

COMMENT

Commentary on Steinley and Brusco (2011): Recommendations and Cautions

Geoffrey J. McLachlan
University of Queensland St. Lucia

I discuss the recommendations and cautions in Steinley and Brusco's (2011) article on the use of finite models to cluster a data set. In their article, much use is made of comparison with the K -means procedure. As noted by researchers for over 30 years, the K -means procedure can be viewed as a special case of finite mixture modeling in which the components are in equal (fixed) proportions and are taken to be normal with a common spherical covariance matrix. In this commentary, I pay particular attention to this link and to the use of normal mixture models with arbitrary component-covariance matrices.

Keywords: normal mixture models, expectation–maximization algorithm, mixtures of factor analyzers

The aim of Steinley and Brusco's (2011) article was to compare mixture modeling clustering (MMC) relative to the classical and easily implemented K -means procedure. As noted by Steinley and Brusco, it is well-known in the statistical literature that the K -means procedure can be viewed as a special case of MMC. An important reference on this point that was not included by Steinley and Brusco is Scott and Symons's (1971) article. Scott and Symons considered the clustering problem, using what has become known as the classification likelihood approach. With this approach, the likelihood function, L_C , for a specific number of clusters, g , is formed as if the cluster labels of the observations z_1, \dots, z_N are known. Here, z_j is a g -dimensional vector of zeros or ones, where the k th element is one if the j th observation vector, x_j , belongs to the k th cluster and is zero otherwise ($k = 1, \dots, g; j = 1, \dots, N$). On adopting some parametric form, $f_k(x_j; \theta_k)$, for the density of x_j in the k th cluster, the unknown cluster labels are treated as unknown parameters to be estimated along with the cluster-specific parameter vectors, θ_k , and the proportions, α_k ($k = 1, \dots, g$), in which the clusters occur. This estimation is achieved by maximization of L_C . Scott and Symons showed that many of the commonly used clustering procedures correspond to application of this classification likelihood approach with various restrictions on the cluster-specific covariance matrices, Σ_k , and the cluster proportions, α_k . In the special case where the latter are specified to be equal and the cluster densities are taken to be multivariate normal with a common spherical covariance matrix ($\Sigma_k = \sigma^2 I$), the K -means clustering procedure is obtained.

After the introduction of the expectation–maximization (EM) algorithm by Dempster, Laird, and Rubin (1977), which is commonly used to fit a mixture model to the data, McLachlan (1982) noted that L_C is equivalent to the so-called *complete-data likelihood* that is formed in the EM framework. It follows that K -means can also be viewed as a special case of MMC where (a) the normal

mixture model is taken to consist in known equal proportions of normal component densities with a common spherical covariance matrix and (b) on the E-step of the implementation of the EM algorithm, the current (fractional) estimates $\hat{\pi}_{kj}$ of the posterior probabilities of component membership of the mixture for the data point x_j are replaced by the \hat{z}_{kj} , where

$$\hat{z}_{kj} = 1, \text{ if } k = \arg \max_i \hat{\pi}_{ij}, \\ = 0, \text{ otherwise.}$$

Here $\hat{\pi}_{kj}$ is the estimated posterior probability that the j th data point belongs to the k th component of the mixture given the feature vector x_j ($k = 1, \dots, g; j = 1, \dots, N$).

Over the years, there has been increasing emphasis on a mixture likelihood-based approach to clustering, that is, on the use of MMC. Aitkin, Anderson, and Hinde (1981), in their reply to the discussion of their article, remarked that “when clustering samples from a population, no cluster method is, a priori believable without a statistical model” (p. 460). They also noted that “clustering methods based on such mixture models allow estimation and hypothesis testing within the framework of standard statistical theory” (Aitkin et al., 1981, p. 460). Previously, Marriott (1974) had noted that the mixture likelihood-based approach “is about the only clustering technique that is entirely satisfactory from the mathematical point of view. It assumes a well-defined mathematical model, investigates it by well-established statistical techniques, and provides a test of significance for the results” (p. 70).

Thus, against this background, I do not follow Steinley and Brusco's (2011) statement that “broad based simulation studies have yet to be implemented in this area to determine the ability of MMC procedures to correctly assign observations to clusters” (p. 64). If the model adopted for the component densities is valid, and there are sufficient data available to calculate the maximum likelihood estimates of the parameters, then it is clear from results on a decision-theoretic approach to classification that the MMC is using a plug-in version of the Bayes (optimal) rule for the allocation of the observations to the clusters. That is, MMC is providing an asymptotic optimal allocation of the data. In this sense, there is

Correspondence concerning this article should be addressed to Geoffrey J. McLachlan, Department of Mathematics and Institute for Molecular Bioscience, University of Queensland, St. Lucia, Brisbane 4072, Queensland, Australia. E-mail: g.mclachlan@uq.edu.au

no need for simulations to justify the use of MMC, at least in the case where the sample sizes are not small. If the subpopulations in the data are fairly widely separated, then MMC should still be quite useful in partitioning the data into clusters corresponding to the subpopulations; see, for example, Ganesalingam and McLachlan (1979) and other references in McLachlan and Peel (2000a, Chapter 3). To provide a useful clustering, it suffices that MMC is able to estimate the subpopulation boundaries with sufficient precision, although not necessarily the parameters themselves.

