Some mathematical problems in symbolic data analysis

Richard Emilion
University of Orléans, Orléans, France - richard.emilion@univ-orleans.fr

Standard data analysis deals with a n rows by p columns table of data which are considered as n identically and independently distributed observations coming from a unique probability distribution $\mathbb{P}_X$ of a random vector $X = (X_1, \ldots, X_p)$ in $\mathbb{R}^p$.

First, Symbolic data analysis, as introduced by E. Diday (see e.g. [2]), consists in extending standard data analysis methods to the case where the observed data come independently from K sources, classes or groups, that is from K probability distributions on $\mathbb{R}^p$, say $\mathbb{P}_1, \ldots, \mathbb{P}_K$, respectively.

Next, as multivariate joint distributions on $\mathbb{R}^p$ are estimable with difficulty, several types of studies are possible. The most simple case, consists in dealing with the $K \times p$ marginal distributions $\mathbb{P}_{k,j}, j = 1, \ldots, p$ of the $\mathbb{P}_k$'s, $k = 1, \ldots, K$, while the most complex case directly deals with the joint distributions $\mathbb{P}_k$. As distributions are actually handled through their estimators, such studies raise some consistency problems. It can also be decided to deal with the random variables $X_{k,j}$, respectively their summarization such as their range.

The present talk paper will present some recent results addressing a few topics: Sources identification [1], Principal Component Analysis, Classification [4] and Concept Lattices [3] . It will be shown that interesting, but rather non trivial, tools and notions, such as copulas, Dirichlet random distributions [5], stochastic order [3], are needed. Some open questions will also be mentioned.

Keywords: copulas; Dirichlet distributions; symbolic; stochastic order.

References

[1] Benaglia T., Chauveau D., and Hunter D.R., An EM-like algorithm for semi-and nonparametric estimation in multivariate mixtures. *Journal of Computational and Graphical Statistics* 18.2 (2009): 505-526.

[2] E. Diday, The symbolic approach in clustering and related methods of Data Analysis, in *Proc. Classification and Related Methods of Data Analysis, IFCS*, H. Bock Ed., North-Holland, 1987.

[3] E. Diday and R. Emilion, Maximal and stochastic Galois lattices. *Discrete Applied Math. Journal*, vol. 127, pp. 271-284, 2003.

[4] R. Emilion, Unsupervised Classification and Analysis of objects described by nonparametric probability distributions. *Statistical Analysis and Data Mining* (SAM), Vol 5, 5, 388-398, 2012.

[5] J. F. Kingman, Random discrete distributions, *J. Roy. Statist. Soc. B* 37, 1-22, 1975.