Accepted Manuscript

Comparing classical criteria for selecting intra-class correlated features in Multimix

Lynette A. Hunt, Kaye E. Basford

PII: S0167-9473(16)30130-X
DOI: http://dx.doi.org/10.1016/j.csda.2016.05.018
Reference: COMSTA 6280

To appear in: *Computational Statistics and Data Analysis*

Received date: 1 May 2015
Revised date: 26 May 2016
Accepted date: 27 May 2016

Please cite this article as: Hunt, L.A., Basford, K.E., Comparing classical criteria for selecting intra-class correlated features in Multimix. *Computational Statistics and Data Analysis* (2016), http://dx.doi.org/10.1016/j.csda.2016.05.018

Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 00 (2016) 1–21
csdalogo
~~Computational~~ Statistics & Data ~~Analysis~~

Comparing classical criteria for selecting intra-class correlated features in Multimix

Lynette A. Hunt[a,*], Kaye E. Basford[b]

[a] *University of Waikato, Hamilton, New Zealand.*
[b] *University of Queensland, Brisbane, Australia.*

Abstract

The mixture approach to clustering requires the user to specify both the number of components to be fitted to the model and the form of the component distributions. In the Multimix class of models, the user also has to decide on the correlation structure to be introduced into the model. The behaviour of some commonly used model selection criteria is investigated when using the finite mixture model to cluster data containing mixed categorical and continuous attributes. The performance of these criteria in selecting both the number of components in the model and the form of the correlation structure amongst the attributes when fitting the Multimix class of models is illustrated using simulated data and a real medical data set. It is found that criteria based on the integrated classification likelihood have the best performance in detecting the number of clusters to be fitted to the model and in selecting the form of the component distributions. The performance of the Bayesian information criterion in detecting the correct model depends on the partitioning structure among the attributes while the Akaike information criterion and classification likelihood criterion perform in a less satisfactory way.

Keywords: model selection criteria, finite mixture models, mixed data, Multimix

1. Introduction

Finite mixture models are widely used in a variety of applications to model the distributions of various events and to cluster data sets, see for example McLachlan and Peel (2000), McLaclan and Chang (2004), Everitt *et al* (2011), Stahl and Sallis (2012) and Melnykov (2013). This paper focuses on the use of the mixture model approach to clustering which provides a formal statistical framework on which the clustering can be based. The procedure gives a probabilistic clustering that allows for overlapping clusters which correspond to the components in the model, and where each component in the finite mixture model corresponds to a cluster in the data. The probability that an observation belongs to each of the clusters can be obtained from the estimates of the posterior probabilities of cluster membership. A definitive partitioning of the observations into components (clusters) is obtained by assigning each observation to the component to which it has highest probability of belonging. The finite mixture model requires the specification of both the form of the density function of each of the underlying components and the number of components to be fitted in the model.

*Corresponding author
Email addresses: lah@waikato.ac.nz (Lynette A. Hunt), k.e.basford@uq.edu.au (Kaye E. Basford)

1.1. Number of components

Prior knowledge concerning the number of components, K, in the mixture reduces the complexity of the analysis. However there are many situations where there is no *a priori* knowledge of the number of components to be fitted, and thus finding the number of components present in the data becomes part of the clustering problem.

An obvious way of approaching this problem is to use the likelihood ratio statistic λ to test for the smallest value of K compatible with the data. However when testing for the number of components in a mixture, the usual regularity conditions do not hold for $-2\log\lambda$ to have its standard asymptotic null distribution of χ^2 with the degrees of freedom equal to the difference between the number of parameters under the full and reduced models. Accounts of the breakdown of the regularity conditions are given for example, by Hartigan (1977, 1985a, b), Titterington (1981), Titterington, Smith and Makov (1985), Ghosh and Sen (1985), McLachlan and Basford (1988), and McLachlan and Peel (2000).

An alternative procedure is to use a bootstrap approach. McLachlan (1987) proposed a resampling procedure that involves a bootstrapped likelihood ratio test. Bootstrap samples are generated from the finite mixture model fitted under the null hypothesis of K components, where the parameters of the mixture are the likelihood estimates after fitting a K component model to the original sample. The value of the likelihood ratio statistic is computed for each of the bootstrap samples generated after fitting mixtures with K and K' components, where $K' > K$. The process is repeated independently B times. The replicated values of $-2\log\lambda$ formed from the successive bootstrap samples provide an assessment of the true null distribution of $-2\log\lambda$: see also Feng and McCulloch (1996), and McLachlan and Peel (1997). However, the problem with bootstrap methods is that they can be computationally intensive when the number of components is large, and little is known about the performance of the test when the distributional and model assumptions are violated (see Nyland *et al* 2007). See also Lo *et al* (2001) for details on another approach called the Lo−Mendell−Rubin likelihood ratio test which uses an approximation of the distribution of the difference of the two log likelihoods.

The use of information criteria to estimate the number of components of a finite mixture has become increasingly popular in model based cluster analysis. Information criteria allow the user to quantify the differences between a candidate set of models and help determine the number of components to be fitted to the mixture model. Many criteria have been proposed with some criteria derived within a Bayesian framework. The authors of this papaer have used criteria that are Bayesian based, information criteria and classification criteria. See for example, McLachlan and Peel (2000, chapter 6), Frayley and Raftery(2002), Miloslavski and Van der Laan (2003), McLachlan and Rathnayake (2014) plus the references therein for discussions on other approaches to the problem of determining the number of components.

The specification of the component distributions is also required in the fitting of a mixture model. There has been extensive use of mixtures where the component distributions are multivariate normal and there has been much interest in determining the number of components to be fitted to this model. McLachlan and Ng (2000) report Monte Carlo simulations to compare the performance of some criteria with that of classical criteria such as the Akaike information criterion (AIC) and the Bayesian information criterion (BIC), for determining the number of components in mixtures of multivariate normals.

1.2. Covariance structure for models

While it is common to take the component distributions to be multivariate normal, decisions still need to be made on the structure of the components' covariance matrices. There is the unrestricted case where the component covariances Σ_k are unequal, however this may be too general for many situations in practice. Often the component covariances are restricted to being the same ($\Sigma_k = \Sigma$ for $k = 1,\ldots,K$), but this can have an adverse effect on the resulting clustering (Chapter 3, McLachlan and Peel, 2000).

Another way of proceeding is to adopt some model that is intermediate between homoscedasity and the general unrestricted heteroscedastic case. Several authors (Banfield and Raftery, 1993; Celeux and Govaert, 1995; Bensmail, Celeux, Raftery and Robert, 1997) have used the eigenvalue decomposition of the component covariance matrices in Gaussian mixtures to propose models for clustering. The covariance matrix Σ_k can be written in the form $\Sigma_k = \lambda_k D_k A_k D_k'$ where D_k is the matrix of eigenvectors of Σ_k, A_k is a diagonal matrix with the normalised eigenvectors of Σ_k on the diagonal in decreasing order with $|A_k| = 1$, and $\lambda_k = |\Sigma_k|^{\frac{1}{d}}$ where d denotes the number of variables. The volume, orientation and shape of the k^{th} component are determined by λ_k, D_k and A_k respectively. Celeux and Govaert

(1995) and Bensmail and Celeux (1996) consider 14 different models corresponding to different assumptions on the components' covariance matrices.

Biernacki and Govaert (1999) performed Monte Carlo simulations using two component bivariate Gaussian mixtures with different covariance matrices to compare the performance of several classical criteria in selecting a relevant and parsimonious model. The covariance matrices for the component distributions were determined using the 14 models relating to different assumptions on the component covariance matrix. They performed simulations using small ($n = 40$) and larger ($n = 200$) samples where the components were mixed in both equal and different proportions.

Hunt (1996) and Hunt and Jorgensen (1999) proposed a set of models that they termed the Multimix class of mixture models. The Multimix approach uses a form of conditional independence within the components, and can be used to cluster data containing both categorical and continuous attributes. When using the Multimix approach to clustering, Hunt (1996) suggested that a form of forward selection of covariates be used for selecting the correlation structure in the model.

Galimberti and Soffritti (2013) also used conditional independence within the components when clustering using the finite mixture model. These authors' approach imposed constaints on the component covariance matrices and only applied to quantitative attributes. Their approach cannot cope with attributes that are categorical.

The behaviour of some commonly used model selection criteria is investigated when the finite mixture model is used to cluster mixed data with categorical and continuous attributes. In Section 2, the mixture model framework for clustering is reviewed and the differences between the likelihood and the classification likelihood is emphasized. The Multimix approach is reviewed in Section 3, and Section 4 examines the criteria that will be used to assess the fitted models. In Section 5, the performance of the criteria is assessed using data simulated from multivariate normal and discrete distributions and a real medical data set.

2. The Mixture model approach to clustering data

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be the observed values of a random sample from a mixture of K underlying populations in unknown proportions $\pi_1, \ldots, \pi_K$, and where $0 < \pi_k < 1$, for $k = 1, \ldots, K$ and $\sum_{k=1}^{K} \pi_k = 1$. The density of the i^{th} observation $\mathbf{x}_i$ for $i = 1, \ldots, n$, in the sample is a $p \times 1$ vector, that can be represented as the finite mixture

$$f(\mathbf{x}_i; \boldsymbol{\phi}) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{x}_i; \boldsymbol{\theta}_k,) \tag{1}$$

where the vector of unknown parameters $\boldsymbol{\phi} = (\boldsymbol{\theta}', \boldsymbol{\pi}')'$, for $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)'$, and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)'$, and where $f_k(\mathbf{x}_i; \boldsymbol{\theta}_k)$ is the density of $\mathbf{x}_i$ in component k, and $\boldsymbol{\theta}_k$ is the parameter vector for component k. For clustering purposes, each component in the mixture model corresponds to a cluster.

2.1. Likelihood approach

The log-likelihood function for $\boldsymbol{\phi}$ can be formed from the observed data by

$$\log L(\boldsymbol{\phi}) = \sum_{i=1}^{n} \log f(\mathbf{x}_i; \boldsymbol{\phi}) = \sum_{i=1}^{n} \log\{\sum_{k=1}^{K} \pi_k f_k(\mathbf{x}_i; \boldsymbol{\theta}_k)\}, \tag{2}$$

where $L(\boldsymbol{\phi})$ denotes the likelihood for $\boldsymbol{\phi}$.

