**New challenges for modelling Big Data**

Maurizio Vichi1

1 Sapienza University of Rome, Department of Statistics,

maurizio.vichi@uniroma1.it

**Abstract**. *Big Data* represented by data matrices **X** with a huge number of rows (objects, statistical units) and columns (variables) are generally analysed to synthesize the relevant information and to obtain a reduced data structure formed by prototype (mean profiles) objects and latent variables. This is achieved by the simultaneous grouping of rows and columns of **X** so that the results are informative and easy to interpret, because denote a compressed, but relevant representation of the big data, while trying to preserve most of the original information. The reduction could represent a strong synthesis to be directly interpreted in order to identify the most important characteristics in term of ideal objects and ideal variables. Alternatively, the reduction could be soft to obtain a light compression of the multivariate data in order to allow the successive application of other multivariate computationally complex methods on the compressed data matrix.

In this presentation starting from an extension of standard *K*-means to simultaneously clustering of observations and features, namely Double *K*-Means (DKM) (Vichi, 2001), the model is developed in a probabilistic framework, an efficient coordinate ascent algorithm is proposed and the advantages of using this approach are discussed. DKM treats symmetrically the two modes of the data matrix (rows and columns) by producing a compressed set of prototype objects and prototype variables. The model is generalized to include an asymmetric treatment of the two modes so as to comprise both clustering and disjoint principal component analysis (Vichi Saporta, 2009) and its extension based on a probabilistic principal component analysis. A new coordinate ascent algorithm is developed and its performance is tested via simulation studies and real data sets. Finally, the results obtained on the real data are validated by building resampling confidence intervals for block centroids.

References

MARTELLA F. and VICHI M. (2012) : Clustering microarray data using model based double K-means, *Journal of Applied Statistics*,

VICHI M, (2001) : Double k-means clustering for simultaneous classification of objects and variables, in *Advances in Classification and Data Analysis*, S. Borra, R. Rocci, M. Vichi, and M. Schader, eds., Springer, Heidelberg, 43–52.

VICHI M. and Saporta G. (2009) : Clustering and Disjoint Principal Component, *Computational Statistics & Data Analysis*, vol. 53, 8, 3194-3208.

Keywords

Double K-means, Two-mode clustering, Coordinate Ascent Algorithm.