Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://amstat.tandfonline.com/loi/uasa20

Inference on the Order of a Normal Mixture

Jiahua Chen [a] , Pengfei Li [b] & Yuejiao Fu [c]

[a] Department of Statistics, University of British Columbia, Vancouver, BC, Canada, V6T 1Z2

[b] Department of Statistics and Actuarial Sciences, University of Waterloo, Waterloo, ON, Canada, N2L 3G1

[c] Department of Mathematics and Statistics, York University, Toronto, ON, Canada, M3J 1P3

Accepted author version posted online: 04 Jun 2012.Version of record first published: 08 Oct 2012.

To cite this article: Jiahua Chen, Pengfei Li & Yuejiao Fu (2012): Inference on the Order of a Normal Mixture, Journal of the American Statistical Association, 107:499, 1096-1105

To link to this article: http://dx.doi.org/10.1080/01621459.2012.695668

Inference on the Order of a Normal Mixture

Jiahua CHEN, Pengfei LI, and Yuejiao FU

Finite normal mixture models are used in a wide range of applications. Hypothesis testing on the order of the normal mixture is an important yet unsolved problem. Existing procedures often lack a rigorous theoretical foundation. Many are also hard to implement numerically. In this article, we develop a new method to fill the void in this important area. An effective expectation-maximization (EM) test is invented for testing the null hypothesis of arbitrary order m_0 under a finite normal mixture model. For any positive integer $m_0 \geq 2$, the limiting distribution of the proposed test statistic is $\chi^2_{2m_0}$. We also use a novel computer experiment to provide empirical formulas for the tuning parameter selection. The finite sample performance of the test is examined through simulation studies. Real-data examples are provided. The procedure has been implemented in R code. The p-values for testing the null order of $m_0 = 2$ or $m_0 = 3$ can be calculated with a single command. This article has supplementary materials available online.

KEY WORDS: Chi-squared limiting distribution; Computer experiment; EM test; Likelihood ratio test; Order selection; Tuning parameter; Unequal variance.

1. INTRODUCTION

Normal mixture models have been used for a wide range of scientific investigations. The number of components in, or the order of, the finite normal mixture model often has important scientific implications; see chapter 6 of McLachlan and Peel (2000). Hypothesis testing for the order of a normal mixture is an important problem. The order under null hypothesis is often proposed to represent some default proposition with scientific significance; the rejection of which usually leads to propositions of greater interest. We give two motivating examples as follows.

Example 1. Finite normal mixture is often used to assess the impact of possible underlying genotypes that display continuous or quantitative variation in the population (Schork, Allison, and Thiel 1996; McLachlan and Peel 2000). If the phenotype or quantitative trait is mainly influenced by a gene with two alleles A and a, there are three possible genotypes that an individual can possess: AA, Aa, and aa. Suppose the trait values associated with individuals possessing AA, Aa, and aa are distributed as $N(\theta_1, \sigma_1^2)$, $N(\theta_2, \sigma_2^2)$, and $N(\theta_3, \sigma_3^2)$, respectively. Then, the population quantitative trait has a three-component normal mixture distribution

$$\sum_{i=1}^{3} \alpha_i N\left(\theta_i, \sigma_i^2\right),$$

where α_1, α_2, and α_3 are the proportions of individuals possessing the genotypes AA, Aa, and aa, respectively. A competing genetic model arises when the genotypes AA and Aa have the same phenotypes; this leads to a two-component normal mixture model. See Roeder (1994) for a genetic example on sodium-lithium countertransport (SLC) activity in red blood cells.

Geneticists may suspect the existence of a major gene and if it exists, whether or not it is dominant. They may also want to know whether or not the quantitative trait is actually affected by several genes (Schork, Allison, and Thiel 1996). The above questions can be addressed through hypothesis tests. The null order of 1 represents the default proposition that the suspected gene does not exist; the rejection of this supports the existence of a major gene. Another null order could be 2; rejection suggests that the gene is not dominant. The rejection of a null order of 3 or less suggests the existence of several relevant genes.

Example 2. Miloslavsky and van der Laan (2003) and Pavlic, Brand, and Cummings (2001) provided an example where a normal mixture model is used to describe the observed change in the characteristic of interest in treated patients. It is suspected that the treatment has no effect on a subgroup of the patients. At the same time, one or more groups potentially respond to the treatment differentially.

The questions of interest include whether the nonresponsive group really exists; if yes, whether the people in this group respond to the treatment in a homogeneous way. The latter information will be useful to estimate the proportion of the nonresponsive group, which is of clinical importance. Again these questions can be addressed by hypothesis tests. A null order of 1 represents the nonexistence of nonresponsive groups and a null order of 2 implies that the responsive subgroup is homogeneous.

Designing effective hypothesis test methods for the order of finite mixture models goes back to Hartigan (1985) and Ghosh and Sen (1985). Both investigated the use of a likelihood ratio test for a null hypothesis of order 1, namely homogeneity. The former article reveals the impact of nonregularity of the mixture models, and the latter gives the limiting distribution of the likelihood ratio test statistic under a separation condition. Chernoff and Lander (1995) obtained the limiting distribution of the likelihood ratio test statistic for homogeneity under a binomial mixture model without separation conditions. Dacunha-Castelle and Gassiat (1999) and Liu and Shao (2003) obtained

Jiahua Chen is Professor, Department of Statistics, University of British Columbia, Vancouver, BC, Canada V6T 1Z2 (E-mail: *jhchen@stat.ubc.ca*). Pengfei Li is Assistant Professor, Department of Statistics and Actuarial Sciences, University of Waterloo, Waterloo, ON, Canada N2L 3G1 (E-mail: *pengfei.li@uwaterloo.ca*). Yuejiao Fu is Associate Professor, Department of Mathematics and Statistics, York University, Toronto, ON, Canada M3J 1P3 (E-mail: *yuejiao@mathstat.yorku.ca*). The authors thank the editor, the associate editor, and three referees for constructive comments and suggestions that lead to significant improvements in the article. The research is supported by the Natural Sciences and Engineering Research Council of Canada and by a start-up grant from the University of Waterloo.

Journal of the American Statistical Association
September 2012, Vol. 107, No. 499, Theory and Methods
DOI: 10.1080/01621459.2012.695668

the limiting distribution of the likelihood ratio test statistic for general-order null hypotheses under a compact parameter space and other assumptions. However, their elegant theoretical results are not applicable to finite normal mixture models because of the violation of several crucial conditions. Under the normal mixture model, the likelihood function is unbounded (Hathaway 1985), the Fisher information on mixing proportion can be infinity (Chen and Li 2009), and the model is not strongly identifiable (Chen 1995). Charnigo and Sun (2004) developed a D-test for homogeneity. Li and Chen (2010) developed an expectation-maximization (EM) test for assessing the order of finite mixture models with single-parameter component distributions. It is effective and easy to use for Poisson mixture models or for normal mixture models with known component variance.

Chen and Chen (2003) obtained a specific result for the likelihood ratio test under an unknown but equal component variance finite normal mixture model for homogeneity. Yet the stochastic limiting distribution is hard to implement in applications. Lo, Mendell, and Rubin (2001) aimed to develop a general theory that can be used to test the number of components (or the order) in finite normal mixture models. However, Jeffries (2003) observed that the conditions required by Lo, Mendell, and Rubin (2001) are generally not met when the null hypothesis holds. Therefore, their result is not proven and may be incorrect, according to Jeffries (2003). McLachlan (1987) proposed a resampling approach to the assessment of the p-value of the likelihood ratio test in testing the number of components. The idea cannot be directly applied to finite normal mixture models because of the unbounded likelihood function. The method in Chen and Li (2009) is applicable only to the test of homogeneity in normal mixture models.

In sharp contrast to the hypothesis test, there have been important developments on the order selection procedure for estimating the order of normal mixture model; see Leroux (1992); Richardson and Green (1997); Keribin (2000); Ishwaran, James, and Sun (2001); James, Priebe, and Marchette (2001); Miloslavsky and van der Laan (2003); Woo and Sriram (2006); and Chen and Khalili (2008).

In general, order selection procedures search for a simple model that adequately describes the real world. Hypothesis test is used to check the validity of scientific claims. For instance, an insurance company may be interested in dividing drivers into homogeneous subgroups to device a most profitable insurance product for each group. Thus, searching for the most suitable number of subgroups is the problem of interest. It fits perfectly into the order selection framework. In Examples 1 and 2, the order is linked to scientific propositions. In this case, conclusion of the hypothesis test on the order of the mixture model has direct scientific interpretations. It fits perfectly into the hypothesis test framework.

