 © Physica-Verlag/Springer 2004

A GENERAL PARTITION CLUSTER ALGORITHM

Daniel Peña, Julio Rodríguez and George C. Tiao

Key words: Predictive distribution, robust estimation, SAR procedure.
COMPSTAT 2004 section: Clustering.

Abstract: A new cluster algorithm based on the SAR procedure proposed by Peña and Tiao (2002) is presented. The method splits the data into more homogeneous groups by putting together observations which have the same sensitivity to the deletion of extreme points in the sample. As the sample is always split by this method the second stage is to check if observations outside each group can be recombined one by one into the groups by using the distance implied by the model. The performance of this algorithm is compared to some well known cluster methods.

1 Introduction

Finding groups in data is a key activity in many scientific fields. Gordon (1999) is a good general reference. Classical Partition and Hierarchical algorithms have been very useful in many problems but they have some four main limitations. First, the criteria used are not affine equivariant and therefore the results obtained depend on the changes of scale and/or rotation applied to the data. Second, the usual heterogeneity measures based on the Euclidian metric do not work well for highly correlated observations forming elliptical clusters or when the clusters overlap. Third, we have to specify the number of clusters or decide about the criteria for choosing them. Fourth, there is no general procedure to deal with outliers. Some advances have been made to solve these problems, see Cuesta-Albertos, Gordaliza and Matrán (1997), Cuevas et al (2000) and Tibshirani et al (2001).

An alternative approach to cluster is to fit mixture models. This idea has been explored both from the classic and Bayesian point of view. Banfield and Raftery (1993) and DasGupta and Raftery (1998) have proposed a model-based approach to clustering which finds an initial solution by hierarchical clustering and then assumes a mixture of normals model and uses the EM algorithm to estimate the parameters. A clear advantage of fitting normal mixtures is that the implied distance is the Mahalanobis distance, which is affine equivariant. From the Bayesian point of view the parameters of the mixture are estimated by Markov Chain Monte Carlo methods and several procedures have been proposed to allow for an unknown number of components in the mixture, see Richarson and Green (1997) and Stephens (2000). A promising approach to cluster analysis, that can avoid the curse of dimensionality, is projection pursuit, where low-dimensional projections of the multivariate data are used to provide the most interesting views of the

full-dimensional data. Peña and Prieto (2001) have proposed an algorithm where the data is projected on the directions of maximum heterogeneity defined as those directions in which the kurtosis coefficient of the projected data is maximized or minimized. Then they used the spacings to search for clusters on the univariate variables obtained by these projections.

Finally, Peña and Tiao (2003) propose the SAR (split and recombine) procedure for detecting heterogeneity in a sample with respect to a given model. This procedure is general, affine equivariant, does not require to specify a priori the number of clusters and it is well suited for finding the components in a mixture of models. The idea of the procedure is first to split the sample into more homogeneous groups and second recombine the observations one by one in order to form homogeneous clusters. The SAR procedure has two important properties, that are not shared by many of the most often used cluster algorithms, (i) it does not require an initial starting point, (ii) each homogeneous group is obtained independently from the others, so that each group does not compete with the others to incorporate an observation. The first property implies that the algorithm we propose can be used as a first solution for any other cluster algorithm, the second, that the procedure may work well even if the groups are not well separated. This paper analyzes the application of the SAR procedure to cluster analysis and it is organized as follows. Section 2 presents the main ideas of the procedure. Section 3 compares it in a Monte Carlo study to Mclust (Fraley and Raftery, 1999), k-means, pam (Struyf, Hubert and Rousseeuw, 1997) and Kpp (Peña and Prieto, 2001).

