View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by INRIA a CCSD electronic archive server

Combining Mixture Components for Clustering

Jean-Patrick Baudry, Adrian E. Raftery, Gilles Celeux, Kenneth Lo, Raphael Gottardo

To cite this version:

Jean-Patrick Baudry, Adrian E. Raftery, Gilles Celeux, Kenneth Lo, Raphael Gottardo. Combining Mixture Components for Clustering. Journal of Computational and Graphical Statistics, Taylor & Francis, 2010, 19, pp.332-353. inria-00321090

HAL Id: inria-00321090

https://hal.inria.fr/inria-00321090

Submitted on 12 Sep 2008

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Combining Mixture Components for Clustering

Jean-Patrick Baudry — Adrian E. Raftery — Gilles Celeux — Kenneth Lo — Raphael Gottardo

N° 6644

Septembre 2008

Thème COG

Rapport de recherche

ISSN 0249-6399 ISRN INRIA/RR--6644--FR+ENG

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
SACLAY – ÎLE-DE-FRANCE

Combining Mixture Components for Clustering

Jean-Patrick Baudry [*], Adrian E. Raftery [†], Gilles Celeux [‡] , Kenneth Lo [§] , Raphael Gottardo [§]

Thème COG — Systèmes cognitifs
Équipes-Projets SELECT

Rapport de recherche n° 6644 — Septembre 2008 — 18 pages

Abstract: Model-based clustering consists of fitting a mixture model to data and identifying each cluster with one of its components. Multivariate normal distributions are typically used. The number of clusters is usually determined from the data, often using BIC. In practice, however, individual clusters can be poorly fitted by Gaussian distributions, and in that case model-based clustering tends to represent one non-Gaussian cluster by a mixture of two or more Gaussian distributions. If the number of mixture components is interpreted as the number of clusters, this can lead to overestimation of the number of clusters. This is because BIC selects the number of mixture components needed to provide a good approximation to the density, rather than the number of clusters as such. We propose first selecting the total number of Gaussian mixture components, K, using BIC and then combining them hierarchically according to an entropy criterion. This yields a unique soft clustering for each number of clusters less than or equal to K; these clusterings can be compared on substantive grounds. We illustrate the method with simulated data and a flow cytometry dataset.

Key-words: BIC, entropy, flow cytometry, mixture model, model-based clustering, multivariate normal distribution.

[*] Université Paris-Sud
[†] University of Washington
[‡] INRIA
[§] University of British Columbia

Centre de recherche INRIA Saclay – Île-de-France
Parc Orsay Université
4, rue Jacques Monod, 91893 ORSAY Cedex
Téléphone : +33 1 72 92 59 00

Agréger des composantes d'un mélange pour la classification non supervisée

Résumé : La classification non supervisée par les modèles de mélange consiste à ajuster un modèle de mélange aux données, puis à identifier une classe à chacune des composantes obtenues. Le nombre de classes est habituellement choisi d'après les données, typiquement par le critère BIC. Cependant, en pratique, certains groupes peuvent être mal approchés par une distribution gaussienne et cette démarche mène souvent à représenter un tel groupe par un mélange de plusieurs composantes gaussiennes. Lorsque le nombre de composantes est effectivement interprété comme le nombre de classes, ce dernier risque d'être surestimé. La cause en est que BIC estime le nombre de composantes gaussiennes nécessaire pour que le mélange obtenu approche correctement la densité des données, et pas le nombre de classes en tant que tel. Nous proposons de sélectionner dans un premier temps le nombre total de composantes gaussiennes du mélange à ajuster, K, par BIC, puis de combiner hiérarchiquement ces composantes d'après un critère d'entropie. Nous obtenons ainsi une classification pour chaque nombre de classes inférieur ou égal à K; la nature des données et les connaissances a priori les concernant peuvent alors permettre de comparer ces différentes classifications. Nous illustrons cette méthode sur des données simulées et par une application à des données de cytologie.

Mots-clés : BIC, entropie, cytologie, modèles de mélange pour la classification non supervisée, distribution normale multivariée.

1 INTRODUCTION

Model-based clustering is based on a finite mixture of distributions, in which each mixture component is taken to correspond to a different group, cluster or subpopulation. For continuous data, the most common component distribution is a multivariate Gaussian (or normal) distribution. A standard methodology for model-based clustering consists of using the EM algorithm to estimate the finite mixture models corresponding to each number of clusters considered and using BIC to select the number of mixture components, taken to be equal to the number of clusters (Fraley and Raftery 1998). The clustering is then done by assigning each observation to the cluster to which it is most likely to belong *a posteriori*, conditionally on the selected model and its estimated parameters. For reviews of model-based clustering, see McLachlan and Peel (2000) and Fraley and Raftery (2002).

Biernacki, Celeux, and Govaert (2000) argued that the goal of clustering is not the same as that of estimating the best approximating mixture model, and so BIC may not be the best way of determining the number of clusters, even though it does perform well in selecting the number of components in a mixture model. Instead they proposed the ICL criterion, whose purpose is to assess the number of mixture components that leads to the best clustering. This turns out to be equivalent to BIC penalized by the entropy of the corresponding clustering.

We argue here that the goal of selecting the number of mixture components for estimating the underlying probability density is well met by BIC, but that the goal of selecting the number of *clusters* may not be. Even when a multivariate Gaussian mixture model is used for clustering, the number of mixture components is not necessarily the same as the number of clusters. This is because a cluster may be better represented by a mixture of normals than by a single normal distribution.

We propose a method for combining the points of view underlying BIC and ICL to achieve the best of both worlds. BIC is used to select the number of components in the mixture model. We then propose a sequence of possible solutions by hierarchical combination of the components identified by BIC. The decision about which components to combine is based on the same entropy criterion that ICL implicitly uses. In this way, we propose a way of interpreting the mixture model in clustering terms by identifying a set of mixture components with each cluster. Finally, ICL could be helpful for identifying the number of clusters among the solutions provided by the designed hierarchy. This number of clusters can be different from the number of components chosen with BIC.

Often the number of clusters identified by ICL is smaller than the number of components selected by BIC, raising the question of whether BIC tends to overestimate the number of groups. On the other hand, in almost all simulations based on assumed true mixture models, the number of components selected by BIC does not overestimate the true number of components (Biernacki et al. 2000; McLachlan and Peel 2000; Steele 2002). Our approach resolves this apparent paradox.

In Section 2 we provide background on model-based clustering, BIC and ICL, and in Section 3 we describe our proposed methodology. In Section 4 we give results for simulated data, and in Section 5 we give results from the analysis of a flow cytometry dataset. There, one of the sequence of solutions from our method is clearly indicated substantively, and seems better than either the original BIC or ICL solutions. In Section 6 we discuss issues relevant to our method and other methods that have been proposed.