The maximum likelihood estimate $\hat{\boldsymbol{\phi}}$ of $\boldsymbol{\phi}$ is the global maximiser of the likelihood function assuming that this function is bounded, and is obtained as a solution of the log-likelihood (score) equation given by $\frac{\partial}{\partial \boldsymbol{\phi}} \log L(\boldsymbol{\phi}) = 0$. In the case where the likelihood is unbounded, the estimator of $\boldsymbol{\phi}$ can be taken to be an appropriate root of the likelihood equation, corresponding to a local maximum. In this situation, $\hat{\boldsymbol{\phi}}$ is usually taken to correspond to the largest of the maxima located. See, for example, McLachlan and Peel (2000) for a discussion on maximum likelihood estimation for mixture models.

The EM algorithm of Dempster, Laird and Rubin (1977) is used for the computation of the maximum likelihood estimates of $\boldsymbol{\phi}$ by viewing the data as incomplete (see for example, McLaclan and Peel (2000), McLachlan and

Krishnan (2008)). For the finite mixture model, the 'missing' data $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})'$, are the unobserved indicators of component membership defined by

$$z_{ik} = \begin{cases} 1 & \text{if individual } i \in \text{component } k; \\ 0 & \text{if individual } i \notin \text{component } k, \end{cases}$$

where $\mathbf{z}_i, i = 1, \ldots, n$, are independently and identically distributed according to a multinomial distribution generated by a single trial of an experiment with K mutually exclusive outcomes having probabilities $\pi_1, \ldots, \pi_K$. The complete data, in EM terminology, consists of the $n \times p$ array of observed data $\mathbf{x}_1, \ldots, \mathbf{x}_n$, and the conceptual $n \times K$ array $\{z_{ik}\}$ of class membership indicators. The EM algorithm iterates between the E step and the M step until convergence of the likelihood.

The posterior probability that an observation with measurements $\mathbf{x}_i$ belongs to the k^{th} component is given by

$$\tau_k(\mathbf{x}_i; \boldsymbol{\phi}) = pr(z_{ik} = 1 | \mathbf{x}_i) = \frac{\pi_k f_k(\mathbf{x}_i; \boldsymbol{\theta}_k)}{f(\mathbf{x}_i; \boldsymbol{\phi})} \tag{3}$$

for $i = 1, \ldots, n$; $k = 1, \ldots, K$.

2.2. Classification Likelihood approach

The classification likelihood approach is another likelihood based approach to clustering. With this approach, $\boldsymbol{\phi}$ and the unknown indicators of component membership $\mathbf{z}_i, i = 1, \ldots, n$, are chosen to maximise $\log L_C(\boldsymbol{\phi})$, the complete data log-likelihood for $\boldsymbol{\phi}$ given by

$$\log L_C(\boldsymbol{\phi}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \{\log \pi_k + \log f_k(\mathbf{x}_i; \boldsymbol{\theta}_k)\}. \tag{4}$$

This is also known as the classification log-likelihood. The classification likelihood approach treats the $\mathbf{z}_i$ as unknown parameters that are estimated along with $\boldsymbol{\phi}$. This approach to clustering has been shown to be equivalent to some commonly used clustering criteria under varying assumptions on the component densities. Further details on the classification likelihood approach to clustering may be found in for example, Basford and McLachlan (1988), Section 1.2, Celeux and Govaert (1991, 1993, 1995) and Banfield and Raftery (1993).

2.3. Relationship between the Likelihood and Classification Likelihood approaches

As noted by Hathaway (1986), and others, the relationship between the log-likelihood for the mixture model (1) and the classification log-likelihood (4) can be written as

$$\log L(\boldsymbol{\phi}) = \log L_C(\boldsymbol{\phi}) - \log g(\boldsymbol{\phi}), \tag{5}$$

where

$$\log g(\boldsymbol{\phi}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log \tau_{ik},$$

$\tau_{ik} = \tau_k(\mathbf{x}_i; \boldsymbol{\phi})$ is the posterior probability that $\mathbf{x}_i$ comes from the k^{th} component defined by (3), and $g(\boldsymbol{\phi})$ is the conditional density of $\mathbf{z} = (\mathbf{z}_1', \ldots, \mathbf{z}_n')'$ given the observed data $\mathbf{x} = (\mathbf{x}_1', \ldots, \mathbf{x}_n')'$.

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

$$EN(\boldsymbol{\tau}) = -\sum_{i=1}^{n} \sum_{k=1}^{k} \tau_{ik} \log \tau_{ik}$$

is the entropy of the fuzzy classification matrix composed of elements τ_{ik}, and where $\boldsymbol{\tau} = (\boldsymbol{\tau}_1', \ldots, \boldsymbol{\tau}_n')'$ and $\boldsymbol{\tau}_i = (\tau_{i1}, \ldots, \tau_{ik})'$ is the vector of posterior probabilities of component membership of $\mathbf{x}_i$ $(i = 1, \ldots, n)$. It follows from (5) that if $\mathbf{z}$ is replaced by $\hat{\boldsymbol{\tau}}$ in $\log L_C(\boldsymbol{\phi})$, then

$$\log L_C(\hat{\boldsymbol{\phi}}) = \log L(\hat{\boldsymbol{\phi}}) - EN(\hat{\boldsymbol{\tau}}), \tag{6}$$

where $\hat{\boldsymbol{\tau}}$ is the maximum likelihood estimate of $\boldsymbol{\tau}$ formed by replacing τ_{ik} with $\hat{\tau}_{ik} = \tau_k(\mathbf{x}_i; \hat{\boldsymbol{\phi}})$ for $i = 1, \ldots, n; k = 1, \ldots, K$.

3. The Multimix approach

When clustering real multivariate data sets that have a large number of attributes, it is rare to find all attributes being either categorical or continuous, as some approaches based on mixture models require. Hunt (1996) and Hunt and Jorgensen (1999) proposed the Multimix class of models. They suggest an approach based on a form of local independence by partitioning the attribute vector $\mathbf{x} = (x_1, x_2, \ldots, x_p)'$ into L subvectors of varying sizes such that

$$\mathbf{x} = (\check{\mathbf{x}}_1', \check{\mathbf{x}}_2', \ldots, \check{\mathbf{x}}_L')'$$

where the attributes within subvector $\check{\mathbf{x}}_l$ are independent of the attributes in subvector $\check{\mathbf{x}}_{l'}$, for $l, l' = 1, \ldots, L$, and $l \neq l'$, within each of the K components. Thus if individual i belongs to component k, then we write

$$f_k(\mathbf{x}_i) = \prod_{l=1}^{L} f_{kl}(\check{\mathbf{x}}_{il}).$$

The Multimix approach allows the clustering of mixed data containing both categorical and continuous attributes. Although it is possible to use other distributions for the partition cells, only the following two distributions are considered for the subvector partitions l:

1. Discrete distributions where $\check{\mathbf{x}}_l$ is a one dimensional discrete attribute that can take the values $1, \ldots, M_l$ with probability $\lambda_{kl1}, \ldots, \lambda_{klM_l}$. This distribution will be denoted by $D(\lambda_{kl1}, \ldots, \lambda_{klM_l})$.
2. Multivariate Normal Distributions where $\check{\mathbf{x}}_l$ is a p_l-dimensional vector with a $N_{p_l}(\mu_{kl}, \Sigma_{kl})$ distribution.

The subvectors resulting from the partitioning, are usually formed with vectors of the same type, categorical or continuous. When a subvector contains only a single variable, that variable is independent of all other variables within each component. As described by Hunt and Jorgensen (1999), if all attributes are continuous, then the distribution given by (1) is a mixture of multivariate normal distributions. The form of the matrix of covariance parameters in each component distribution f_k is determined by the way in which the set of attributes is partitioned into subvectors. The form is block diagonal with a square block corresponding to each subvector. If all the attributes are discrete, the model is the usual latent class model. If strong within-cluster associations between two discrete attributes are detected after a preliminary clustering, the two attributes may be combined into a single discrete attribute with a level for each cell of the two-way table (or fewer, if some cells are pooled). However, this has the disadvantage of increasing the number of parameters that need to be estimated.

To cope with the possibility of within-cluster associations between a discrete attribute and several continuous attributes, a location model distribution (Krzanowski (1983)) could be used for a subvector partition. Although this possibility is introduced in the general MULTIMIX model (Hunt and Jorgensen (1999)), it has the disadvantage that it greatly increases the number of parameters that need to be estimated. We do not consider location model distributions for the subvector partitions.

When using the Multimix approach to clustering, Hunt and Jorgensen (1996, 1999) suggested that once it is decided that a mixture model is appropriate, the model of complete local independence should firstly be fitted for various values of K. These authors then used i) the approximation for the likelihood ratio test suggested by Wolfe (1971), and ii) the estimates of the posterior probabilities of component membership as guides for testing the number of components to be fitted to the mixture. For the selected value (or values) of K, the observations are assigned to the component to which they have greatest posterior probability of belonging. Using the component assignment from this model, modifications to the model are made by examining correlations, scatter plots and two-way tables within each of the components formed. Attributes with strong within-component associations are grouped together in a subvector partition for the next series of fits. This process is repeated with the model modified as necessary. Note that this is a form of forward selection of attributes. Correlations are introduced into the model sparingly as this increases the number of parameters that need to be estimated.

Hunt and Jorgensen (1996,1999) recommend that subject knowledge should also be used to see if the components produced are meaningful. For a specified value of K, they use the likelihood ratio test for nested models. For testing between competing non-nested models, they do not specify any criteria to aid in whether a correlation should be incorporated in the model – this is left to the judgement of the user.

The use of information criteria to quantify the differences between a set of candidate models has become increasingly popular in model based cluster analysis (Stahl and Sallis (2012)). The fitting strategy suggested by Hunt and Jorgensen (1999) is followed, but various clustering criteria are also calculated and used to guide in the selection of the final form of the model.

4. Criteria for assessing the mixture model

4.1. Akaike Information Criterion

The Akaike Information Criterion (*AIC*) (Akaike (1973, 1974)) is given by

$$AIC = -2\log L(\hat{\boldsymbol{\phi}}) + 2d,$$

where $L(\hat{\boldsymbol{\phi}})$ is the maximised likelihood function for $\boldsymbol{\phi}$, the unknown parameters of the model, and d is equal to the total number of free parameters in the model. This criterion selects the model and the number of components K in the model, to be the combination that has the smallest *AIC*.

It is well known (see for example, Aitkin and Rubin (1985)) that the regularity conditions on which the *AIC* criterion relies, do not hold for tests on the number of components in a mixture model. Studies have shown that the *AIC* criterion tends to overestimate the number of components in the model (see for example, Celeux and Soromenho (1996)). However, this criterion is often used to assess the number of components to be fitted.