2 The SAR procedure

Suppose we define a measure $H(x, X)$ of the heterogeneity between an observation, x, and a set of data, X. We are going to use this measure to split the sample iteratively into homogeneous groups and to recombine observations into the groups. We assume that the heterogeneity measure $H(x, X)$ is equivariant, that is invariant to linear transformations, and is coherent with the assumed model. As the true structure of the data is unknown, we start the process by assuming that the data is homogeneous, and have been generated by a normal distribution, $N_p(\boldsymbol{\mu}, \mathbf{V})$. Then we propose a heterogeneity measure based on out of sample prediction as follows. The predictive distribution for a new observation $\mathbf{x}_f$ generated by a normal model using a Jeffrey's prior $p(\boldsymbol{\mu}, \mathbf{V}) \propto |\mathbf{V}|^{-(p+1)/2}$ is (see for instance, Box and Tiao, 1973) $p(\mathbf{x}_f, \mathbf{X}) \propto \left(1 + \frac{Q_f}{n-p}\right)^{-n/2}$, where $Q_f = \frac{n}{n+1}(\mathbf{x}_f - \bar{\mathbf{x}})'\hat{\mathbf{V}}^{-1}(\mathbf{x}_f - \bar{\mathbf{x}})$ and $\bar{\mathbf{x}}$ is the sample mean and $\hat{\mathbf{V}}$ the sample covariance matrix, given by $\hat{\mathbf{V}} = (\mathbf{X} - \mathbf{1}\bar{\mathbf{x}})'(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}})/(n-p)$. Following Peña and Tiao (2003) we will use as measure of heterogeneity of a data $\mathbf{x}_i$ with respect to a group $\mathbf{X}_{(i)}$ which

does not contain this observation, the standardized predictive value given by

$$H(\mathbf{x}_i, \mathbf{X}_{(i)}) = -2 \ln \left\{ \frac{p(\mathbf{x}_i|\mathbf{X}_{(i)})}{p(\hat{\mathbf{x}}_{i(i)}|\mathbf{X}_{(i)})} \right\} = (n-1) \ln \left\{ 1 + \frac{Q_{i(i)}}{(n-1)-p} \right\}, \quad (1)$$

where $Q_{i(i)} = \frac{n-1}{n}(\mathbf{x}_i - \bar{\mathbf{x}}_{(i)})' \hat{\mathbf{V}}_{(i)}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_{(i)})$, and $\hat{\mathbf{V}}_{(i)}$ and $\bar{\mathbf{x}}_{(i)}$ are the covariance matrix and the mean computed using the sample $\mathbf{X}_{(i)}$ without the case i*th*. Note that $H(\mathbf{x}_i, \mathbf{X}_{(i)})$ is a monotonic function of the Mahalanobis distance $Q_{i(i)}$, which is usually used to check the heterogeneity of a point $\mathbf{x}_i$ with respect to the sample $\mathbf{X}_{(i)}$.

The splitting of the sample is made as follows. For each observation, $\mathbf{x}_i$, we define the discriminator of this point as the observation which, when deleted from the sample, makes the point $\mathbf{x}_i$ as heterogeneous as possible with the rest of the data. The discriminator of $\mathbf{x}_i$ is the point $\mathbf{x}_j$ if

$$\mathbf{x}_j = \arg\max_{x_k} H(\mathbf{x}_i, \mathbf{X}_{(ik)}) = \arg\max_{x_k} (\mathbf{x}_i - \bar{\mathbf{x}}_{(ik)})' \hat{\mathbf{V}}_{(ik)}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_{(ik)}).$$

Each sample point must have a unique discriminator, but several sample points may share the same discriminator. It can be proved (see Peña, Rodriguez and Tiao, 2004) that the discriminators are members of the convex hull of the sample. That is, a discriminator must be an extreme point. An intuitive procedure to split the sample into groups is to put together observations with share the same discriminators, as they are affected in the same way to modifications of the sample by deleting some extreme values. It is obvious that if two observations are identical they will have the same discriminator and if they are close they also will have the same discriminator. The number of points in the sample which share the same discriminator is called the order of the discriminator. We consider as special points discriminators of order larger than K, where $K = f(p, n)$ and we will put them in a special group of extreme observations. However, discriminators of order smaller than K are considered as usual points and are assigned to the group defined by all the observations that share a common discriminator. We need to define the minimum size of a set of data to be considered as a group. We will say that we have a group if we could compute the mean and covariance matrix of the group and, therefore, the minimum group size must be $n_0 = p + h$, where $h > 0$, and p is the number of variables. Usually $h = f(p, n)$ and in the examples we have taken $h = \log(n - p)$. In the procedure which follows we have considered as special points to those discriminators of order larger that K, where $K = p + h - 1$. This value seems to work well in the simulations we have made. Based on these considerations the sample is split as follows: 1) Observations which have the same discriminator are put in the same group, the discriminator is only included in the group if it has order smaller than K; 2) Discriminators of order bigger that K are allocated to a specific group of isolated points; 3) if two groups formed by the previous rules have any observation in common the two groups are joined into one group. This three

rules split the sample into more homogeneous groups. Each group is now considered as a new sample and the three rules are applied again until splitting further the sample will lead to isolated points because the groups obtained are all of them of size smaller than the minimum group size n_0. A group of data is called basic group if when split will lead to subgroups of size smaller than the minimum size, $p + h$.

