Multiple Imputation and Multidimensional Scaling applied to a K-means Method

Ana Lorga da Silva[1], Gilbert Saporta[2] and Helena Bacelar-Nicolau[3]

[1] Faculdade de Economia e Gestão - ULHT
Campo Grande 376, 1749-024 Lisboa, Portugal, *ana.lorga@ulusofona.pt*
[2] Conservatoire National des Arts et Metiers
292 rue Saint Martin, 75141 Paris cedex 03, France, *gilbert.saporta@cnam.fr*
[3] FPCE, Universidade de Lisboa, *hbacelar@fpce.ul.pt*

Abstract. The effects of missing data (MD)and imputation methods(IM) in cluster analysis have been studied in, Silva (2005) and Silva et al. (2006), for some hierarchical classification methods and partition methods, in the case of variables clustering. As in Silva et al (2006) the partition method is the following: we start by finding a dissimilarity matrix between variables; a multidimensional scaling technique (MDS)-PROXSCAL-provides components which are used as inputs in a k-means method. In this communication, when there are MD, we evaluate the effect of IM combined with the PROXSCAL MDS procedure (Commandeur and Heiser (1993)): for a data matrix with missing data; m imputations are realized; m dissimilarity matrices are then obtained from each imputed matrix; PROXSCAL without constraints over these m dissimilarity matrices provides components; k-means is performed on these components and finally the partitions is compared with the original one ie with the complete data by means of the Rand index as in Youness and Saporta (2004) and an affinity coefficient as in Sousa (2006). The simulation study consists in generating different patterns of partitions from twenty-five variables following multinormal distributions. As in Silva (2005) data are deleted in increasing proportions to create MD patterns and several IM are compared.

Keywords: k-means, missing data, imputation methods, PROXSCAL

References

Commandeur, J. and Heiser, W.J. (1993). Mathematical Derivations in the Proximity Scaling (PROXSCAL) of Symmetric Data Matrices. In: Research Report RR-93-04, *DDT, Leiden University.*

Silva, A. L. (2005): *Tratamento de dados Omissos e Metodos de Imputacão em Classificacão*, Ph Thesis, CNAM, Paris and ISEG/UTL, Lisboa.

Sousa, A. (2005): *Contribuicões a Metodologia VL e indices de validacão para dados de natureza complexa*, Phd Thesis, Universidade dos Acores.

Silva, A. L., Saporta, G. and Bacelar-Nicolau (2006) Dealing with missing data in a k-means method - A simulation based approach. In: Book of Abstracts, *COMPSTAT 2006.*

Youness, G. and Saporta, G., (2004): Une Mthodologie pour la Comparaison de Partitions. *Revue de Statistique Applique, vol. 52(1), pp. 97-120.*