In summary, testing the order of the finite normal mixture model is an important yet arguably unsolved problem. In this article, we propose a new likelihood-based EM test for the order of finite normal mixture models. The new method first applies a penalty function on the component variance to obtain a bounded penalized likelihood. It then assesses the improvement of prespecified higher-order models over the null models. By tactically placing these higher-order models and using the EM iteration, the new method quickly locates the direction and degree of improvements in terms of the penalized likelihood. The evidence against the null order is hence assessed and quantified. We show that the new test statistic has a simple chi-squared limiting distribution. The penalty function contains a tuning parameter that affects the precision of the test. We solve the tuning problem via a novel computer experiment and provide an easy-to-use data-dependent formula. Simulation results show that the new method has accurate Type I errors and adequate power. Software implementing the test has been developed in the R language (R Development Core Team 2008) and is available as the online supplementary materials.

The organization of the article is as follows. In Section 2, we set up the testing problem, introduce the new EM test procedure, and present asymptotic results. In Section 3, we give empirical formulas for the tuning parameter through designed computer experiments. The simulation results are presented in Section 4, and the application examples are in Section 5. Section 6 presents a discussion, and the proofs are given in the online supplementary materials.

2. TEST FOR THE ORDER OF A NORMAL MIXTURE MODEL

Let $f(x;\theta,\sigma)$ be the normal density function with mean θ and variance σ^2. The finite normal mixture model with order m has density function

$$f(x;\Psi)=\sum_{h=1}^{m}\alpha_h f(x;\theta_h,\sigma_h)=\int f(x;\theta,\sigma)d\Psi(\theta,\sigma), \quad (1)$$

where the α_h's are the mixing proportions with $\sum_{h=1}^{m}\alpha_h=1$, the θ_h's and σ_h's are component parameters, and Ψ is called the mixing (or latent) distribution and takes the form

$$\Psi(\theta,\sigma)=\sum_{h=1}^{m}\alpha_h I(\theta_h\le\theta,\sigma_h\le\sigma).$$

Here $I(\cdot)$ is the indicator function. Suppose we have a random sample $X_1,\ldots,X_n$ of size n from the above normal mixture model. The goal of this article is to develop an effective procedure for testing the following null and alternative hypotheses:

$$H_0: m=m_0 \text{ versus } H_A: m>m_0 \quad (2)$$

for some given positive integer m_0. The true order of a finite mixture model is defined as the smallest number of components such that all component densities are different and all mixing proportions are nonzero. We restrict our attention to the most practical situation where all the component mean parameters are different. The testing problem where two normal mixture components have the same mean is less interesting in applications but more challenging technically. Our asymptotic result is applicable only to finite normal mixture models with distinct component means.

One hypothesis test problem of Example 1 is to test $H_0: m=2$ versus $H_A: m=3$ under normal mixture models. Our new method designed for $H_0: m=2$ versus $H_A: m>2$ is still applicable and effective. A more powerful test tailor-made specifically for $H_0: m=2$ versus $H_A: m=3$ could be possible, but we are not aware of any such methods in the literature.

2.1 The New EM Test Statistic

We start by constructing a likelihood-based consistent estimator $\hat{\Psi}_0$ of the mixing distribution under the null hypothesis. Given a random sample, the log-likelihood function of the mixing distribution is

$$l_n(\Psi) = \sum_{i=1}^{n} \log f(X_i; \Psi).$$

This likelihood function diverges to positive infinity when some component variance goes to 0. Thus, the maximum likelihood estimator of Ψ is known to be inconsistent. This is a crucial difference between the finite normal mixture model and other finite nonnormal mixture models. To overcome this technical difficulty, a penalty function is often introduced. Let $\bar{X} = n^{-1}\sum_{i=1}^{n} X_i$ and $s_n^2 = n^{-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. Define

$$\ell_n(\Psi) = l_n(\Psi) + \sum_{h=1}^{m} p_n\big(\sigma_{1h}^2; s_n^2\big)$$

for some $\hat{\sigma}^2$-dependent smooth penalty function $p_n(\sigma^2; \hat{\sigma}^2)$ of σ^2, so that it goes to negative infinity when σ goes to either 0 or infinity. Let $\hat{\Psi}_0$ be the maximum point of $\ell_n(\Psi)$ under the null hypothesis. According to Chen, Tan, and Zhang (2008), $\hat{\Psi}_0$ is a consistent estimator of Ψ under H_0, with some mild conditions on $p_n(\cdot;\cdot)$. One recommended choice is

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

We use $a_n = 1/n$ to obtain $\hat{\Psi}_0$ and this choice satisfies the conditions in Chen, Tan, and Zhang (2008).

In principle, one could compute the maximum point $\hat{\Psi}_A$ of $\ell_n(\Psi)$ under the alternative model, and construct a penalized likelihood ratio test statistic $2\{\ell_n(\hat{\Psi}_A) - \ell_n(\hat{\Psi}_0)\}$ for the purpose of testing the order. This approach does not work as expected for several reasons. First, it is technically challenging to find its finite sample or limiting distributions. Second, because the Fisher information with respect to the mixing proportion can be infinity (Chen and Li 2009), the limiting distribution may not even exist.

To overcome these difficulties, we introduce a new EM test. The test statistic has to be defined in several steps that may appear complex and unmotivated. With the help of the computer software, the actual data analysis is simple. In particular, the test statistic has a simple limiting distribution that is essential for obtaining the p-value. We will explain the motivation behind each step after the statistic is completely defined.

Let $\hat{\theta}_{0h}$ and $\hat{\sigma}_{0h}$ be the constituent entries of $\hat{\Psi}_0$ defined earlier. Without loss of generality, assume $\hat{\theta}_{01} \le \hat{\theta}_{02} \le \cdots \le \hat{\theta}_{0m_0}$. We first give the four key ingredients of the EM test.

The first ingredient is m_0 intervals defined as $I_h = (\eta_{h-1}, \eta_h]$, $h = 1, \ldots, m_0$, where $\eta_0 = -\infty$, $\eta_{m_0} = \infty$, and $\eta_h = (\hat{\theta}_{0h} + \hat{\theta}_{0\,h+1})/2$, $h = 1, \ldots, m_0 - 1$.

The second ingredient is a special class of mixing distributions of order $2m_0$ defined as

$$\Omega_{2m_0}(\boldsymbol{\beta}) = \left\{ \sum_{h=1}^{m_0} \{\alpha_h\beta_h I(\theta_{1h} \le \theta, \sigma_{1h} \le \sigma) + \alpha_h(1-\beta_h) I(\theta_{2h} \le \theta, \sigma_{2h} \le \sigma)\} : \theta_{1h}, \theta_{2h} \in I_h \right\}$$

for some vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{m_0})^\tau$ such that $\beta_h \in (0, 0.5]$.

The third ingredient is a modified penalized likelihood function defined on $\Omega_{2m_0}(\boldsymbol{\beta})$:

$$pl_n(\Psi) = l_n(\Psi) + \sum_{h=1}^{m_0} \big\{p_n\big(\sigma_{1h}^2; \hat{\sigma}_{0h}^2\big) + p_n\big(\sigma_{2h}^2; \hat{\sigma}_{0h}^2\big)\big\} + \sum_{h=1}^{m_0} p(\beta_h),$$

where $p_n(\sigma^2, \hat{\sigma}^2)$ was defined earlier, and $p(\beta)$ will be chosen as a unimodal continuous function that goes to $-\infty$ when β goes to 0. We recommend $p(\beta) = \log(1 - |1 - 2\beta|)$. Note that the penalty term $p_n(\sigma_h^2; \hat{\sigma}_{0h}^2)$ depends on $\hat{\Psi}_0$, and prevents $l_n(\Psi)$ from attaining its maximum value at mixing distributions with some $\sigma_h^2 = 0$. Further, the level of penalty, a_n, in $p_n(\sigma_h^2; \hat{\sigma}_{0h}^2)$ will be experimentally tuned so that the related statistical procedures have desirable properties. This issue will be discussed in detail in Section 3. To clarify, the a_n value used to get $\hat{\Psi}_0$ is not tuned because it leads to a sufficiently accurate $\hat{\Psi}_0$ for our purpose.

