Ridgeline plot and clusterwise stability as tools for merging Gaussian mixture components

Christian Hennig[1]

Department of Statistical Science, UCL, Gower St., London WC1E 6BT, United Kingdom, chrish@stats.ucl.ac.uk

Summary. The problem of merging Gaussian mixture components is discussed in situations where a Gaussian mixture is fitted but the mixture components are not separated enough from each other to interpret them as "clusters". Two methods are introduced, corresponding to two different "cluster concepts" (separation by gaps and "data patterns"). A visualisation of the modality of a density of a mixture of two Gaussians is proposed and the stability of the unmerged Gaussian mixture is compared to that of clusterings obtained by merging components.

Key words: model-based cluster analysis, multilayer mixture, unimodality

1 Introduction

The Gaussian mixture model is often used for cluster analysis [1]. $I\!R^p$-valued observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are modelled as i.i.d. with density

$$f(\mathbf{x}) = \sum_{j=1}^{s} \pi_j \varphi_{\mathbf{a}_j, \Sigma_j}(\mathbf{x}), \tag{1}$$

where $\pi_j > 0\ \forall j,\ \sum_{j=1}^{s} \pi_j = 1,\ \varphi_{\mathbf{a},\Sigma}$ is the density of the p-dimensional Gaussian distribution $\mathcal{N}(\mathbf{a}, \Sigma)$ with mean vector $\mathbf{a}$ and covariance matrix Σ. Given a fixed s, the parameters can be estimated by Maximum Likelihood using the EM algorithm. The data points can then be classified to the mixture components by maximizing the estimated a posteriori probability that $\mathbf{x}_i$ was generated by mixture component j,

$$\hat{P}(\gamma_i = j | \mathbf{x}_i = \mathbf{x}) = \frac{\hat{\pi}_j \varphi_{\hat{\mathbf{a}}_j, \hat{\Sigma}_j}(\mathbf{x}))}{\sum_{l=1}^{s} \hat{\pi}_l \varphi_{\hat{\mathbf{a}}_l, \hat{\Sigma}_l}(\mathbf{x})}, \tag{2}$$

where γ_i is defined by the two-step version of the mixture model where

$$P(\gamma_i = j) = \pi_j,\ \mathbf{x}_i | (\gamma_i = j) \sim \varphi_{\mathbf{a}_j, \Sigma_j},\ i = 1, \ldots, n, \text{ i.i.d.} \tag{3}$$

Estimators are denoted by “hats”. A standard method to estimate the number of components s is the Bayesian Information Criterion (BIC). This can also be used to estimate suitable constraints on the covariance matrices [1]. For the present paper, the Gaussian mixture model has been fitted using the default options of the add-on package MCLUST version 3 [2] of the statistical software R (`www.R-project.org`).

In cluster analysis usually every mixture component is interpreted as a cluster, and pointwise maximization of (2) defines the clustering. However, this is often not justified. Some mixtures of more than one Gaussian distribution are unimodal, and in reality, model assumptions are never precisely fulfilled and Gaussian mixtures are a very flexible tool to fit all kinds of densities. But this means that a population that can be interpreted as “homogeneous” could be fitted by a mixture of more than one Gaussian mixture component.

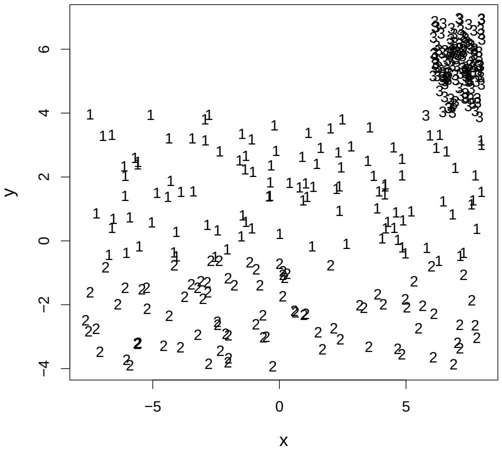

Fig. 1. Data from unimodal mixture of uniforms with 3-cluster solution by MCLUST

Figure 1 shows an artificial dataset generated from a mixture of two uniform distributions. MCLUST estimates $s = 3$ for this dataset. Note that the “correct” number of clusters is not well defined and one may see two “true” clusters here if “distiguishable patterns” are interpreted to be clusters or one “true” cluster if clusters are associated with density modes or are required to be separated by gaps. However, three is difficult to justify as the number of clusters here.