2 MODEL SELECTION IN MODEL-BASED CLUSTERING

Model-based clustering assumes that observations $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ in $\mathbf{R}^{nd}$ are a sample from a finite mixture density

$$p(\mathbf{x}_i \mid K, \theta_K) = \sum_{k=1}^{K} p_k \phi(\mathbf{x}_i \mid \mathbf{a}_k), \tag{1}$$

where the p_k's are the mixing proportions ($0 < p_k < 1$ for all $k = 1, \ldots, K$ and $\sum_k p_k = 1$), $\phi(. \mid \mathbf{a}_k)$ denotes a parameterized density, and $\theta_K = (p_1, \ldots, p_{K-1}, \mathbf{a}_1, \ldots, \mathbf{a}_K)$. When the data are multivariate continuous observations, the component density is usually the d-dimensional Gaussian density with parameter $\mathbf{a}_k = (\mu_k, \Sigma_k)$, μ_k being the mean and Σ_k the variance matrix of component k.

For estimation purposes, the mixture model is often expressed in terms of complete data, including the groups to which the data points belong. The complete data are

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

where the missing data are $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$, with $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})$ being binary vectors such that $z_{ik} = 1$ if $\mathbf{x}_i$ arises from group k. The $\mathbf{z}_i$'s define a partition $P = (P_1, \ldots, P_K)$ of the observed data $\mathbf{x}$ with $P_k = \{\mathbf{x}_i \text{ such that } z_{ik} = 1\}$.

From a Bayesian perspective, the selection of a mixture model can be based on the integrated likelihood of the mixture model with K components (Kass and Raftery 1995), namely

$$p(\mathbf{x}|K) = \int p(\mathbf{x}|K, \theta_K)\pi(\theta_K)d\theta_K, \tag{2}$$

where $\pi(\theta_K)$ is the prior distribution of the parameter θ_K. Here we use the BIC approximation of Schwarz (1978) to the log integrated likelihood, namely

$$\mathrm{BIC}(K) = \log p(\mathbf{x}|K, \hat{\theta}_K) - \frac{\nu_K}{2}\log(n), \tag{3}$$

where $\hat{\theta}_K$ is the maximum likelihood estimate of θ_K and ν_K is the number of free parameters of the model with K components. (Keribin 2000) has shown that under certain regularity conditions the BIC consistently estimates the number of mixture components, and numerical experiments show that the BIC works well at a practical level (Fraley and Raftery 1998; Biernacki et al. 2000; Steele 2002).

There is one problem with using this solution directly for clustering. Doing so is reasonable if each mixture component corresponds to a separate cluster, but this may not be the case. In particular, a cluster may be both cohesive and well separated from the other data (the usual intuitive notion of a cluster), without its distribution being Gaussian. This cluster may be represented by two or more mixture components, if its distribution is better approximated by a mixture than by a single Gaussian component. Thus the number of clusters in the data may be different from the number of components in the best approximating Gaussian mixture model.

To overcome this problem, Biernacki et al. (2000) proposed estimating the number of *clusters* (as distinct from the number of mixture components) in model-based clustering using the integrated complete likelihood (ICL), defined as the integrated likelihood of the complete data $(\mathbf{x}, \mathbf{z})$. ICL is defined as

$$p(\mathbf{x}, \mathbf{z} \mid K) = \int_{\Theta_K} p(\mathbf{x}, \mathbf{z} \mid K, \theta)\pi(\theta \mid K)d\theta, \tag{4}$$

where

$$p(\mathbf{x}, \mathbf{z} \mid K, \theta) = \prod_{i=1}^{n} p(\mathbf{x}_i, \mathbf{z}_i \mid K, \theta)$$

with

$$p(\mathbf{x}_i, \mathbf{z}_i \mid K, \theta) = \prod_{k=1}^{K} p_k^{z_{ik}} \left[\phi(\mathbf{x}_i \mid \mathbf{a}_k)\right]^{z_{ik}} .$$

To approximate this integrated complete likelihood, Biernacki et al. (2000) proposed using a BIC-like approximation, leading to the criterion

$$\mathrm{ICL}(K) = \log \mathbf{p}(\mathbf{x}, \hat{\mathbf{z}} \mid K, \hat{\theta}_K) - \frac{\nu_K}{2} \log n, \tag{5}$$

where the missing data have been replaced by their most probable values, given the parameter estimate $\hat{\theta}_K$.

Roughly speaking, ICL is equal to BIC penalized by the mean entropy

$$\mathrm{Ent}(K) = -\sum_{k=1}^{K} \sum_{i=1}^{n} t_{ik}(\hat{\theta}_K) \log t_{ik}(\hat{\theta}_K) \geq 0, \tag{6}$$

where t_{ik} denotes the conditional probability that $\mathbf{x}_i$ arises from the kth mixture component ($1 \leq i \leq n$ and $1 \leq k \leq K$), namely

$$t_{ik}(\hat{\theta}_K) = \frac{\hat{p}_k \phi(\mathbf{x}_i | \hat{\mathbf{a}}_k)}{\sum_{j=1}^{K} \hat{p}_j \phi(\mathbf{x}_i | \hat{\mathbf{a}}_j)}.$$

Thus the number of clusters, K', favored by ICL tends to be smaller than the number K favored by BIC because of the additional entropy term. ICL aims to find the number of clusters rather than the number of mixture components. However, if it is used to estimate the number of mixture components it can underestimate it, particularly in data arising from mixtures with poorly separated components. In that case, the fit is worsened.

Thus the user of model-based clustering faces a dilemma: do the mixture components really all represent clusters, or do some subsets of them represent clusters with non-Gaussian distributions? In the next section, we propose a methodology to help resolve this dilemma.

3 METHODOLOGY

The idea is to build a sequence of clusterings, starting from a mixture model that fits the data well. Its number of components is chosen using BIC. We design a sequence of candidate soft clusterings with $\hat{K}^{\mathrm{BIC}}$, $\hat{K}^{\mathrm{BIC}} - 1, \ldots, 1$ clusters by successively merging the components in the BIC solution.

At each stage, we choose the two mixture components to be merged so as to minimize the entropy of the resulting clustering. Let us denote by $t_{i,1}^K, \ldots, t_{i,K}^K$ the conditional posterior probabilities that $\mathbf{x}_i$ arises from cluster $1, \ldots, K$ with respect to the K-cluster solution. If clusters k and k' from the K-cluster solution are combined, the $t_{i,j}$'s remain the same for every j except for k and k'. The new cluster $k \cup k'$ then has the following conditional probability:

$$t_{i,k \cup k'}^K = t_{i,k}^K + t_{i,k'}^K.$$

Then the resulting entropy is:

$$-\sum_{i=1}^{n}\left(\sum_{j\neq k,k'} t_{ij}^{K}\log t_{ij}^{K} + (t_{ik}^{K}+t_{ik'}^{K})\log\left(t_{ik}^{K}+t_{ik'}^{K}\right)\right). \tag{7}$$

Thus, the two clusters k and k' to be combined are those maximizing the criterion:

$$-\sum_{i=1}^{n}\{t_{ik}^{K}\log(t_{ik}^{K}) + t_{ik'}^{K}\log(t_{ik'}^{K})\} + \sum_{i=1}^{n} t_{i\,k\cup k'}^{K}\log t_{i\,k\cup k'}^{K}$$

among all possible pairs of clusters (k,k'). Then $t_{i,k}^{K-1},\ i=1,\dots,n,\ k=1,\dots,K-1$ can be updated.

