COMPSTAT'2004 Symposium

EXPLORING THE STRUCTURE OF MIXTURE MODEL COMPONENTS

Friderich Leisch

Key words: Finite mixture models, regression, cluster analysis, visualization.
COMPSTAT 2004 section: Clustering, Statistical software.

Abstract: Model-based cluster analysis and latent class regression are popular methods for grouping observations into unobserved segments. In many applications it is of great interest to the practitioner to assess the relationships between those segments, especially which segments are close to each other and which are markedly different from the rest. We present several new tools for the R statistical computing environment that allow the user to visually explore the component structure of arbitrary mixture models and do computations using a graph representation of the model.

1 Introduction

Finite mixture models have been used for more than 100 years, but have seen a real boost in popularity over the last decade due to the tremendous increase in available computing power. The areas of application of mixture models range from biology and medicine to physics, economics and marketing. On the one hand these models can be applied to data where observations originate from various groups and the group affiliations are not known, and on the other hand to provide approximations for multi-modal distributions [4], [12], [9].

In the 1990s finite mixture models have been extended by mixing standard linear regression models as well as generalized linear models [14]. An important area of application of mixture models and also of these extensions are in market segmentation [15], where finite mixture models replace more traditional cluster analysis and cluster-wise regression techniques as state of the art. Finite mixture models with a fixed number of components are usually estimated with the EM algorithm within a maximum likelihood framework [2] and with MCMC sampling [3] within a Bayesian framework.

The R environment for statistical computing [10] features several packages for finite mixture models, including `mclust` for mixtures of multivariate Gaussian distributions [6, 5], `fpc` for mixtures of linear regression models [7] and `mmlcr` for mixed-mode latent class regression [1]. All of those primarily target one or more special cases of mixture models. Package `flexmix` implements an extensible framework for mixture modelling where users can easily create new models by supplying their own M-step for the EM algorithm [8].

Efficient estimation of mixture models has received a lot of attention over the last years, however model diagnostics and general visualization techniques are scarcely available. E.g., the confidence ellipses commonly used to visualize low-dimensional Gaussians cannot be used for regression models. In this

paper we present several new tools implemented in `flexmix` that can be used to graphically explore the structure of the components of any finite mixture model. Of special interest in all applications where mixtures are used to group observations is which components are overlapping or "close" to each other. If the mixture model is used for market segmentation it is important to know for the practitioner which clusters are distinct market niches and which clusters are parts of larger consumer groups.

2 The posterior class probabilities

Consider finite mixture models with K components of form

$$h(y|x,w) = \sum_{k=1}^{K} \pi_k f(y|x,\theta_k) \tag{1}$$

$$\pi_k \geq 0, \quad \sum_{k=1}^{K} \pi_k = 1$$

where y is a (possibly multivariate) dependent variable with conditional density h, x is a vector of independent variables, π_k is the prior probability of component k, and θ_k is the component specific parameter vector for the density function f.

If f is a normal density with component-specific mean $\beta_k' x$ and variance σ_k^2, we have $\theta_k = (\beta_k', \sigma_k^2)'$ and Equation (1) describes a mixture of standard linear regression models, also called latent class regression. A special case is $x \equiv 1$, which gives a mixture of Gaussians without a regression part. If f is a member of the exponential family, we get a mixture of generalized linear models (GLMs).

The posterior probability that observation (x, y) belongs to class j is given by

$$\P(j|x,y) = \frac{\pi_j f_j(y|x,\theta_j)}{\sum_k \pi_k f_k(y|x,\theta_k)}$$

Histograms or rootograms of the posterior class probabilities can be used to assess the cluster structure [11], this is now the default plot method for `"flexmix"` objects. Rootograms are very similar to histograms, the only difference is that the height of the bars correspond to square roots of counts rather than the counts themselves, hence low counts are more visible and peaks less emphasized.

Usually in each component a lot of observations have posteriors close to zero, resulting in a high count for the corresponing bin in the rootogram which obscures the information in the other bins. To avoid this problem, all probabilities with a posterior below a threshold are ignored (we use 0.0001). A peak at probability 1 indicates that a mixture component is well seperated from the other components, while no peak at 1 and/or significant mass in the middle of the unit interval indicates overlap with other components.

