Overview

On the number of components in a Gaussian mixture model

Geoffrey J. McLachlan* and Suren Rathnayake

Mixture distributions, in particular normal mixtures, are applied to data with two main purposes in mind. One is to provide an appealing semiparametric framework in which to model unknown distributional shapes, as an alternative to, say, the kernel density method. The other is to use the mixture model to provide a probabilistic clustering of the data into g clusters corresponding to the g components in the mixture model. In both situations, there is the question of how many components to include in the normal mixture model. We review various methods that have been proposed to answer this question.

How to cite this article:
WIREs Data Mining Knowl Discov 2014, 4:341–355. doi: 10.1002/widm.1135

INTRODUCTION

Finite mixture models are being increasingly used to model the distributions of a wide variety of random phenomena and to cluster data sets; see, for example, McLachlan and Peel.[1] Let

$$\boldsymbol{Y} = \left(Y_1, \ldots, Y_p\right)^T \tag{1}$$

be a p-dimensional vector of feature variables. For continuous features Y_j, the density of $\boldsymbol{Y}$ can be modeled by a mixture of a sufficiently large enough number g of multivariate normal component distributions,

$$f\left(\boldsymbol{y};\boldsymbol{\Psi}_g\right) = \sum_{i=1}^{g}\pi_i\phi\left(\boldsymbol{y};\mu_i,\Sigma_i\right), \tag{2}$$

where $\phi(\boldsymbol{y};\mu,\Sigma)$ denotes the p-variate normal density function with mean μ and covariance matrix Σ. Here the vector $\boldsymbol{\Psi}_g$ of unknown parameters consists of the mixing proportions π_i, the elements of the component means μ_i, and the distinct elements of the component–covariance matrices $\Sigma_i (i = 1, \ldots, g)$. We have inserted the subscript 'g' on $\boldsymbol{\Psi}$ to explicitly denote that it pertains to a g-component mixture model.

The parameter vector $\boldsymbol{\Psi}_g$ can be estimated by maximum likelihood. For an observed random sample, $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$, the log likelihood function for $\boldsymbol{\Psi}_g$ is given by

$$\log L\left(\boldsymbol{\Psi}_g\right) = \sum_{j=1}^{n}\log f\left(\boldsymbol{y}_j;\boldsymbol{\Psi}_g\right). \tag{3}$$

The maximum likelihood estimate (MLE) of $\boldsymbol{\Psi}_g$, $\hat{\boldsymbol{\Psi}}_g$, is given by an appropriate root of the likelihood equation,

$$\partial \log L\left(\boldsymbol{\Psi}_g\right)/\partial\boldsymbol{\Psi}_g = 0. \tag{4}$$

Solutions of (4) corresponding to local maximizers of $\log L(\boldsymbol{\Psi}_g)$ can be obtained via the expectation–maximization (EM) algorithm[2]; see also McLachlan and Peel[1] (Chapter 1).

Besides providing an estimate of the density function of $\boldsymbol{Y}$, the normal mixture model (2) provides a probabilistic clustering of the observed data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ into g clusters in terms of their estimated posterior probabilities of component membership of the mixture. The posterior probability $\tau_i(\boldsymbol{y}_j;\boldsymbol{\Psi}_g)$ that the jth feature vector with observed value $\boldsymbol{y}_j$ belongs to the ith component of the mixture can be expressed by Bayes' theorem as

$$\tau_i\left(\boldsymbol{y}_j;\boldsymbol{\Psi}_g\right) = \frac{\pi_i\phi\left(\boldsymbol{y}_j;\mu_i,\Sigma_i\right)}{\sum_{h=1}^{g}\pi_h\phi\left(\boldsymbol{y}_j;\mu_h,\Sigma_h\right)} \quad (i = 1, \ldots, g; \quad j = 1, \ldots, n). \tag{5}$$

An outright assignment of the data is obtained by assigning each data point $\boldsymbol{y}_j$ to the component to which

*Correspondence to: gjm@maths.uq.edu.au

Department of Mathematics, University of Queensland, Brisbane, QLD, Australia

Conflict of interest: The authors have declared no conflicts of interest for this article.

it has the highest estimated posterior probability of belonging.

DENSITY ESTIMATION

The normal mixture model (2) can be used to estimate an unknown density function. This is because the set of all normal mixture densities is dense in the set of all density functions under the L1 metric; see, for example, Li and Barron.[3] In this context of density estimation, the commonly used criterion Bayesian information criterion (BIC) of Schwarz[4] would appear to be adequate for choosing the number of components g for a suitable density estimate. In particular, Leroux[5] established under mild conditions that certain penalized log likelihood criteria, including Akaike's information criterion[6] (AIC) and BIC, do not underestimate the true number of components, asymptotically. Roeder and Wasserman[7] have shown that when a normal mixture model is used to estimate a density 'nonparametrically', the density estimate that uses BIC to select the number of components in the mixture is consistent.

ORDER OF A MIXTURE MODEL

In the sequel, we concentrate on the choice of the number of components in the situation where the mixture model is used in a clustering context. In this context, the choice of the number of components arises with the question of how many clusters there are in the data.

As discussed in McLachlan and Peel[1] (Chapter 6), a mixture density with g components might be empirically indistinguishable from one with either fewer than g components or more than g components. It is therefore sensible in practice to approach the question of the number of components in a mixture model in terms of an assessment of the smallest number of components in the mixture compatible with the data. To this end, the true order g_o of the g-component normal mixture model (2) is defined to be the smallest value of g such that the model is compatible with the data, with the model having all normal components different and all the associated mixing proportions π_i nonzero.

In some applications, available information needs to be used in addressing the order of the normal mixture model. This is because the specification of a parametric family for the component densities may have a major impact on the clustering so obtained, in particular, the number of components. For example, a mixture of univariate normal components with unequal variances will generally require fewer components to provide an adequate fit than a mixture of normal components with equal variances. But for the application at hand, it might not be reasonable to have components with disparate variances as in the case study of the 1872 Hidalgo stamp issue of Mexico.[8]

Another practical issue arises with the parametric specification of the component densities when the number of components in a mixture model are being taken to reflect the number of distinct groups in a population. Normal mixture densities can play a useful role in modeling the distribution of continuous multivariate data that have asymmetrical distributions. Indeed, any continuous distribution can be approximated arbitrarily well by a finite mixture of normal densities with common covariance matrices. Thus if a normal mixture model is being used to detect the presence of grouping in some data, then there may not be a one-to-one correspondence between the mixture components and the groups if the data have a skewed distribution within some of the groups. This is because more than one normal component may be needed to model a skewed group-conditional distribution.[9,10]

The estimation of the order of a mixture model has been discussed mainly by consideration of the likelihood, using two main ways. One way is based on a penalized form of the log likelihood. As the likelihood increases with the addition of a component to a mixture model, the likelihood (usually, the log likelihood) is penalized by the subtraction of a term that 'penalizes' the model for the number of parameters in it. This leads to a penalized log likelihood, yielding what are called information criteria for the choice of g.

The other main way for deciding on the order of a mixture model is to carry out a hypothesis test, using a likelihood ratio test (LRT). Unfortunately, the standard regularity conditions do not hold for the null distribution of the likelihood ratio test statistic (LRTS) to have its usual chi-squared distribution with degrees of freedom equal to the difference between the number of parameters under the null and alternative hypotheses.

In practice, the latter is often estimated by a resampling approach in order to produce a P-value. Thus penalized likelihood criteria, like AIC and BIC, are less demanding than the LRT. However, they produce no number that quantifies the confidence in the result, such as a P-value.

Several of the information-based criteria have been derived within a Bayesian framework for model selection, but can be applied also in a non-Bayesian framework. Hence they can be applied to choose the

number of components in mixture models considered from either a Bayesian or frequentist perspective. There are also approaches that apply only within a Bayesian framework, such as the procedure of Richardson and Green[11] who used reversible jump Markov chain Monte Carlo methods to handle the case where the dimension of the parameter space is of varying dimension. The effect of the prior structure especially with respect to the mixing proportions and to g itself is an important aspect of a Bayesian analysis of mixtures. The reader is referred to Richardson and Green[11] and the contributions of the many discussants of their paper on this issue.

LIKELIHOOD RATIO TEST

An obvious way of approaching the problem of testing for the smallest value of the number of components in a mixture model is to use the LRTS, $-2\log\lambda$, where λ denotes the likelihood ratio. Suppose we wish to test the null hypothesis,

$$H_0 : g = g_0 \tag{6}$$

versus an alternative,

$$H_1 : g = g_1, \tag{7}$$

for some $g_1 > g_0$.