The fourth ingredient is a finite set of numbers from (0, 0.5], denoted by $\mathcal{B}$. For example, $\mathcal{B} = \{0.1, 0.3, 0.5\}$. If $\mathcal{B}$ contains J elements, then $\mathcal{B}^{m_0}$ contains J^{m_0} vectors of $\boldsymbol{\beta}$. For each $\boldsymbol{\beta}_0 \in \mathcal{B}^{m_0}$, we compute

$$\Psi^{(1)}(\boldsymbol{\beta}_0) = \arg\max\{pl_n(\Psi) : \Psi \in \Omega_{2m_0}(\boldsymbol{\beta}_0)\},$$

where the maximization is with respect to $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{m_0})^\tau$, $\boldsymbol{\theta}_1 = (\theta_{11}, \ldots, \theta_{1m_0})^\tau$, $\boldsymbol{\theta}_2 = (\theta_{21}, \ldots, \theta_{2m_0})^\tau$, $\boldsymbol{\sigma}_1 = (\sigma_{11}, \ldots, \sigma_{1m_0})^\tau$, and $\boldsymbol{\sigma}_2 = (\sigma_{21}, \ldots, \sigma_{2m_0})^\tau$. The EM algorithm with multiple initial values are used to search for $\Psi^{(1)}(\boldsymbol{\beta}_0)$. Note that $\Psi^{(1)}(\boldsymbol{\beta}_0)$ is a member of $\Omega_{2m_0}(\boldsymbol{\beta}_0)$.

When all the four ingredients are ready, the EM iteration leads to the EM test as follows. Let $\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}_0$. Suppose we have $\Psi^{(k)}(\boldsymbol{\beta}_0)$ already calculated with $\boldsymbol{\alpha}^{(k)}$, $\boldsymbol{\theta}_1^{(k)}$, $\boldsymbol{\theta}_2^{(k)}$, $\boldsymbol{\sigma}_1^{(k)}$, $\boldsymbol{\sigma}_2^{(k)}$, and $\boldsymbol{\beta}^{(k)}$ available. The calculation for $k = 1$ has been illustrated with $\boldsymbol{\alpha}^{(1)}$, $\boldsymbol{\theta}_1^{(1)}$, $\boldsymbol{\theta}_2^{(1)}$, $\boldsymbol{\sigma}_1^{(1)}$, $\boldsymbol{\sigma}_2^{(1)}$, and $\boldsymbol{\beta}^{(1)}$ being constituent entities of $\Psi^{(1)}(\boldsymbol{\beta}_0)$. A more appropriate notation might be

$$\Psi\big(\boldsymbol{\alpha}^{(1)}, \boldsymbol{\theta}_1^{(1)}, \boldsymbol{\theta}_2^{(1)}, \boldsymbol{\sigma}_1^{(1)}, \boldsymbol{\sigma}_2^{(1)}, \boldsymbol{\beta}^{(1)}; \boldsymbol{\beta}_0\big).$$

For simplicity, we use the compact notation $\Psi^{(1)}(\boldsymbol{\beta}_0)$ and more generally, $\Psi^{(k)}(\boldsymbol{\beta}_0)$.

For each $i = 1, \ldots, n$ and $h = 1, \ldots, m_0$, let

$$w_{i1h}^{(k)} = \frac{\alpha_h^{(k)}\beta_h^{(k)} f\big(X_i; \theta_{1h}^{(k)}, \sigma_{1h}^{(k)}\big)}{f\big(X_i; \Psi^{(k)}(\boldsymbol{\beta}_0)\big)}$$

and

$$w_{i2h}^{(k)} = \frac{\alpha_h^{(k)}\big(1 - \beta_h^{(k)}\big) f\big(X_i; \theta_{2h}^{(k)}, \sigma_{2h}^{(k)}\big)}{f\big(X_i; \Psi^{(k)}(\boldsymbol{\beta}_0)\big)}.$$

We then proceed to obtain $\Psi^{(k+1)}(\boldsymbol{\beta}_0)$ by setting

$$\alpha_h^{(k+1)} = n^{-1}\sum_{i=1}^{n} \big\{w_{i1h}^{(k)} + w_{i2h}^{(k)}\big\},$$

$$\big(\theta_{jh}^{(k+1)}, \sigma_{jh}^{(k+1)}\big) = \arg\max_{\theta,\sigma} \left\{ \sum_{i=1}^{n} w_{ijh}^{(k)} \log f(X_i; \theta, \sigma) + p_n\big(\sigma^2; \hat{\sigma}_{0h}^2\big) \right\}, \quad j = 1, 2,$$

and

$$\beta_h^{(k+1)} = \arg\max_{\beta} \left\{ \sum_{i=1}^{n} w_{i1h}^{(k)} \log(\beta) + \sum_{i=1}^{n} w_{i2h}^{(k)} \log(1-\beta) + p(\beta) \right\}.$$

The computation is iterated a prespecified number of times, K.

For each $\boldsymbol{\beta}_0 \in \mathcal{B}^{m_0}$ and k, we define

$$M_n^{(k)}(\boldsymbol{\beta}_0) = 2\{pl_n(\Psi^{(k)}(\boldsymbol{\beta}_0)) - l_n(\hat{\Psi}_0)\}.$$

The retooled EM test statistic, for a prespecified K, is then defined to be

$$\mathrm{EM}_n^{(K)} = \max\{M_n^{(K)}(\boldsymbol{\beta}_0) : \boldsymbol{\beta}_0 \in \mathcal{B}^{m_0}\}. \tag{3}$$

The new EM test rejects the null hypothesis when $\mathrm{EM}_n^{(K)}$ exceeds some critical value.

We now give the motivation behind the complex definition. In the usual likelihood ratio test, the likelihood will be maximized over the whole parameter space. The degree of improvement in the log-likelihood value over the best possible null model is used for the test. The literature tells us that such a statistic has complex stochastic behavior even in simple situations (Dacunha-Castelle and Gassiat 1999; Liu and Shao 2003). Based on this observation, our first two ingredients confine the primary candidate alternative models to a relatively simple subset of mixing distributions $\Omega_{2m_0}(\boldsymbol{\beta})$. The optimal mixing distribution within $\Omega_{2m_0}(\boldsymbol{\beta}_0)$ has simple asymptotic properties when the null hypothesis is true for any fixed $\boldsymbol{\beta}_0$. This leads to $M_n^{(1)}(\boldsymbol{\beta}_0)$, which also has a simple limiting distribution. In principle, $M_n^{(1)}(\boldsymbol{\beta}_0)$ can be directly used for testing the order of the finite normal mixture model.

However, $M_n^{(1)}(\boldsymbol{\beta}_0)$ is overly dependent on an arbitrary choice of $\boldsymbol{\beta}_0$. This leads to the fourth ingredient, $\mathcal{B}^{m_0}$, which contains a number of $\boldsymbol{\beta}$ vectors that fill the space evenly. Hence, $\mathrm{EM}_n^{(1)}$ measures the amount of improvement in the log-likelihood over many representative alternative models. Any specific alternative model is likely reasonably approximated by one mixing distribution in $\cup\{\Omega_{2m_0}(\boldsymbol{\beta}) : \boldsymbol{\beta} \in \mathcal{B}^{m_0}\}$.

The EM iteration further expands the range of alternative models being investigated. It checks the amount of possible further improvement in the log-likelihood from a few iterations. With the help of the third ingredient, the simple limiting distribution of $\mathrm{EM}_n^{(1)}$ is not destroyed by a finite number of iterations. Thus, we obtain a new test that is highly effective yet simple to implement.

2.2 Asymptotic Distribution

The asymptotic distribution of $\mathrm{EM}_n^{(K)}$ is obtained with the careful choice of two penalty functions $p(\beta)$ and $p_n(\cdot;\cdot)$:

C 1. $p(\beta)$ is a continuous function such that it is maximized at $\beta = 0.5$ and goes to negative infinity as β goes to 0 or 1. Further, $p(0.5) = 0$.

C 2. For any given $\sigma_2^2 > 0$, $p_n(\sigma_1^2;\sigma_2^2)$ is a smooth function of σ_1^2 and is maximized at $\sigma_1^2 = \sigma_2^2$. Further, $p_n(\sigma_2^2;\sigma_2^2) = 0$.

C 3. For any given $\sigma_1^2 > 0$ and $\sigma_2^2 > 0$, $p_n(\sigma_1^2;\sigma_2^2) = o(n)$.

C 4. For any given $\sigma_2^2 > 0$, there exists a $c > 0$ such that $p_n(\sigma_1^2;\sigma_2^2) \le 4(\log n)^2 \log(\sigma_1)$, when $\sigma_1 \le c/n$ and n is large.