4.2. Bayesian Information Criterion

The Bayesian Information Criterion (*BIC*) of Schwarz (1978) is another criterion that is commonly used for model selection. This criterion was derived in a Bayesian framework but it can also be used in a non-Bayesian framework for model selection in mixture models (Fraley and Raftery (2002), Steele and Raftery (2010)). The *BIC* is given by

$$BIC = -2\log L(\hat{\boldsymbol{\phi}}) + d\log n, \tag{7}$$

where d and $L(\hat{\boldsymbol{\phi}})$ are as defined above, and n is the number of individuals. It uses an approximation to the exact Bayes solution, with the number of components K to be fitted to the model chosen to be that value that has minimum *BIC*. Similarly to the *AIC*, the *BIC* also depends on regularity conditions that do not hold for assessing the number of components. However several authors, for example Fraley and Raftery (1998), Dasgupta and Raftery (1998), Steele and Raftery (2010), Everitt *et al* (2011) report that there is appreciable support for use of *BIC* in this context. Choosing the model with the minimum *BIC* is equivalent to choosing the model with the largest posterior probability, asymptotically.

The *BIC* can also be used to also to compare models with differing parameterizations (Biernacki and Govaert (1999) and Raftery and Dean (2006)). It can determine the relative merits of each of the M models considered (Hastie *et al* (2001)) by computing the *BIC* for each of these models $(BIC_1, \ldots, BIC_M)$, and then estimating the posterior probability of each model as

$$\frac{e^{-\frac{1}{2}BIC_m}}{\sum_{m=1}^{M} e^{-\frac{1}{2}BIC_m}}$$

for $m = 1, \ldots, M$.

It can be seen from equation (7) that the penalty term in the *BIC* penalizes complex models more heavily than *AIC*, and that the criterion gives preference to simpler models. However, for small samples where the model for the component densities is valid, the *BIC* has been found to fit models that have too few components (Celeux and Soromenho (1996)). If the correct model is not one of the models being considered, the *BIC* will tend to fit too many components in the model (Biernacki and Govaert (1999)).

4.3. Classification Likelihood Information Criterion

Biernacki and Govaert (1997, 1999) suggested an approach based on the classification likelihood of Symons (1981). They proposed a criterion that makes use of the relationship that exists between the log-likelihood $L(\boldsymbol{\phi})$ and the classification log-likelihood $L_C(\boldsymbol{\phi})$ (see equation (6)). This criterion is known as the classification likelihood information criterion (CLC). The number of components in the mixture model is chosen by minimizing

$$CLC = -2\log L(\hat{\boldsymbol{\phi}}) + 2EN(\hat{\boldsymbol{\tau}}), \tag{8}$$

where the entropy $EN(\hat{\boldsymbol{\tau}})$ is a term that penalizes the standard likelihood and measures the quality of the partition (Celeux and Soromenho (1996)). The entropy measures the overlap of the mixture components and can be regarded as a measure of the ability of the mixture model to provide well separated clusters. If the mixture components are well separated, $EN(\hat{\boldsymbol{\tau}})$ will be close to its minimum value of zero, whereas if the components are poorly separated $EN(\hat{\boldsymbol{\tau}})$ will have a large value.

Biernacki and Govaert (1997, 1999) suggested four strategies for using the classification likelihood to determine the number of components in the mixture model. They compared the performance of these strategies with some standard criteria using simulated data and Fisher's Iris data. Biernacki *et al* (1999) found that the classification likelihood criterion worked well when the mixing proportions were restricted to be equal, but when no restrictions were placed on the mixing proportions, the criterion tended to overestimate the correct number of components. Biernacki *et al* (1998, 2000) suggested that this occurred because the classification likelihood does not penalize the complexity of the mixture model.

4.4. Integrated Completed Likelihood Criterion

The integrated completed likelihood (ICL) criterion was proposed by Biernacki *et al* (1998, 2000) in an attempt to overcome the shortcomings of BIC and CLC. This criterion is based on the integrated likelihood of the complete data $(\mathbf{x}, \mathbf{z})$. These authors use a Jeffrey's non informative prior where $\alpha = 0.5$. Twice the negative of the log integrated classification likelihood can be approximated by

$$-2\log L(\hat{\boldsymbol{\phi}}) + 2EN(\hat{\boldsymbol{\tau}}) + 2n\sum_{k=1}^{K}\hat{\pi}_k\log\hat{\pi}_k + d_1\log n - 2\Upsilon(n\hat{\pi}_1,\ldots,n\hat{\pi}_K),$$

where d_1 is the number of unknown parameters, and

$$\Upsilon(n\hat{\pi}_1,\ldots,n\hat{\pi}_K) = \sum_{k=1}^{K}\log\Gamma(n\hat{\pi}_k+\alpha) - \log\Gamma(n+K\alpha) - K\log\Gamma(\alpha) + \log\Gamma(K\alpha).$$

The ICL criterion can be used for choosing the number of components to be fitted to the mixture model, and also for choosing the form of a relevant model by selecting the model that has minimum value of ICL. Further details on the derivation of this criterion may be found for example, in Biernacki *et al* (2000), McLachlan and Ng (2000) and McLachlan and Peel (2000). Biernacki *et al* (2000) compared the behaviour of BIC, ICL and an approximation that they termed the Cheeseman Stutz approximation of the integrated complete likelihood for choosing the form of the component distributions and the number of components in the mixture using simulated and real data sets. They restricted their attention to Gaussian mixtures. They found that the ICL performed well in assessing the number of components to be fitted to the mixture model and the form of the component distributions.

Biernacki *et al* (1998) also derived an approximation to the ICL that holds in situations with large cluster sizes. Similarly to McLachlan and Ng (2000), we refer to this version of the ICL as $ICL - BIC$. This criterion is given by

$$ICL - BIC = -2\log L(\hat{\boldsymbol{\phi}}) + 2EN(\hat{\boldsymbol{\tau}}) + d\log n. \tag{9}$$

Biernacki *et al* (1998) found that the performance of this criterion differed little from that of the ICL even though the approximation is only appropriate for large cluster sizes.

In contrast to the use of the likelihood ratio test statistic, $-2\log\lambda$, for the determination of the number of components K, the likelihood ratio test statistic can be used when comparing two nested models and it should not give

misleading results (see for example, Wolfe (1971), Hunt and Jorgensen, 1999). The test is based on the approximate distribution for -2 log λ being χ^2 with the degrees of freedom equal to the difference in the number of parameters in the two models.

McLachlan and Peel (2000, Chapter 6), give a discussion on assessing the number of components in mixture models where the model has been used for clustering. McLachlan and Ng (2000) describe some of the criteria proposed for determining the number of components in the mixture model. They report the results of three simulated data sets for comparing the performance of these criteria with the classical criteria, *AIC* and *BIC*. They show that some of the more recent criteria performed better than the classical criteria and correctly determined the true number of components in all three simulated data sets.

5. The analysis

Various criteria for model selection will be calculated when we consider the clustering of cases using the pretrial variables from the prostate cancer clinical trial data reported by Byar and Green (1980). The dataset is available at *http://lib.stat.cmu.edu/datasets/Andrews/T46.1*. These data were obtained from a randomized clinical trial comparing four treatments for 506 patients with prostatic cancer. Twelve pre-trial attributes (Table 1) had been measured on each patient. Seven of these attributes may be taken to be continuous, four to be discrete, and one attribute ('Index of tumour stage and histolic grade', SG) may be taken as either categorical or continuous. This measurement is considered as a continuous attribute. Similarly to Hunt and Jorgensen (1999), two covariates, 'Size of primary tumour' (SZ) and 'Serum prostatic acid phosphatase' (AP), have been transformed to make their distributions more symmetric, SZ with a square root transformation and AP by a logarithmic transformation.

Attribute	Abbreviation	Number of levels if categorical
Age	Age	
Weight index	WT	
Performance rating	PF	4
Cardiovascular disease history	HX	2
Systolic blood pressure	SBP	
Diastolic blood pressure	DBP	
Electrocardiogram code	EKG	7
Serum haemoglobin	HG	
Size of primary tumour	SZ	
Index of tumour stage and histolic grade	SG	
Serum prostatic acid phosphatase	AP	
Bone metastasis	BM	2

Table 1. Pre-trial attributes. Adapted from: Hunt and Jorgensen (1999)

A complete case clustering of the twelve pretrial attributes is reported where individuals that had missing values in any of these pretrial attributes were omitted from further analysis, leaving 475 out of the original 506 individuals available. A classification of the data is available as physicians had classified the patients using clinical criteria as having either Stage 3 or Stage 4 prostatic cancer. However this information shall be excluded from the fitting of the models as in most cluster analyses that are performed, a classification of the data is not available. The fitting strategy of Hunt and Jorgensen (1996) will be followed.

We regard the data as a random sample from the distribution given by equation (1). Under the model of complete local independence, the component distributions will be of the form $\prod_{l=1}^{12} f_{kl}(\mathbf{x}_{il}; \theta_{kl})$ where f_{kl} is the $N_{kl}(\mu_{kl}, \sigma^2_{kl})$ distribution for each of the eight continuous attributes and $D(\lambda_{kl1}, ..., \lambda_{klM_l})$ for each of the four discrete attributes. The partitioning in the local independence model will be referred to as Attribute Partition 1 (AP1). To determine the number of components to be fitted to the mixture model, models with partition AP1 were fitted for $K = 1, \ldots, 4$. Clustering criteria were calculated for each of the fitted models.

The models were fitted iteratively using the EM algorithm from various different starting values generated by splitting the data into clusters both randomly and using various criteria. Several local maxima were found, and the solution of the likelihood equation was taken to be the one corresponding to the largest of the local maxima. Each individual was assigned to the component to which it had highest estimated posterior probability of belonging.

