K-MEANS CLUSTER ANALYSIS AND MAHALANOBIS METRICS: A PROBLEMATIC MATCH OR AN OVERLOOKED OPPORTUNITY?

Andrea Cerioli

Dipartimento di Economia, Sezione di Statistica e Informatica, Università di Parma Via Kennedy 6, 43100 – Parma, Italy.e-mail: andrea.cerioli@unipr.it

Abstract

In this paper we consider the performance of the widely adopted K*-means clustering algorithm when the classification variables are correlated. We measure performance in terms of recovery of the true data structure. As expected, performance worsens considerably if the groups have elliptical instead of spherical shape. We suggest some modifications to the standard* K*-means algorithm which considerably improve cluster recovery. Our approach is based on a combination of careful seed selection techniques and use of Mahalanobis instead of Euclidean distances. We show that our method performs well in a number of examples where the standard algorithm fails. In such applications our nonparametric technique is seen to be competitive when compared to parametric model-based clustering methods. Hence, our conclusion is that use of the Mahalanobis distance should become a standard option of the available* K*-means routines for non-hierarchical cluster analysis. This goal can be achieved by minor modifications in popular commercial software.*

1. INTRODUCTION

The performance of clustering algorithms can be measured through their ability to recover clusters that are known in advance to exist. If there is correlation among the classification variables, so that groups exhibit elliptical rather than spherical shape, the performance of many widely used methods often worsens to a considerable extent (Everitt, 1993). For example, consider the two-population simulated data set shown in Figure 1 and taken from Atkinson, Riani and Cerioli (2004). More details on it are given in Section 4.1. To this data set we apply the popular convergent K-means algorithm as implemented in *SPSS*®, called the *Quick Cluster* algorithm, with $K = 2$. It is very unfortunate to see that even in such a clear-cut example, where we know the true number of groups, the algorithm misclassifies five units when applied to standardized variables, and seven units when run on the raw data. Figure 1 also shows the misclassified units from the best of these standard solutions.

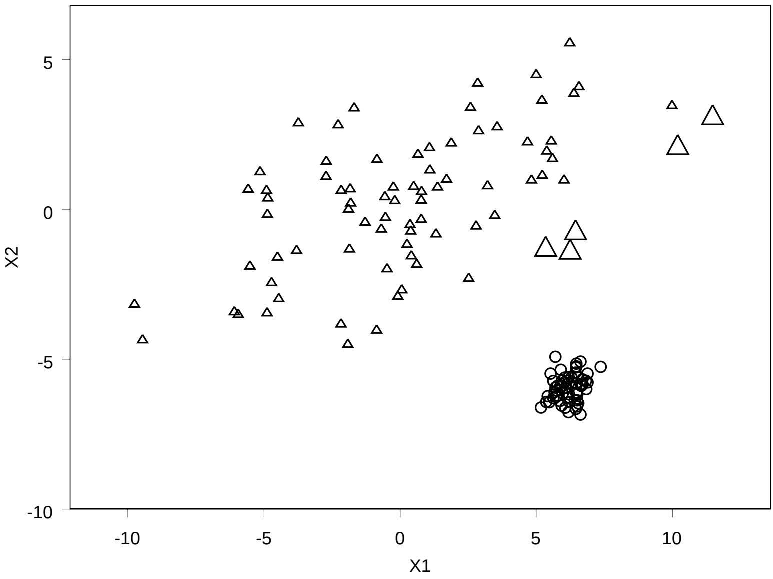

Fig. 1: Example 1: simulated bivariate data set with two clusters. Larger symbols indicate the five units misclassified by the *Quick Cluster* algorithm of *SPSS*® run on standardized variables.

The failure of the standard *K*-means method is motivated by the fact that cluster analysis is performed with a distance function which is chosen *a priori*, without reference to the actual shape of the clusters. Traditionally, the *K*-means routine has been implemented using the Euclidean distance, which ensures convergence of the algorithm. Some proprietary implementations (e.g. *SAS*® and *S-Plus*®) have also included the options for the city-block metric and for other robust clustering techniques (Struyf, Hubert and Rousseeuw, 1997). However, even these robust methods are usually unable to improve the recovery of elliptical clusters.

