Supplementary materials for this article are available online.
Please click the JCGS link at http://pubs.amstat.org.

Combining Mixture Components for Clustering

Jean-Patrick BAUDRY, Adrian E. RAFTERY, Gilles CELEUX, Kenneth LO, and Raphaël GOTTARDO

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, and we also describe an automatic way of selecting the number of clusters via a piecewise linear regression fit to the rescaled entropy plot. We illustrate the method with simulated data and a flow cytometry dataset. Supplemental materials are available on the journal web site and described at the end of the article.

Key Words: BIC; Entropy; Flow cytometry; Mixture model; Model-based clustering; Multivariate normal distribution.

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

Jean-Patrick Baudry is Doctorant and Gilles Celeux is Directeur de Recherche, both at INRIA Saclay Île-de-France, Université Paris-Sud, Bâtiment 425, 91405 Orsay Cedex, France. Jean-Patrick Baudry is also Researcher, Laboratoire MAP5, Université Paris Descartes and CNRS. Adrian E. Raftery is Blumstein–Jordan Professor of Statistics and Sociology, Department of Statistics, Box 354322, University of Washington, Seattle, WA 98195-4322. Kenneth Lo is Postdoctoral Senior Fellow, Department of Microbiology, Box 358070, University of Washington, Seattle, WA 98195-8070. Raphaël Gottardo is Research Unit Director in Computational Biology, Institut de recherches cliniques de Montréal (IRCM), 110, avenue des Pins Ouest, Montréal, Canada H2W 1R7.

Journal of Computational and Graphical Statistics, Volume 19, Number 2, Pages 332–353
DOI: 10.1198/jcgs.2010.08111

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 the works of McLachlan (1982), McLachlan and Basford (1988), 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 terms of clustering by identifying a subset of the mixture components with each cluster. We suggest assessing all the resulting clusterings substantively. We also describe an automatic method for choosing the number of clusters based on a piecewise linear regression fit to the rescaled entropy plot. The number of clusters selected, either substantively or automatically, 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, Celeux, and Govaert 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

$$p(\mathbf{x}_i|K, \theta_K) = \sum_{k=1}^{K} p_k \phi(\mathbf{x}_i|\mathbf{a}_k), \tag{2.1}$$

where the p_k's are the mixing proportions ($0 < p_k < 1$ for all $k = 1, \ldots, K$ and $\sum_k p_k = 1$), $\phi(\cdot|\mathbf{a}_k)$ 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

$$p(\mathbf{x}|K) = \int p(\mathbf{x}|K, \theta_K)\pi(\theta_K)\, d\theta_K, \tag{2.2}$$

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

$$\mathrm{BIC}(K) = \log p(\mathbf{x}|K, \hat{\theta}_K) - \frac{\nu_K}{2}\log(n), \tag{2.3}$$

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. This was first applied to model-based clustering by Dasgupta and Raftery (1998). Keribin (1998, 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, Celeux, and Govaert 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 of Gaussians 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, Celeux, and Govaert (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

$$p(\mathbf{x}, \mathbf{z}|K) = \int_{\Theta_K} p(\mathbf{x}, \mathbf{z}|K, \theta)\pi(\theta|K)\, d\theta, \tag{2.4}$$

where

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

with

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

To approximate this integrated complete likelihood, Biernacki, Celeux, and Govaert (2000) proposed using a BIC-like approximation, leading to the criterion

$$\mathrm{ICL}(K) = \log p(\mathbf{x}, \hat{\mathbf{z}}|K, \hat{\theta}_K) - \frac{\nu_K}{2} \log n, \tag{2.5}$$

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

$$\mathrm{Ent}(K) = -\sum_{k=1}^{K} \sum_{i=1}^{n} t_{ik}(\hat{\theta}_K) \log t_{ik}(\hat{\theta}_K) \geq 0, \tag{2.6}$$

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 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

$$-\sum_{i=1}^{n}\Bigg(\sum_{j\neq k,k'} t_{ij}^K \log t_{ij}^K + (t_{ik}^K + t_{ik'}^K)\log(t_{ik}^K + t_{ik'}^K)\Bigg). \tag{3.1}$$

Thus, the two clusters k and k' to be combined are those that maximize the criterion

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

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

At the first step of the combining procedure, $K = \hat{K}^{\mathrm{BIC}}$ and t_{ik}^K is the conditional probability that $\mathbf{x}_i$ arises from the kth mixture component ($1 \le i \le n$ and $1 \le k \le K$). But as soon as at least two components are combined in a cluster k (hence $K < \hat{K}^{\mathrm{BIC}}$), t_{ik}^K is the conditional probability that observation $\mathbf{x}_i$ belongs to one of the combined components in cluster k.

Our method is a soft clustering one that yields probabilities of cluster membership rather than cluster assignments. However, it can be used as the basis for a hard clustering method, simply by assigning the maximum a posteriori cluster memberships. Note that this will not necessarily be a strictly hierarchical clustering method. For example, an observation that was not assigned to either cluster k or k' by the K-cluster solution might be assigned to cluster $k \cup k'$ by the $(K-1)$-cluster solution.

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 as 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 possibilities 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. This experiment was repeated with 100 different such datasets, but we first present a single one of them to illustrate

