Working Group on Statistical Learning Seminar

Title: Clustering with the Multivariate Normal Inverse Gaussian Distribution

Speaker: Adrian O'Hagan

Date: Fri 18th March 2011 at 1:00PM

Location: Statistics Seminar Room- L550 Library building

Abstract: mclust is a popular model-based clustering methodology. It fits a Gaussian mixture model to the data. This implies that the data, within each group, is elliptically contoured. Hence any non-elliptical group must be modeled by more than one component, often resulting in over-fitting. This may render the clustering rule more ambiguous than if the correct (lower) number of groups is identified; ultimately this can result in higher misclassification rates.

Karlis Santourian (2008) use a mean-variance mixture of multivariate normal distributions with an inverse Gaussian mixing distribution (MNIG) in place of the Gaussian, to yield a more flexible family of distributions. The mixing distribution may be skewed and has fatter tails than the normal distribution. The model is fitted using the EM algorithm and the complete data likelihood requires two latent variables. The first is the standard group indicator variable and the second a mixing distribution value.

We extend the MNIG based approach to the full range of eigenvalue decomposed covariance structures considered in mclust. Furthermore we consider the family of

MNIG models where distributional parameters are constrained to be equal across groups, or omitted completely.

BIC is used to identify the optimal model and number of components. Disparities in clustering solutions under the mixture of MNIG and mixture of Gaussian (mclust) approaches are highlighted. Finally we detail the fitting of mixtures of MNIG distributions in the package MNIGclust to be made available in R.