Projection onto an orthogonal space having the same dimensions as the original data set does not necessarily make elliptical clusters more apparent. For our simulated data, application of the *K*-means algorithm to the principal component scores computed from the correlation matrix leads to the same results as in Figure 1. An additional problem arising here, as well as in most classification studies, relates to the question of whether or not variables should be standardized before the analysis. It is well known that allocation is not invariant to this choice and no general result is available to provide definite guidance towards the best solution (see, e.g., Gordon, 1999, pp. 23-26).

The purpose of this paper is to investigate the performance with elliptical clusters of a modified *K*-means algorithm using Mahalanobis instead of Euclidean distances. The Mahalanobis distance is a basic ingredient of many multivariate techniques, but has rarely been adopted for clustering purposes. One reason is that the most familiar definition of the Mahalanobis distance is for a single population, and its computation can markedly worsen the recovery of clusters, as Figure 2 below shows. An additional reason is that even the cluster-specific distances supported in this paper can be negatively affected by poor initialization of the algorithm. Hence our proposal is to combine computation of adaptive Mahalanobis distances with identification of an effective starting point for the iterative *K*-means algorithm.

The paper is organized as follows. In Section 2 we establish notation and give a brief review of the use of Mahalanobis metrics for clustering purposes. The proposed modification of the *K*-means method is then described in Section 3. In Section 4 we compare the performance of our algorithm to that of standard available software in a number of data sets. The paper ends with some concluding remarks in Section 5.

2. PRELIMINARIES ON MAHALANOBIS METRICS

The common way of dealing with correlation and differential weighting of variables in many multivariate techniques is through the Mahalanobis distance. Let $\mathbf{x}_i = [x_{i1},\ldots, x_{ip}]'$ and $\mathbf{x}_j = [x_{j1},\ldots, x_{jp}]'$ denote two multivariate observations from a population with p variables $X_1,\ldots, X_p$ and covariance matrix Σ. Suppose that Σ is estimated by its unbiased estimator $\mathbf{S} = \mathbf{C}'\mathbf{C}/(n - 1)$, where $\mathbf{C}$ is the "centred" data matrix. The (sample) Mahalanobis distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ is then defined as

$$D(\mathbf{x}_i, \mathbf{x}_j) = \{(\mathbf{x}_i - \mathbf{x}_j)'\mathbf{S}^{-1}(\mathbf{x}_i - \mathbf{x}_j)\}^{1/2}. \tag{1}$$

The Mahalanobis distance (1) has the major advantage of taking correlations into account when $\mathbf{S}$ is not diagonal. Clearly, $D(\mathbf{x}_i, \mathbf{x}_j)$ is equivalent to an ordinary Euclidean distance after orthonormal transformation of $X_1,\ldots, X_p$. That is

$$D(\mathbf{x}_i, \mathbf{x}_j) = \{(\mathbf{z}_i - \mathbf{z}_j)'(\mathbf{z}_i - \mathbf{z}_j)\}^{1/2}, \tag{2}$$

where $\mathbf{z}_i = \mathbf{S}^{-1/2}\mathbf{x}_i$ and similarly for $\mathbf{z}_j$. The orthonormal transformation matrix $\mathbf{S}^{-1/2}$ can easily be computed through the singular value decomposition of $\mathbf{S}$.

Introductory properties and traditional applications of the Mahalanobis distance are reviewed in many textbooks (e.g. Seber, 1984; Rencher, 2002). Recently, its use has been advocated in various ways as an important tool for the purpose of detecting multivariate outliers (Rousseeuw and van Zomeren, 1990; Hadi, 1992; Rocke and Woodruff, 1996; Atkinson, Riani and Cerioli, 2004). However, despite the recognition of its importance in multivariate analysis, the Mahalanobis distance has never gained much popularity as a dissimilarity measure among classification practitioners.

A basic reason why use of $D(\mathbf{x}_i, \mathbf{x}_j)$ has been strongly discouraged in cluster analysis is that definition (1) is adequate only for units coming from the same population. In its influential book, Hartigan (1975, p. 63) wrote that "The Mahalanobis distance based on the full data covariance matrix is even worse than the "equal variance" scale in decreasing the clarity of clusters". This shortcoming is particularly evident for the simulated data set introduced in Section 1. Figure 2 shows the classification output provided for such data by the convergent K-means algorithm of *SPSS*®, when applied after the orthonormal transformation of equation (2). It is seen that performance is very poor in this application, with as many as twenty misclassified units.

