Robust Clustering Method for the Detection of Outliers: using AIC to Select the Number of Clusters

Carla M. Santos-Pereira and Ana M. Pires

Abstract In [14] we proposed a method to detect outliers in multivariate data based on clustering and robust estimators. To implement this method in practice it is necessary to choose a clustering method, a pair of location and scatter estimators, and the number of clusters, *k*. After several simulation experiments it was possible to give a number of guidelines regarding the first two choices. However the choice of the number of clusters depends entirely on the structure of the particular data set under study. Our suggestion is to try several values of *k* (e.g. from 1 to a maximum reasonable *k* which depends on the number of observations and on the number of variables) and select *k* minimizing an adapted AIC. In this paper we analyze this AIC based criterion for choosing the number of clusters *k* (and also the clustering method and the location and scatter estimators) by applying it to several simulated data sets with and without outliers.

1 Methodology

The procedure most commonly used to detect outliers in multivariate data sets is based on the Mahalanobis distances, $(\mathbf{x}_i - \hat{\mu})^T \hat{\Sigma}^{-1} (\mathbf{x}_i - \hat{\mu})$, $i = 1, \ldots, n$. To avoid the masking effect it is recommended to use robust estimates, $\hat{\mu}$ and $\hat{\Sigma}$, instead of the classical estimates, i.e. the sample mean vector and the sample covariance matrix (see e.g. [12, 5]). However the performance of that procedure is still highly dependent of multivariate normality of the bulk of the data [2], or on the data being elliptically contoured. To avoid this dependency, a method to detect outliers in

Carla M. Santos-Pereira
CEMAT, IST and Departamento de Engenharia Civil, Faculdade de Engenharia da Universidade do Porto, Rua Roberto Frias, 4200-465 Porto, Portugal, e-mail: carlasp@fe.up.pt

Ana M. Pires
Departamento de Matemática and CEMAT, Instituto Superior Técnico, Av. Rovisco Pais 1, 1049-001, Lisboa, Portugal, e-mail: apires@math.ist.utl.pt

multivariate data based on clustering and robust estimators was introduced in [14]. A somehow similar method designed to work with non-overlapping clusters was proposed later in [4]. Both [14] and [4] have been referenced recently in relation to robust clustering [8, 3].

Consider a multivariate data set with n observations in p variables. The basic ideas of the method proposed in [14] are described in the following steps.

1. Segment the n points cloud (of perhaps complicated shape) in k smaller subclouds using a partitioning clustering method with the hope that each subcloud (cluster) looks “more normal” than the original cloud.
2. Then apply a simultaneous multivariate outlier detection rule to each cluster by computing Mahalanobis-type distances from all the observations to all the clusters. An observation is considered an outlier if it is an outlier for every cluster. All the observations in a cluster may also be considered outliers if the size of that cluster is small taking into account the number of variables (our proposal is less than $2p+2$, since in that case the covariance matrix estimates are very unreliable).
3. Remove the observations detected in 2 and repeat 1 and 2 until no more observations are detected.
4. The final decision on whether all the observations belonging to a given cluster (not previously removed, that is with size at least $2p+2$) are outliers is based on a table of between clusters Mahalanobis-type distances.

In [14] we presented results from a simulation study with several distributional situations, three clustering methods (k-means, *pam* and *mclust*) and three pairs of location and scatter estimators (classical and two robust), from which it was possible to conclude that for normal data all the methods behave well, whereas for non-normal data the best performance is usually achieved by *mclust*, without large differences between the classical and the robust estimators of location and scatter. A general conclusion from [14] is that the exploratory method proposed for outlier detection works well both under elliptical and non-elliptical data configurations.

The aim of this paper is to propose a criterion for selecting an appropriate number of clusters, k, to use in the above algorithm, and to assess the robustness of that criterion. In the next section we introduce the new criterion, in Section 3 we present the results of a simulation study and in Section 4 we state some conclusions.

