MATEMATIKA, 2014, Volume 30, Number 1a, 59-70

Doubtful Outliers with Robust Regression of an M-estimator In Cluster Analysis

[1]Muhamad Alias Md Jedi, [2]Robiah Adnan, and [3]Sayed Ehsan Saffari

[1,2]Universiti Teknologi Malaysia
Department of Mathematics, Faculty of Science, UTM,81310, Skudai Johor
[3]Sabzevar University of Medical Sciences,
Centre of Education, Iran
Email: [1]muhamad_jedi@yahoo.com, [1]robiaha@utm.my, [2]ehsanreiki@yahoo.com

Abstract Doubtful outlier between clusters may show some meaningful data. In some cases for example it may explain the potential or the unique pattern within the data. However, there is still no further analysis to show how this data (doubtful) connected to one another. In the simulation, we use different threshold values to detect how many doubtful outliers exist between clusters. For these cases we will use 1%, 5%, 10%, 15% and 20% of threshold values. For real data, we fit a linear model using an M estimator with the existences of doubtful data with 10% threshold value. The objective is to determine if doubtful data affect the parameter of M estimator. By comparing using linear model with the deletion of outliers we can conclude that doubtful outlier affect the parameter of M estimator make it less robust towards doubtful outliers in the present of 10% of threshold value.

Keywords Doubtful outlier, Cluster Analysis, Robust Regression, M estimator

2010 Mathematics Subject Classification 46N60, 92B99.

1 Introduction

The presence of certain amount of outlying observations is common in many practical statistical applications. In cluster analysis methods those outlying observations may lead to unsatisfactory clustering results. Methods for cluster analysis are basically aimed at detecting homogeneity with large heterogeneity among them. For non-robust methods, clustering may be heavily influenced by even a small fraction of outlying data. Thus application of robust clustering methods is very appropriate. Certain technique related with cluster analysis and

robust methods may lead for the interest of robust clustering techniques. For instance, robust clustering method can be used to handle data of highly concentrated outliers. The TCLUST package in R statistical computing implements different robust non-hierarchical clustering. In TCLUST, trimming plays a key role as it allows the removal of a fraction, α of the most outlying data and therefore the influence of outlying data can be avoided and robustness naturally arises. This trimming approach to clustering has been introduced in Cuesta-Abertos et.al (1997), Gallegos (2002), Gallegos and Ritter (2005) and Garcia-Escudero et.al. (2008) [1,2,6]. In TCLUST analysis, selecting the number of groups and trimming size perhaps the most complex problems when applying cluster analysis. In some cases, researcher might have an idea on the number of clusters or the trimming proportion of true contamination level.

In TCLUST, some additional exploratory graphical tool can be applied in order to evaluate the quality of the cluster assignment and the trimming decision. This is done by applying the function of discriminant factor $DF_{(i)}$. The use of this type of discriminant factors was suggested in Van Aelst *et al*. (2006) [7] in a cluster problem without trimming. Silhouette plots (Rousseeuw, 1987) can be used for summarizing the obtained order discriminant factors. Clusters in silhouette plot with indicated value of discriminant factor would suggest the existence of well-determined cluster. The most doubtful assignment with certain degree of discriminant value will allow us to determine if the cluster is well defined or not. Besides that, an argument may arise whether the doubtful assignment came from the outlier or from the some doubtful trimmed observation. Furthermore, the existence of outlying data may strongly influence the cluster assignment. Therefore, the outlier detection in cluster analysis is required in order to make the cluster assignment to be well classified according to its own cluster. This may help the data not to be in the overlapping cluster. Good clustering explained the data to describe real life condition especially applied analysis for example in medical data namely cancer and blood transfusion for blood group. Therefore the outlier detection methods in cluster analysis are required.

2 Method and Data

2.1 Past Studies

In clustering analysis, trimming the data simply means removing the outlying observations and do not intend to fit all of them. Researchers sometimes view isolated data or small groups of outliers as cluster. This is quite logical because cluster is obviously heterogeneous to other data structures. In model-based clustering, the methods intend to find clusters formed around different types of objects. For example these objects were initially centers of cluster, the number of clusters and the constraints on scatter matrices of the clusters. Measuring

robustness in cluster analysis is a very difficult task because the measure, for example the location scatter estimation or regression model sometimes may be insufficient. In most cases we do not know the true proportion of outliers that exist in the data. However by trial and error the trimming proportion in cluster analysis for example in robust k-means, partition around mediod (PAM), and TCLUST may give rise to a close estimate of the percentages of true outliers that exists. Gallegos (2002) and Gallegos and Ritter (2005) [2] suggest that we may assume the doubtful assignment in cluster analysis as indication of outliers or bad trimming proportion (false trimming). The question is, if we know exactly the proportion of outliers and but still can detect the doubtful assignment, what does this tell us? Can we use doubtful assignment as a tool to detect outliers?