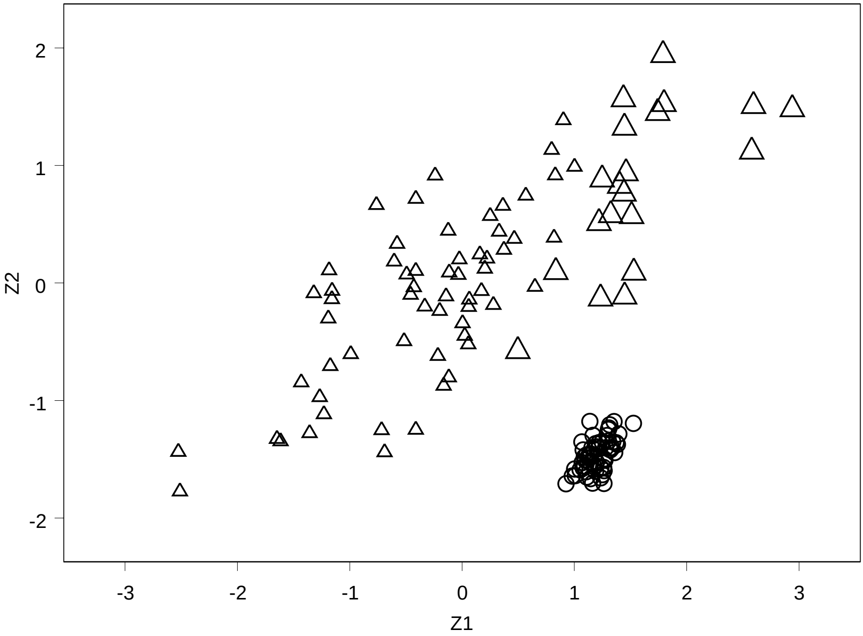

Fig. 2: Example 1: simulated bivariate data set with two clusters. Larger symbols indicate the twenty units misclassified by the *Quick Cluster* algorithm of *SPSS*® run on orthonormal variables (see equation (2)). The orthonormal transformation matrix S is computed on all observations.

An apparently more sensible approach would be to define Σ as the pooled within groups covariance matrix. However, also this approach does not result in a satisfactory solution, because the groups are not known in advance and estimation of Σ becomes not trivial. Gnanadesikan, Harvey and Kettenring (1993) suggested an iterative procedure for the computation of the within groups covariance matrix that does not require prior knowledge of the group structure, but their technique is cumbersome and not intuitive for the practitioner. Furthermore, computing a pooled covariance matrix implies the often unreasonable assumption of a common covariance matrix for all clusters.

In the next section we describe a more flexible approach to *K*-means clustering, which can be applied with group-specific covariance matrices. Although covariance estimation still requires iteration, our method fits nicely within the iterative nature of non-hierarchical techniques. Hence the resulting procedure is much simpler than the algorithm of Gnanadesikan, Harvey and Kettenring (1993) and it can be implemented through minor modifications of the available software. We also address the relevant issue of selecting an accurate starting point for the iterative procedure. We argue that this step is even more important if clusters have elliptical instead of spherical shape.

A similar algorithm was also explored by Maronna and Jacovkis (1974), but with results different from ours. Indeed, the examples of Section 4 show that the adoption of the Mahalanobis metric with careful choice of the preliminary seeds can significantly improve the performance of cluster analysis. In these applications the performance of our technique is comparable to that of some recent model-based clustering algorithms based on mixtures of multivariate normal distributions (Fraley and Raftery, 1999; 2002). On the contrary, our nonparametric method does not make any distributional assumption about the structure of the data. Differential weighting of variables in the Euclidean space was also recommended by Milligan (1989) in the related context of ultrametric hierachical clustering.

3. THE PROPOSED CLUSTERING ALGORITHM

Our starting point is the convergent *K*-means algorithm of Anderberg (1973, p. 163), which underlies many familiar non-hierarchical clustering procedures, such as *Quick Cluster* of *SPSS*® and *Fastclus* of *SAS*®. The basic idea, which might be traced back at least to Chernoff (1972), is to modify the standard method in order to let each group determine its own metric. Chernoff suggestion and a few variations were examined by Maronna and Jacovkis (1974), who concluded that the adoption of cluster-specific Mahalanobis distances like those in equation (3) below was not worth considering. Diday et Govaert (1977) also investigated the convergence properties of a similar iterative algorithm and the geometry of the clusters identified by it under different settings. A brief review of this material is given by Gordon (1999, p. 48).