2 AIC based criterion

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 and scatter estimators. In [14] it is suggested to try 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 decide after a careful analysis of the results. A less subjective way for choosing k (and also the clustering method and the location and scatter estimators) is to minimize an adapted AIC (see [13]):

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

The full specification of AIC needs $\hat{f}$. This can be either a nonparametric estimate or the density from a parametric model with estimated parameters. The model we consider in this paper is a finite mixture of multivariate normal densities:

$$\hat{f}(\mathbf{x}) = \sum_{j=1}^{k} \frac{n_j}{n_T} f_N(\mathbf{x}; \hat{\mu}_j, \hat{\Sigma}_j), \text{ and } n_T = \sum_{j=1}^{k} n_j, \tag{2}$$

where

$$f_N(\mathbf{x}; \hat{\mu}, \hat{\Sigma}) \text{ is the density of } \mathbf{N}_p(\hat{\mu}, \hat{\Sigma}). \tag{3}$$

The number of components of the mixture (i.e., the number of clusters), k, is limited in practice ($K_{\max}$). As a generic guidance we can take the advise given in [6], that one should have at least 5 to 10 observations per variable. This means to choose $k_{\max}$ somewhere between $0.1n/p$ and $0.2n/p$.

In this paper we assess the robustness of the AIC based criterion (1) for choosing the number of clusters, k. Thisis done by comparing results of simulations with and without outliers, for some non-normal distributional situations described in [14].

3 Simulation Study

In order to evaluate the robustness of this AIC based criterion (1) for choosing the number of clusters, k, we conducted a simulation study with:

- Three clustering methods, k-means, *pam* (partitioning around medoids [7]) and *mclust* (model based clustering for gaussian distributions [1]), each of them with $k = 2,3,4,5,6$. The case k=1, for which the clustering method is irrelevant was also considered.
- Three pairs of location and scatter estimators: classical ($\bar{\mathbf{x}}, \mathbf{S}$) with asymptotic detection limits; RMCD25 [11] and OGK$_{(2)}(0.9)$ [9] with detection limits determined previously by simulation with 10000 normal data sets.
- Four distributional situations:

1. Non-normal ($p = 2$) without outliers, 50 observations from $N_2(\boldsymbol{\mu_1}, \Sigma_1)$, 50 observations from $N_2(\boldsymbol{\mu_2}, \Sigma_2)$ and 50 observations from $N_2(\mathbf{0}, \Sigma_1)$, with $\mu_1 = (0,12)^T$, Σ_1 =diag(1,0.3), $\mu_2 = (1.5,6)^T$ and Σ_2= diag(0.2,9).
2. Non-normal ($p = 2$) with outliers, 150 observations as in the previous case plus 10 outlying observations from $N_2((-2,6)^T, 0.01\mathbf{I})$.
3. Non-normal ($p = 2$) without outliers, 75 observations from $N_2(\mathbf{0}, \Sigma_3)$ and 75 observations from $N_2(\mathbf{0}, \Sigma_4)$, with Σ_3 = diag(1,81) and Σ_4= diag(81,1).

4. Non-normal ($p = 2$) with outliers, 150 observations as in the previous case plus 20 outlying observations from $N_2(\mathbf{10}, 0.1\mathbf{I})$.

We have not considered normal data in this simulation study because we have concluded in [14] that in that case the choice of k is not critical. For each distributional situation one hundred data sets were generated.

In each distributional situation we recorded (in each simulation) the chosen k for each clustering × estimator combination (that is, the value of k minimizing AIC), and also the overall minimizing combination (that is, the specific values of (clustering, estimator, k) which minimizes AIC, at each simulation). Tables 1 to 4 give, for the four distributional situations, respectively, the proportion of simulations for which each k was chosen (within each clustering × estimator combination).

Table 1 Proportion of simulations for which each k was chosen within each clustering × estimator combination (distributional situation 1).

	k	MCD	Classical	OGK
k-means	1	0.00	0.00	0.00
	2	0.01	0.00	0.02
	3	0.28	0.01	0.32
	4	0.26	0.28	0.14
	5	0.15	0.26	0.19
	6	**0.30**	**0.45**	**0.33**
pam	1	0.00	0.00	0.00
	2	0.00	0.00	0.00
	3	0.29	0.02	0.27
	4	0.20	0.23	0.19
	5	0.14	0.23	0.11
	6	**0.37**	**0.52**	**0.43**
mclust	1	0.00	0.00	0.00
	2	0.00	0.00	0.00
	3	**0.61**	**0.48**	**0.66**
	4	0.30	0.28	0.24
	5	0.06	0.15	0.08
	6	0.03	0.09	0.02

The overall minimizing combination was always the *mclust* × classical, which agrees with the simulations in [14] and shows that this choice can be recommended irrespective of the characteristics of the data sets. This conclusion, which may look unexpected can be justified as follows: the algorithm either isolates or removes the outliers, leaving almost exclusively “good” observations, and it is well known that in this case the classical estimators are more efficient.

