On choosing a mixture model for clustering

Joseph Ngatchou-Wandji, Jan Bulla

On choosing a mixture model for clustering

Joseph Ngatchou-Wandji
Département INFOBIOSTAT, EHESP, 35043 Rennes and INSERM 954, Université Henri Poincaré, 54505 Vandoeuvre-lès-Nancy, France
Jan Bulla
LMNO and Département de mathématiques, Université de Caen, Campus II, 14032 Caen, France

Summary

Two methods for both clustering data and choosing a mixture model are proposed. First, the unknown clusters are assessed. Then, the likelihood conditional to these clusters is written as the product of likelihoods from each cluster. AIC and BIC type-approximations are then applied, and the resulting criteria turn out to be the sum of the AIC or BIC relative to each cluster. The performances of our methods are evaluated on real data examples and numerical simulations.

Key Words: Mixtures models, clustering, AIC, BIC, ICL.

1 Introduction

Because of their ability to represent relationships in data, finite mixture models are commonly used for summarizing distributions. In cluster analysis, they can provide a framework for assessing the partitions of the data, and for choosing the number of clusters. A finite mixture model is characterized by its form (m), and the number of components K, which can be interpreted as the number of species in the population from which the data has been

collected. For optimizing a mixture, one often uses a scoring function on which the comparison between the competing models with different values of K. Such scoring functions are, for example, penalized likelihoods computing the likelihood on a single training set and provide a penalty for model complexity. The AIC (Akaike 1973, 1974) and the BIC (Schwarz 1978) criteria base on such likelihoods, as well as the algorithm provided by Figueiredo et al. (1993) for estimating a mixture model.

For assessing the number of clusters arising from a Gaussian mixture model, Biernacki & Govaert (1997, 1999) used a penalized completed likelihood (CL). However, the associated criterion tends to overestimate the correct number of clusters when there is no restriction on the mixing proportions. The reason for this shortcoming is that the CL does not penalize the number of parameters in a mixture model. A penalization is provided in a Bayesian framework by Biernacki & Govaert (2000), who proposed a criterion based on the integrated completed likelihood (ICL). Their method consists in approximating the integrated completed likelihood by a BIC. This approximation, however, suffers a lack of a theoretical justification, although their numerical simulations show satisfactory performance. Other procedures for choosing the clusterings of the data and a mixture model can be found, for instance, in Kazakos (1977), Engelman & Hartigan (1969), Bozdogan (1992), Medvedovic & Dixon (2001), Fraley & Raftery (2006), or McCullagh & Yang (2008).

In this paper, we propose two alternative approaches, based on AIC and BIC criteria applied to the classification likelihood. In a certain sense, these are close to Fraley & Raftery (2006), whose method is rather based on BIC criterion applied to the mixture likelihood. Concretely, we first construct a new classification algorithm allowing to assess the clusters of the data. On the basis of this classification, we define two new criteria based on AIC- and BIC-like-approximations, which turn out to be the sum of the AIC or BIC approximations relative to each cluster plus an entropy term. On the one hand, these methods avoid a number of technical difficulties encountered by

ICL. On the other hand, the iterative estimation algorithm converges quickly in general, and thus the computational load is rather low.

This paper is organized as follows. In Section 2, we summarize some methods for clusterings. Subsequently, Section 3 recalls some existing methods for choosing a mixture, and we describe our new approaches. Finally, Section 4 contains a real-data example and a simulation study to evaluate the performance of our methods.

2 Model-based clustering

A d-variate finite mixture model assumes that the data $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathbb{R}^{dn}$ are a sample from a probability distribution with density of the form

$$f(\mathbf{u}|m, K, \theta) = \sum_{k=1}^{K} p_k \phi_k(\mathbf{u}|\mathbf{a}_k), \quad \mathbf{u} \in \mathbb{R}^d, \tag{1}$$

where the p_k represent mixing proportions, the $\phi_k(\cdot|\mathbf{a}_k)$ are density functions, each with a known form and depending on the parameter vector $\mathbf{a}_k$. Finally, $\theta := (\theta_1, \theta_2) := ((p_1, \ldots, p_K), (\mathbf{a}_1, \ldots, \mathbf{a}_K))$ represents the full parameter vector of the mixture (m, K) at hand. The most popular mixture is the Gaussian mixture model, where the $\phi_k(\cdot|\cdot)$ are the same Gaussian density with mean μ_k and covariance matrix Σ_k. More precisely, $\phi_k(\cdot|\mathbf{a}_k) = \phi(\cdot|\mathbf{a}_k)$ is a d-variate Gaussian density with $\mathbf{a}_k = (\mu_k, \Sigma_k)$ for $k = 1, \ldots, K$.

It is well known that the mixture model can be seen as an incomplete data structure model, where the complete data is

$$\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n) = ((\mathbf{x}_1, \mathbf{z}_1), \ldots, (\mathbf{x}_n, \mathbf{z}_n)),$$

with $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ standing for the missing data Titterington et al. (1985). Note that $\mathbf{z}_i = (\mathbf{z}_{i1}, \ldots, \mathbf{z}_{iK})$ is a K-dimensional vector such that $\mathbf{z}_{ik}$ takes

the value 1 if $\mathbf{x}_i$ arises from the component k, and takes the value 0 if not for $i = 1, \ldots, n$. It is clear that the vector $\mathbf{z}$ defines a partition $C = \{C_1, \ldots, C_K\}$ of the data $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, with $C_k = \{\mathbf{x}_i | \mathbf{z}_{ik} = 1, i = 1, \ldots, n\}$. If $\mathbf{z}$ was observed, the clusters would be known and the data in each class C_k could be assumed to be drawn from a distribution with density $\phi_k(\cdot; \mathbf{a}_k)$. Therefore, the likelihood conditional on $\mathbf{z}$ would have a form allowing for easy inference. Unfortunately, $\mathbf{z}$ is in general not observed and has to be estimated.