We let $F_{\boldsymbol{\Psi}_g}$ denote the distribution function of the g-component mixture (2) of p-variate normal distributions, where $\boldsymbol{\Psi}_g$ denotes the vector of all unknown parameters in this model. We let $\widehat{\boldsymbol{\Psi}}_{g_i}$ denote the MLE of $\boldsymbol{\Psi}_{g_i}$ calculated under $H_i (i = 0, 1)$. The likelihood ratio λ is then given by

$$\lambda = L\left(\widehat{\boldsymbol{\Psi}}_{g_0}\right) / L\left(\boldsymbol{\Psi}_{g_1}\right), \tag{8}$$

and so the LRTS, $-2\log\lambda$, can be expressed as

$$-2\log\lambda = 2\left\{\log L\left(\widehat{\boldsymbol{\Psi}}_{g_1}\right) - \log L\left(\widehat{\boldsymbol{\Psi}}_{g_0}\right)\right\}; \tag{9}$$

that is, twice the increase in the log likelihood or the decrease in deviance.

The evidence against H_0 will be strong if λ is sufficiently small, or equivalently, if $-2\log\lambda$ is sufficiently large. Usually, $g_1 = g_0 + 1$ in practice as it is common to keep adding components until the increase in the log likelihood starts to fall away as g exceeds some threshold. The value of this threshold is often taken to be the g_0 in H_0. Of course it can happen that the log likelihood may fall away for some intermediate values of g only to increase sharply at some larger value of g.

Failure of Regularity Conditions to Hold

As remarked above, regularity conditions do not hold for $-2\log\lambda$ to have its usual asymptotic null distribution. To briefly explain why this is so, suppose that the component densities are completely specified. Then the parameter vector $\boldsymbol{\Psi}_g$ consists of just the mixing proportions. Thus, as $g_0 < g_1$ in Eqs (6) and (7), the null hypothesis is specified by the true value of $\boldsymbol{\Psi}_g$ being on the boundary of the parameter space (with one or more of the mixing proportions specified as zero). Further, if the component densities belong to, say, the same parametric family as with the normal family here, then H_0 will hold also if an appropriate number of the component distributions are not distinct. That is, H_0 corresponds to a nonidentifiable subset of the parameter space. Thus with the true value of the parameter vector under H_0 lying on the boundary of the parameter space and also in a nonidentifiable subset if the component densities depend on unknown parameters, the classic regularity conditions[12] about the asymptotic properties of the MLE are not valid under the null hypothesis H_0. In particular, the asymptotic distribution of the MLE in the nonidentifiable case under H_0 is unknown. The lack of identifiability leads to a degeneracy in the information matrix when considering the asymptotic null distribution of the (normalized) log likelihood formed under the alternative distribution H_1. As a consequence of the Fisher-information matrix being singular, the log likelihood function does not admit a large-sample approximation by a quadratic form.

In an attempt to overcome the shortcomings of the LRT for the number of components in a mixture model in a frequentist framework, Bayesian approaches have been suggested. For example, Aitkin and Rubin[13] adopted an approach which places a prior distribution on the vector of mixing proportions π. An advantage of this proposal is that any null hypothesis about the number of components is specified in the interior of the parameter space. However, Quinn et al.[14] showed that the asymptotic null distribution of $-2\log\lambda$ will not necessarily be chi-squared, as regularity conditions still do not hold. In particular, for the test of $H_0 : g = 1$ versus $H_1 : g = g_1 (> 1)$ for component densities belonging to the same parametric family, they showed that $1/n$ times the negative of the Hessian matrix of the log likelihood at the null value of the parameter vector has negative eigenvalues with nonzero probability under H_0, as $n \to \infty$. That is, they showed that

$$\lim_{n\to\infty} \text{pr}\left\{A_n > 0\right\} = \text{pr}\left\{\sum_{k=1}^{p} |Z_k| < \sqrt{p}\right\}, \tag{10}$$

where $Z_1, \ldots, Z_p$ are i.i.d. $N(0,1)$ and where the matrix $\boldsymbol{A}_n$ denotes the negative of the Hessian of the log likelihood after integrating out the prior probabilities over some prior distribution for them. In the left-hand side of (10), $\boldsymbol{A}_n > 0$ implies that the matrix $\boldsymbol{A}_n$ is positive definite. The value of the right-hand side of Eq. (10) is equal to 0.683, 0.466, and 0.288 for $p = 1$, 2, and 3, respectively.

Some Distributional Results for the LRTS

Over the years, a number of theoretical and simulation-based results have been published on the null distribution of the LRTS, $-2\log\lambda$, for inference on the number of components in a finite mixture model. We very briefly consider here some of the theoretical results that have been derived; a fuller account may be found in McLachlan and Peel[1] (Chapter 6).

Ghosh and Sen[15] provided a comprehensive account of the breakdown in regularity conditions for the classical asymptotic theory to hold for the LRTS, $-2\log\lambda$. For a mixture of two known but general univariate densities in unknown proportions, Titterington[16] and Titterington et al.[17] considered the LRT of $H_0 : g = 1 (\pi_1 = 1)$ versus $H_1 : g = 2 (\pi_1 < 1)$. They showed asymptotically under H_0 that $-2\log\lambda$ is zero with probability 0.5 and, with the same probability, is distributed as chi-squared with one degree of freedom. Another way of expressing this is that the asymptotic null distribution of $-2\log\lambda$ is the same as the distribution of

$$\{\max(0, W)\}^2, \tag{11}$$

where W is a standard normal random variable. A further way of expressing this is to say that

$$-2\log\lambda \sim \frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$$

under H_0, where χ_0^2 denotes the degenerate distribution that puts mass 1 at zero. In his monograph, Lindsay[18](Section 4.2) referred to this distribution as a chi-bar squared; that is, a mixture of chi-squared distributions.

Hartigan[19,20] obtained the same result for the asymptotic null distribution of $-2\log\lambda$ in the case of the two-component normal mixture with unspecified π_1 but known common variance and known means μ_1 and μ_2 where, as in the previous example, the null hypothesis $H_0 : g = 1$ was specified by $\pi_1 = 1$. This example was considered also by Ghosh and Sen[15] in the course of their development of asymptotic theory for the distribution of the LRTS for mixture models. They were able to derive the limiting null distribution of $-2\log\lambda$ for unknown but identifiable μ_1 and μ_2, where μ_2 lies in a compact set. They showed in the limit, that $-2\log\lambda$ is distributed as a certain functional,

$$\left[\max\left\{0, \sup_{\mu_2} W(\mu_2)\right\}\right]^2, \tag{12}$$

where $W(\cdot)$ is a Gaussian process with zero mean and covariance kernel depending on the true value of μ_1 under H_0, and the variance of $W(\mu_2)$ is unity for all μ_2.

Ghosh and Sen[15] established a similar result for component densities from a general parametric family under certain conditions. For the case where the vector of parameters $\boldsymbol{\Psi}_g$ was not assumed to be identifiable, they imposed a separation condition on the values of $\boldsymbol{\Psi}_g$ under H_0 and H_1. Hartigan[19,20] showed that if μ_2 is unknown with no restrictions on it, then $-2\log\lambda$ is asymptotically unbounded above in probability at a very slow rate $(1/2\log(\log n))$ when H_0 is true. Also, Bickel and Chernoff[21] investigated the null behavior of the LRTS for this model.

The removal of the separation condition imposed in Ghosh and Sen[15] presented a major challenge to researchers; see, for example, Dacunha–Castelle and Gassiat,[22] Chen and Chen,[23] and Liu and Shao.[24] Garel[25] subsequently showed it was possible to remove the separation condition with assumptions that involve only the second derivatives of the mixture density.

Other work on the distribution of the LRTS or modifications to it so that its limiting distribution exists includes the papers by Ruck,[26,27] Seidel et al.,[28] Seidel et al.,[29] Lo et al.,[30] Lo,[31,32] and Hall and Stewart.[33] In Jeffries,[34] it is noted that the conditions required for the result derived in Lo et al.[30] to hold are generally not met under the null hypothesis.

Chen et al.[35,36] modified the LRTS and derived its limiting distribution. Li et al.[37] and Chen and Li[38] proposed an EM test in the case of $g_0 = 1$ (that is, a single normal distribution under the null hypothesis), while it was further developed by Li and Chen[39] and Chen et al.,[40] including an extension to the case of $g_0 > 1$.

RESAMPLING APPROACH

Although as noted above the LRTS has been shown under certain conditions to be stochastically unbounded, the null distribution of the LRTS $-2\log\lambda$ does exist for finite sample size n. This was overlooked in Chen et al.[40] who stated that 'The idea cannot be directly applied to finite normal mixture models because of the unbounded likelihood function'. Although it is true that the likelihood function

is unbounded for mixtures of normal with unequal covariance matrices, the LRTS is bounded if one takes the ML solution to correspond to a local maximum, for example, to be the maximizer corresponding to the largest of the local maxima located.

We focus here on a resampling approach to the assessment of P-values associated with the use of the LRTS $-2\log\lambda$ to test for the smallest number of components in the mixture model compatible with the data. It can be viewed as a particular application of the general bootstrap approach of Efron[41,42]; see also Efron and Tibshirani.[43] Aitkin et al.[44] had adopted a resampling approach in the context of a latent class analysis, while McLachlan[45] investigated the use this approach in the context of normal mixture models.