C 5. For any given $\sigma_1^2 > 0$ and $\sigma_2^2 > 0$, $p_n'(\sigma_1^2;\sigma_2^2) = o_p(n^{1/4})$. Here, $p_n'(\sigma_1^2;\sigma_2^2)$ is the partial derivative of $p_n(\sigma_1^2;\sigma_2^2)$ with respect to σ_1^2.

Examples of functions satisfying the above mathematical conditions were given earlier. Since the user has the freedom to choose the penalty functions, these conditions are not restrictive as long as such functions exist. The utility of $p(\beta)$ is to restore some level of identifiability to finite mixture models. The penalty $p_n(\sigma^2;\hat{\sigma}^2)$ prevents fitted mixing distributions with degenerate component variances. Conditions C2–C4 make $\hat{\Psi}_0$ a consistent estimator of Ψ as shown by Chen, Tan, and Zhang (2008). Condition C5 allows a particularly simple limiting distribution to be presented below. Its proof is given in the online supplementary materials.

Theorem 1. Let $\mathrm{EM}_n^{(K)}$ be defined as in (3) based on a random sample of size n from the finite normal mixture model (1). Assume that the penalty functions in the definition of $\mathrm{EM}_n^{(K)}$ satisfy C1–C5, and the set $\mathcal{B}$ in the definition of $\mathrm{EM}_n^{(K)}$ contains the real number 0.5. Under the null hypothesis H_0 (2) that the order of the finite normal mixture model $m = m_0$ and for any fixed finite positive integer K,

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

in distribution as the sample size $n \to \infty$.

The definition of $\mathrm{EM}_n^{(K)}$ involves a few tuning parameters; they must be determined for each application. In the next section, we give some recommendations based on experience and computer experiments. The results are implemented in the R code so that an approximate p-value and other statistics can be obtained with a single command. The code is also given in the online supplementary materials.

3. PENALTY FUNCTION RECOMMENDATIONS

To apply the EM test, we must specify the set $\mathcal{B}$, the number of iterations K, and the penalty functions $p(\beta)$ and $p_n(\sigma^2;\hat{\sigma}^2)$. Based on our experience, we recommend choosing $\mathcal{B} = \{0.1, 0.3, 0.5\}$, $K = 3$, and $p(\beta) = \log(1 - |1 - 2\beta|)$. For the penalty function $p_n(\sigma^2;\hat{\sigma}^2)$, we recommend

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

This function satisfies Conditions C2–C5 when $a_n = o_p(n^{1/4})$. That is, any choice of a_n of this order does not change the first-order asymptotic property given in Theorem 1. Tuning a_n further for computing $\Psi^{(K)}(\boldsymbol{\beta}_0)$ improves the precision of the level of the EM test.

In the spirit of the Bartlett correction (Bartlett 1953), we wish to choose the value of a_n such that

$$E\{\mathrm{EM}_n^{(K)}\} = 2m_0,$$

when the sample is from a null model. This can be challenging or even impossible because a_n may depend on the unknown distribution of the sample. Instead, we develop through computer experiments an empirical formula for a_n based on the sample size, the data, and the null hypothesis, so that $E\{\mathrm{EM}_n^{(K)}\}$ and $2m_0$ are close.

Given a mixing distribution Ψ_0 and a sample size, we simulate the value of $E\{\mathrm{EM}_n^{(K)}\}$ as a function of a_n and find the value

$\hat{a}_n$ that solves $E\{\mathrm{EM}_n^{(K)}\} = 2m_0$. We regard $\hat{a}_n$ or its function as a dependent variable and (n, Ψ_0) as explanatory variables. Through exploratory data analysis, we build a regression model between $\hat{a}_n$ and some covariates based on (n, Ψ_0). We therefore obtain an empirical formula in the form of

$$a_n = g(n, \Psi_0).$$

In applications, we first compute $\hat{\Psi}_0$ and then choose a tuning parameter according to $a_n = g(n, \hat{\Psi}_0)$ for the EM test.

We present the results of the computer experiments for the two most important cases, $m_0 = 2$ and $m_0 = 3$, in the following two subsections. They cover most application examples we are aware of in the literature, including those given in Section 1. If necessary, formulas can be obtained for $m_0 = 4$ or higher in the same way. Although the tuning formulas are obtained via deliberate computer experiments and careful design, their application is no more complex than a few lines of R code.

3.1 Empirical Formulas for a_n With $m_0 = 2$

We carry out pilot experiments for many choices of representative normal mixture models of order $m_0 = 2$. We find that when $a_n > 0.35$, the averages of the EM test are smaller than $2m_0 = 4$. For the designed experiment, we choose three levels for the sample size n: 100, 300, 500; two levels for the mixing proportions: $(\alpha_1, \alpha_2) = (0.25, 0.75)$, $(0.5, 0.5)$; three levels for the component means: $(\theta_1, \theta_2) = (-1.5, 1.5), (-2, 2), (-2.5, 2.5)$; and two levels for the component variances: $(\sigma_1, \sigma_2) = (1, 1), (1.5, 0.75)$. A full factorial design with $3 \times 2 \times 3 \times 2 = 36$ level combinations is implemented. We use 1000 repetitions at each level combination to obtain $\hat{a}_n$. The results are reported in Table 1.

We transform $\hat{a}_n$ into $y = \log\{\hat{a}_n/(0.35 - \hat{a}_n)\}$ to restrict the fitted value of a_n to the interval $(0, 0.35)$. We then search for sensible functions of n and Ψ_0 to serve as covariates. It turns out that n^{-1} and the average misclassification rate ω_{12} discussed by Maitra and Melnykov (2010) are good choices. Based on a two-component normal mixture model, an observation X from component 2 is misclassified into component 1 with probability

$$\omega_{1|2} = \Pr(\alpha_1 f(X; \theta_1, \sigma_1) > \alpha_2 f(X; \theta_2, \sigma_2)).$$

Similarly, let $\omega_{2|1}$ be the opposite misclassification rate and ω_{12} be the average misclassification rate. Regressing y with respect to the covariates n^{-1} and ω_{12} gives

$$\hat{y} = -1.859 - 0.577 \log\{\omega_{12}/(1 - \omega_{12})\} - 60.453/n,$$

Table 1. Simulated $\hat{a}_n$ values under normal mixtures with $m_0 = 2$

	$(\sigma_1, \sigma_2) = (1, 1)$			$(\sigma_1, \sigma_2) = (1.5, 0.75)$		
(θ_1, θ_2)	$n = 100$	$n = 300$	$n = 500$	$n = 100$	$n = 300$	$n = 500$
			$(\alpha_1, \alpha_2) = (0.5, 0.5)$			
(−1.5,1.5)	0.107	0.134	0.183	0.093	0.128	0.169
(−2.0,2.0)	0.160	0.179	0.226	0.145	0.179	0.166
(−2.5,2.5)	0.233	0.301	0.230	0.181	0.196	0.259
			$(\alpha_1, \alpha_2) = (0.25, 0.75)$			
(−1.5,1.5)	0.082	0.098	0.139	0.087	0.095	0.093
(−2.0,2.0)	0.128	0.170	0.153	0.108	0.123	0.134
(−2.5,2.5)	0.190	0.210	0.241	0.174	0.164	0.211

Table 2. Simulated $\hat{a}_n$ values under normal mixtures with $m_0 = 3$

	$(\sigma_1, \sigma_2, \sigma_3) = (1, 1, 1)$			$(\sigma_1, \sigma_2, \sigma_3) = (0.75, 1.5, 0.75)$		
$(\theta_1, \theta_2, \theta_3)$	$n = 100$	$n = 300$	$n = 500$	$n = 100$	$n = 300$	$n = 500$
(−4,0,4)	0.098	0.155	0.16	0.092	0.12	0.131
(−4,0,5)	0.119	0.197	0.21	0.106	0.14	0.159
(−5,0,5)	0.144	0.225	0.259	0.123	0.164	0.181
(−4,0,6)	0.131	0.213	0.232	0.119	0.164	0.179
(−5,0,6)	0.153	0.236	0.276	0.138	0.212	0.237
(−6,0,6)	0.167	0.268	0.289	0.112	0.194	0.252

with $R^2 = 0.78$. From this, we derive an empirical formula for a_n with $m_0 = 2$:

$$a_n = \frac{0.35 \exp(-1.859 - 0.577 \log\{\omega_{12}/(1 - \omega_{12})\} - 60.453/n)}{1 + \exp(-1.859 - 0.577 \log\{\omega_{12}/(1 - \omega_{12})\} - 60.453/n)}. \quad (4)$$

Its effectiveness will be illustrated in simulation studies.

