Robust Cluster Analysis

L.A.García-Escudero, A.Gordaliza, C.Matrán and A.Mayo
Departamento de Estadística e I.O.
Universidad de Valladolid

Introduction

A new method for performing robust clustering is proposed. The method allows for different covariance matrices and weights for the groups at the price of handling a restriction for the ratio between the maximum and the minimum eigenvalues of the groups' scatter matrices.

The problem

We start from a "spurious-outlier" model like that considered in Gallegos and Ritter (2005) and Gallegos (2002). But, apart from considering different covariance matrices for the groups as in Gallegos (2002), we also assume the existence of different weights π_j. Let $f(x_i;\mu,\Sigma)$ denote the p.d.f. of a p-variate normal distribution with mean μ and covariance matrix Σ and $g_{\psi_i}(x_i)$ denote other p.d.f.'s generating the "nonregular" part of the data. Given a p-dimensional sample $\{x_1, x_2, ..., x_n\}$, we will be interested in the maximization of

$$\prod_{j=1}^{k}\prod_{i\in R_j}\pi_j f(x_i;\mu_j,\Sigma_j)\prod_{i\notin R} g_{\psi_i}(x_i), \qquad (1)$$

with $R=\bigcup_{j=1}^{k} R_j$ corresponding to the indexes of the regular data and such that $\#R = n(1-\alpha)$ (a proportion α of nonregular observations is considered). We define some "assignation" functions $z_1,...,z_g$ such that $z_j(x)=1$ when a point $x\in\Re^p$ is assigned to group *j* and *0* otherwise. Analogously, we consider a function $z_0(x)=1$ whenever *x* should be trimmed off. The view of the previous sample problem in terms of these functions defined for all $x\in\Re^p$ allows us to define a new population problem. This extension provides us a theoretical framework for Robust Clustering. In order these problems to be well-defined, we consider an additional condition. If $\lambda_l(\Sigma_j)$ for j=1,...,k and l=1,...,p are the eigenvalues of the groups' scatter matrices, we fix a constant c>0 such that

$$\frac{M_n}{m_n}\le c \quad \text{with} \quad M_n=\max_{j=1..k}\max_{l=1..p}\lambda_l(\Sigma_j). \qquad (2)$$

This restriction may be seen as an extension of the (so-called) "Hathaway-type" restrictions. A wide range of problems may be declared depending on the strength of the restriction (2). In the extreme case of c=1 we have a weighted version of the trimmed k-means method (Cuesta-Albertos, Gordaliza and Matrán 1997). However, when relaxing this restriction, experience shows that it has not perverse effects and it merely prevents us from pathological cases that the generality of the assumed clustering model entails.

Existence and consistency

Under some mild conditions on the underlying distribution, the problem previously stated always has a solution for every fixed constant *c* chosen in (2). Moreover, the optimal solution of the sample case converges to the optimal solution of the population case. The eigenvalues-ratio restriction plays a very important role in the proofs of these two results.

Algorithm

An algorithm for obtaining approximate solutions for the sample problems is proposed. The proposed algorithm follows an EM methodology where we take advantage of the well-known Dykstra's algorithm for satisfying the eigenvalues-ratio restrictions.

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.

R.L.Dykstra(1993) An algorithm for restricted least squares regression, Journal of the American Statistical Association, 78, 837-842.

M.T.,Gallegos and G.Ritter (2005), A robust method for cluster analysis, The Annals of Statistics, 33, 347-380.

M.T.,Gallegos (2002), Maximum likelihood cluster with outliers, in Jajuga, K., Sokowski, A. and Bock, H.-H. (eds) Classification, Clustering and Data Analysis, Springer-Verlag, Berlin, 247-255.