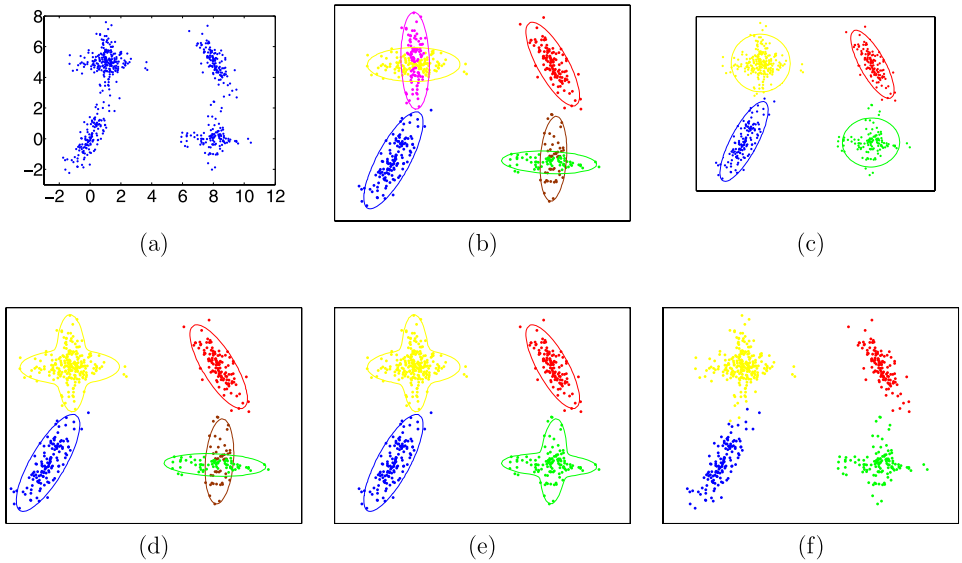


Figure 1. Simulated Example 1. (a) Simulated data from a six-component two-dimensional Gaussian mixture ($n = 600$). (b) BIC solution with six components ($K = 6, \text{Ent} = 122$). (c) ICL solution with four clusters ($K = 4, \text{Ent} = 3$). (d) Combined solution with five clusters ($K = 5, \text{Ent} = 41$). (e) Combined solution with four clusters ($K = 4, \text{Ent} = 5$). (f) The true labels for a four-cluster solution. In (b) and (c) the entropy, Ent, is defined by (2.6) with respect to the maximum likelihood solution, and in (d) and (e) Ent is defined by (3.1).

the method. Although all the features of our approach cannot be tabulated, results illustrating the stability of the method are reported and discussed at the end of this subsection.

For the dataset at hand, the BIC selected a six-component mixture model, which was the correct model; this is shown in Figure 1(b). ICL selected a four-cluster model, as shown in Figure 1(c). The four clusters found by ICL are well separated.

Starting from the BIC six-component solution, we combined two components to get the five-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 four-cluster solution, shown in Figure 1(e). This is the number of clusters identified by ICL. Note that there is no conventional formal 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.

However, the decrease of the entropy at each step of the procedure may help guide the choice of the number of clusters, or of a small number of solutions to be considered. 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 together with the choice of ICL, this leads us to focus on this solution.

A finer examination of those graphics gives more information about the merging process. The first merging (from six to five clusters) is clearly necessary, since the decrease in entropy is large (with respect, for example, to the minimal decreases, when merging from two to one cluster, say). The second merging (from five to four clusters) also seems to be necessary for the same reason, although it results in a smaller decrease of the entropy (about half of the first one). This is far from zero, but indicates either that the components involved in this merging overlap less than the first two to be merged, or that this merging involves only about half as many observations as the first merging.

To further analyze the situation, we suggest changing the scale of the first of those graphics so that the difference between the abscissas of two successive points is propor-

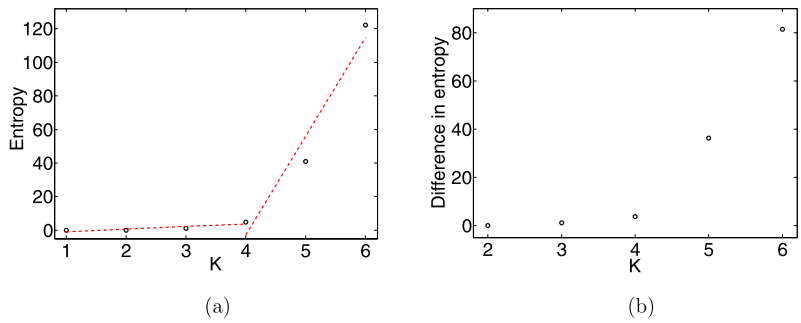


Figure 2. (a) Entropy values for the K-cluster combined solution, as defined by (3.1), for Simulated Example 1. The dashed line shows the best piecewise linear fit, with a breakpoint at $K = 4$ clusters. (b) Differences between successive entropy values. A color version of this figure is available in the electronic version of this article.

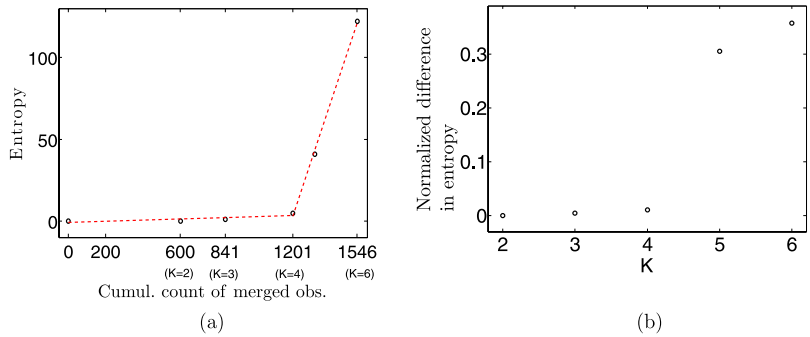


Figure 3. Simulated Example 1: (a) Entropy values for the K-cluster combined solution, as defined by (3.1), plotted against the cumulative sum of the number of observations merged at each step. The dashed line shows the best piecewise linear fit, with a breakpoint at $K = 4$ clusters. (b) Rescaled differences between successive entropy values: $\frac{\text{Ent}(K+1)-\text{Ent}(K)}{\text{Number of merged obs.}}$. A color version of this figure is available in the electronic version of this article.

tional to the number of observations involved in the corresponding merging step: see Figure 3(a). This plot leads to the conclusion that the reason why the second merging step gives rise to a smaller entropy decrease than the first one is that it involves fewer observations. The mean decrease in entropy for each observation involved in the corresponding merging step is about the same in both cases, since the last three points of this graphic are almost collinear. The same result can be seen in a slightly different way by plotting the differences of entropies divided by the number of observations involved at each step, as shown in Figure 3(b). These new graphical representations accentuate the elbow at $K = 4$.

In the four-cluster solution, the clusters are no longer all Gaussian; now two of them are modeled as mixtures of two Gaussian components 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 three of the 600 points were misclassified.

It will often be scientifically useful to examine the full sequence of clusterings that our method yields and assess them on substantive grounds, as well as by inspection of the entropy plots. However, in some cases an automatic way of choosing the number of clusters may be desired. A simple approach to this was proposed by Byers and Raftery (1998) in a different context, namely to fit a two-part piecewise linear regression model to the values in the entropy plot, and use the estimated breakpoint as the selected number of clusters.

For Simulated Example 1, this is shown as the dashed line in Figure 2(a) for the raw entropy plot and in Figure 5(a) (Section 4.2) for the rescaled entropy plot. The method chooses $K = 4$ using both the raw and rescaled entropy plots, but the fit of the piecewise linear regression model is better for the rescaled entropy plot, as expected.