A formal test of the null hypothesis $H_0: g = g_0$ versus the alternative $H_1: g = g_1 (g_1 > g_0)$ can be undertaken using a resampling method, as described in McLachlan.[45] With this approach, bootstrap samples are generated from the mixture model fitted under the null hypothesis of g_0 components. That is, the bootstrap samples are generated from $F_{\widehat{\Psi}_{g_0}}$, the g_0-component mixture model with the vector of unknown parameters replaced by its ML estimate $\widehat{\Psi}_{g_0}$ computed by consideration of the log likelihood formed from the original data under H_0. The value of $-2\log\lambda$, where λ is the LRTS, is computed for each bootstrap sample after fitting mixture models for $g = g_0$ and g_1 to it in turn. The process is repeated independently B times, and the replicated values of $-2\log\lambda$ formed from the successive bootstrap samples provide an assessment of the bootstrap, and hence of the true, null distribution of $-2\log\lambda$. It enables an approximation to be made to the achieved level of significance P corresponding to the value of $-2\log\lambda$ evaluated from the original sample. The rth-order statistic of the B bootstrap replications can be used to estimate the quantile of order $r/(B+1)$. A preferable alternative would be to use the rth-order statistic as an estimate of the quantile of order $(3r-1)/(3B+1)$; see Hoaglin.[46]

If a very accurate estimate of the P-value were required, then B may have to be very large.[43] Usually, however, there is no interest in estimating a P-value with high precision. Even with a limited replication number B, the amount of computation involved is still considerable, in particular for values of g_0 and g_1 not close to one. However, as noted by Smyth,[47] the process can be easily and efficiently implemented on parallel computing hardware, for example, by using B parallel processors.[47]

In the narrower sense where the decision to be made concerns solely the rejection or retention of the null hypothesis at a specified significance level α, Aitkin et al.[44] noted how, analogous to the Monte Carlo test procedure of Barnard[48] and Hope,[49] the bootstrap replications can be used to provide a test of approximate size α. The test that rejects H_0 if $-2\log\lambda$ for the original data is greater than the rth smallest of its B bootstrap replications has size

$$\alpha = 1 - r/(B+1) \tag{13}$$

approximately. For if any difference between the bootstrap and true null distributions of $-2\log\lambda$ is ignored, then the original and subsequent bootstrap values of $-2\log\lambda$ can be treated as the realizations of a random sample of size $B+1$, and the probability that a specified member is greater than r of the others is $1-r/(B+1)$. For some hypotheses the null distribution of λ will not depend on any unknown parameters, and so then there will be no difference between the bootstrap and true null distribution of $-2\log\lambda$. An example is the case of normal populations with all parameters unknown where $g_0 = 1$ under H_0. The normality assumption is not crucial in this example.

In general, the use of the estimate $\widehat{\Psi}_{g_0}$, in place of the unknown value of Ψ_g under the null hypothesis, will affect the accuracy of the P-values assessed on the basis of the bootstrap replications of $-2\log\lambda$. McLachlan and Peel[50] performed some simulations to demonstrate this effect. They observed that there was a tendency for the resampling approach using bootstrap replications to underestimate the upper percentiles of the null distribution of $-2\log\lambda$, and hence overestimate the P-value of tests based on this statistic.

Resampling Approaches for High-Dimensional Data

In the case of high-dimensional data, one might proceed in the first instance by first reducing the number of variables p to a manageable level by using the LRTS to screen each variable individually. That is, one selects the top p_1 variables ranked on the basis of the value of the LRTS in the test of a single t-component distribution versus a mixture of $g = 2$ t-components. Concerning the choice of p_1, one might try for instance two different levels of p_1, say, $p_1 = 100$ and 200. Or one might adopt some criterion to choose p_1; for example, one might wish to choose all variables for which the LRTS is greater than some specified threshold C. The value of $C = 8$ was used in the examples considered in McLachlan et al.[51] Another way would be to convert the P-value corresponding to the value of the LRTS for each variable to a z-score,

and then to fit a two-component normal mixture model to these z-scores. It provides an estimate of the estimated posterior probability that an individual variable with a given z-score is a 'null' variable (that is, its distribution can be modeled adequately by a single normal distribution); see McLachlan et al.[52] In the latter case, p_1 could be taken to be equal to the number of 'null' variables, assuming that the value of p_1 is manageable.[53]

Concerning the generation of the bootstrap samples, we can generate a bootstrap sample of p_1-dimensional observations, namely

$$\boldsymbol{Y}^*_{1,p_1}, \ldots, \boldsymbol{Y}^*_{n,p_1} \overset{\text{i.i.d.}}{\sim} F_{\widehat{\boldsymbol{\Psi}}_{g_0,p_1}}, \tag{14}$$

where we now let $\widehat{\boldsymbol{\Psi}}_{g_0,p_1}$ denote the estimate of $\widehat{\boldsymbol{\Psi}}_{g_0}$ under the null version of the mixture model with $g = g_0$, using the selected p_1 variables.

But this is ignoring the fact that the p_1 variables were selected according to some criterion from a set of p variables. So we should really generate the bootstrap sample in two steps. On Step 1, we generate the p-dimensional bootstrap sample,

$$\boldsymbol{Y}^*_1, \ldots, \boldsymbol{Y}^*_n \overset{\text{i.i.d.}}{\sim} F_{\widehat{\boldsymbol{\Psi}}_{g_0,p}}. \tag{15}$$

Then on Step 2 we obtain the p_1-dimensional bootstrap sample by selecting p_1 variables in the same way that the p_1 variables were selected from the original p-dimensional observations. We denote this bootstrap sample by

$$\boldsymbol{Y}^*_{1,p_1^*}, \ldots, \boldsymbol{Y}^*_{n,p_1^*}. \tag{16}$$

On contrasting the bootstrap sample (16) with (15), it raises the question on whether the use of the latter will lead to a biased estimation of the P-value associated with the value of the LRTS $-2\log\lambda$. In the supervised classification case, it is well known that there can be an appreciable selection bias in the estimate of a classifier when it is formed on a subset of p_1 variables selected in some optimal manner from a much larger set of p variables.[54]

We let $\widehat{P}(p_1)$ denote the P-value estimated from B bootstrap replications of the LRTS $-2\log\lambda$ generated according to (14), while we let $\widehat{P}(p_1^*)$ denote the estimated P-value using the two-step generation procedure defined by (15) and (16). McLachlan and Rathnayake[53] used mixtures of factor common analyzers[55] to estimate $\widehat{\boldsymbol{\Psi}}_{g_0}$ for use in the generation of the original data under the null hypothesis. They concluded that there was no serious bias in the assessment of the P-value obtained by performing the resampling from the mixtures of factor analyzers model fitted to the data with reduced dimension rather than from the data of full dimension. As the time in fitting a mixture of factor analyzers to a data set with dimensions p in the thousands is quite significant, there is much to be gained computationally if the estimate $\widehat{P}(p_1)$ can be used; that is, if it does not have a significant bias compared to $\widehat{P}(p_1^*)$.

INFORMATION CRITERIA IN MODEL SELECTION

Model selection can be approached in terms of the Kullback–Leibler[56] information of the true model with respect to the fitted model. If $f(\boldsymbol{y})$ denotes the true density, then the Kullback–Leibler information of $f(\boldsymbol{y})$ with respect to an estimate $f\left(\boldsymbol{y};\widehat{\boldsymbol{\Psi}}_g\right)$ is

$$I\left\{f(\boldsymbol{y}); f\left(\boldsymbol{y};\widehat{\boldsymbol{\Psi}}_g\right)\right\} = \int f(\boldsymbol{y})\log f(\boldsymbol{y})\,d\boldsymbol{y} - \int f(\boldsymbol{y})\log f\left(\boldsymbol{y};\widehat{\boldsymbol{\Psi}}_g\right)d\boldsymbol{y}, \tag{17}$$

which is a measure of the divergence of $f(\boldsymbol{y})$ relative to $f\left(\boldsymbol{y};\widehat{\Psi}_g\right)$. The aim is to make the Kullback–Leibler information (17) small. As the first term on the right-hand side of (17) does not depend on the model, only the second term is relevant. It can be expressed as

$$\eta\left(\boldsymbol{y}_T; F\right) = \int f(\boldsymbol{y})\log f\left(\boldsymbol{y};\widehat{\Psi}_g\right)d\boldsymbol{y} = \int \log f\left(\boldsymbol{y};\widehat{\boldsymbol{\Psi}}_g\right)dF(\boldsymbol{y}), \tag{18}$$

where F denotes the true distribution and $\boldsymbol{y}_T = \left(\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T\right)^T$ contains the (totally) observed data.

