Adv Data Anal Classif (2015) 9:447–460
DOI 10.1007/s11634-015-0220-z

REGULAR ARTICLE

Improved initialisation of model-based clustering using Gaussian hierarchical partitions

Luca Scrucca[1] · **Adrian E. Raftery**[2]

Received: 30 November 2014 / Revised: 3 October 2015 / Accepted: 12 October 2015 /
Published online: 26 October 2015
© Springer-Verlag Berlin Heidelberg 2015

Abstract Initialisation of the EM algorithm in model-based clustering is often crucial. Various starting points in the parameter space often lead to different local maxima of the likelihood function and, so to different clustering partitions. Among the several approaches available in the literature, model-based agglomerative hierarchical clustering is used to provide initial partitions in the popular `mclust` R package. This choice is computationally convenient and often yields good clustering partitions. However, in certain circumstances, poor initial partitions may cause the EM algorithm to converge to a local maximum of the likelihood function. We propose several simple and fast refinements based on data transformations and illustrate them through data examples.

Keywords Model-based clustering · Model-based agglomerative hierarchical clustering · `mclust` · Data transformation

Mathematics Subject Classification 62H30 (Classification and discrimination; cluster analysis)

✉ Luca Scrucca
luca@stat.unipg.it

Adrian E. Raftery
raftery@u.washington.edu

1 Dipartimento di Economia, Università degli Studi di Perugia, Via A. Pascoli 20, 06123 Perugia, Italy

2 Department of Statistics, University of Washington, Box 354322, Seattle, Washington 98195-4322, USA

1 Introduction

Model-based clustering is an increasing popular method for unsupervised learning. In contrast to classical heuristic methods, such as the original k-means proposal and hierarchical clustering, model-based clustering methods rely on a probabilistic assumption about the data distribution. According to the main underlying assumption, data are generated from a mixture distribution, where each cluster is described by one or more mixture components. Maximum likelihood estimation of parameters is usually carried out via the EM algorithm (Dempster et al. 1977). The EM algorithm is an iterative, strictly hill-climbing procedure whose performance can be very sensitive to the starting point because the likelihood surface tends to have multiple modes, although it usually produces sensible results when started from reasonable starting values. Thus, good initialisation is crucial for finding MLEs, although no method suggested in the literature uniformly outperforms the others.

In the case of Gaussian model-based clustering (Fraley and Raftery 2002), several approaches are available, both stochastic and deterministic, for selecting an initial partition of the observations, or an initial estimate of the parameters. In the `mclust` R package (Fraley et al. 2012, 2015), the EM algorithm is initialised using the partitions obtained from model-based agglomerative hierarchical clustering. Efficient numerical algorithms exist for approximately maximising the classification likelihood with multivariate normal models. However, in certain circumstances, poor initial partitions may cause the EM algorithm to converge to a local maximum of the likelihood function.

In this contribution we discuss cases where an initial partition may lead to sub-optimal maximum likelihood estimates when applied to coarse data with ties (e.g. discrete data or continuous data that are rounded in some form when measured), and we present some possible refinements to improve the fitting of such finite mixture models.

The outline of this article is as follows. Section 2 gives a brief review of background material on model-based clustering, with special attention devoted to some of the proposals available in the literature for the initialisation of EM algorithm. Section 3 discusses the model-based hierarchical agglomerative clustering method used for starting the EM algorithm. This is a very convenient and efficient algorithm, but in certain circumstances presents a serious drawback. Section 4 contains some simple transformation-based methods to refine the EM initialisation step derived from model-based agglomerative hierarchical clustering. The behaviour of these methods is illustrated through the use of real data examples in Sect. 5. The final section provides some concluding remarks.

2 Background

2.1 Model-based clustering overview

Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ be a sample of n independent identically distributed observations. The distribution of every observation is specified by a probability mass or density function through a finite mixture model of G components, which takes the following form

$$f(\boldsymbol{x};\boldsymbol{\Psi}) = \sum_{k=1}^{G} \pi_k f_k(\boldsymbol{x};\boldsymbol{\theta}_k), \tag{1}$$

where $\boldsymbol{\Psi} = \{\pi_1, \ldots, \pi_{G-1}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G\}$ are the parameters of the mixture model, $f_k(\boldsymbol{x};\boldsymbol{\theta}_k)$ is the kth component density at $\boldsymbol{x}$ with parameter(s) $\boldsymbol{\theta}_k$, $(\pi_1, \ldots, \pi_{G-1})$ are the mixing weights or probabilities (such that $\pi_k > 0$, $\sum_{k=1}^{G} \pi_k = 1$), and G is the number of mixture components.

Assuming G fixed, mixture model parameters $\boldsymbol{\Psi}$ are usually unknown and must be estimated. The log-likelihood function corresponding to Eq. (1) is given by $\ell(\boldsymbol{\Psi}; \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \sum_{i=1}^{n} \log(f(\boldsymbol{x}_i;\boldsymbol{\Psi}))$. Direct maximisation of the log-likelihood function is often complicated, so MLE of finite mixture models is usually carried out via the EM algorithm (McLachlan and Peel 2000).