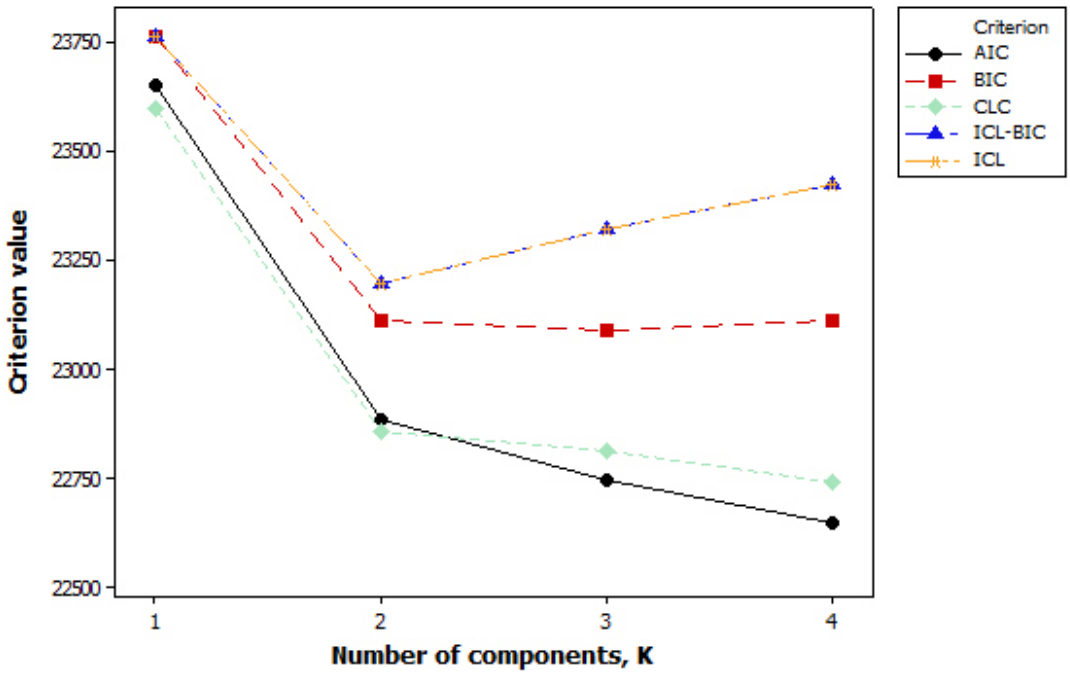

Figure 1. Clustering criteria, *AIC*, *BIC*, *CLC*, $ICL - BIC$ and *ICL*, for Attribute Partition 1 (AP1) for $K = 1, \ldots, 4$ for clustering the prostate cancer clinical trial data.

It can be seen from Figure 1 that the two criteria, $ICL - BIC$ and *ICL*, had minimum value at K=2, indicating that there were two components in the data, whereas the *BIC* for the models with partion AP1, had minimum value at $K = 3$, indicating three components. Both the *CLC* and *AIC* had minimum value at $K = 4$ components. It is known that *AIC* traditionally overestimates the number of components to be fitted to the model, and that *CLC* does not perform well when the mixing proportions are unequal. For the four component model, there was increased sensitivity to starting values and more tendency to converge to local maxima. Note that for the four component models, we are unsure whether we have achieved the best endpoint because of this difficulty. As more components were added to the model, there was an increasing tendency to converge to local maxima. This was not unexpected as each additional component added to the model required an additional 28 parameters to be estimated. It was therefore decided not to investigate the model with $K = 5$ components.

The log-likelihood ratio test was also considered as a guide to the possible number of components to be fitted to the model. To test the hypothesis of $K = 1$ versus $K = 2$ components in the models with AP1, the likelihood ratio test statistic (LRTS) $-2 \log \lambda = 823.202$. Thus we can be confident that there is not a single component. The LRTS for the AP1 models fitted with $K = 2$ versus $K = 3$ had $-2 \log \lambda = 195.474$. While for the models with $K = 3$ versus $K = 4$, the LRTS $-2 \log \lambda = 150.136$. These are both significant values. Clearly there are at least two components.

The posterior probabilities were also examined to guide the number of components to be fitted. Table 2 shows that as the number of components fitted to the model increased, there was a decrease in the number of individuals that were decisively assigned ($\hat{z}_{ik} \geq 0.95$) to a component. It can also be seen that as the number of components in the model increased, there were also more individuals that had appreciable membership in more than one component. The two component model had more individuals that were decisively assigned.

$\max_k \hat{z}_{ik}$	Components K 2	3	4
0.25− < 0.80	33	100	132
0.80− < 0.95	44	99	142
0.95− < 0.99	46	75	85
0.99 – 1.0	352	201	116

Table 2. Maximum Posterior probabilities for models with partition AP1 for clustering the prostate cancer clinical trial data.

The observations were assigned to the component of greatest posterior probability for $k = 2, \ldots, 4$, and the 'statistical diagnosis' was compared with the clinical classification. As detailed in Hunt (1996), the two component model also has an interpretation that agrees with the clinical classification of Stage 3 and Stage 4 prostatic cancer, with the statistical and clinical classifications only differing for 41 individuals. For the three component model, 90.7% of the Stage 3 patients were divided among two components and most (89.6%) of the Stage 4 patients were assigned to the third component. In the four component model, the two components of the Stage 3 patients were virtually identical to those found in the three component model, however the component corresponding to the Stage 4 patients had been divided among two components. This type of division indicates the stability of the clustering structure under the models fitted.

We then investigated models for finite mixtures fitted with both two and three components. The 'forward selection of attributes' strategy given by Hunt and Jorgensen (1999) was used, commencing with the model of complete local independence. After fitting this model for the fixed number of components i.e. either $K = 2$ or $K = 3$, each individual was assigned to the component of greatest posterior probability. The within cluster correlation structure was examined and local correlations were added progressively to the model by considering partitions with more parameters.

5.1. Attribute Partitions for Two Components

When the data were placed into two clusters using the posterior assignments from the model fitted with partitioning AP1, it was found that both clusters exhibited a moderate correlation between systolic blood pressure SBP and diastolic blood pressure DBP, 0.629 for component one and 0.623 for component two. This correlation was incorporated into a subvector partition, Attribute Partition 2 (AP2), by placing these two attributes together into a subvector, with all other attributes independent. The local independence assumption was weakened slightly by adding this covariance parameter in each component.

The next partitioning of attributes chosen had two bivariate subvectors, with SBP and DBP in one subvector, Weight (Wt) and Serum Haemoglobin (Hg) in another subvector, and all other attributes were independent. This is Attribute Partition 3 (AP3). The process of examining the within component correlation structure and only introducing correlations into the model as they were forced in, was repeated, resulting in AP4, AP5, AP6 and AP7. Table 3 shows the attribute partitions, the correlation structure within the partitions, and the number of parameters to be estimated for the considered models.

It can be seen from Figure 2 that the criteria *AIC* and *CLC* both selected the model with the correlation structure in AP7, whereas the criteria *BIC*, $ICL - BIC$ and *ICL* all selected the model with the correlation structure in AP6 as the best model for the data. In fact, there was not much difference between the *BIC*, *ICL* and $ICL - BIC$ criteria values for models with the correlation structure in AP3 and AP6. Hunt and Jorgensen (1999) studied models with the correlation structure in AP2 in more detail whereas Hunt (1996), Jorgensen and Hunt (1996) analysed models with the correlation structure in AP4. These authors originally selected AP4 as the preferred partitioning because they thought that on physical grounds there would be correlations between a patient's weight and the two types of blood pressure.

5.1.1. Comparison of Attribute Partitions for Two Component Models

The likelihood ratio test was also used as a guide for the preferred model. The model with attribute partition AP2 had two extra parameters in comparison to the model with attribute partition AP1, *i.e.* the covariance between the two blood pressures for each component. The log-likelihood ratio $-2\log\lambda = 235.08$, clearly a definite improvement. The attribute partition AP3 added covariances between Hg and Wt to the attribute partiton AP2, with - $2\log\lambda = 29.27$

Attribute Partitions	Correlation Structure	No. Parameters	Log-likelihood
AP1	Local Independence	55	-11386.27
AP2	{SBP, DBP}	57	-11268.72
AP3	{SBP, DBP}, {Wt, Hg}	59	-11254.09
AP4	{SBP, DBP, Wt}	61	-11254.74
AP5	{SBP, DBP, Wt, Hg }	67	-11235.96
AP6	{SBP, DBP}, {AP, Wt, Hg}	63	-11236.85
AP7	{AP, SBP, DBP, Wt, Hg }	75	-11217.10

Table 3. Attribute Partitions, correlation structure within a partition, number of parameters (including the mixing proportions) and log-likelihoods for the two component models fitted to the prostate cancer clinical trial data.

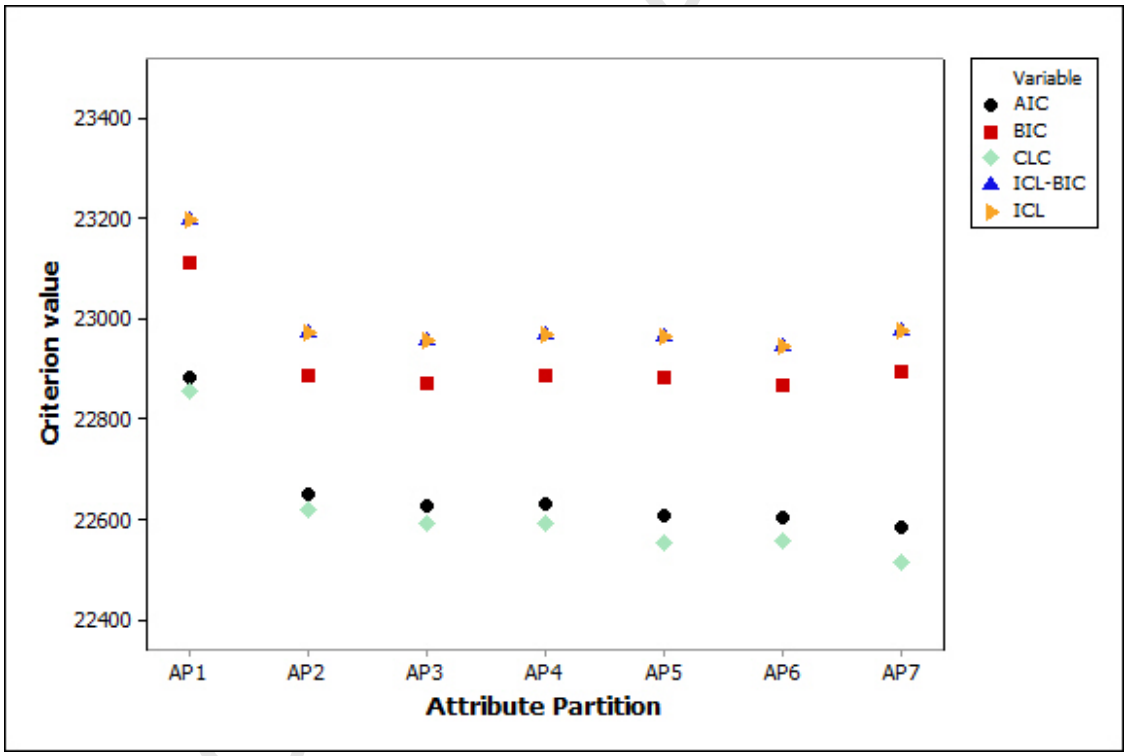

Figure 2. Clustering criteria, *AIC*, *BIC*, *CLC*, *ICL* − *BIC* and *ICL*, for partitions AP1 to AP7 for $K = 2$ for the prostate cancer clinical trial data.

for the cost of two extra parameters. The partition AP5 added covariances between Hg and Wt, and SBP and DBP to partition AP3, for the cost of eight parameters with $-2\log\lambda = 36.26$. Partition AP4 added four extra parameters to AP2 with $-2\log\lambda = 27.96$. Partition AP5 added six extra parameters to AP4, with $-2\log\lambda = 37.56$. Partition AP6 added four extra parameters to AP4, with $-2\log\lambda = 34.48$. Attribute partitions AP2 to AP6 were all better fitting models than the fully local independence model AP1, at a cost of a modest number of extra parameters. Partitions AP3 and AP4 were better fitting models than AP2, again for a small number of extra parameters.

