Classification through reduced-dimensional mixture of multivariate Gaussians

Cinzia Viroli and Angela Montanari

Department of Statistics, University of Bologna, Italy

Model based clustering assumes that the data come from a finite mixture model with each component corresponding to a cluster. For quantitative data each mixture component is usually modeled as a multivariate Gaussian distribution.
When the number of observed variables is large, it is well known that Gaussian mixture models represent an over-parameterized solution as, besides the mixing weights, it is required to estimate the mean vector and the variance-covariance matrix for each component. This issue has been widely and variously addressed in the statistical literature (see Banfield and Raftery, 1993, McLachlan and Peel, 2000, for some of the most relevant proposals).
In this work we propose to address the dimension reduction issue in model based clustering by assuming that the mean centered p observed continuous variables have been generated according to the linear factor model

$$\mathbf{y} = \mathbf{\Lambda z} + \mathbf{u}. \tag{1}$$

where $\mathbf{z}$ is a r-dimensional vector of latent variables, $\mathbf{\Lambda}$ is the factor loading matrix and $\mathbf{u}$ is a p-dimensional Gaussian term which includes the so called specific factors with zero mean and diagonal covariance matrix $\mathbf{\Psi}$. We further assume that the vector of latent variables $\mathbf{z}$ can be modeled according to a finite mixture of multivariate Gaussians

$$\mathbf{z} \sim \sum_{i=1}^{k} w_i \phi_i^{(r)}(\boldsymbol{\mu}_i, \mathbf{\Sigma}_i) \tag{2}$$

where w_i are the unknown mixing proportions, $\phi_i^{(r)}$ is the r-dimensional Gaussian density with component mean and variance-covariance matrix $\boldsymbol{\mu}_i$ and $\mathbf{\Sigma}_i$ respectively. Modeling the factors as a multivariate Gaussian mixture amounts to model the observed variables as a particular multivariate Gaussian mixture model too:

$$\mathbf{y} \sim \sum_{i=1}^{k} w_i \phi_i^{(p)}(\mathbf{\Lambda}\boldsymbol{\mu}_i, \mathbf{\Lambda}\mathbf{\Sigma}_i\mathbf{\Lambda}^\top + \mathbf{\Psi}) \tag{3}$$

which allows for heteroscedastic mixture components, sharing the same $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ matrices, thus yielding a remarkable reduction in the number of free parameters either in the mean vectors and in the variance-covariance matrices. We evaluate the performance of the proposed classified method on simulated and real data.

References

[1] Banfield, J.D. and Raftery, A.E. (1993) Model-based Gaussian and non-Gaussian Clustering, *Biometrics, 49, 803-821*.

[2] Fraley, C. and Raftery, A.E., (2002) Model-based clustering, discriminant analysis, and density estimation, *Journal of the American Statistical Association, 97, 611-631*.

[3] McLachlan, G.J., Peel, D., (2000) *Finite Mixture Models*, John Wiley & Sons INC, New York.

[4] McLachlan, G.J., Peel, D., Bean, R.W., (2003) Modelling high-dimensional data by mixtures of factor analyzers, *Computational Statistics and Data Analysis*, 41, 379-388.