As example we use a 2-component mixture of Poisson regression models with one independent variable, parameters $\theta_1 = (2, -0.2)'$ and $\theta_2 = (1, 0.1)'$, and the exponential link function. Hence, given x the response y in group k has a Poisson distribution with mean $\exp((1, x) \cdot \theta_k)$. A sample with 100 observations in each group is shown in Figure 1. For data stored in an R data frame `mydata` the mixture model can be estimated using the commands

```
R> model1 = flexmix(y ~ x, data = mydata, k = 2,
+     model = FLXglm(family = "poisson"))

Classification: weighted
  10 Log-likelihood:    -458.3680
  20 Log-likelihood:    -458.1333
  24 Log-likelihood:    -458.1307
converged
```

The estimated parameters are

```
     (Intercept)      x
[1,]       1.922 -0.181
[2,]       0.997  0.106
```

which is close to the true parameters. The corresponding clusters can be seen in the right panel of Figure 1.

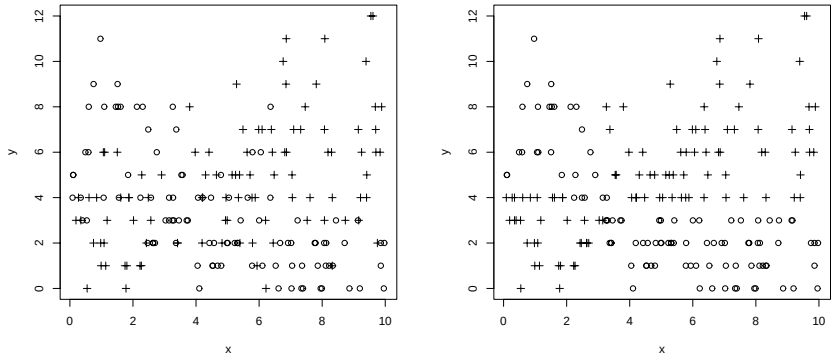

Figure 1: Poisson regression mixture with 2 components: true groups (left) and groups found by `model1` (right).

Issuing the command `plot(model1)` gives the rootograms shown in the left panel of Figure 2. The obvious overlap between the clusters is easily identified, the posteriors have almost a uniform distribution over the interval $[0, 1]$.

Now assume that instead of 200 independent observations we have 2 measurements each from 100 persons and that column `id` of `mydata` contains a factor identifying the 100 persons. If we use the additional information the EM algorithm needs only half the number of iterations to converge:

```
R> model2 = flexmix(y ~ x | id, data = mydata, k = 2,
+     model = FLXglm(family = "poisson"))

Classification: weighted
  10 Log-likelihood:    -889.0594
  13 Log-likelihood:    -889.0556
converged
```

The `model2` parameter estimates

```
     (Intercept)      x
[1,]        1.96 -0.201
[2,]        1.04  0.101
```

are only slightly better than for `model1`, but now we can assign the observations with more confidence into the two classes as the posteriors are shifted towards 0 and 1 (middle panel of Figure 2). If we have 4 repeated measurements from 50 persons, this effect is of course even much more pronounced (right panel of Figure 2) and there are only very few observations with posteriors close to 0.5.

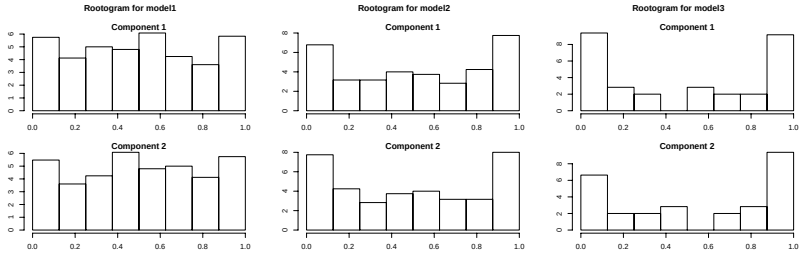

Figure 2: Rootograms for models with no repeated measurements (left), 2 (middle) and 4 (right) measurements per person.

3 Kullbach-Leibler divergence between component

Histograms or rootograms of posteriors visualize with how much confidence observations are assigned to clusters, but can not help to identify relationships between clusters in case of more than 2 components. Consider the smiley data from R package `mlbench` shown in Figure 3. Although only the "eyes" are really Gaussian, we can use model-based clustering with Gaussians to approximate the multimodal density (similar to a density estimate using a Gaussian kernel).