In the model-based approach to clustering, each component of a finite mixture of density functions belonging to a given parametric class is associated with a group or cluster. Most applications assume that all component densities arise from the same parametric distribution family, although this need not be the case in general. A popular model assumes a Gaussian distribution for each component, i.e. $f_k(\boldsymbol{x};\boldsymbol{\theta}_k) \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. Thus, clusters are ellipsoidal, centred at the mean vector $\boldsymbol{\mu}_k$, and with other geometric features, such as volume, shape and orientation, determined by $\boldsymbol{\Sigma}_k$. Parsimonious parameterisations of the covariance matrices can be defined by means of eigen-decomposition in the form $\boldsymbol{\Sigma}_k = \gamma_k \boldsymbol{O}_k \boldsymbol{A}_k \boldsymbol{O}_k^\top$, where γ_k is a scalar controlling the volume of the ellipsoid, $\boldsymbol{A}_k$ is a diagonal matrix specifying the shape of the density contours, and $\boldsymbol{O}_k$ is an orthogonal matrix which determines the orientation of the corresponding ellipsoid (Banfield and Raftery 1993; Celeux and Govaert 1995). Fraley et al. (2012, Table 1) summarized some parameterisations of within-group covariance matrices available in the `mclust` software, and the corresponding geometric characteristics.

The number of mixture components and the parameterisation of the component covariance matrices can be selected on the basis of model selection criteria, such as the Bayesian information criterion (BIC; Schwartz 1978; Fraley and Raftery 1998) or the integrated complete-data likelihood criterion (ICL; Biernacki et al. 2000).

2.2 Initialisation of EM algorithm

The EM algorithm is an easy to implement, numerically stable algorithm, which has, under fairly general conditions, reliable global convergence. However, it may converge slowly and, like any other Newton-type method, does not guarantee convergence to the global maximum when there are multiple maxima (McLachlan and Krishnan 2008, p. 29). Further, in the case of finite mixture modelling, the estimates obtained depend on the starting values. Thus initialisation of EM is crucial because the likelihood surface tends to have multiple modes, although it usually produces sensible results when started from reasonable starting values (Wu 1983; Everitt et al. 2011, p. 150).

Several approaches are available, both stochastic and deterministic, for initialising the EM algorithm. Broadly speaking, there are two general approaches. The first one starts from some initial values for the parameters to be estimated. A simple strategy

is based on generating several candidates by drawing parameter values uniformly at random over the feasible parameters regions. Since the random-starts strategy has a fair chance of not providing good initial starting values, a common suggestion to alleviate this problem is to run the EM algorithm with several random starts and to choose the best solution. However, such a strategy can be quite time consuming and is not always practical, especially for high-dimensional datasets.

Two other stochastic initialisation schemes are the so-called *emEM* and *rndEM*. The former approach, proposed by Biernacki et al. (2003), uses several short runs of the EM initialised with valid random starts as parameter estimates until an overall number of total iterations is exhausted. Then, the solution with the highest log-likelihood is chosen to be the initialiser for the long EM, which runs until the usual strict convergence criteria are met. This approach is computationally intensive and the same comments made above about random starts apply to it also.

Two related approaches were proposed by Maitra (2009), one called *Rnd-EM* where the short EM stage is replaced by choosing multiple starting points and evaluating the log-likelihood at these values without running any EM iterations. Then, the best obtained solution serves as an initialiser for the long EM stage. The second proposal is a staged approach based on finding a large number of local modes of the dataset, and then to choose representatives from the most widely-separated ones. This approach is reported to be very time-consuming for high-dimensional data, and Melnykov and Maitra (2010) found the method to be outperformed by *emEM* and *RndEM*. Recently, Melnykov and Melnykov (2012) have proposed a strategy for initialising mean vectors by choosing points with higher concentrations of neighbours and using a truncated normal distribution for the preliminary estimation of dispersion matrices.

A second kind of approach for initialising the EM algorithm is based on the partition obtained from another clustering algorithm, e.g. k-means or hierarchical agglomerative clustering (HAC). In this case, the final classification is used to start the EM algorithm from the M-step. Unfortunately, most of these partitioning algorithm have several drawbacks, such as the need to be properly initialised or the tendency to impose specific shapes and patterns on clusters. In the popular `mclust` package for R, the EM algorithm is initialised using the partitions obtained from model-based hierarchical agglomerative clustering (MBHAC). In this approach, k clusters are obtained from a large number of smaller clusters by recursively merging the two clusters that have the smallest dissimilarity in a model-based sense, i.e. the dissimilarity used for agglomeration is derived from a probabilistic model. Banfield and Raftery (1993) proposed a dissimilarity based on a Gaussian mixture model, which is equal to the decrease in likelihood resulting by the merging of two clusters. Fraley (1998) showed how the structure of some specific Gaussian models can be exploited to yield efficient algorithms for agglomerative hierarchical clustering. More details about this approach are discussed in the following section.