For the *mclust* cases, the value of k chosen more often is the expected according to the distributional situation (see Fig. 1 and Fig. 2). Note that k must be increased by 1 when the outliers are introduced and this is captured by the AIC criterion.

Table 2 Proportion of simulations for which each k was chosen within each clustering × estimator combination (distributional situation 2)

	k	MCD	Classical	OGK
k-means	1	0.00	0.00	0.00
	2	0.00	0.00	0.00
	3	0.03	0.00	0.00
	4	0.17	0.18	0.12
	5	0.31	0.31	0.33
	6	**0.49**	**0.51**	**0.55**
pam	1	0.00	0.00	0.00
	2	0.00	0.00	0.00
	3	0.00	0.00	0.00
	4	0.27	0.03	0.31
	5	**0.43**	0.44	0.30
	6	0.30	**0.53**	**0.39**
mclust	1	0.00	0.00	0.00
	2	0.00	0.00	0.00
	3	0.13	0.07	0.14
	4	**0.46**	**0.40**	**0.56**
	5	0.27	0.21	0.14
	6	0.14	0.32	0.16

Table 3 Proportion of simulations for which each k was chosen within each clustering × estimator combination (distributional situation 3)

	k	MCD	Classical	OGK
k-means	1	0.00	0.00	0.00
	2	0.00	0.00	0.00
	3	0.04	0.00	0.01
	4	0.16	0.09	0.10
	5	**0.41**	**0.47**	0.38
	6	0.39	0.44	**0.51**
pam	1	0.00	0.00	0.00
	2	0.00	0.00	0.00
	3	0.14	0.02	0.03
	4	0.13	0.04	0.02
	5	0.30	0.36	0.47
	6	**0.43**	**0.58**	**0.48**
mclust	1	0.00	0.00	0.00
	2	**0.68**	**0.46**	**0.56**
	3	0.12	0.12	0.18
	4	0.06	0.16	0.12
	5	0.09	0.11	0.07
	6	0.05	0.15	0.07

Table 4 Proportion of simulations for which each *k* was chosen within each clustering × estimator combination (distributional situation 4)

	k	MCD	Classical	OGK
k-means	1	0.00	0.00	0.00
	2	0.01	0.00	0.03
	3	0.07	0.00	0.02
	4	0.05	0.03	0.04
	5	0.19	0.25	0.25
	6	**0.68**	**0.72**	**0.66**
pam	1	0.00	0.00	0.00
	2	0.00	0.00	0.00
	3	0.02	0.00	0.01
	4	0.02	0.00	0.00
	5	0.16	0.05	0.07
	6	**0.80**	**0.95**	**0.92**
mclust	1	0.00	0.00	0.00
	2	0.02	0.02	0.01
	3	**0.68**	**0.47**	**0.60**
	4	0.17	0.21	0.21
	5	0.08	0.15	0.12
	6	0.05	0.15	0.06

4 Conclusions

The results of the limited simulation study presented in Section 3 show that the adapted AIC criterion (1) for selecting *k* and the clustering method is a useful tool. Moreover, we can also conclude that this criterion is, in association with the present algorithm, robust, since it works well both with and without outliers. An explanation for this robust behavior is that the outliers are either deleted or isolated in their own clusters, before computing the AIC. We this conclude that in this setup there is no

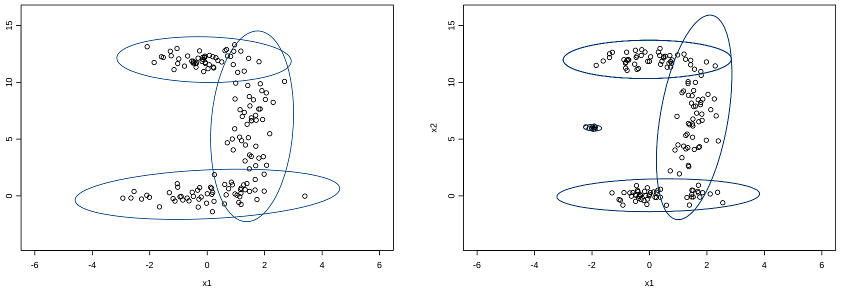

Fig. 1 Distributional situations 1 and 2 with contours.

need to consider other more complicated criteria such as the adapted AIC with M-estimators, introduced in [10].