We repeated this experiment 100 times to assess the stability of the method, simulating new data from the same model each time. The piecewise linear regression model fit to the rescaled entropy plot selected $K = 4$, 95 times out of 100.

We carried out an analogous experiment in dimension 6. The "crosses" involved two components each, with four discriminant directions between them and two noisy direc-

tions. The proportions of the components were equal. Our piecewise linear regression model method almost always selected four clusters.

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 10 (Figure 4(a)). This is because the true 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 4(b)).

In the hierarchical merging process, the two components of one of the "crosses" were combined first (Figure 4(c)), followed by the components of the other cross (Figure 4(d)). 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 4(e)). Next, two of the three mixture components representing the upper right cluster were combined (Figure 4(f)). After the next step there were five clusters, and all three mixture components representing the lower left cluster had been combined (Figure 4(g)).

The next step got us to four clusters, the number identified by ICL (Figure 4(h)). 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 5 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.

In this example, the elbow graphics (Figure 5(a) and (c)) exhibit three different stages in the merging process (a two-change-point piecewise line is necessary to fit them well):

- The two first merging steps (from ten to eight clusters) correspond to a large decrease in entropy (Figure 5(a)). They are clearly necessary. The mean entropy is equivalent in each one of those two steps (Figure 5(c)). Indeed, Figure 4 shows that they correspond to the formation of the two crosses.

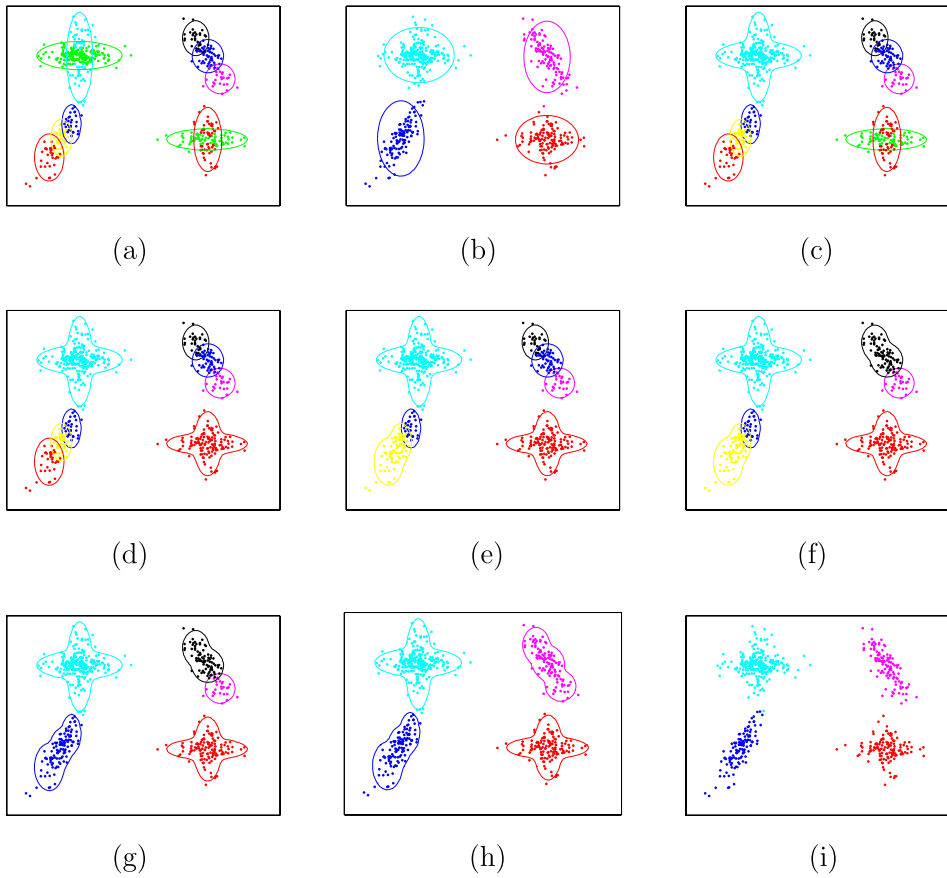


Figure 4. Simulated Example 2. The data are the same as in Simulated Example 1, but the model space is more restrictive, as only Gaussian mixture models with diagonal covariance matrices are considered. See Figure 1 legends for explanations about Ent. (a) BIC solution with ten mixture components ($K = 10, \text{Ent} = 179$). (b) ICL solution with four clusters ($K = 4, \text{Ent} = 3$). (c) Combined solution with nine clusters ($K = 9, \text{Ent} = 100$). (d) Combined solution with eight clusters ($K = 8, \text{Ent} = 52$). (e) Combined solution with seven clusters ($K = 7, \text{Ent} = 37$). (f) Combined solution with six clusters ($K = 6, \text{Ent} = 23$). (g) Combined solution with five clusters ($K = 5, \text{Ent} = 10$). (h) Combined solution with four clusters ($K = 4, \text{Ent} = 3$). (i) True labels with four clusters.

- The four following merging steps (from eight to four clusters) correspond to smaller decreases in entropy (Figure 5(a)). They have a comparable common mean decrease of entropy, but it is smaller than that of the first stage: a piece of the line would be fitted for them only (as appears in Figure 5(c)). They correspond to the merging of components which overlap in a different way than those merged at the first stage (Figure 4).
- The four last merging steps should not be applied.

In this case the user can consider the solutions with four and eight clusters, and take a final decision according to the needs of the application. The automatic rule in Section 4.1 (see Figure 5(d)) selects $K = 6$ clusters, which splits the difference between the two solu-

