Bibliography

2005).

- I. E. Frank and J. H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–148, 1993. 12
- K. Jørgensen, B.-H. Mevik, and T. Næs. Combining designed experiments with several blocks of spectroscopic data. Submitted, 2005. 17
- J. H. Kalivas. Two data sets of near infrared spectra. *Chemometrics and Intelligent Laboratory Systems*, 37: 255–259, 1997. 13
- P. A. Lachenbruch and M. R. Mickey. Estimation of error rates in discriminant analysis. *Technometrics*, 10(1):1–11, 1968. 13
- H. Martens and T. Næs. *Multivariate Calibration*. Wiley, Chichester, 1989. 12

- B.-H. Mevik and H. R. Cederkvist. Mean squared error of prediction (MSEP) estimates for principal component regression (PCR) and partial least squares regression (PLSR). *Journal of Chemometrics*, 18(9):422–429, 2004. 13
- T. Næs and H. Martens. Principal component regression in NIR analysis: Viewpoints, background details and selection of components. *Journal of Chemometrics*, 2:155–167, 1988. 12
- M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B—Methodological,* 36:111–147, 1974. 13
- S. Wold, M. Sjöström, and L. Eriksson. PLSregression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2):109– 130, 2001. 12

Bjørn-Helge Mevik IKBM, Norwegian University of Life Sciences, Ås, Norway. bjorn-helge.mevik@umb.no

Some Applications of Model-Based Clustering in Chemistry

by Chris Fraley and Adrian E. Raftery

Interest in clustering has experienced a recent surge due to the emergence of new areas of application. Prominent among these is the analysis of images resulting from new technologies involving chemical processes, such as microarray or proteomics data, and contrast-enhanced medical imaging. Clustering is applied to the image data to produce segmentations that are appropriately interpretable. Other applications include minefield detection (Dasgupta and Raftery 1998; Stanford and Raftery 2000), finding flaws in textiles (Campbell et al. 1997; 1999), grouping coexpressed genes (Yeung et al. 2001), *in vivo* MRI of patients with brain tumors (Wehrens et al. 2002), and statistical process control (Thissen et al. 2005).

The use of clustering methods based on probability models rather than heuristic procedures is becoming increasingly common due to recent advances in methods and software for model-based clustering, and the fact that the results are more easily intepretable. Finite mixture models (McLachlan and Peel, 2000), in which each component probability corresponds to a cluster, provide a principled statistical approach to clustering. Models that differ in the number of components and/or component distributions can be compared using statistical criteria. The clustering process estimates a model for the data that allows for overlapping clusters, as well as a probabilistic clustering that quantifies the uncertainty of observations belonging to components of the mixture.

The R package mclust (Fraley and Raftery 1999, 2003) implements clustering based on normal mixture models. The main clustering functionality is provided by the function EMclust, together with its associated summary and plot methods. Users can specify various parameterizations of the variance or covariance of the normal mixture model, including spherical and diagonal models in the multivariate case, along with the desired numbers of mixture components to consider. The mixture parameters are estimated via the EM algorithm (Dempster et al. 1977; McLachlan and Krishnan 1997), initialized by model-based hierarchical clustering (Banfield and Raftery 1993; Fraley 1998). The best model is selected according to the Bayesian Information Criterion or BIC (Schwarz 1978), a criterion that adds a penalty to the loglikelihood that increases with the number of parameters in the model.

In this article, we discuss an application of model-

17

based clustering to diabetes diagnosis from glucose and insulin levels in blood plasma. We also discuss two applications in image segmentation. In the first, model-based clustering is used to give an initial segmentation of microarray images for signal extraction. In the second, model-based clustering is used to segment a dynamic breast MR image to reveal possible tumors.

Model-based Clustering

In model-based clustering, the data *x* are viewed as coming from a mixture density $f(x) = \sum_{k=1}^{G} \tau_k f_k(x)$, where f_k is the probability density function of the observations in group *k*, and τ_k is the probability that an observation comes from the *k*th mixture component $(0 < \tau_k < 1 \text{ for all } k = 1, ..., G \text{ and } \sum_k \tau_k = 1)$.