For each of the models fitted, the observations were assigned to their cluster of greater posterior probability, and the clusters were compared with the clinical classification. The models with partitions AP4 and AP6, detected the clinical classification of Stage correctly for 91.8% of the observations, whilst models with partitions AP3, AP7, and AP5 detected the clinical classification of Stage correctly for 91.6% of the observations. Models with partitions AP1 and AP2, respectively, detected the clinical classification of Stage correctly for 91.4% and 91.1% of the observations. Investigation of the clusters formed for each of the fitted models found that they only differed in the assignment of a maximum of four observations.

5.2. *Attribute Partitions for Three Components*

The fitting strategy described above was followed for models with three components. Models with the attribute partitions shown in Table 3 were fitted. Note that following the fitting strategy described above, the attribute partitions, AP6 and AP7, only showed a weak correlation in two of the components and would not be considered as possible partitioning.

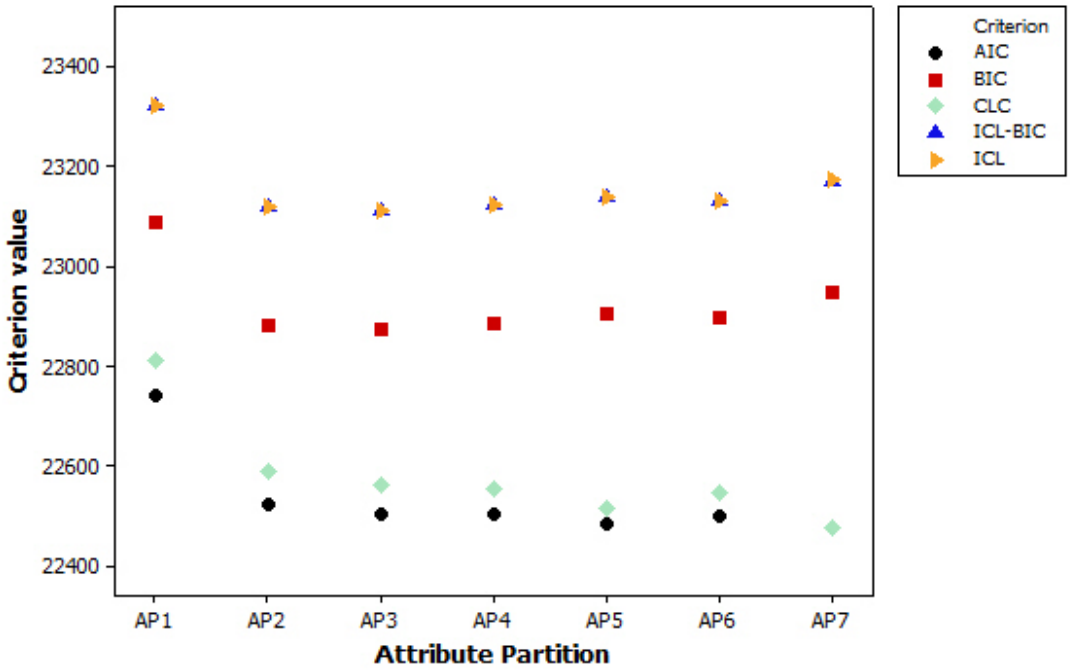

Figure 3. Clustering criteria, *AIC*, *BIC*, *CLC*, *ICL* − *BIC* and *ICL*, for partitions AP1 to AP7 for $K = 3$ for the prostate cancer clinical trial data.

We see from Figure 3 that the criteria, *BIC*, *ICL* − *BIC*, and *ICL*, all picked attribute partition AP3 as the best model whereas the criteria *AIC* and *CLC*, both picked the model with partition AP5. This was not unexpected as *AIC* has a tendency to pick a model with more parameters. *CLC* does not work as well when the mixing proportions are not equal, ($\hat{\pi} = (0.308, 0.278, 0.415)$). Note that if the attribute partitions AP6 and AP7, were included, both *AIC* and *CLC* would pick the model with the partitioning given in AP7.

5.2.1. Comparison of Attribute Partitions for Three component Models

In comparison to a three component model with partition AP1, AP2 has three extra parameters to be estimated. Twice the difference in the log-likelihoods is 224.58, clearly a definite improvement. The attribute partition AP3 added the covariances between Hg and Wt to attribute partition AP2 (at a cost of the estimation of three extra parameters), with $-2log\lambda = 25.962$. Attribute partition AP5 added 12 extra parameters to partition AP3 with $-2\log\lambda = 42.944$. The partition AP4 added six extra parameters to partition AP2 with $-2\log\lambda = 33.384$. Attribute partition AP5 added 9 extra parameters to AP4 with $-2\log\lambda = 35.522$. Models with attribute partitions AP2 to AP5, are all better fitting models than the local independence model (AP1) with the cost of a modest number of parameters. Models with attribute partitions AP3 and AP4 are better fitting models than partition AP2, again for a small number of parameters.

For each of the fitted models with $K = 3$, the observations were assigned to the cluster of greatest posterior probability. The clusters were compared to the clusters from the comparable two component model. When a three component model was fitted, most (96%) of the observations in the cluster that corresponded to a clinical classification of Stage 4, were assigned to a single cluster, whilst the bulk (98.5%) of the remaining cluster was divided into two other clusters. Investigation of the clusters formed in the three component models with different partitioning of the attributes, found that they differed in the assignment of a maximum of six observations, indicating the stability of the grouping structure under the models fitted.

5.3. The fitted model with attribute partition AP3

The number of components to be fitted to the model was then assessed using the attribute partition AP3.

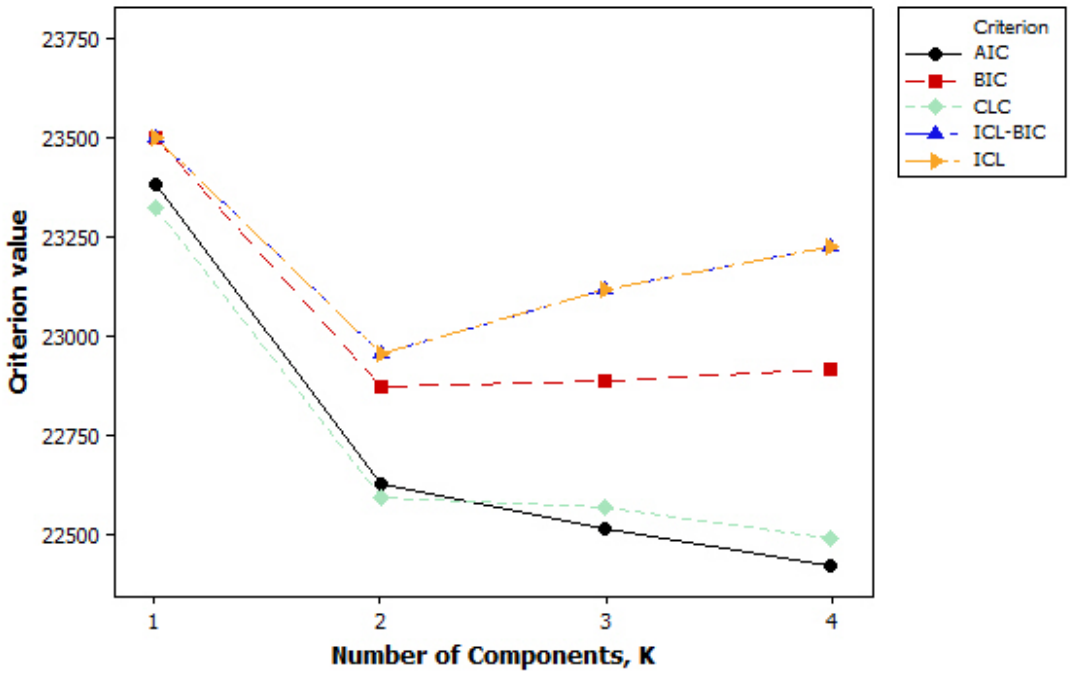

Figure 4. Clustering criteria, *AIC*, *BIC*, *CLC*, *ICL* − *BIC* and *ICL*, for model selection for investigating the number of components to be fitted using the partition AP3 for the prostate cancer clinical trial data.

It can be seen from Figure 4 that both *AIC* and *CLC* overestimated the number of components to be fitted to the mixture model. For the model with four components, the estimates of the mixing proportions were $\hat{\pi} = (0.278, 0.283, 0.264, 0.176)$. Criterion CLC performs better when the mixing proportions are equal (Biernacki and Govaert (1997)). The remaining criteria all have their minimum value at $K = 2$ components.

The criteria were then examined for all two component (AP1 to AP7) and three component (AP1 to AP5) models that had been fitted. Criteria, AIC and CLC, both had minimum value for partition AP5 for $K = 3$ components, whereas criteria, BIC, ICL and $ICL - BIC$, had minimum value for partition AP6 for $K = 2$ components.

6. Simulation studies

In this section, the results of some simulation studies to assess the various criteria in determining the optimal form of distributions and the number of components in the mixture model are reported. It was decided to assess the criteria on mixed data that had a known classification and a known correlation structure within each component, as although the cancer data had an available classification, the correlation structure within each of the components was unknown.