A simple estimator of $\eta(\boldsymbol{y}_T; F)$ is given by

$$\eta\left(\boldsymbol{y}_T;\widehat{F}_n\right) = \frac{1}{n}\sum_{j=1}^{n}\log f\left(\boldsymbol{y}_j;\widehat{\boldsymbol{\Psi}}_g\right) = \frac{1}{n}\log L\left(\widehat{\boldsymbol{\Psi}}_g\right), \tag{19}$$

obtained by replacing F in (18) by the empirical distribution function $\widehat{F}_n$, which places mass $1/n$ at each observation $\boldsymbol{y}_j (j = 1, \ldots, n)$. Usually this provides an overestimate of the expected log density

$$\int \log f(\boldsymbol{y})\,dF(\boldsymbol{y}), \tag{20}$$

as the empirical distribution function $\widehat{F}_n$ is generally closer to the fitted distribution function $F_{\boldsymbol{\Psi}_g}$ than the

true one F. The bias of $\eta\left(\boldsymbol{y}_T;\hat{F}_n\right)$ as an estimator of (20) is the functional

$$b(F) = E_F\left\{\eta\left(\boldsymbol{Y}_T;\hat{F}_n\right) - \eta\left(\boldsymbol{Y}_T;F\right)\right\}$$
$$= E_F\left\{\frac{1}{n}\sum_{j=1}^{n}\log f\left(\boldsymbol{Y}_j;\hat{\boldsymbol{\Psi}}_g\right) - \int \log f\left(\boldsymbol{y};\hat{\boldsymbol{\Psi}}_g\right) dF(\boldsymbol{y})\right\}, \quad (21)$$

where E_F denotes expectation using F as the common distribution function of the (independent) $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$.

An information criterion for model selection can be based on the bias-corrected log likelihood given by

$$\log L\left(\hat{\boldsymbol{\Psi}}_g\right) - b(F), \quad (22)$$

using an appropriate estimate of the bias term $b(F)$. The intent is to select the model (that is, the number of components in the present context) to maximize (22), and thus to minimize the Kullback–Leibler information (17).

In the literature, the information criteria so formed are generally expressed in terms of twice the negative value of this difference, so that they are of the form

$$-2\log L\left(\hat{\boldsymbol{\Psi}}_g\right) + 2C, \quad (23)$$

where the first term on the right-hand side of (23) measures the lack of fit and the second term C is the penalty term that measures the complexity of the model. The intent therefore is to choose a model to minimize the criterion (23).

Akaike's Information Criterion

Akaike[6,57] showed that $b(F)$ is asymptotically equal to d, where d is equal to the total number of parameters in the model. Thus from (22), AIC selects the model that minimizes

$$-2\log L\left(\hat{\boldsymbol{\Psi}}_g\right) + 2d; \quad (24)$$

see Bozdogan and Sclove[58] and Sclove[59] on the use of AIC in the present context of selecting the number of components in a mixture.

Konishi and Kitagawa[60] derived the corresponding asymptotic bias where the true density $f(\boldsymbol{y}_j)$ does not belong to the postulated parametric family and where the parameter vector is not necessarily estimated by maximum likelihood. However, the validity of these asymptotic expansions for $b(F)$ depend on the same regularity conditions needed for the usual asymptotic theory for the null distribution of the LRTS to hold.[17] As discussed in the previous section, these conditions break down for tests on the number of components in a mixture model.

However, in spite of this, the AIC criterion is still often used to assess the order of a mixture model. Many authors (for example, Koehler and Murphee[61]) observed that AIC is order inconsistent and tends to overfit models. In the mixture context, it means that AIC tends to overestimate the correct number of components.[62,63]

Bozdogan[64,65] proposed the informational complexity (ICOMP) criterion in an attempt to improve on the performance of AIC.

Bootstrap-Based Information Criterion

Ishiguro et al.[66] proposed that the bias term in Eq. (22) be estimated using Efron's[41] bootstrap; see also Pan.[67] Their Efron (bootstrapped) information criterion, which they called EIC, chooses the number of components g on the basis of

$$-2\log L\left(\hat{\boldsymbol{\Psi}}_g\right) + 2b\left(\hat{F}_n\right), \quad (25)$$

where the (nonparametric) bootstrap bias $b\left(\hat{F}_n\right)$ is approximated by Monte Carlo methods on the basis of B bootstrap samples. From Eq. (21),

$$b\left(\hat{F}_n\right) = E_{\hat{F}_n}\left\{\frac{1}{n}\sum_{j=1}^{n}\log f\left(\boldsymbol{Y}_j^*;\hat{\boldsymbol{\Psi}}_g^*\right) - \frac{1}{n}\sum_{j=1}^{n}\log f\left(\boldsymbol{Y}_j;\hat{\boldsymbol{\Psi}}_g^*\right)\right\}, \quad (26)$$

where $\boldsymbol{\Psi}_g^*$ denotes the MLE formed from the bootstrap sample

$$\boldsymbol{Y}_1^*, \ldots, \boldsymbol{Y}_n^* \overset{\text{i.i.d.}}{\sim} \hat{F}_n.$$

We can approximate this bootstrap bias on the basis of B independent bootstrap samples

$$\boldsymbol{Y}_{1b}^*, \ldots, \boldsymbol{Y}_{nb}^* \overset{\text{i.i.d.}}{\sim} \hat{F}_n \quad (b = 1, \ldots, B),$$

where we let $\hat{\boldsymbol{\Psi}}_{g,b}^*$ denote the MLE formed from the bth bootstrap sample ($b = 1, \ldots, B$). This gives

$$b\left(\hat{F}_n\right) \approx \frac{1}{B}\sum_{b=1}^{B}\left\{\frac{1}{n}\sum_{j=1}^{n}\log f\left(\boldsymbol{y}_{jb}^*;\boldsymbol{\Psi}_{g,b}^*\right) - \frac{1}{n}\sum_{j=1}^{n}\log f\left(\boldsymbol{y}_j;\hat{\boldsymbol{\Psi}}_{g,b}^*\right)\right\}. \quad (27)$$

Konishi and Kitagawa[60] showed that the number of bootstrap samples can be greatly reduced by using a variance–reduction technique in the bootstrap simulation.

Minimum Information Ratio Criterion

The rate of convergence of the EM algorithm is determined by the largest eigenvalue of the rate matrix,

$$\boldsymbol{I}_d - \boldsymbol{I}_c^{-1}\left(\widehat{\boldsymbol{\Psi}}_g; \boldsymbol{y}_T\right) \boldsymbol{I}\left(\widehat{\boldsymbol{\Psi}}_g; \boldsymbol{y}_T\right), \tag{28}$$

or, equivalently, by the smallest eigenvalue of the information rate matrix,

$$\boldsymbol{I}_c^{-1}\left(\widehat{\boldsymbol{\Psi}}_g; \boldsymbol{y}_T\right) \boldsymbol{I}\left(\widehat{\boldsymbol{\Psi}}_g; \boldsymbol{y}_T\right), \tag{29}$$

where $\boldsymbol{I}_c\left(\widehat{\boldsymbol{\Psi}}_g; \boldsymbol{y}_T\right)$ denotes the complete-data expected information matrix and where $\boldsymbol{I}\left(\widehat{\boldsymbol{\Psi}}_g; \boldsymbol{y}_T\right)$ denotes the observed information matrix; see, for example, McLachlan and Krishnan[68] (Chapter 5).

With the minimum information ratio (MIR) criterion of Windham and Cutler,[69] the choice of the number of components is based on the magnitude of the smallest eigenvalue e_g of the information rate matrix, with g chosen to maximize e_g over g. The value of e_g can be computed making use of the result that it is equal to one minus the rate of convergence of the EM algorithm, which can be calculated numerically McLachlan and Krishnan[68] (Chapter 5).

Polymenis and Titterington[70] proposed a modification of the MIR criterion, which was motivated by the remark of Windham and Cutler[69] that as soon as a mixture model with too many components is fitted, the observed information matrix $\boldsymbol{I}\left(\widehat{\boldsymbol{\Psi}}_g; \boldsymbol{y}\right)$ will be close to singular with the result that the corresponding e_g will be close to zero. The idea of Polymenis and Titterington[70] therefore is to detect the smallest value of $g(g_o)$ for which e_g is 'close to zero', and select g to be $g_o - 1$. In order to quantify at what point an observed value of e_g is close to zero, a Monte Carlo approach is used.[70]

Cross-Validation-Based Information Criterion

The bias correction of the log likelihood can be undertaken using cross-validation as in Smyth.[47] This cross-validation-based information criterion (CVIC) chooses g on the basis of the cross-validated log likelihood,

$$\sum_{j=1}^{n} \log f\left(\boldsymbol{y}_j; \widehat{\boldsymbol{\Psi}}_{g(j)}\right), \tag{30}$$

where $\widehat{\boldsymbol{\Psi}}_{g(j)}$ denotes the MLE of $\boldsymbol{\Psi}_g$ formed from the observed sample $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$, after first deleting the jth observation $\boldsymbol{y}_j (j = 1, \ldots, n)$. The use of cross-validation in this role can be viewed as an alternative method to evaluating the fitted model on a test sample of the same size as the original one (the training sample) on which $\widehat{\boldsymbol{\Psi}}_g$ is based.