An additional issue is that many validation studies have shown the importance of careful seed choice at a preliminary stage. Random selection of seeds may be largely inefficient, and more effective procedures are usually implemented (e.g. Gordon, 1999, p. 41). Furthermore, the convergent K-means algorithm has the unappealing feature of being potentially sensitive to the order in which the units are listed (Anderberg, 1973, p. 163), a drawback often overlooked in applications. In this paper we adopt either the refined seed selection technique of Cerioli (1999), or the exploratory approach of Cerioli and Zani (2001). The first method drastically reduces the list order effect on the K-means algorithm through an improvement of the seed choice rule implemented in the *Fastclus* procedure of *SAS*®. The second technique aims at detecting modes in the data cloud by the use of some tools from exploratory spatial statistics. Both methods outperform standard seed selection techniques implemented in popular packages such as *SAS*® or *SPSS*®.

Given n multivariate observations $\mathbf{x}_1, ..., \mathbf{x}_n$ and the value of K, the steps of the modified convergent K-means algorithm are summarized as follows.

1. Compute K preliminary centroids using Euclidean distances and careful seed selection methods. For this purpose, adopt either the technique of Cerioli (1999) or that of Cerioli and Zani (2001).
2. Assign each unit to the nearest centroid cluster according to the Euclidean distance criterion. Let n_k denote the number of units assigned to cluster k.
3. For $k = 1, ..., K$, compute the centroid, say $\mathbf{c}_k$, of the data units belonging to cluster k. If $n_k > 1$ also compute the corresponding unbiased estimate of the covariance matrix, say $\mathbf{S}_k$.
4. Take observation $\mathbf{x}_i$ and compute the Mahalanobis distances

   $$D(\mathbf{x}_i, \mathbf{c}_k) = \{(\mathbf{x}_i - \mathbf{c}_k)'\mathbf{S}^{-1}{}_k(\mathbf{x}_i - \mathbf{c}_k)\}^{1/2} \tag{3}$$

   to all centroids of clusters such that $n_k > 1$ and $\mathbf{S}_k$ is non-singular. If $n_k > 1$ and $\mathbf{S}_k$ is singular, define $D(\mathbf{x}_i, \mathbf{c}_k)$ to be the modified Mahalanobis distance of Hadi (1992, eq. 2.3). If $n_k = 1$, define $D(\mathbf{x}_i, \mathbf{c}_k)$ to be the Euclidean distance between $\mathbf{x}_i$ and $\mathbf{c}_k$.
5. Assign unit i to the nearest centroid cluster according to $D(\mathbf{x}_i, \mathbf{c}_k)$. If this step moves unit i from one cluster to another, update the corresponding centroids and covariance matrices as in step 3.

6. Repeat steps 4 and 5 for $i = 1, ..., n$.
7. Iterate steps 4 through 6 until convergence or until the allowed maximum number of iterations is reached.

A FORTRAN program for performing steps 1 through 7 has been developed by the author and can be made available upon request. This program is based on simple modifications of the standard Anderberg's algorithm that could easily be implemented in commercial software for cluster analysis.

The modified *K*-means algorithm described above is able to take correlations into account, once that a tentative classification with non-trivial clusters has been set up. However, at a preliminary stage Euclidean distances are still used. Hence a possible refinement would be to replace the Euclidean distances, say $d(\mathbf{x}_i, \mathbf{c}_k)$, in steps 1 and 2 with the weighted distances

$$d^*(\mathbf{x}_i, \mathbf{c}_k) = d(\mathbf{x}_i, \mathbf{c}_k)/(1 + w_{i(k)}), \tag{4}$$

where $w_{i(k)}$ is a function of the number of observations within a circle of radius $d(\mathbf{x}_i, \mathbf{c}_k)$ centred at $\mathbf{c}_k$ that point broadly to the same direction as $(\mathbf{x}_i - \mathbf{c}_k)$.