When the sample cannot be split further the recombining process is applied starting from any of the basic groups obtained. The recombining process is the one suggested by Peña and Tiao (2003). Each group is enlarged by incorporating observations one by one. For a given group, we begin by testing the observation outside the group which is the closest to the group in terms of the measure $H(y_f, X_g)$, where y_f is the observation outside the group formed by data X_g. If $H(y_f, X_g)$ is smaller than some cut-off value, this observation is incorporated into the group and the process of testing the closest observation to the group is repeated for the enlarged group. The enlarging process will continue until either the threshold is crossed or the entire sample is included. A similar idea of recombining points has been used for robust estimation (see for instance, Atkinson, 1994). We may have one of the three possible cases. First, the enlarging of all the basic groups leads to the same group which include all the observations apart from some outliers. Then we have a homogeneous sample with some isolated outliers and the procedure ends. Second, the enlarging of the basic groups leads to a partition of the sample into disjoint groups and we conclude we have some groups in the data and again the procedure ends. Third, we obtain more than a possible solution because the partition obtained is different when starting from different basic groups. Then we have more than one possible solution and the final solutions found are called possible data configurations, PDC. The selection among them is made by a model selection criterion.

3 Monte Carlo results

The properties of the algorithm have been studied in a Monte Carlo experiment, similar to the one used by Peña and Prieto (2001) to illustrate the behavior of their cluster procedure. Sets of $10 \times p \times k$ random observations in dimension $p = 2, 4, 8$ have been generated from a mixture of $k = 2, 4$ components of a multivariate distributions. In all data sets the number of observations from each distribution has been determined randomly, but ensuring that each cluster contains a minimum of $p+1$ observations. The mean for each distribution is chosen at random from the multivariate normal distribution $N_p(\mathbf{0}, f\mathbf{I})$. The factor f (see Table 1) is selected to be as small as possible while ensuring that the probability of overlapping between groups is roughly equal to 0.01. We generated data sets in six different scenarios.

a) Mixture of k multivariate normal distributions. In each group the covariance matrix is generated as $\mathbf{S} = \mathbf{UDU}'$, from a random orthogonal

matrix $\mathbf{U}$ and a diagonal matrix $\mathbf{D}$ with entries generated from a uniform distribution (a1): $[10^{-3}, 5\sqrt{p}]$, so that the covariance matrices are well conditioned, and (a2): $[10^{-3}, 10\sqrt{p}]$, so that the covariance matrices are ill-conditioned.

b) Mixture of k multivariate uniform distributions with (b1) covariance generated as (a1) and (b2) covariance generated as (a2).

c) Mixture of k multivariate normal distributions generated as indicated in scenario a1), but 10% of the data are outliers (c1): generated by $N_p(\mathbf{0}, f\mathbf{I})$ and (c2): for each cluster in the data, 10% of its observations have been generated as a group of outliers at a distance $4\chi^2_{p,0.99}$ in a group along a random direction, and a single outlier along another random direction.

To provide better understanding of the behavior of the new procedure, in each table we compare the proposed method with Kpp, k-means, Mclust and the `pam` algorithm. The Mclust algorithm has been run with the function 'EMclust' with models EI, VI, EEE, VVV, EEV and VEV and number of cluster between 1 to 8 and the final configuration is selected by the BIC (see Fraley and Raftery, 1999, for a description of different models used in the function 'EMclust'). The rule to select the number of clusters in the algorithm `pam` is the maximum of the silhouette statistic for $k = 1, \ldots, 8$ and in k-means the stopping rule used is the one proposed by Calinski and Harabasz.

Table 1 gives the average percentage of observations which have been labeled incorrectly in scenarios a1) and a2), obtained from 200 replications for each value in the same data sets in all procedures. In scenario a1) the SAR procedure has the best performance, and Kpp and Mclust are second having a similar behavior. In the scenario a2) when the covariance matrix is ill-conditioned, the SAR procedure is again the best followed by Kpp and Mclust. This result is quite consistent as the SAR procedure is the best in eight out of the twelve comparison included in the two scenarios of Table 1 and in the four cases in which it is not the best it is not far from the best one. The k-means and `pam` show a poor result.