2.2 Strength of Cluster Assignments

For a given TCLUST clustering solution, we now introduce some confirmatory graphical tools that will help us to evaluate the quality of the cluster assignments and the strength of trimming decisions. Let us consider an optimal solution $\hat{R} = \left\{ \hat{R}_0, \hat{R}_1, ..., \hat{R}_k \right\}, \hat{q} = \left\{ \hat{q}_0, ..., \hat{q}_k \right\}$ and $\hat{p} = \left\{ \hat{p}_1, ..., \hat{p}_k \right\}$ returned by the TCLUST for some *k*, a and values. Given an observation x_i, let us define

$$D_j\left(x_i : \hat{q}, \hat{p}\right) = \hat{p}_j f\left(x_i, \hat{q}_j\right) \text{ for } j = 1,2,...,k \tag{1}$$

$R_i\left(i = 0,1,...,k\right)$ is a set of indices of x_i, $\hat{q}_i\left(i = 0,1,...,k\right)$ is a covariance matrices of x_i and $p_i\left(i = 0,1,...,k\right)$ is a weight of each cluster. The values in equation (1) can be sorted as $D_{(1)}\left(x_i; \hat{q}, \hat{p}\right) £ ... £ D_{(k)}\left(x_i; \hat{q}, \hat{p}\right)$. A non-trimmed observation x_i would be assigned to group *j* if $D_{(j)}\left(x_i; \hat{q}, \hat{p}\right) £ ... £_{(k)}\left(x_i; \hat{q}, \hat{p}\right)$ (Garcia et.al. 2008). Therefore, we can measure the strength of the assignment of x_i to group *j* by analyzing the size of $D_{(k)}\left(x_i; \hat{q}, \hat{p}\right)$ with respect to the second largest value $D_{(k-1)}\left(x_i; \hat{q}, \hat{p}\right)$. We thus define the discriminant factors $DF_{(i)}$'s as

$$DF_{(i)} = \log\left(D_{(k-1)}\left(x_i; \hat{q}, \hat{p}\right) / D_{(k)}\left(x_i; \hat{q}, \hat{p}\right)\right) \tag{2}$$

The idea of using posterior probabilities (equation 1) to measure assignment strengths is not new in clustering. The use of these $DF_{(i)}$'s was already suggested in Van Aelst et.al. 2006. Let us consider $d_{(i)} = D_{(k)}\left(x_i;\hat{q},\hat{p}\right)$ for all the observations in the sample and sort them in $d_{(1)} \leq \ldots \leq d_{(n)}$. The TCLUST trims off a proportion a, of observations with smallest assignment strengths. In other words, the trimmed observations are $\hat{R}_0 = \left\{ i \in \{1,\ldots,n\} : d_{(i)} \leq d_{\lceil na \rceil} \right\}$. Therefore, we can quantify the certainty of the trimming decision for the trimmed observation x_i through

$$DF_{(i)} = \log\left(d_{\lceil na \rceil+1} \Big/ D_{(k)}\left(x_i;\hat{q},\hat{p}\right) \right) \tag{3}$$

Large values of $DF_{(i)}$ for example $DF_{(i)} > \log(y)$ can be explain where *y* is a comparison value to indicate doubtful assignments or trimming decisions. Of course, this $\log(y)$ threshold value is a subjective choice. With this in mind, different summaries of the discriminant factors may be obtained. For instance silhouette plot (Rousseeuw 1987) can be made that can indicate the value of $DF_{(i)}$. The larger value tells that the obtained solution includes some groups having not enough strength (the existences of doubtful data is high). Moreover we can also plot observation having large $DF_{(i)}$ values and these observations correspond to doubtful assignment or trimming decisions. In graphical view, the observations in the frontier between clusters that appear when splitting one of the main groups are labeled as doubtful assignments that are referred as data in overlapping clusters. Some trimmed observations in the boundaries of the main groups may be considered as doubtfully trimmed ones.