The rationale behind (4) is that we want to downweight Euclidean distances along directions of high correlation, and hence of high density in the data cloud, in a way similar to what is done by $D(\mathbf{x}_i, \mathbf{c}_k)$. The individual weight given to observation $\mathbf{x}_j \neq \mathbf{x}_i$ in the computation of $w_{i(k)}$ is taken to be proportional to $exp(\varphi \cos(2\theta_{ij}))$, where θ_{ij} denotes the angle between the vectors $(\mathbf{x}_i - \mathbf{c}_k)$ and $(\mathbf{x}_j - \mathbf{c}_k)$, and φ is a smoothing parameter. This weight corresponds to the most important example of a kernel available for spherical data. It gives the probability density for θ_{ij} under a von Mises – Fisher distribution with concentration parameter φ and support on the arc $[-\pi/2, +\pi/2]$ (Mardia and Jupp, 2000). A detailed comparison of standard Euclidean distances with the weighed distances (4) is under investigation and will be reported elsewhere.

4. EXAMPLES

In the examples that follow we focus on cluster summaries. Performance of clustering algorithms is measured through the misclassification error rate, that is the rate of units allocated to a wrong cluster. With G populations, the notation $(\nu_{1(g)}, \nu_{2(g)}, ..., \nu_{G(g)})$ for cluster g means that ν_1 units classified in that cluster actually come from population 1, ν_2 units come from population 2, etc. Since the ordering of cluster labels is unimportant, we report results for cluster g as being related to population g. The misclassification error rate is then defined as

$$R = \frac{1}{n}\sum_{g=1}^{G}\sum_{g'=1}^{G} v_{g'(g)} I\left\{g \neq g'\right\},$$

where $I\{g \neq g'\}$ is the indicator function taking value 1 if $g \neq g'$ and 0 otherwise.

Actual group sizes are denoted by $m_1, ..., m_G$ ($m_1 + ... + m_G = n$). Since the purpose of the following examples is to compare group recovery for alternative methods, we assume that G is known and set $K = G$. We take as basic K-means the *Quicksort* algorithm of *SPSS*® 11.0 for Windows, while robust K-means refers to the *Pam* algorithm as implemented in *S-Plus*® 6 for Windows (Struyf, Hubert and Rousseeuw, 1997). We do not claim that these methods are the "gold" classification standards. However, they provide a useful benchmark for comparison, due to their popularity among applied researchers. Data are always standardized before applying the scale variant methods of *SPSS*® and *S-Plus*®.

4.1 EXAMPLE 1: TWO-POPULATION BIVARIATE DATA SET

This simulated data set is shown in Figure 1. The data are obtained by simulation of $m_1 = 80$ observations (displayed as triangles) from a bivariate normal distribution with var(X_1) = 25, var(X_2) = 5 and cov(X_1; X_2) = 6, and of $m_2 = 60$ observations (shown as circles) from a bivariate normal distribution with var(X_1) = var(X_2) = 0.2 and cov(X_1; X_2) = 0.01. The full data set is given in Table A.13 of Atkinson, Riani and Cerioli (2004).

For these data basic K-means gives (75, 0) and (5, 60). The error rate of the standard method is then $R = 5/140 = 0.036$. Robust K-means slightly improves cluster recovery, yielding an error rate of $R = 3/140 = 0.021$. On the contrary, basic K-means run after the orthonormal transformation of equation (2) results in the disastrous performance of Figure 2 ($R = 20/140 = 0.143$).

The modified convergent K-means algorithm described in Section 3 provides the correct classification ($R = 0$).

4.2 EXAMPLE 2: MARONNA AND JACOVKIS (1974) PATTERN

For ease of comparison with previous literature, we have simulated a sample of 25 units from each of four bivariate normal populations as in Maronna and Jacovkis (1974). Of these populations, three have non spherical covariance matrices. The data are displayed in Figure 3, where misclassified units according to our method are highlighted.

Maronna and Jacovkis reported of an unsatisfactory behaviour of the cluster-specific Mahalanobis metric (3). Our findings, however, disagree with their results. In fact, the modified convergent *K*-means algorithm yields (23, 0, 3, 0), (2, 25, 0, 0), (0, 0 22, 0) and (0, 0, 0, 25). The misclassification rate is then $R = (3 + 2)/100 = 0.05$, which is about half the values for basic and robust *K*-means. In this application an orthonormal transformation based on all observations gives essentially the same result as our modified convergent *K*-means algorithm.