A two component mixture of 500 observations was generated where the proportion of observations in each component was 0.574 and 0.426. Note that this is similar to the proportion of observations in the cancer data set that were classified as having Stage 3 or Stage 4 prostate cancer. Each observation consisted of 12 attributes, $A, B, C, \ldots, L$, where 8 attributes, $A, B, C, \ldots, H$, were continuous and 4 attributes, $I, \ldots, L$, were categorical. Similarly to the cancer data, one categorical attribute had 7 levels, another had 4 levels and the remaining two attributes each had 2 levels. Each continuous attribute was generated from a normal distribution. Within each component, the parameters for the continuous attributes and the probability of being at a particular level for the categorical attributes were similar to those for the observations classified as having Stage 3 or Stage 4 prostate cancer. Fifty data sets were generated for each set of simulations.

The Multimix program of Hunt (1996) was used to fit various models to the simulated data sets. Twenty random starts were performed for each data set to initiate the algorithm. The solution corresponding to the largest of the local maximum was taken to be the solution of the likelihood equation, and the values of each of the five criteria, $AIC, BIC, CLC, ICL - BIC$ and ICL, were noted for this solution.

The fitting strategy given by Hunt and Jorgensen (1999) was followed and K component ($K = 1, \ldots, 4$) models were fitted to the data. The value of K that had minimum value over the four components for each of the criteria calculated, was noted for each data set. For the selected models, each observation was assigned to the component which it has highest estimated posterior probability of belonging. The within component correlation structure was examined and variables with strong within component correlations were put together into a partition cell for the next series of fits. This process was repeated until there were no further associations to be incorporated into the model.

6.1. Simulation Set 1

In the first set of 50 simulations, the data were generated such that each of the attributes was independent within the two sub populations. The model of complete local independence was fitted for $K = 1, \ldots, 4$, and the five clustering criteria were calculated for each model.

The percentage of times each criterion had its minimum value at K components for $K = 2, \ldots, 4$ is listed in Table 4. This helped identify the number of components K to be fitted to the local independence model. We see from Table 4 that the criteria $BIC, ICL - BIC$ and ICL did not differ in the number of components K to be fitted to the model. The criteria BIC, $ICL - BIC$ and ICL correctly identified the number of components in the mixture 96% of the times whereas both AIC and CLC did not perform as well. Examination of the within component correlation structure after assigning each observation to its component of greater posterior probability for $K = 2$ did not indicate the need to incorporate any local associations in the model for any of the simulation sets analysed. It can therefore be concluded that the local independence model with two components could be used for clustering the data.

6.2. Simulation Set 2

Data were simulated from a two component distribution with the partitioning $\{A, B\}$, with a correlation of 0.5 existing between the two attributes A and B within each component, while all other 10 attributes were independent.

It can be seen in Table 4 that when the partitioning of the correct model is not included in the model fitted to the data, AIC overfits the data and picks the model with the highest number of components. The criteria BIC, $ICL - BIC$ and ICL correctly picked the number of components to be fitted to the data in all simulations.

Examination of the within cluster correlation structure after assigning each observation to its component of greater posterior probability for $K = 2$ indicated a within component correlation between the two variables A and B, in both

Simulation Set	Underlying Attribute Partition	Fitted Attribute Partition	Number of Components, K	AIC	CLC	BIC	$ICL-BIC$	ICL
1	Local Independence	Local Independence	2	78	56	96	96	96
			3	20	42	4	4	4
			4	2	2	0	0	0
2	$\{A, B\}$	Local Independence	2	0	44	100	100	100
			3	20	56	0	0	0
			4	80	0	0	0	0
		$\{A, B\}$	2	78	44	100	100	100
			3	22	56	0	0	0
			4	0	0	0	0	0
3	$\{A, B\}$,$\{C, D\}$	Local Independence	2	0	56	100	100	100
			3	20	20	0	0	0
			4	80	4	0	0	0
		$\{A, B\}$	2	0	72	100	100	100
			3	36	24	0	0	0
			4	64	4	0	0	0
		$\{A, B\}$, $\{C, D\}$	2	60	40	100	100	100
			3	20	20	0	0	0
			4	20	40	0	0	0
4	$\{A, B, C\}$	Local Independence	2	0	0	54	100	100
			3	0	0	32	0	0
			4	100	100	14	0	0
		$\{A, B, C\}$	2	0	24	100	100	100
			3	0	36	0	0	0
			4	100	40	0	0	0
5	$\{A, B\}$, $\{C, D, E\}$,	Local Independence	2	0	0	48	100	100
			3	0	8	48	0	0
			4	100	92	4	0	0
		$\{A, B\}$	2	0	0	60	100	100
			3	0	6	38	0	0
			4	100	94	2	0	0
		$\{A, B\}$, $\{C, D, E\}$	2	58	28	100	100	100
			3	28	20	0	0	0
			4	16	52	0	0	0
6	$\{A, B, C, D, E\}$,	$\{A, B\}$	2	0	0	0	0	0
			3	0	0	0	30	30
			4	100	100	100	70	70
		$\{A, B, C\}$	2	0	0	12	100	100
			3	0	0	48	0	0
			4	100	100	40	0	0
		$\{A, B, C, D\}$	2	0	0	100	100	100
			3	4	28	0	0	0
			4	96	72	0	0	0
		$\{A, B, C, D, E\}$	2	62	18	100	100	100
			3	32	42	0	0	0
			4	6	40	0	0	0

Table 4. The percentage of times each of the five criteria, AIC, CLC, BIC, $ICL-BIC$ and ICL, had minimum value at K components for each fitted AP for a given underlying AP for the Simulation Sets.

components. The attribute partition $\{A, B\}$ was incorporated into the model. After fitting this model, the within component correlation structure was examined. There were no further correlations to be incorporated into the model. The number of components to be fitted to this model was then investigated.

Table 4 indicates that when the partition $\{A, B\}$ was included in the partitioning of the attributes, the criteria BIC, $ICL - BIC$ and ICL all had minimum value for models with $K = 2$ components. The criterion AIC selected two components in the model for 78% of the simulations whereas CLC did not perform as well with a model with three components being predicted in 56% of the simulations. Examination of the within component correlation structure after each observation had been assigned to its component of greater posterior probability for $K = 2$ did not indicate the need for any further associations to be incorporated into the model.

Each of the five criteria calculated for the local independence model was compared with the one calculated for the fitted model with attribute partition $\{A, B\}$ for all simulations, $K = 1, \ldots, 4$. The minimum value of each criterion in all simulations was that for the fitted model with attribute partition $\{A, B\}$. It can therefore be concluded that the model with the partitioning $\{A, B\}$ and $K = 2$ components could be used for clustering the data.

6.3. Simulation Set 3

Data were simulated from a two component distribution with the partitioning $\{A, B\}$, $\{C, D\}$, where there was a correlation of 0.5 existing between the two attributes A and B, and the attributes C and D, within each of the two components, while all other 8 attributes were independent, (Model 3, Table 4).

It can be seen from Table 4 that when the model of complete local independence was fitted, AIC picked four components in 80% of the simulations and selected three components in the remaining simulations. CLC picked two components in 56% of the simulations, and picked three or four components respectively in 20% and 24% of the simulations. The criteria BIC, ICL and $ICL - BIC$ all selected two components in the model. Examination of the within component correlation structure after assigning each observation to its component of greater posterior probability for $K = 2$, showed moderate correlations existing between attributes A and B and also between C and D in each of the two components. This indicated that some attribute partitions needed to be incorporated into the model.

A model with the partitioning $\{A, B\}$, with all other attributes independent was fitted, (Model 2, Table 4), and the number of components K to be fitted to the model was investigated. It can be seen from Table 4 that the criteria BIC, ICL and $ICL - BIC$ all selected two components to be fitted in the model. AIC picked four components in 64% of the simulations and selected three components in the remaining simulations. CLC picked two components in 72% of the simulations, three components in 24% of the simulations and four components for 4% of the simulations. For each of the Simulation Set 3 datasets, each of the five criteria calculated for the local independence model was compared with the one calculated for the fitted model with attribute partition $\{A, B\}$ for $K = 1, \ldots, 4$. The minimum value of each criterion in all simulations was that for the fitted model with attribute partition $\{A, B\}$. Examination of the within component correlation structure after the observations had been assigned to their component of greater posterior probability for $K = 2$ indicated that a further partitioning of the attributes was needed in the model to incorporate the within-component correlation that existed between attributes C and D.

A model with the attribute partitions $\{A, B\}$,$\{C, D\}$, (Model 3, Table 4), all other attributes being independent, was then fitted. Examination of the within-component correlations for $K = 2$ did not indicate the need to incorporate any further partitions in the model. This model was then fitted for $K = 1, \ldots, 4$ components. It can be seen from Table 4 that when the fitted attribute partition is the same as that of the generated data, AIC selected two components in 60% of the simulations and fitted too many components in the remaining 40% of the simulations, whereas CLC overfitted in 60% of the simulations. The criteria BIC, ICL and $ICL - BIC$, selected two components in all simulations. It can be seen that for Simulation Set 3, the criteria BIC, ICL and $ICL - BIC$ selected two components regardless of the partitioning that was incorporated in the model.

For each of the Simulation Set 3 datasets, each of the criteria calculated for the model with attribute partition $\{A, B\}$ was compared with the one calculated for the fitted model with attribute partition $\{A, B\}$,$\{C, D\}$ for $K = 1, \ldots, 4$. The minimum value of each criterion in all simulations was that for the fitted model with attribute partition $\{A, B\}$,$\{C, D\}$. For all models, (local independence, attribute partitions $\{A, B\}$ and $\{A, B\}$,$\{C, D\}$), fitted to the Simulation Set 3 datasets, the minimum value of each criteria in all simulations was that for the model with fitted attribute partition $\{A, B\}$,$\{C, D\}$. It can therefore be concluded that the best model has two components, with attribute partition $\{A, B\}$,$\{C, D\}$.

6.4. Simulation Set 4

Data were simulated from a two component distribution with the attribute partition $\{A, B, C\}$, where there was a correlation of 0.5 existing between the three attributes A, B, C within each of the two components, while all other attributes were independent, (Model 4, Table 4).

Table 4 shows that when the fitted attribute partition is local independence, the criteria AIC and CLC, always selected a model with four components to be fitted, whereas the criteria ICL and $ICL - BIC$ always selected two components. The criterion BIC correctly identified two components 54% of the time, selected three components 32% of the time and selected four components 14% of the time.

Examination of the within component correlation structure for $K = 2$ indicated a within component correlation of approximately 0.5 existing between each of A and B, B and C, and A and C. This indicated that the attribute partition $\{A, B, C\}$ should be incorporated into the model for $K = 2$. Examination of the within component correlation structure for $K = 3$ showed a within component association of approximately 0.5 existing between all pairs of variables A, B, C in only one component, whilst the correlations between all pairs of variables in the other two components were small. This indicated that there was no need to incorporate any partitions of the attributes into the model for $K = 3$.