There are many ways for estimating $\mathbf{z}$. For instance, Rayment (1972), Kazakos (1977), Scott & Symons (1971), Symons (1981) treat the vector $\mathbf{z}$ as a parameter, which is estimated jointly with K and θ by maximizing the likelihood function

$$f(\mathbf{x}, \mathbf{z} | m, K, \theta) = \prod_{i=1}^{n} f(\mathbf{x}_i, \mathbf{z}_i | m, K, \theta), \tag{2}$$

where

$$f(\mathbf{x}_i, \mathbf{z}_i | m, K, \theta) = \prod_{k=1}^{K} p_k^{\mathbf{z}_{ik}} [\phi_k(\mathbf{x}_i | \mathbf{a}_k)]^{\mathbf{z}_{ik}}, \ i = 1, \ldots, n. \tag{3}$$

The drawback of this method is that all possible clusterings of the data in K groups are considered, which may be computationally costly. Additionally, Marriott (1975) points out an inconsistency of the parameter estimates, and $\mathbf{z}$ is formally treated as a parameter rather than a vector of missing observations. A Bayesian estimator of $\mathbf{z}$ is also defined in Symons (1981). Another more popular method is the so-called MAP (maximum a posteriori) method, described as follows. For $i = 1, \ldots, n$ and $k = 1, \ldots, K$, let $t_{ik}(\theta)$ denote the conditional probability that $\mathbf{x}_i$ arises from the kth mixture component. Then, one can easily show that

$$t_{ik}(\theta) = \frac{p_k \phi_k(\mathbf{x}_i | \mathbf{a}_k)}{\sum_{\ell=1}^{K} p_\ell \phi_\ell(\mathbf{x}_i | \mathbf{a}_\ell)}. \tag{4}$$

Let $\hat{\theta}$ be the maximum likelihood estimate of θ. Under some regularity

conditions the so-called EM algorithm Titterington et al. (1985) allows the computation of this estimator, by means of which, $\mathbf{z}_{ik}$ can be derived by

$$\hat{\mathbf{z}}_{ik} = \begin{cases} 1 & \text{if } \arg\max_{\ell \in \{1,\ldots,K\}} t_{i\ell}(\hat{\theta}) = k \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

for $i = 1, \ldots, n$ and $k = 1, \ldots, K$. For more approaches for estimating $\mathbf{z}$, see, e.g., McLachlan (1992), Medvedovic & Dixon (2001), Fraley & Raftery (2006), or McCullagh & Yang (2008). The estimates $\hat{\mathbf{z}}$ provided by either of these methods serve for determining the clusters of the data. Based on these clusters, it is possible to express likelihood for further inference. In the following section, we propose a new clustering algorithm based on the so-called classification likelihood.

3 Choosing a mixture model

3.1 Existing methods

Several methods exist for choosing a mixture model among a given number of models. One of these, consisting in maximizing (2) has already been recalled and commented in the previous section (see Symons 1981, for details). However, the most popular approaches base on the AIC and BIC criteria, as well as their extensions or other criteria such as presented by Figueiredo et al. (1993). In a Bayesian framework, one selects the model having the largest posterior probability. This is tantamount to choosing the model with the largest integrated completed likelihood (ICL), provided that all the models have equal prior probabilities (see, e.g., Biernacki & Govaert 2000). This corresponds to the model $(\hat{m}, \hat{K})$ such that

$$(\hat{m}, \hat{K}) = \arg\max_{m,K} f(\mathbf{x}, \mathbf{z}|m, K),$$

where

$$f(\mathbf{x}, \mathbf{z}|m, K) = \int_{\Theta_{m,K}} f(\mathbf{x}, \mathbf{z}|m, K, \theta)\pi(\theta|m, K)d\theta, \tag{6}$$

with parameter space $\Theta_{m,K}$, $\pi(\theta|m, K)$ a non-informative or weakly informative prior distribution on $\theta \in \Theta_{m,K}$ for the same model, and $f(\mathbf{x}, \mathbf{z}|m, K, \theta)$ the likelihood function defined by (2). From a BIC-like approximation of the right-hand side of (6), the Biernacki & Govaert (2000) propose to select the model which maximizes

$$\log f(\mathbf{x}, \mathbf{z}|m, K, \hat{\theta}^*) - \frac{d_{m,K}}{2}\log(n), \tag{7}$$

where $d_{m,K}$ stands for the dimension of the space $\Theta_{m,K}$, and $\hat{\theta}^* = \arg\max_\theta f(\mathbf{x}, \mathbf{z}|m, K, \theta)$. Since $\mathbf{z}$ is not observed, it is substituted by $\hat{\mathbf{z}}$ given by (5), and $\hat{\theta}$ is utilized instead of $\hat{\theta}^*$ in the above formula. Thus, their ICL criterion selects $(\hat{m}, \hat{K})$ maximizing

$$\text{ICL}(m, K) = \log f(\mathbf{x}, \hat{\mathbf{z}}|m, K, \hat{\theta}) - \frac{d_{m,K}}{2}\log(n). \tag{8}$$

It is important to note that approximation (7) is not valid in general for mixture models. Moreover, even if this approximation was valid, the accuracy of (8) obtained by substituting $\mathbf{z}$ for $\hat{\mathbf{z}}$ and $\hat{\theta}$ for $\hat{\theta}^*$ may be hard to quantify.

3.2 Some new approaches

In the sequel, we adopt different techniques for finding the mixture model leading to the greatest evidence of clustering given data $\mathbf{x}$. Our approaches consist in first assessing the clusters, and secondly applying AIC-/BIC-like criteria to the likelihood derived from these clusters. More precisely, we consider the likelihood defined by equation (2), given that the vector $\mathbf{z}$ and thus $\theta_1 = (p_1, \ldots, p_K)$ are assumed to be known. Indeed, with this assumption it is easy to derive that the resulting conditional likelihood can be expressed

as a product of the likelihoods of each component of the mixture model to which AIC or BIC approximations can be applied. Moreover, the estimators of the parameters involved are consistent if the number of observations per cluster is large enough, which follows directly from MLE theory.