This 'leave-one-out' cross-validated form would be very time-consuming, given that only one observation is deleted at a time. Hence consideration might be given to v-fold cross-validation where $v > 1$ observations are deleted at a time. The data set is divided up into v disjoint subsets each of size n/v. Another way known as Monte Carlo cross-validation generates B independent partitions of the data set into a test sample of size γn and a training sample of size $(1-\gamma)n$ for the estimation of $\boldsymbol{\Psi}_g$ for some fixed value of γ. The main difference between this method and the conventional v-fold cross-validation method is that each data point may be used more than once in the test set. Smyth[47] suggests that the choice of $\gamma = 0.5$ appears to be reasonably robust across a variety of problems, while values of B between 20 and 50 appear to be adequate for most applications. Smyth[47] reported some simulation results in which CVIC implemented via Monte Carlo methods was comparable with the Autoclass algorithm of Cheeseman and Stutz,[71] but the tenfold cross-validation version was often unreliable.

Clest Method

Dudoit and Fridlyand[72] proposed a prediction-based method for assessing the number of clusters in the data, which they called Clest. It is concerned with the reproducibility or predictability of the clusters. In the spirit of the Monte Carlo version of the cross-validation approach discussed above, it proceeds for a fixed number of clusters g by repeatedly dividing the original sample into two sets, a training or learning set and a test set on a given bootstrap replication b. A clustering of the learning set is obtained and a classifier is found on the basis of this clustering as if the cluster labels were the true class labels. This classifier is then applied to the test set and the predicted group labels are compared using some external index. This procedure is repeated B times to give an estimate $a_1, \ldots, a_B$ and their median m_g. The null distribution of m_g is approximated by the bootstrap under the uniformity hypothesis whereby the data are sampled from a uniform distribution in p-dimensional space.

BAYESIAN-BASED INFORMATION CRITERIA

Bayesian Approach

We now consider some criteria that have been derived within a Bayesian framework for model selection, but

can be applied also in a non-Bayesian framework, and hence to the choice of the number of components in mixture models considered from either a Bayesian or frequentist perspective.

The main Bayesian-based information criteria use an approximation to the integrated likelihood, as in the original proposal by Schwarz[4] leading to his BIC. The usual theoretical justifications of this approximation rely on the same regularity conditions that break down for inference on the number of components in a frequentist framework.

Laplace's Method of Approximation

The Bayes factor for one model against another model is the posterior odds for that model against the other when neither model is favored over the other *a priori*. It is thus equal to the ratio of the marginal or integrated likelihood for each model; see Kass and Raftery.[73] An alternative to the use of Bayes factors is to use the posterior distribution of the deviance as advocated in the monograph of Aitkin.[74]

We let $p(\boldsymbol{\Psi}_g)$ denote the prior density for $\boldsymbol{\Psi}_g$. The integrated likelihood $p(\boldsymbol{y}_T)$ is defined to be

$$\begin{aligned} p\left(\boldsymbol{y}_T\right) &= \int p\left(\boldsymbol{\Psi}_g, \boldsymbol{y}_T\right) d\boldsymbol{\Psi}_g, \\ &= \int \exp\left\{\log p\left(\boldsymbol{\Psi}_g, \boldsymbol{y}_T\right)\right\} d\boldsymbol{\Psi}_g, \end{aligned} \tag{31}$$

where

$$p\left(\boldsymbol{\Psi}_g, \boldsymbol{y}_T\right) = p\left(\boldsymbol{\Psi}_g\right) L\left(\boldsymbol{\Psi}_g\right).$$

We let $\widetilde{\boldsymbol{\Psi}}_g$ denote the posterior mode, satisfying

$$\partial \log p\left(\widetilde{\boldsymbol{\Psi}}_g, \boldsymbol{y}_T\right) / \partial \boldsymbol{\Psi}_g = 0, \tag{32}$$

where $\partial \log p\left(\widetilde{\boldsymbol{\Psi}}_g, \boldsymbol{y}_T\right) / \partial \boldsymbol{\Psi}_g$ denotes the gradient of $\log p(\boldsymbol{\Psi}_g, \boldsymbol{y}_T)$ evaluated at $\boldsymbol{\Psi}_g = \widetilde{\boldsymbol{\Psi}}_g$. The negative Hessian matrix of $\log p(\boldsymbol{\Psi}_g, \boldsymbol{y}_T)$ evaluated at $\boldsymbol{\Psi}_g = \widetilde{\boldsymbol{\Psi}}_g$ is denoted by $\boldsymbol{H}\left(\widetilde{\boldsymbol{\Psi}}_g\right)$.

Using a second-order Taylor series about the point $\boldsymbol{\Psi}_g = \widetilde{\boldsymbol{\Psi}}_g$ it can be approximated to give

$$\begin{aligned} \log p\left(\boldsymbol{y}_T\right) \approx{}& \log L\left(\widetilde{\boldsymbol{\Psi}}_g\right) + \log p\left(\widetilde{\boldsymbol{\Psi}}_g\right) \\ &- \frac{1}{2} \log \left|\boldsymbol{H}\left(\widetilde{\boldsymbol{\Psi}}_g\right)\right| + \frac{1}{2} d \log (2\pi). \end{aligned} \tag{33}$$

This approximation is known as Laplace's method or the saddle-point approximation.

Laplace's method may be applied in alternative forms by omitting part of the integrand from the exponent when performing the expansion; see Kass and Raftery.[73] An important variant on Eq. (33) is

$$\begin{aligned} \log p\left(\boldsymbol{y}_T\right) = \log L\left(\widehat{\boldsymbol{\Psi}}_g\right) + \log p\left(\widehat{\boldsymbol{\Psi}}_g\right) \\ - \frac{1}{2} \log \left|\boldsymbol{I}\left(\widehat{\boldsymbol{\Psi}}_g; \boldsymbol{y}_T\right)\right| + \frac{1}{2} d \log (2\pi), \end{aligned} \tag{34}$$

where the posterior mode is replaced by the MLE $\boldsymbol{\Psi}_g$ and $\boldsymbol{H}(\boldsymbol{\Psi}_g)$ is replaced by the observed information matrix $\boldsymbol{I}(\boldsymbol{\Psi}_g; \boldsymbol{y}_T)$. This approximation thus assumes that the prior is very diffuse so that its effect can be effectively ignored. As cautioned by Ripley[75] (Section 2.6), the assumption that the prior can be neglected is a strong one.

Bayesian Information Criterion

The BIC of Schwarz[4] is obtained by ignoring terms of $O(1)$ in Eq. (34) and noting that

$$\left|\boldsymbol{I}\left(\widehat{\boldsymbol{\Psi}}_g; \boldsymbol{y}_T\right)\right| = O\left(n^d\right) \tag{35}$$

to give

$$-2 \log L\left(\widehat{\Psi}_g\right) + d \log n \tag{36}$$

as twice the negative penalized log likelihood to be minimized in model selection, including the present situation for the number of components g in a normal mixture model.

Note that BIC can be used not only to choose the number of components in the mixture model, but also to decide on the adopted model, say, for the component–covariance matrices in the normal component densities; see, for example, Biernacki and Govaert.[76]

The approximation (34) requires the parameters to be identifiable. Hence both this approximation and the expansion (33) depend on regularity conditions that do not hold for finite mixture models. However, as Fraley and Raftery[77] note, there is considerable support for use of BIC in this context. As mentioned previously, Leroux[5] has shown that BIC does not underestimate the true number of components, asymptotically. And Roeder and Wasserman[7] have shown that when a normal mixture model is used to estimate a density 'nonparametrically', the density estimate that uses BIC to select the number of components in the mixture is consistent. They also reported a simulation study in which BIC performed very well. Also, Campbell et al.[78] and Dasgupta and Raftery[79] have reported encouraging results for BIC applied to mixture models.

More recently, under certain conditions, Keribin[80] has shown that BIC performs consistently

in choosing the true number of components in a mixture model; see also Drton[81] who has derived a modified criterion called BICS for singular models.

The criterion BIC has been derived also by Rissanen[82,83] from another perspective based on coding theory. Also, criteria based on the approximation (33) are very similar to the criterion based on the Minimum Message Length (MML) principle Wallace and Freeman[84] and Wallace and Dowe,[85] whereby g is chosen to minimize the minimum message length. For further discussion on the MML and related approaches to the choice of the number of components, the reader is referred to Figueiredo and Jain[86] who have given an excellent account of these approaches.

We have presented BIC in a non-Bayesian framework in the above. Roeder and Wasserman[7] used it in a Bayesian framework to construct an estimate of $\text{pr}\{g|\boldsymbol{y}_T\}$.

Choosing a penalty term for the log likelihood function is a challenging but clearly a crucial problem. An alternative to well-known penalized criteria with fixed penalties such as AIC and BIC, Baudry et al.[87] changed to considered slope heuristics for the choice of the number of components in a mixture model, using the ideas of Birgé and Massart.[88,89]

CLASSIFICATION-BASED INFORMATION CRITERIA

We consider now some criteria that have been developed by consideration either from a frequentist or Bayesian perspective of the so-called classification likelihood $L_c(\boldsymbol{\Psi}_g)$, which is the complete-data likelihood within the EM framework for the fitting of a mixture model.

