Robustness of AIC Based Criterion for Selecting the Number of Clusters

C.M. Santos-Pereira[1] and A.M. Pires[2]

[1] CEMAT/IST and Department of Civil Engineering, Faculty of Engineering, Oporto University
Rua Dr. Roberto Frias, 4200-465 Porto, Portugal, *carlasp@fe.up.pt*
[2] Department of Mathematics and CEMAT, IST, Technical Univ. of Lisbon.
Avenida Rovisco Pais, 1049-001, Lisboa, Portugal, *apires@math.ist.utl.pt*

Abstract A method to detect outliers in multivariate data based on clustering and robust estimators was introduced in Santos-Pereira and Pires (2002). In order to evaluate the performance of the method, we conducted a simulation study with several distributional situations, three clustering methods (K-means, pam and mclust) and three pairs of location-scatter estimators. One of the difficulties encountered in the implementation of the method, was the choice of the number of clusters, k, as well as the clustering method and the location-scatter estimators. At that time we suggested to apply several values of k (e.g. from 1 to a maximum possible k which depends on the number of observations and on the number of variables) and select k minimizing

$$AIC = -2\sum_{i=1}^{n} \log \hat{f}(\mathbf{x}_i) + 2k\left(p + \frac{p(p+1)}{2}\right),$$

with

$$\hat{f}(\mathbf{x}) = \sum_{j=1}^{k} \frac{n_j}{n_T} f_N(\mathbf{x}; \hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j), \quad \text{and} \quad n_T = \sum_{j=1}^{k} n_j.$$

(see Sakamoto et al. (1988) and Ronchetti (1997)).

In this communication we discuss the robustness of this AIC based criterion for choosing the number of clusters k, by using some distributional situations described in Santos-Pereira and Pires (2002) with and without outliers.

Keywords: AIC, outliers, robust estimators

References

RONCHETTI, E. (1997): Robustness aspects of model choice. *Statistica Sinica 7, 327–338.*

SAKAMOTO, Y., ISHIGURO, M., and KITAGAWA, G. (1988): *Akaike Information Criterion Statistics.* Kluwer Academic Publishers, New York.

SANTOS-PEREIRA, C.M. and PIRES, A.M. (2002): Detection of outliers in multivariate data: a method based on clustering and robust estimators. In: W. Härdle and B. Rönz (Eds.): *Proc. in Computational Statistics.* Physica-Verlag, Heidelberg, 291–296.