Assume $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ being given, so the data are partitioned into K classes $C_1, C_2, \ldots, C_K$. Moreover, let $n_k = \sum_{i=1}^{n} z_{ik} = |C_k|$ for all $k = 1, \ldots, K$, $i = 1, \ldots, n$, where z_{ik} is the kth component of $\mathbf{z}_i$. Then, the p_k can be consistently estimated by the natural estimators $\hat{p}_k = n_k/n$, which are also asymptotically normal. Thus, a consistent and asymptotically normal estimator of θ_1 is given by $\hat{\theta}_1 = (\hat{p}_1, \ldots, \hat{p}_K)$. Then, the likelihood and log-likelihood functions of θ_2 given $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ can be approximated by

$$\ell(m, K, \theta_2|\theta_1, \mathbf{z}) = \prod_{k=1}^{K} \prod_{\mathbf{x}_j \in C_k} \hat{p}_k \phi_k(\mathbf{x}_j|\mathbf{a}_k), \tag{9}$$

$$L(m, K, \theta_2|\theta_1, \mathbf{z}) = \sum_{k=1}^{K} \left(\sum_{\mathbf{x}_j \in C_k} \log(\phi_k(\mathbf{x}_j|\mathbf{a}_k) + n_k \log \hat{p}_k \right). \tag{10}$$

It remains the estimation of $\theta_2 = (\mathbf{a}_1, \ldots, \mathbf{a}_K)$ from (9) or (10). Note that the estimator of $\mathbf{a}_k$ depends only on the n_k observations within the kth group C_k for $k = 1, \ldots, K$. Henceforth, we denote $\ell(m, K, \theta_2|\theta_1, \mathbf{z})$ by $\ell(m, K, \theta_2)$ and $L(m, K, \theta_2|\theta_1, \mathbf{z})$ by $L(m, K, \theta_2)$.

Let $d_{\mathbf{a}_k}$ denote the length of the vector $\mathbf{a}_k$, and $\Theta_{m,K}^{(k)} \subset \mathbb{R}^{d_k}$ for all $k = 1, \ldots, K$. In what follows, we suppose that $\theta_2 \in \Theta_{m,K}^{*} = \Theta_{m,K}^{(1)} \times \ldots \times \Theta_{m,K}^{(K)}$, and that the $\phi_k(\cdot|\mathbf{a}_k)$ are identifiable and differentiable up to order 2. Then, the integrated likelihood is defined by

$$\begin{aligned} \ell(m, K) &= \int_{\Theta_{m,K}^{*}} \ell(m, K, \theta_2) \pi(\theta_2|m, K) d\theta_2 \\ &= \prod_{k=1}^{K} \hat{p}_k \int_{\Theta_{m,K}^{(k)}} \prod_{\mathbf{x}_j \in C_k} \phi_k(\mathbf{x}_j|\mathbf{a}_k) \pi_k(\mathbf{a}_k|m, K) d\mathbf{a}_k, \end{aligned} \tag{11}$$

which follows from the likelihood function of θ_2 defined by (9).

Theorem 1 *Assume that* $\mathbf{z}$ *is known, and that the* n_k*'s are large enough for* $k = 1, 2, \ldots, K$*. Then, the following approximation for the log-likelihood function holds:*

$$L(m, K, \theta_2) \approx \sum_{k=1}^{K} \left(\sum_{\mathbf{x}_j \in C_k} \log \phi_k(\mathbf{x}_j | \hat{\mathbf{a}}_k) - d_{\mathbf{a}_k} \right) + \sum_{k=1}^{K} n_k \log \hat{p}_k. \tag{12}$$

Proof. Given $\mathbf{z}$, the deviance of the model can be approximated by

$$\begin{aligned} & L(m, K, \theta_2) - \sum_{k=1}^{K} \left(\sum_{\mathbf{x}_j \in C_k} \log \phi_k(\mathbf{x}_j | \hat{\mathbf{a}}_k) + n_k \log \hat{p}_k \right) \\ = & \sum_{k=1}^{K} \sum_{\mathbf{x}_j \in C_k} \Big[\log \phi_k(\mathbf{x}_j | \mathbf{a}_k) - \log \phi_k(\mathbf{x}_j | \hat{\mathbf{a}}_k) \Big], \end{aligned}$$

which is the sum of the deviances relative to the components of the mixture. As n_k is large, it follows

$$\sum_{\mathbf{x}_j \in C_k} \Big[\log \phi_k(\mathbf{x}_j | \mathbf{a}_k) - \log \phi_k(\mathbf{x}_j | \hat{\mathbf{a}}_k) \Big] \approx -d_{\mathbf{a}_k}.$$

for each $k = 1, \ldots, K$ (see, e.g., Akaike 1974). □

Theorem 2 *Assume that* $\mathbf{z}$ *is known, that the* n_k *are large enough for* $k = 1, 2, \ldots, K$*, and that the prior on* θ_2 *has the form*

$$\pi(\theta_2 | m, K) = \pi_1(\mathbf{a}_1 | m, K) \times \ldots \times \pi_K(\mathbf{a}_K | m, K). \tag{13}$$

Then, the the logarithm of the integrated likelihood can be approximated by

$$\log \ell(m, K) \approx \sum_{k=1}^{K} \left(\sum_{\mathbf{x}_j \in C_k} \log \phi_k(\mathbf{x}_j | \hat{\mathbf{a}}_k) - \frac{1}{2} d_{\mathbf{a}_k} \log(n_k) \right) + \sum_{k=1}^{K} n_k \log \hat{p}_k. \tag{14}$$

Proof. See the appendix.