In the present paper, two methods are introduced that can be used to decide whether and which Gaussian mixture components should be merged. A method based on detecting density gaps is introduced in Section 2. A method based on estimating misclassification probabilities is introduced in Section 3. Detailed background for both methods along with some alternatives is given in [5], so they are only explained very briefly here. The following two tools are introduced exclusively in the present paper: the idea of Section 2 can be used to define "ridgeline plots" that show how strongly mixture components are separated, and in Section 4 the clusterwise stability assessment method introduced in [4] is proposed to compare the stability of the MCLUST and the merged clustering solution.

A real dataset from musicology is analysed in Section 5 and Section 6 concludes the paper.

Further methods for merging Gaussian mixture components are given in [8] and [6]. Both of them are discussed in [5].

2 The Ridgeline Method

In [7] it is shown that for any mixture f of s Gaussian distributions on $\mathbb{R}^p$ there is an $s-1$-dimensional manifold of $\mathbb{R}^p$ so that all extremal points of f lie on this manifold.

For $s = 2$, this manifold is defined by the so-called "ridgeline",

$$\mathbf{x}^*(\alpha) = [(1-\alpha)\Sigma_1^{-1} + \alpha\Sigma_2^{-1}]^{-1}[(1-\alpha)\Sigma_1^{-1}\mathbf{a}_1 + \alpha\Sigma_2^{-1}\mathbf{a}_2], \tag{4}$$

and all density extrema (and therefore all modes, which may be more than 2 in some situations) can be found for $\alpha \in [0, 1]$. The following algorithm ("ridgeline method") can be used to merge mixture components:

1. Choose a tuning constant $r^* < 1$.
2. Start with all components of the initially estimated Gaussian mixture as current clusters.
3. Using the mean vectors and covariance matrices of the current clusters, for any pair of current clusters compute, from (4), $r = \frac{\min_{0\leq\alpha\leq 1} f(\mathbf{x}^*(\alpha))}{m_2(f(\mathbf{x}^*(\alpha)))}$, where m_2 denotes the second largest mode; let $r = 1$ if there is only one mode.
4. If $r < r^*$ for all pairs of current clusters, use the current clustering as the final one.
5. Otherwise, merge the pair of current clusters with maximum r and go to step 3.

Following [5], $r^* = 0.2$ is used here. Note that it is not advisable to demand precise unimodality ($r^* = 1$), because the probability is high that MCLUST estimates multimodal mixtures even in unimodal situations. For example, for the (unimodal) dataset in Figure 1, the first two components are merged at $r = 0.58$ and their union is merged with the third one at $r = 0.21$.

The separation of the estimated Gaussian mixture components can be visualised by plotting the ridgeline density $f(\mathbf{x}^*(\alpha))$ vs. α. The results for the dataset in Figure 1 are shown in Figure 2.

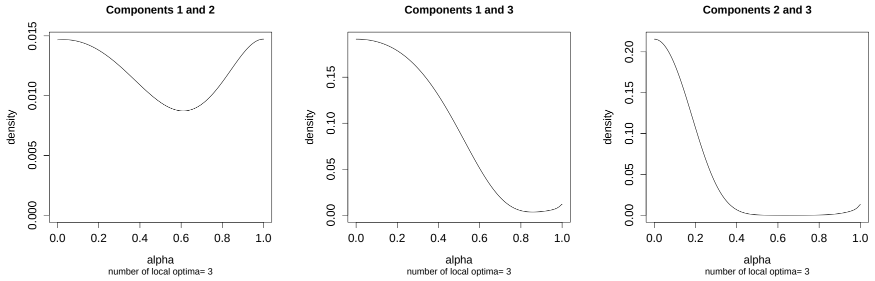

Fig. 2. Ridgeline plot for the three Gaussian components estimated by MCLUST for the data in Figure 1.