3.2 Empirical Formulas for a_n With $m_0 = 3$

For $m_0 = 3$, we develop an empirical formula for a_n by considering the following three factors: component mean, component variance, and sample size. We include only one set of mixing proportions, $(\alpha_1, \alpha_2, \alpha_3) = (0.33, 0.33, 0.34)$. We use 1000 repetitions at each level combination to obtain the $\hat{a}_n$ values. The specification of the $6 \times 2 \times 3 = 36$ full factorial design and the results are given in Table 2.

Next, we regress the response $y = \log\{\hat{a}_n/(0.35 - \hat{a}_n)\}$ with respect to three covariates: n^{-1}, $\log\{\omega_{12}/(1 - \omega_{12})\}$, and $\log\{\omega_{23}/(1 - \omega_{23})\}$, where ω_{12} and ω_{23} are the average of the misclassification probabilities between the first two components and the last two components, respectively. The fitted coefficients for $\log\{\omega_{12}/(1 - \omega_{12})\}$ and $\log\{\omega_{23}/(1 - \omega_{23})\}$ are not significantly different. They are hence replaced by a single covariate $\log\{\omega_{12}\omega_{23}/(1 - \omega_{12})(1 - \omega_{23})\}$, and the resulting model becomes

$$\hat{y} = -1.602 - 0.240 \log\{\omega_{12}\omega_{23}/(1 - \omega_{12})(1 - \omega_{23})\} - 130.394/n,$$

with $R^2 = 0.89$. Therefore, the empirical formula for a_n with $m_0 = 3$ is

$$a_n = \{0.35 \exp(-1.602 - 0.240 \log\{\omega_{12}\omega_{23}/(1 - \omega_{12}) \times (1 - \omega_{23})\} - 130.394/n)\}/\{1 + \exp(-1.602 - 0.240 \times \log\{\omega_{12}\omega_{23}/(1 - \omega_{12})(1 - \omega_{23})\} - 130.394/n)\}.$$

We next illustrate the effectiveness of the empirical formulas by simulation.

4. SIMULATION STUDY

The purpose of the simulation is two-fold: to assess the accuracy of the proposed asymptotic approximation in finite samples and to examine the power of the EM test. The EM test is calculated based on the recommendations for $\mathcal{B}$, K, and the two penalty functions $p(\beta)$ and $p_n(\sigma^2; \hat{\sigma}^2)$.

We first test $H_0 : m = 2$. In total, we choose 12 null models with order 2 as specified in Table 3. For each null model

Table 3. Parameter specifications for 12 null models with order 2

(α_1, α_2)	(0.5,0.5), (0.2,0.8)
(θ_1, θ_2)	(−1.25,1.25), (−1.75,1.75), (−2.25,2.25)
(σ_1, σ_2)	(1,1), (1.2,0.6)

under two sample sizes, 200 and 400, we calculate the simulated Type I error rates based on 5000 repetitions. The simulation results for two significance levels, 5% and 1%, are summarized in Figure 1. The plots show that the observed levels are close to their targets in most cases. To check the power of the EM test, we choose eight alternative models as specified in Table 4. The power of the EM test under each alternative model is calculated based on 1000 repetitions. The simulation results are also summarized in Table 4. The power of the EM test is much larger than the level, and increases as the sample size increases.

Next, we test $H_0 : m = 3$. In total, we choose 12 null models with order 3 as specified in Table 5. Figure 2 shows that the observed levels are close to their nominal levels in most cases. We again choose eight alternative models as specified in Table 6. The power of the EM test is calculated based on 1000 repetitions and the results are summarized in Table 6. Clearly, as the sample size increases, the power increases. Further, when the component means under the alternative models become far away from one another, the power of the test increases.

5. APPLICATION EXAMPLES

5.1 SLC Data

Geneticists often study SLC activity in red blood cells, since it relates to blood pressure and the prevalence of hypertension. Furthermore, SLC activity is easier to study than blood pressure; see Example 1 and Roeder (1994).

Suppose the SLC activity is determined by a simple mode of inheritance compatible with the action of a single gene with two alleles. If each observation was composed of the sum of a genetic component and a normally distributed measurement error, then the SLC measurements would follow one of two competing genetic models, namely the simple dominance model or the additive model, corresponding to either a two-component or a three-component normal mixture model. Hence, there is a need to test the null hypothesis $m_0 = 2$.

The data consist of red blood cell SLC activity measured for 190 individuals. We apply the EM test for $H_0 : m = m_0 = 2$. The constituent entries of $\hat{\Psi}_0$ are

$$(\hat{\alpha}_{01}, \hat{\alpha}_{02}, \hat{\theta}_{01}, \hat{\theta}_{02}, \hat{\sigma}_{01}, \hat{\sigma}_{02}) = (0.654, 0.346, 2.194, 3.455, 0.556, 1.081).$$

With $\hat{\Psi}_0$, the estimated average overlap probability ω_{12} between the two groups is 0.211. Applying the empirical formula in (4), we get $a_n = 0.068$ and $\mathrm{EM}_n^{(1)} = 4.595$, $\mathrm{EM}_n^{(2)} = 4.637$, and $\mathrm{EM}_n^{(3)} = 4.657$, with corresponding p-values 0.331, 0.327, and 0.324, respectively. Therefore, the simple dominance model is not rejected.

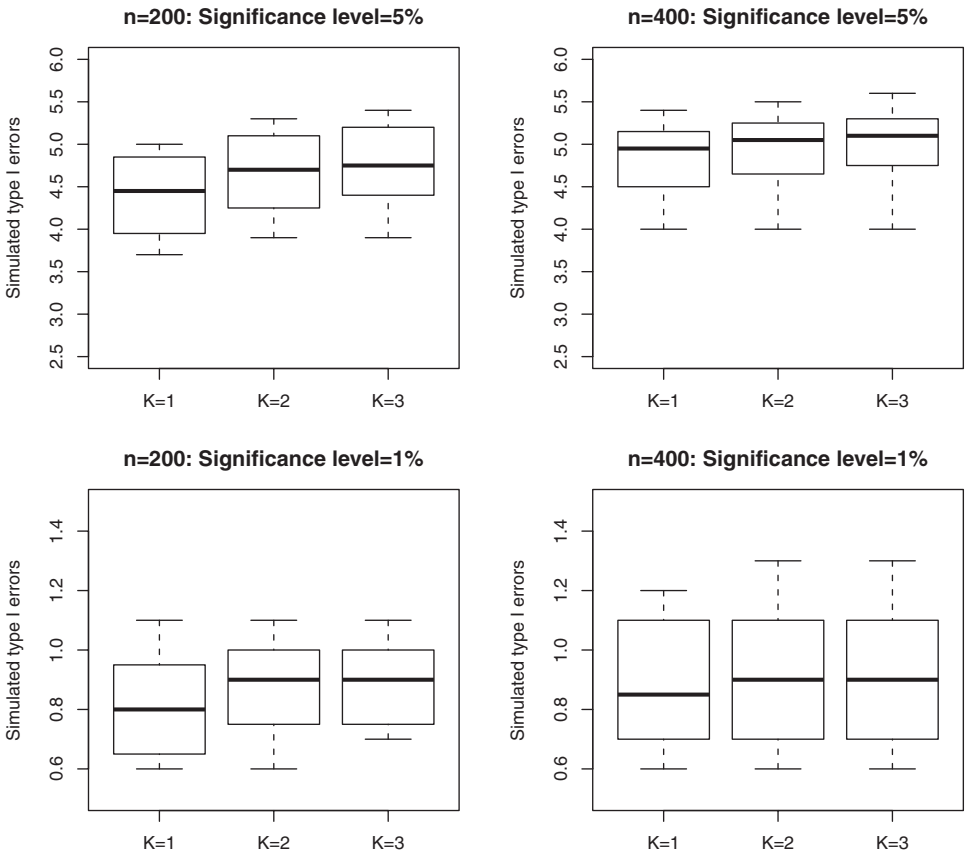

Figure 1. Simulated Type I errors of $\mathrm{EM}_n^{(K)}$ for $m_0 = 2$.