The first terms on the right-hand sides of (13) and (14) look like sums of AIC and BIC respectively, and depend on $\mathbf{z}$ and θ_1. Therefore, we denote these quantities by $\text{SAIC}(m, K|\theta_1, \mathbf{z})$ and $\text{SBIC}(m, K|\theta_1, \mathbf{z})$, which stands for "Sum of AIC/BIC". They can be written as

$$\text{SAIC}(m, K|\theta_1, \mathbf{z}) = \sum_{k=1}^{K} \left(\sum_{\mathbf{x}_j \in C_k} \log \phi_k(\mathbf{x}_j | \hat{\mathbf{a}}_k) + n_k \log \hat{p}_k - d_{\mathbf{a}_k} \right) \tag{15}$$

$$\text{SBIC}(m, K|\theta_1, \mathbf{z}) = \sum_{k=1}^{K} \left[\sum_{\mathbf{x}_j \in C_k} \log \phi_k(\mathbf{x}_j | \hat{\mathbf{a}}_k) + n_k \log \hat{p}_k - \frac{d_{\mathbf{a}_k}}{2} \log(n_k) \right]. \tag{16}$$

Before describing a technique for model selection based on equation (16) respectively (16), we first provide an algorithm for parameter estimation given number of clusters, say $\tilde{k}$. Denote the corresponding missing data by $\mathbf{z}_K$, and the corresponding parameter vector θ_1 by θ_{1K}. The algorithm is given by

- Define ϕ_k, $k = 1, \ldots, K$
- Initialize $\mathbf{z}$ (for example, by the K-means algorithm)
- **Repeat**

- for $k = 1, \ldots, K$, compute $n_k = \sum_{i=1}^{n} z_{ik}$ and $p_k = \frac{n_k}{n}$, thus $\theta_{1K} = (p_1, \ldots, p_K)$
- maximize the log-likelihood (10) with respect to $\theta_2 = (\mathbf{a}_1, \ldots, \mathbf{a}_K)$ and denote θ_{2K} the vector for which the likelihood reaches the maximum.
- for $i = 1, \ldots, n$ and $k = 1, \ldots, K$, compute

$$\mathbf{z}_{ik} = \begin{cases} 1 & \text{if } \arg\max_{\ell \in \{1,\ldots,K\}} t_{i\ell}(\theta_K) = k \\ 0 & \text{otherwise} \end{cases},$$

where $\theta_K = (\theta_{1K}, \theta_{2\tilde{k}})$ and t_{ik} is defined by (4)

Until the log-likelihood remains constant

- Return $\mathbf{z}_K$ and θ_{1K}.

For choosing the model, and thus determining the number of clusters and their parameters, we propose to proceed as follows.

- Set the maximum number of components $K_{\max}$
- For $K = 2, \ldots, K_{\max}$
  - Compute θ_{1K} and $\mathbf{z}_K$ (with the above algorithm)
  - Compute $\mathrm{SAIC}(m, K|\theta_{1K}, \mathbf{z}_K)$ or $\mathrm{SBIC}(m, K|\theta_{1K}, \mathbf{z}_K)$
- Select $(\hat{m}, \hat{K})$ and $\mathbf{z}_{\hat{K}}$ by

$$(\hat{m}, \hat{K}) = \arg\max_{m,K} \mathrm{SAIC}(m, K|\theta_{1K}, \mathbf{z}_K).$$

or

$$(\hat{m}, \hat{K}) = \arg\max_{m,K} \mathrm{SBIC}(m, K|\theta_{1K}, \mathbf{z}_K).$$

Table 1: Model selection by SAIC/SBIC

This table displays log-likelihood, SAIC, and SBIC of the estimated models with 2, 3, and 4 components, initialized by K-Means or random paths.

	kmeans			random		
comp.	2	3	4	2	3	4
logL	-1131	-1152	-1148	-1131	-1125	-1120
SAIC	-1141	-1167	-1168	-1141	-1140	-1140
SBIC	-1155	-1186	-1188	-1155	-1157	-1158

4 Examples

In the first part of this section, we present and application of our clustering algorithm and model selection criteria to data from the Old Faithful Geyser (Yellowstone National Park, USA). After this, a simulation study compares the performance of our algorithm to a second model-based clustering approach.

4.1 Old Faithful Geyser

The data analyzed are waiting times between eruptions and the durations of the eruption for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA. This data set with 272 observations is included in the `datasets` package of the `R` software (R Development Core Team 2010). In order to initialize our clustering algorithm, termed mb1 in the following, we followed two approaches. On the one hand, we initialized the **z** by the K-Means algorithm (function `kmeans` in `R`, started by 100 different random sets). On the other hand, we simply generated 1000 random paths for initialization and ran mb1 for each of them. The models estimated have 2, 3, and 4 components, Table 1 summarizes the resulting values of log-likelihood, SAIC, and SBIC.