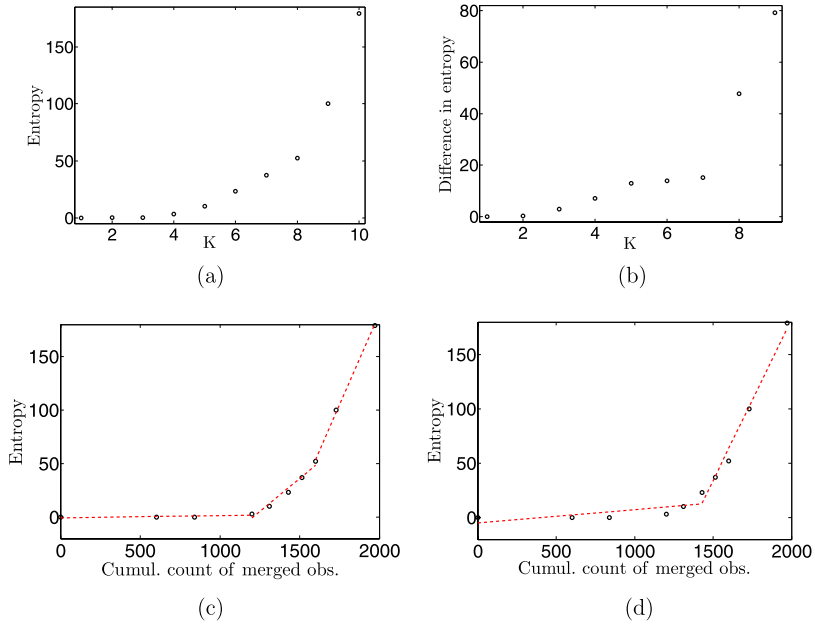


Figure 5. (a) Entropy values for the K-cluster combined solution, as defined by (3.1), for Simulated Example 2. (b) Differences between successive entropy values. (c) Entropy values with respect to the cumulative sum of the number of observations merged at each step $K+1 \to K$. Two-change-points piecewise linear regression. (d) Entropy values with respect to the cumulative sum of the number of observations merged at each step $K+1 \to K$. Single-change-point piecewise linear regression with minimum least squares choice of the change-point. A color version of this figure is available in the electronic version of this article.

tions we identified by inspection of the plot. This seems reasonable if a single automatic choice is desired, but either four or eight clusters might be better in specific contexts.

4.3 CIRCLE/SQUARE EXAMPLE

The data shown in Figure 6(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, Celeux, and Govaert (2000), BIC chose two components in only 60% of the simulated cases. Here we show one simulated dataset in which BIC approximated the underlying non-Gaussian density using a mixture of five normals (Figure 6(b)). ICL always selected two clusters (Figure 6(c)).

The progress of the combining algorithm is shown in Figure 6(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

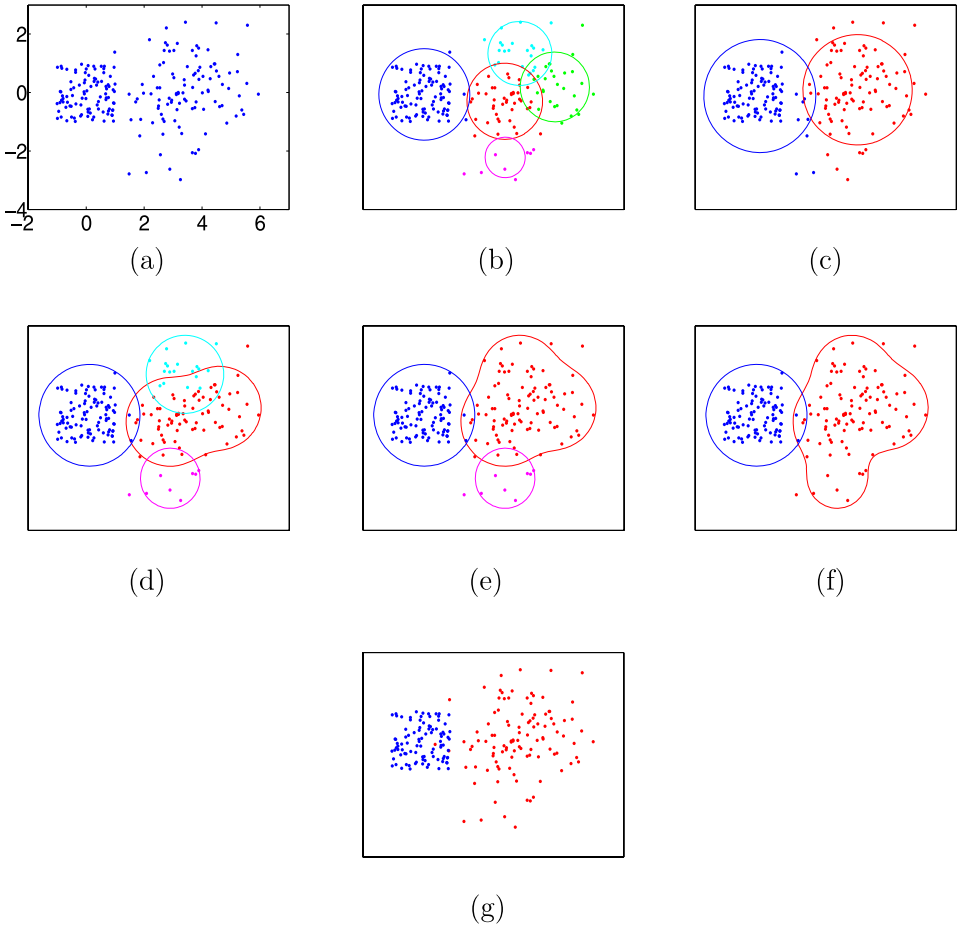


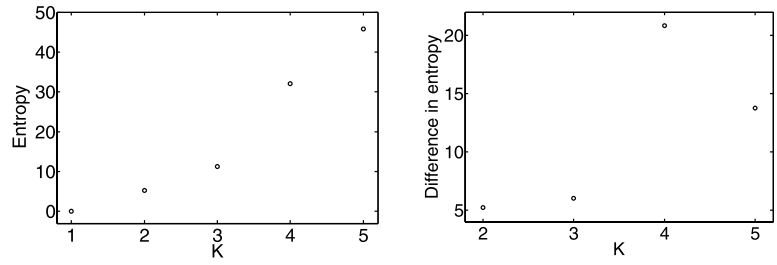
Figure 6. Circle-Square Example. See Figure 1 legends for explanations about Ent. (a) Observed data simulated from a mixture of a uniform distribution on a square and a spherical Gaussian distribution ($n = 200$). (b) The BIC solution, with five components ($K = 5, \text{Ent} = 46$). (c) The ICL solution with two clusters ($K = 2, \text{Ent} = 14$). (d) The combined solution with four clusters ($K = 4, \text{Ent} = 32$). (e) The combined solution with three clusters ($K = 3, \text{Ent} = 11$). (f) The final combined solution, with two clusters ($K = 2, \text{Ent} = 5$). (g) The true labels.

true labels are shown in Figure 6(g). The entropy plot in Figure 7 does not have a clear elbow.

4.4 COMPARISON WITH LI'S METHOD

In this section, our methodology is compared with the related method of Li (2005). Similarly to our approach, Li proposed modeling clusters as Gaussian mixtures, starting with the BIC solution, and then merging mixture components. However, unlike us, Li assumed that the true number of clusters is known in advance. The author also used k-means clustering to merge components; this works well when the mixture components are spherical but may have problems when they are not.