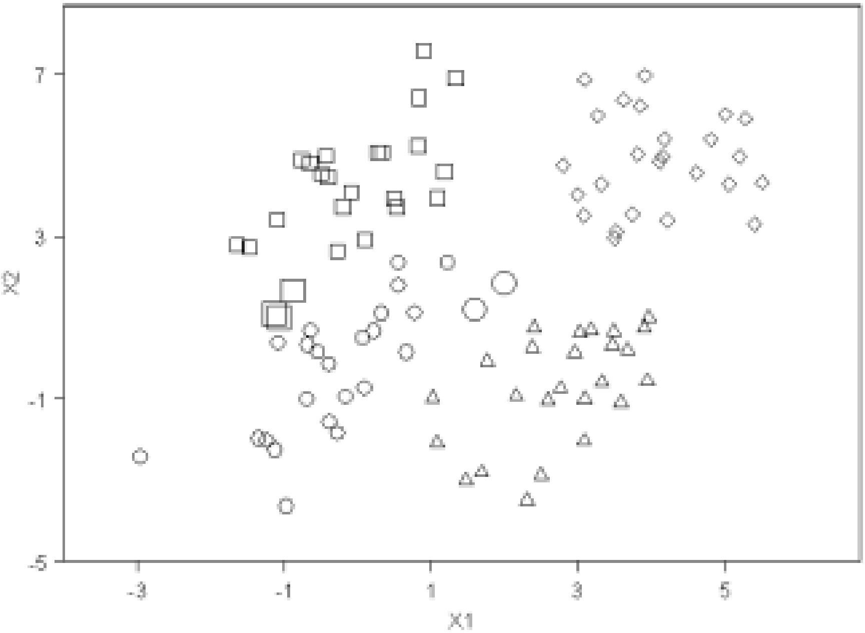

Fig. 3: Example 2: simulated bivariate data set with four clusters. Larger symbols indicate the five units misclassified by the modified convergent *K*-means algorithm based on the Mahalanobis distance.

4.3 IRIS DATA

As a final example, we analyze the famous Fisher's Iris data set, with four variables, three groups and $m_1 = m_2 = m_3 = 50$. Standard (Euclidean) clustering methods are prone to high misclassification rates, due to the large overlap between two groups, corresponding to the *Versicolor* and *Virginica* species. However, a simple exploratory analysis following the approach of Cerioli and Zani (2001) clearly shows three modes in the underlying density. Figure 4 is a three-dimensional view of Cerioli and Zani's measure of local clustering in the data cloud. The base axes represent the indexes of quadrats obtained by superimposing a regular grid onto the scatterplot of the first two principal component scores. The histogram highlights the modes in the corresponding density cloud. See Cerioli and Zani (2001, pp. 14-15) for further details.

We use information from Figure 4 to compute the preliminary centroids in step 1 of our algorithm. Specifically, we consider the peaks in the plot of the clustering measure as indicators of clusters. Then we pick out the units which give rise to such peaks, as well as the units belonging to quadrats close to the peak and for which the diagnostic measure is still high. This exploratory step ends up with $n_1 = 22$, $n_2 = 23$ and $n_3 = 19$ observations, which form our tentative initial classification. We compute the centroids of such observations and use them to start up the algorithm. Careful choice of preliminary seeds is crucial in this application, due to the large overlap between two of the three groups.

Our modified convergent K-means algorithm misclassifies only five observations from the *versicolor* group. This error rate ($R = 5/150 = 0.033$) is very similar to those obtained for Fisher's iris data by modern classification techniques making stronger distributional assumptions, such as model based clustering (Coleman *et al*., 1999; Fraley and Raftery, 1999) and classification through a local fit (Loader,