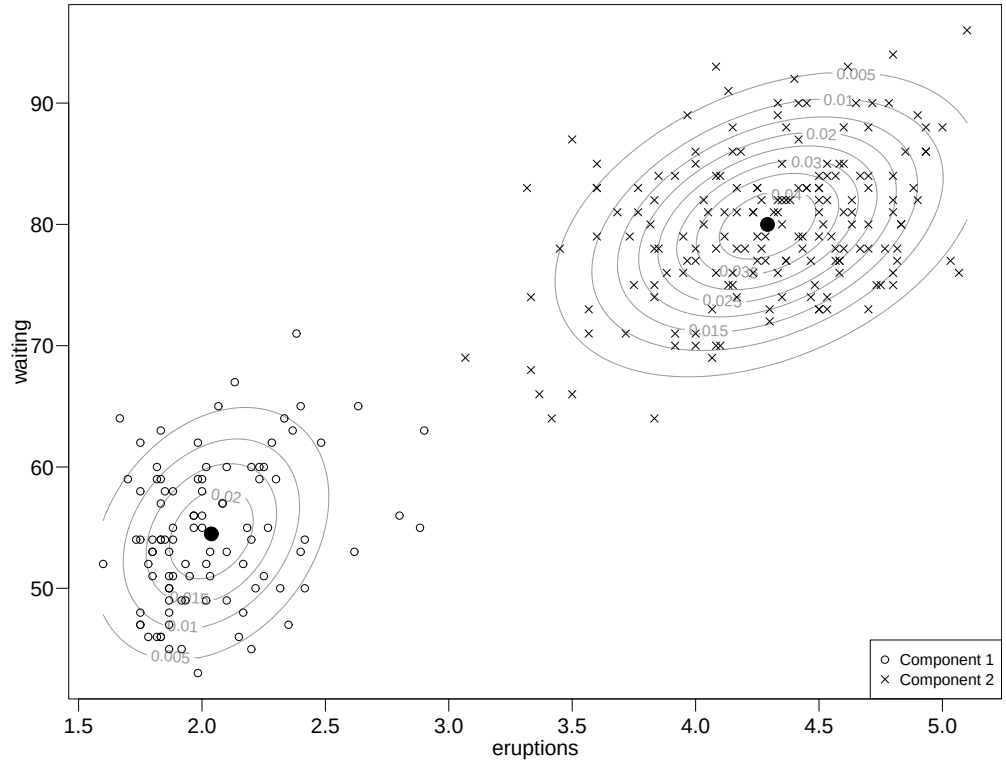
The SAIC/SBIC values of those models initialized by K-means show a clear preference for the model with two components, which also exhibits the highest likelihood. As to the models initialized by the random samples, similar tendencies arise. The SAIC is almost constant and thus the parsimonious two components may be selected, and the SBIC prefers two components. For comparison with a standard algorithm for model-based clustering, we also fitted models using the `R` package `mclust` (Fraley & Raftery 2006). This algorithm, termed mb2 in the following, initializes by hirarchichal clustering and selects an appropriate model by the BIC. The result of the mb2 is a model with 3 components, which might be attributed to "model deviation to normality in the two obvious groups rather than a relevant additional group" (Biernacki & Govaert 2000). Thus, the selection of the SBIC corresponds to that of the ICL criterion of the before mentioned authors, selecting 2 components as well. Figure 1 displays the data, the estimated densities of the two components and the mapping of the observations to the components.

The estimated parameters are $\mu_1 = (2.04, 54.5)'$, $\mu_2 = (4.29, 80.0)'$, $\sigma_1^2 = \begin{pmatrix} 0.0712 & 0.452 \\ 0.452 & 34.1 \end{pmatrix}$, and $\sigma_2^2 = \begin{pmatrix} 0.169 & 0.918 \\ 0.918 & 35.9 \end{pmatrix}$. The estimated values of $\mathbf{z}$ indicate that 35.7% and 64.3% of the observations belong to the respective components.

Additionally to the model selection, it may be noted that the number of iterations required by the algorithm is rather manageable in the majority of cases. Considering the random initializations, the third quartile of the numberer of iterations lies at 14, 16, and 15 for models with 2, 3, and 4 components, respectively. For models with two components, mb1 failed to converge in 12% of the cases, which can be attributed to very poor initialization of the components. The algorithm converged to the maximum likelihood of -1131 in 69.6% of the cases. The results are less satisfactory for 3/4 componentes, which may be viewed keeping in mind "Garbage in, garbage out". First, almost all estimated models are (slightly) different to each other. Moreover,

Figure 1: Clustering of Old Faithful Geyser data

The figure shows bivariate data from the Old Faithful Geyser, clustered by mb1. The preferred model has two components, the centers of which are marked by filled circles. Contours result from the two estimated Gaussian densities.



in 48%/76% of the samples the algorithm does not converge properly, determines components with very few observation (< 10), or estimates two or more components with (almost) identical parameters. This behavior may, however, underline the preference for the model with two components.

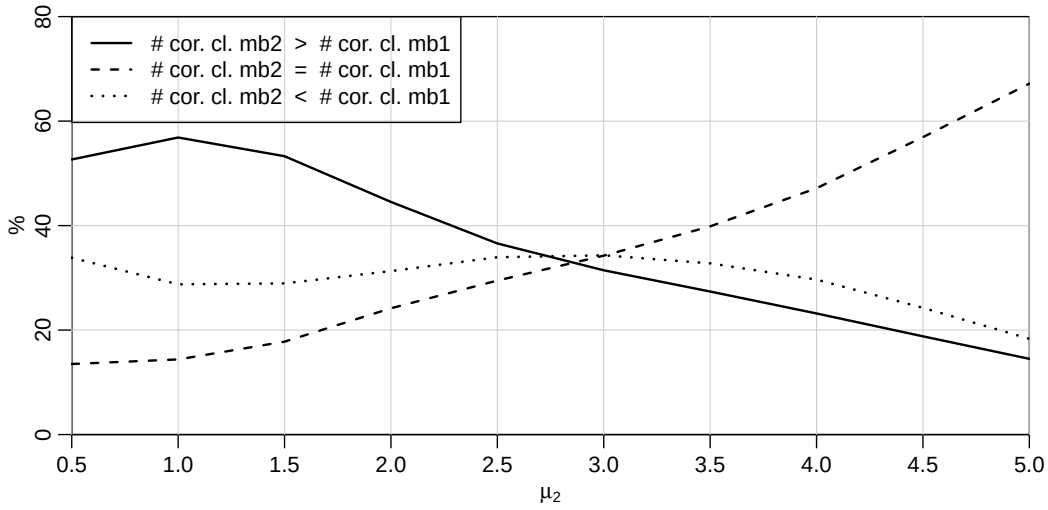
4.2 Simulation study

The aim of this simulation study is to compare the classification performance of our algorithm and the mb2 by Fraley & Raftery (2006). The two main questions addressed are: Firstly, is mb1 in general or under certain conditions able to correctly classify a higher number of observations than mb2? Secondly, is there a significant difference between the number of observations correctly classified by mb1 and mb2?