3 Model-based hierarchical agglomerative clustering

The objective of MBHAC is to obtain a hierarchical ordering of clusters of n objects on the basis of the some measure of similarity among them. The result is a tree-

like structure, which proceeds from n clusters containing one object to one cluster containing all n objects by successively merging objects and clusters.

Given n objects or observations $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, let $(z_1, \ldots, z_n)^\top$ denote the classification labels, i.e. $z_i = k$ if $\boldsymbol{x}_i$ is assigned to cluster k. The unknown parameters are obtained by maximising the classification likelihood

$$\mathcal{L}_C(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G, z_1, \ldots, z_n | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \prod_{i=1}^{n} f_{z_i}(\boldsymbol{x}_i | \boldsymbol{\theta}_{z_i}).$$

Assuming the multivariate Gaussian distribution for $f_k(\boldsymbol{x}_i | \boldsymbol{\theta}_k)$, the parameters are the mean vector $\boldsymbol{\mu}_k$ and the covariance matrix $\boldsymbol{\Sigma}_k$. By imposing different covariance structures over $\boldsymbol{\Sigma}_k$, different criterion can be derived (see Fraley 1998). For instance, when the covariance is allowed to be different among clusters, the criterion to be minimised at each hierarchical stage is

$$\mathcal{W}_G = \sum_{k=1}^{G} n_k \log \left| \frac{\boldsymbol{W}_k}{n_k} \right|, \tag{2}$$

where n_k is the number of observations in group k and $\boldsymbol{W}_k$ is the sample cross-product matrix for the kth group ($k = 1, \ldots, G$). Efficient numerical algorithms are available for approximately maximising the classification likelihood (or, equivalently, minimising the corresponding criterion) with multivariate Gaussian models have been discussed.

As mentioned, the above MBHAC approach is used for starting the EM algorithm in the `mclust` R package. This is particularly convenient because the underlying probabilistic model can be shared by the initialisation step and the model fitting step. MBHAC is also computationally convenient because a single run provides the basis for initialising the EM algorithm for any number of mixture components and parameterisations of the component covariance matrices.

However, a serious problem for the MBHAC approach may arise when, at any stage, two pairs of objects attain the same minimum value for the criterion in (2). In the presence of coarse data, resulting from the discrete nature of the data or from continuous data that are rounded in some way when measured, ties must be broken by choosing the pair of entities that will be merged. This is often done at random but, regardless of which method is adopted for breaking ties, this choice can have important consequences because it changes the clustering of the remaining observations. In this case the final EM solution may depend on the ordering of the variables, and to a lesser extent on permutation of the observations (the latter case is not studied further in this paper).

This difficulty is known as the *ties in proximity problem* in the hierarchical clustering literature (see, for example, Jain and Dubes 1988, Sec. 3.2.6). This problem can also arise in other contexts, such as k-means clustering (Gordon 1999, p. 42) or partition around medoids (PAM; Kaufman and Rousseeuw 1990, p. 104).

4 Transformation-based approaches for obtaining starting partitions in model-based hierarchical agglomerative clustering

In this section we describe some simple proposals for starting the EM algorithm using the partitions obtained with MBHAC. Ideally, we would like to retain the positive aspects of such approach, but, at the same time, reduce the chance that a poor initial partition causes the EM algorithm to converge to a local maximum of the likelihood function.

The idea is to project the data through a suitable transformation before applying the MBHAC at the initialisation step. Once a reasonable hierarchical partition is obtained, the EM algorithm is run using the data on the original scale.

Let $\boldsymbol{X}$ be the $(n \times p)$ data matrix, and $\mathbb{X} = (\boldsymbol{X} - \mathbf{1}_n \bar{\boldsymbol{x}}^\top)$ be the corresponding centred matrix, where $\bar{\boldsymbol{x}} = (\bar{x}_1, \ldots, \bar{x}_p)^\top$ is the vector of sample means, and $\mathbf{1}_n$ is the unit vector of length n. Let $\widehat{\boldsymbol{\Sigma}} = \{s_{ij}\} = \mathbb{X}^\top \mathbb{X}/n$ be the $(p \times p)$ sample covariance matrix. Consider the singular value decomposition (SVD) of the centred data matrix,

$$\mathbb{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top = \sum_{i=1}^{r} \lambda_i \boldsymbol{u}_i \boldsymbol{v}_i^\top,$$

where $\boldsymbol{u}_i$ are the right singular vectors, $\boldsymbol{v}_i$ the left singular vectors, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r > 0$ the singular values, and r the rank of matrix $\mathbb{X}$. Similarly, the centred and scaled data matrix can be decomposed as

$$\mathbb{X}\boldsymbol{S}^{-1/2} = \boldsymbol{U}^*\boldsymbol{D}^*\boldsymbol{V}^{*\top} = \sum_{i=1}^{r} \lambda_i^* \boldsymbol{u}_i^* \boldsymbol{v}_i^{*\top},$$

where $\boldsymbol{S} = \mathrm{diag}(s_1^2, \ldots, s_p^2)$ is the diagonal matrix of sample variances. In both SVD decompositions the rank r is the rank of the data matrix, i.e. $r \leq \min(n, p)$, with equality when there are no singularities.