Any combined solution fits the data as well as the BIC solution, since it is based on the same Gaussian mixture; the likelihood does not change. Only the number and definition of clusters are different. Our method yields just one suggested set of clusters for each K, and the user can choose between them on substantive grounds. our flow cytometry data example in Section 5 provides one instance of this.

If a more automated procedure is desired for choosing a single solution, one possibility is to select, among the possible solutions, the solution providing the same number of clusters that ICL. An alternative is to use an elbow rule on the graphic displaying the entropy variation against the number of clusters. Both these strategies are illustrated in our examples.

The algorithm implementing the suggested procedure is given in the Appendix.

4 SIMULATED EXAMPLES

We first present some simulations to highlight the posssibilities of our methodology. They have been chosen to illustrate cases where BIC and ICL do not select the same number of components.

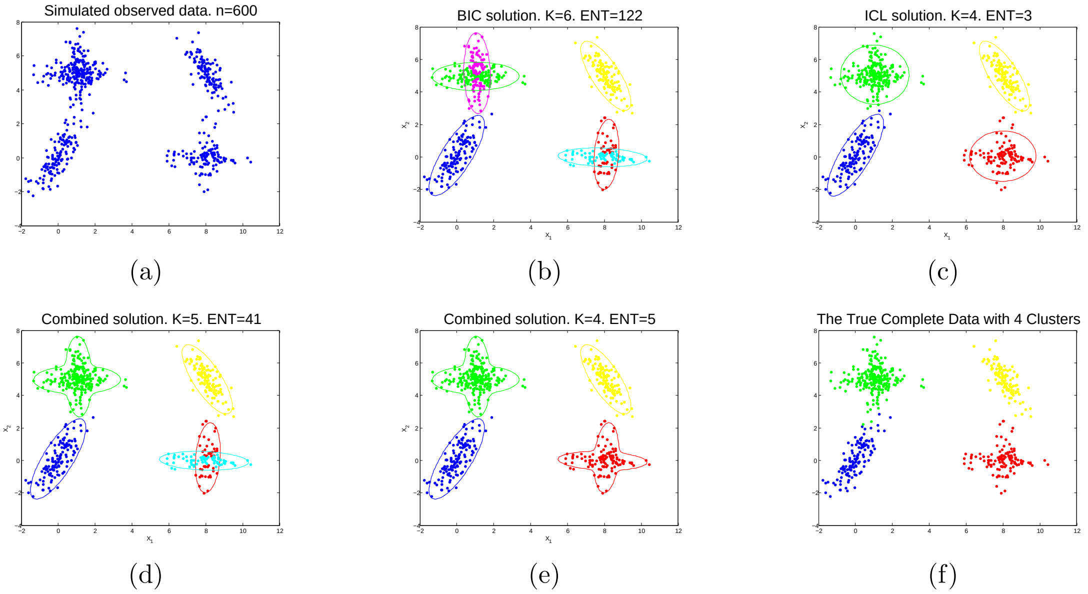
4.1 Simulated Example with Overlapping Components

The data, shown in Figure 1(a), were simulated from a two-dimensional Gaussian mixture. There are six components, four of which are axis-aligned with diagonal variance matrices (the four components of the two “crosses”), and two of which are not axis-aligned, and so do not have diagonal variance matrices. There were 600 points, with mixing proportions 1/5 for each non axis-aligned component, 1/5 for each of the upper left cross components, and 1/10 for each of the lower right cross components.

We fitted Gaussian mixture models to this simulated dataset. BIC selected a 6-component mixture model, which was the correct model; this is shown in Figure 1(b). ICL selected a 4-cluster model, as shown in Figure 1(c). The four clusters found by ICL are well separated.

Starting from the BIC 6-component solution, we combined two components to get the 5-cluster solution shown in Figure 1(d). To decide which two components to merge, each pair of components was considered, and the entropy after combining these components into one cluster was computed. The two components for which the resulting entropy was the smallest were combined.

The same thing was done again to find a 4-cluster solution, shown in Figure 1(e). This is the number of clusters identified by ICL. The entropies of the combined solutions are shown in Figure 2, together with the differences between successive entropy values. There seems to be an elbow in the plot at $K=4$, and so we focus on this solution. Note that there is no conventional formal



Simulated observed data. n=600

BIC solution. K=6. ENT=122

ICL solution. K=4. ENT=3

Combined solution. K=5. ENT=41

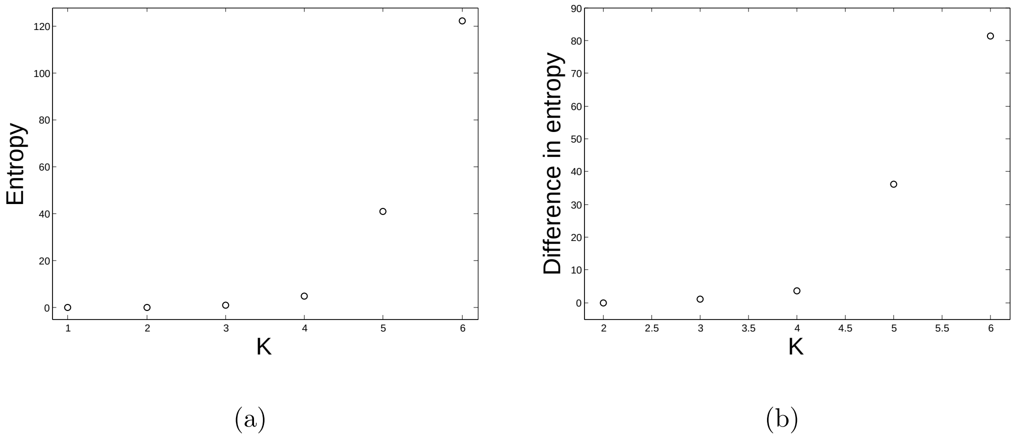
Combined solution. K=4. ENT=5

The True Complete Data with 4 Clusters

(a) (b) (c)

(d) (e) (f)

Figure 1: Simulated Example 1. (a) Simulated data from a 6-component 2-dimensional Gaussian mixture. (b) BIC solution with 6 components. (c) ICL solution with 4 clusters. (d) Combined solution with 5 clusters. (e) Combined solution with 4 clusters. (f) The true labels for a 4-cluster solution. In (b) and (c) the entropy, ENT, is defined by equation (6) with respect to the the maximum likelihood solution, and in (d) and (e) ENT is defined by equation (7).



(a) (b)

Figure 2: (a) Entropy values for the K-cluster Combined Solution, as defined by equation (7), for Simulated Example 1. (b) Differences between successive entropy values.

statistical inferential basis for choosing between different numbers of clusters, as the likelihood and the distribution of the observations are the same for all the numbers of clusters considered.

In the four-cluster solution, the clusters are no longer all Gaussian; now two of them are modeled as mixtures of two Gaussians each. Note that this four-cluster solution is not the same as the four-cluster solution identified by ICL: ICL identifies a mixture of four Gaussians, while our method identifies four clusters of which two are not Gaussian. Figure 1(f) shows the true classification. Only 3 of the 600 points were misclassified.