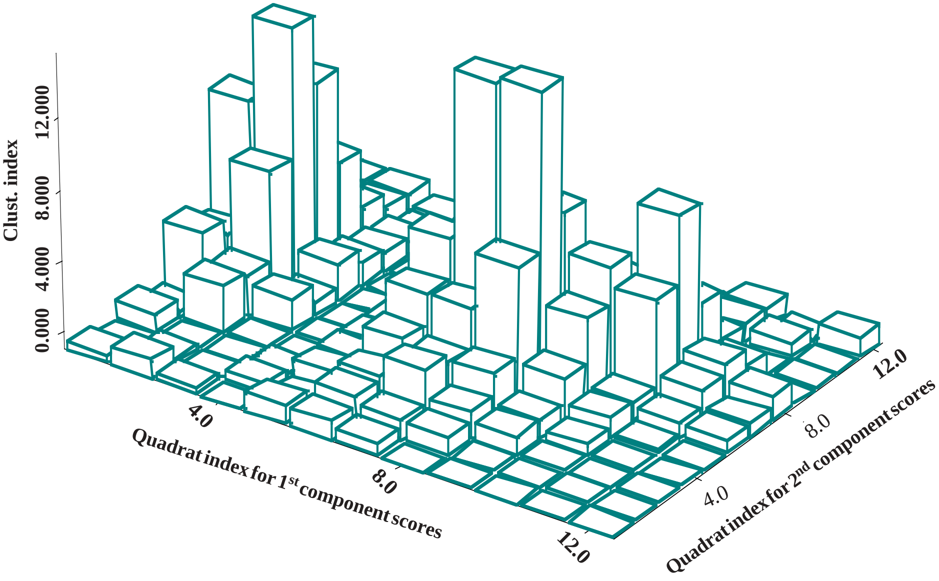

Fig. 4: Example 3: Iris data. Plot of the local clustering index of Cerioli and Zani (2001) computed on principal component scores. Three modes are apparent.

1999, p. 147). On the contrary, both basic and robust K-means with $K = 3$ yield an unsatisfactory $R \cong 0.15$. The orthonormal transformation described in (2) has a disastrous effect here, as in our first example, leading to a much larger error rate than algorithms based on the Euclidean distance.

It is important to remark that the performance of standard methods does not improve even when they are started from the same preliminary seeds as our algorithm. Therefore, it is the combination of a proper initial classification and use of the Mahalanobis metric (3) that leads to good cluster recovery in this application.

As a further refinement, we might also compute the covariance matrices of the 3 tentative clusters identified in step 1. Step 2 could be modified consequently, and the Mahalanobis distance could then be introduced from the beginning of the algorithm. Although this development is likely to improve the clustering performance of our technique in many applications, it does not modify the final allocation of the iris data set.

5. CONCLUSIONS

In this paper we have addressed the issue whether the Mahalanobis distance should have a role as a distance function in non-hierarchical cluster analysis. Mahalanobis metrics have been introduced a long time ago in the statistical literature, but have never gained much popularity within the classification community. On the contrary, we have shown that a sound combination of careful seed selection procedures and use of cluster-specific Mahalanobis distances can significantly improve the performance of standard clustering algorithms when groups have elliptical instead of spherical shape.

We do not contend that our approach gives the best performance in all situations. However, we make the point that it should be considered as a valuable option in routine cluster analysis problems. An additional bonus is that the adoption of Mahalanobis distances yields a scale invariant classification, thus overcoming the usual dilemma about variable standardization.

Finally, we remark that our approach is simple to implement and easy to understand, because it relies on minor modifications of the convergent K-means algorithm as implemented in many popular statistical packages. We hope that our results will stimulate further research about the Mahalanobis metric (3) and a wider appreciation of its usefulness in applications. An emerging research area that has not been covered in this paper, but where computation of (3) is an important step, is the development of robust classification techniques. We refer to Atkinson, Riani and Cerioli (2004) for further details on this topic.

REFERENCES

ANDERBERG M. R. (1973) *Cluster Analysis for Applications*, Academic Press, New York.

ATKINSON A. C., RIANI M., CERIOLI A. (2004) *Exploring Multivariate Data with the Forward Search*, Springer, New York.

CERIOLI A. (1999) Measuring the influence of individual observations and variables in cluster analysis, in: *Classification and Data Analysis*, Vichi, M. & Opitz, O. (Eds.), Springer, Berlin, 3-10.

CERIOLI A., ZANI S. (2001) Exploratory methods for detecting high density regions in cluster analysis, in: *Advances in Classification and Data Analysis*, Borra S., Rocci R., Vichi, M. & Schader, M. (Eds.), Springer, Berlin, 11-18.

