore

$$l_n(\alpha_j^{(k)}, \theta_{j1}^{(k)}, \theta_{j2}^{k)}) - l_n(\alpha_j, \theta_0, \theta_0) \geqslant p(\alpha_j) - p(\alpha_j^{(k)}) \geq p(\alpha_j) - p(0.5) > -\infty.$$

By Lemma A1 and $\alpha_j^{(0)} = \alpha_j$, we have shown that $\theta_{j1}^{(0)}$ and $\theta_{j2}^{0)}$ are consistent for θ_0. As a result, the conclusions of Lemmas A2 and A3 apply. Hence, we find

$$\alpha_j^{(1)} - \alpha_j = o_p(1),\ \ \theta_{j1}^{(1)} - \theta_0 = O_p(n^{-1/4}),\ \ \theta_{j2}^{(1)} - \theta_0 = O_p(n^{-1/4}).$$

The above results for $k = 1$ are then used to show the same conclusions for $k = 2$. By mathematical induction, the conclusion of the theorem is true for all finite k. □

Proof of Theorem 2. By the properties proved in Theorem 1, the inequality (A3) is applicable. Hence, for any (j, k), we have

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

It is obvious that

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

Hence, we have

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

It is simple to show that

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

when $\alpha_j = 0.5$. Thus,

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

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

References

Anaya-Izquierdo K. A. & Marriott P. (2007a). Local mixture models of exponential families. *Bernoulli* **13**, 623–40.

Anaya-Izquierdo K. A. & Marriott P. (2007b). Local mixtures of the exponential distribution. *Ann. Inst. Statist. Math.* **59**, 111–34.

Bickel, P. & Chernoff, H. (1993). Asymptotic distribution of the likelihood ratio statistic in a prototypical non regular problem, Eds. J.K. Ghosh, S.K. Mitra, K.R. Parthasarathy, B.L.S. PrakasaRao, pp. 83-96. *Statistics and Probability*: A Raghu Raj Bahadur Festschrift. Wiley Eastern, New Delhi.

Charnigo, R. & Sun J. (2004). Testing homogeneity in a mixture distribution via the L^2-distance between competing models. *J. Am. Statist. Assoc.* **99**, 488–98.

Chen, H. & Chen, J. (2001). The likelihood ratio test for homogeneity in the finite mixture models. *Can. J. Statist.* **29**, 201–15.

Chen, H., Chen, J. & Kalbfleisch, J. D. (2001). A modified likelihood ratio test for homogeneity in finite mixture models. *J. R. Statist. Soc.* B **63**, 19-29.

Chen, H., Chen, J. & Kalbfleisch, J. D. (2004). Testing for a finite mixture model with two components. *J. R. Statist. Soc.* B **66**, 95-115.

Chen, J. (1998). Penalized likelihood ratio test for finite mixture models with multinomial observations. *Can. J. Statist.* **26**, 583–99.

Chen, J. & Cheng, P. (1995). The Limit distribution of the restricted likelihood ratio statistic for finite mixture models. *Northeast. Math. J.* **11**, 365–74.

Dacunha-Castelle, D. & Gassiat, E. (1999). Testing the order of a model using locally conic parametrization: population mixtures and stationary ARMA processes. *Ann. Statist.* **27**, 1178–209.

Davies R. B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **64**, 247–54.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via EM algorithm (with Discussion). *J. R. Statist. Soc.* B **39**, 1-38.

Giné E., Götze F. & Mason D. M. (1997). When is the Student t-statistic asymptotically standard normal? *Ann. Prob.* **25**, 1514–31.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer Series in Statistics. New York, Springer.

Hartigan, J. A. (1985). A failure of likelihood asymptotics for normal mixtures, In *Proc. Berkeley Conf. in Honor of J. Neyman and Kiefer, Volume 2*, Eds L. LeCam and R. A. Olshen, Wadsworth, Belmont, CA. 807–10.

Lemdani, M. & Pons, O. (1995). Tests for genetic linkage and homogeneity. *Biometrics* **51**, 1033–41.

Liang, K. Y. & Rathouz, P. J. (1999). Hypothesis testing under mixture models: application to genetic linkage analysis. *Biometics* **55**, 65-74.

Liu, X. & Shao, Y. Z. (2003). Asymptotics for likelihood ratio tests under loss of identifiability. *Ann. Statist.* **31**, 807–32.

Liu, X. & Shao, Y. Z. (2004). Asymptotics for the likelihood ratio test in a two-component normal mixture model. *J. Statist. Plan. Infer.* **123**, 61-81.

Marriott, P. (2007). Extending local mixture models. *Ann. Inst. Statist. Math.* **59**, 95-110.

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

Proschan, F. (1963). Theoretical explanation of observed decreasing failure rate. *Technometrics* **5**, 375–83.

Serfling, R. J. (1980). *Approximation Theorem of Mathematical Statistics*. New York: Wiley.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B **58**, 267–88.

Titterington, D. M., Smith, A. F. M. & Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist.* **20**, 595-601.

Wilks S. S. (1938). The large sample distribution of the likelihood ratio for testing composition hypotheses. *Ann. Math. Statist.* **9**, 60–2.

[*Received ??* 2008. *Revised ??* 2008]