Each component is usually modeled by the normal or Gaussian distribution. In the univariate case, component distributions are characterized by the mean μ_k and the variance σ_k^2 , and have the probability density function

$$\phi(x_i;\mu_k,\sigma_k^2) = \frac{1}{\sqrt{(2\pi\sigma_k^2)}} \exp\left\{-\frac{(x_i-\mu_k)^2}{2\sigma_k^2}\right\}.$$
 (1)

In the multivariate case, component distributions are characterized by the mean μ_k and the covariance matrix Σ_k , and have the probability density function

$$\phi(x_i; \mu_k, \Sigma_k) = \frac{\exp\{-\frac{1}{2}(x_i - \mu_k)^T \Sigma_k^{-1}(x_i - \mu_k)\}}{\sqrt{\det(2\pi\Sigma_k)}}.$$
(2)

The likelihood for data consisting of n observations assuming a Gaussian mixture model with G multivariate mixture components is

$$\prod_{i=1}^{n}\sum_{k=1}^{G}\tau_{k}\phi(x_{i};\mu_{k},\Sigma_{k}).$$
(3)

For reviews of model-based clustering, see McLachlan and Peel (2000) and Fraley and Raftery (2002).

For a fixed number of components *G*, the model parameters τ_k , μ_k , and Σ_k can be estimated using the EM algorithm initialized by hierarchical modelbased clustering (Dasgupta and Raftery 1998; Fraley and Raftery 1998). Data generated by mixtures of multivariate normal densities are characterized by groups or clusters centered at the means μ_k , with increased density for points nearer the mean. The corresponding surfaces of constant density are ellipsoidal.

Geometric features (shape, volume, orientation) of the clusters are determined by the covariances Σ_k , which may also be parametrized to impose crosscluster constraints. There are a number of possible parameterizations of Σ_k , many of which are implemented in **mclust**. Common instances include Σ_k = λI , where all clusters are spherical and of the same size; $\Sigma_k = \Sigma$ constant across clusters, where all clusters have the same geometry but need not be spherical; and unrestricted Σ_k , where each cluster may have a different geometry.

Banfield and Raftery (1993) proposed a general framework for geometric cross-cluster constraints in multivariate normal mixtures by parametrizing covariance matrices through eigenvalue decomposition in the following form:

$$\Sigma_k = \lambda_k D_k A_k D_k^T, \tag{4}$$

where D_k is the orthogonal matrix of eigenvectors, A_k is a diagonal matrix whose elements are proportional to the eigenvalues, and λ_k is an associated constant of proportionality. Their idea was to treat λ_k , A_k and D_k as independent sets of parameters, and either constrain them to be the same for each cluster or allow them to vary among clusters. When parameters are fixed, clusters will share certain geometric properties: D_k governs the orientation of the *k*th component of the mixture, A_k its shape, and λ_k its volume, which is proportional to $\lambda_k^d \det(A_k)$. The model options available in **mclust** are summarized in Table 1.

A 'best' model for the data can be estimated by fitting models with differing parameterizations and/or numbers of clusters to the data by maximum likelihood, and then applying a statistical criterion for model selection. The Bayesian Information Criterion or BIC (Schwarz 1978) is the model selection criterion provided in the **mclust** software; the 'best' model is taken to be the one with the highest BIC value.

Example 1: Diabetes Diagnosis from Glucose and Insulin Levels

We first illustrate the use of **mclust** on the diabetes dataset (Reaven and Miller 1979) giving three measurements for each of 145 subjects:

glucose	 plasma glucose respons 			
		to oral glucose		
insulin	-	plasma insullin response		
		to oral glucose		
sspg	-	steady-state plasma glucose		
		(measures insulin resistance)		

This dataset is included in the **mclust** package. The subjects were clinically diagnosed into three groups: normal, chemically diabetic, and overtly diabetic. The diagnosis is given in the first column of the diabetes dataset, which is excluded from the cluster analysis.