CHERNOFF H. (1972) Metric considerations in cluster analysis, in: *Proceedings of the VIth Berkeley Symposium on Mathematical Statistics and Probability*, Le Cam L. M., Neyman J. & Scott E. L. (Eds.), vol. 1, 621-629.

COLEMAN D., DONG X., HARDIN J., ROCKE D. M., WOODRUFF D. L. (1999) Some computational issues in cluster analysis with no a priori metric, *Computational Statistics and Data Analysis*, 31, 1-11.

DIDAY E., GOVAERT G. (1977) Classification automatique avec distances adaptives, *R.A.I.R.O. Informatique/Computer Science*, 11, 329-349.

EVERITT B. S. (1993) *Cluster Analysis*, 3rd edition, Arnold, London.

FRALEY C., RAFTERY A. E. (1999) MCLUST: Software for model-based cluster and discriminant analysis, *Journal of Classification*, 16, 297-306.

FRALEY C., RAFTERY A. E. (2002) Model-based clustering, discriminant analysis, and density estimation, *Journal of the American Statistical Association*, 97, 611-631.

GNANADESIKAN R., HARVEY J. W., KETTENRING J. R. (1993) Mahalanobis metrics for cluster analysis, *Sankhya*, 55, A, 494-505.

GORDON A.D. (1999) *Classification*, 2nd edition, Chapman & Hall/CRC, Boca Raton.

HADI A.S. (1992) Identifying multiple outliers in multivariate data, *Journal of the Royal Statistical Society* B, 54, 761-771.

HARTIGAN J. A. (1975) *Clustering Algorithms*, Wiley, New York.

LOADER C. (1999) *Local Regression and Likelihood*, Springer, New York.

MARONNA R., JACOVKIS P. M. (1974) Multivariate clustering procedures with variable metrics, *Biometrics* 30, 499-505.

MARDIA K.V., JUPP P. E. (2000) *Directional Statistics*, Wiley, New York.

MILLIGAN G.W. (1989) A validation study of a variable weighting algorithm for cluster analysis, *Journal of Classification*, 6, 53-71.

RENCHER A.C. (2002) Methods of Multivariate Analysis, 2nd edition, Wiley, New York.

ROCKE D., WOODRUFF D. (1996) Identification of outliers in multivariate data, *Journal of the American Statistical Association*, 91, 1047-1061.

ROUSSEEUW P.J., VAN ZOMEREN B.C. (1990) Unmasking multivariate outliers and leverage points, *Journal of the American Statistical Association*, 85, 633-639.

SEBER G. A. F. (1984) *Multivariate Observations*, Wiley, New York.

STRUYF A., HUBERT M., ROUSSEEUW P. J. (1997) Integrating robust clustering techniques in S-plus, *Computational Statistics and Data Analysis*, 26, 17-37.

METODO DELLE K-MEDIE E DISTANZA DI MAHALANOBIS: UN’UNIONE DIFFICILE O UN’OPPORTUNITÀ SOTTOVALUTATA?

Riassunto

In questo lavoro si studia il comportamento, definito in termini di capacità di ricostruzione delle strutture di gruppo presenti nei dati, del ben noto metodo delle K*-medie nel caso in cui le variabili di classificazione siano correlate tra loro. Come è logico attendersi, la capacità di ricostruzione di tale metodo peggiora drasticamente nel caso di gruppi di forma ellittica anziché sferica. Proponiamo pertanto alcune semplici ma efficaci modifiche al metodo tradizionale che consentono di ovviare ad un simile inconveniente. Il nostro approccio si fonda sull’impiego combinato di criteri accurati per la scelta della partizione iniziale e della distanza di Mahalanobis in luogo di quella euclidea nei passi successivi dell’algoritmo di classificazione. Attraverso alcuni esempi, mostriamo come l’approccio seguito in questo articolo possa migliorare sensibilmente l’accuratezza delle partizioni ottenute attraverso il metodo classico delle* K*-medie, risultando così competitivo con le tecniche di classificazione che seguono un approccio modellistico. La conclusione del presente lavoro è dunque che l’impiego della distanza di Mahalanobis dovrebbe diventare un’opzione abituale a disposizione degli utilizzatori delle procedure di classificazione non gerarchica. Tale obiettivo può essere agevolmente raggiunto mediante alcune semplici modifiche ai software di uso corrente.*