Table 2 shows the outcome for scenarios b1) and b2) where we analyze the same structure that in scenarios a1) and a2) but now using mixtures of uniform distributions. Table 2 shows the percentages of mislabeled observations for both scenarios b1) and b2). The behavior of the SAR procedure is again the best as an average and the best in ten of the twelve cases. The second best behavior corresponds to Kpp, that is better than Mclust in eleven out of the twelve cases.

A final simulation study has been conducted (see Table 3) to determine the behavior of the methods in the presence of outliers. Scenarios c1) and c2) contain 10% of data contaminated by first, a non concentrate contamination, and second, a concentrated contamination defined in scenario c). The

a1) Covariance matrices well conditioned							
p	k	f	SAR	Kpp	k-means	Mclust	`pam`
2	2	55	**1.65**	7.33	45.35	16.73	34.98
	4	140	1.29	**0.95**	24.90	1.54	1.86
4	2	14	**4.83**	9.90	47.15	12.38	32.11
	4	20	**5.58**	9.39	27.20	6.75	10.76
8	2	12	15.43	13.13	43.29	**12.28**	55.61
	4	18	7.52	12.58	15.81	**3.75**	14.42
Average			**6.05**	8.88	33.95	8.90	24.96
a2) Covariance matrices ill-conditioned							
p	k	f	SAR	Kpp	k-means	Mclust	`pam`
2	2	55	**1.58**	9.38	46.38	14.23	33.95
	4	140	1.00	0.61	25.14	**0.60**	1.83
4	2	14	**0.99**	4.96	48.54	11.64	32.89
	4	20	**1.39**	5.07	30.99	6.55	5.38
8	2	12	**0.64**	5.19	44.83	0.66	50.94
	4	18	**0.87**	6.01	22.92	4.36	11.01
Average			**1.08**	5.20	36.47	6.34	22.66

Table 1: Percentages of mislabeled observations for the SAR, the Kpp, the k-means, the Mclust and the `pam` procedures. Normal observations with: (a1) covariance matrices well conditioned, (a2) covariance matrices ill-conditioned. The best method in each case is indicated in boldface.

criterion to obtain the mislabeled observation is based only in the 90% of observations not contaminated. Table 3 shows the percentage of mislabeled observations for the scenarios c1) and c2). The maximum number of clusters k have been increase to ten in the algorithms k-means, Mclust and `pam` so that the concentrated contamination can be considered as isolated clusters. In the scenario c1) the best methods, as an average, are, with very small difference, the `pam` algorithm and the SAR procedure. However, for concentrated contamination, scenario c2), the SAR procedure is again clearly the best followed by Kpp. As a summary of this Monte Carlo study we may conclude that the SAR procedure has the smallest error classification rate in 22 out of the 36 situations considered and the best average number of mislabeled observations in 5 scenarios out of the six considered. The only scenario in which the SAR is not the best is in scenario c1) but the difference with respect to the best method, `pam` , is very small: misclassification percentage of 6.4% versus 6.32% for pam. The Kpp is the second best in five out of the six scenarios. Ordering the methods for average classification errors in all the scenarios from better to worse, the order would be: SAR, Kpp, Mclust, `pam` and k-means.

b1) Covariance matrices well conditioned							
p	k	f	SAR	Kpp	k-means	Mclust	`pam`
2	2	55	**0.45**	11.53	51.40	21.08	44.75
	4	140	**0.58**	0.38	29.25	0.84	1.16
4	2	14	**0.85**	4.81	51.71	12.48	51.41
	4	20	**1.58**	4.33	33.15	9.11	7.68
8	2	12	6.24	**5.45**	41.83	7.38	60.80
	4	18	**2.33**	4.93	20.07	5.58	16.93
Average			**2.00**	5.24	37.90	9.41	30.46
b2) Covariance matrices ill-conditioned							
p	k	f	SAR	Kpp	k-means	Mclust	`pam`
2	2	55	**1.55**	11.78	48.65	20.53	41.95
	4	140	**0.56**	0.99	34.30	1.75	2.06
4	2	14	**0.79**	4.06	53.23	6.00	46.45
	4	20	**0.38**	3.13	34.39	7.54	7.28
8	2	12	0.34	5.76	45.96	**0.00**	62.13
	4	18	**0.46**	4.21	27.32	4.74	12.61
Average			**0.68**	4.99	40.64	6.76	28.75

Table 2: Percentages of mislabeled observations for the SAR, the Kpp, the k-means, the Mclust and the `pam` procedures. Uniform observations with: (b1) covariance matrices well conditioned, (b2) covariance matrices ill-conditioned.