2.3 The Spurious Outliers Model

The discussion about simulated examples goes back to Gallegos (2002) and Gallegos and Ritter (2005) [1,2] who proposed the spurious outliers model. This model is defined through likelihoods in equation (4).

$$\left[\prod_{j=1}^{k} \prod_{i \in R_j} f\left(x_i;\mu_j,\Sigma_j\right) \right] \left[\prod_{i \in R_0} g_i\left(x_i\right) \right] \tag{4}$$

where $f\left(x;\mu_j,\Sigma_j\right)$ is a probability density function of the p-variate normal distribution with mean μ and covariance matrix Σ. From (equation 4) $\{R_0,...,R_k\}$ being partitioned to the set of indices $\{1,2,...,n\}$ such that $R_0=\lceil n\alpha\rceil$. R_0 are the indices of non regular observation generated by probability function of g_i. The search of $\{R_0,...,R_k\}$, vector μ_j and matrices Σ_j maximizing (equation 4) can be simplified such that equation 5

$$\sum_{j=1}^{k}\sum_{i\in R_j}\log f\left(x_i;\mu_j,\Sigma_j\right) \tag{5}$$

The maximum of equation 5 implicitly assumes equal cluster weight and alternatively cluster weight can be considered as $\pi_j \in [0,1]$ maximizing

$$\sum_{j=1}^{k}\sum_{i\in R_j}\left(\log\pi_j+\log f\left(x_i;\mu_j,\Sigma_j\right)\right) \tag{6}$$

For doubtful assignment in cluster analysis we may use theorem from Gallegos (2002) and Gallegos and Ritter (2005) to assume that there are outliers between the overlapping areas of two artificially found clusters. Observations with large $DF_{(i)}$ values indicate doubtful assignments or trimming decision. In clustering problem, the use of discriminant factors was already suggested by Van Aelst et.al (2006). Silhouette plot (Rousseeuw 1987) can be used to summarize the order of discriminant factors. Figure 1 shows the result after applying the $DF_{(i)}$ function to a clustering solution found for the real data (real data will be explain in section 2.6). The most doubtful assignments with $DF_{(i)}$ larger than a log (threshold) value are highlighted in such

$$DF_{(i)} \geq \log(threshold) \tag{7}$$

Threshold = 0.1 means that a decision on a particular observation is considered as doubtful if the quality of the second best possible decision is smaller than one tenth of the quality of the actual made decision. All observations with $DF_{(i)} \geq \log(0.1)$ are highlighted in darker color in Figure 1.

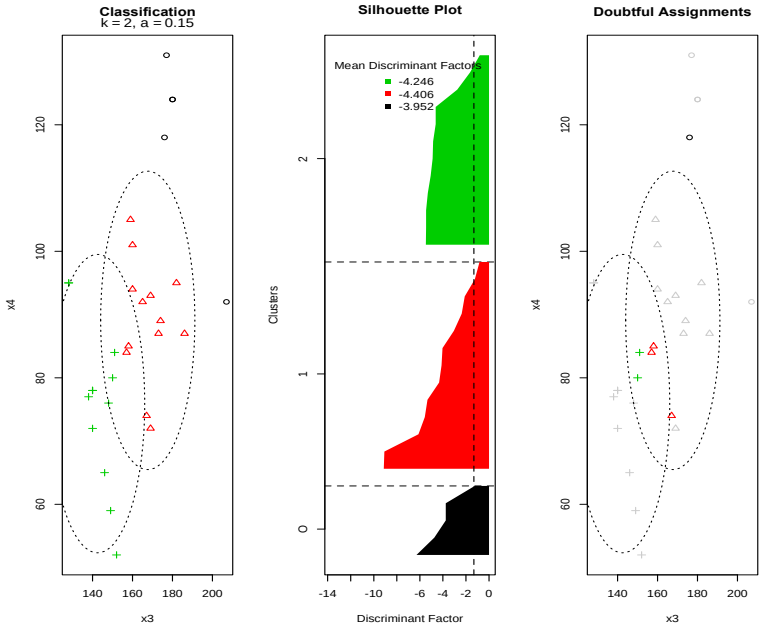

Figure 1: Result of Cluster analysis using doubtful assignment (+ is a data for cluster 1, Δ is a data for cluster 2)

