The Canadian Journal of Statistics
Vol. 44, No. 1, 2016, Pages 82–101
La revue canadienne de statistique

Sample-size calculation for tests of homogeneity

Jiahua CHEN[1,2], Pengfei LI[3] and Yukun LIU[4]*

[1] *School of Mathematics and Statistics, Yunnan University, Kunming, Yunnan 650091, China*
[2] *Department of Statistics, University of British Columbia, Vancouver, BC, Canada V6T 1Z2*
[3] *Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, Canada N2L 3G1*
[4] *School of Statistics, East China Normal University, Shanghai 200241, China*

Key words and phrases: Calibration; $C(\alpha)$ test; contiguity theory; EM-test; exponential family; homogeneity; local alternative; locally most powerful test; mixture model.

MSC 2010: Primary 62P99; 62P10; secondary 62F03

Abstract: Mixture models are widely used to explain excessive variation in observations that is not captured by standard parametric models, and they lead to suggestive latent structures. The hypothetical latent structure often needs critical examination based on experimental data. It is therefore important to know the sample size needed to ensure a reasonable chance of success. We investigate this issue for the EM-test and the $C(\alpha)$ test. They are shown to be asymptotically equivalent and have simple limiting distributions under two sets of local alternatives for commonly used mixture models. We obtain a simple sample-size formula and an associated simulation-based calibration procedure, and we demonstrate via data examples and simulation studies that they provide useful guidance for several common mixture models. *The Canadian Journal of Statistics* 44: 82–101; 2016 © 2016 Statistical Society of Canada

Résumé: Les modèles de mélange sont largement utilisés pour expliquer des variations que les modèles paramétriques habituels n'arrivent pas à capturer, et ils suggèrent ainsi une structure latente qui doit souvent faire l'objet d'un examen critique basé sur des données. Il importe donc de connaître la taille d'échantillon nécessaire afin de garantir une probabilité de succès raisonnable pour un tel examen. Les auteurs étudient ce problème pour les tests EM et C(α). Ils montrent que ces deux tests sont asymptotiquement équivalents et qu'ils présentent une distribution limite simple sous des hypothèses locales pour des modèles de mélange fréquemment utilisés. Les auteurs obtiennent une formule simple pour la taille d'échantillon pourvue d'une procédure de calibration basée sur la simulation. Ils démontrent à l'aide d'exemples de données et de simulations que leur approche s'avère utile pour de nombreux modèles de mélange communs. *La revue canadienne de statistique* 44: 82–101; 2016 © 2016 Société statistique du Canada

1. INTRODUCTION

Mixture models are widely used to explain excessive variation in observations that is not captured by parametric models, and they therefore reveal potential latent structures. In genetic applications, for instance, the presence of a latent structure is indicative of some disease-causing genes. Detecting the presence of a latent structure, usually through a homogeneity test, is often the first step in data analysis. A scientific claim, such as a latent structure, must be critically examined based on experimental data. It is therefore important to know the sample size needed to ensure a reasonable chance of success. There is an abundant literature on sample-size calculation for the two-sample test and for case-control studies. Sample-size calculation for the homogeneity test has not been investigated, and it is the focus of this paper. Clearly sample-size calculation is

* *Author to whom correspondence may be addressed.*
E-mail: ykliu@sfs.ecnu.edu.cn

test-specific. The less efficient a method, the larger the sample required to achieve a given power. Hence our investigation starts by screening homogeneity test methods.

The likelihood ratio test for homogeneity has received the most attention in the literature. The results in Hartigan (1985), Dacunha-Castelle & Gassiat (1999), Chen & Chen (2001), and Liu & Shao (2003) are insightful but arguably hard to implement in applications. The paper of Chernoff & Lander (1995) on binomial mixtures lacks generality. We therefore do not consider the likelihood ratio test for sample-size calculation.

The $C(\alpha)$ test of Neyman & Scott (1966) is convenient for homogeneity testing. The method has a simple limiting distribution and possesses local optimality within this class. In addition some over-dispersion testing methods coincide with this $C(\alpha)$ test. The modified likelihood ratio test proposed in Chen (1998) and Chen, Chen, & Kalbfleisch (2001) is another potential candidate. However it has been overshadowed by the EM-test developed more recently by Chen & Li (2009) and Li, Chen, & Marriott (2009). Although the modified likelihood ratio test precedes the EM-test and provides important insight for its development, the latter has several advantages: it is more generic and valid under weaker conditions on the mixture models. In addition the issue of the tuning parameter selection has been addressed by Chen & Li (2011) via computer experiments. In summary we take the $C(\alpha)$ test and the EM-test as the basis for the sample-size calculation.

If the data are generated from an alternative model, any sensible test will reject the null hypothesis with probability one as the sample size increases to infinity. At what intermediate sample size will the power attain a specific target? Extensive computer simulations may be able to provide an answer, but an answer based on one set of simulations may be applicable to only one setting. A more general approach is to determine a formula for the sample size and to support it via a follow-up simulation study. In this paper we provide a simple sample-size formula based on the concept of power calculation under local alternative models. We find that the EM-test and the $C(\alpha)$ test are asymptotically equivalent under two sets of local alternative models of interest. Given a potential alternative model we insert it into the sequence of local alternative models to determine the sample size needed. Simulation studies indicate that our sample-size formulas lead to tests with powers close to the targets under the commonly used normal, binomial, and Poisson mixtures. In applications, simulation should be used to examine the finite-sample power. If the deviation from the target power is too large, a calibration formula can be used to refine the sample size.

The organization of the paper is as follows. In Section 2 we review the EM-test and the $C(\alpha)$ homogeneity test. In Section 3 we obtain the limiting distributions of these tests for local alternative models under simple conditions. In Section 4 we determine a sample-size formula for commonly used exponential family distributions, together with an explicit form of the $C(\alpha)$ test statistic. We also develop a calibration formula. In Section 5 we apply the formula to several real examples and examine its validity via simulations. In Section 6 we summarize our results and discuss potential future topics. The proofs are given in the Appendix.

2. EM-TEST AND $C(\alpha)$ TEST FOR HOMOGENEITY

A two-component mixture model is defined through its density function

$$f(x; \Psi) = \int f(x; \theta) d\Psi(\theta) = (1-\gamma) f(x; \theta_1) + \gamma f(x; \theta_2), \tag{1}$$

where $f(x; \theta_1)$ and $f(x; \theta_2)$ are kernel densities from some parametric family, $1-\gamma$ and γ are the mixing proportions, and θ_1 and θ_2 are the component parameters. We consider only the case where θ is one-dimensional. We use the mixing distribution $\Psi(\theta) = (1-\gamma) I(\theta_1 \le \theta) + \gamma I(\theta_2 \le \theta)$ with $0 \le \gamma \le 1$ to record the mixture structure. The null hypothesis under the homogeneity test is

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

i.e., the mixture structure degenerates. We consider the sample-size calculation problem for the EM-test and the $C(\alpha)$ test.

2.1. EM-Test

The EM-test is a likelihood-based method. Given a random sample $X_1, X_2, \ldots, X_n$ from (1), the log-likelihood function is given by

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

To partially restore the regularity of the likelihood under mixture models, a penalized log-likelihood function is used to construct the EM-test:

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

for some penalty function $p(\gamma)$ on γ to be specified later. The EM-test for the homogeneity hypothesis (Li, Chen, & Marriott, 2009) is best described via its calculation procedure:

Step 1. Select a number of initial values $\gamma_1, \ldots, \gamma_J$, and a number of iterations K.

Step 2. For $j = 1, \ldots, J$, do the following:

Step 2.1. Let $k = 1$ and $\gamma_j^{(k)} = \gamma_j$, and compute $(\theta_{j1}^{(k)}, \theta_{j2}^{(k)}) = \arg\max_{\theta_1, \theta_2} pl_n(\gamma_j^{(k)}, \theta_1, \theta_2)$.

Step 2.2. For $i = 1, \ldots, n$, compute the conditional expectations:

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

Then compute

$$\begin{aligned}
\gamma_j^{(k+1)} &= \arg\max_{\gamma} \left\{ \left(n - \sum_{i=1}^{n} w_{ij}^{(k)}\right) \log(1-\gamma) \right. \\
&\qquad \left. + \sum_{i=1}^{n} w_{ij}^{(k)} \log(\gamma) + p(\gamma) \right\}, \\
\theta_{j1}^{(k+1)} &= \arg\max_{\theta_1} \left\{ \sum_{i=1}^{n} (1 - w_{ij}^{(k)}) \log f(X_i; \theta_1) \right\}, \\
\theta_{j2}^{(k+1)} &= \arg\max_{\theta_2} \left\{ \sum_{i=1}^{n} w_{ij}^{(k)} \log f(X_i; \theta_2) \right\}.
\end{aligned}$$

Step 2.3. Let $k = k + 1$. Repeat Step 2.2 until $k > K$.