In the framework of the so-called multilayer mixture model, Li (2005) proposed two methods for partitioning the components of a mixture model into a fixed number of clus-



(a) (b)

Figure 7. (a) Entropy values for the K-cluster combined solution, as defined by (3.1), for the Circle-Square Example. (b) Differences between successive entropy values.

ters. They are both initialized with the same double-layer k-means procedure. Then the first method consists of computing the maximum likelihood estimator of a Gaussian mixture model with a greater number of components than the desired number of clusters. The components are then merged by minimizing a within-cluster inertia criterion (sum of squares) on the mean vectors of the mixture components. The second method consists of fitting the Gaussian mixture model through a CEM-like algorithm (Celeux and Govaert 1992), to maximize the classification likelihood, where the clusters are taken as mixtures of components. The total number of components (for each method) and the number of components per cluster (for the CEM method) are selected through either BIC or ICL.

4.4.1 First Experiment: Gaussian Mixture

We simulated 100 samples of size $n = 800$ of a four-component Gaussian mixture in $\mathbf{R}^2$. An example of such a sample is shown in Figure 8(a).

Since Li's method imposes a fixed number of clusters, we fixed it to three and stopped our algorithm as soon as it yielded three clusters. For each simulated sample we always obtained the same kind of result for both methods. They are depicted in Figure 8(b) for our method, which always gave the same result. Figure 8(c) shows the results for the four variants of Li's method. Li's method with the CEM-like algorithm always gave rise to the solution in Figure 8(c). Li's method with the k-means on the means and the selection through BIC found the same solution in 93 of the 100 cases. The method with the k-means on the means and the selection through ICL found such a solution in 27 cases, but in most other cases found a different solution whose fit was poorer (Figure 8(d)).

4.4.2 Second Experiment: 3D Uniform Cross

We simulated data in $\mathbf{R}^3$ from a mixture of two uniform components; see Figure 9. One is a horizontal thick cross (red in Figure 9) and has proportion 0.75 in the mixture, while the other is a vertical pillar (black in Figure 9) and has proportion 0.25. We simulated 100 datasets of size 300, and we applied Li's procedures, Ward's sum of squares method, and ours. We fixed the number of clusters to be designed at its true value (two), and we then fitted general Gaussian mixture models.

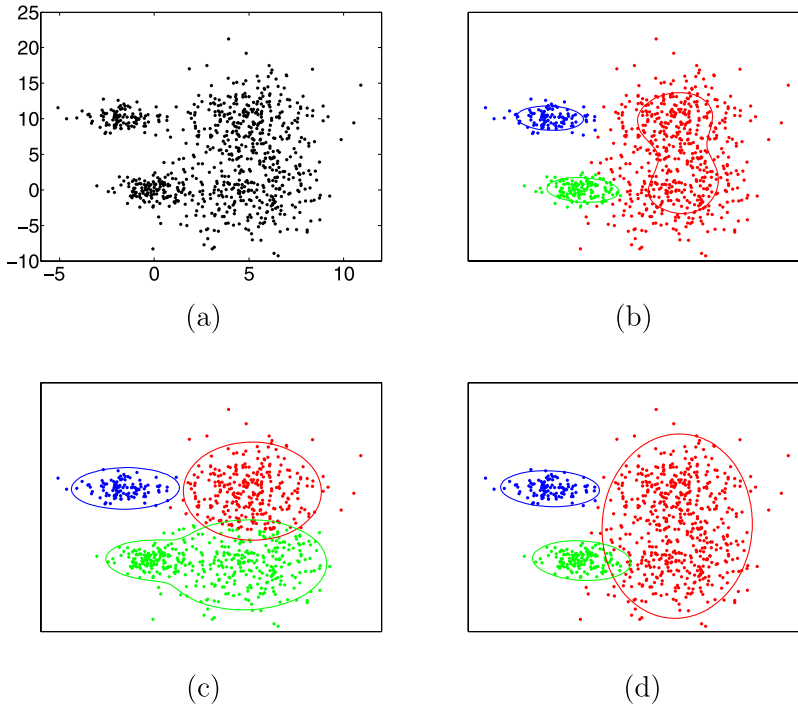


Figure 8. Comparison with Li's method. (a) A simulated dataset to compare Li's method with ours. (b) The three-cluster solution with our method. (c) The three-cluster solution with most of Li's methods. (d) The typical three-cluster solution with Li's "k-means on the means + ICL" method.

BIC selected four components for 69 of the 100 datasets, and three components for 18 of them. ICL selected four components for 60 of the datasets, and three components for 29 of them.

As in the preceding example, Li's approach did not recover the true clusters. Li's CEM-like methods always yielded a bad solution: sometimes one of the arms of the cross merged to the pillar, and sometimes two, as in Figure 10. Li's BIC + k-means method recovered

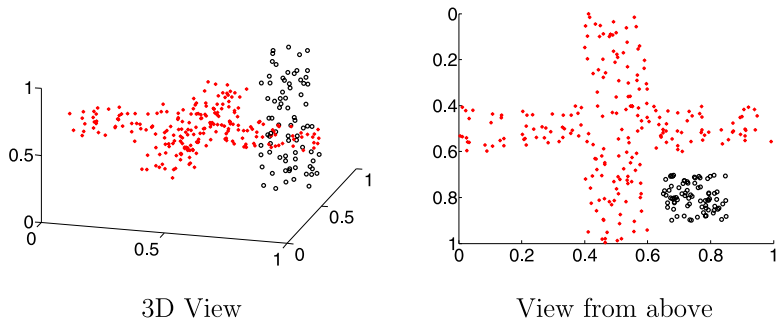


Figure 9. Simulated Example 2: There are two clusters, the 3D cross (red) and the uniform pillar (black). The true cluster memberships are shown here. A color version of this figure is available in the electronic version of this article.