2.4 The Quality of the Actual Made Decision

The actual made decision is a threshold value, which the existence of doubtful data is as minimum as possible. Figure 2 illustrates the number of outlying observations in the simulated data. The data generated was 1000 observations together with 5% of trimming proportion and two clusters. By setting the threshold value (x1), therefore for $DF(i)^{3} \log(x1)$ the doubtful data may arise when the strength of cluster assignment is low. Furthermore, the doubtful observation might tell us about the overall cluster in TCLUST. Figure 2 for example, we let threshold to be 0.1 because we assume that only less than 10% of the quality of actual made decision for observation belongs to the second best possible cluster. For the next analysis we tried using different values of threshold that are 1%, 5%, 15%, 20% and 25%

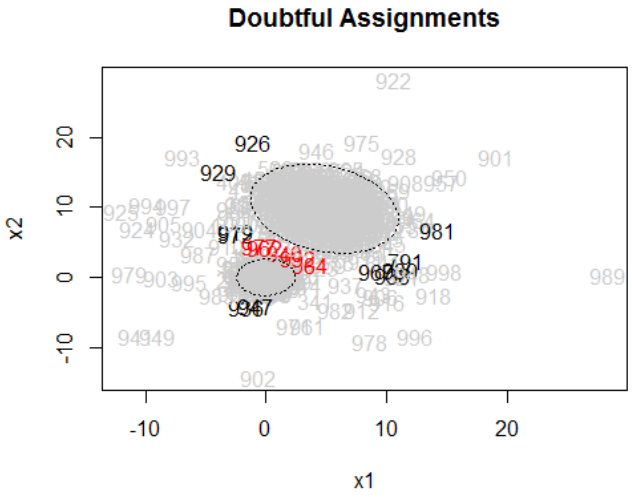

Figure 2: Graphical displays based on $DF(i)$ values for the number of observations. The dark colour of observation indicate as normal outlier whereas the coloured one consider as doubtful observations

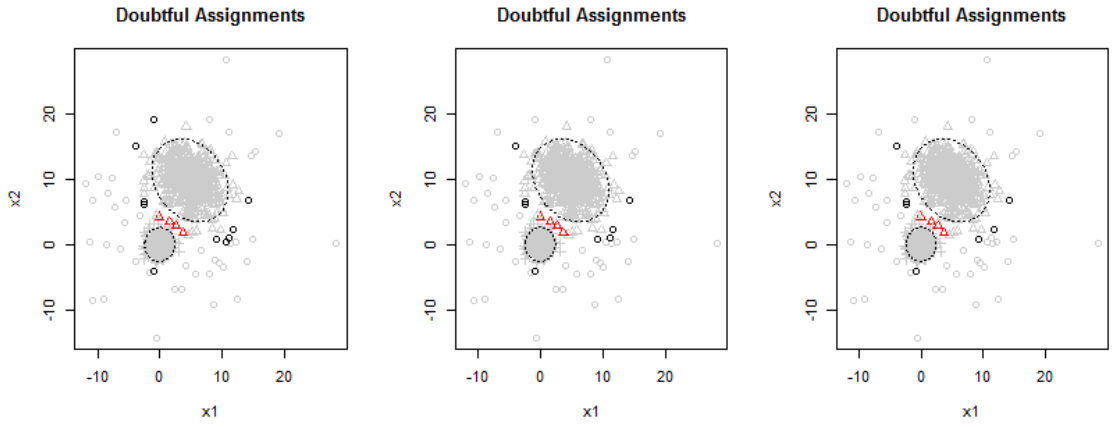

Figure 3: Graphical displays: Threshold 15%(Left), Threshold 20%(middle) and Threshold 25%(Right) (o is a real outlier, Δ is a doubtful observations)

Figure 3 represent the threshold values for 15%, 20%, and 25% whereas Figure 4 represent the threshold values for 1% and 5%. These values were chosen to find the best estimation for the threshold value therefore we can find the best possible percentages of outliers exist. The summary of threshold value and doubtful data from Figure 3 and Figure 4 are summaries in Table 1. From Table 1, as the threshold values increases, we can see the doubtful data decrease. In this study we try to show how important it is to estimate the threshold value so that we can know precisely the number of true outliers that exist in doubtful assignment. Table 1 shows that the suitable threshold value should be greater than 15%. Since there is no change in doubtful assignment when threshold value increases, we decided that the suitable value of threshold is 15%. There are 4 most doubtful decision (threshold value of 15%, 20% and 25%) and perhaps can be considered as outliers. If we let threshold value to be 15%, we can conclude that there are 4 most doubtful assignments with $DF(i) \geq \log(0.15)$. Which means, a decision of 4 doubtful assignments on observation, x_i are considered as doubtful because of the quality of second best possible decision is smaller than 15% of the quality of the actual made decision. The summary is in Table 1.