Step 3. Calculate

$$M_n^{(K)}(\gamma_j) = 2\{pl_n(\gamma_j^{(K)}, \theta_{j1}^{(K)}, \theta_{j2}^{(K)}) - pl_n(0.5, \hat{\theta}, \hat{\theta})\},$$

with $\hat{\theta} = \arg\max_\theta pl_n(0.5, \theta, \theta)$, and export the value of the EM-test statistic as

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

The limiting distribution of $\mathrm{EM}_n^{(K)}$ under H_0 is $0.5\chi_0^2 + 0.5\chi_1^2$ under some regularity conditions that will be given in the Appendix. Here $0.5\chi_0^2 + 0.5\chi_1^2$ denotes an equal mixture of a distribution with point mass at zero and a χ_1^2 distribution. Given a significance level $\alpha < 0.5$, the EM-test rejects H_0 when $\mathrm{EM}_n^{(K)} > z_\alpha^2$ where z_α is the $(1-\alpha)$th quantile of the standard normal distribution. Li, Chen, & Marriott (2009) further recommend $K = 3$, $\{\gamma_1, \ldots, \gamma_J\} = \{0.1, 0.3, 0.5\}$, and $p(\gamma) = C \log(1 - |1 - 2\gamma|)$. Based on computer experiments, Chen & Li (2011) recommend $C = 0.54$ for normal, Poisson, and binomial kernels and

$$C = \frac{\exp(0.74 + 82n^{-1})}{1 + \exp(0.74 + 82n^{-1})},$$

for the exponential kernel to achieve accurate sizes for the EM-test.

2.2. $C(\alpha)$ Test

The general $C(\alpha)$ test is designed to test a specific null value of a parameter of interest in the presence of nuisance parameters. In addition the test statistic is derived from a class of zero-mean functions with some regularity properties. Within this class, the function based on the projected score function is optimal: it has the highest asymptotic power against local alternatives. Neyman & Scott (1966) regard the variance of the mixing distribution Ψ as the parameter of interest, and the mean of Ψ as a nuisance parameter, when $C(\alpha)$ is applied to test for homogeneity. Furthermore the resulting statistic is the same for any Ψ under some conditions. Hence the test is also model-robust in some sense.

The $C(\alpha)$ test for homogeneity can be motivated from several angles that will not be covered here. Under the null model $f(x;\theta)$, define

$$Y_i(\theta) = \frac{f'(X_i;\theta)}{f(X_i;\theta)}, \quad Z_i(\theta) = \frac{f''(X_i;\theta)}{2f(X_i;\theta)}. \tag{2}$$

The mean θ of Ψ is a nuisance parameter. It can be argued that $\sum_{i=1}^n Y_i(\theta)$ and $\sum_{i=1}^n Z_i(\theta)$ are score functions for the mean and variance of Ψ, respectively. Projecting $Z_i(\theta)$ into the space of $Y_i(\theta)$ we have the residual $W_i(\theta) = Z_i(\theta) - \beta(\theta)Y_i(\theta)$ with $\beta(\theta) = \mathbb{E}\{Y_1(\theta)Z_1(\theta)\}/\mathbb{E}\{Y_1^2(\theta)\}$. Unless otherwise specified, throughout this paper, $\mathbb{E}$ and $\mathbb{V}\mathrm{ar}$ denote the expectation and variance operators, respectively under $f(x;\theta_0)$, the true or the perceived null distribution. We will clarify the notion of the perceived null distribution later. The projection makes $W_i(\theta)$ uncorrelated with $Y_i(\theta)$. If the true value θ_0 of θ is known, a test statistic would be the standardized $\sum_{i=1}^n W_i(\theta_0)$, in which the standard error also involves θ_0. If θ_0 is replaced by a root-n consistent estimator $\hat\theta$ under the null model, this test statistic is still available. Under mild conditions, the replacement does not change the limiting distribution.

When $\hat\theta$ is the maximum-likelihood estimator under $f(x,\theta)$, the $C(\alpha)$ statistic has a simpler form:

$$T_n = \frac{\sum_{i=1}^n W_i(\hat\theta)}{\sqrt{n\nu(\hat\theta)}} = \frac{\sum_{i=1}^n Z_i(\hat\theta)}{\sqrt{n\nu(\hat\theta)}},$$

with $\nu(\theta) = \mathbb{E}_*\{W_1^2(\theta)\}$, where $\mathbb{E}_*$ indicates that the expectation is taken with respect to the homogeneous $f(x;\theta)$ distribution.

It is found that T_n is asymptotically standard normal under the null distribution. At a given significance level α we reject H_0 when $T_n > z_\alpha$.

3. LOCAL ALTERNATIVE MODELS AND ASYMPTOTIC POWER

The power function of a test for a finite sample size does not have a simple analytical form except in a few special cases. The power function of any sensible test at $n = \infty$ is typically constant 1: the null model will be rejected 100% if the alternative is true and the sample size is infinite. Consequently, power calculations are feasible only via computer simulation, but even extensive simulations have a limited scope. One may instead investigate the local power, the limit of the power function at a sequence of alternative models approaching the null model when $n \to \infty$. We use the term *local alternative* for such an alternative model sequence that approaches a null model as the sample size increases. Local asymptotic results are useful for power comparisons of competing tests or for sample-size calculation.

In the context of the homogeneity test, the local alternative models are mixtures with mixing distributions approaching a point mass. In this paper we focus on two-component mixture models. Thus our local alternatives have the following two forms:

$$H_{A1}^n : \gamma = \gamma_0, \quad \theta_1 = \theta_0 - n^{-1/4}\{\gamma_0/(1-\gamma_0)\}^{1/2}\tau, \quad \theta_2 = \theta_0 + n^{-1/4}\{(1-\gamma_0)/\gamma_0\}^{1/2}\tau$$
$$H_{A2}^n : \gamma = n^{-1/2}\eta, \quad \theta_1 = \theta_0 - n^{-1/4}\{\gamma/(1-\gamma)\}^{1/2}\tau, \quad \theta_2 = \theta_0 + \{(1-\gamma)/\eta\}^{1/2}\tau.$$

Hereafter, unless otherwise stated, γ_0, η, and τ are assumed to be constants not depending on n. As $n \to \infty$, both θ_1 and θ_2 in H_{A1}^n tend to θ_0 at the rate $n^{-1/4}$ while the mixing proportion is stationary. In comparison, in H_{A2}^n the mixing proportion $\gamma \to 0$, but $\theta_1 \to \theta_0$ and $\theta_2 \to \theta_0 + \tau/\sqrt{\eta} \neq \theta_0$ all at the rate $n^{-1/2}$. The dynamics in θ_1 and θ_2 of H_{A2}^n are not intrinsic to the local alternative but embedded to fix the means of the mixing distribution at θ_0. As a consequence, both alternative distribution sequences converge to $f(x; \theta_0)$, which is regarded as the perceived null distribution under which $\mathbb{E}$ and $\mathbb{V}\mathrm{ar}$ are computed.

These two sequences of local alternatives represent two types of loss-of-identifiability. The first corresponds to the case in which the two-component distributions are nearly identical, and the second to the case in which the second component is barely present in the mixture. Chen, Chen, & Kalbfleisch (2001) investigated the limiting distribution of the modified likelihood ratio test under H_{A1}^n. There has been no investigation under H_{A2}^n. Under both local alternative sequences, the mean and variance of Ψ are θ_0 and $\delta^2 = n^{-1/2}\tau^2$, respectively. Neyman & Scott (1966) also choose $\delta = O(n^{-1/4})$ for their local alternatives. Chen (1995) showed that the best possible rate for estimating Ψ is $n^{-1/4}$, which explains why we work on the local asymptotic in the $n^{-1/4}$ neighbourhood.

The next theorem gives the limiting distributions of $\mathrm{EM}_n^{(K)}$ and T_n under H_{A1}^n and H_{A2}^n. As notational preparation we define another Z function,

$$Z_i^*(\theta) = \frac{f(X_i; \theta) - f(X_i; \theta_0) - f'(X_i; \theta_0)(\theta - \theta_0)}{(\theta - \theta_0)^2 f(X_i; \theta_0)}. \tag{3}$$

Taking the continuous limit we have $Z_i^*(\theta_0) = Z_i(\theta_0)$, which was defined earlier in (2). We use $\xrightarrow{d}$ for convergence in distribution and, if there is no possibility of confusion we use Φ for both a standard normal random variable and its cumulative distribution function.

Theorem 1. *Suppose $p(\gamma)$ and $f(x; \theta)$ satisfy the regularity conditions specified in the Appendix. For any finite K we have $EM_n^{(K)} = (T_n^+)^2 + o_p(1)$ under H_{A1}^n. As $n \to \infty$, $T_n \xrightarrow{d} \Phi + \Delta_1$ and hence*

$$EM_n^{(K)} \xrightarrow{d} \{(\Phi + \Delta_1)^+\}^2,$$

where $\Delta_1 = \tau^2 \mathbb{E}\{W_1(\theta_0) Z_1(\theta_0)\} / \sqrt{\mathbb{E}\{W_1^2(\theta_0)\}}$.