Classification Likelihood Criterion

Biernacki and Govaert[90] made use of the relationship linking the likelihood $L(\boldsymbol{\Psi}_g)$ for the mixture model and the complete-data likelihood $L_c(\boldsymbol{\Psi}_g)$ to propose a criterion for selecting the number of clusters arising from the fitting of a normal mixture model. In the EM framework, the complete-data log likelihood function $\log L_c(\boldsymbol{\Psi}_g)$ is given by the log of the likelihood function $L_c(\boldsymbol{\Psi}_g)$ formed on the basis of the complete-data; that is, on the basis of the observed data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ and the unobservable component labels given by $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$, where $z_{ij} = (\boldsymbol{z}_j)_i$ is one or zero according as to whether $\boldsymbol{y}_j$ belongs to the ith component of the mixture $(i = 1, \ldots, g; j = 1, \ldots, n)$. For a normal mixture model, we have that

$$\log L_c\left(\boldsymbol{\Psi}_g\right) = \sum_{i=1}^{g}\sum_{j=1}^{n} \times z_{ij}\left\{\log \pi_i - \frac{p}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_i| - \frac{1}{2}\left(\boldsymbol{y}_j - \mu_i\right)^T \Sigma_i^{-1}\left(\boldsymbol{y}_j - \mu_i\right)\right\}. \quad (37)$$

Although $L_c(\boldsymbol{\Psi}_g)$ is referred to as the complete-data likelihood within the EM framework, it is sometimes called the classification likelihood in a classification context.[91]

As noted by Hathaway,[92] among others, we can express the mixture log likelihood, $\log L(\boldsymbol{\Psi}_g)$, as

$$\log L\left(\boldsymbol{\Psi}_g\right) = \log L_c\left(\boldsymbol{\Psi}_g\right) - \log k\left(\boldsymbol{\Psi}_g\right), \quad (38)$$

where

$$\log k\left(\boldsymbol{\Psi}_g\right) = \sum_{i=1}^{g}\sum_{j=1}^{n} z_{ij}\log \tau_{ij}$$

and where $\tau_{ij} = \tau_i(\boldsymbol{y}_j; \boldsymbol{\Psi}_g)$ is the posterior probability of ith component membership defined by Eq. (5). That is, $k(\boldsymbol{\Psi}_g)$ is the conditional density of the vector of component-indicator variables

$$\boldsymbol{z} = \left(\boldsymbol{z}_1^T, \ldots, \boldsymbol{z}_n^T\right)^T,$$

given the observed data $\boldsymbol{y}_T = \left(\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T\right)^T$.

The conditional mean of $\log k(\boldsymbol{\Psi}_g)$ given the observed data $\boldsymbol{y}$ is equal to $-EN(\tau)$, where

$$EN(\tau) = -\sum_{i=1}^{g}\sum_{j=1}^{n}\tau_{ij}\log \tau_{ij}$$

is the entropy of the fuzzy classification matrix $\boldsymbol{C} = ((\tau_{ij}))$ and where

$$\tau = \left(\tau_1^T, \ldots, \tau_n^T\right)^T, \quad (39)$$

and

$$\tau_j = (\tau_1\left(\boldsymbol{y}_j; \boldsymbol{\Psi}_g\right), \ldots, \tau_g(\boldsymbol{y}_j; \boldsymbol{\Psi}_g))^T \quad (40)$$

is the vector of posterior probabilities of component membership of $\boldsymbol{y}_j (j = 1, \ldots, n)$. We now write the complete-data likelihood as $L_c(\boldsymbol{\Psi}_g; \boldsymbol{z})$ to explicitly denote that it is formed on the basis of $\boldsymbol{z}$ containing the component indicators, in addition to $\boldsymbol{y}$. Then it follows from Eq. (38) that if we put $\boldsymbol{z} = \hat{\tau}$ in $L_c(\boldsymbol{\Psi}_g; \boldsymbol{z})$, we have

$$\log L_c\left(\widehat{\Psi}_g; \hat{\tau}\right) = \log L\left(\widehat{\Psi}_g\right) - EN(\tau), \quad (41)$$

where τ is the MLE of τ formed by replacing τ_{ij} by

$$\hat{\tau}_{ij} = \tau_i\left(\boldsymbol{y}_j; \widehat{\boldsymbol{\Psi}}_g\right) \quad (i = 1, \ldots, g; j = 1, \ldots, n) \quad (42)$$

in Eq. (40). From Eq. (41), we can form the classification likelihood information criterion (CLC), where g is chosen to minimize

$$-2\log L\left(\hat{\Psi}_g\right)+2EN\left(\hat{\tau}\right), \tag{43}$$

where the estimated entropy $EN(\tau)$ is used as the term that penalizes the model for its complexity.

If the components of the mixture are well separated, then $EN\left(\hat{\tau}\right)$ will be close to its minimum value of zero. But if the mixture components are poorly separated, then $EN\left(\hat{\tau}\right)$ will have a large value. Hence how severely this criterion penalizes the log likelihood depends on how well separated the fitted components are. According to Biernacki et al.,[76] this criterion works well when the mixing proportions are restricted to being equal. But it tends to overestimate the correct number of clusters when no restriction is placed on the mixing proportions.

Banfield and Raftery[93] suggested an approximate Bayesian solution to the choice of the number of clusters using the classification ML approach. Their approximation, which is a crude approximation to twice the log Bayes factor for g clusters, leads to the approximate weight of evidence (AWE) criterion having the form

$$AWE(g)=-2\log L_c\left(\hat{\Psi}_g;\hat{z}\right)+2d\left(3/2+\log n\right).$$

When the mixture components are well separated, we have seen above that $L_c\left(\hat{\Psi}_g\right)\approx L\left(\hat{\Psi}_g\right)$, and thus it can then be expected to be similar to BIC. When the clusters are not well separated, it has been noted that the classification likelihood approach to model fitting leads to severely biased estimates of the parameters.[91]

Normalized Entropy Criterion

Celeux and Soromenho[63] proposed using the estimated entropy $EN\left(\hat{\tau}\right)$ (after normalization) as a criterion in its own right for choosing the number of clusters. This criterion is known as the normalized entropy criterion (NEC). The estimated entropy $EN\left(\hat{\tau}\right)$ cannot be used directly to assess the number of components in a mixture model, since $\log L\left(\hat{\Psi}_g\right)$ is an increasing function of g. The normalized form is given by

$$NEC(g)=\frac{EN\left(\hat{\tau}\right)}{\log L\left(\hat{\Psi}_g\right)-\log L\left(\hat{\Psi}_g^*\right)}, \tag{44}$$

where $\hat{\Psi}_g^*$ denotes the MLE of Ψ_g in the case of a single $(g=1)$ component. The entropy for $g=1$ is zero. As it stands, this criterion is unable to decide between $g=1$ and a value of g greater than one. Celeux and Soromenho[63] proposed a rule of thumb, but their procedure was restricted to normal mixtures and had performed disappointingly.[76] In the latter paper, a general procedure was proposed to deal with this problem. Effectively, they define $NEC(g)$ to be one for $g=1$. The modified criterion simply then consists of choosing g to minimize $NEC(g)$. According to Biernacki et al.[76] this improved version of the NEC criterion corrects for the tendency of the original version to prefer $g>1$ clusters when the true number is one.

A similar type criterion is the partition coefficient (PC) of Bezdek,[94] where

$$PC(g)=\sum_{i=1}^{g}\sum_{j=1}^{n}\hat{\tau}_{ij}^2.$$

Numerical experiments reported by Windham and Cutler[69] clearly show that the PC criterion tends to underestimate the order of the mixture model.

Integrated Classification Likelihood Criterion

As noted above, so far as assessing the number of clusters, it has been observed that BIC tends to favor models with enough components in order to provide a good estimate of the mixture density. This led Biernacki et al.[95] to develop the integrated classification criterion (ICL). An approximation to this criterion is given by

$$-2\log L\left(\hat{\Psi}_g\right)+d\log n+EN\left(\hat{\tau}\right), \tag{45}$$

where $EN\left(\hat{\tau}\right)$ is the entropy of the fuzzy classification matrix $\left(\left(\hat{\tau}_i\left(\boldsymbol{y}_j\right)\right)\right)$. That is, the ICL criterion uses the entropy term $EN\left(\hat{\tau}\right)$ to penalize the model for its complexity (too many components and hence clusters).

Another approach to refining the number of clusters has been given recently by Baudry et al.,[96] who have suggested a way in which the components can be recombined. Also, Hennig[97] has considered ways of merging components of a Gaussian mixture model where the fitted components are not separated enough from each other to be interpreted as 'clusters'.