Table 4. Parameters and powers of EM test in eight alternative models for testing against $H_0 : m = 2$ at the 5% level

Alternative models		$n = 200$			$n = 400$		
		$EM_n^{(1)}$	$EM_n^{(2)}$	$EM_n^{(3)}$	$EM_n^{(1)}$	$EM_n^{(2)}$	$EM_n^{(3)}$
$(\alpha_1, \alpha_2, \alpha_3)$	$(\sigma_1, \sigma_2, \sigma_3)$	$(\theta_1, \theta_2, \theta_3) = (-2.5,0,2.5)$					
(1/3,1/3,1/3)	(1,1,1)	25.3	25.5	25.8	64.5	64.8	64.8
(0.4,0.2,0.4)	(1,1,1)	22.7	23.1	23.2	51.1	51.2	51.3
(1/3,1/3,1/3)	(0.6,1.2,0.6)	99.1	99.1	99.1	100.0	100.0	100.0
(0.4,0.2,0.4)	(0.6,1.2,0.6)	98.8	98.8	98.8	100.0	100.0	100.0
$(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$	$(\sigma_1, \sigma_2, \sigma_3, \sigma_4)$	$(\theta_1, \theta_2, \theta_3, \theta_4) = (-3,-1,1,3)$					
(0.25,0.25,0.25,0.25)	(1,1,1,1)	19.6	19.9	20.0	43.7	43.9	44.1
(0.35,0.15,0.15,0.35)	(1,1,1,1)	32.9	33.2	33.5	70.2	70.2	70.2
(0.25,0.25,0.25,0.25)	(0.6,1.2,1.2,0.6)	40.1	40.4	40.5	59.7	60.0	60.2
(0.35,0.15,0.15,0.35)	(0.6,1.2,1.2,0.6)	100.0	100.0	100.0	100.0	100.0	100.0

The above analysis was accomplished with an R function that we have developed

```
> emtest.norm(x, 2)
```

where x denotes the SLC data vector and 2 is the null order; see the online supplementary materials for more details.

Roeder (1994) analyzed the data and concluded that a three-component normal mixture is most suitable. The difference between the two conclusions may be due to the equal-component variance assumption in Roeder's analysis. We may examine the equal-variance assumption by testing the hypothesis $H_0 : \sigma_1 = \sigma_2$ versus $H_A : \sigma_1 \neq \sigma_2$. Under the null hypothesis, the penalized likelihood ratio test statistic $R_n = 2\{\sup_{H_A} \ell_n(\Psi) - \sup_{H_0} \ell_n(\Psi)\}$ converges in distribution to χ_1^2 (Chen, Tan, and Zhang 2008). For the SLC data, we find that the penalized likelihood ratio test statistic for testing the variance equality in the two-component normal mixture is 6.37. Calibrated by the χ_1^2 distribution, the p-value is 1.2%, which suggests strong evidence against the equal-variance assumption.

If $\sigma_2^2 \gg \sigma_1^2$ in a two-component mixture, a fitting based on the equal-variance assumption may split the second component into several to compensate for the overdispersion. This is probably why Roeder's method favors a three-component mixture while the EM test favors a two-component mixture. This is visually supported by Figure 3, where we have included the two-component fitting with unequal variances and the three-component fitting with equal variances. The two-component fitting with unequal variances provides a slightly better fit, especially in the region where the SLC measurement is in the neighborhood of 4.

The correctness of the simple dominant mode cannot be determined solely by statistical analysis. However, for the first time, we have a way to quantify the support of this mode.

Table 5. Parameter specifications for 12 null models with order 3

$(\alpha_1, \alpha_2, \alpha_3)$	(1/3,1/3,1/3), (0.25,0.5,0.25)
$(\theta_1, \theta_2, \theta_3)$	(−3.5,0,4.5), (−4.5,0,4.5)
$(\sigma_1, \sigma_2, \sigma_3)$	(1,1,1), (0.6,1.2,0.6), (0.6,0.6,1.2)

5.2 Adulteration in Wine Production

A normal mixture model is used by Monetti et al. (1996) to estimate the proportion of adulterated musts and establish classification regions for the acceptance or rejection of a given wine sample. Experience shows that the characteristics of the genuine samples have a normal distribution. There is no prior information on the distribution of the characteristics of the adulterated samples. A two-component normal mixture model was found to be suitable for three characteristics of the wine samples. Possible heterogeneity within the adulterated samples was not explored by these authors. This information can be important for forming the classification regions and estimating the prevalence of adulterated samples.

One would like to investigate whether or not a two-component normal mixture model is suitable for the characteristics of wine samples. Monetti et al. (1996) gave data for four variables (*scyllo*-inositol, *myo*-inositol, two D/H ratios) suitable for discovering adulterations via added sugar from plants other than grapes. The data consist of 344 observations. We use the EM test for the null hypotheses $m_0 = 2$ and $m_0 = 3$ for log-transformed *scyllo*-inositol and *myo*-inositol measurements and untransformed D/H measurements. The results are summarized in Table 7. Our results show that the order $m_0 = 3$ is most appropriate for the first three characteristics. Monetti et al. (1996) suggested $m_0 = 2$ on the basis of the nature of the problem, not rigorous hypothesis tests. They did not discuss the possibility of modeling the data with $m_0 = 3$. We find that $m_0 = 1$ is most appropriate for the fourth characteristic, which is consistent with their analysis. Using this characteristic to test $m_0 = 2$ is for illustration only, and testing $m_0 = 3$ is not necessary and omitted.

According to our result, we fit a three-component multivariate normal mixture model based on the first three characteristics: log-transformed *scyllo*-inositol and *myo*-inositol measurements and untransformed D/H_I measurement. The component with the largest mean value in each characteristic corresponds to the genuine samples. It has the proportion 69.2%, which is different to the value 75.8% in Monetti et al. (1996) obtained from a two-component multivariate mixture. Since the three-component mixture is more suitable for the data, we expect that $1-69.2\% = 30.8\%$ is a more precise estimate of the prevalence of

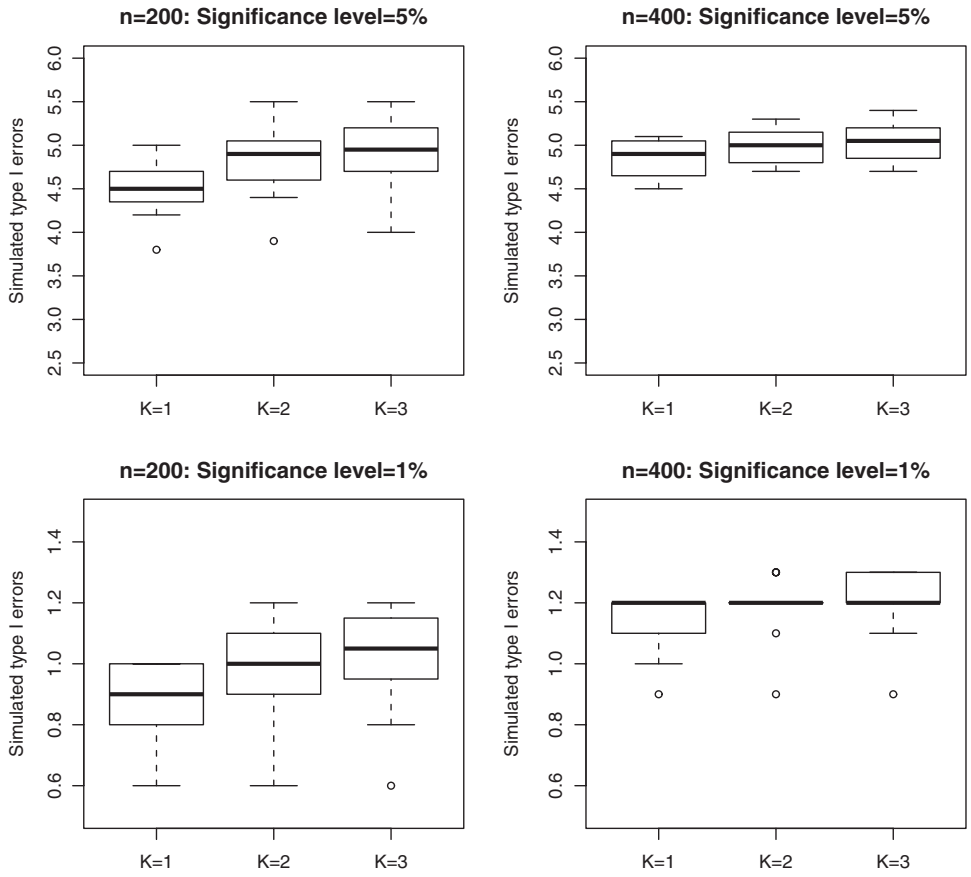

Figure 2. Simulated Type I errors of $\mathrm{EM}_n^{(K)}$ for $m_0 = 3$.

adulteration. Based on the fitted three-component model, we can further calculate the posterior probability that a given sample belongs to a genuine sample, which can be used to establish classification regions for the acceptance or rejection of a given wine sample.