Assume further that $\mathbb{E}[\{Z_i^*(\theta)\}^2] < \infty$ *for* θ *in a neighbourhood of* $\theta_0 + \tau/\eta^{1/2}$. *We also have* $EM_n^{(K)} = (T_n^+)^2 + o_p(1)$ *under* H_{A2}^n. *As* $n \to \infty$, $T_n \xrightarrow{d} \Phi + \Delta_2$ *and hence*

$$EM_n^{(K)} \xrightarrow{d} \{(\Phi + \Delta_2)^+\}^2,$$

where $\Delta_2 = \tau^2 \mathbb{E}\{W_1(\theta_0) Z_1^*(\theta_0 + \tau/\eta^{1/2})\}/\sqrt{\mathbb{E}\{W_1^2(\theta_0)\}}$.

The additional finite second moment condition under H_{A2}^n deserves attention. Under exponential mixtures, this condition is not satisfied for some $\theta_0 + \tau/\eta^{1/2}$. The effectiveness of the first-order asymptotic deteriorates when this value gets closer to the region where this condition is not satisfied. The sample-size formula under an exponential mixture thus has poorer precision, as will be seen.

The EM-test is essentially a likelihood ratio test with some modifications. Thus it should possess the same local power as the optimal likelihood ratio test. The $C(\alpha)$ homogeneity test is locally the most powerful within its class (Neyman & Scott, 1966). It is not surprising that these test statistics are closely linked under local alternatives. When the alternative model is some distance from the closest null model, the EM-test is superior, as observed in many simulation studies.

Suppose a potential alternative model (with a fixed δ^2) is specified in an application, and we insert this model into the local sequence of either H_{A1}^n or H_{A2}^n (with $\tau^2 = n^{1/2}\delta^2$ dependent on n). The power of the two tests for detecting this alternative model can then be assessed based on the limiting distribution under the local alternative given any potential n. A larger n leads to a larger Δ_1 or Δ_2. Hence we need a minimum sample size n to attain a target probability of rejection.

In conclusion, the above theorem has already provided a simple recipe for sample-size calculation. Given a target alternative model, one first decides to have it embedded into either H_{A1}^n or H_{A2}^n. Subsequently, the user can obtain the value of either Δ_1 or Δ_2 given n and therefore the asymptotic power via, for $j = 1$ or $j = 2$,

$$P\{(\Phi + \Delta_j) > z_\alpha\}.$$

The sample size needed to achieve the target power is therefore obtained, for instance, by a grid search. In applications, it is desirable to conduct a simulation study to be absolutely sure. If the simulated power deviates from the target more than can be tolerated, a calibration formula can be used to refine the sample-size calculation. The exact formula will be given later.

In many applications, the kernel of the mixture model belongs to a natural exponential family with a quadratic variance function (NEF-QVF; Morris, 1982). For these mixture models we find straightforward analytical sample-size formulas. We present the results in the next section.

4. EXPLICIT FORMULAS UNDER NEF-QVF

The NEF-QVF, first investigated by Morris (1982), covers most commonly used distribution families, such as normal, Poisson, binomial, and exponential. The density function in the one-parameter natural exponential family has a unified analytical form,

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

where $\theta = A'(\phi)$ represents the mean parameter. Let $\sigma^2 = A''(\phi)$ be the variance under $f(x;\theta)$. For a member of the natural exponential family with a quadratic variance function, there must exist constants a, b, and c such that

$$\sigma^2 = A''(\phi) = a\{A'(\phi)\}^2 + bA'(\phi) + c = a\theta^2 + b\theta + c. \tag{4}$$

Thus the variance is a quadratic function of the mean.

The following theorem presents the simple forms of the C(α) test statistic, Δ_1, and Δ_2 when $f(x;\theta)$ is a member of the natural exponential family with a quadratic variance function.

Theorem 2. *If the kernel $f(x;\theta)$ is a member of the natural exponential family with a quadratic variance function, then*

$$T_n = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 - n\hat{\sigma}^2}{\sqrt{2n(a+1)}\hat{\sigma}^2},$$

where $\bar{X} = \sum_{i=1}^n X_i/n$ and $\hat{\sigma}^2 = a\bar{X}^2 + b\bar{X} + c$ with coefficients given by (4) are the maximum-likelihood estimators of θ and σ^2, respectively.

In addition the noncentrality parameters defined in Theorem 1 are given by

$$\Delta_1 = \Delta_2 = \sqrt{0.5(1+a)}\tau^2\sigma_0^{-2},$$

where σ_0^2 is the variance under $f(x;\theta_0)$.

The analytical form of the $C(\alpha)$ test statistics for the normal, Poisson, binomial, and exponential kernels is included in Table 1 for easy reference.

We are now ready to work on the sample-size calculation. Given an alternative model with density function $(1-\gamma)f(x;\theta_1) + \gamma f(x;\theta_2)$, we insert it into the local alternative sequence with $\tau^2 = n^{1/2}\delta^2 = n^{1/2}\gamma(1-\gamma)(\theta_1-\theta_2)^2$. We here fix δ^2 and allow τ^2 to be dependent on n. For n such that τ^2 is of moderate size, the local powers of the EM-test and the $C(\alpha)$ test are approximated by

$$\Phi(\Delta_1 - z_\alpha) = \Phi\left(\sqrt{0.5n(1+a)}\delta^2\sigma_0^{-2} - z_\alpha\right).$$

TABLE 1: T_n and $n_{\alpha,\beta}$ under the normal, Poisson, binomial, and exponential kernels.

Kernel	θ	ϕ	a	σ^2	T_n	$n_{\alpha,\beta}$
$N(\theta,1)$	θ	θ	0	1	$\frac{\sum_{i=1}^n (X_i-\bar{X})^2 - n}{\sqrt{2n}}$	$\frac{2(z_\alpha+z_\beta)^2}{\delta^4}$
Poi(θ)	θ	$\log\theta$	0	θ	$\frac{\sum_{i=1}^n (X_i-\bar{X})^2 - n\bar{X}}{\sqrt{2n}\bar{X}}$	$\frac{2(z_\alpha+z_\beta)^2\theta_0^2}{\delta^4}$
Bin(m,p)	mp	$\log\frac{p}{(1-p)}$	$-\frac{1}{m}$	$\theta - \frac{\theta^2}{m}$	$\frac{\sum_{i=1}^n (X_i-\bar{X})^2 - n\bar{X}(m-\bar{X})/m}{\sqrt{2n(1-1/m)}\bar{X}(m-\bar{X})/m}$	$\frac{2(z_\alpha+z_\beta)^2m^3(1-p_0)^2p_0^2}{(m-1)\delta^4}$
exp(θ)	θ	-1/θ	1	θ^2	$\frac{\sum_{i=1}^n (X_i-\bar{X})^2 - n\bar{X}^2}{\sqrt{4n}\bar{X}^2}$	$\frac{(z_\alpha+z_\beta)^2\theta_0^4}{\delta^4}$

Here $\delta^2 = \gamma(1-\gamma)(\theta_1-\theta_2)^2$.

Therefore if the target power is $1-\beta$ at a significance level α, the required sample size for both tests approximately satisfies

$$\sqrt{0.5n(1+a)}\delta^2\sigma_0^{-2} - z_\alpha = z_\beta.$$

In other words, the minimum sample size is

$$n_{\alpha,\beta} = \left\{ \frac{(z_\alpha + z_\beta)\sigma_0^2}{\delta^2\sqrt{0.5(1+a)}} \right\}^2.$$

To use this formula, one needs to find the value of σ_0^2 from $\theta_0 = (1-\gamma)\theta_1 + \gamma\theta_2$. For easy reference, the explicit expressions for $n_{\alpha,\beta}$ under the normal, Poisson, binomial, and exponential kernels are given in the last column of Table 1.

4.1. Calibration of the Sample-Size Formula

As will be seen in the next section, the sample-size formula works well for the normal, binomial, and Poisson mixtures. For these models, the average simulated power is 78% with standard deviation 4.7%. There are a few cases in which the simulated powers differ from the target by as much as 10%.

It is always good practice to conduct a simulation at the suggested sample size before one fully commits to this size. If the deviation from the target is not acceptable, one may calibrate the sample size as follows, based on a suggestion from the associate editor.

