

Mixture models for simultaneous reduction and classification

Roberto Rocci

Dept. SEFeMQ, University of Rome "Tor Vergata", roberto.rocci@uniroma2.it

Maurizio Vichi

Dept. of Statistics, University "La Sapienza", maurizio.vichi@uniroma1.it

Keywords: mixture model, classification, principal component analysis.

Abstract

In many statistical applications, principal component analysis is applied to reduce the dimension of the data before clustering. This procedure, named "tandem analysis", has been criticized by several authors (Chang, 1983; De Soete & Carroll, 1994), because in the reduction phase some information relevant for the successive classification can be lost. To overcome this problem, some authors proposed methodologies for simultaneous classification and reduction of two-way data (De Soete & Carroll, 1994; Vichi & Kiers, 2001) or three-way data (Rocci & Vichi, 2001), in the context of hard or fuzzy partitioning. The main purpose of this paper is to reformulate these previous works using a mixture maximum likelihood approach. It is assumed that the observed data to be clustered are sampled from a finite mixture of Gaussians, i.e., each observation is taken to be a realization of a mixture density, where the components correspond to underlying groups. The groups-constrained covariance matrix depends on occasions and variables following Browne's (1984) direct product model. This allows to decompose the within-group variability into parts due to variables and occasions, respectively. The mean vectors of the groups are constrained to lie in a reduced subspace according to a Tucker model. In this way, the latent factors for variables and occasions that best explain the between-group variability in the data are identified. A computationally efficient ECM algorithm (Meng & Rubin, 1993) to compute the maximum likelihood estimates of the model parameters is also presented.

References

- Browne, M. W. (1984). The decomposition of multitrait-multimethod matrices, *British Journal of Mathematical and Statistical Psychology*, 37, 1-21.
- Chang, W. (1983). On using principal components before separating a mixture of two multivariate normal distributions. *Applied Statistics*, 32, 267-275.
- De Soete, G., & Carroll, J. D. (1994). K-means clustering in a low-dimensional Euclidean space. In E. Diday et al. (Eds), *New Approaches in Classification and Data Analysis* (pp. 212-219). Springer, Heidelberg.
- Meng, X. L., & Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80, 267-278.
- Rocci, R., Vichi, M. (2001). Simultaneous three-mode component and cluster analysis. *Proceedings of the Italian Statistical Society meeting on: Processi e metodi statistici di valutazione*. Roma.
- Vichi, M., & Kiers, H. A. L. (2001). Factorial k-means analysis for two-way data. *Computational Statistics and Data Analysis*, 37, 49-64.