We now provide the details of the transformations investigated and some remarks.

4.1 Data sphering

Sphering (or whitening) the data (SPH) is obtained by applying the following transformation:

$$\boldsymbol{Z}_{\mathrm{SPH}} \leftarrow \mathbb{X}\boldsymbol{V}\boldsymbol{D}^{-1}\sqrt{n} = \boldsymbol{U}\sqrt{n},$$

where $\boldsymbol{V}$ and $\boldsymbol{D}^{-1}\sqrt{n} = \mathrm{diag}(\sqrt{n}/\lambda_i)$ are, respectively, the matrix of eigenvectors and the diagonal matrix of square root inverse of eigenvalues from the spectral decomposition of the sample marginal covariance, $\widehat{\boldsymbol{\Sigma}}$. For this transformation, $\mathrm{E}(\boldsymbol{Z}_{\mathrm{SPH}}) = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{Z}_{\mathrm{SPH}}) = \boldsymbol{I}$, so the features are centred at zero, with unit variances and uncorrelated. Thus, this transformation converts an elliptically shaped symmetric cloud of points into a spherically shaped cloud.

4.2 PCA scores from covariance matrix

The principal component transformation from the covariance matrix (PCS) is obtained as:

$$\boldsymbol{Z}_{\text{PCS}} \leftarrow \mathbb{X}\boldsymbol{V} = \boldsymbol{U}\boldsymbol{D},$$

for which $\text{E}(\boldsymbol{Z}_{\text{PCS}}) = \boldsymbol{0}$ and $\text{Var}(\boldsymbol{Z}_{\text{PCS}}) = \boldsymbol{D}^2/n = \text{diag}(\lambda_i^2/n)$, so the features are centred, uncorrelated and with decreasing variances equal to the eigenvalues of $\widehat{\boldsymbol{\Sigma}}$. Usually the first few components account for most of the dispersion in the data. A similar idea was proposed by McLachlan (1988), who discussed the use of principal component analysis in a preliminary exploratory analysis for the selection of suitable starting values.

4.3 PCA scores from correlation matrix

The principal component transformation from the correlation matrix (PCR) is defined as:

$$\boldsymbol{Z}_{\text{PCR}} \leftarrow \mathbb{X}\boldsymbol{S}^{-1/2}\boldsymbol{V}^* = \boldsymbol{U}^*\boldsymbol{D}^*,$$

for which $\text{E}(\boldsymbol{Z}_{\text{PCR}}) = \boldsymbol{0}$ and $\text{Var}(\boldsymbol{Z}_{\text{PCR}}) = \boldsymbol{D}^{*2}/n = \text{diag}(\lambda_i^{*2}/n)$, so the features are centred, uncorrelated and with decreasing variances equal to the eigenvalues of the marginal sample correlation matrix, $\boldsymbol{S}^{-1/2}\widehat{\boldsymbol{\Sigma}}\boldsymbol{S}^{-1/2}$.

4.4 Scaled SVD projection

The scaled SVD transformation (SVD) is computed as:

$$\boldsymbol{Z}_{\text{SVD}} \leftarrow \mathbb{X}\boldsymbol{S}^{-1/2}\boldsymbol{V}^*\boldsymbol{D}^{*-1/2} = \boldsymbol{U}^*\boldsymbol{D}^{*1/2},$$

for which $\text{E}(\boldsymbol{Z}_{\text{SVD}}) = \boldsymbol{0}$ and $\text{Var}(\boldsymbol{Z}_{\text{SVD}}) = \boldsymbol{D}^*/n = \text{diag}(\lambda_i^*/n)$. Again the features are centred, uncorrelated and with decreasing variances equal to the square root of the eigenvalues of the marginal sample correlation matrix, $\boldsymbol{S}^{-1/2}\widehat{\boldsymbol{\Sigma}}\boldsymbol{S}^{-1/2}$. In this case the features' dispersion presents a gentle decline compared to the PCR case.

4.5 Remarks

All the above transformations allow one to remove the "ordering" effect, so that even in the presence of ties the partitions obtained by applying MBHAC are invariant to permutations of the input variables. Furthermore, as shown in the next section through examples with real data, most of them allow one to achieve better clustering results when used for initialising the EM algorithm in Gaussian model-based clustering.