5.3 Differential Gene Expression

In microarray experiments, the expression levels of a large number of genes are obtained to identify those differentially expressed over usually two samples. A t-test can be used to identify individual differentially expressed genes. Because the number of genes in such high-throughput experiments is huge, controlling the Type I error familywise is no longer sensible. Instead, geneticists favor the notion of controlling the false discovery rate (Benjamini and Hochberg 1995). Among many recipes for controlling this rate, Efron (2004) used a finite normal mixture to classify the genes into null and alternative subgroups based on the z-scores derived from the individual t-tests. Determining the order of the normal mixture model is a necessary step of such analysis. Often, the order is based on the genetic background (Efron 2004; McLachlan, Bean, and Ben-Tovim Jones 2006). In addition, an order selection procedure can be used (Chen and Khalili 2008).

Table 6. Parameters and powers of EM test in eight alternative models for testing against $H_0 : m = 3$ at the 5% level

Alternative models		$n = 200$			$n = 400$		
		$\mathrm{EM}_n^{(1)}$	$\mathrm{EM}_n^{(2)}$	$\mathrm{EM}_n^{(3)}$	$\mathrm{EM}_n^{(1)}$	$\mathrm{EM}_n^{(2)}$	$\mathrm{EM}_n^{(3)}$
$(\theta_1, \theta_2, \theta_3, \theta_4)$	$(\sigma_1, \sigma_2, \sigma_3, \sigma_4)$	$(\alpha_1, \alpha_2, \alpha_3, \alpha_4) = (0.25,0.25,0.25,0.25)$					
(−4.5,−1.5,1.5,4.5)	(1,1,1,1)	17.5	19.0	19.2	52.4	52.6	52.8
(−6,−2,2,6)	(1,1,1,1)	93.2	93.7	94.0	100.0	100.0	100.0
(−4.5,−1.5,1.5,4.5)	(0.6,1.2,0.6,1.2)	84.6	85.3	85.7	99.6	99.6	99.6
(−6,−2,2,6)	(0.6,1.2,0.6,1.2)	99.9	99.9	99.9	100.0	100.0	100.0
$(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5)$	$(\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5)$	$(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5) = (0.2,0.2,0.2,0.2,0.2)$					
(−5,−2.5,0,2.5,5)	(1,1,1,1,1)	9.1	10.1	10.4	27.5	28.2	28.4
(−6,−3,0,3,6)	(1,1,1,1,1)	38.4	40.2	40.7	84.3	84.5	84.6
(−5,−2.5,0,2.5,5)	(0.6,1.2,0.6,1.2,1)	42.3	44.5	44.8	82.6	82.7	83.0
(−6,−3,0,3,6)	(0.6,1.2,0.6,1.2,1)	80.7	81.9	82.3	99.5	99.5	99.5

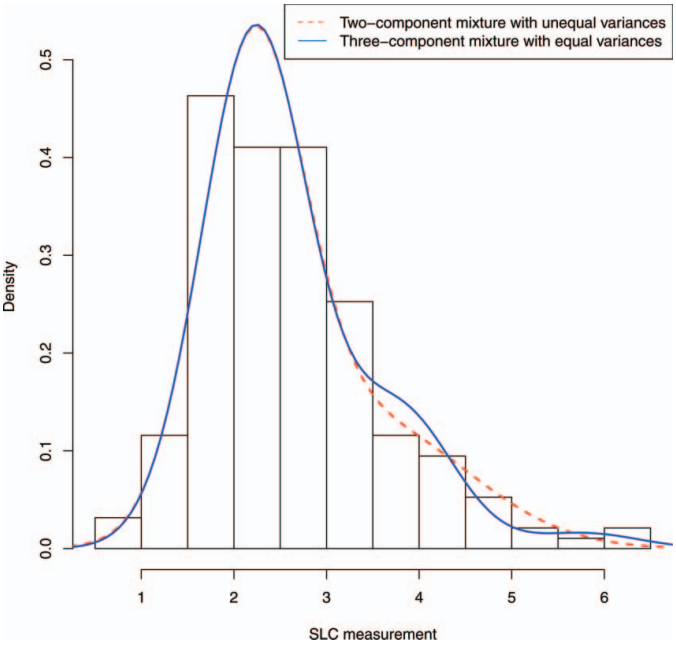

Figure 3. Histogram for 190 SLC measurements: two-component mixture fitting with unequal variances and three-component mixture fitting with equal variances. The online version of this figure is in color.

The EM test provides another rigorous approach. We analyze the prostate cancer dataset of Singh et al. (2002). This data consists of the gene expression levels of 6033 genes for 52 prostate cancer patients and 50 normal control subjects. The main objective of the study was to find the genes that are differentially expressed between the prostate cancer patients and the normal control subjects. The 6033 z-scores, transformed from the two-sample t-test statistics, can be downloaded from *http://www-stat.stanford.edu/~brad*. Positive z-scores indicate higher expression levels among prostate cancer patients, whereas negative z-scores indicate higher expression levels among normal control subjects. If the gene is not differentially expressed, theoretically, the corresponding z should follow $N(0, 1)$.

Following Efron (2004), we apply the normal mixture to model the 6033 z-scores. Two-component or three-component normal mixture models have a natural interpretation here. The two-component normal mixture implies the existence of nondifferentially expressed genes and one of the overexpressed genes (large positive component mean) and underexpressed genes (small negative component mean), while the three-component normal mixture implies the existence of all three gene types.

Table 7. EM statistics and p-values for adulteration data analysis

	$m_0 = 2$		$m_0 = 3$	
Characteristic	$EM^{(3)}$	p-value	$EM^{(3)}$	p-value
scyllo-inositol	15.96	0.003	10.78	0.096
myo-inositol	24.35	0.000	7.06	0.315
D/H_I	14.04	0.007	4.51	0.609
D/H_{II}	1.79	0.775	–	–

We use the EM test to check if the two-component normal mixture provides an adequate fit to the z-scores. We first test the null hypothesis of order 2. The EM test statistics are $EM_n^{(1)} = 12.997$, $EM_n^{(2)} = 12.998$, and $EM_n^{(3)} = 12.999$. The corresponding p-values are around 0.011. Therefore, there is strong evidence to reject the null hypothesis of order 2. We further apply the EM test to test the null hypothesis of order 3. The constituent entries of $\hat{\Psi}_0$ are

$$\begin{aligned}(\hat{\alpha}_{01}, \hat{\alpha}_{02}, \hat{\alpha}_{03}, \hat{\theta}_{01}, \hat{\theta}_{02}, \hat{\theta}_{03}, \hat{\sigma}_{01}, \hat{\sigma}_{02}, \hat{\sigma}_{03}) \\ = (0.006, 0.985, 0.009, -3.340, \\ -0.005, 3.232, 0.477, 1.064, 0.770),\end{aligned}$$

which results in $\omega_{12} = 0.154$ and $\omega_{23} = 0.228$, and further $a_n = 0.100$. The EM test statistics are $EM_n^{(1)} = 9.334$, $EM_n^{(2)} = 9.361$, and $EM_n^{(3)} = 9.380$. The corresponding p-values are 0.15. Hence, the null hypothesis of $m_0 = 3$ is not rejected at the 5% level, and a three-component normal mixture is adequate for the data.

According to the fitting of the three-component normal mixture model, the 6033 genes can be classified into three groups: underexpressed (around 0.6%), nondifferentially expressed (around 98.5%), and overexpressed (around 0.9%). The estimate of the proportion of nondifferentially expressed genes is 0.985, which is close to the estimate 0.984 obtained by Efron (2010, p. 85). Strong evidence for the existence of both overexpressed and underexpressed genes may help prostate cancer researchers to devise validation experiments accordingly.

6. DISCUSSION

Hypothesis testing on the order of normal mixture is an important but unsolved problem. This article comes up with a novel recipe to create an effective solution. The carefully designed recipe is possible only with a thorough understanding of a large number of existing ingredients. The assembly of these ingredients is itself a formidable task because of the many unyielding properties of the finite normal mixture models. The significance of the contribution is that this is the first valid and effective hypothesis test for this long-standing problem. The procedure comes with an automated tuning strategy and an easy to use R function. A single command can give the users the test statistics and the corresponding asymptotic p-values.

