GJ McLachlan

Given the extensive use of finite mixture models in the clustering of data sets from a wide variety of fields, it is instructive to examine from time to time the basic assumptions underlying this approach to clustering. This paper serves a useful role in this respect.

It is focussed on the case of mixed data where perhaps mixture models have not been applied to the same extent as with continuous data. However, various researchers have studied the use of mixture models for mixed data, concentrating on the location model, which formed the basis of the MULTIMIX procedure of Hunt and Jorgensen (1999); see, for example, Chapter 5 of McLachlan and Peel (2000) where mixture models are considered for discrete and mixed data.

The authors state that "an obvious problem with the model-based approach is that statisticians usually do not believe models to be true." Firstly, I would contend that statisticians like working with models (see the paper by Breiman (2001) with discussions from statisticians that included Professors Cox and Efron). Secondly, it reminds me of the well-known saying that "All models are wrong but some are useful" as attributed to the distinguished statistician Professor George Box. It is in the latter spirit that I have approached the use of mixture models for clustering. But also, I consider there are many situations where the component distributions in the mixture model being used to effect the clustering are relevant to describe the distribution of the data in the clusters so imposed, particularly in the biological, medical, and physical sciences. A generic example concerns the situation where the data on the phenomenon under study can be modelled adequately by a single normal distribution but when the system is perturbed, an additional normal component is needed. Although any distributional density for continuous data can be modelled with sufficiently high accuracy by a mixture of normals as noted by the authors, the key point in the aforementioned example is that it is the smallest number of normal components needed to model the distribution that is of interest.

Finally, the authors acknowledge that model-assumptions are helpful for making explicit the cluster-distributional assumptions being implicitly imposed with the use of so-called distribution-free methods of clustering. This is important since procedures such k-means are often claimed to be model-free whereas they are based on some particular assumption of the cluster distributions (normal distributions in equal proportions with common spherical component covariance matrices in the case of k-means).

References:

Breiman, L. (2001). Statistical Modeling: the two cultures.(with discussion). *Statistial Science* **16** 199-231.

Hunt, L.A. and Jorgensen, M.A. (1999). Mixture model clustering: a brief introduction to the MULTIMIX program. *Australian and New Zealand Journal of Statistics* **40**, 153–171.

McLachlan, G.J. and Peel, D. (2000). *Finite Mixture Models*. New York: Wiley.