Under the null model and asymptotically, the EM-test rejects the null model when $\Phi^+ > z_\alpha$, where z_α is the upper α quantile of the standard normal. With noncentrality parameter $\Delta(n) = D\sqrt{n}$, the asymptotic power of the test is

$$P\{(\Phi + \Delta(n))^+ > z_\alpha\} = P(\Phi > z_\alpha - \Delta(n)),$$

when $\alpha < 0.5$. If the target power is $1-\beta > \alpha$ we look for $n_{\alpha,\beta}$ such that

$$z_\alpha - \Delta(n_{\alpha,\beta}) = z_{1-\beta}.$$

When the asymptotic does not work precisely, a correction may help. Suppose

$$P(\Phi > z_\alpha - \Delta(n_{\alpha,\beta})) = 1 - \beta' \neq 1 - \beta.$$

If so we must have $\Delta(n_{\alpha,\beta}) = z_{\beta'} + z_\alpha$. Suppose $n_{\rm cal}$ is such that $\Delta(n_{\rm cal}) = z_\beta + z_\alpha$, then we get

$$n_{\rm cal} = n_{\alpha,\beta}\left(\frac{z_\alpha + z_\beta}{z_\alpha + z_{\beta'}}\right)^2. \tag{5}$$

Consider, for example, $\alpha = 0.05$, $\beta' = 0.75$, $\beta = 0.8$, $\delta^2 = 0.155$, $\sigma_0^2 = 1$, and $a = 0$ so that $n = 514$. The calibrated sample size is 590. As will be seen, the power at the calibrated sample size is very accurate.

5. EXAMPLES AND SIMULATION

The sample-size formula is based on local asymptotic results under either H_{A1}^n or H_{A2}^n. In this section we use simulations to examine its accuracy for three examples.

5.1. Example 1: Children's Nutritional Status

Standardized anthropometric scores based on an international reference population are recommended for the evaluation of the nutritional status of children in developing countries. One such score, HE/AGE, is computed as (HEIGHT−M)/SD, where HEIGHT is the child's height, M is the median height, and SD is the standard deviation in the reference population according to the age and sex class of the children.

To detect subclinical malnourishment, Böhning, Schlattmann, & Lindsay (1992) suggest modelling these scores by a two-component normal mixture with equal component variances 1. In applications, the doctor would like to ascertain the pattern of malnourishment to gauge the nutritional status of an individual child. Böhning, Schlattmann, & Lindsay (1992) analyzed the anthropometric measurements of 708 preschool children in northeast Thailand. They suggested that the following two-component normal mixture

$$0.995N(-1.63, 1) + 0.005N(-6.19, 1),$$

is suitable for the HE/AGE scores. The children in the second group should be carefully considered.

Is the evidence strong enough to reject the homogenous model? The answer is no because the P-value based on the EM-test for homogeneity is 1. Note that the nonhomogeneous model fits well into the H^n_{A2} local alternative sequence with $\theta_0 = -1.6528$ and $\delta^2 = 0.1034$. From the formula in Table 1 for the normal kernel, the minimum sample size is $n_{0.05,0.20} = 1,155$ to detect a departure of this magnitude with probability 80%. The current sample size has a power of 62%. If the original researchers had access to this paper, they would be able to make an informed decision on the cost of collecting more data and the scientific significance of establishing the potentially highly valued alternative model.

According to Kim & Lindsay (2015), there typically exists a limited range of confidence levels where the likelihood region has a natural partition into identifiable subsets for any given data set and mixture model. The identifiable region generally expands with the sample size. Therefore the information helps to determine the size of the additional sample needed to confirm the suggested departure.

How accurate is this sample size? We generated 10,000 random samples of size 1,155 from the above normal mixture model, based on which the simulated powers of $\mathrm{EM}_n^{(3)}$ and the $C(\alpha)$ test are 83.6 and 69.2%, respectively. Clearly our sample-size formula works well for the EM-test.

How accurate is this formula more generally? We conducted additional simulations under a number of mixture models. We choose three sets of mixing distributions with mean $\theta_0 = -1.6528$. For the first set we choose γ from $\{0.1, 0.25, 0.4\}$ and δ^2 from $\{0.1034, 1.5 \times 0.1034\}$. In this set, the mixing proportions are distant from zero and the sizes of δ^2 are moderate. Hence they fit well into H^n_{A1}. For the second set we choose γ from $\{0.01, 0.02, 0.03\}$ and δ^2 from $\{0.1034, 1.5 \times 0.1034\}$. For this set of alternatives, the γ values are close to 0, and they fit well into H^n_{A2}. For the third set, the γ value is chosen from $\{0.1, 0.25, 0.4\}$ and $\delta^2 = 3 \times 0.1034$. Both the mixing proportions are distant from zero, and the value of δ^2 is large for this set of alternatives. Hence neither H^n_{A1} nor H^n_{A2} fit well. They may be loosely referred to as "nonlocal."

There are 15 alternative models in the three sets. From each we generated 10,000 random samples of size $n_{0.05,0.2}$ according to the sample-size formula for the normal mixture from Table 1. The model parameters, the sample sizes, and the simulated powers of the EM-test and the $C(\alpha)$ test are given in Table 2.

The simulation results indicate that the sample-size formula is slightly liberal but reasonable for both tests when the data are generated from models that fit H^n_{A1} well. The formula from Table 1 provides a good starting point. When the data are generated from models that fit H^n_{A2} well, the sample-size formula is conservative for the EM-test but slightly liberal for the $C(\alpha)$

TABLE 2: Simulated powers of the EM-test and $C(\alpha)$ test with the calculated sample size $n_{0.05,0.20}$ under 15 normal mixture models.

Alternative model	$n_{0.05,0.2}$	$\mathrm{EM}_n^{(1)}$	$\mathrm{EM}_n^{(2)}$	$\mathrm{EM}_n^{(3)}$	$C(\alpha)$ test
$0.90N(-1.521, 1) + 0.10N(-2.835, 1)$	514	0.765	0.765	0.765	0.749
$0.75N(-1.425, 1) + 0.25N(-2.335, 1)$	514	0.752	0.752	0.752	0.755
$0.60N(-1.331, 1) + 0.40N(-2.135, 1)$	514	0.745	0.745	0.745	0.750
$0.90N(-1.546, 1) + 0.10N(-2.618, 1)$	1,155	0.781	0.781	0.781	0.770
$0.75N(-1.467, 1) + 0.25N(-2.210, 1)$	1,155	0.768	0.768	0.768	0.769
$0.60N(-1.390, 1) + 0.40N(-2.047, 1)$	1,155	0.763	0.763	0.763	0.763
$0.99N(-1.630, 1) + 0.01N(-5.572, 1)$	514	0.805	0.807	0.810	0.682
$0.98N(-1.597, 1) + 0.02N(-4.410, 1)$	514	0.824	0.826	0.827	0.717
$0.97N(-1.584, 1) + 0.03N(-3.893, 1)$	514	0.806	0.808	0.809	0.731
$0.99N(-1.620, 1) + 0.01N(-4.853, 1)$	1,155	0.855	0.856	0.857	0.724
$0.98N(-1.607, 1) + 0.02N(-3.904, 1)$	1,155	0.834	0.835	0.836	0.748
$0.97N(-1.596, 1) + 0.03N(-3.482, 1)$	1,155	0.823	0.823	0.824	0.758
$0.90N(-1.467, 1) + 0.10N(-3.324, 1)$	128	0.721	0.722	0.722	0.695
$0.75N(-1.331, 1) + 0.25N(-2.618, 1)$	128	0.716	0.716	0.716	0.719
$0.60N(-1.198, 1) + 0.40N(-2.335, 1)$	128	0.707	0.707	0.708	0.716

test. The simulation results show that the EM-test is generally more efficient in this situation. When the data are generated from models that do not fit into either local alternative setting, the sample-size formula is an underestimate. In this case the above confirmation simulation can be used with the calibration formula to obtain a more accurate sample size. We will report the performance of the calibration formula shortly.

5.2. Example 2: Sex Ratios in German Families

Geissler (1889) compiled a vast record of sex ratios for German families. Sokal & Rohlf (1973) analyzed a portion of the data, consisting of the number of males among the first 12 children in sibships of size 13 for 6,115 families from Saxony, Germany, 1876–1885. The 13th child was ignored in an effort to discount the effects of stopping rules. One may consider sex determination to be a sequence of Bernoulli trials with some constant probability p of having a male child. If so, the number of male children in a family with 12 children would be distributed as $\mathrm{Bin}(12, p)$. However most analyses have found that a single $\mathrm{Bin}(12, p)$ is not suitable for the data because there is a substantial overdispersion. There are a number of possible explanations, and the most plausible is that the value of p varies among families. Hence a binomial mixture is a good choice for modelling this data set. Lindsay & Roeder (1992) and Lindsay (1995) applied geometric diagnostic methods to determine the appropriate order of the mixture model. They found that a binomial mixture of order two gave an adequate fit to the data.