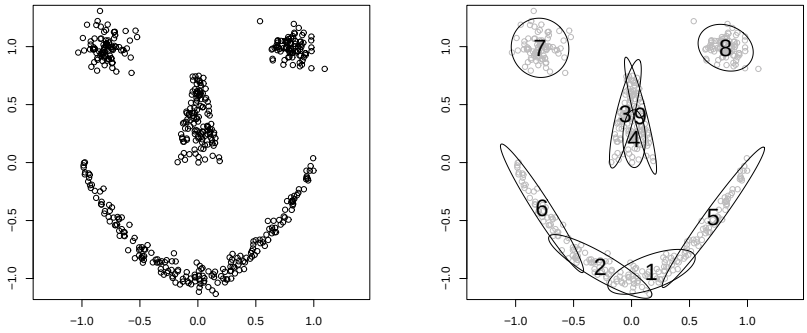

Figure 3: The smiley data (right) and a 9 component partition.

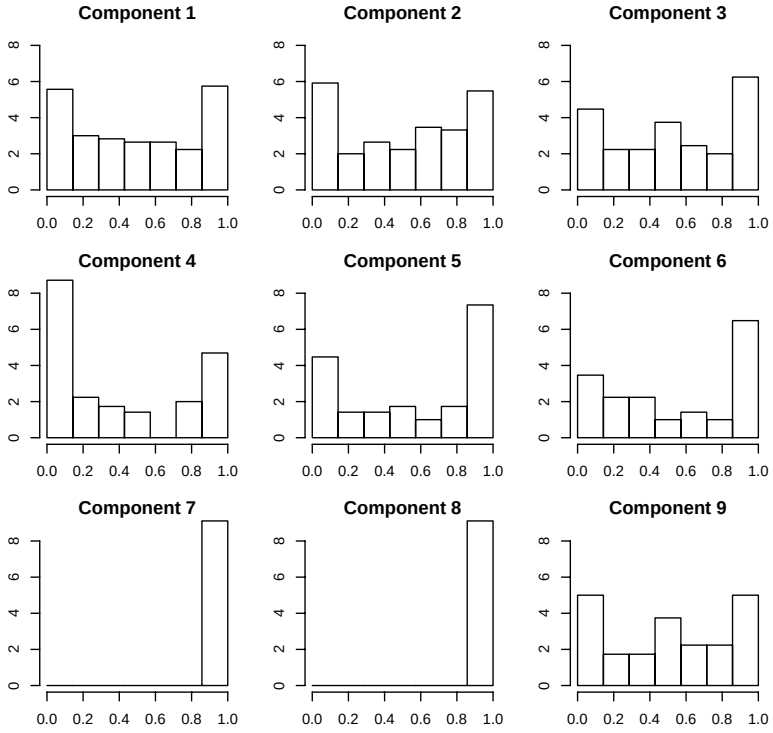

Figure 4: Rootograms of the 9 components for the smiley data.

The corrsponding rootograms are shown in Figure 4. Only components 7 and 8 (the "eyes") do not overlap with any other cluster, all others have a lot of posteriors much smaller than 1. One possibility to explore which pairs of clusters overlap would be to use the brushing facilities of interactive histograms as provided by the R package `iplots` [13]. Another way is to compute pairwise distances between the clusters. The most common distance measure for two distributions with densities f and g is the Kullbach-Leibler (KL) divergence

$$KL(f, g) = \int f(x) \left(\log f(x) - \log g(x)\right) \, dx$$

(for discrete distributions the integral is replaced by a sum), which cannot be solved analytically in many cases. However, the KL divergence between mixture components k and l can be estimated as

$$\begin{aligned} KL(f_k, f_l) &\approx \sum_{n=1}^{N} p_{nk} \left(\log p_{nk} - \log p_{nl}\right) \\ p_{nk} &= \pi_k f(y_n | x_n, \theta_k) \end{aligned}$$

Evaluating the sum is numerically problematic because most p_{nk} are almost zero. To get a stable estimate we remove all terms in the sum involving densities below a threshold of $\epsilon = 0.01$, which results in the KL divergence matrix

	1	2	3	4	5	6	7	8	9
1	0	14	.	.	26	.	.	.	.
2	28	0	.	.	.	18	.	.	.
3	.	.	0	139	.	.	.	.	41
4	.	.	20	0	.	.	.	.	15
5	18	.	.	.	0	.	.	.	.
6	.	19	.	.	.	0	.	.	.
7	.	.	.	.	.	.	0	.	.
8	.	.	.	.	.	.	.	0	.
9	.	.	18	109	.	.	.	.	0

for the smiley data (rounded to integers). Dot entries correspond to components where the regions with densities larger than ϵ do not overlap.