To get an idea of this, consider Fig. 1 where we report a scatterplot matrix with pairs of plots for the original variables in the Crabs dataset (to be discussed in Sect. 5.1) and

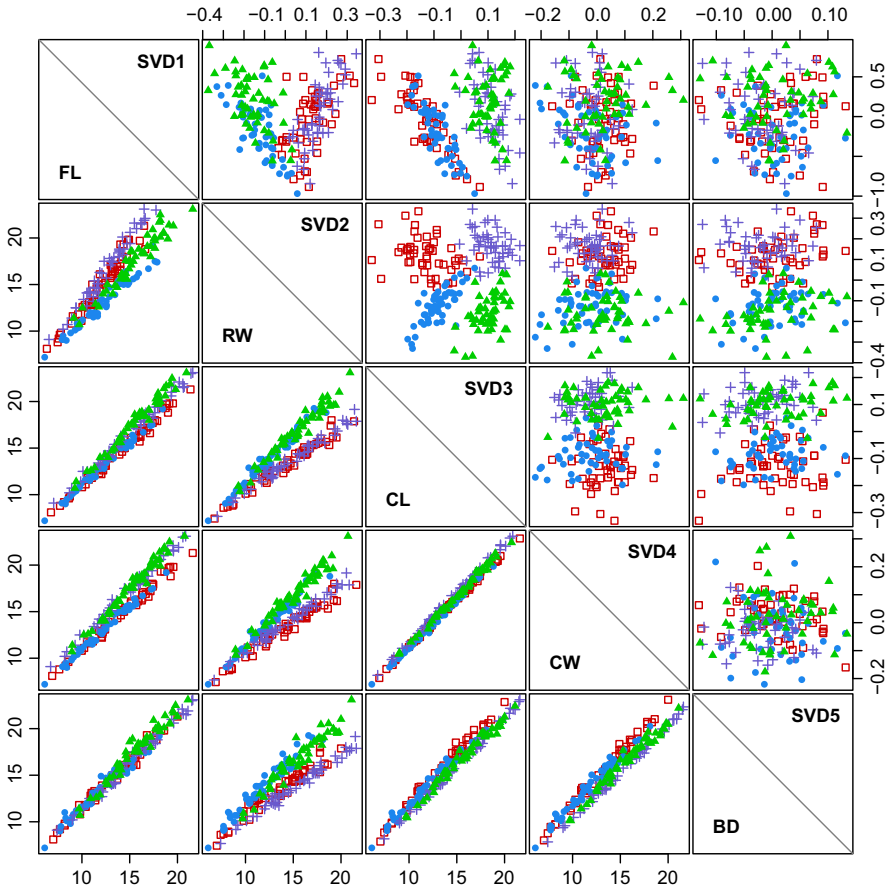

Fig. 1 Scatterplot matrix for the Crabs data: *lower panels* show scatterplots for pairs of variables in the original scale; *upper panels* show the features obtained by applying the scaled SVD transformation. For all graphs points are marked according to the true classes

for scaled SVD-transformed features in, respectively, the lower and upper panels, and with points marked according to the true classes. In the graphs for the variables in the original scale is hard to detect any clustering structure, whereas in the first three SVD features there appears to be a certain degree of separation among classes. Applying MBHAC on this features space should provide better starting points for applying the EM algorithm.

5 Data analyses

In this section we present some examples using real data. We compare the behaviour of the proposed transformations against the usual MBHAC for starting the EM algorithm, and two common strategies for GMMs initialisation. The first strategy is to start from the best partition obtained out of 50 runs of the k-means algorithm. The second is the *emEM* strategy as implemented in the `mixmod` software (Biernacki et al. 2006;

Auder et al. 2014), which, by default, uses 50 short runs of EM, each made of 5 iterations, followed by a long run of EM from the solution maximising the log-likelihood. The tolerance used for assessing the log-likelihood relative convergence of the EM algorithm is set to 10^{-5}.

The comparison among the different initialising strategies is based on two measures, the BIC and the adjusted Rand index (ARI). The former is used to select the best Gaussian finite mixture model with respect to both the number of components and the component-covariances decomposition. The BIC for model $\mathcal{M}$ with k components has the following form

$$\mathrm{BIC}_{\mathcal{M},k} = 2\ell(\widehat{\boldsymbol{\Psi}}; \boldsymbol{x}) - \nu_{\mathcal{M},k} \log(n),$$

where $\ell(\widehat{\boldsymbol{\Psi}}; \boldsymbol{x})$ is the maximised log-likelihood, $\nu_{\mathcal{M},k}$ is the number of independent parameters to be estimated in model $\mathcal{M}$, and k is the number of mixture components. This criterion depends on the starting partition through the log-likelihood at the MLEs $\widehat{\boldsymbol{\Psi}}$, penalised by the complexity of the model. It is the default criterion used in `mclust` for selecting a model, so the larger the value of the BIC the stronger the evidence for the corresponding model and number of components.

The ARI (Hubert and Arabie 1985) is used for evaluating the clustering obtained with a given mixture model. This is a measure of agreement between two partitions, one estimated by a statistical procedure independent of the labelling of the groups, and one being the true classification. The ARI has zero expected value in the case of a random partition, and it is bounded above by 1, with higher values representing better partition accuracy. Furthermore, it can be applied to compare partitions having different numbers of parts. The ARI is the index recommended by Milligan and Cooper (1986) for measuring the agreement between an external reference partition and a clustering partition. In the following data analysis examples, we take advantage of the knowledge of the true classes for measuring clustering accuracy, but not for model fitting.