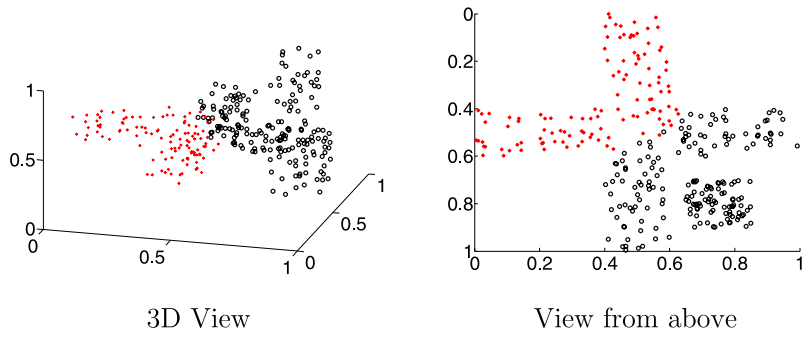


Figure 10. Example solution with Li's procedures. A color version of this figure is available in the electronic version of this article.

the true clusters in 19 cases out of 100, and Li's ICL + k-means method did so in 33 cases out of 100. This occurred almost every time the number of Gaussian components was 3 (two for the cross, which then have almost the same mean, and one for the pillar). When the number of fitted components is higher, the distance between the means of the components is no longer a relevant criterion, and those methods yielded clusterings such as Figure 10.

Our merging procedure almost always (95 times out of 100) recovered the true clusters.

4.4.3 Conclusions on the Comparisons With Other Methods

Here are some comments on the comparison between Li's methods and ours based on these simulations. Our method takes into account the overlap between components to choose which ones to merge, whereas Li's method is based on the distances between the component means, through the initialization step in each method, and also through the merging procedure in the first method. This sometimes leads to mergings that are not relevant from a clustering point of view.

Our method is appreciably faster since only one EM estimation has to be run for each considered number of components, whereas numerous runs are needed with Li's method. For the examples we considered, and with codes which should still be optimized, the time has been multiplied by a factor of at least 2.

Our procedure can also be applied when the number of clusters is unknown, unlike Li's method.

We also compared our results with those of a non-model-based clustering method: Ward's hierarchical method (Ward 1963). We used Matlab's *datacluster* function to apply this procedure in each of the experiments described in this section. Ward's method always found irrelevant solutions, close to Li's ones, for each of the 200 ($= 2 \times 100$) datasets.

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 dis-

ease (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 9083 and 6809 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 280 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 six CD3$^+$ cell sub-populations (Brinkman et al. 2007). ICL selected nine clusters, five of which correspond to the CD3$^+$ population (Figure 11(b)). Compared with the result shown in the work of Lo, Brinkman, and Gottardo (2008), the CD4$^+$ CD8β^- region located at the bottom right of the graph is missing.

BIC selected 12 components to provide a good fit to the positive sample, six of which are labeled CD3$^+$ (Figure 11(a)). The CD4$^+$ CD8β^+ region seems to be encapsulated by the cyan, green, and red components. Starting from this BIC solution, we repeatedly combined two components causing maximal reduction in the entropy. The first three combinations all occurred within those components originally labeled CD3$^-$, and the CD4 versus CD8β projection of the CD3$^+$ sub-populations remains unchanged.

The combined solution with nine clusters, in which six are labeled CD3$^+$, provides the most parsimonious view of the positive sample while retaining the six important CD3$^+$ cell sub-populations. However, when the number of clusters is reduced to eight, the magenta cluster representing the CD3$^+$ CD4$^+$ CD8β^- population is combined with the big CD3$^-$ cluster, resulting in an incomplete representation of the CD3$^+$ population (Figure 11(c)). Note that the entropy of the combined solution with nine clusters (1474) was smaller (i.e., better) than that of the ICL solution (3231). The entropy plot along with the piecewise regression analysis (Figure 12) suggests an elbow at $K = 9$ clusters, agreeing with the number of clusters returned by the ICL as well as 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 seven clusters, three of which correspond to the CD3$^+$ population (Figure 13(b)). The red cluster on the left of the graph represents the CD4$^-$ region. The blue cluster at the bottom right of the graph represents the CD4$^+$ CD8β^- region. It seems that it misses a part of this cluster near the red cluster. In addition, contrary to previous findings in which CD4$^+$ CD8β^+ cell sub-populations were found only in positive samples but not in control samples, a cyan cluster is used to represent the observations in the CD4$^+$ CD8β^+ region. These suggest that the ICL solution could be improved.

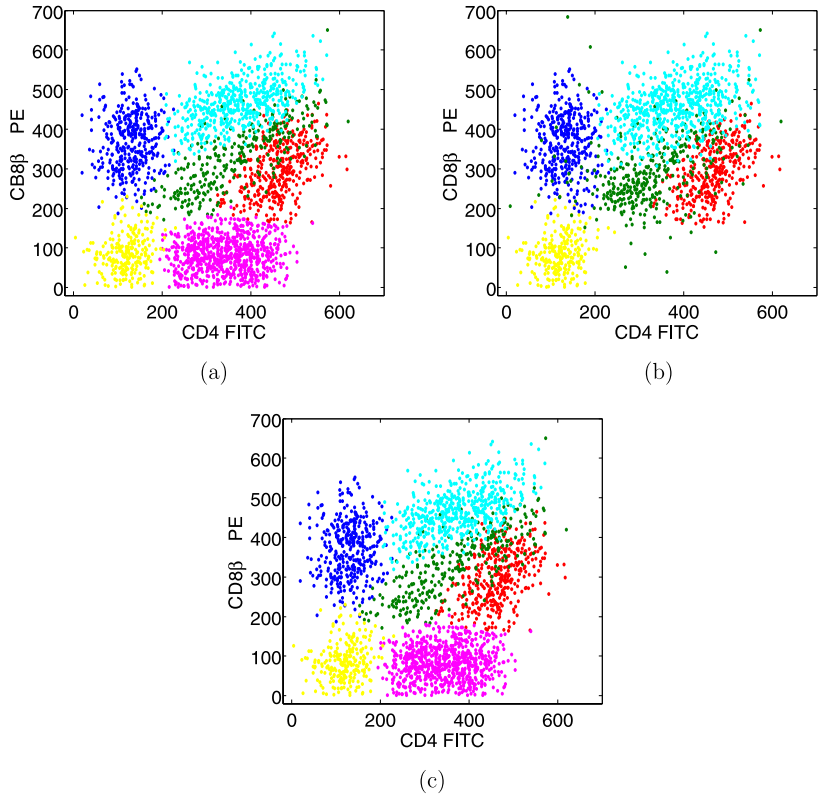


Figure 11. GvHD positive sample. Only components labeled $CD3^+$ are shown. (a) BIC solution ($K = 12$, Ent $= 4782$). The combined solutions for $K = 11$ and $K = 10$ are almost identical for these $CD3^+$ components. (b) ICL solution ($K = 9$, Ent $= 3235$). (c) Combined solution ($K = 9$, Ent $= 1478$).