The KL divergence matrix can be represented by a directed graph with one node for each component as shown in Figure 5. Overlapping clusters correspond to connected nodes and modes of the mixture density to cliques of the graph. For our 2-dimensional example data without covariates a natural positioning of the graph nodes are the centers of the clusters. For higher-dimensional data or regression mixtures this is not possible and we have to restrict ourselves to general graph layout algorithms. Sammon mapping of the KL divergences results in the right panel of Figure 5, where especially the linear structure of the "mouth" is preserved correctly. Of course all unconnected components of the graph are placed randomly with respect to each other (and could as well be projected seperately).

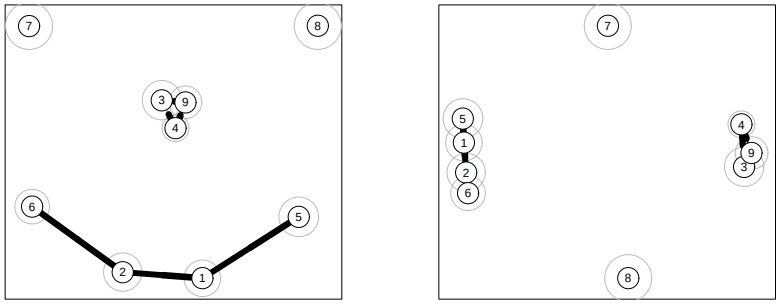

Figure 5: Graph corresponding to the KL divergences: node positions according to cluster centers (left) and Sammon mapping (right). The circles around the nodes are proportional to the cluster sizes.

4 Summary

Finite mixture models have become increasingly popular in many domains of applications, yet diagnostic tools for fitted models (and especially corresponding software) are much less developed. We have developed several new tools available in the R package `flexmix` which allow the user to explore the relationships between components of fitted mixture models.

All methods presented in this paper work off the densities or posterior probabilities of the observations and thus do not depend on the dimensionality of the input space. While we have used simple 2-dimensional examples to demonstrate the techniques, they can easily be used on high-dimensional data sets or models with complicated covariate structures.

As a next step we will integrate the graph representation of the model into more interactive visualization systems such that the user can easily explore the distribution of background variables. E.g., if each mixture component corresponds to a market segment, clicking on a node in the graph could show the distribution of sales and sociodemographic data of the consumers in the respective segment.

References

[1] Buyske S. (2003). *R Package `mmlcr`: mixed-mode latent class regression.* version 1.3.2.

[2] Dempster A., Laird N., Rubin D. (1977). *Maximum likelihood from incomplete data via the EM-alogrithm.* Journal of the Royal Statistical Society, B **39**, 1 – 38.

[3] Diebolt J., Robert C.P. (1994). *Estimation of finite mixture distributions through Bayesian sampling.* Journal of the Royal Statistical Society, Series B **56**, 363 – 375.

[4] Everitt B.S., Hand D.J. (1981). *Finite mixture distributions.* London: Chapman and Hall.

[5] Fraley C., Raftery A.E. (2002). *MCLUST: Software for model-based clustering, discriminant analysis and density estimation.* Technical Report 415, Department of Statistics, University of Washington, Seattle, WA, USA.

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

[7] Hennig C. (2000). *Identifiability of models for clusterwise linear regression.* Journal of Classification **17**, 273 – 296.

[8] Leisch F. (2003). FlexMix: *A general framework for finite mixture models and latent class regression in R.* Report 86, SFB "Adaptive Information Systems and Modeling in Economics and Management Science".

[9] McLachlan G., Peel D. (2000). *Finite mixture models.* John Wiley and Sons Inc.

[10] R Development Core Team. (2003). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria.

[11] Tantrum J., Murua A., 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, 197 – 205. ACM Press, New York, NY, USA.

[12] Titterington D., Smith A., Makov U. (1985). *Statistical analysis of finite mixture distributions.* Chichester: Wiley.

[13] Urbanek S., Theus M. (2003). *iPlots — high interaction graphics for R.* In Hornik K., Leisch F., Zeileis A., (eds), Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vienna, Austria.

[14] Wedel M., DeSarbo W.S. (1995). newblock *A mixture likelihood approach for generalized linear models.* Journal of Classification **12**, 21 – 55.

[15] Wedel M., Kamakura W.A. (2001). *Market segmentation - conceptual and methodological foundations.* Kluwer Academic Publishers, Boston, MA, USA, 2 edition.

Address: F. Leisch, Department of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstraße 8-10, 1040 Wien, Austria

E-mail: `Friedrich.Leisch@ci.tuwien.ac.at`