Discussion of: *Clustering of objects on subsets of attributes* **by J.H. Friedman and J.J. Meulman**

Professor G.J. McLachlan and Dr. R.W. Bean, University of Queensland, Brisbane

The authors are to be congratulated on tackling the challenging problem of clustering very high-dimensional data. Their approach to find subgroups of objects that cluster on subsets of the attributes (variables) rather than on all of them simultaneously is relevant in many practical situations, such as in the example they give on the clustering of a limited number of tissue samples on the expression levels of thousands of genes.

The approach is similar to that adopted by McLachlan, Bean, and Peel (2002) in considering the latter problem via their EMMIX-GENE procedure. In order to cluster the tissue samples, McLachlan et al. (2002) clustered the variables into groups, on the basis of Euclidean distance, so that highly correlated variables tend to be put into the same group. Prior to this step of grouping the variables, the latter are standardized and there is also the provision of eliminating variables considered to be of little use (individually) in the clustering the data (in terms of fitted mixture models). Having grouped the variables into, say, n_o groups, the EMMIX-GENE program displays heat maps of the expression levels for the tissues for the genes within a group, which provides a visual aid in assessing which groups of genes will lead to similar clusterings of the tissues. This question can be addressed more formally by fitting mixture models on the basis of the genes within a group using, if necessary, mixtures of factor analyzers to handle the problem of a large number of variables relative to a limited number of objects. Alternatively, one can procceed by working with a representative of each group such as its sample mean (a metagene). We applied the EMMIX-GENE procedure to the mitochondrial RNA data set. A reduced number of genes (4978) was clustered into $n_o = 40$ groups, and then the tissues clustered on the basis of the top ten metagenes. Seven of the implied clusters for a nine-component mixture model corresponded to the experiments designated as Hol, Cho, Spe_alpha, Spe_elut, Spe_cdc, Chu, and Der, while the other two corresponded to Vel and to Myers and Roth combined.

Finally, in their discussion of non-distance-based modelling methods, the authors provide some references in the nineties on product density mixture modelling, which they note is the closest in spirit to their approach. Some earlier references on the mixture modelling approach to clustering are Wolfe (1967), Day (1969), and Ganesalingam and McLachlan (1979), and McLachlan and Basford (1988).

References in Discussion (not previously cited in paper)

Day, N.E. (1969). Estimating the components of a mixture of two normal distributions. *Biometrika* **56**, 463–474.

Ganesalingam, S. and McLachlan, G.J. (1979). *Statistica Neerlandica* **33**, 81–90.

McLachlan, G.J. and Basford, K.E. (1988). *Mixture Models: Inference and Applications to Clustering*. New York: Marcel Dekker.

McLachlan, G.J., Bean, R.W., and Peel, D. (2002). A mixture model-based approach to

the clustering of microarray expression data. *Bioinformatics* **18**, 413–422.

Wolfe, J.H. (1967). NORMIX: Computational methods for estimating the parameters of multivariate normal mixtures of distributions. *Research Memo. SRM 68–2*. San Diego: U.S. Naval Personnel Research Activity.