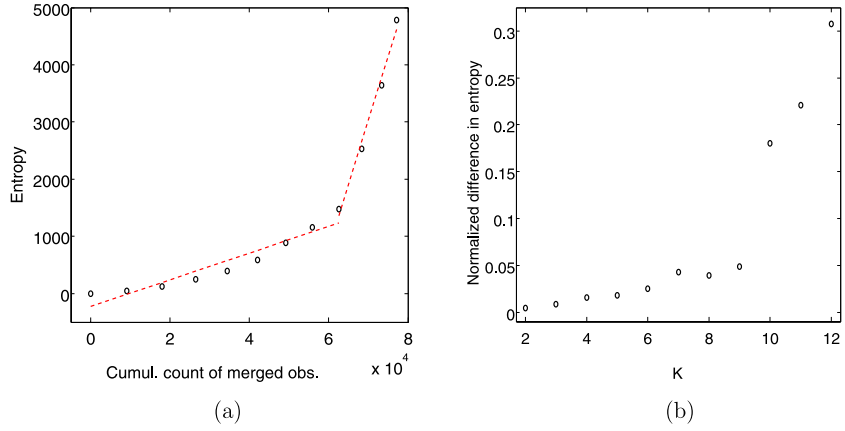


Figure 12. (a) Entropy values for the GvHD positive sample. The piecewise regression analysis suggests choosing $K = 9$ clusters. (b) Differences between successive entropy values. A color version of this figure is available in the electronic version of this article.

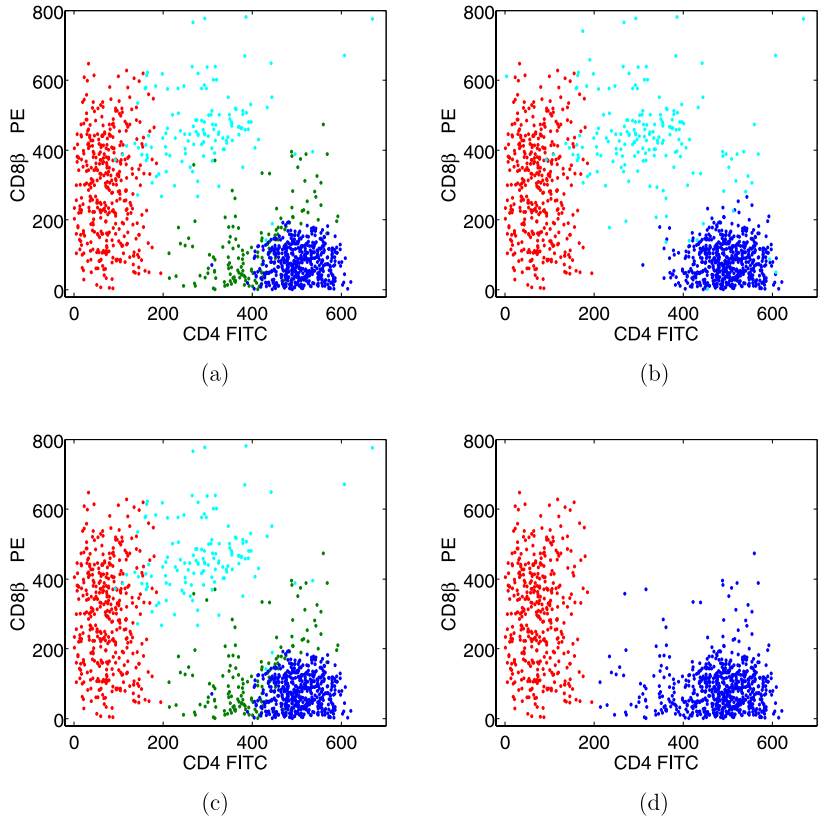


Figure 13. GvHD control sample. Only components labeled $CD3^+$ are shown. (a) BIC solution ($K = 10, \text{Ent} = 3733$). (b) ICL solution ($K = 7, \text{Ent} = 1901$). (c) Combined solution ($K = 6, \text{Ent} = 367$). (d) Combined solution ($K = 3, \text{Ent} = 58$).

BIC selected 10 components, four of which are labeled $CD3^+$ (Figure 13(a)). A green component is found next to the blue component, filling in the missing part in the ICL solution and resulting in a more complete representation of the $CD4^+$ $CD8\beta^-$ region. Meanwhile, similarly to the ICL solution, a cyan component is used to represent the observations scattered within the $CD4^+$ $CD8\beta^+$ region.

When we combined the components in the BIC solution, the first few combinations took place within those components initially labeled $CD3^-$, similarly to the result for the positive sample. Going from $K = 5$ to $K = 4$, the blue and green components in Figure 13(a) were combined, leaving the $CD3^+$ sub-populations to be represented by three clusters.

After one more combination ($K = 3$), the cyan component merged with a big $CD3^-$ cluster. Finally we had a "clean" representation of the $CD3^+$ population with no observations from the $CD3^+$ $CD4^+$ $CD8\beta^+$ region, consistent with the results of Brinkman et al. (2007) and Lo, Brinkman, and Gottardo (2008). This solution results in the most parsimonious view of the control sample with only three clusters but showing all the relevant features (Figure 13(d)). Once again, the entropy of the combined solution (58) was much smaller than that of the ICL solution (1895). Note that in this case we ended up with a com-

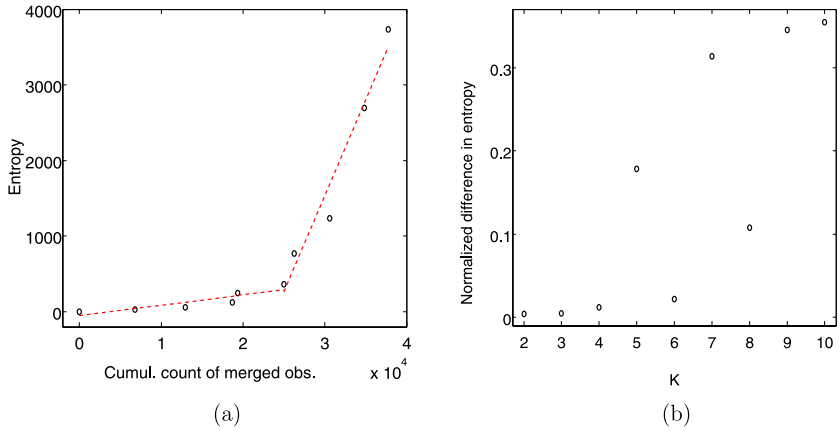


Figure 14. (a) Entropy values for the K-cluster combined solution for the GvHD control sample. The piecewise regression analysis suggests choosing $K = 6$ clusters. (b) Differences between successive entropy values. A color version of this figure is available in the electronic version of this article.

bined solution that has fewer clusters than the ICL solution. The entropy plot along with the piecewise regression analysis (Figure 14) suggests an elbow at $K = 6$, but substantive considerations suggest that we can continue merging past this number.