The posterior probabilities were examined for both $K = 2$ and $K = 3$. These showed that there was a slight increase in the number of observations that had appreciable membership in two components for $K = 3$ in comparison to $K = 2$. The observations were assigned to their component of greatest posterior probability for both models and the resulting component assignments were compared. An examination of the assignments found that when going from $K = 2$ to $K = 3$ components, one component was identical in both models whilst the second component for $K = 2$ had been split into two components for $K = 3$. This indicated the stability of the component structure under the models fitted.

The number of components to be fitted to a model with the partitioning $\{A, B, C\}$ was then investigated. Table 4 shows that the criterion AIC overestimated the number of components to be fitted to the model even when the correct partitioning is included in the model. The criterion CLC also did not perform well in selecting the number of components to be fitted. The criteria BIC, $ICL - BIC$ and ICL all selected two components to be fitted to the model. The observations were assigned to their component of greater posterior probability and the within cluster association was examined for $K = 2$. No further associations were required to be incorporated into the model.

For each of the simulation set 4 datasets, each of the five criteria calculated for the local independence model was compared with the one calculated for the fitted model with attribute partition $\{A, B, C\}$ for $K = 1, \ldots, 4$. The minimum value of each of the criteria in all simulations was that for the fitted model with attribute partition $\{A, B, C\}$. It can therefore be concluded that the partition structure $\{A, B, C\}$ with all other attributes independent, should be fitted to the two component model.

Overall for this simulation set, we see that the BIC was able to detect the correct number of components to be fitted to the model when the correct partitioning structure was included in the model, however when the correct partitioning was not included in the model to be fitted, BIC tended to overestimate the number of components to be fitted to the model. The criteria $ICL - BIC$ and ICL, always selected two components regardless of the partition structure used in the model whereas AIC always fitted too many components regardless of the partitioning.

6.5. Simulation Set 5

Data were simulated from a two component distribution with the partitioning $\{A, B\}$, $\{C, D, E\}$, with a correlation of 0.5 existing between A and B, and also between all pairs of C, D and E, within each of the two components, while the other seven attributes were independent, (Model 6, Table 4).

The local independence model was initially fitted for $K = 1, 2, 3, 4$. Table 4 shows that the criterion AIC always selected four components to fit to the model. CLC also tended to fit too many components to the model. The criteria $ICL - BIC$ and ICL always selected two components whereas BIC selected both two components and three components in 48% of the simulations. Examination of the within component correlation structure for $K = 2$ showed correlations existing between A and B in both components and also all between all pairs of attributes C, D and E. The largest of these existing in components 1 and 2 respectively was 0.557 and 0.555 between attributes A and B. This correlation was incorporated into the model by putting the two attributes A and B into one partition, with all other attributes independent. This model was then investigated.

Table 4 shows that when the fitted attribute partition is $\{A, B\}$, criteria AIC and CLC both recommended fitting more components to the model than the other criteria. BIC had minimum value at $K = 2$ in 60% of the simulations, whereas $ICL - BIC$ and ICL picked two components in all simulations. For each of the Simulation Set 5 datasets, each criterion calculated for the local independence model was compared with the one calculated for the fitted model with attribute partition $\{A, B\}$. The minimum value of each criterion in all simulations was that for the fitted model with attribute partition $\{A, B\}$.

The within component correlation structure was investigated for $K = 2$ components. Correlations of approximately 0.5 were found between all pairs of attributes C, D and E in both components. These correlations were incorporated into the model by putting the attributes C, D, E into a partition cell.

Table 4 shows that with fitted attribute partition $\{A, B\}$, $\{C, D, E\}$, criteria BIC, $ICL - BIC$, and ICL always selected two components to fit to the model, AIC selected two components in 58% of the simulations whilst CLC selected four components in 52% of the simulations. The observations were assigned to their component of greater posterior probability for $K = 2$ and the within component correlation structure was examined. There were no further attribute partitions to be incorporated into the model.

For each of the 50 simulations, each criterion calculated for the fitted attriubute partition $\{A, B\}$ model was compared with the analogous criterion calculated for the fitted model with attribute partition $\{A, B\}$, $\{C, D, E\}$ for $K = 1, \ldots, 4$. The minimum value for the criteria in all simulations were those for the model with fitted attribute partition $\{A, B\}$, $\{C, D, E\}$. Thus it could be concluded that the partitioning $\{A, B\}$, $\{C, D, E\}$ should be incorporated into the model, and the model is likely to have two componentsl.

For this set of simulations, it can be seen that the criteria $ICL - BIC$, and ICL always selected $K = 2$ components regardless of the partitioning of the attributes. When the correct partition structure was incorporated into the model, BIC always detected the correct number of components to be fitted to the model, however when the correct partitioning was not in the model, BIC varied in the number of components to be fitted. The CLC tended to overestimate the number of components to be fitted and AIC has variable performance.

6.6. Simulation set 6

Data were simulated from a two component distribution with the partitioning $\{A, B, C, D, E\}$, with a correlation of 0.5 existing between the five attributes, A, B, C, D and E, within each of the two components, while the other seven attributes were independent (Model 7, Table 4).

When the local independence model was fitted, it was found that the criteria AIC, BIC, CLC, ICL and $ICL - BIC$ all selected $K = 4$ components to be fitted to the model for all 50 data sets. These results also include those obtained from starting the algorithm using the known component assignment for $K = 2$. Examination of the within component correlation structure for $K = 4$ showed that most correlations between pairs of variables were low. The highest correlation between a pair of variables B, C was 0.437 in one component, however the correlations between the variables B, C ranged between 0.143 and 0.211 in the other three components. This indicated that there were no correlations to be incorporated into the model. Hence, it could be concluded that the local independence model with $K = 4$ components could be used for this data.

Although the within component correlation structure did not indicate the need for any correlation structure to be fitted to the model, it was decided to increasingly introduce associations into the local independence model to observe the effect this would have on the criteria. The model with the partition $\{A, B\}$ and all other attributes independent, was fitted for $K = 1, \ldots, 4$. It can be seen from Table 4 that AIC, CLC, BIC had minimum value at $K = 4$ components for all simulations, whilst both $ICL - BIC$ and ICL had minimum value at $K = 4$ components for 70% of the simulations and $K = 3$ for the remaining simulations. Each criteria calculated for the local independence model was compared with the one calculated for the model with partition $\{A, B\}$ for $K = 1, \ldots, 4$ and for all simulations. The BIC selected the model of local independence model with four components in 8% of the simulations, and the four component model with partition $\{A, B\}$ in 96% of the simulations, whilst ICL and $ICL - BIC$ selected the local independence model with four components in 16% of the simulations, and a four component model with partition $\{A, B\}$ in the remaining simulations. AIC and CLC selected a four component model with partition $\{A, B\}$ in all simulations. The observations were assigned to their component of greatest probability for $K = 4$ and the within component correlations were examined. This did not indicate the need to incorporate any further partitioning into the model.

The next set of models fitted had the partition $\{A, B, C\}$ and all other attributes were independent. This model was fitted for $K = 1, \ldots, 4$. It can be seen from Table 4 that the criteria AIC and CLC, selected four components to

be fitted, whereas the criteria $ICL - BIC$ and ICL selected two components in all simulations. The criterion BIC, selected three components in 48% of the simulations, four components in 40% of the simulations and two components in the remaining 12% of the simulations. For each of the Simulation Set 5 datasets, each of the criteria calculated for the models with partition $\{A, B, C\}$ was compared with the one calculated for the fitted models with partition $\{A, B\}$. The minimum value of each criterion in all simulations was that for the fitted model with partition $\{A, B\}$. Each observation was assigned to its component of greatest posterior probabilty for $K = 2$ and $K = 4$. Examination of the within component correlation structure for $K = 2$ showed a correlation of 0.5 existing between five attributes, A, B, C, D and E within each of the two components. For $K = 4$ components, it was found that there was no need to incorporate any further correlations into the model.

The number of components to be fitted to models with the partitioning $\{A, B, C, D\}$ and all other attributes independent, was then investigated for Simulation Set 6. Table 4 shows that the criteria AIC and CLC both tended to recommend fitting models with four components whereas the criteria BIC, $ICL - BIC$ and ICL selected two components in all simulations. For each of the Simulation Set 5 datasets, each criterion calculated for the models with partition $\{A, B, C, D\}$ was compared with the one calculated for the fitted models with partition $\{A, B, C\}$. Criteria AIC selected the model with partition {A,B,C} with $K = 4$ in all simulations whereas all other criteria selected the models with partition $\{A, B, C, D\}$ for $K = 2, \ldots, 4$. Examination of the within component correlation structure for $K = 2$ showed correlations of approximately 0.5 existing between all pairs of attributes A, B, C, D and E in both components, indicating that the partitioning $\{A, B, C, D, E\}$ needed to be incorporated into the model.

The model with the partitioning {A, B, C, D, E} and all other attributes independent, was fitted for $K = 1, \ldots, 4$. This is the same partitioning that was used to generate the data. It can be seen in Table 4 that the criteria BIC, $ICL - BIC$ and ICL selected $K = 2$ components to be fitted to the data in all simulations. The criterion AIC selected $K = 2$ components in 62% of the simulations and selected more than two components in the remaining simulations. The criterion CLC fitted more than two components in 82% of the simulations. For each of the Simulation Set 5 datasets, each criterion calculated for the models with partition $\{A, B, C, D, E\}$ was compared with the one calculated for the fitted models with partition $\{A, B, C, D\}$. All criteria selected the fitted model with partition $\{A, B, C, D, E\}$ for $K = 2, \ldots, 4$.

For this set of simulations, it can be seen that when the correct partitioning was used in the model, the criteria BIC, $ICL - BIC$, ICL detected the correct number of components to be fitted to the data. However, when the fitted attribute partition structure is far from the underlying attribute partition, all criteria fitted too many components to the model. Even though AIC selected a mixture model with too many components in all the fitted attribute partitions, AIC performed better than CLC at detecting the structure in the generated data.

7. Discussion

The finite mixture model is a model based clustering approach that is characterized by the form of the component densities and the number K of components. An important task when using the mixture model is choosing an appropriate form for the component distributions and assessing the number of components in the model.