The simulations undertaken by Steinley and Brusco (2011) are of interest as they provide some results on the performance of MMC when the adopted model is not valid. However, as conceded by the authors themselves, the conclusions of their simulation studies are limited by the fact “that the results found herein are generalizable only to data sets that have the conditions that were studied” (p. 76). Indeed, this is why caution needs to be exercised if there is any attempt to use the simulation results of Steinley and Brusco to make general recommendations about the use of MMC. Moreover, the simulations for uniform component distributions, for example, might have limited applicability as potential users of MMC might have reservations about applying this method to such data. In applying MMC to cluster a data set, an implicit assumption is that the adopted mixture model is identifiable; that is, distinct values of the parameter vector in the adopted model correspond to different distributions apart from a permutation of the component labels in the case where the component densities belong to the same parametric family, for example, the normal. However, with uniform component densities, there is an identifiability problem, because a mixture of two overlapping uniform densities can be represented by a single uniform density (see Titterton, Smith, & Makov, 1985, p. 36).

Steinley and Brusco (2011) noted that “high-dimensional data appear to be somewhat of an Achilles’ heel for MMC” (p. 76). It is true that high-dimensional data present challenges for MMC, but this applies to any statistical technique that attempts to allow for correlations among many variables. As noted in Steinley and Brusco, Banfield and Raftery (1993) introduced a parameterization of each component-covariance matrix, Σ_k , based on a variant of the standard spectral decomposition. This approach can be used to consider whether it is reasonable to reduce the number of parameters in the specification of the Σ_k . However, if the number of variables, V , is large relative to the sample size, N , it may not be possible to use this decomposition to infer an appropriate model for the component-covariance matrices, Σ_k . Even if it is possible, the results may not be reliable because of potential problems with near-singular estimates of the component-covariance matrices when V is large relative to N . A singular estimate of a component-covariance matrix occurs when the observations identified with a cluster lie in a space of lower dimension than V . One way in which this can happen is if there are fewer than $V + 1$ observations belonging to a cluster.

However, such problems can be avoided by limiting the number of parameters in the specification of the component-covariance matrices by the adoption of a factor-analytic representation of them as, for example, with mixtures of factor analyzers; see McLachlan and Peel (2000a, Chapter 8), McLachlan and Peel (2000b), and more recently, Baek, McLachlan, and Flack (in press). However, if the number of variables is extremely large

relative to the sample size, n , as with microarray gene-expression data, then some form of variable selection may have to be performed first, as in McLachlan, Bean, and Peel (2002), before the fitting mixtures of factor analyzers. With the latter approach, the variables are first clustered with what is essentially a soft version of K -means, and then the variables are replaced by one or a few representatives of each variable cluster (so-called *metavariables*); for example, the mean or the first few principal components.

On some additional references concerning the question of the number of components to include in the model, there is a resampling approach, as considered by McLachlan (1987). Also, Keribin (2000) has shown the consistency of Bayesian information criterion in mixture problems.

References

- Aitkin, M., Anderson, D., & Hinde, J. (1981). Statistical modelling of data on teaching styles (with discussion). *Journal of the Royal Statistical Society A*, 144, 419–461.
- Baek, J., McLachlan, G. J., & Flack, L. (2010). Mixtures of factor analyzers with common factor loadings: Applications to the clustering and visualisation of high-dimensional data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32, 1298–1309.
- Banfield, J. D., & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49, 803–821.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B*, 39, 1–38.
- Ganesalingam, S., & McLachlan, G. J. (1979). Small sample results for a linear discriminant function estimated from a mixture of normal populations. *Journal of Statistical Computation and Simulation*, 9, 151–158.
- Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhya A*, 62, 49–66.
- Marriott, F. H. C. (1974). *Interpretation of multiple observations*. London, United Kingdom: Academic Press.
- McLachlan, G. J. (1982). The classification and mixture maximum likelihood approaches to cluster analysis. In P. R. Krishnaiah & L. Kanal (Eds.), *Handbook of statistics* (Vol. 2, pp. 199–208). Amsterdam, the Netherlands: North-Holland.
- McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics*, 36, 318–324.
- McLachlan, G. J., Bean, R. W., & Peel, D. (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, 18, 413–422.
- McLachlan, G. J., & Peel, D. (2000a). *Finite mixture models*. New York, NY: Wiley.
- McLachlan, G. J., & Peel, D. (2000b). Mixtures of factor analyzers. In P. Langley (Ed.), *Proceedings of the Seventeenth International Conference on Machine Learning* (pp. 599–606). San Francisco, CA: Morgan Kaufmann.
- Scott, A. J., & Symons, M. J. (1971). Clustering methods based on likelihood ratio criteria. *Biometrics*, 27, 387–397.
- Steinley, D., & Brusco, M. J. (2011). Evaluating mixture modeling for clustering: Recommendations and cautions. *Psychological Methods*, 16, 63–79.
- Titterton, D. M., Smith, A. F. M., & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. New York, NY: Wiley.

Received January 13, 2010

Revision received June 24, 2010

Accepted June 28, 2010 ■