The study deals with samples of different sizes, that is, $n \in \{100, 250, 500, 1000, 5000\}$. For every sample size, we generated samples from a mixture of two Gaussian distributions with mixing probabilities equal to 0.5. The mean μ_1 of the first density always equals 0, whereas the mean μ_2 of the second density takes values in $\{0.5, 1, \ldots, 5\}$; both variances are equal to 1. For each combination of n and μ_2, we generated 10000 samples and carried out mb1 and mb2. The algorithm mb2 was executed utilizing the default settings of the procedure `Mclust`. Thus the initialization is by hierarchical clustering, once without and once with the true means as prior. The initialization of mb1 based on the K-means algorithm, division of the sample by the median, and the classification results of mb2 (with and without means as prior). For both algorithms, the highest log-likelihood determined the preferred classification result.

Note that splitting up the subsequent analysis by the usage/non-usage of prior information on the means has no effect on the results and is therefore omitted. For this study, the version of `R` is 2.9.2, and `Mclust` has version number 3.3.1. All code is available from the authors on request.

Figure 2: Comparing classification performance of mb1 and mb2
The figure displays the relative classification performance of mb1 and mb2. Averaged over all samples and sample sizes, the solid line represents the proportion samples in which mb2 correctly classifies a higher number of observations than mb1 for $\mu_2 = 0.5, 1, \ldots, 5$. The dashed line concerns the converse case, and the dotted lines represents an identical number of correct classifications.



As to the first question, Figure 2 sheds light on comparing the classification performance of mb1 and mb2. The solid line represents the average proportion of samples which for mb1 correctly classifies a higher number of observations than mb2, for all sample sizes. The solid line indicates an identical number of correct classification for mb1 and mb2, whereas the dashed line represents the case that mb2 correctly classifies a higher number of observations than mb1.

The tendency is clear: For smaller values of μ_2, i.e., means lying closer together, mb1 seems to correctly classify more than or the same number of observations as mb2. The further the densities separate, the more often mb2 attains a higher number of classifications than mb1. This holds true for samples of all sizes considered, as Table 2 shows.

Additionally to the findings above, it becomes clear that the algorithms tend

Table 2: Comparing classification performance of mb1 and mb2

The table presents the proportion of samples for which the algorithm mb1 and mb2, respectively, correctly classify the highest number of observations in the sample. For every sample size $n = 100, 250, \ldots, 5000$, the number of samples equals 10000.

n	alg.	μ_2= 0.5	1	1.5	2	2.5	3	3.5	4	4.5	5
100	mb1	49.4	55.1	52.8	45.2	35.8	27.1	19.9	13.5	9.3	6.2
	mb2	29.1	23.1	22.7	24.2	25.6	25.9	24.0	20.1	14.1	9.6
250	mb1	53.7	57.4	55.0	46.0	36.7	30.1	24.8	19.2	14.3	10.2
	mb2	31.5	26.5	25.4	28.5	31.6	31.7	29.8	26.4	19.7	14.1
500	mb1	54.6	58.7	54.8	44.7	36.2	31.6	28.3	23.5	18.3	12.9
	mb2	32.6	28.2	28.7	31.4	34.9	35.3	33.1	29.8	24.1	17.8
1000	mb1	54.1	58.0	53.6	43.3	37.0	32.7	29.4	26.8	22.3	17.5
	mb2	35.8	30.5	31.1	35.1	37.4	38.1	36.4	33.2	28.1	21.3
5000	mb1	51.5	55.1	50.1	43.4	37.2	35.7	34.6	32.8	29.8	25.8
	mb2	40.3	35.5	36.7	37.1	40.2	40.5	40.6	38.8	35.3	28.8

to produce different classification results for an increasing number of observations, which can be expected.

With respect to the second question for a significant difference between the number of correctly classified observations by mb1 and mb2, Figure 3 shows the results for sample sizes 100, 250, and 500 in the upper, middle, and lower panel, respectively. In each of the panels, two box plots display the proportion of correctly classified observations by mb1 and mb2 for every value of μ_2. The left box plot results from classifications obtained by mb1 and the right box plot refers to mb2. Additionally, the Wilcoxon signed-rank test with alternative of a positive location shift is applied with 0.1%-level for every value of μ_2. Grey box plots indicate that the hypothesis of no shift can be rejected, which is the case for all $\mu_2 \leq 2$. For the smaller samples with 100 and 250 observations the shift is even present for $\mu_2 = 2.5$. We do not provide the corresponding figures for $n = 1000$ and $n = 5000$, because the visual difference is too small. However, for both sample sizes a highly

significant positive shift is present for all $\mu_2 \leq 2$.

Summarizing, mb1 seems to provide a classification performance which is particularly interesting in cases where the separation of the mixing distributions is not too obvious, that is small sample sizes and/or close mean values.