We obtained the maximum-likelihood estimate based on a two-component binomial mixture; the fitted model is given by

$$0.72\mathrm{Bin}(12, 0.48) + 0.28\mathrm{Bin}(12, 0.62).$$

TABLE 3: Simulated powers of the EM-test and $C(\alpha)$ test with the calculated sample size $n_{0.05,0.20}$ under 13 binomial mixture models.

Alternative model	$n_{0.05,0.2}$	$\text{EM}_n^{(1)}$	$\text{EM}_n^{(2)}$	$\text{EM}_n^{(3)}$	$C(\alpha)$ test
0.90Bin(12, 0.498) + 0.10Bin(12, 0.708)	374	0.777	0.777	0.777	0.751
0.75Bin(12, 0.483) + 0.25Bin(12, 0.628)	374	0.755	0.755	0.755	0.754
0.60Bin(12, 0.468) + 0.40Bin(12, 0.596)	374	0.744	0.744	0.744	0.748
0.90Bin(12, 0.502) + 0.10Bin(12, 0.673)	841	0.774	0.774	0.774	0.760
0.75Bin(12, 0.490) + 0.25Bin(12, 0.608)	841	0.768	0.768	0.768	0.769
0.60Bin(12, 0.477) + 0.40Bin(12, 0.582)	841	0.760	0.760	0.760	0.762
0.98Bin(12, 0.510) + 0.02Bin(12, 0.959)	374	0.862	0.865	0.868	0.722
0.97Bin(12, 0.508) + 0.03Bin(12, 0.877)	374	0.841	0.843	0.845	0.728
0.98Bin(12, 0.512) + 0.02Bin(12, 0.878)	841	0.865	0.867	0.868	0.755
0.97Bin(12, 0.510) + 0.03Bin(12, 0.811)	841	0.837	0.838	0.838	0.758
0.90Bin(12, 0.494) + 0.10Bin(12, 0.750)	166	0.764	0.764	0.765	0.726
0.75Bin(12, 0.475) + 0.25Bin(12, 0.653)	166	0.735	0.735	0.735	0.731
0.60Bin(12, 0.456) + 0.40Bin(12, 0.613)	166	0.727	0.727	0.728	0.733

Note that the component parameters are close to each other, which fits well into the local alternative sequence H_{A1}^n. Suppose our interest is to detect the above heterogeneous model with 80% power at the 5% level. With $(\gamma, p_1, p_2) = (0.28, 0.48, 0.62)$ and $m = 12$, the formula in Table 1 shows that the required sample size for the EM-test or $C(\alpha)$ test is $n_{0.05,0.20} \approx 374$. Is $n = 374$ large enough? We generated 10,000 data sets of this size from the above binomial mixture model, and the rate of rejection of the homogeneous model was 74.4% for $\text{EM}_n^{(3)}$ and 75.4% for the $C(\alpha)$ test. Both rates are slightly lower than but close to the 80% target. We conclude that the sample-size formula is reasonably accurate.

We also conducted a simulation study under three sets of mixing models. The imaginary null model is binomial with $m = 12$ and $p_0 = 0.72 \times 0.48 + 0.28 \times 0.62 = 0.5192$. We fix $m = 12$ and $\gamma p_1 + (1-\gamma)p_2 = 0.5192$. In the first set we select three γ values, 0.1, 0.25, and 0.4, and two $\gamma(1-\gamma)(p_1 - p_2)^2$ values from $\{0.00395, 0.00395/1.5\}$ with $0.00395 = 0.72 \times 0.28 \times (0.48 - 0.62)^2$. In the second set we select two γ values 0.02 and 0.03, and two $\gamma(1-\gamma)(p_1 - p_2)^2$ values from $\{0.00395, 0.00395/1.5\}$. In the third set, γ takes values 0.1, 0.25, and 0.4, and $\gamma(1-\gamma)(p_1 - p_2)^2 = 1.5 \times 0.00395$. In total we obtain 13 alternative models and the corresponding $n_{0.05,0.2}$ values. The simulated powers of the EM-test and the $C(\alpha)$ test based on 10,000 repetitions together with other details are given in Table 3.

The simulation results in Table 3 are similar to those for the normal mixture models. For the H_{A1}^n alternatives, the sample-size formula is liberal yet reasonable; for the H_{A2}^n alternatives, it is mildly conservative for the EM-test but liberal for the $C(\alpha)$ test; for less local alternatives, it is rather liberal.

Simulations were also done on Poisson mixture models, and the results are similar. They are therefore omitted. The performance of the calibration formula will be reported shortly.

5.3. Example 3: Failure Times of Air Conditioning Systems

We now consider the data studied in Proschan (1963). They consist of the times of successive failures for the air conditioning system of each plane in a fleet of 13 Boeing 720 jet aircraft. After

TABLE 4: Simulated powers of the EM-test and $C(\alpha)$ test with the calculated sample size $n_{0.05,0.20}$ under 12 exponential mixture models.

Alternative model	$n_{0.05,0.2}$	$\text{EM}_n^{(1)}$	$\text{EM}_n^{(2)}$	$\text{EM}_n^{(3)}$	$C(\alpha)$ test
0.80exp(107.444) + 0.20exp(35.854)	692	0.834	0.834	0.834	0.726
0.65exp(114.139) + 0.35exp(54.102)	692	0.758	0.758	0.758	0.718
0.50exp(121.762) + 0.50exp(64.490)	692	0.722	0.722	0.723	0.709
0.80exp(103.250) + 0.20exp(52.629)	2,766	0.800	0.801	0.801	0.766
0.65exp(107.984) + 0.35exp(65.532)	2,766	0.758	0.758	0.758	0.748
0.50exp(113.375) + 0.50exp(72.877)	2,766	0.745	0.745	0.745	0.748
0.80exp(113.369) + 0.20exp(12.126)	173	0.964	0.965	0.966	0.687
0.65exp(122.843) + 0.35exp(37.938)	173	0.765	0.766	0.767	0.660
0.50exp(133.623) + 0.50exp(52.629)	173	0.676	0.677	0.678	0.639
0.95exp(97.771) + 0.05exp(4.864)	2,766	0.991	0.992	0.993	0.777
0.95exp(97.281) + 0.05exp(14.182)	4,322	0.933	0.933	0.934	0.776
0.95exp(96.919) + 0.05exp(21.061)	6,244	0.900	0.901	0.901	0.786

careful analysis, Proschan (1963) concluded that the failure time distribution for each aircraft was exponential, but the rate varied. A mixture of exponential distributions is therefore a reasonable model for the data. Li, Chen, & Marriott (2009) applied the EM-test for homogeneity to the aircraft data and found significant evidence for heterogeneity. The evidence for the order of the mixture being 3 or higher based on the EM-test (Li & Chen, 2010) is insignificant with a P-value of 0.6.

The fitted two-component exponential mixture model via the maximum-likelihood method is given by

$$0.43\exp(46.5) + 0.57\exp(128.3).$$

Suppose our interest is to detect the above heterogeneous model with 80% power at the 5% level. Setting $(\gamma, \theta_1, \theta_2) = (0.57, 128.3, 46.5)$ and applying the sample-size formula in Table 1, we get the required sample size for the EM-test and the $C(\alpha)$ test: $n_{0.05,0.20} \approx 173$. A subsequent simulation study in the spirit of Examples 1 and 2 found that the powers of $\text{EM}_n^{(3)}$ and $C(\alpha)$ are 71.3 and 64.8%. This unsatisfactory outcome should clearly be investigated more thoroughly.

We start with more simulations. We fix $(1 - \gamma)\theta_1 + \gamma\theta_2 = 0.43 \times 46.5 + 0.57 \times 128.3$ and construct three sets of alternative models similarly. In the first three sets, γ takes values 0.2, 0.35, and 0.5. We then let $\delta^2 = \gamma(1 - \gamma)(\theta_1 - \theta_2)^2$ take values in $\{1/2, 1/4, 1\} \times 1{,}640$, where the last number is from $1{,}640 = 0.43 \times 0.57 \times (46.5 - 128.3)^2$. In the fourth and the last set, $\gamma = 0.05$ and δ^2 takes three values in $\{1/4, 1/5, 1/6\} \times 1{,}640$. For each alternative model we generated 10,000 random samples of size $n_{0.05,0.2}$ from the model and computed the simulated powers of the EM-test and the $C(\alpha)$ test. The results are in Table 4.

We find that the performance of the sample-size formula under the exponential kernel is quite different. The formula is satisfactory for the first six models but poor for the rest. Consider the extra condition in Theorem 1 under H_{A2}^n. This condition is irrelevant under normal, binomial, and Poisson mixtures. However under exponential mixtures, it translates into $-2(1/\theta - 1/\theta_0) < 1/\theta_0$ or $\theta < 2\theta_0$ for all component parameter values. None of the models in Table 4 violate this condition. However a small $2\theta_0 - \theta$ value leads to a slower rate of the $o_p(1)$ term in (A2) for Λ_{n2} and therefore

TABLE 5: Simulated powers of the EM-test at calibrated sample sizes.