The following code computes the BIC curves using the function EMclust and then plots them (see Figure 1, upper left):

```
> data(diabetes)
```

```
> diBIC <- EMclust(diabetes[,-1])</pre>
```

```
> plot(diBIC)
```

identifier	Model	# covariance parameters	Distribution	Volume	Shape	Orientation
EII	λI	1	Spherical	=	=	NA
VII	$\lambda_k I$	G	Spherical		=	NA
EEI	λA	d	Diagonal	=	=	axes
VEI	$\lambda_k A$	G + (d-1)	Diagonal		=	axes
EVI	λA_k	1 + G(d-1)	Diagonal	=		axes
VVI	$\lambda_k A_k$	Gd	Diagonal			axes
EEE	λDAD^T	d(d+1)/2	Ellipsoidal	=	=	=
EEV	$\lambda D_k A D_k^T$	1 + (d-1) + G[d(d-1)/2]	Ellipsoidal	=	=	
VEV	$\lambda_k D_k A D_k^T$	G + (d-1) + G[d(d-1)/2]	Ellipsoidal		=	
VVV	$\lambda_k D_k A_k D_k^T$	G[d(d+1)/2]	Ellipsoidal			

Table 1: Parameterizations of the multivariate Gaussian mixture model available in **mclust**. In the column labeled '# covariance parameters', *d* denotes the dimension of the data, and *G* denotes the number of mixture components. The total number of parameters for each model can be obtained by adding *Gd* parameters for the means and G - 1 parameters for the mixing proportions.

```
EII VII EEI VEI EVI VVI EEE EEV VEV VVV
"A" "B" "C" "D" "E" "F" "G" "H" "I" "J"
```

The model parameters can then be extracted via the summary function, and results can be plotted using the function coordProj as follows:

The summary object diS contains the parameters and classification for the best (highest BIC) model. The function coordProj can be used to plot the data and **mclust** classification, marking the means and drawing ellipses (with axes) corresponding to the variance for each group (see Figure 1, lower left).

For this data, model-based clustering chooses a model with three components, each having a different covariance. Moreover, the corresponding three-group classification matches the three clinically diagnosed groups with 88% accuracy.

The uncertainty of a classification can also be assessed in model-based clustering. The function uncerPlot can be used to display the uncertainty of misclassified objects when there is a known classification for comparison. More generally, the function coordProj can be used to display the relative uncertainty of a classification:

The resulting plots are shown in Figure 1, upper right and lower right. In this case, the misclassified data points tend to be among the most uncertain.

Example 2: Microarray Image Segmentation

Microarray technology is now a widely-used tool in a number of large-scale assays. While many array platforms exist, a common method for making DNA arrays consists of printing the single-stranded DNA representing the genes on a solid substrate using a robotic spotting device. In the two-color array, the cDNA extracted from the experimental and control samples are first labelled using the Cy3 (green) and Cy5 (red) fluorescent dyes. Then they are mixed and hybridized with the arrayed DNA spots. After hybridization, the arrays are scanned at the corresponding wavelengths separately to obtain the images corresponding to the two channels. The fluorescence measurements are used to determine the relative abundance of the mRNA or DNA in the samples.

The quantification of the amount of fluorescence from the hybridized sample can be affected by a variety of defects that occur during both the manufacturing and processing of the arrays, such as perturbations of spot positions, irregular spot shapes, holes in spots, unequal distribution of DNA probe within spots, variable background, and artifacts such as dust and precipitates. Ideally these events should be automatically recognized in the image analysis, and the estimated intensities adjusted to take account of them.

Li et al. (2005) proposed a robust model-based method for processing microarray images so as to estimate foreground and background intensities. It starts with an automatic gridding algorithm that uses a sliding window to find the peaks and valleys. Then model-based clustering is applied to the (univariate) sum of the intensities of the two channels measuring the red and green signals to provide an initial segmentation. Based on known information about the data, it is assumed there can be no more than three groups in the model (background, fore-



Figure 1: Upper left: BIC computed by EMclust for the 10 available model parameterizations and up to 9 clusters for the diabetes dataset. Different letters encode different model parameterizations, as output from the plot method. The 'best' model is taken to be the one with the highest BIC among the fitted models. Lower left: A projection of the diabetes data, with different symbols indicating the classification corresponding to the best model as computed by EMclust. The component means are marked and ellipses with axes are drawn corresponding to their covariances. In this case there are three components, each with a different covariance. Upper right: Uncertainty of the classification of each observation in the best model. Observations are ordered by increasing uncertainty along the horizontal axis. Vertical lines indicate misclassified observations, which in this case tend to be among the most uncertain. Lower right: A projection of the diabetes data showing classification uncertainty. Larger symbols indicate the more uncertain observations.

ground, uncertain). If there is more than one group, connected components below a certain threshold in size are removed (designated as unknown) from the brightest group as a precaution against artifacts. The procedure is depicted in Figure 2.