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 also describe two automatic ways of selecting the number of clusters, one based on a piecewise linear regression fit to the rescaled entropy plot, the other 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 dataset. 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.

Our merging procedure generally improves the entropy over the ICL solution. This highlights the better fit of the clusters that result from the merging procedure. 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, sec. 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, Murua, and Stuetzle (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), while that of Tantrum, Murua, and Stuetzle (2003) is a hard clustering method that merges groups of data points. Second, the merging criterion is different.

As discussed earlier, 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.

SUPPLEMENTAL MATERIALS

Appendix: An appendix describing the algorithm we used to apply the merging method. (Appendix.pdf)

Codes: The computer code and the datasets we used to illustrate the article. This computer code is implemented in Matlab. We used the MIXMOD software (*http://www.mixmod.org*) to run the EM algorithm for the estimation of the mixture parameters. (Codes.zip)

ACKNOWLEDGMENTS

Raftery's research was supported by NIH-NICHD grant HD-54511, NSF grant llS-0534094, and NSF grant ATM-0724721. The research of Lo and Gottardo was supported by a discovery grant of the Natural Sciences and Engineering Research Council of Canada and by NIH grant R01-EB008400. The authors are grateful to Naiysin Wang for suggesting the idea of using entropy as the criterion for the mixture component merging procedure, which is fundamental to the article. They also thank Christian Hennig for helpful discussions.

[Received August 2008. Revised March 2010.]

REFERENCES

Banfield, J. D., and Raftery, A. E. (1993), "Model-Based Gaussian and Non-Gaussian Clustering," *Biometrics*, 49, 803–821. [351]

Biernacki, C., Celeux, G., and Govaert, G. (2000), "Assessing a Mixture Model for Clustering With the Integrated Completed Likelihood," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22, 719–725. [333-335,342]

Brinkman, R. R., Gasparetto, M., Lee, S.-J. J., Ribickas, A. J., Perkins, J., Janssen, W., Smiley, R., and Smith, C. (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. [346,347,349]

Byers, S. D., and Raftery, A. E. (1998), "Nearest Neighbor Clutter Removal for Estimating Features in Spatial Point Processes," *Journal of the American Statistical Association*, 93, 577–584. [339]

Celeux, G., and Govaert, G. (1992), "A Classification EM Algorithm for Clustering and Two Stochastic Versions," *Computational Statistics & Data Analysis*, 14, 315–332. [344]

Dasgupta, A., and Raftery, A. E. (1998), "Detecting Features in Spatial Point Processes With Clutter via Model-Based Clustering," *Journal of the American Statistical Association*, 93, 294–302. [334]

Fraley, C., and Raftery, A. E. (1998), "How Many Clusters? Answers via Model-Based Cluster Analysis," *The Computer Journal*, 41, 578–588. [333,334]

——— (2002), "Model-Based Clustering, Discriminant Analysis and Density Estimation," *Journal of the American Statistical Association*, 97, 611–631. [333]

Hagenaars, J. A., and McCutcheon, A. L. (2002), *Applied Latent Class Analysis*, Cambridge, U.K.: Cambridge University Press. [351]

Hartigan, J. A., and Hartigan, P. M. (1985), "The Dip Test of Unimodality," *The Annals of Statistics*, 13, 78–84. [351]

Jörnsten, R., and Keleş, S. (2008), "Mixture Models With Multiple Levels, With Application to the Analysis of Multifactor Gene Expression Data," *Biostatistics*, 9, 540–554. [352]

Kass, R. E., and Raftery, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795. [334]

Keribin, C. (1998), "Consistent Estimate of the Order of Mixture Models," *Comptes Rendus de l'Acamdémie des Sciences, Série I-Mathématiques*, 326, 243–248. [334]

——— (2000), "Consistent Estimation of the Order of Mixture Models," *Sankhyā*, 62, 49–66. [334]

Lazarsfeld, P. F. (1950), "The Logical and Mathematical Foundations of Latent Structure Analysis," in *Measurement and Prediction. The American Soldier: Studies in Social Psychology in World War II*, Vol. IV, eds. S. A. Stouffer et al., Princeton, NJ: Princeton University Press, Chapter 10. [351]

Li, J. (2005), "Clustering Based on a Multilayer Mixture Model," *Journal of Computational and Graphical Statistics*, 14, 547–568. [343,351]

Lo, K., Brinkman, R. R., and Gottardo, R. (2008), "Automated Gating of Flow Cytometry Data via Robust Model-Based Clustering," *Cytometry A*, 73, 321–332. [347,349]

McLachlan, G. (1982), "The Classification and Mixture Maximum Likelihood Approaches to Cluster Analysis," in *Handbook of Statistics*, Vol. 2, eds. P. Krishnaiah and L. Kanal, Amsterdam: North-Holland, pp. 199–208. [333]

McLachlan, G., and Basford, K. (1988), *Mixture Models: Inference and Applications to Clustering*, New York: Dekker. [333]

McLachlan, G. J., and Peel, D. (2000), *Finite Mixture Models*, New York: Wiley. [333]

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464. [334]

Steele, R. J. (2002), "Importance Sampling Methods for Inference in Mixture Models and Missing Data," Ph.D. thesis, University of Washington, Dept. of Statistics, Seattle, WA. [333,334]

Tantrum, J., Murua, A., and Stuetzle, W. (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: Association for Computing Machinery, pp. 197–205. [351]

Tibshirani, R., Walther, G., and Hastie, T. (2001), "Estimating the Number of Clusters in a Data Set via the Gap Statistic," *Journal of the Royal Statistical Society, Ser. B*, 63, 411–423. [350]

Walther, G. (2002), "Detecting the Presence of Mixing With Multiscale Maximum Likelihood," *Journal of the American Statistical Association*, 97, 508–513. [351]

Wang, N., and Raftery, A. E. (2002), "Nearest Neighbor Variance Estimation (NNVE): Robust Covariance Estimation via Nearest Neighbor Cleaning" (with discussion), *Journal of the American Statistical Association*, 97, 994–1019. [351]

Ward, J. H. (1963), "Hierarchical Grouping to Optimize an Objective Function," *Journal of the American Statistical Association*, 58, 236–244. [346]