Model	n	Power	n_{cal}	Power$_{\text{cal}}$
$0.90N(-1.521, 1) + 0.10N(-2.835, 1)$	514	0.765	565	0.805
$0.75N(-1.425, 1) + 0.25N(-2.335, 1)$	514	0.752	586	0.804
$0.60N(-1.331, 1) + 0.40N(-2.135, 1)$	514	0.745	599	0.807
$0.90N(-1.546, 1) + 0.10N(-2.618, 1)$	1,155	0.781	1,220	0.794
$0.75N(-1.467, 1) + 0.25N(-2.210, 1)$	1,155	0.768	1,263	0.796
$0.60N(-1.390, 1) + 0.40N(-2.047, 1)$	1,155	0.763	1,278	0.806
$0.99N(-1.630, 1) + 0.01N(-5.572, 1)$	514	0.805	508	0.796
$0.98N(-1.597, 1) + 0.02N(-4.410, 1)$	514	0.824	475	0.800
$0.97N(-1.584, 1) + 0.03N(-3.893, 1)$	514	0.806	500	0.800
$0.99N(-1.620, 1) + 0.01N(-4.853, 1)$	1,155	0.855	971	0.802
$0.98N(-1.607, 1) + 0.02N(-3.904, 1)$	1,155	0.834	1,039	0.806
$0.97N(-1.596, 1) + 0.03N(-3.482, 1)$	1,155	0.823	1,076	0.790
$0.90N(-1.467, 1) + 0.10N(-3.324, 1)$	128	0.721	159	0.798
$0.75N(-1.331, 1) + 0.25N(-2.618, 1)$	128	0.716	161	0.788
$0.60N(-1.198, 1) + 0.40N(-2.335, 1)$	128	0.707	165	0.792

less effective first-order asymptotics. Interestingly, the exponential mixture has always been an aberration, as pointed out in Li, Chen, & Marriott (2009).

A new theory is the best solution for exponential mixtures with unbalanced mixing proportions. Fortunately the calibration formula seems to give a satisfactory solution.

5.4. Calibration

We maintain that the sample-size formula has satisfactory precision for normal, binomial, and Poisson mixtures. In many applications, there are many factors influencing the eventual outcome. A deviation of around 5% is likely tolerable, but it never hurts to simulate the power at the sample size suggested by a formula. We always perform confirmation simulations in our consulting projects regardless of the reliability of a sample-size formula.

If the deviation from the target power is unacceptable, the calibration formula given in (5) may be used. In this section we examine the effectiveness of the calibration. Table 5 contains information on the calibrated sample sizes and resulting simulated power based on the results in Table 2. In the table the sample size and power of the EM-test ($\text{EM}_n^{(3)}$) are denoted by n and power; the calibrated sample size and power are denoted by n_{cal} and power$_{\text{cal}}$.

Clearly after the calibration, the simulated power of the EM-test is extremely close to the target 80%. The average simulated power in the table is 79.9% with standard deviation 0.6%.

How accurate are the sample-size formula and its calibration formula when the target power is 90 or 95%? We avoid presenting an excessive number of tables. When the target is 90%, without calibration the average simulated power is 87.9% with standard error 3.5%. After calibration, these numbers are 90.0 and 0.4%. When the target is 95%, these two numbers are 93.4 and 2.4% before calibration and 94.9 and 0.4% after calibration. Clearly the formulas have satisfactory precision both before and after calibration.

We do not give detailed results for binomial models here. After calibration, the average simulated powers and corresponding standard deviations are 80.0 and 0.6%, 90.2 and 0.4%, and 94.9 and 0.2% for the targeted powers 80, 90, and 95%, respectively.

Finally does calibration work for the unruly exponential mixture? After calibration, the average power becomes 80.3% with standard deviation 1.9% when the target power is 80%. However the power for 0.95exp(97.77) + 0.05exp(4.86) is still as high as 86.1%. If we calibrate this size once more, the power becomes 80.7%, and the overall average becomes 80.0% with standard deviation 0.7%. Clearly although the original sample-size formula is not satisfactory, calibration works well even for exponential mixtures. When the target power is 90%, the average simulated power is 90.5% with standard deviation 1.3%. When the target power is 95%, the average simulated power after calibration is 95.0% with standard deviation 1.2%. In both cases, the model 0.95exp(97.77) + 0.05exp(4.86) contributes large shares of the standard deviations. If it is calibrated twice, the standard deviations would be 0.6% in both cases.

6. DISCUSSION AND FUTURE TOPICS

We successfully utilized Le Cam's contiguity theory to determine the limiting distributions of the EM-test and $C(\alpha)$ statistics under local alternative models. The result enabled us to derive a sample-size formula for the homogeneity test. The formula has respectable precision in achieving the target powers. Furthermore it can be refined via a calibration formula. Simulation shows that the power at the calibrated sample size is very close to the target in all the mixture models investigated.

The need for a sample-size formula does not stop at homogeneity tests with a one-dimensional mixing parameter. The sample-size problems related to testing for the order of a mixture model and to multidimensional mixing parameters are important and challenging. We plan to make meaningful progress in that direction in the future.

APPENDIX

The proofs are based on the following regularity conditions on the kernel density function and the penalty function $p(\gamma)$. In our proofs we work on the asymptotic properties of the likelihood ratio under local alternatives. This is a simpler task than proving a general result for the EM-test. Therefore many of the following conditions may appear unnecessary. However if these conditions are violated, the general conclusions for the EM-test are endangered.

Assumption 1 *The penalty function $p(\gamma)$ is a continuous function such that it is maximized at $\gamma = 0.5$ and goes to negative infinity as γ goes to 0 or 1. Further, $p(0.5) = 0$.*

Assumption 2 (Wald's integrability conditions) *(i) $\mathbb{E}|\log f(X;\theta_0)| < \infty$. (ii) For sufficiently small ρ and for sufficiently large r, the expected values $\mathbb{E}\log\{1 + f(X;\theta,\rho)\} < \infty$ for $\theta \in \Theta$ and $\mathbb{E}\log\{1 + \varphi(X, r)\} < \infty$, where $f(x;\theta,\rho) = \sup_{|\theta'-\theta|\le\rho} f(x;\theta')$ and $\varphi(x;r) = \sup_{|\theta|\ge r} f(x;\theta)$. (iii) $\lim_{|\theta|\to\infty} f(x;\theta) = 0$ for all x except on a set with probability zero.*

Assumption 3 (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 4 (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 5 (Uniform boundedness) *Let*

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

For some neighbourhood of θ_0, there exists a g with finite expectation such that $|Y_i^*(\theta)|^3 \leq g(X_i)$, $|Z_i^*(\theta)|^3 \leq g(X_i)$, and $|Z_i^{*''}(\theta)|^2 \leq g(X_i)$, where $Z_i^*(\theta)$ is defined in (3).

Assumption 6 (Positive definite) *The covariance matrix of $\{Y_i^*(\theta_0), Z_i^*(\theta_0)\}$ is positive definite.*

We finally require that for the EM-test the initial values of the mixing probability include $\gamma = 0.5$. Together with Assumption 1, this requirement allows the limiting distribution of the EM-test for homogeneity to have the simplest expression.

We begin with two results from Le Cam's contiguity theory (Le Cam, 1953), which are the basis of our proof of Theorem 1.

Lemma 1. *Let $X_1, X_2, \ldots, X_n$ be a random sample and define $\Lambda_n = \sum_{i=1}^n \log\{f_n(X_i)/f_0(X_i)\}$, where $f_0(x)$, $f_1(x)$, $f_2(x)$, ... is a series of density functions with respect to some sigma finite measure. If $\Lambda_n \xrightarrow{d} N(-h^2/2, h^2)$ for some $h^2 > 0$ under $f_0(x)$, where $\xrightarrow{d}$ denotes convergence in distribution, then for a statistic $T_n \equiv T_n(X_1, X_2, \ldots, X_n)$, $T_n = o_p(1)$ under $f_0(x)$ implies that $T_n = o_p(1)$ under $f_n(x)$.*

Remark: By "under $f_n(x)$," we mean that $X_1, \ldots, X_n$ is a random sample from $f_n(x)$.

Lemma 2. *Assume the same conditions as in Lemma 1. For a statistic $S_n \equiv S_n(X_1, X_2, \ldots, X_n)$, if*

$$\begin{pmatrix} S_n \\ \Lambda_n \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} \mu \\ -h_{22}/2 \end{pmatrix}, \begin{pmatrix} h_{11} & h_{12} \\ h_{12} & h_{22} \end{pmatrix} \right),$$

under $f_0(x)$, then $S_n \xrightarrow{d} N(\mu + h_{12}, h_{11})$ under $f_n(x)$.