5.1 Crabs data

We now consider a dataset consisting of data on five morphological measurements for 200 Leptograpsus crabs, with 50 crabs for each of two colour forms (blue and orange) and both sexes. Figure 2 shows the marginal distribution for each variable. In each stripchart data points are stacked to avoid overlapping, so the presence of several ties in the data is evident.

Overall, there are $5! = 120$ possible different ordering of the variables, of which 105 lead to selection of the (EEE,9) model, and 15 the (EEV,3) model. Table 1 shows the corresponding BIC and ARI obtained for these models and for models initialised using the transformations discussed in Sect. 4. As it can be seen, initialisation with the scaled SVD transformation returns the model with both the largest BIC and the most accurate partition (ARI $= 0.7938$). This transformation is the only one that selects the correct number of mixture components. Initialisation based on k-means also yields the right number of components, but its fit and accuracy are substantially worse. The *emEM* initialisation strategy yields a better fit, but it selects the wrong number of clusters.

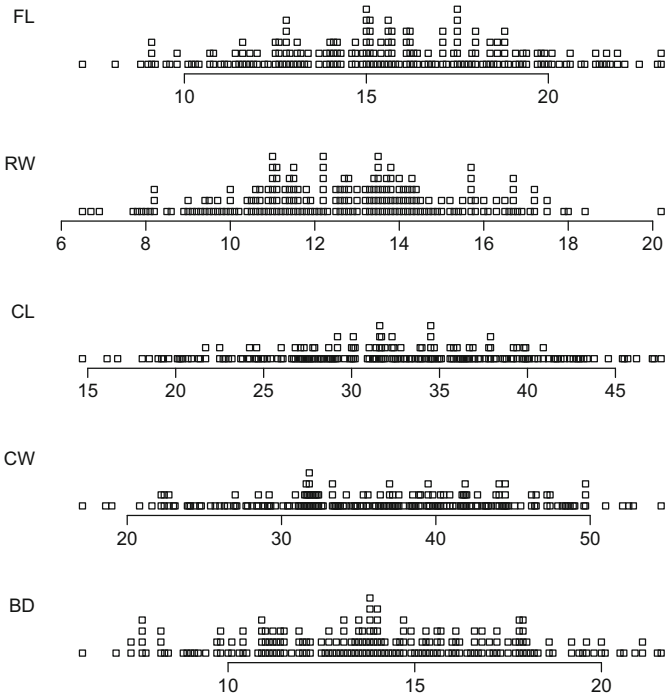

Fig. 2 Stripcharts with stacked data points for the Crabs data showing the marginal distribution for each variable. From this plot the presence of several ties is clearly visible

Table 1 Results from different initialisation strategies for the Crabs data

Initialisation	$\mathcal{M}$	k	BIC	ARI	System time (s)
MBHAC					
Default					
(105)	EEE	9	−2883.68	0.4831	2.61
(15)	EEV	3	−2925.59	0.4715	2.47
SPH	EEE	8	−2861.27	0.5248	2.40
PCS	EEE	9	−2883.68	0.4831	2.69
PCR	EEE	10	−2894.65	0.4572	2.72
SVD	EEV	4	**−2842.30**	**0.7938**	1.88
k-means	EEV	4	−2916.68	0.5926	2.78
emEM	EEE	6	−2866.61	0.6305	11.08

All possible orderings of the variables are reported for the default MBHAC initialisation; values among parentheses provide the number of such orderings which converge to the same solution. Larger values of both BIC and ARI are better. The best values are shown in bold font. System times are averaged over 10 runs of the estimation algorithm

Raftery and Dean (2006) selected as optimal subset for clustering purposes the variables `(FL, RW, CW, BD)` (in the identified ordering). Also for such a subset, different orderings give different results in 7 out of 24 possible arrangements. Table 2

Table 2 Results from different initialisation strategies for the Crabs data using the optimal subset of variables identified by Raftery and Dean (2006)

Initialisation	$\mathcal{M}$	k	BIC	ARI	System time (s)
MBHAC					
Default					
(17)	EEV	4	−2609.89	0.8154	2.54
(5)	EEV	5	−2645.19	0.7487	2.40
(1)	EEV	4	−2609.90	0.8276	2.40
(1)	EEE	11	−2674.40	0.4106	2.46
SPH	EEV	4	−2609.79	0.8154	1.89
PCS	EEV	4	−2609.89	0.8154	2.55
PCR	EEE	8	−2652.58	0.4860	2.52
SVD	EEV	4	**−2609.78**	**0.8400**	2.06
k-means	EEE	5	−2661.07	0.5608	2.62
emEM	EEV	4	−2609.87	0.8154	8.70

All possible orderings of the variables are reported for the default MBHAC initialisation; values among parentheses provide the number of such orderings which converge to the same solution. Larger values of both BIC and ARI are better. The best values are shown in bold font. System times are averaged over 10 runs of the estimation algorithm

shows the clustering results obtained with different initialisation strategies using the above mentioned optimal subset of variables. Again, initialisation with the scaled SVD transformation yields the largest BIC and the highest ARI among the considered strategies, with a slight improvement over the default initialisation with the identified ordering of the variables. Note that *emEM* has results analogous to the best default MBHAC initialisation, but k-means results are markedly worse.