3 A Method Based on Misclassification Probabilities

Misclassification probabilities $p_{ij} = P(\tilde{\gamma}_1^* = i|\gamma_1^* = j) = \frac{P(\tilde{\gamma}_1^*=i,\ \gamma_1^*=j)}{\pi_j^*}$ between components of a mixture distribution can be estimated directly from the results of the EM algorithm. γ_k^* here denotes the membership indicator of observation k in a Gaussian mixture model, but where a cluster may be identified with a single mixture component (in which case $\gamma_k^* = \gamma_k$ in (3)) or a mixture of several Gaussian components. π_k^* denote the corresponding prior probabilities.

Here γ_1^* denotes the cluster number that generated the first data point (or any other point according to the i.i.d. assumption, as long as only probabilities are of interest), and $\tilde{\gamma}_1^*$ is the mixture component to which the point is classified by the Bayes rule with true parameters.

$\hat{\pi}_k^*$ can be obtained straightforward by summing up the $\hat{\pi}_m$ of the member components of cluster k. Note that

$$\hat{P}(\tilde{\gamma}_1^* = i,\ \gamma_1^* = j) = \frac{1}{n}\sum_{h=1}^{n} \hat{P}(\gamma_h^* = j|\mathbf{x}_h)1(\hat{\gamma}_h^* = i) \qquad (5)$$

is a consistent estimator of $P(\tilde{\gamma}_1^* = i,\ \gamma_1^* = j)$, where $\hat{\gamma}_h^*$ denotes the data based classification of data point $\mathbf{x}_h$, estimating $\tilde{\gamma}_h^*$, which also defines $\hat{P}(\gamma_h^* = j|x_h)$. $1(\bullet)$ denotes the indicator function.

Therefore,

$$\hat{p}_{ij} = \frac{\hat{P}(\tilde{\gamma}_1^* = i,\ \gamma_1^* = j)}{\hat{\pi}_j^*}$$

is a consistent estimator of p_{ij}. This works regardless of whether the mixture components are Gaussian distributions or mixtures of Gaussians. Here is the method of directly estimated misclassification probabilities (DEMP):

1. Choose a tuning constant $q^* < 1$.
2. Start with all components of the initially estimated Gaussian mixture as current clusters.

3. Compute $q = \max(\hat{p}_{ij}, \hat{p}_{ji})$ for all pairs of current clusters.
4. If $q < q^*$ for all pairs of current clusters, use the current clustering as the final one.
5. Otherwise, merge the pair of current clusters with maximum q and go to step 3.

$q^* = 0.025$ is used here [5].

For the data in Figure 1, DEMP merges components 1 and 2 at $q = 0.078$. For the cluster of these two and component 3, $q = 0.01$, so with $q^* = 0.025$, two clusters are found, namely mixture components 1/2 together and 3. This makes sense if clusters refer to "patterns" in the data but are not required to be separated by gaps.

4 Bootstrap Stability Assessment

In [4], the following idea has been introduced for checking the stability of a cluster:

- Draw B nonparametric bootstrap samples from the dataset (when using with MCLUST, discard the copies of points drawn more than once, so that the bootstrap sample is actually smaller than the original dataset; further schemes to generate datasets are discussed in [4], but they don't deliver very different results for the datasets treated here).
- Cluster the bootstrapped datasets by the same method that was used for the original dataset.
- For every cluster in the original dataset, find the most similar one in every bootstrapped dataset. Similarity is measured according to the Jaccard similarity between sets $C, D : j(C, D) = \frac{|C \cap D|}{|C \cup D|}$.
- For every cluster i, compare the mean Jaccard similarity $\bar{j}_i$ to the most similar cluster from each of the B bootstrap samples.

The Jaccard similarity is between 0 and 1. It makes sense to consider clusters with $\bar{j}_i < 0.5$ as "dissolved" [4] and a meaningful stable cluster should have $\bar{j}_i \gg 0.5$, better above 0.7 or 0.8 (though not every stable cluster is meaningful).

In the given situation it is interesting to apply the idea to the MCLUST clusterings and compare them to the stability achieved by the clusters yielded by the merging methods, i.e., to consider "do MCLUST first and apply ridgeline or DEMP method to the solution" as a clustering method in its own right.