4.2 Simulated Example with Overlapping Components and Restrictive Models

We now consider the same data again, but this time with more restrictive models. Only Gaussian mixture models with diagonal variance matrices are considered. This illustrates what happens when the mixture model generating the data is not in the set of models considered.

BIC selects more components than before, namely ten (Figure 3a). This is because the generating model is not considered, and so more components are needed to approximate the true distribution. For example, the top right non-axis-aligned component cannot be represented correctly by a single Gaussian with a diagonal variance matrix, and BIC selects three diagonal Gaussians to represent it. ICL still selects four clusters (Figure 3b).

In the hierarchical merging process, the two components of one of the “crosses” were combined first (Figure 3c), followed by the components of the other cross (Figure 3d). The nondiagonal cluster on the lower left was optimally represented by three diagonal mixture components in the BIC solution. In the next step, two of these three components were combined (Figure 3e). Next, two of the three mixture components representing the upper right cluster were combined (Figure 3f). After the next step there were five clusters, and all three mixture components representing the lower left cluster had been combined (Figure 3g).

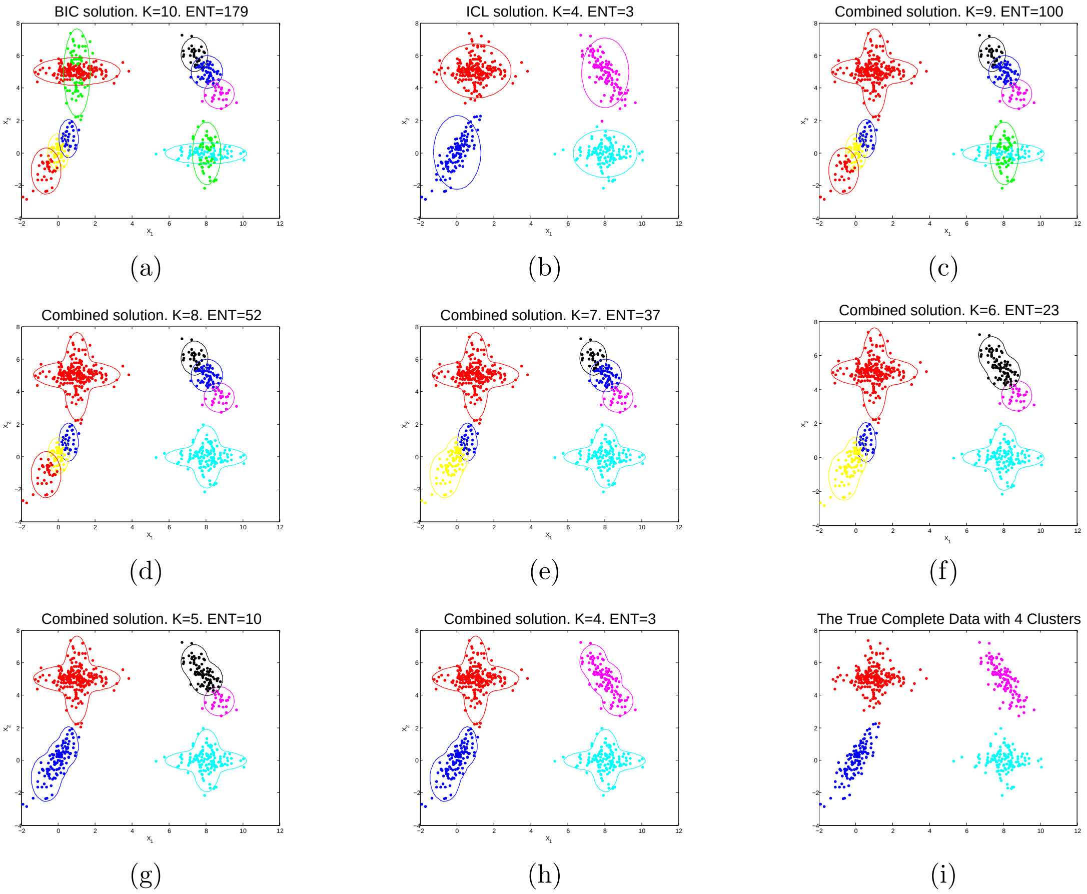
The next step got us to four clusters, the number identified by ICL (Figure 3h). After this last combination, all three mixture components representing the upper right cluster had been combined. Note that this four-cluster solution is not the same as the four-cluster solution got by optimizing ICL directly. Strikingly, this solution is almost identical to that obtained with the less restrictive set of models considered in Section 4.1.

The plot of the combined solution entropies against the number of components in Figure 4 suggests an elbow at $K = 8$, with a possible second, less apparent one at $K = 4$. In the $K = 8$ solution the two crosses have been merged, and in the $K = 4$ solution all four visually apparent clusters have been merged. Recall that the choice of the number of clusters is not based on formal statistical inference, unlike the choice of the number of mixture components. Our method generates a small set of possible solutions that can be compared on substantive grounds. The entropy plot is an exploratory device that can help to assess separation between clusters, rather than a formal inference tool.

4.3 Circle/Square Example

This example was presented by Biernacki et al. (2000). The data, shown in Figure 5(a), were simulated from a mixture of a uniform distribution on a square and a spherical Gaussian distribution. Here, for illustrative purposes, we restricted the models considered to Gaussian mixtures with spherical variance matrices with the same determinant. Note that the true generating model does not belong to this model class.

In the simulation results of Biernacki et al. (2000), BIC chose two components in only 60% of the simulated cases. Here we show one simulated dataset in which BIC approximated



BIC solution. K=10. ENT=179

(a)

ICL solution. K=4. ENT=3

(b)

Combined solution. K=9. ENT=100

(c)

Combined solution. K=8. ENT=52

(d)

Combined solution. K=7. ENT=37

(e)

Combined solution. K=6. ENT=23

(f)

Combined solution. K=5. ENT=10

(g)

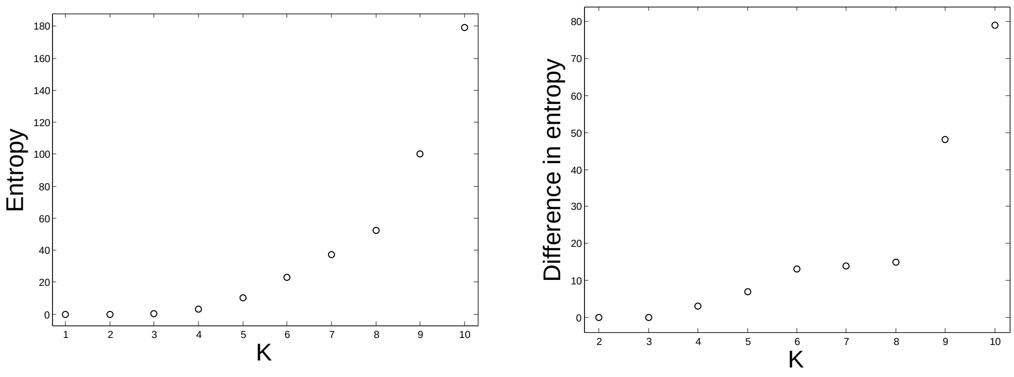
Combined solution. K=4. ENT=3

(h)

The True Complete Data with 4 Clusters

(i)

Figure 3: Simulated Example 2. The data are the same as in Simulated Example 1, but the model space is more restrictive, as only Gaussian mixture models with diagonal covariance matrices are considered. See Fig.1 legends for explanations about ENT. (a) BIC solution with 10 mixture components. (b) ICL solution with 4 clusters. (c) Combined solution with 9 clusters. (d) Combined solution with 8 clusters. (e) Combined solution with 7 clusters. (f) Combined solution with 6 clusters. (g) Combined solution with 5 clusters. (h) Combined solution with 4 clusters. (i) True labels with 4 clusters.