Finally, note that, using both the full set of the variables and the optimal subset, analysis with the initialisation using the scaled SVD transformation used the smallest computing time. This may appear counterintuitive at first sight, because computing the SVD is time consuming. However, this is more than balanced by the fact that having good starting values allows the EM algorithm to converge in fewer steps.

5.2 Female voles data

Flury (1997) reported the data for a sample of 86 female voles from two species, 41 from Microtus californicus and 45 from Microtus ochrogaster, and seven variables describing various body measurements in units of 0.1mm, so several ties are contained in this data set.

There are $7! = 5040$ possible orderings of the variables. When the default initialisation is used, there is considerable variation among the resulting solutions: 41 final models with different BIC values (up to 5 significant digits) are estimated, leading to 38 different partitions. From Table 3, about 67 % of the solutions give a single compo-

Table 3 Results from different initialisation strategies for the female voles data

Initialisation	$\mathcal{M}$	K	BIC	ARI	System time (s)
MBHAC					
Default					
(3395)	XXX	1	−3874.9	0.0000	0.19
(996)	EEE	3	−3862.8	0.7720	0.29
(381)	EEE	2	**−3844.2**	**0.9081**	0.23
(253)	EEE	4	−3872.6	0.6015	0.24
SPH	EEE	5	−3871.0	0.4129	0.26
PCS	XXX	1	−3874.9	0.0000	0.22
PCR	EEE	2	**−3844.2**	**0.9081**	0.22
SVD	EEE	2	**−3844.2**	**0.9081**	0.19
k-means	EEE	4	−3872.6	0.6015	0.25
emEM	EEE	2	**−3844.2**	**0.9081**	2.43

The results for the default MBHAC initialisation are reported for some of all possible orderings of the variables; values among parentheses provide the number of such orderings which converge to the same pair $(\mathcal{M}, k)$. Larger values of both BIC and ARI are better. The best values are shown in bold font. System times are averaged over 10 runs of the estimation algorithm

nent model, about 20 % have three components, and only about 7.5 % of the solutions correctly identify the correct number of clusters. On the contrary, three transformation-based initialisation strategies are able to achieve the best fit in term of BIC, which in turn provides the largest ARI. The same optimal solution is also achieved by the *emEM* strategy, whereas results from the k-means initialisation are inferior. Note that also in this case the scaled SVD transformation is among the best strategies, with no increase in computing time.

5.3 Italian wines data

Forina et al. (1986) reported data on several chemical and physical properties of 178 wines grown in the same region in Italy but derived from three different cultivars (Barolo, Grignolino, Barbera). We consider 27 of the original 28 variables that were described in the Forina et al. paper and are available in the `pgmm` R package (McNicholas et al. 2015). In what follows, the data are analysed on the standardised scale and, given the large number of features, only the common full covariance matrix model (EEE) is examined. Table 4 shows the results obtained using different MBHAC initial partitions, k-means and *emEM* initialisation strategies. Except for one case, the MBHAC initialisations allow one to achieve good clustering accuracy. In particular, SVD has the highest BIC and attains a clustering that perfectly matches the real classes. By comparison, the models initialised by k-means and *emEM* show somewhat worse values of both BIC and ARI.

Table 4 Results from different initialisation strategies for the Italian wines data

Initialisation	$\mathcal{M}$	K	BIC	ARI	System time (s)
MBHAC					
Default	EEE	3	−12309.90	0.9489	0.36
SPH	EEE	2	−12370.80	0.5820	0.78
PCS	EEE	3	−12309.90	0.9489	0.36
PCR	EEE	3	−12309.90	0.9489	0.37
SVD	EEE	3	**−12306.75**	**1.0000**	0.35
k-means	EEE	3	−12314.72	0.9637	0.74
emEM	EEE	4	−12312.04	0.8355	3.99

Larger values of both BIC and ARI are better. The best values are shown in bold font. System times are averaged over 10 runs of the estimation algorithm

6 Final comments

The mixture model-based approach to clustering provides a firm statistical framework for unsupervised learning. The EM algorithm is typically used for parameter estimation. However, because of the many local maxima of the likelihood function, an important problem for getting sensible maximum likelihood estimates is to obtain reasonable starting points. A computationally efficient approach to EM initialisation is based on the partitions obtained from agglomerative hierarchical clustering.