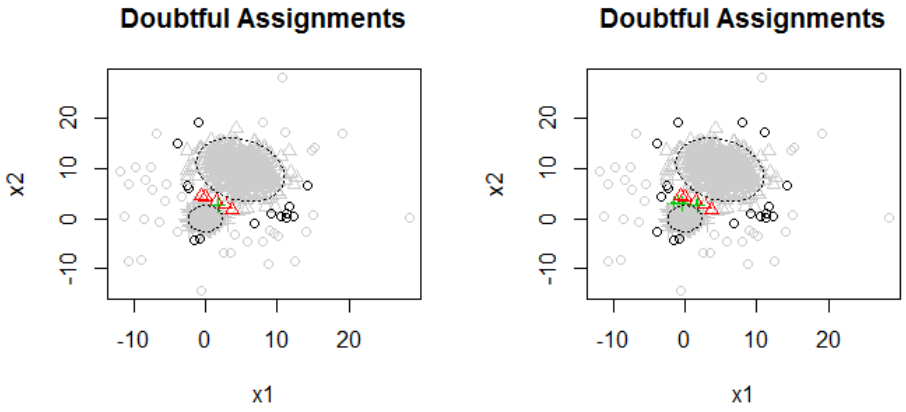

Figure 4: Graphical displays: Threshold 1% (Left) and Threshold 5%(Right) (o is a real outlier, Δ is a doubtful observations for cluster 1, + is a doubtful observations for cluster 2)

Table 1: Summary of threshold value and doubtful data detected

Threshold Value (%)	Number of Doubtful Data
1	8
5	6
10	5
15	4
20	4
25	4

2.5 Robust Linear Model, M Estimator

It is known that the median is the best robust estimator for the parameter b (the weighted mean). The median is neither very efficient nor can it be well generalized to the regression model. The M-estimator here is another family of estimator for the location model. M-estimator plays a role not only in fitting of location models but also in fitting of regression models. From practical point of view, the M-estimator is essentially a weighted mean, where the weights are designed to prevent the influence of outliers of the estimator as much as possible. The weighted is defined,

$$\hat{b}_M = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} \tag{8}$$

2.6 Data

A total of 30 participants to undergo the treadmill exercise are used. 15 of these participants are healthy and have no family history of hypertension or any cardiovascular disease, these groups are categorized as ‘Control Group’. On the other hand, the other 15 is a healthy participant with a family history of hypertension. This group is categorized as ‘High Risk Group’. The data is collected from Faculty of Biomedical Sciences, University of Selangor (Unisel).

3 Results and discussion

3.1 Result of Cluster Analysis

The systolic and diastolic of data before the exercise are used to indicate the existence of doubtful data. The output data of blood pressure before exercise are calculated so that the discriminant factor and doubtful data can be measure and plotted. The result in Figure 5 indicated the systolic against diastolic blood pressure before having exercise. The result shows that groups of people can be divided into two. Mean of discriminant factor shows that the clusters are well defined. In doubtful assignment, there are four doubtful data exist. Let threshold value 0.1, therefore for $DF_{(i)}^{3}\ \log(0.1)$ the observations that can be considered as doubtful are number 14, 15, 16, 29 and number 30 (Figure 5). With this simulation output, the doubtful observations or outliers may arise when the strength of cluster assignment is low. Furthermore, the doubtful observation might tell us about the overall cluster. For the case in Figure 5, we just set the threshold to be 0.1 because we assume that only less than 10% of the quality of actual made decision for observation belongs to the second best possible cluster.