(a)

(b)

Figure 4: (a) Entropy values for the K-cluster Combined Solution, as defined by equation (7), for Simulated Example 2. (b) Differences between successive entropy values.

the underlying non-Gaussian density using a mixture of five normals (Figure 5b). ICL always selected two clusters (Figure 5c).

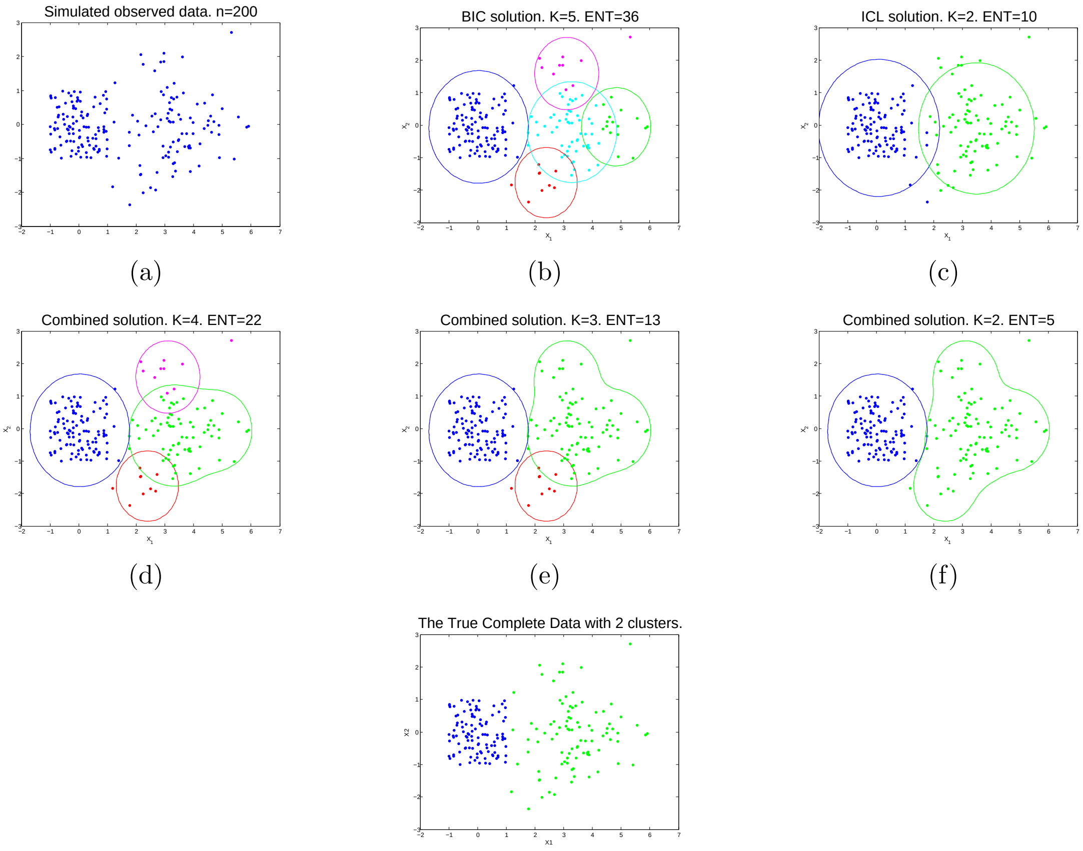
The progress of the combining algorithm is shown in Figure 5(d-f). The final two-cluster solution, obtained by hierarchical merging starting from the BIC solution, is slightly different from the clustering obtained by optimizing ICL directly. It also seems slightly better: ICL classifies seven observations into the uniform cluster that clearly do not belong to it, while the solution shown misclassifies only three observations in the same way. The true labels are shown in Figure 5(g). The entropy plot in Figure 6 does not have a clear elbow.

5 FLOW CYTOMETRY EXAMPLE

We now apply our method to the GvHD data of Brinkman et al. (2007). Two samples of this flow cytometry data have been used, one from a patient with the graft-versus-host disease (GvHD), and the other from a control patient. GvHD occurs in allogeneic hematopoietic stem cell transplant recipients when donor-immune cells in the graft attack the skin, gut, liver, and other tissues of the recipient. GvHD is one of the most significant clinical problems in the field of allogeneic blood and marrow transplantation.

The GvHD positive and control samples consist of 9,083 and 6,809 observations respectively. Both samples include four biomarker variables, namely, CD4, CD8β, CD3 and CD8. The objective of the analysis is to identify CD3$^+$ CD4$^+$ CD8β^+ cell sub-populations present in the GvHD positive sample. In order to identify all cell sub-populations in the data, we use a Gaussian mixture model with unrestricted covariance matrix. Adopting a similar strategy to that described by Lo, Brinkman, and Gottardo (2008), for a given number of components, we locate the CD3$^+$ sub-populations by labeling components with means in the CD3 dimension above 270 CD3$^+$. This threshold was based on a comparison with a negative control sample as explained by Brinkman et al. (2007).

We analyze the positive sample first. A previous manual analysis of the positive sample suggested that the CD3$^+$ cells could be divided into approximately five to six CD3$^+$ cell sub-populations (Brinkman et al. 2007). ICL selected nine clusters, five of which correspond to the

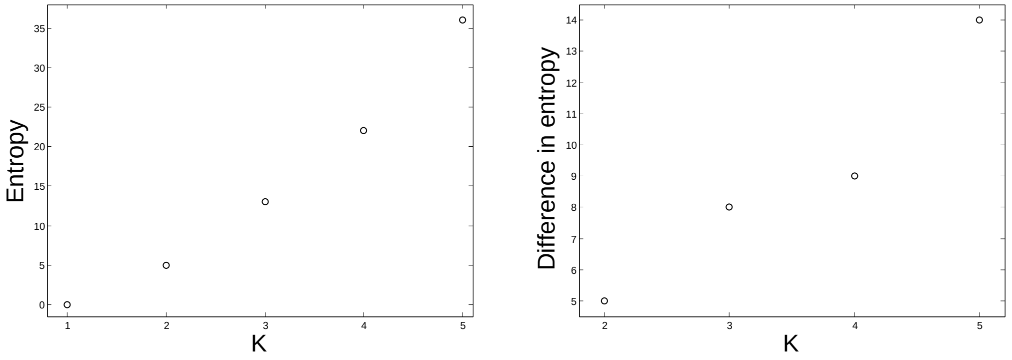


(a) (b) (c)

(d) (e) (f)

