nag®

Gaussian mixture model

Modelling data drawn from an unknown statistical distribution with a weighted sum of distributions defines a finite mixture model, also known as a latent class method. The most common example incorporates a given number, say k, of Gaussian (i.e., Normal) distributions to model data. Mixture models can be applied to a wide range of applications to grouped data, such as density estimation and clustering.

The routine G03GAF is new to Mark 24 of the NAG Fortran Library and fits a Gaussian k-mixture model with the following (co)variance structures:

a. Group-wise covariances: the data are assumed to be drawn from k Normal distributions with different means and covariances.

b. Pooled covariances: the data are assumed to be drawn from k Normal distributions with different means but equal covariances.

c. Group-wise variances: the data are assumed to be drawn from k Normal distributions with different means and variances.

d. Pooled variances: the data are assumed to be drawn from k Normal distributions with different means and equal variances.

e. Overall variance: the data are assumed to be drawn from k Normal distributions with different means and equal (single) variance.

Where option (a) gives the most flexible model as it requires the most parameters, and the flexibility of models reduces progressing towards (e).

Given a data set assumed to arise from a known number of mixtures of Gaussians, G03GAF estimates the free parameters of a mixture model: the means and (co)variances. Examples include modelling buyer behaviour by identifying different customer groups or the distribution of power loads over a network.

For the purpose of this example, we simulate data from a population consisting of two bivariate Normal distributions and compare G03GAF's parameter estimates with the known values. If the parameter estimates are "close" to their true values, the model reflects the true nature of the population. These distributions are defined by:

$$\textit{Group } 1.\ \text{Mean} \begin{matrix}\textit{Variable } 1\\ \textit{Variable } 2\end{matrix}\begin{pmatrix}1\\2\end{pmatrix} \text{ and covariance } \begin{matrix}\\ \textit{Variable } 1\\ \textit{Variable } 2\end{matrix}\begin{matrix}\textit{Variable } 1 & \textit{Variable } 2\\ \end{matrix}\begin{pmatrix}1.1 & 0.2\\ 0.2 & 1.3\end{pmatrix};$$

$$\textit{Group } 2.\ \text{Mean} \begin{matrix}\textit{Variable } 1\\ \textit{Variable } 2\end{matrix}\begin{pmatrix}-1\\1\end{pmatrix} \text{ and covariance } \begin{matrix}\\ \textit{Variable } 1\\ \textit{Variable } 2\end{matrix}\begin{matrix}\textit{Variable } 1 & \textit{Variable } 2\\ \end{matrix}\begin{pmatrix}0.6 & 0.4\\ 0.4 & 0.3\end{pmatrix}.$$

Figure 1 shows that *Group* 1 is drawn from an almost circular distribution and *Group* 2 from a rotated ellipse, with an overlap between the groups.

Fitting $k = 2$ Gaussians to this data using G03GAF's mixture model, we obtain the following results:

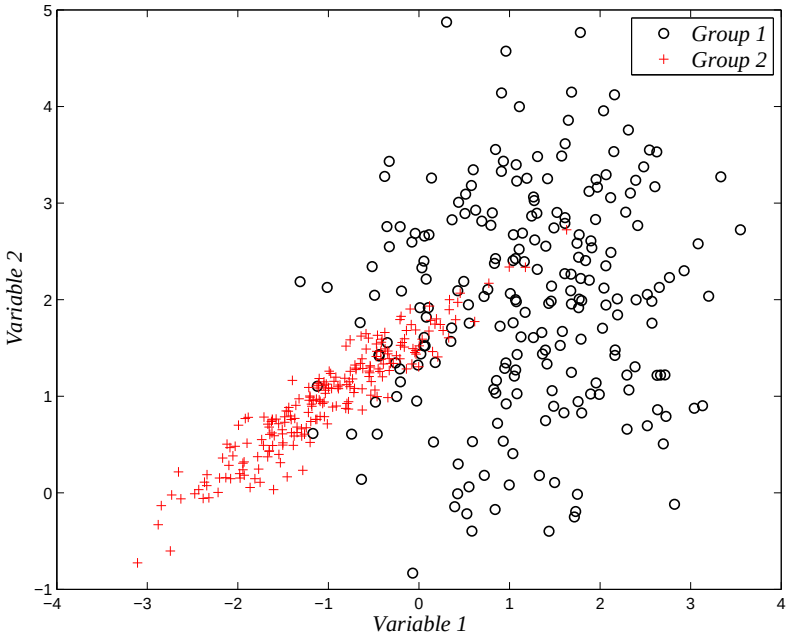

Figure 1: Scatter plot of 222 variates from each bivariate population.

- Group means:
$$\begin{array}{c} \\ \textit{Variable } 1 \\ \textit{Variable } 2 \end{array} \begin{array}{c} \textit{Group } 1 \quad \textit{Group } 2 \\ \begin{pmatrix} 1.2794 & -0.9558 \\ 2.0123 & 1.0042 \end{pmatrix} \end{array}.$$

- Group 1 covariance matrix:
$$\begin{array}{c} \\ \textit{Variable } 1 \\ \textit{Variable } 2 \end{array} \begin{array}{c} \textit{Variable } 1 \quad \textit{Variable } 2 \\ \begin{pmatrix} 0.8976 & 0.0781 \\ 0.0781 & 1.2407 \end{pmatrix} \end{array};$$

- Group 2 covariance matrix:
$$\begin{array}{c} \\ \textit{Variable } 1 \\ \textit{Variable } 2 \end{array} \begin{array}{c} \textit{Variable } 1 \quad \textit{Variable } 2 \\ \begin{pmatrix} 0.6765 & 0.4421 \\ 0.4421 & 0.3285 \end{pmatrix} \end{array}.$$