The referees and the associate editor raised some important issues regarding the use of the EM test. Clarification of which may also help general readers. We address three of them as follows.

The first one is on the multiple test issue. In applications, an inference goal should be unambiguously established before the data analysis. Our examples clearly violate this principle due to illustrative nature. Suppose the inference goal is to test $m = 2$ against the alternative $m > 2$ at the 5% level period. Then, we should conclude with an EM test for this pair of hypotheses. If instead, one is curious on the lowest order with adequate fit and $m < 2$ is ruled out a priori. One should then spend the level of the test in a sequential manner. For example, one may use 3% to test $m = 2$ against $m > 2$. If the test is significant, we then use 2% to test $m = 3$ against $m > 3$.

The second one is on scientific implication. A m_0-component model could be rejected for various reasons, not necessarily because the actual distribution is a normal mixture with more components. A call for accepting a specific alternative model should only be made based on scientific principles. The EM test provides valuable numerical evidence to support such conclusions.

The third one is the choice of K. The purpose of the usual EM algorithm is more on accurately locating the maximum point of the likelihood function, less on its maximum value. The slow rate of convergence is often observed because the likelihood function is flat near the maximum point, especially in finite mixture model applications. In this article, we focus on its maximum value and the EM algorithm achieves a very good precision after a few iterations. Empirical evidences suggest that $K = 3$ is satisfactory. A slightly larger K does not improve the power property because $\mathrm{EM}_n^{(K)}$ may merely increase in the third decimal place. Iterate until convergence invalidates the asymptotic conclusion.

[Received October 2011. Revised April 2012.]

REFERENCES

Bartlett, M. S. (1953), "Approximate Confidence Intervals. I," *Biometrika*, 40, 12–19. [1099]

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society,* Series B, 57, 289–300. [1103]

Charnigo, R., and Sun, J. (2004), "Testing Homogeneity in a Mixture Distribution via the L^2-Distance Between Competing Models," *Journal of the American Statistical Association*, 99, 488–498. [1097]

Chen, H., and Chen, J. (2003), "Tests for Homogeneity in Normal Mixtures With Presence of a Structural Parameter," *Statistica Sinica*, 13, 351–365. [1097]

Chen, J. (1995), "Optimal Rate of Convergence in Finite Mixture Models," *The Annals of Statistics*, 23, 221–234. [1097]

Chen, J., and Khalili, A. (2008), "Order Selection in Finite Mixture Models With a Non-Smooth Penalty," *Journal of the American Statistical Association*, 103, 1674–1683. [1097,1103]

Chen, J., and Li, P. (2009), "Hypothesis Test for Normal Mixture Models: The EM Approach," *The Annals of Statistics*, 37, 2523–2542. [1097,1098]

Chen, J., Tan, X., and Zhang, R. (2008), "Inference for Normal Mixtures in Mean and Variance," *Statistica Sinica*, 18, 443–465. [1098,1099,1102]

Chernoff, H., and 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. [1096]

Dacunha-Castelle, D., and Gassiat, E. (1999), "Testing the Order of a Model Using Locally Conic Parametrization: Population Mixtures and Stationary ARMA Processes," *The Annals of Statistics*, 27, 1178–1209. [1096,1099]

Efron, B. (2004), "Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis," *Journal of the American Statistical Association*, 99, 96–104. [1103,1104]

——— (2010), *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction* (Institute of Mathematical Statistics Monographs, Vol. I), Cambridge: Cambridge University Press. [1104]

Ghosh, J. K., and Sen, P. K. (1985), "On the Asymptotic Performance of the Log-Likelihood Ratio Statistic for the Mixture Model and Related Results," in *Proceedings of the Berkeley Conference in Honor of J. Neyman and J. Kiefer* (Vol. 2), eds. L. LeCam and R. A. Olshen, Monterey, CA: Wadsworth, pp. 789–806. [1096]

Hartigan, J. A. (1985), "A Failure of Likelihood Asymptotics for Normal Mixtures," in *Proceedings of the Berkeley Conference in Honor of J. Neyman and J. Kiefer* (Vol. 2), eds. L. LeCam and R. A. Olshen, Monterey, CA: Wadsworth, pp. 807–810. [1096]

Hathaway, R. J. (1985), "A Constrained Formulation of Maximum-Likelihood Estimation for Normal Mixture Distributions," *The Annals of Statistics*, 13, 795–800. [1097]

Ishwaran, H., James, L. F., and Sun, J. (2001), "Bayesian Model Selection in Finite Mixtures by Marginal Density Decompositions," *Journal of the American Statistical Association*, 96, 1316–1332. [1097]

James, L. F., Priebe, C. E., and Marchette, D. J. (2001), "Consistent Estimation of Mixture Complexity," *The Annals of Statistics*, 29, 1281–1296. [1097]

Jeffries, N. O. (2003), "A Note on Testing the Number of Components in a Normal Mixture," *Biometrika*, 90, 991–994. [1097]

Keribin, C. (2000), "Consistent Estimation of the Order of Mixture Models," *Sankhya*, 62, 49–66. [1097]

Leroux, B. (1992), "Consistent Estimation of a Mixture Distribution," *The Annals of Statistics*, 20, 1350–1360. [1097]

Li, P., and Chen, J. (2010), "Testing the Order of a Finite Mixture Model," *Journal of the American Statistical Association*, 105, 1084–1092. [1097]

Liu, X., and Shao, Y. (2003), "Asymptotics for Likelihood Ratio Tests Under Loss of Identifiability," *The Annals of Statistics*, 31, 807–832. [1096,1099]

Lo, Y., Mendell, N. R., and Rubin, D. B. (2001), "Testing the Number of Components in a Normal Mixture," *Biometrika*, 88, 767–778. [1097]

Maitra, R., and Melnykov, V. (2010), "Simulating Data to Study Performance of Finite Mixture Modeling and Model-Based Clustering Algorithms," *Journal of Computational and Graphical Statistics*, 19, 354–376. [1100]

McLachlan, G. J. (1987), "On Bootstrapping the Likelihood Ratio Test Statistic for the Number of Components in a Normal Mixture," *Journal of the Royal Statistical Society,* Series C, 36, 318–324. [1097]

McLachlan, G. J., Bean, R. W., and Ben-Tovim Jones, L. (2006), "A Simple Implementation of a Normal Mixture Approach to Differential Gene Expression in Multiclass Microarrays," *Bioinformatics*, 22, 1608–1615. [1103]

McLachlan, G. J., and Peel, D. (2000), *Finite Mixture Models*, New York: Wiley. [1096]

Miloslavsky, M., and van der Laan, M. J. (2003), "Fitting of Mixtures With Unspecified Number of Components Using Cross-Validation Distance Estimate," *Computational Statistics and Data Analysis*, 41, 413–428. [1096,1097]

Monetti, A., Versini, G., Dalpiaz, G., and Reniero, F. (1996), "Sugar Adulterations Control in Concentrated Rectified Grape Musts by Finite Mixture Distribution Analysis of the Myo- and Scyllo-Inositol Content and the D/H Methyl Ratio of Fermentative Ethanol," *Journal of Agricultural and Food Chemistry*, 44, 2194–2201. [1102]

Pavlic, M., Brand, R., and Cummings, S. M. (2001), "Estimating Probability of Non-Response to Treatment Using Mixture Distributions," *Statistics in Medicine*, 20, 1739–1753. [1096]

R Development Core Team (2008), *R: A Language and Environment for Statistical Computing*, Vienna: R Foundation for Statistical Computing. Available at *http://www.R-project.org*. [1097]

Richardson, S., and Green, P. J. (1997), "On Bayesian Analysis of Mixtures With an Unknown Number of Components," *Journal of the Royal Statistical Society,* Series B, 59, 731–792. [1097]

Roeder, K. (1994), "A Graphical Technique for Determining the Number of Components in a Mixture of Normals," *Journal of the American Statistical Association*, 89, 487–500. [1096,1101,1102]

Schork, N. J., Allison, D. B., and Thiel, B. (1996), "Mixture Distributions in Human Genetics," *Statistical Methods in Medical Research*, 5, 155–178. [1096]

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P., Lander, E. S., Loda, M., Kantoff, P. W., Golub, T. R., and Sellers, W. R. (2002), "Gene Expression Correlates of Clinical Prostate Cancer Behavior," *Cancer Cell*, 1, 203–209. [1104]

Woo, M., and Sriram, T. N. (2006), "Robust Estimation of Mixture Complexity," *Journal of the American Statistical Association*, 101, 1475–1485. [1097]