Lemma 1 is an application of Example 6.5 of van der Vaart (2000) (p. 89) and Le Cam's first lemma on contiguity theory (van der Vaart, 2000; p. 90). This result facilitates quadratic approximations of the EM-test statistics under local alternative models. Lemma 2 is an application of Le Cam's third lemma on contiguity theory (van der Vaart, 2000; p. 90), based on which we derive the asymptotic distribution of the EM-test statistics under local alternative models.

Proof of Theorem 1. We now prove Theorem 1 under the local alternative H_{A1}^n. Let

$$\Lambda_{n1} = \sum_{i=1}^n \log \frac{(1-\gamma_0)f(X_i;\theta_0 - n^{-1/4}\tau_{11}) + \gamma_0 f(X_i;\theta_0 + n^{-1/4}\tau_{12})}{f(X_i;\theta_0)},$$

with $\tau_{11} = \{\gamma_0/(1-\gamma_0)\}^{1/2}\tau$ and $\tau_{12} = \{(1-\gamma_0)/\gamma_0\}^{1/2}\tau$. Under the regularity conditions on $f(x;\theta_0)$ we easily find the quadratic approximation

$$\Lambda_{n1} = \tau^2 n^{-1/2}\sum_{i=1}^n Z_i - 0.5\tau^4 n^{-1}\sum_{i=1}^n Z_i^2 + o_p(1),$$

under $f(x;\theta)$. Therefore $\Lambda_{n1} \xrightarrow{d} N(-h^2/2, h^2)$ where $h^2 = \tau^4 \mathbb{E}\{Z_1^2(\theta_0)\}$. This result lays the foundation for applying Lemmas 1 and 2 under the local alternative H_{A1}^n.

Let $V_n(\theta) = \sum_{i=1}^n W_i(\theta)/\sqrt{n\nu(\theta)}$ and $V_n = V_n(\theta_0)$, where $\nu(\theta) = \mathbb{E}_*\{W_1^2(\theta)\}$ with the expectation taken under $f(x;\theta)$. It can be seen that under $f(x;\theta_0)$ we have

$$\begin{pmatrix} V_n \\ \Lambda_{n1} \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ -h^2/2 \end{pmatrix}, \begin{pmatrix} 1 & \Delta_1 \\ \Delta_1 & h^2 \end{pmatrix} \right),$$

where Δ_1 is the noncentrality parameter defined in Theorem 1. By Lemma 2 we find $V_n \xrightarrow{d} N(\Delta_1, 1)$ under H_{A1}^n.

According to Li, Chen, & Marriot (2009), the EM-test statistic has the following expansion under $f(x;\theta_0)$:

$$\mathrm{EM}_n^{(K)} = \frac{[\{\sum_{i=1}^n W_i(\theta_0)\}^+]^2}{n\mathbb{E}\{W_1^2(\theta_0)\}} + o_p(1) = (V_n^+)^2 + o_p(1).$$

From Lemma 1, this approximation or the order assessment of the remainder is also valid under the alternative model H_{A1}^n. Therefore the distribution of $\mathrm{EM}_n^{(K)}$ is that of $(V_n^+)^2$ under H_{A1}^n. This result suffices for the conclusion on the limiting distribution of the EM-test.

The conclusion for the $C(\alpha)$ test is proved using Theorem 1 of Neyman (1959). We first highlight this result explicitly. Note that $T_n = V_n(\hat{\theta})$. Neyman (1959) proved that $V_n(\hat{\theta}) - V_n(\theta_0) = o_p(1)$ when $\sum_{i=1}^n W_i(\theta)$ is uncorrelated with $\sum_{i=1}^n Y_i(\theta)$, the score function for the nuisance parameter θ, and $\hat{\theta}$ is root-n consistent with θ_0. The lack of correlation is true by the definition of $W_i(\theta)$, and the root-n consistency is due to the choice of MLE and the regularity of the kernel distribution. Therefore the $C(\alpha)$ test statistic $T_n = V_n + o_p(1)$, and it has the claimed limiting distribution.

Under the alternative model H_{A2}^n, the log likelihood ratio becomes

$$\Lambda_{n2} = \sum_{i=1}^n \log \frac{(1 - n^{-1/2}\eta) f(X_i;\theta_{1n}) + n^{-1/2}\eta f(X_i;\theta_{2n})}{f(X_i;\theta_0)},$$

with

$$\theta_{1n} = \theta_0 - n^{-1/2}\tau\left(\frac{\eta}{1 - n^{-1/2}\eta}\right)^{1/2}, \quad \theta_{2n} = \theta_0 + \tau\left(\frac{1 - n^{-1/2}\eta}{\eta}\right)^{1/2}.$$

When $n \to \infty$ we have $\theta_{1n} \to \theta_0$ and $\theta_{2n} \to \theta_0 + \tau\eta^{-1/2}$ at the rate $n^{-1/2}$. The mixing proportions also approach 1 and 0 at this rate. Using the regularity conditions on $f(x;\theta)$ we can then easily show that

$$\Lambda_{n2} = n^{-1/2}\tau^2 \sum_{i=1}^n Z_i^*(\theta_2) - (1/2)n^{-1}\tau^4 \sum_{i=1}^n \left\{Z_i^*(\theta_2)\right\}^2 + o_p(1). \qquad \text{(A2)}$$

Therefore under $f(x;\theta_0)$,

$$\Lambda_{n2} \xrightarrow{d} N(-\tilde{h}^2/2, \tilde{h}^2),$$

with $\tilde{h}^2 = \tau^4 \mathbb{E}[\{Z_1^*(\theta_0 + \tau\eta^{-1/2})\}^2]$. This establishes contiguity and implies that under the alternative model H_{A2}^n,

$$\mathrm{EM}_n^{(K)} = (V_n^+)^2 + o_p(1).$$

Note that under $f(x; \theta_0)$,

$$\begin{pmatrix} V_n \\ \Lambda_{n2} \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ -\tilde{h}^2/2 \end{pmatrix}, \begin{pmatrix} 1 & \Delta_2 \\ \Delta_2 & \tilde{h}^2 \end{pmatrix} \right).$$

Therefore under H_{A2}^n, $V_n \xrightarrow{d} N(\Delta_2, 1)$. Hence under H_{A2}^n,

$$\mathrm{EM}_n^{(K)} = (V_n^+)^2 + o_p(1) \xrightarrow{d} \{(\Phi + \Delta_2)^+\}^2.$$

The proof of the result for the $C(\alpha)$ test is identical to that under H_{A1}^n. ■

Proof of Theorem 2. Form of $C(\alpha)$ test statistic. The quadratic variance function under the natural exponential family is characterized by its density function $f(x; \theta) = h(x)\exp\{x\phi - A(\phi)\}$ and $A''(\phi) = a\{A'(\phi)\}^2 + bA'(\phi) + c$ for some constants a, b, and c. Recall also that the mean and variance are given by $\theta = A'(\phi)$ and $\sigma^2 = A''(\phi)$. More moment relationships can easily be obtained. Taking derivatives with respect to ϕ on the quadratic relationship we find that

$$\begin{aligned} A'''(\phi) &= \{2aA'(\phi) + b\}A''(\phi) = (2a\theta + b)\sigma^2, \\ A^{(4)}(\phi) &= 2a\{A''(\phi)\}^2 + \{2aA'(\phi) + b\}A'''(\phi) = 2a\sigma^4 + (2a\theta + b)^2\sigma^2. \end{aligned}$$

Because of the regularity of the exponential family we have

$$\mathbb{E}\left\{ \frac{d^k f(X; \theta_0)/d\phi^k}{f(X; \theta_0)} \right\} = 0,$$

for $k = 1,\ 2,\ 3,\ 4$. This implies

$$\begin{aligned} \mathbb{E}\{(X - \theta_0)^3\} &= A'''(\phi_0) = (2a\theta_0 + b)\sigma_0^2, \\ \mathbb{E}\{(X - \theta_0)^4\} &= 3\{A''(\phi_0)\}^2 + A^{(4)}(\phi_0) = (2a + 3)\sigma_0^4 + (2a\theta_0 + b)^2\sigma_0^2, \end{aligned}$$

where ϕ_0 is the value of the natural parameter corresponding to θ_0.