In spite of the good results of this promising technique, one shall not forget that outlier detection in multivariate data is a very difficult task and will always remain an open problem.

References

1. Banfield, J., Raftery, A.: Model-based Gaussian and non-Gaussian clustering. Biometrics. **49**, 803–822 (1992)
2. Cerioli A.: Multivariate outlier detection with high breakdown estimators. Journal of the American Statistical Association. **105**, 147–156 (2010)
3. Garcia-Escudero, L., Gordaliza, A., Matrán, C. and Mayo-Iscar, A.: A review of robust clustering methods. Advances in Data Analysis and Classification, **4**, 89–109 (2010)
4. Hardin, J. and Rocke, D.: Outlier detection in multiple cluster setting using the minimum covariance determinant estimator. Computational Statistics and Data Analysis, **44**, 625–638 (2004)
5. Hubert, M., Rousseeuw, P.J. and Van Aelst, S.: High-breakdown robust multivariate methods. Statistical Science, **23**, 92–119 (2008)
6. Jain, A. and Chandrasekaran, B.: Dimensionality and sample size considerations in pattern recognition practice. In: Krishnainh, P., Kanal, L. (eds.) Handbook of Statistics, Vol 2, pp. 835–855. North Holland, Amsterdam (1982)
7. Kaufman, L., Rousseeuw, P.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York (1990)
8. Kumar, M. and Orlin, J.B.: Scale-invariant clustering with minimum volume ellipsoids. Computers and Operations Research, 35, 1017–1029 (2008)
9. Maronna, R., Zamar, R: Robust estimates of location and dispersion for high dimensional data sets. Technometrics. **44**, 307–317 (2002)
10. Ronchetti, E.: Robustness aspects of model choice. Statistica Sinica. **7**, 327–338 (1997)

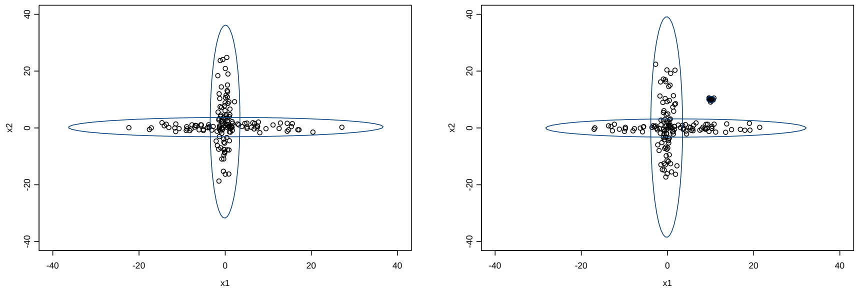

Fig. 2 Distributional situations 3 and 4 with contours.

11. Rousseeuw, P.J.: Multivariate estimation with high breakdown point. In: Grossman,W., Pflug, G., Vincze, I., Werz, W. (eds.) Multivariate estimation with high breakdown point, Vol B, pp. 283–297. Reidel, Dordrecht (1985)
12. Rousseeuw, P.J., von Zomeren, B.C.: Unmasking multivariate outliers and leverage points. Journal of the American Statistical Association. **85**, 633–639 (1990)
13. Sakamoto, Y., Ishiguro, M., Kitagawa, G.: Akaike Information Criterion Statistics. Kluwer, New York (1988)
14. Santos-Pereira, C., Pires, A.: Detection of outliers in multivariate data: a method based on clustering and robust estimators. In: Härdle,W., Rönz, B. (eds.) Computational Statistics, pp. 291-296. Physica-Verlag, Heidelberg (2002)