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

[2] Department of Mathematics and CEMAT, IST, Technical University of Lisbon.

email: carlasp@fe.up.pt, apires@math.ist.utl.pt

Methodology

The performance of the usual methods to identify outliers is highly dependent of multivariate normality of the bulk of the data, or on the data being elliptically contoured. To reduce this dependency, a method to detect outliers in multivariate data based on clustering and robust estimators was introduced in Santos-Pereira and Pires (2002). The basic ideas of the method can be described in the steps bellow. Consider a multivariate data set with n observations in p variables.

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 relative size of that cluster is small (our proposal is less than $2p + 2$, since for smaller number of observations 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 greater than $2p + 1$) are outliers is based on a table of between clusters Mahalanobis-type distances.

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 (classical and two robust). After this simulation study we concluded that for normal data all the methods behave well, except for the masking with the classical Mahalanobis distance (which is not surprising). 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. Generally we have concluded that the exploratory method proposed for outlier detection works well both under elliptical and non-elliptical data configurations.

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-scatter estimators. In Santos-Pereira and Pires (2002) it is 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 decide after a careful analysis of the results. A less subjective way for choosing k (and also the clustering method and the location-scale estimators) is to minimize an adapted AIC (see Sakamoto et al. (1988)):

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

with

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

where

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

An adapted AIC, with M-estimators, was also introduced by 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.

Simulation study

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

- Three clustering methods k-means, *pam* (partitioning around medoids, from Kaufman and Rousseeuw, 1990) and *mclust* (model based clustering for gaussian distributions, from Banfield and Raftery, 1992), 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-scatter estimators: classical $(\bar{\mathrm{x}}, \mathrm{S})$ with asymptotic detection limits; RMCD25 (Rousseeuw, 1985) and OGK$_{(2)}(0.9)$ (Maronna and Zamar, 2002) 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 $N_2(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, 50 observations from $N_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ and 50 observations from $N_2(\mathbf{0}, \boldsymbol{\Sigma}_1)$, with $\boldsymbol{\mu}_1 = (0, 12)^T$, $\boldsymbol{\Sigma}_1$ =diag(1,0.3), $\boldsymbol{\mu}_2 = (1.5, 6)^T$ and $\boldsymbol{\Sigma}_2$= 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 $N_2(\mathbf{0}, \boldsymbol{\Sigma}_3)$ and 75 observations from $N_2(\mathbf{0}, \boldsymbol{\Sigma}_4)$, with $\boldsymbol{\Sigma}_3$ = diag(1,81) and $\boldsymbol{\Sigma}_4$= diag(81,1).
  4. Non-normal ($p = 2$) with outliers, 150 observations as in the previous case plus 20 outlying observations from $N_2(\mathbf{10}, 0.1\mathbf{I})$.

In each combination we recorded, for each clustering × estimator combination, the chosen k, and also the overall minimizing combination (clustering × estimator × k). Tables 1 to 4 give, for the 4 distributional situations respectively, the proportion of simulations for which each k was chosen (within each clustering × estimator combination).
The overall minimizing combination was always the *mclust* × classical.
For the *mclust* cases, the value of k chosen more often is the expected according to the distributional situation (Figures 1 and 2). Note that k must be increased by 1 when the outliers are introduced and this is captured by the AIC criterion.