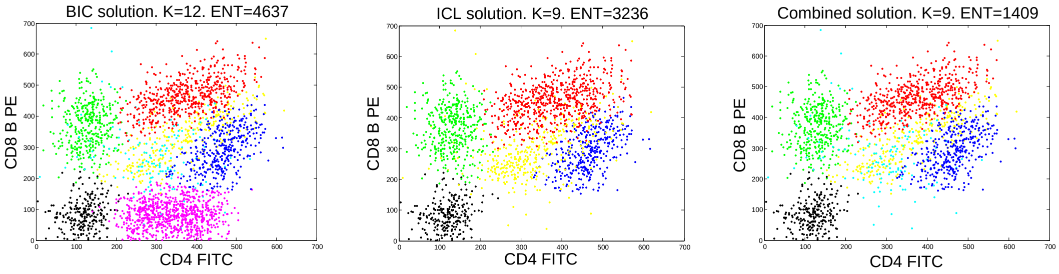
(g)

Figure 5: Circle-Square Example. See Fig.1 legends for explanations about ENT. (a) Observed data simulated from a mixture of a uniform distribution on a square and a spherical Gaussian distribution. (b) The BIC solution, with 5 components. (c) The ICL solution with 2 clusters. (d) The combined solution with 4 clusters. (e) The combined solution with 3 clusters. (f) The final combined solution, with 2 clusters. (g) The true labels.



(a) (b)

Figure 6: (a) Entropy values for the K-cluster Combined Solution, as defined by equation (7), for the Circle-Square Example. (b) Differences between successive entropy values.



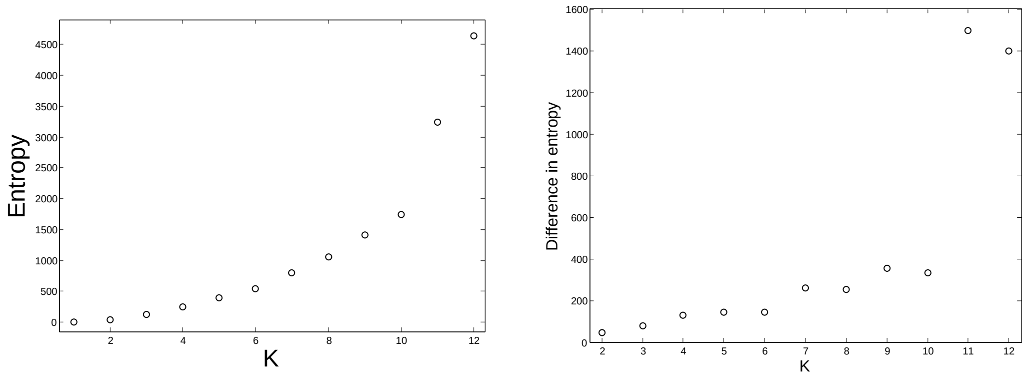
(a) (b) (c)

Figure 7: GvHD positive sample. Only components labeled CD3$^+$ are shown. (a) BIC Solution ($K = 12$). The combined solutions for $K = 11$ and $K = 10$ are almost identical for these CD3$^+$ components. (b) ICL Solution (K=9). (c) Combined Solution (K=9).

CD3$^+$ population (Figure 7(b)). Compared with the result shown in Lo et al. (2008), the CD4$^+$ CD8β^- region located at the bottom right of the graph is not represented.

BIC selected 12 components to provide a good fit to the positive sample, seven of which are labeled CD3$^+$ (Figure 7(a)). The CD4$^+$ CD8β^+ region seems to be encapsulated by the cyan, blue, yellow and red components. Starting from this BIC solution, we repeatedly combined two components causing maximal reduction in the entropy. The first two combinations all occurred within those components originally labeled CD3$^-$, and the CD4 vs CD8β projection of the CD3$^+$ sub-populations remains unchanged.

However, when the number of clusters was reduced to nine, the purple cluster representing the CD3$^+$ CD4$^+$ CD8β^- population was combined with the big CD3$^-$ cluster, resulting in an incomplete representation of the CD3$^+$ population (Figure 7 (c)). Hence, the combined solution with ten clusters, in which seven are labeled CD3$^+$, seems to provide the most parsimonious view of the positive sample whilst retaining the seven important CD3$^+$ cell sub-populations. Note that the entropy of the combined solution with nine clusters (1409) was smaller than that of the



(a) (b)

Figure 8: (a) Entropy values for the GvHD positive sample. (b) Differences between successive entropy values.

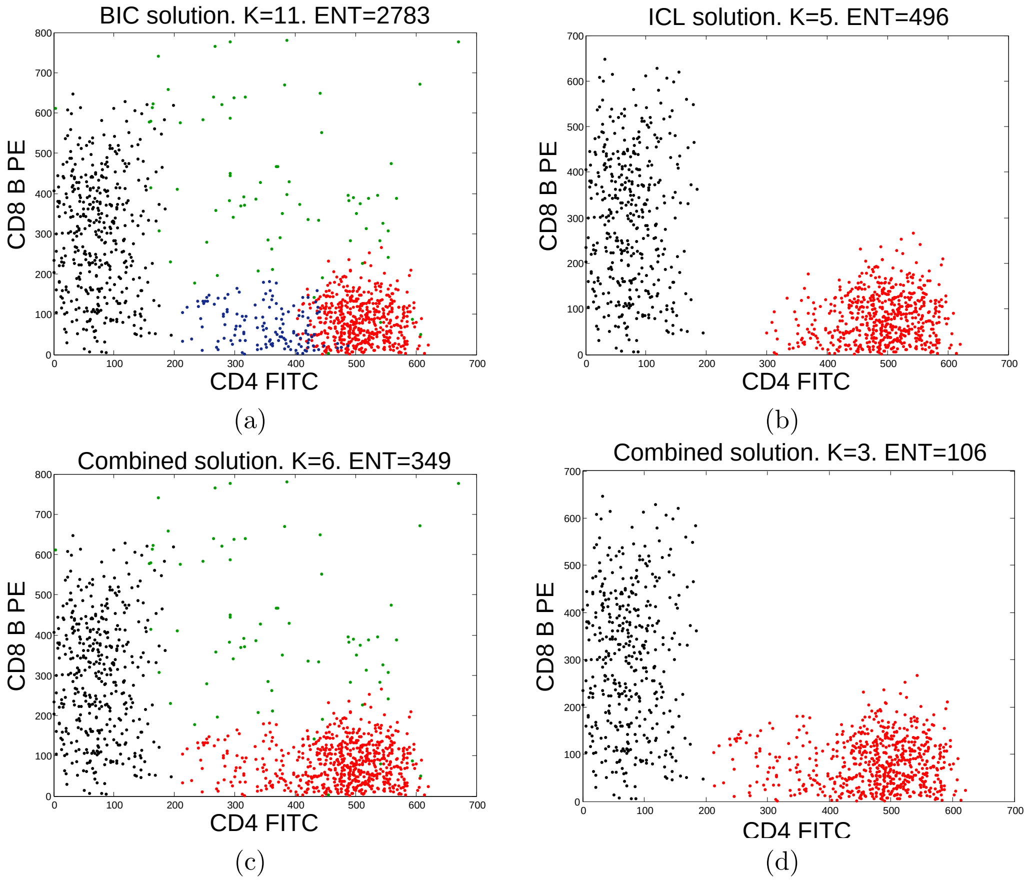
ICL solution (3236). The entropy plot (Figure 8) suggests an elbow at $K = 10$ clusters, agreeing with our more substantively-based conclusion.