In this paper we have presented and experimented with simple transformation-based methods to refine the EM initialisation step derived from model-based agglomerative hierarchical clustering. The proposed transformation-based strategies allow one to remove the dependence on the ordering of the variables when selecting starting partitions. They also often lead to improved model fitting and more accurate clustering results. Among the investigated transformations, the scaled SVD transformation performed the best in our experiments. The proposed approach may be applicable to other mixture modelling contexts, but we have not explored this possibility yet. Future studies will be devoted to this aspect.

As mentioned by Biernacki et al. (2003, p. 567) and Melnykov and Maitra (2010), we cannot expect an initialisation strategy to work uniformly well in all cases. Therefore, it is important to explore different strategies and to choose the solution with the highest log-likelihood value, but also to consider those sub-optimal solutions on a more subject specific considerations.

The transformation-based initialisation strategies discussed in this paper are available in the R package `mclust` (version >= 4.4), which can be downloaded from CRAN at http://cran.r-project.org/web/packages/mclust/index.html.

Acknowledgments The authors are grateful to the Coordinating Editor and two referees for their very helpful comments. This work was supported by NIH Grants R01-HD054511, R01-HD070936 and U54-HL127624, and by Science Foundation Ireland Walton Research Fellowship Number 11/W.1/I2079.

References

Auder B, Lebret R, Lovleff S, Langrognet F (2014) Rmixmod: an interface for MIXMOD. http://CRAN.R-project.org/package=Rmixmod, R package version 2.0.2

Banfield J, Raftery AE (1993) Model-based Gaussian and non-Gaussian clustering. Biometrics 49:803–821

Biernacki C, Celeux G, Govaert G (2000) Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Trans Pattern Anal Mach Intell 22(7):719–725

Biernacki C, Celeux G, Govaert G (2003) Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. Comput Stat Data Anal 41(3):561–575

Biernacki C, Celeux G, Govaert G, Langrognet F (2006) Model-based cluster and discriminant analysis with the MIXMOD software. Comput Stat Data Anal 51:587–600

Celeux G, Govaert G (1995) Gaussian parsimonious clustering models. Pattern Recognit 28:781–793

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). J R Stat Soc Series B Stat Methodol 39:1–38

Everitt B, Landau S, Leese M, Stahl D (2011) Cluster analysis, 5th edn. Wiley, Chichester, UK

Flury B (1997) A first course in multivariate statistics. Springer, New York

Forina M, Armanino C, Castino M, Ubigli M (1986) Multivariate data analysis as a discriminating method of the origin of wines. Vitis 25:189–201

Fraley C (1998) Algorithms for model-based Gaussian hierarchical clustering. SIAM J Sci Compu 20(1):270–281

Fraley C, Raftery AE (1998) How many clusters? Which clustering method? Answers via model-based cluster analysis. Comput J 41:578–588

Fraley C, Raftery AE (2002) Model-based clustering, discriminant analysis, and density estimation. J Am Stat Assoc 97(458):611–631

Fraley C, Raftery AE, Murphy TB, Scrucca L (2012) MCLUST version 4 for R: normal mixture modeling for model-based clustering, classification, and density estimation. Technical Report 597, Department of Statistics, University of Washington

Fraley C, Raftery AE, Scrucca L (2015) mclust: normal mixture modelling for model-based clustering, classification, and density estimation. http://CRAN.R-project.org/package=mclust, R package version 5.0.1

Gordon AD (1999) Classification, 2nd edn. Chapman & Hall/CRC

Hubert L, Arabie P (1985) Comparing partitions. J Classif 2:193–218

Jain AK, Dubes RC (1988) Algorithms for clustering data. Prentice-Hall, Inc

Kaufman L, Rousseeuw PJ (1990) Finding groups in data: an introduction to cluster analysis. Wiley, UK

Maitra R (2009) Initializing partition-optimization algorithms. IEEE/ACM Trans Comput Biol Bioinform 6(1):144–157

McLachlan G, Krishnan T (2008) The EM algorithm and extensions, 2nd edn. Wiley-Interscience, Hoboken, New Jersey

McLachlan G, Peel D (2000) Finite mixture models. Wiley, New York

McLachlan GJ (1988) On the choice of starting values for the EM algorithm in fitting mixture models. Statistician 37(4/5):417

McNicholas PD, ElSherbiny A, McDaid AF, Murphy TB (2015) pgmm: Parsimonious Gaussian Mixture Models. http://CRAN.R-project.org/package=pgmm, R package version 1.2

Melnykov V, Maitra R (2010) Finite mixture models and model-based clustering. Stat Surv 4:80–116

Melnykov V, Melnykov I (2012) Initializing the EM algorithm in Gaussian mixture models with an unknown number of components. Comput Stat Data Anal 56(6):1381–1395

Milligan GW, Cooper MC (1986) A study of the comparability of external criteria for hierarchical cluster analysis. Multivar Behav Res 21(4):441–458

Raftery AE, Dean N (2006) Variable selection for model-based clustering. J Am Stat Assoc 101(473):168–178

Schwartz G (1978) Estimating the dimension of a model. Ann Stat 6:31–38

Wu CJ (1983) On the convergence properties of the EM algorithm. Ann Stat 11(1):95–103