5 Appendix: Proof of Theorem 2

Let $k \in \{1, \ldots, K\}$, and denote

$$\Lambda^{(k)}(m, K) = \int_{\Theta^{(k)}_{m,K}} \prod_{\mathbf{x}_j \in C_k} \phi_k(\mathbf{x}_j|\mathbf{a}_k)\pi_k(\mathbf{a}_k|m, K)d\mathbf{a}_k$$

and

$$g(\mathbf{a}_k) = \sum_{\mathbf{x}_j \in C_k} \log \phi_k(\mathbf{x}_j|\mathbf{a}_k) + \log \pi_k(\mathbf{a}_k|m, K).$$

Moreover, define the vector $\mathbf{a}^*_k$ and the matrix $A_{\mathbf{a}^*_k}$ as follows:

$$\mathbf{a}^*_k = \arg \max_{\mathbf{a}_k \in \Theta^{(k)}_{m,K}} \left(\frac{1}{n_k} g(\mathbf{a}_k) \right)$$

and

$$A_{\mathbf{a}^*_k} = -\frac{1}{n_k} \left(\frac{\partial^2 g(\mathbf{a}^*_k)}{\partial \mathbf{a}^{(i)}_k \partial \mathbf{a}^{(j)}_k} : 1 \leq i, j \leq d_{\mathbf{a}_k} \right).$$

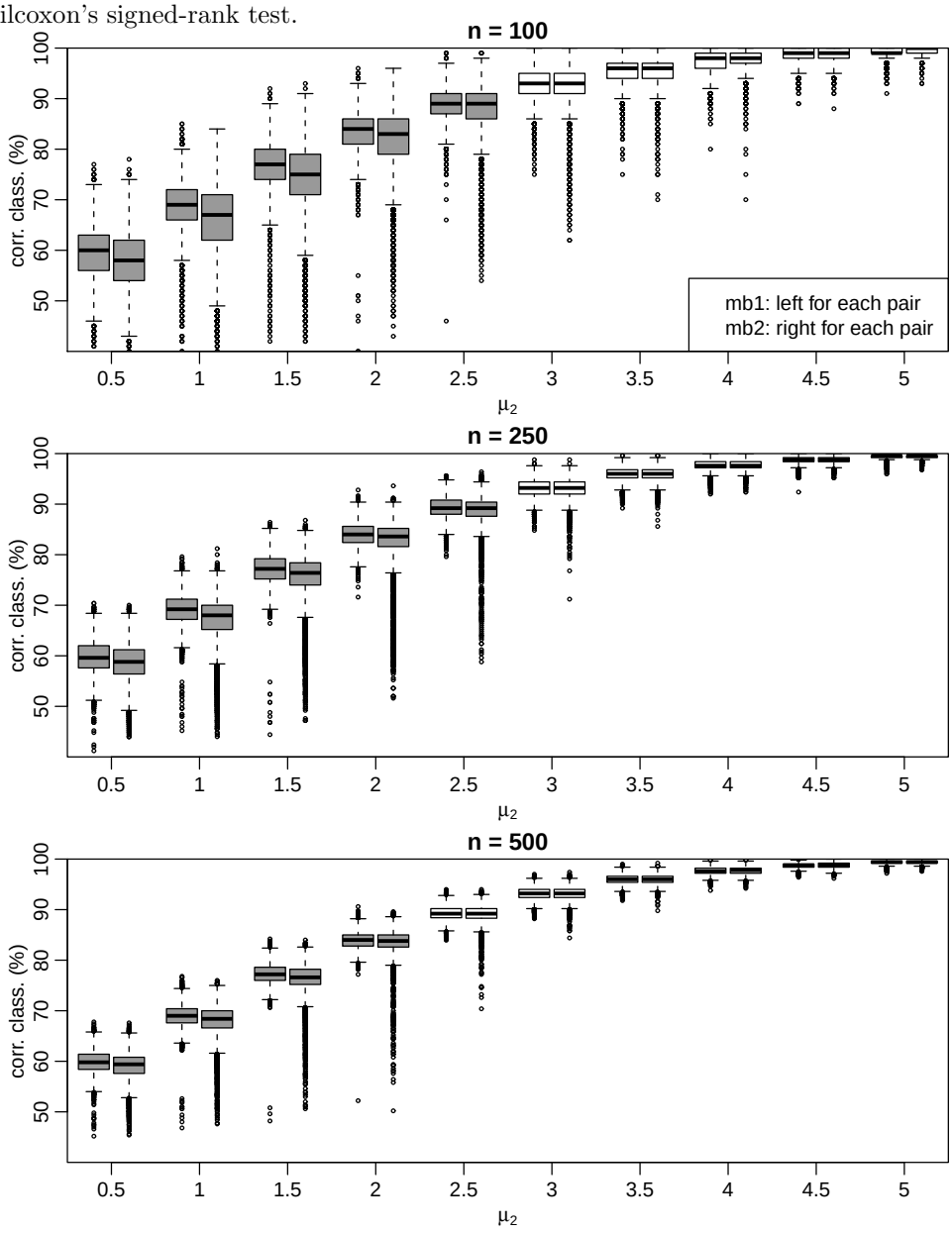
Then we obtain

$$\begin{aligned} \Lambda^{(k)}(m, K) &= \int_{\Theta^{(k)}_{m,K}} \exp[g(\mathbf{a}_k)]d\mathbf{a}_k \\ &= \exp[g(\mathbf{a}^*_k)] \left(\frac{2\pi}{n_k} \right)^{d_{\mathbf{a}_k}/2} |A_{\mathbf{a}^*_k}|^{-1/2} + O(n_k^{-1/2}). \end{aligned}$$

by the Laplace transformation (see, e.g., Kass & Raftery 1995). Since $\log g(\mathbf{a}_k)$ behaves like the product $\prod_{\mathbf{x}_j \in C_k} \phi_k(\mathbf{x}_j|\mathbf{a}_k)$ for all $k = 1, \ldots, K$, which in-

Figure 3: Relative classification performance of mb1 and mb2

The figure shows the relative classification performance of mb1 and mb2, measured by the proportion of correctly classified observations. The upper panel, middle, and lower panel concern the sample sizes 100, 250, and 500, respectively. For every $\mu_2 = 0.5, 1, \ldots, 5$, the box plot on the left hand side summarizes the proportion of correctly classified observations in 10000 samples by mb1. The box plot on the right hand side results from mb2. Grey shaded boxes indicate that the hypothesis of equal distributions is rejected at 1% level by Wilcoxon's signed-rank test.