Next we analyze the control sample. A satisfactory analysis would show an absence of the CD3$^+$ CD4$^+$ CD8β^+ cell sub-populations. ICL chose five clusters, two of which correspond to the CD3$^+$ population (Figure 9 (b)). The black cluster at the left of the graph represents the CD4$^-$ region. The red cluster at the rigth of the graph represents the CD8$^-$ region. It seems that it misses a part of this cluster near the black cluster. This suggests that the ICL solution could be improved.

BIC selected 11 components, four of which are labeled CD3$^+$ (Figure 9(a)). The black component on the left side does not extend to the CD4$^+$ region. However, contrary to previous findings in which CD4$^+$ CD8β^+ cell sub-populations were found only in positive samples and not in control samples, a green component is used to represent the observations scattered within the CD4$^+$ CD8β^+ region.

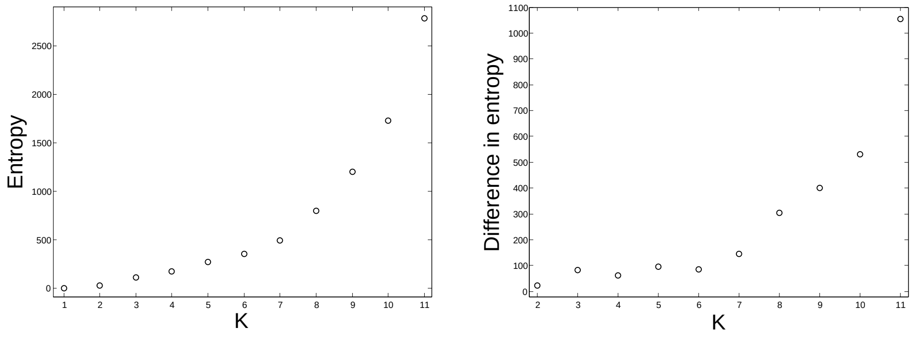
Similar to the result for the positive sample, when we combined the components in the BIC solution, the first few combinations took place within those components initially labeled CD3$^-$. When only six clusters remained, the blue and red components in Figure 9(a) combined, leaving the CD3$^+$ sub-populations to be represented by three clusters (Figure 9(c)).

After two more combinations (K=4), the green component merged with a big CD3$^-$ cluster. Finally we had a "clean" representation of the CD3$^+$ population with no observations from the CD3$^+$ CD4$^+$ CD8β^+ region. One last combination (K=3) merged the two remaining CD3$^-$ clusters, resulting in the most parsimonious view of the control sample with only three clusters but showing all the relevant features (Figure 9(d)). Once again, the entropy of the combined solution (106) was much smaller than that of the ICL solution (496). Note that in this case we ended up with a combined solution that has fewer clusters than the ICL solution. The plot of the entropy of the combined solution against the number of clusters (Figure 10) suggests an elbow at $K = 7$, but substantive considerations suggest that we can continue merging past this number.



(a) (b) (c) (d)

Figure 9: GvHD control sample. Only components labeled CD3$^+$ are shown. (a) BIC Solution (K=11). (b) ICL Solution (K=5). (c) Combined Solution (K=6). (d) Combined Solution (K=3).



(a) (b)

Figure 10: (a) Entropy values for the K-cluster Combined Solution for the GvHD Control Sample. (b) Differences between successive entropy values.

6 DISCUSSION

We have proposed a way of addressing the dilemma of model-based clustering based on Gaussian mixture models, namely that the number of mixture components selected is not necessarily equal to the number of clusters. This arises when one or more of the clusters has a non-Gaussian distribution, which is approximated by a mixture of several Gaussians.

Our strategy is as follows. We first fit a Gaussian mixture model to the data by maximum likelihood estimation, using BIC to select the number of Gaussian components. Then we successively combine mixture components using the entropy of the conditional membership distribution to decide which components to merge at each stage. This yields a sequence of possible solutions, one for each number of clusters, and in general we expect that users would consider these solutions from a substantive point of view.

The underlying statistical model is the same for each member of this sequence of solutions, in the sense that the likelihood and the modeled probability distribution of the data remain unchanged. What changes is the interpretation of this model. Thus standard statistical testing or model selection methods cannot be used to choose the preferred solution.

If a data-driven choice is required, however, we suggest inspecting the entropy plot and looking for an elbow, or choosing the number of clusters selected by ICL. An inferential choice could be made, for example using the gap statistic (Tibshirani, Walther, and Hastie 2001). However, the null distribution underlying the resulting test does not belong to the class of models being tested, so that it does not have a conventional statistical interpretation in the present context. It could still possibly be used in a less formal sense to help guide the choice of number of clusters.

Our method preserves the advantages of Gaussian model-based clustering, notably a good fit to the data, but it allows us to avoid the overestimation of the number of clusters that can occur when some clusters are non-Gaussian. The mixture distribution selected by BIC allows us to start the hierarchical procedure from a good summary of the data set. The resulting hierarchy is easily interpreted in relation to the mixture components. We stress that the whole hierarchy from K to 1 clusters might be informative. Note that our method can also be used when the number of clusters K^* is known, provided that the number of mixture components in the BIC solution is at least as large as K^*.

One attractive feature of our method is that it is computationally efficient, as it uses only the conditional membership probabilities. Thus it could be applied to any mixture model, and not just to a Gaussian mixture model, effectively without modification. This includes latent class analysis (Lazarsfeld 1950; Hagenaars and McCutcheon 2002), which is essentially model-based clustering for discrete data.

Several other methods for joining Gaussian mixture components to form clusters have been proposed. Walther (2002) considered the problem of deciding whether a univariate distribution is better modeled by a mixture of normals or by a single, possibly non-Gaussian and asymmetric distribution. To our knowledge, this idea has not yet been extended to more than one dimension, and it seems difficult to do so. Our method seems to provide a simple alternative approach to the problem addressed by Walther (2002), in arbitrary dimensions.

Wang and Raftery (2002, Section 4.5) considered the estimation of elongated features in a spatial point pattern with noise, motivated by a minefield detection problem. They suggested first clustering the points using Gaussian model-based clustering with equal spherical covariance matrices for the components. This leads to the feature being covered by a set of “balls” (spherical components), and these are then merged if their centers are close enough that the components are likely to overlap. This works well for joining spherical components, but may not work well if the components are not spherical, as it takes account of the component means but not their shapes.

Tantrum, Murua, and Stuetzle (2003) proposed a different method based on the hierarchical model-based clustering method of Banfield and Raftery (1993). Hierarchical model-based clustering is a "hard" clustering method, in which each data point is assigned to one group. At each stage, two clusters are merged, with the likelihood used as the criterion for deciding which clusters to merge. Tantrum et al. (2003) proposed using the dip test of Hartigan and Hartigan (1985) to decide on the number of clusters. This method differs from ours in two main ways. Ours is a probabilistic ("soft") clustering method that merges mixture components (distributions), and while that of Tantrum et al. (2003) is a hard clustering method that merges groups of data points. Secondly, the merging criterion is different.