An implementation is available in Bioconductor (see http://www.bioconductor.org). The package is called **spotSegmentation**, and consists of two basic functions:

- spotgrid: determines spot locations in blocks
 within microarray slides
- spotseg: determines foreground and background
 signals within individual spots

- 1. Automatic gridding.
- 2. Model-based clustering for \leq 3 groups.
- 3. Threshold connected components.
- 4. Foreground / background determination:
 - If there is more than one group, the foreground is taken to be the group of highest mean intensity and the background the group of lowest mean intensity.
 - If there is only one group, it is assumed that no foreground signal is detected.

Figure 2: Basic Procedure for Model-based Segmentation of Microarray Blocks.

These functions will be illustrated on the spotSegTest dataset supplied with the package, which consists of a portion of the first block from the first microarray slide image from van't Wout et al. (2003). This data set is a data frame, with two columns, one from each of the two channels of absorption intensities. The spotSegTest dataset can be obtained via the data command once the **spotSegmentation** package is installed.

```
> data(spotSegTest)
```

Because the data are encoded for compact storage, they need to be transformed as follows in order to extract the intensities:

Note that this transformation is specific to this data; in general stored image data must be converted as needed to image intensities. The function spotgrid can be used to divide the microarray image block into a grid separating the individual spots.

Here we have used the knowlege that there are 4 rows and 6 columns in this subset of spots from the microarray image. The show option allows display of the gridded image.

The individual spots can now be segmented using the function spotseg, which does model-based clustering for up to 3 groups via **mclust** followed by a connected component analysis. The following segments all spots in the block:

The corresponding plot is shown in Figure 3.

Example 3: Dynamic MRI Segmentation

Dynamic contrast-enhanced magnetic resonance imaging (MRI) is emerging as a powerful tool for the diagnosis of breast abnormalities (e.g. Hylton 2005). Because of the high reactivity of breast carcinomas after gadolinium injection, this technology has the potential to allow differentiation between malignant and benign tissues. Its unique ability to provide morphological and functional information can be used to assist in the differential diagnosis of lesions that other methods find questionable. It is currently used as a complementary diagnostic modality in breast imaging. However, data acquistion, postprocessing, image analysis and interpretation of dynamic breast MRI are still active areas of research. Forbes et al. (2006) developed a region of interest (ROI) selection method that combines model-based clustering with Bayesian morphology (Forbes and Raftery 1999), to produce a classification of the data for potential use in diagnosis.

Each dynamic MR image consists of 25 sequential images recording signal intensity after gadolinium injection. Instead of working directly with the image data, they are summarized in terms of five derived variables considered to be of significance in cancer diagnosis:

- **Time to peak:** the time at which the signal peaks.
- **Difference at peak:** absolute increase of intensity between the beginning of the signal and the time at which the signal peaks.
- Enhancement slope: in units of intensity/time.
- **Maximum step:** maximum change between two adjacent dynamic samples.
- Washout slope: in units of intensity/time.

Model-based classification for up to four groups is then applied to this data to segment the image. The choice of four groups is based on knowledge about the data: the main distinguishable components in breast tissue are blood vessels, air, fat, and possibly lesions or tumors. Figure 4 gives an example of model-based clustering applied to multivariate data derived as decribed above from dynamic contrastenhanced breast MRI. Further steps using Bayesian morphology may then be applied to smooth the resulting image.

Summary

The contributed R package **mclust** implements parameter estimation for normal mixture models with and without constraints, with higher-level functions for model-based clustering and discriminant analysis. It includes functions for displaying the fitted models and clustered data.

The Bioconductor package **spotSegmentation** uses **mclust** to determine foreground and background of spots in microarray images.



Figure 3: Left: The sum of channel signals from a portion of a microarray block containing HIV data, with the grid produced by spotgrid superimposed. Right: the corresponding segmented spots produced by spotseg, based on the grid produced by spotgrid. The color scheme is as follows: *black* denotes the spots, *yellow* denotes background, *gray* denotes pixels of uncertain classification.