The investigations have found that caution is needed when following the procedure given by Hunt and Jorgensen (1999) and using criteria to select your model. For the cancer data, the criteria ICL and its approximation $ICL - BIC$ performed in an identical manner in selecting both the number of components to be fiited to the model and the partitioning structure of the attributes. With the exception of the local independence model initially fitted to the data, criterion BIC performed similar to the two criteria, ICL and $ICL - BIC$ in selecting the number of components and the partitioning structure of the attributes. For all models fitted to this data set, AIC and CLC tended to overfit the data with more complex models. For the models fitted to the two component simulation sets, the criteria based on the integrated classification likelihood criterion ICL and its approximation $ICL - BIC$ performed in an identical manner and generally detected the correct number of components to be fitted to the model, even when the form of the component distributions was not the same as that of the generated data. For models with a simple partitioning stucture as in Simulation Sets 1 to 3, the BIC criterion always selected the same models as the criteria based on the integrated classification likelihood. However, for models where there were more than two attributes in a partition subvector, the criterion BIC could overestimate the number of components to be fitted to the model when the partitioning differed from that of the generated data. When the partition structure to be fitted in the model was identical to that of the

generated data, *BIC* always detected the correct number of components to be fitted in the mixture model. Both the AIC and the CLC tended to overfit the data with more complex models.

The investigations reported here show that the criteria based on the integrated classification likelihood have the best overall performance in detecting the form of the component distributions and the number of components to be fitted when clustering with the Multimix class of mixture models. The performance of the Bayesian information criterion in detecting the correct model to be fitted was variable, and this criterion tended to overestimate the number of components to be fitted to the model when the partition structure incorporated into the model differed greatly from that of the generated data. The Akaike information criterion and classification criterion performed in a less satisfactory way.

8. Acknowledgements

We would like to acknowledge the reviewers' helpful comments.

References

[1] Aitkin, M. and Rubin, D. B. (1985). Estimation and Hypothesis Testing in Finite Mixture Models. *J.R. Statist. Soc.* B, 47, 67–75.

[2] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, in *Second International Symposium on Information Theory*, B.N.Pretov and F. Csaki (eds.). Budapest:Academiai Kiado, 267–281.

[3] Akaike, H. (1974). A new look at statistical model identification. *IEEE Trans. Automat. Contr.*, AC–19, 716–723.

[4] Banfield, J. D. and Raftery, A. E. 1993. Model-Based Gaussian and non Gaussian Clustering. *Biometrics*, 49, 803-821.

[5] Bensmail, H. and Celeux, G. (1996) Regularised Gaussian Discriminant Analysis through Eigenvalue Decomposition. *Journal of the American Statistical Association*, 91, 1743-1748.

[6] Bensmail, H., Celeux, G., Raftery, A., and Robert, C.P. 1997. Inference in Model -based Cluster Analysis. *Statistics and Computing*, 7, 1-10.

[7] Biernacki, C., Celeux, G. and Govaert, G. (1998). *Assessing a Mixture Model for Clustering with the Integrated Classification Likelihood*. Technical Report 3521, Inria.

[8] Biernacki, C., Celeux, G. and Govaert, G. (2000). Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 22, No. 7, 719–725.

[9] Biernacki, C. and Govaert, G. (1997). Using the Classification Likelihood to Choose the Number of Clusters. *Computing Science and Statistics*, 29(2), 451-457.

[10] Biernacki, C. and Govaert, G. (1999). Choosing Models in Model-Based Clustering and Discriminant Analysis. *Journal of Statististical Computation and Simulation*, 64, 49–71.

[11] Byar, B. P. and Green, S. B., (1980). The Choice of Treatment for Cancer Patients Based on Covariate Information: Application to Prostate Cancer. *Bull. Cancer* 67, 477–490.

[12] Celeux, G. and Govaert, G. (1991). Clustering Criteria for Discrete Data and Latent Class Models, *Journal of Classification*, 8, 157–176.

[13] Celeux, G. and Govaert, G. (1995). Gaussian Parsimonious Clustering Models, *Pattern Recognition*, 28, 781-793.

[14] Celeux, G. and Soromenho, G. (1996). An Entropy Criterion for Assessing the Number of Clusters in a Mixture Model, *Journal of Classification*, 13, 195–212.

[15] Dasgupta, A. and Raftery, A.E. (1998). Detecting features in spatial point processes with clutter via model based clustering. *Journal of the American Statistical Association*, 93, 294-302.

[16] Dempster, A. P., Laird, N. M., and Rubin, D. B., (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm (with Discussion). *J. R. Statist. Soc. B* 39, 1–38.

[17] Everitt, B.S., Landau, S., Leese, M., and Stahl, D. (2011). *Cluster Analysis*. 5th ed. Chicester, West Sussex: John Wiley and Sons.

[18] Feng, .D. and McCulloch, C.E. (1996). Using Bootstrap Likelihood Ratios in Finite Mixture Models. *Journal of Royal Statistical Society Series B*,58, 609–617.

[19] Fraley, C. and Raftery, A.E. (1998). How many clusters? Which Clustering Method? Answers via Model-Based Cluster Analysis. *Computer Journal*, 41, 578–588.

[20] Fraley, C. and Raftery, A.E. (2002). Model based clustering, Discriminant Analysis and Density Estimation. *Journal of the American Statistical Association*, 97, 611–631.

[21] Galimberti, G. and Soffritti, G. (2013). Using conditional independence for parsimonious model-based Gaussian clustering. *Statistics and Computing*, 23, 625–638.

[22] Ghosh, J. K. and Sen, P. K. (1985). On the asymptotic performance of the log likelihood ratio test statistic for the mixure model and related results. *Proc. Berkeley Conference in Honor of Jerzy Newman and Jack Kiefer* (Vol. II), L.M. Le Cam and R.A. Olshen (Eds.). Monterey: Wadsworth, 789–806.

[23] Hartigan, J. A. (1977). Distribution problems in clustering. In *Classification and Clustering*, J. Van Ryzin (Ed.). New York: Academic Press, 45–71.

[24] Hartigan, J. A. (1985a). A failure of likelihood asmptotics for normal mixtures. *Proc. Berkeley Conference in Honor of Jerzy Newman and Jack Kiefer* (Vol. II), L. M. Le Cam and R. A. Olshen (Eds.). Monterey: Wadsworth. 807–810.

[25] Hartigan, J. A. (1985b). Statistical theory in clustering. *J. Classification*, 2, 63–76.

[26] Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer-Verlag New York.

[27] Hathaway, R. J. (1986). Another interpretation of the EM algorithm for Mixture Distributions. *Statistics & Probability Letters*, 4, 53–56.

[28] Hunt, L.A. 1996. *Clustering using Finite Mixture Models*. PhD thesis, Dept. of Statistics, University of Waikato, New Zealand.

[29] Hunt, L. A. and Jorgensen, M. A., 1999. Mixture Model Clustering Using the Multimix Program. *Austral. & New Zealand J. Statist*. 41, 153–171.

[30] Jorgensen, M. A. and Hunt, L. A., 1996. Mixture Model Clustering of Data Sets with Categorical and Continuous Variables. In *Proceedings of the Conference on Information, Statistics and Induction in Science, Melbourne, 1996*, 375–384.

[31] Krzanowski, W.J. (1983). Distance between populations using mixed continuous and categorical variables. *Biometrika*, 70, 235–243.

[32] Lo, Y.T., Mendell, N.R., and Rubin, D.B. (2001). Testing the number of components in a normal mixture. *Biometrika*, 88, 767–778.

[33] McLachlan, G.J. (1987). On Bootstrapping the Likelihood Ratio Test Statistic for the Number of components in a Normal Mixture. *Applied Statistics*, 36, 318324.

[34] McLachlan, G.J. and Basford, K.E. (1988). *Mixture Models: Inference and Applications to Clustering*. New York : Dekker.

[35] McLachlan, G. J. and Chang, S.U. (2004). Mixture modelling for cluster analysis. *Statistical Methods in Medical Research* 13, 5 347–61.

[36] McLachlan, G.J. and Krishnan, T. (2008). *The EM algorithm and Extensions.* 2nd Edn., New Jersey: Wiley.

.

[37] McLachlan, G. J. and Ng, S. K. (2000). Assessing the Number of Components in Mixture Models, *University of Queensland Research Report*, 99. 1-17.

[38] McLachlan, G. J. and Peel, D. (1997). On a resampling approach to choosing the number of components in normal mixture models. In: Billard, L., Fisher, N.I. (Eds.), *Computing Science and Statistics: Graph-Image-Vision : Proceedings of the 28th Symposium on the Interface, Sydney, Australia*, 28, Fairfax Station, Virginia:Interface Foundation of North America, 260266.

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

[40] McLachlan, G.J. and Rathnayake, S. (2014). On the number of components in a Gaussian Mixture Model. *WIREs Data Mining Knowl. Discov.* doi: 10.1002/widm.1135

[41] Melnykov, V., (2013) Challenges in model-based clustering. *WIREs Comput. Stat.*, 5: 135–148. doi: 10.1002/wics.1248

[42] Miloslavsky, M. & van der Laan, M.J.(2003). Fitting of mixtures with unspecified number of components using cross validation distance estimate. *Computational Statistics and Data Analysis*, vol. 41, no. 3, 413–428.

[43] Nyland, K.L., Asparoutiov, T., and Muthen, B.O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: a Monte Carlo simulation study. *Struct. Equation Model A Multidiscip. J.* 25, 14: 535–569.

[44] Raftery, A.E. and Dean, N. (2006). Variable Selection for Model-Based Clustering. *Journal of the American Statistical Association*, **101**, 168178.

[45] Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics*, 6, 461–464.

[46] Stahl, D. and Sallis, H. (2012). Model Based Cluster Analysis. *WIRES Comput. Stat.*, 4:341–358. doi:10.1002/wics. 1204.

[47] Steele, R.J., and Raftery, A.E. (2010). Performance of Bayesian Model Selection Criteria for Gaussian Mixture Models. In *Frontiers of Statistical Decision Making and Bayesian Analysis*, edited by M.-H. Chen *et al*, 113–130, New York: Springer.

[48] Symons, M. J. (1981). Clustering criteria and multivariate normal mixtures. *Biometrics*, **37**, 35–43.

[49] Titterington, D. M. (1981). Contribution to the discussion of paper by M. Aitkin, D. Anderson and J. Hinde. *J.R.Statist. Soc.* A **144**, 459.

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

[51] Wolfe, J.H. (1971). A Monte Carlo Study of the Sampling Distribution of the Likelihood Ratio for Mixtures of Multinormal Distributions. *Technical Bulletin STB 72-2.* San Diego: U.S. Naval Personnel and Training Research Laboratory.