Li (2005) assumed that the number of clusters K is known in advance, used BIC to estimate the number of mixture components, and joined them using k-means clustering applied to their means. This works well if the clusters are spherical, but may not work as well if they are elongated, as the method is based on the means of the clusters but does not take account of their shape. The underlying assumption that the number of clusters is known may also be questionable in some applications. Jörnsten and Keleş (2008) extended Li's method so as to apply it to multifactor gene expression data, allowing clusters to share mixture components, and relating the levels of the mixture to the experimental factors.

A Algorithm

Choose a family of mixture models: $\{\mathcal{M}_{K_{\min}},\dots,\mathcal{M}_{K_{\text{max}}}\}$. Complete Gaussian mixture models are suggested: $\mathcal{M}_K$ contains any mixture with K Gaussian components. Here is the algorithm we work with:

1. Compute MLE(K) for each model using the EM algorithm:

$$\forall K \in \{K_{\min},\dots,K_{\max}\},\quad \hat{\theta}_K = \arg\max_{\theta_K \in \Theta_K} \log p(\mathbf{x} \mid K, \theta_K)$$

2. Compute the BIC solution:

$$\hat{K}^{\text{BIC}} = \underset{K \in \{K_{\min},\dots,K_{\max}\}}{\text{argmin}} \left\{ -\log p(\mathbf{x} \mid K, \hat{\theta}_K) + \frac{\nu_K}{2}\log n \right\}$$

3. Compute the density f_k^K of each combined cluster k for each K from $\hat{K}^{\text{BIC}}$ to $K_{\min}$:

$$\forall k \in \{1,\dots,\hat{K}^{\text{BIC}}\},\quad f_k^{\hat{K}^{\text{BIC}}}(\cdot) = \hat{p}_k^{\hat{K}^{\text{BIC}}} \phi\left(\cdot \mid \hat{a}_k^{\hat{K}^{\text{BIC}}}\right).$$

For $K = \hat{K}^{\text{BIC}},\dots,(K_{\text{min}}+1)$:

- Choose the clusters l and l' to be combined at step $K \to K-1$:

$$(l,l') = \underset{(k,k')\in\{1,\dots,K\}^2,\ k\neq k'}{\text{argmax}} \left\{ -\sum_{i=1}^n \left\{ t_{ik}^K \log(t_{ik}^K) + t_{ik'}^K \log(t_{ik'}^K) \right\} + \sum_{i=1}^n (t_{ik}^K + t_{i\,k'}^K)\log(t_{ik}^K + t_{i\,k'}^K) \right\},$$

  where $t_{ik}^K = \frac{f_k^K(x_i)}{\sum_{j=1}^K f_j^K(x_i)}$ is the conditional probablity of component k given the K-cluster combined solution.

- Define the densities of the combined clusters for the (K-1) clusters solution by combining l and l':
  for $k = 1, \ldots, (l \wedge l' - 1), (l \wedge l' + 1), \ldots, (l \vee l' - 1)$ $\{f_k^{K-1} = f_k^K\}$
  $f_{l \wedge l'}^{K-1} = f_l^K + f_{l'}^K$
  for $k = l \vee l', \ldots, (K-1)$ $\{f_k^{K-1} = f_{k+1}^K\}$

4. To select the number of clusters through ICL:

$$\hat{K}^{\text{ICL}} = \underset{K \in \{K_{\min}, \ldots, K_{\max}\}}{\operatorname{argmin}} \left\{ -\log p(\mathbf{x} \mid K, \hat{\theta}_K) - \sum_{i=1}^{n} \sum_{k=1}^{K} t_{ik}(\hat{\theta}_K) \log t_{ik}(\hat{\theta}_K) + \frac{\nu_K}{2} \log n \right\},$$

where $t_{ik}(\hat{\theta}_K) = \frac{\hat{p}_k^K \phi(x_i | \hat{a}_k^K)}{\sum_{j=1}^{K} \hat{p}_j^K \phi(x_i | \hat{a}_j^K)}$. is the conditional probability of component k given the MLE for the model with K Gaussian components.

References

Banfield, J. D. and A. E. Raftery (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics 49*, 803–821.

Biernacki, C., G. Celeux, and G. Govaert (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*, 719–725.

Brinkman, R. R., M. Gasparetto, S.-J. J. Lee, A. J. Ribickas, J. Perkins, W. Janssen, R. Smiley, and C. Smith (2007). High-content flow cytometry and temporal data analysis for defining a cellular signature of graft-versus-host disease. *Biology of Blood and Marrow Transplantation 13*, 691–700.

Fraley, C. and A. E. Raftery (1998). How many clusters? Answers via model-based cluster analysis. *The Computer Journal 41*, 578–588.

Fraley, C. and A. E. Raftery (2002). Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association 97*, 611–631.

Hagenaars, J. A. and A. L. McCutcheon (2002). *Applied Latent Class Analysis*. Cambridge, U.K.: Cambridge University Press.

Hartigan, J. A. and P. M. Hartigan (1985). The dip test of unimodality. *Annals of Statistics 13*, 78–84.

Jörnsten, R. and S. Keleş (2008). Mixture models with multiple levels, with application to the analysis of multifactor gene expression data. *Biostatistics 9*, 540–554.

Kass, R. E. and A. E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association 90*, 773–795.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhya 62*, 49–66.

Lazarsfeld, P. F. (1950). The logical and mathematical foundations of latent structure analysis. In S. A. Stouffer (Ed.), *Measurement and Prediction, Volume IV of The American Soldier: Studies in Social Psychology in World War II*, pp. Chapter 10. Princeton University Press.

Li, J. (2005). Clustering based on a multilayer mixture model. *Journal of Computational and Graphical Statistics 14*, 547–568.

Lo, K., R. R. Brinkman, and R. Gottardo (2008). Automated gating of flow cytometry data via robust model-based clustering. *Cytometry A 73*, 321–32.

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

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics 6*, 461–464.

Steele, R. J. (2002). *Importance sampling methods for inference in mixture models and missing data.* Ph. D. thesis, Department of Statistics, University of Washington, Seattle, Wash.

Tantrum, J., A. Murua, and W. Stuetzle (2003). Assessment and pruning of hierarchical model based clustering. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, pp. 197–205. Association for Computing Machinery.

Tibshirani, R., G. Walther, and T. Hastie (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society, Series B 63*, 411–423.

Walther, G. (2002). Detecting the presence of mixing with multiscale maximum likelihood. *Journal of the American Statistical Association 97*, 508–513.

Wang, N. and A. E. Raftery (2002). Nearest neighbor variance estimation (NNVE): Robust covariance estimation via nearest neighbor cleaning (with discussion). *Journal of the American Statistical Association 97*, 994–1019.

Centre de recherche INRIA Saclay – Île-de-France
Parc Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 Orsay Cedex (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Grenoble – Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399