References

[1] Atkinson, A. C. (1994). *Fast very robust methods for detection of multiple outliers.* Journal of the American Statistical Association **89**, 1329-1339.

[2] Box, G.E.P. and Tiao, G.C. (1973), *Bayesian Inference in Statistical Analysis*, Addison-Wesley.

[3] Banfield, J.D., and Raftery, A. (1993). *Model-Based Gaussian and Non-Gaussian Clustering.* Biometrics **49**, 803-821.

[4] Cuesta-Albertos, J. A., Gordaliza, A. C. and Matrán, C. (1997). *Trimmed k-means: An attempt to robustify quantizers.* The Annals of Statistics **25**, 553-576.

[5] Cuevas, A. , Febrero, M. and Fraiman, R. (2000). *Estimating the number of clusters.* Canadian Journal of Statistics **28**, 367-382.

[6] Dasgupta, A., and Raftery, A. E. (1998). *Detecting Features in Spatial Point Processes With Clutter via Model-Based Clustering.* Journal of the American Statistical Association **93**, 294-302.

c1) Non concentrated contaminations							
p	k	f	SAR	Kpp	k-means	Mclust	`pam`
2	2	55	1.25	**0.68**	3.00	6.47	0.69
2	4	140	**0.83**	1.30	12.31	3.50	2.85
4	2	14	8.58	9.46	14.55	**6.71**	7.21
4	4	20	5.66	11.89	22.64	**5.27**	6.13
8	2	12	12.64	14.48	16.88	**12.58**	16.46
8	4	18	9.47	16.67	44.08	6.78	**4.59**
Average			6.40	9.08	18.91	6.89	**6.32**
c2) Concentrated contaminations							
p	k	f	SAR	Kpp	k-means	Mclust	`pam`
2	2	55	**0.98**	4.03	26.25	12.61	17.50
2	4	140	**0.40**	0.65	12.88	0.49	2.04
4	2	14	**3.58**	6.29	35.46	17.90	28.46
4	4	20	**3.21**	10.01	17.69	15.47	7.50
8	2	12	15.03	**13.41**	38.66	23.42	53.08
8	4	18	8.15	13.73	17.72	**6.93**	14.71
Average			**5.22**	8.02	24.78	12.80	20.55

Table 3: Percentages of mislabeled observations for the SAR, the Kpp, the k-means, the Mclust and the `pam` procedures. Normal observations with 10% the outliers: (c1) non concentrated contaminations, (c2) concentrated contaminations.

[7] Fraley, C., and Raftery, A.E. (1999). *MCLUST: Software for Model-Based Cluster Analysis*. Journal of Classification **16**, 297-306.

[8] Gordon, A,. (1999). *Classification*. 2nd edn. London: Chapman and Hall-CRC.

[9] Peña, D., and Tiao, G.C. (2003). *The SAR Procedure: A Diagnostic Analysis of Heterogeneous Data*. (manuscript submitted for publication).

[10] Peña, D., Rodriguez, J. and Tiao, G.C. (2004). *Cluster Analysis by the SAR Procedure* (manuscript submitted for publication).

[11] Peña, D. and Prieto, J. (2001). *Cluster Identification using Projections*. Journal of the American Statistical Association **96**, 1433-1445.

[12] Richarson, S., and Green, P.J. (1997). *On Bayesian analysis of mixtures with an unknown number of components*. Journal of the Royal Statistical Society B **59**, 731-758.

[13] Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust Regression and Outlier detection*, New York: John Wiley.

[14] Stephens, M. (2000). *Bayesian Analysis of Mixture Models with an Unknown Number of Components–An Alternative to Reversible Jump Methods*. The Annals of Statistics **28**, 40-74.

[15] Stuyf, A. Hubert, M. and Rousseeuw, P. J. (1997). *Integrating robust clustering techniques in S-PLUS*. Computational statistics and data analysis **26**, 17-37.

[16] Tibshirani, R., Walther, G. and Hastie, T. (2001). *Estimating the number of clusters in a data set via the gap statistic*. Journal of the Royal Statistical Society B **63**, 411-423.

Address: Daniel Peña: Departamento de Estadística, Universidad Carlos III de Madrid, Spain. Julio Rodríguez: Laboratorio de Estadística, Universidad Politécnica de Madrid, Spain. George C. Tiao: Graduate School of Business, University of Chicago, USA.

E-mail: dpena@est-econ.uc3m.es