The ingredients of the $C(\alpha)$ statistics are

$$\begin{aligned} Y_i(\theta_0) &= \frac{f'(X; \theta_0)}{f(X_i; \theta_0)} = \frac{(X_i - \theta_0)}{\sigma_0^2}, \\ Z_i(\theta_0) &= \frac{f''(X; \theta_0)}{2f(X_i; \theta_0)} = \frac{(X_i - \theta_0)^2 - (2a\theta_0 + b)(X_i - \theta_0) - \sigma_0^2}{2\sigma_0^4}. \end{aligned}$$

We then have

$$\begin{aligned} \mathbb{E}\{Y_i(\theta_0)Z_i(\theta_0)\} &= \frac{\mathbb{E}\{(X_i - \theta_0)^3\} - (2a\theta_0 + b)\mathbb{E}\{(X_i - \theta_0)^2\} - \sigma_0^2\mathbb{E}(X_i - \theta_0)}{2\sigma_0^6} \\ &= 0. \end{aligned}$$

Therefore the regression coefficient of $Z_i(\theta_0)$ against $Y_i(\theta_0)$ is $\beta(\theta_0) = 0$. This leads to the projection $W_i(\theta_0) = Z_i(\theta_0) - \beta(\theta_0)Y_i(\beta_0) = Z_i(\theta_0)$ and

$$\begin{aligned}
4\sigma_0^8 \mathbb{V}\mathrm{ar}\{W_i(\theta_0)\} &= \mathbb{V}\mathrm{ar}\{(X_i - \theta_0)^2\} - 2(2a\theta_0 + b)\mathbb{E}\{(X_i - \theta_0)^3\} \\
&\quad + (2a\theta_0 + b)^2 \mathbb{E}\{(X_i - \theta_0)^2\} \\
&= (2a+3)\sigma_0^4 + (2a\theta_0 + b)^2\sigma_0^2 - \sigma_0^4 \\
&\quad - 2(2a\theta_0 + b)^2\sigma_0^2 + (2a\theta_0 + b)^2\sigma_0^2 \\
&= (2a+2)\sigma_0^4.
\end{aligned}$$

Hence $\nu(\theta_0) = \mathbb{V}\mathrm{ar}\{W_i(\theta_0)\} = 0.5(a+1)\sigma_0^{-4}$.

Because the maximum-likelihood estimator $\hat{\theta} = \bar{X}$ we find that

$$\sum_{i=1}^{n} W_i(\hat{\theta}) = \sum_{i=1}^{n} Z_i(\hat{\theta}) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 - \hat{\sigma}^2}{2\hat{\sigma}^4},$$

with $\hat{\sigma}^2 = a\bar{X}^2 + b\bar{X} + c$ due to invariance. The $C(\alpha)$ test statistic, $T_n = \sum_{i=1}^{n} W_i(\hat{\theta})/\sqrt{n\nu(\hat{\theta})}$, is therefore given by the simplified expression in Theorem 2.

Form of Δ_1. Theorem 1 implies $\Delta_1 = \tau^2\mathbb{E}\{W_1(\theta_0)Z_1(\theta_0)\}/\sqrt{\mathbb{E}\{W_1^2(\theta_0)\}}$. We have just shown that $W_1(\theta_0) = Z_1(\theta_0)$. Therefore

$$\Delta_1 = \tau^2\sqrt{\mathbb{E}\{W_1^2(\theta_0)\}} = \tau^2\sqrt{\nu(\theta_0)} = \tau^2\sqrt{0.5(a+1)}\sigma_0^{-2}.$$

Form of Δ_2. Recall that

$$\Delta_2 = \frac{\tau^2\mathbb{E}\{W_1(\theta_0)Z_1^*(\theta_0 + \tau/\eta^{1/2})\}}{\sqrt{\mathbb{E}\{W_1(\theta_0)^2\}}} = \frac{\tau^2\mathbb{E}\{Z_1(\theta_0)Z_1^*(\theta_0 + \tau/\eta^{1/2})\}}{\sqrt{\nu(\theta_0)}}.$$

According to the definitions of $Y_i^*(\theta)$ and $Z_i^*(\theta)$ in (A1) and (3), respectively, we have

$$Z_i^*(\theta) = \frac{Y_i^*(\theta) - Y_i^*(\theta_0)}{\theta - \theta_0}.$$

In the following we first calculate $\mathbb{E}\{Z_1(\theta_0)Y_1^*(\theta)\}$, which can be rewritten as

$$\mathbb{E}\{Z_1(\theta_0)Y_1^*(\theta)\} = \mathbb{E}\left\{Z_1(\theta_0)\frac{f(X_1;\theta) - f(X_1;\theta_0)}{(\theta - \theta_0)f(X_1;\theta_0)}\right\} = \frac{\mathbb{E}_*\{Z_1(\theta_0)\} - \mathbb{E}\{Z_1(\theta_0)\}}{\theta - \theta_0}.$$

Here $\mathbb{E}_*$ means that the expectation is taken under the density $f(x;\theta)$. Note that $\mathbb{E}\{Z_1(\theta_0)\} = 0$. Hence

$$\begin{aligned}
\mathbb{E}\{Z_1(\theta_0)Y_1^*(\theta)\} &= \frac{\mathbb{E}_*\{Z_1(\theta_0)\}}{\theta - \theta_0} \\
&= \frac{\mathbb{E}_*\{(X_i - \theta_0)^2 - (2a\theta_0 + b)(X_i - \theta_0) - \sigma_0^2\}}{2\sigma_0^4(\theta - \theta_0)} \\
&= \frac{\sigma^2 + (\theta - \theta_0)^2 - (2a\theta_0 + b)(\theta - \theta_0) - \sigma_0^2}{2\sigma_0^4(\theta - \theta_0)}.
\end{aligned}$$

With $\sigma^2 = a\theta^2 + b\theta + c$ and $\sigma_0^2 = a\theta_0^2 + b\theta_0 + c$, we can simplify $\mathbb{E}\{Z_1(\theta_0)Y_1^*(\theta)\}$ to

$$\mathbb{E}\{Z_1(\theta_0)Y_1^*(\theta)\} = \frac{(a+1)(\theta-\theta_0)}{2\sigma_0^4}.$$

Hence

$$\mathbb{E}\{Z_1(\theta_0)Z_1^*(\theta)\} = \mathbb{E}\left\{Z_1(\theta_0)\frac{Y_1^*(\theta)-Y_1^*(\theta_0)}{\theta-\theta_0}\right\} = \frac{a+1}{2\sigma_0^4}.$$

With $\nu(\theta_0) = 0.5(a+1)\sigma_0^{-4}$ as shown before, we have

$$\Delta_2 = \frac{\tau^2\mathbb{E}\{Z_1(\theta_0)Z_1^*(\theta_0+\tau/\eta^{1/2})\}}{\sqrt{\nu(\theta_0)}} = \tau^2\sqrt{0.5(a+1)}\sigma_0^{-2}.$$

This completes the proof. ■

ACKNOWLEDGEMENTS

We thank the referees and the Associate Editor for their constructive criticisms of the original version of this paper. In particular, the calibration idea is a direct consequence of an insightful suggestion from the Associate Editor. Drs. Chen and Li's work are partially supported by the Natural Sciences and Engineering Research Council of Canada grant Nos. RGPIN-2014-03743 and RGPIN-2015-06592, respectively. Dr. Liu's research is supported by grant from the National Natural Science Foundation of China (11371142), the Program of Shanghai Subject Chief Scientist (14XD1401600), and the 111 Project (B14019).

BIBLIOGRAPHY

Böhning, D., Schlattmann, P., & Lindsay, B. (1992). Computer-assisted analysis of mixtures (C.A.MAM): Statistical algorithms. *Biometrics*, 48, 283–303.

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

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

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. *The Canadian Journal of Statistics*, 26, 583–599.

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

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

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, 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, 1178–1209.

Geissler, A. (1889). Beiträge zur Frage des Geschlechtsverhältnisses der Geborenen. *Zeitschrift des K. Sächsischen Statistischen Bureaus*.

Hartigan, J. A. (1985). A failure of likelihood asymptotics for normal mixtures. In *Proc. Berkeley Conf. in Honor of J. Neyman and J. Kiefer, Volume 2*, LeCam, L. & Olshen, R. A., editors, 807–810.

Kim, D. & Lindsay, B. G. (2015). Empirical identifiability in finite mixture models. *Annals of the Institute of Statistical Mathematics*, 67, 745–772.

Le Cam, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes procedures. *University of California Publications in Statistics*, 1, 277–330.

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

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

Lindsay, B. G. (1995). *Mixture Models: Theory, Geometry and Applications*. Institute for Mathematical Statistics, Hayward.

Lindsay, B. G. & Roeder, K. (1992). Residual diagnostics in the mixture model. *Journal of the American Statistical Association*, 87, 785–795.

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

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

Neyman, J. (1959). Optimal asymptotic tests of composite statistical hypotheses. *Probability and Statistics*, Grenander, U., editor, *The Harold Cramir Volume*. Almqvist and Wiksell, Uppsala, pp. 213–234.

Neyman, J. & Scott, E. L. (1966). On the use of $C(\alpha)$ optimal test of composite hypotheses. *Bulletin de l'Institut International de Statistique*, 41, 447–497.

Proschan, F. (1963). Theoretical explanation of observed decreasing failure rate. *Technometrics*, 5, 375–383.

Sokal, R. R. & Rohlf, F. J. (1973). *Introduction to Biostatistics*, W. H. Freeman, San Francisco.

van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press, New York.

Received 21 December 2014
Accepted 8 November 2015