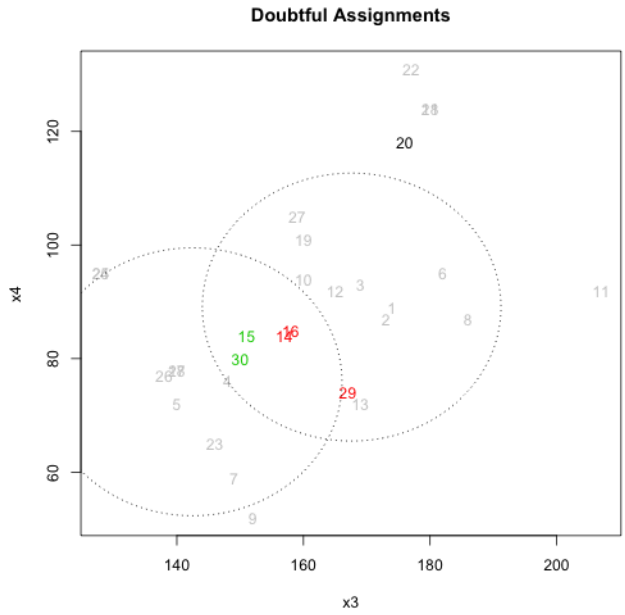

Figure 5: The observation of doubtful data

3.2 Fitting Linear Model Using M-estimator

The use of M-estimator in this paper is to fit linear model of robust regression allowing robust inference for parameters and robust model selection. It is because outliers in both respond variables and explanatory variable minimally influence robust fit. In the case of outlier in clusters, there are only doubtful observations that are exist between overlapping clusters. We try to figure out if robust regression having any affect with the presences of doubtful data in clustering. We run result for regression with deletion of doubtful data and on the other hand we run result for robust regression (M-estimator) with and without the deletion of doubtful data. Table 2 is a summary of regression analysis for linear and robust (m-estimator). Table 2 shows the coefficient of linear regression with intercept 14.89 where the slope is 0.4598. After the deletion of doubtful data, the intercept increase to 15.33 and the slope is 0.4639. For M-estimator, the data before the deletion, result shows the intercept is high which 30.47 and its slope is 0.3569. After the deletion of doubtful data, intercept is 17.62 and slope is 0.4498. This significant result shows that the M-estimator is less robust if there is doubtful data exist. Since we are remove the doubtful data, M-estimator shows that the right value of parameter as almost the same if using normal regression with deletion of doubtful data.

Table 2: Summary of linear regression and robust regression with and without (deletion) doubtful outlier.

	Linear Regression		M-estimator	
	Before deletion	After deletion	Before deletion	After deletion
Intercept	14.89	15.33	30.47	17.62
Slope	0.4598	0.4639	0.3569	0.4498

4 Conclusion

It is found that robust linear regression using M-estimator is affected (no deletion of doubtful data) when there is doubtful data exists. For simulation, result shows that the threshold value of 15% can detect the doubtful data more accurate compare to 10%. However, using real data with n=30 with threshold value of 10% we can verify the numbers of doubtful data exist. In this case we classified it as doubtful outlier. To test if the doubtful data affected a linear

model, comparison (linear model and robust linear model) shows that doubtful data affected the parameter of robust linear regression.

References

[1] Fritz, et.al. A fast algorithm for Robust Constrained Clustering. *Computational Statistics & Data Analysis*, 2013. Vol 61: 124-136

[2] Fritz, et.al. TCLUST: An R Package for a Trimming Approach to Cluster Analysis. *Journal of Statistical Software*, 2012. Vol 47, Issue 12, 1-26

[3] Garcia-Escudero et.al. A Review of Robust Clustering Methods, *Advances in Data Analysis and Classification*, 2010. Vol 4(2-3), 89-109.

[4] L.A., Garcia-Escudero et.al. Exploring the Number of Groups in Robust Model-Based Clustering, *Stat Comput*, 2011. Vol 21: 585-599

[5] Garcia-Escudero et.al. Robust Properties of k-means and trimmed k-means, *J Am Stat Assoc*, 1999 .Vol 94:956-969.

[6] Garcia-Escudero et.al. Trimming Tools in Exploratory Data Analysis. *J Comput Graph Stat*, 2003. Vol 12:434-449

[7] McLachlan GJ et.al. Robust Cluster Analysis via Mixture Models, *Austrian J Stat*, 2006. Vol 35: 157-174

[8] Rousseeuw PJ, Van Driessen K. A Fast Algorithm for the Minimum Covariance determinant Estimator, *Technometrics*, 1999. Vol 41: 212-223