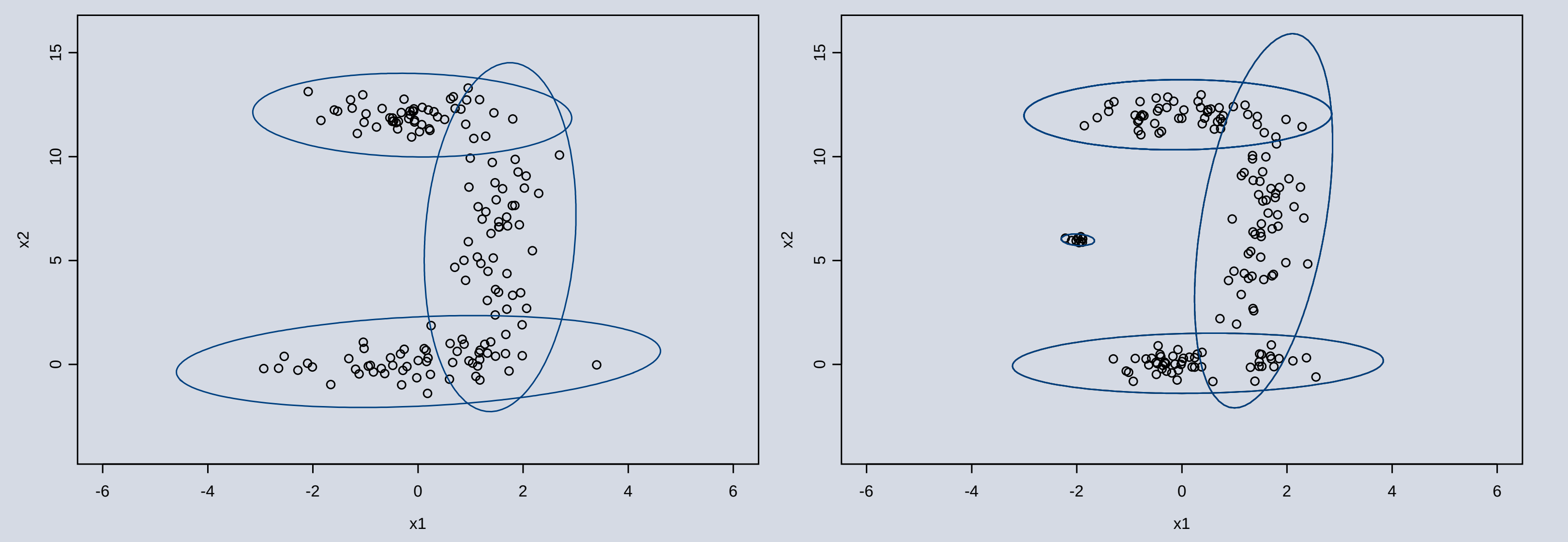

Figure 1: Distributional situations 1 and 2 with detection contours.

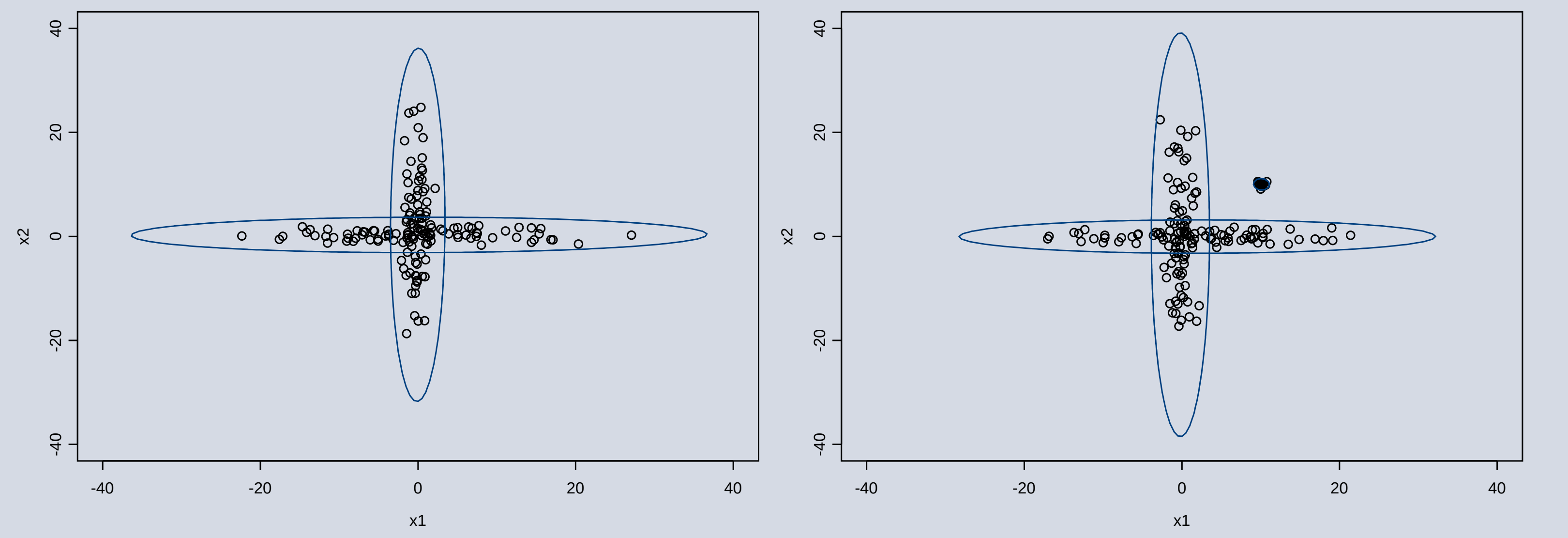

Figure 2: Distributional situations 3 and 4 with detection contours.

Proportion of simulations for which each k was chosen within each clustering × estimator combination:

Table 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

Table 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:

	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:

	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

Conclusions

- The method described for outlier detection works well both under elliptical and non-elliptical data configurations.
- The adapted AIC is a useful tool for selecting k and the clustering method. Due to the cleaning step of the original method, the proposed AIC is robust (the outliers are either deleted or isolated in their own clusters).
- However, one shall not forget that outlier detection in multivariate data is a very difficult task and will always remain an open problem.

References

BANFIELD, J. and RAFTERY, A. (1992): Model-based Gaussian and non-Gaussian clustering. *Biometrics, 49, 803-822.*

KAUFMAN, L. and ROUSSEEUW, P. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. New York: Wiley.

MARONNA, R. and ZAMAR, R. (2002). Robust estimates of location and dispersion for high dimensional data sets. *Technometrics, 44, 307-317.*

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

ROUSSEEUW, P. (1985). Multivariate estimation with high breakdown point. In: W. Grossman, G. Pflug, I. Vincze and W. Werz: *Mathematical Statistics and Applications, Vol B.* Dordrecht, Reidel, 283-297.

SAKAMOTO, Y. and 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.): *Computational Statistics.* Physica-Verlag, Heidelberg, 291-296.