creases whilst $\pi_k(\mathbf{a}_k|m,K)$ is constant, one can substitute the vector $\mathbf{a}_k^*$ by $\hat{\mathbf{a}}_k = \arg\max\{(1/n_k)\prod_{\mathbf{x}_j\in C_k}\phi_k(\mathbf{x}_j|\mathbf{a}_k)\}$ and the matrix $A_{\mathbf{a}_k^*}$ by the Fisher information matrix $I_{\hat{\mathbf{a}}_k}$ defined by

$$\begin{aligned} I_{\hat{\mathbf{a}}_k} &= -\left(\sum_{\mathbf{x}_j\in C_k} E\left[\frac{\partial^2 \log\phi_k(\mathbf{x}_j|\hat{\mathbf{a}}_k)}{\partial\mathbf{a}_k^{(i)}\partial\mathbf{a}_k^{(j)}}\right] : 1\le i,j\le d_{\mathbf{a}_k}\right) \\ &= -\left(n_k E\left[\frac{\partial^2 \log\phi_k(\mathbf{x}_j|\hat{\mathbf{a}}_k)}{\partial\mathbf{a}_k^{(i)}\partial\mathbf{a}_k^{(j)}}\right] : 1\le i,j\le d_{\mathbf{a}_k}\right). \end{aligned}$$

Then follows:

$$\begin{aligned} \log\Lambda^{(k)}(m,K) &= \sum_{\mathbf{x}_j\in C_k}\log\phi_k(\mathbf{x}_j|\hat{\mathbf{a}}_k) + \log\pi_k(\hat{\mathbf{a}}_k|m,K) - \frac{d_{\mathbf{a}_k}}{2}\log(n_k) \\ &\quad + \frac{d_{\mathbf{a}_k}}{2}\log(2\pi) - \frac{1}{2}\log(|I_{\hat{\mathbf{a}}_k}|) + O(n_k^{-1/2}). \end{aligned}$$

Neglecting the $O(n_k^{-1/2})$ and $O(1)$ terms, one obtains the approximation

$$\log\Lambda^{(k)}(m,K) \approx \sum_{\mathbf{x}_j\in C_k}\log\phi_k(\mathbf{x}_j|\hat{\mathbf{a}}_k) - \frac{d_{\mathbf{a}_k}}{2}\log(n_k).$$

Thus, from this approximation and (11) follows

$$\log\ell(m,K) \approx \sum_{k=1}^{K}\left(\sum_{\mathbf{x}_j\in C_k}\log\phi_k(\mathbf{x}_j|\hat{\mathbf{a}}_k) + n_k\log\hat{p}_k - \frac{d_{\mathbf{a}_k}}{2}\log(n_k)\right).$$

□

References

Akaike, H. (1973), *in* 'Proceedings of the Second International Symposium on Information Theory', Budapest: Akademiai Kiado 1973, pp. 267–281.

Akaike, H. (1974), 'A new look at the statistical model identification', *IEEE T. Automat. Contr.* **19**(6), 716–723.

Biernacki, C., C. G. & Govaert, G. (2000), 'Assessing a mixture model for clustering with the integrated completed likelihood', *IEEE T. Pattern Anal.* **22**(7), 719–725.

Biernacki, C. & Govaert, G. (1997), 'Using the classification likelihood to choose the number of clusters', *Comp. Sci. Stat.* **29**(2), 451–457.

Biernacki, C. & Govaert, G. (1999), 'Choosing models in model-based clustering and discriminant analysis', *J. Stat. Comput. Sim.* **64**(1), 49–71.

Bozdogan, H. (1992), *Choosing the number of components clusters in the mixture model using a new informational complexity criterion of the inverse Fisher information matrix. In O. Opitz, B. Lausen, and R. Klar, editors, Information and classification*, Springer-Verlag.

Engelman, L. & Hartigan, J. (1969), 'Percentage points for a test for clusters', *J. Amer. Statist. Assoc.* **64**, 1647.

Figueiredo, M., Leitão, J. & Jain, A. (1993), *On fitting mixture models*, Lecture Notes in Computer Science, Springer Berlin / Heidelberg.

Fraley, C. & Raftery, A. (2006), MCLUST version 3 for R: Normal mixture modeling and model-based clustering, Technical report, Department of Statistics, University of Washington. Technical Report no. 504.

Kass, R. & Raftery, A. (1995), 'Bayes factors', *J. A. Stat. Assoc.* **90**(430), 773–795.

Kazakos, D. (1977), 'Recursive estimation of prior probabilities using a mixture', *IEEE T. Inform. Theory* **23**(2), 203–211.

Marriott, F. (1975), 'Separating mixtures of normal distributions', *Biometrics* **31**(3), 767–769.

McCullagh, P. & Yang, J. (2008), 'How many clusters?', *Bayesian Analysis* **3**(1), 101–120.

McLachlan, G. (1992), *Discriminant Analysis and Statistical Pattern recognition*, Wiley Series in Probability and Statistics, John Wiley & Son.

Medvedovic, M., S. P. S. R. & Dixon, K. (2001), 'Clustering mutational spectra via classification likelihood and markov chain monte carlo algorithms', *J. Agric., Biol., Envir. St.* **6**(1), 19–37.

R Development Core Team (2010), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *http://www.r-project.org*

Rayment, P. (1972), 'The identification problem for a mixture of observations from two normal populations', *Technometrics* **14**(4), 911–918.

Schwarz, G. (1978), 'Estimating the dimension of a model', *Ann. Stat.* **6**(2), 461–464.

Scott, A. & Symons, M. (1971), 'Clustering methods based on likelihood ratio criteria', *Biometrics* **27**(2), 387–397.

Symons, M. (1981), 'Clustering criteria and multivariate normal mixtures', *Biometrics* **37**(1), 35–43.

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