Figure 4: Reference image (left) and four-class **mclust** classification (right). The tumor area is shown in red, with the colors assigned automatically according to the size of the mean difference at peak for pixels within each cluster.

Model-based clustering can be used successfully in a variety of technologies involving chemical processes, including image segmentation for cDNA microarrays and dynamic contrast-enhanced MR.

Bibliography

- J. D. Banfield and A. E. Raftery. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49: 803–821, 1993.
- J. G. Campbell, C. Fraley, F. Murtagh, and A. E. Raftery. Linear flaw detection in woven textiles using model-based clustering. *Pattern Recognition Letters* 18, 1539–1548, 1997.
- J. G. Campbell, C. Fraley, D. Stanford, F. Murtagh, and A. E. Raftery. Model-based methods for realtime textile fault detection. *International Journal of Imaging Systems and Technology* 10, 339–346, 1999.
- A. Dasgupta and A. E. Raftery. Detecting features in spatial point processes with clutter via modelbased clustering. *Journal of the American Statistical Association*, 93:294–302, 1998.
- A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood for incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.
- F. Forbes, N. Peyrard, C. Fraley, D. Georgian-Smith, D. Goldhaber, and A. Raftery. Model-based regionof-interest selection in dynamic breast MRI. *Journal* of Computer Assisted Tomography, 30:576-687, 2006.
- F. Forbes and A. E. Raftery. Bayesian morphology: Fast unsupervised Bayesian image analysis. *Journal of the American Statistical Association*, 94:555– 568, 1999.
- C. Fraley. Algorithms for model-based Gaussian hierarchical clustering. *SIAM Journal on Scientific Computing*, 20:270–281, 1998.
- C. Fraley and A. E. Raftery. How many clusters? Which clustering method? - Answers via modelbased cluster analysis. *The Computer Journal*, 41: 578–588, 1998.
- C. Fraley and A. E. Raftery. MCLUST: Software for model-based cluster analysis. *Journal of Classification*, 16:297–306, 1999.
- C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association*, 97:611–631, 2002.
- C. Fraley and A. E. Raftery. Enhanced software for model-based clustering, density estimation, and discriminant analysis: MCLUST. *Journal of Classification*, 20:263–286, 2003.

- N. Hylton. Magnetic resonance imaging of the breast: Opportunities to improve breast cancer management. *Journal of Clinical Oncology*, 23(8): 1678–1684, March 10 2005.
- Q. Li, C. Fraley, R. E. Bumgarner, K. Y. Yeung, and A. E. Raftery. Donuts, scratches, and blanks: Robust model-based segmentation of microarray images. *Bioinformatics*, 21:2875–2882, 2005.
- G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley, 1997.
- G. J. McLachlan and D. Peel. *Finite Mixture Models*. Wiley, 2000.
- G. M. Reaven and R. G. Miller. An attempt to define the nature of chemical diabetes using a multidimensional analysis. *Diabetologia*, 16:17–24, 1979.
- G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.
- D. Stanford and A. E. Raftery. Principal curve clustering with noise. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22, 601–609, 2000.
- U. Thissen, H. Swierenga, A. P. de Weijer, R. Wehrens, W. J. Melssen and L. M. .C. Buydens, Multivariate statistical process control using mixture modeling. *Journal of Chemometrics*, 19:23-31, 2005.
- R. Wehrens and A. W. Simonetti and L. M. .C. Buydens, Mixture modeling of medical magnetic resonance data. *Journal of Chemometrics*, 16:274-282, 2002.
- A. B. van't Wout, G. K. Lehrman, S. A. Mikeeva, G. C. O'Keefe, M. G. Katze, R. E. Bumgarner, G. K. Geiss, and J. I. Mullins. Cellular gene expression upon human immunodeficiency type 1 infection of CD4(+)-T-cell lines. *Journal of Virology*, 77(2):1392– 1402, January 2003.
- K. Y. Yeung, C. Fraley, A. Murua, A. E. Raftery, and W. L. Ruzzo. Model-based clustering and data transformation for gene expression data. *Bioinformatics* 17, 977–987, 2001.

Chris Fraley Department of Statistics University of Washington fraley@stat.washington.edu

Adrian E. Raftery Department of Statistics University of Washington raftery@stat.washington.edu