Adaptive trimmed k-means for heterogenous groups

L.A. García-Escudero[1] and A. Gordaliza[1]

[1] Departamento de Estadística e I.O., Universidad de Valladolid, Fac. de Ciencias, 47011, Valladolid, Spain

Keywords: Clustering, Robustness, Trimming, MCD, Estimation of scales and shapes.

1 Trimmed k-means

Trimmed k-means methodology was introduced by Cuesta-Albertos, Gordaliza and Matrán (1997) with the aim of robustifying the k-means method (the paradigm of the non-hierarchical cluster analysis) through the use of an "impartial" trimming procedure. The key idea is to allow the data themselves to tell us which observations are to be trimmed off. Its asymptotic properties and a constatation of the robustness gain have been already stated. Additionally, it has been shown that the careful analysis of trimmed k-means, moving k and the trimming level α, can be used as a guidance for the proper choice of the number of groups and the final trimming level.

2 Extension to heterogenous groups

The main drawback found when using trimmed k-means in general clustering problems is that the method is not well suited for dealing with very heterogenous groups. This drawback is completely inherited from classical (untrimmed) k-means where a mixture of spherical, equally sized and scattered clusters underlies.

In order to overcome this drawback, an adaptation of the fast-MCD algorithm in Rousseeuw and van Driessen (1999) to the clustering setting could be tried. An objective function analogous to that leading to the MINO problem in Rocke and Woodruff (2000) is then considered. However, that straight adaptation does not work as the objective function does not always decrease with the iterative algorithm.

Recently, Gallegos (2003) proposes another adaptation where groups are considered on an equal footing with respect to the scales in how distances to centers of groups are measured. Unfortunately, scales play a very important role in this framework (García-Escudero and Gordaliza (2004)) making this approach fail when scatters and weights are quite different.

Taking advantage of the fact that centers and shapes are correctly estimated with Gallegos (2003)'s procedure, we propose an iterative method that provides in each step better estimators of the scale and weight parameters. The procedure starts from an initial high trimming level and proceeds adding outer observations in a smooth way.

References

J.A. Cuesta-Albertos, A. Gordaliza and C. Matrán (1997). Trimmed k-Means: An Attempt to Robustify Quantizers *The Annals of Statistics.* **25**, 553-576.

L.A. García-Escudero and A. Gordaliza (2004). Generalized radius process for elliptically contoured distributions. To appear in *Journal of the American Statistical Association.*

M.T. Gallegos (2003). Robust clustering under general normal assumptions. *Preprint.*

D.M. Rocke and D.M. Woodruff (2000). A Synthesis of Outlier Detection and Cluster Identification. *Preprint.*

P.J. Rousseeuw and K. van Driessen (1999). A Fast Algorithm for the Minimum Covariance Determinant Estimator *Technometrics.* **41**, 212-223.