For the dataset in Figure 1 and the MCLUST solution, $\bar{j}_1 = 0.48, \bar{j}_2 = 0.52, \bar{j}_3 = 0.96$, so the first two clusters are obviously unstable. The ridgeline method yields $\bar{j}_{123} = 0.87$ for the only remaining cluster, which means that in some situations it ends up with more than one cluster (for merging methods, the lower index of $\bar{j}$ refers to the original Gaussian components belonging to the merged cluster). DEMP yields $\bar{j}_{12} = 0.94$, $\bar{j}_3 = 0.98$, which confirms that this is a very stable solution.

5 Real Data Example: Clustering Melody Contours

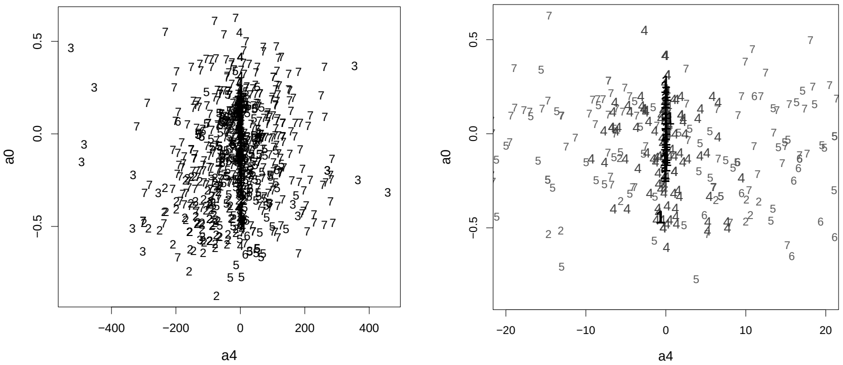

Fig. 3. Variables a4 and a0 of melody data with clusters found by MCLUST (right side: magnified version of the central area; what looks like a "line" along a4≈ 0 is component 1).

The dataset analysed here consists of approximations of the contours of 989 melody phrases taken from commercial pop songs by polynomials of degree 5, as discussed in [3]. The dataset was provided by D. Müllensiefen. Due to a lower intrinsic dimensionality of the dataset, only four coefficients (a4, a3, a1 and a0, indicating the degrees of the terms of the polynomials) were used as variables, and most of the information distinguishing the seven clusters obtained by MCLUST can be seen in the scatterplot of a4 and a0, see Figure 3. Apart from a strong concentration around the value 0 of the first variable (which is represented by two components, no. 1 and 4, in the MCLUST solution, see right side of Figure 3), no clear patterns can be seen.

The stability values for the MCLUST solution are: $\bar{j}_1 = 0.75, \bar{j}_2 = 0.33, \bar{j}_3 = 0.42, \bar{j}_4 = 0.42, \bar{j}_5 = 0.27, \bar{j}_6 = 0.24, \bar{j}_7 = 0.44$, so only the first component is reasonably stable. This in itself is very useful for interpreting the clustering, even before having done any component merging.

Some ridgeline plots are given in Figure 4. Note that the gap between components 2 and 5 on the lower left side is by far the deepest for any pair of components in this dataset, which indicates that the 2-d plots in Figure 3 do not miss any strong separation between any two clusters in 4-d space (particularly taking into account that even in mixtures without clear gaps,

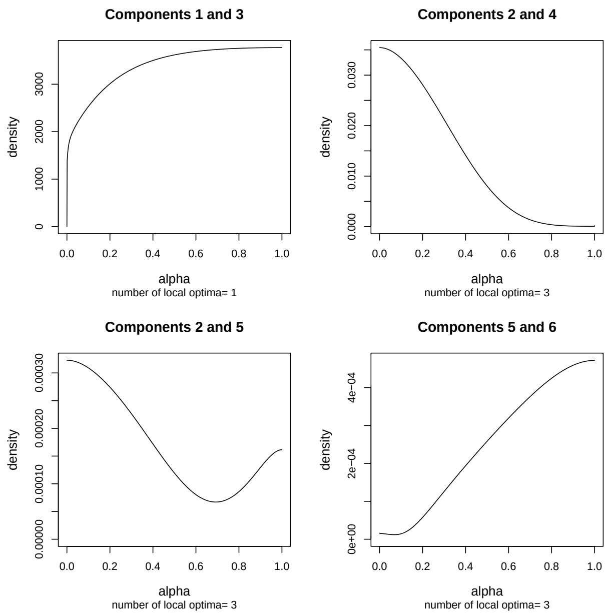

Fig. 4. Selected ridgeline plots for melody data.