OTHER METHODS FOR ASSESSING THE NUMBER OF CLUSTERS

Given that we are assuming a normal mixture model, it is reasonable to make use of this assumption in

addressing the question of the number of components in the model. In situations where it is realistic to assume that the number of components reflects the number of clusters in the data, one might wish to consider nonparametric methods for selecting the number of clusters.[98] Most these methods are based on scalar functions of the between-cluster and within-cluster sums of squares and products matrices; see, for example, Edwards and Cavalli-Sforza,[99] Marriott,[100] Calinski and Harabasz,[101] Hartigan,[102] and Krzanowski and Lai.[103] Other procedures include the silhouette statistic proposed by Kaufman and Rousseeuw,[104] the gap statistic proposed by Tibshirani et al.,[105] the jump statistic of Sugar and James,[106] and cluster-stability based methods such as that proposed by Fang and Wang.[107]

Another approach to the question of the number of clusters is to use a mode seeking method. However, some caution should be exercised in this context as Ray and Ren[108] showed that a two-component normal mixture model can have as many as $p+1$ modes in p dimensions.

CONCLUSION

We have considered the problem of assessing the number of components in a normal mixture model. Attention is given to the breakdown in the usual regularity conditions for the LRTS on the number of components to have its usual asymptotic null distribution of chi-squared. A brief account is given of available results for the large-sample behavior of the LRTS including modifications adopted in order to derive a limiting distribution. For practical applications, the focus is on a resampling approach via the parametric bootstrap and by information-based methods such as BIC.

REFERENCES

1. McLachlan GJ, Peel D. *Finite Mixture Models*. New York: John Wiley & Sons; 2000.
2. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J R Stat Soc Ser B Methodol* 1977, 39:1–38.
3. Li JQ, Barron AR. Mixture density estimation. Technical Report. New Haven, CT: Department of Statistics, Yale University; 2000.
4. Schwarz G. Estimating the dimension of a model. *Ann Stat* 1978, 6:461–464.
5. Leroux BG. Consistent estimation of a mixing distribution. *Ann Stat* 1992, 20:1350–1360.
6. Akaike H. A new look at the statistical model identification. *IEEE Trans Automat Control* 1974, 19:716–723.
7. Roeder K, Wasserman L. Practical Bayesian density estimation using mixtures of normals. *J Am Stat Assoc* 1997, 92:894–902.
8. Basford K, McLachlan G. Modelling the distribution of stamp paper thickness via finite normal mixtures: the 1872 Hidalgo stamp issue of Mexico revisited. *J Appl Stat* 1997, 248:169–179.
9. Lee S, McLachlan G. Finite mixtures of multivariate skew *t*-distributions: some recent and new results. *Stat Comput* 2014, 24:181–202.
10. Pyne S, Hu X, Wang K, Rossin E, Lin TI, Maier L, Baecher-Allan C, McLachlan G, Tamayo P, Hafler D, et al. Automated high-dimensional flow cytometric data analysis. *Proc Natl Acad Sci U S A* 2009, 106:8519–8524.
11. Richardson S, Green PJ. On bayesian analysis of mixtures with an unknown number of components (with discussion). *J R Stat Soc Ser B Methodol* 1997, 59:731–792.
12. Cramér H. *Mathematical Methods of Statistics*. Princeton, NJ: Princeton University Press; 1946.
13. Aitkin M, Rubin DB. Estimation and hypothesis testing in finite mixture models. *J R Stat Soc Ser B Methodol* 1985, 47:67–75.
14. Quinn BG, McLachlan GJ, Hjort NL. A note on the Aitkin-Rubin approach to hypothesis testing in mixture models. *J R Stat Soc Ser B Methodol* 1987, 49:311–314.
15. Ghosh JK, Sen PK. On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results. In: LeCam L, Olshen R, eds. *Proceedings of the Berkeley Conference in Honor of J. Neyman and J. Kiefer*, vol. 2. Monterey, CA: Wadsworth; 1985, 789–806.
16. Titterington DM. Contribution to the discussion of paper by M. Aitkin, D. Anderson and J. Hinde. *J R Stat Soc Ser A Gen* 1981, 144:459.
17. Titterington DM, Smith AFM, Makov UE. *Statistical Analysis of Finite Mixture Distributions*. New York: John Wiley & Sons; 1985.
18. Lindsay BG. Mixture models: theory, geometry and applications. In: *NSF-CBMS Regional Conference Series in Probability and Statistics*, vol. 5. Alexandria, VA: Institute of Mathematical Statistics and the American Statistical Association; 1995.

19. Hartigan JA. Statistical theory in clustering. *J Classif* 1985, 2:63–76.
20. Hartigan JA. A failure of likelihood asymptotics for normal mixtures. In: LeCam L, Olshen R, eds. *Proceedings of the Berkeley Conference in Honor of J. Neyman and J. Kiefer*, vol. 2. Monterey, CA: Wadsworth; 1985, 807–810.
21. Bickel PJ, Chernoff H. Asymptotic distribution of the likelihood ratio statistic in a prototypical non regular problem. In: Ghosh JK, Mitra SK, Parthasarathy KR, Prakasa Rao BLS, eds. *Statistics and Probability: A Raghu Raj Bahadur Festschrift*. New Delhi: Wiley Eastern; 1993, 83–96.
22. Dacunha-Castelle D, Gassiat E. The estimation of the order of a mixture model. *Bernoulli* 1997, 3:279–299.
23. Chen H, Chen J. The likelihood ratio test for homogeneity in finite mixture models. *Can J Stat* 2001, 29:201–215.
24. Liu X, Shao Y. Asymptotics for the likelihood ratio test in a two-component normal mixture model. *J Stat Plann Inference* 2004, 123:61–81.
25. Garel B. Asymptotic theory of the likelihood ratio test for the identification of a mixture. *J Stat Plann Inference* 2005, 131:271–296.
26. Ruck A. Calculating the (asymptotic) distribution of the log-LRT statistic in a contamination mixture model. Discussion papers in statistics and quantitative economics, Univ. der Bundeswehr, Fachberich Wirtschafts- und Organisationswiss; 2001.
27. Ruck A. Calculating the asymptotic distribution of the log-LRT statistic for testing one against two populations in normal mixtures. Discussion papers in statistics and quantitative economics, Univ. der Bundeswehr, Fachbereich Wirtschafts- und Organisationswiss; 2002.
28. Seidel W, Mosler K, Alker M. A cautionary note on likelihood ratio tests in mixture models. *Ann Inst Stat Math* 2000, 52:481–487.
29. Seidel W, Ševčíková H, Sever K. Testing against nonparametric alternatives in mixture models. *J Comput Graph Stat* 2007, 16:350–377.
30. Lo Y, Mendell NR, Rubin DB. Testing the number of components in a normal mixture. *Biometrika* 2001, 88:767–778.
31. Lo Y. Likelihood ratio tests of the number of components in a normal mixture with unequal variances. *Stat Probab Lett* 2005, 71:225–235.
32. Lo Y. A likelihood ratio test of a homoscedastic normal mixture against a heteroscedastic normal mixture. *Stat Comput* 2008, 18:233–240.
33. Hall P, Stewart M. Theoretical analysis of power in a two-component normal mixture model. *J Stat Plann Inference* 2005, 134:158–179.
34. Jeffries NO. A note on 'Testing the number of components in a normal mixture'. *Biometrika* 2003, 90:991–994.
35. Chen H, Chen J, Kalbfleisch JD. A modified likelihood ratio test for homogeneity in finite mixture models. *J R Stat Soc Ser B Methodol* 2001, 63:19–29.
36. Chen H, Chen J, Kalbfleisch JD. Testing for a finite mixture model with two components. *J R Stat Soc Ser B Methodol* 2004, 66:95–115.
37. Li P, Chen J, Marriott P. Non-finite fisher information and homogeneity: an em approach. *Biometrika* 2009, 96:411–426.
38. Chen J, Li P. Hypothesis test for normal mixture models: the EM approach. *Ann Stat* 2009, 37:2523–2542.
39. Li P, Chen J. Testing the order of a finite mixture. *J Am Stat Assoc* 2010, 105:1084–1092.
40. Chen J, Li P, Fu Y. Inference on the order of a normal mixture. *J Am Stat Assoc* 2012, 107:1096–1105.
41. Efron B. Bootstrap methods: another look at the Jackknife. *Ann Stat* 1979, 7:1–26.
42. Efron B. *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia, PA: SIAM; 1981.
43. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. New York: Chapman & Hall; 1994.
44. Aitkin M, Anderson D, Hinde J. Statistical modelling of data on teaching styles. *J R Stat Soc Ser A Gen* 1981, 144:419–461.
45. McLachlan GJ. On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Appl Stat* 1987, 36:318–324.
46. Hoaglin DC, Mosteller F, Tukey JW. Using quantiles to study shape. In: Hoaglin DC, ed. *Exploring Data Tables, Trends, and Shape*. New York: John Wiley & Sons; 2006, 417–460.
47. Smyth P. Model selection for probabilistic clustering using cross-validated likelihood. *Stat Comput* 2000, 10:63–72.
48. Barnard GA. Contribution to the discussion of paper by M.S. Bartlett. *J R Stat Soc Ser B Methodol* 1963, 25:294.
49. Hope ACA. A simplified Monte Carlo significance test procedure. *J R Stat Soc Ser A Gen* 1968, 30:582–598.
50. Mclachlan GJ, Peel D. On a resampling approach to choosing the number of components in normal mixture models. In: Billard L, Fisher N, eds. *Computing Science and Statistics*, vol. 28. Fairfax Station, VA: Interface Foundation of North America; 1997, 260–266.
51. McLachlan GJ, Bean RW, Peel D. A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics* 2002, 18:413–422.
52. McLachlan GJ, Bean RW, Jones LBT. A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics* 2006, 22:1608–1615.
53. McLachlan GJ, Rathnayake SI. Testing for group structure in high-dimensional data. *J Biopharm Stat* 2011, 21:1113–1125.

54. Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci U S A* 2002, 99:6562–6566.
55. Baek J, McLachlan G, Flack L. Mixtures of factor analyzers with common factor loadings: applications to the clustering and visualisation of high-dimensional data. *IEEE Trans Pattern Anal Mach Intell* 2010, 32:1298–1309.
56. Kullback S, Leibler RA. On information and sufficiency. *Ann Math Stat* 1951, 22:79–86.
57. Akaike H. Information theory and an extension of the maximum likelihood principle. In: Parzen E, Tanabe K, Kitagawa G, eds. *Selected Papers of Hirotugu Akaike*. Springer Series in Statistics. New York: Springer; 1998, 199–213.
58. Bozdogan H, Sclove SL. Multi-sample cluster analysis using Akaike's information criterion. *Ann Inst Stat Math* 1984, 36:163–180.
59. Sclove SL. Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika* 1987, 52:333–343.
60. Konishi S, Kitagawa G. Generalised information criteria in model selection. *Biometrika* 1996, 83:875–890.
61. Koehler AB, Murphree ES. A comparison of the Akaike and Schwarz criteria for selecting model order. *J R Stat Soc Ser C Appl Stat* 1988, 37:187–195.
62. Soromenho G. Comparing approaches for testing the number of components in a finite mixture model. *Comput Stat* 1993, 9:65–78.
63. Celeux G, Soromenho G. An entropy criterion for assessing the number of clusters in a mixture model. *J Classif* 1996, 13:195–212.
64. Bozdogan H. On the information-based measure of covariance complexity and its application to the evaluation of multivariate linear models. *Commun Stat Theory Methods* 1990, 19:221–278.
65. Bozdogan H. Choosing the number of component clusters in the mixture-model using a new informational complexity criterion of the inverse-fisher information matrix. In: Opitz O, Lausen B, Klar R, eds. *Information and Classification*. Studies in Classification, Data Analysis and Knowledge Organization. Berlin/Heidelberg: Springer; 1993, 40–54.
66. Ishiguro M, Sakamoto Y, Kitagawa G. Bootstrapping log likelihood and EIC, an extension of AIC. *Ann Inst Stat Math* 1997, 49:411–434.
67. Pan W. Bootstrapping likelihood for model selection with small samples. *J Comput Graph Stat* 1999, 8:687–698.
68. McLachlan GJ, Krishnan T. *The EM Algorithm and Extensions*. 2nd ed. Hoboken, NJ: John Wiley & Sons; 2008.
69. Windham MP, Cutler A. Information ratios for validating mixture analyses. *J Am Stat Assoc* 1992, 87:1188–1192.
70. Polymenis A, Titterington DM. On the determination of the number of components in a mixture. *Stat Probab Lett* 1998, 38:295–298.
71. Cheeseman P, Stutz J. Bayesian classification (autoclass): theory and results. In: *Advances in Knowledge Discovery and Data Mining*. Menlo Park, CA: American Association for Artificial Intelligence; 1996, 153–180.
72. Dudoit S, Fridlyand J. A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biol* 2002, 3:0036.1–0036.21.
73. Kass RE, Raftery AE. Bayes factors. *J Am Stat Assoc* 1995, 90:773–795.
74. Aitken M. *Statistical Inference: An Integrated Bayesian/Likelihood Approach*. Boca Raton, FL: Chapman and Hall; 2010.
75. Ripley BD. *Pattern Recognition and Neural Networks*. Cambridge, UK: Cambridge University Press; 1996.
76. Biernacki C, Celeux G, Govaert G. An improvement of the NEC criterion for assessing the number of clusters in a mixture model. *Non-Linear Anal* 1999, 20:267–272.
77. Fraley C, Raftery AE. How many clusters? Which clustering method? Answers via model-based cluster analysis. *Comput J* 1998, 41:578–588.
78. Campbell JG, Fraley C, Murtagh F, Raftery AE. Linear flaw detection in woven textiles using model-based clustering. *Pattern Recognit Lett* 1997, 18:1539–1548.
79. Dasgupta A, Raftery AE. Detecting features in spatial point processes with clutter via model-based clustering. *J Am Stat Assoc* 1998, 93:294–302.
80. Keribin C. Consistent estimation of the order of mixture models. *Sankhya Indian J Stat Ser A* 2000, 62:49–66.
81. Drton M. Likelihood ratio tests and singularities. *Ann Stat* 2009, 37:979–1012.
82. Rissanen J. Stochastic complexity and modeling. *Ann Stat* 1986, 14:1080–1100.
83. Rissanen J. *Stochastic Complexity in Statistical Inquiry*. Singapore: World Scientific; 1989.
84. Wallace CS, Freeman PR. Estimation and inference by compact coding. *J R Stat Soc Ser B Methodol* 1987, 49:240–265.
85. Wallace CS, Dowe DL. MML clustering of multi-state, Poisson, von Mises circular and Gaussian distributions. *Stat Comput* 2000, 10:73–83.
86. Figueiredo MAT, Jain AK. Unsupervised learning of finite mixture models. *IEEE Trans Pattern Anal Mach Intell* 2002, 24:381–396.
87. Baudry JP, Maugis C, Michel B. Slope heuristics: overview and implementation. *Stat Comput* 2012, 22:455–470.
88. Birgé L, Massart P. Gaussian model selection. *J Eur Math Soc* 2001, 3:203–268.

89. Birgé L, Massart P. Minimal penalties for gaussian model selection. *Probab Theory Relat Fields* 2007, 138:33–73.
90. Biernacki C, Govaert G. Using the classification likelihood to choose the number of clusters. *Comput Sci Stat* 1997, 29:451–457.
91. McLachlan GJ. The classification and mixture maximum likelihood approaches to cluster analysis. In: Krishnaiah P, Kanal L, eds. *Handbook of Statistics*. Amsterdam: North-Holland; 1982, 199–208.
92. Hathaway RJ. Another interpretation of the EM algorithm for mixture distributions. *Stat Probab Lett* 1986, 4:53–56.
93. Banfield JD, Raftery AE. Model-based Gaussian and non-Gaussian clustering. *Biometrics* 1993, 49:803–821.
94. Bezdek JC. *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York: Plenum Press; 1981.
95. Biernacki C, Celeux G, Govaert G. Assessing a mixture model for clustering with the integrated classification likelihood. *IEEE Trans Pattern Anal Mach Intell* 1998, 22:719–725.
96. Baudry JP, Raftery AE, Celeux G, Lo K, Gottardo R. Combining mixture components for clustering. *J Comput Graph Stat* 2009, 9:323–353.
97. Hennig C. Methods for merging gaussian mixture components. *Adv Data Anal Classif* 2010, 4:3–34.
98. McLachlan GJ, Basford KE. *Mixture Models: Inference and Applications to Clustering*. New York: Marcel Dekker; 1988.
99. Edwards AWF, Cavalli-Sforza LL. A method for cluster analysis. *Biometrics* 1965, 21:362–375.
100. Marriott FHC. Practical problems in a method of cluster analysis. *Biometrics* 1971, 27:501–514.
101. Calinski T, Harabasz J. A dendrite method for cluster analysis. *Commun Stat* 1974, 3:1–27.
102. Hartigan JA. *Clustering Algorithms*. New York: John Wiley & Sons; 1975.
103. Krzanowski WJ, Lai YT. A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics* 1988, 44:23–34.
104. Kaufman L, Rousseeuw PJ. *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley & Sons; 1990.
105. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. *J R Stat Soc Ser B Methodol* 2001, 63: 411–423.
106. Sugar CA, James GM. Finding the number of clusters in a dataset: an information-theoretic approach. *J Am Stat Assoc* 2003, 98:750–763.
107. Fang Y, Wang J. Selection of the number of clusters via the bootstrap method. *Comput Stat Data Anal* 2012, 56:468–477.
108. Ray S, Ren D. On the upper bound of the number of modes of a multivariate normal mixture. *J Multivar Anal* 2012, 108:41–52.