gaps are expected to occur in ridgeline ratio plots between “non-neighbouring” components, see Figure 2). Some pairs of components do not yield unimodal mixtures, but one of the modes is usually very weak, as for example for the two component pairs on the right side of Figure 4 (it is hardly visible that the mixture of components 2 and 4 is bimodal but the density goes up a tiny little bit approaching $\alpha = 1$). About half of the pairs of components yield unimodal mixtures such as on the upper left side of Figure 4.

The ridgeline method merges all clusters, justified by the fact that there are no clear gaps. This is very stable ($\bar{\jmath}_{1234567} = 1$). However, demanding estimated unimodality by using $r^* = 1$ in Section 2 merges all components except of component 2 with stability $\bar{\jmath}_2 = 0.12$, indicating again that $r^* = 1$ is not a good idea. DEMP yields three clusters by leaving components 1 and 4 unmerged and merging all the others. These clusters are obviously not separated by gaps, but correspond to visible patterns in Figure 3. The stabilities are $\bar{\jmath}_1 = 0.74, \bar{\jmath}_4 = 0.47, \bar{\jmath}_{23567} = 0.91$. This indicates that it makes sense to distinguish component 1 from the union of components 2, 3, 5, 6, 7 as a stable pattern. Component 4, which lies “between” the other two clusters, cannot be so clearly distinguished from them. This corresponds nicely with the visual impression from Figure 3. In terms of the melodic phrases, there are no groups of phrases that can really be separated from the others by “gaps”, but there is a core pattern of phrases (component 1 and to some extent 4) that can be interpreted as having in common a value of about zero for the fourth degree coefficient (a4) of the contour approximating polynomial, which means that no steep increase/decrease (or decrease/increase) combinations occur in the melody contour.

6 Conclusion

The problem of merging Gaussian mixture components cannot be uniquely solved. Solutions always depend on what kind of clusters the researcher is looking for. For example, clusters can be defined by gaps (rather corresponding to the ridgeline method) or by "patterns" (rather corresponding to the DEMP method). Visualisation of the separation of mixture components and assessment of the stability of clusters can help with the decision whether some of the original mixture components should be merged, and with how the results are to be interpreted. Further methods for merging and visualisation, details, examples and comparisons (including some situations in which DEMP merges stronger than the ridgeline method as opposed to the two examples here) are given in [5].

References

1. C. Fraley, and A. E. Raftery. Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association*, 97:611–631, 2002.
2. C. Fraley, and A. E. Raftery. MCLUST Version 3 for R: Normal Mixture Modeling and Model-Based Clustering. Technical Report no. 504, Department of Statistics, University of Washington, 2006.
3. K. Frieler, D. Müllensiefen, and F. Riedemann. Statistical search for melodic prototypes. In T. Klouche, editor, *Conference on Mathematics in Music*, Staatliches Institut für Musikforschung, Berlin, in press.
4. C. Hennig. Cluster-wise assessment of cluster stability. *Computational Statistics and Data Analysis* 52:258–271, 2007.
5. C. Hennig. Methods for merging Gaussian mixture components. Research Report no. 203, Department of Statistical Science, UCL, 2009. Submitted.
6. J. Li. Clustering based on a multilayer mixture model. *Journal of Computational and Graphical Statistics* 14:547–568, 2004.
7. S. Ray, and B. G. Lindsay. The Topography of Multivariate Normal Mixtures. *Annals of Statistics* 33:2042–2065, 2005.
8. J. Tantrum, A. Murua, and W. Stuetzle. Assessment and Pruning of Hierarchical Model Based Clustering. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, Washington, D.C.*, 197–205, 2003.