Malaysian Journal of Fundamental & Applied Sciences

available online at http://mjfas.ibnusina.utm.my

TCLUST: Trimming Approach of Robust Clustering Method

Muhamad Alias Md.Jedi * and Robiah Adnan

Department of Mathematics, Faculty of Science, UTM, 81310 UTM Skudai, Johor, Malaysia

Received 26 December 2011, Revised 23 June 2012, Accepted 9 July 2012, Available online 25 Agust 2012

ABSTRACT

TCLUST is a method in statistical clustering technique which is based on modification of trimmed k-means clustering algorithm. It is called "crisp" clustering approach because the observation is can be eliminated or assigned to a group. TCLUST strengthen the group assignment by putting constraint to the cluster scatter matrix. The emphasis in this paper is to restrict on the eigenvalues, λ of the scatter matrix. The idea of imposing constraints is to maximize the log-likelihood function of spurious-outlier model. A review of different robust clustering approach is presented as a comparison to TCLUST methods. This paper will discuss the nature of TCLUST algorithm and how to determine the number of cluster or group properly and measure the strength of group assignment. At the end of this paper, R-package on TCLUST implement the types of scatter restriction, making the algorithm to be more flexible for choosing the number of clusters and the trimming proportion.

| TCLUST | Trimmed k-means | Number of Group | Strength of Group-assignments |

1. INTRODUCTION

The presence of outlying observations is a common problem in most statistical analysis. The case is the same when using cluster analysis techniques. Cluster analyses are basically detecting homogeneous clusters with large heterogeneity among them. To deal with outliers, robustness in cluster analysis is needed because outliers appear many times joined together (Garcia-Escudero *et.al.* 2011 [4]). Comparing between robust and non robust clustering procedure, non-robust clustering methods failed to accurately analyses even with the existence of small fraction of outlying data (Fritz *et.al.* 2011 [1]). For this case, robust clustering method always serves better to cluster correctly in the presence of outliers. The term "Spurious" is used by Fritz *et.al.*(2011 [1]) for outlying observation to explain when two or more clusters might be joined together artificially.

TCLUST strengthen the group assignment using constraints on scatter matrices. TCLUST methods are statistical clustering techniques which are based on the modification of trimmed k-means clustering algorithm. The constraints focused on in many literatures are mostly eigenvalues and are applied to TCLUST algorithm on the concentration step. By maximizing the spurious log-likelihood function with constraints on the eigenvalues, H is partitioned according to the number of clusters, k as desire (Garcia-Escudero *et.al.* 2010 [3]).

*Corresponding author at:
E-mail addresses: muhamad_jedi@yahoo.com (muhamad alias)

2. CHARACTERISTICS of TCLUST

2.1 TCLUST with other robust methods

Another robust alternative to k-means is Partitioning Around Medoids (PAM). Compared to TCLUST which is based on k-means, PAM did not well handle outlying data well, Fritz *et.al.* 2011[2] found that small number of outlying data did not affect the clustering result very much. But somehow if the outlier is very remote point or when the number of outliers increases, it will affect the total clustering result.

Forgy's k-means algorithm or called fast-MCD algorithm also play a very important role in cluster analysis for robust methods. This method is similar to trimmed k-means algorithm when we set the trimming level equal to 0. The difference is, trimmed k-means is based on euclidaean distance whereas fast-MCD based on Mahalanobis distance. Mahalanobis distance will update the centers and scatter matrices by computing the sample mean and sample covariance matrices assigned to each cluster. This will lead to insensible clustering result since large cluster sometimes will engulf the smaller ones. Therefore, Garcia-Escudero *et.al.* 2011 [4] introduced TCLUST to the existing method based on relative size constraint on the egenvalues.

2.2 Trimmed k-means

k-means is a simple and widely used in non-hierarchical clustering method. Given $\{x_1,...,x_n\}$ in R^p, k-

means method is based on the search of k point centers $\{m_1,...,m_k\} \subset R^p$ solving the minimization problem

$$\arg\min_{m_1,...,m_k} \sum_{i=1}^{n} \min_{j=1,...,k} \left\| x_i - m_j \right\|^2$$

Observations are then arranged into k-clusters by assigning each observation to the closest k-means center. Garcia and Gordazila 1999 [5] showed that k-means has breakdown point equal to 0. It can be interpreted that even the presented of one single outlier placed far away will completely spoil the k-means method. Later, Garcia and Gordazila 1999 [5] proposed robustness properties of trimmed k-means compared to classical k-means.

As k-means, trimmed k-means is defined through euclidian distances which specially aimed at finding spherical groups with almost the same size. However, when data set contain groups that depart strongly from that assumption, this method will fail and lead to wrong classification results. The search for groups with different size and scatter will lead to the heterogeneous clustering problem, where robustness aspects must also addressed. Gaegos and Ritter 2005 introduced mathematical probability framework for robust clustering problem. They proposed the probability density function $f(\cdot;\mu,\Sigma)$ of p-variate normal distribution with mean μ and covariance matrix Σ. Then the model of outlier called "spurious-outliers model" is defined via likelihood function stated below

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

with $\{R_0,...,R_k\}$ being the partitions of the set of indices {1,2,…,n} such that $\#R_0 = \lceil n\alpha \rceil$. R_0 are the indices of the non-regular observation which are considered as outliers and also the observations generated by other probability density function g_i. By maximizing the likelihood function, the spurious-outlier model can be written as

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

Log-likelihood function of spurious-outlier model will exclude the observation x_i with $i \in R_0$. Notice that the log-likelihood function with k=1 yields the Minimum Covariance Determinant (MCD) estimator. However the log-likelihood function is not well defined for cases k > 1 because this log-likelihood function is unbounded with no constraint on the cluster scatter matrix Σ_j. Since TCLUST algorithm considers different cluster matrix constraints, the different cluster sizes or cluster weights can be determined by searching for a partition $\{R_0,...,R_k\}$ with vectors μ_j, positive definite matrices Σ_j, and weight $\pi_j \in [0,1]$ maximizing

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

The spurious-outlier model will be well defined by maximizing the model function when constraints are applied in the TCLUST algorithm on scatter matrices.

2.3 Constrain on Scatter Matrices Σ_j.

TCLUST implements different algorithms to approximately maximize the well-defined problem of spurious-outlier model under different types of constraints which can be applied on the scatter matrices Σ_j. The strength of the constraint is controlled by the constant *c*.

Based on the eigenvalues of the cluster scatter matrices, a scatter similarity constraint may be defined as

$$M_n = \max_{j=1,...,k}\max_{l=1,...,p} \lambda_l\left(\Sigma_j\right) \text{ and } m_n = \min_{j=1,...,k}\min_{l=1,...,p} \lambda_l\left(\Sigma_j\right)$$

Where, M_n as the maximum and m_n as the minimum eigenvalues. The constraint ratio M_n / m_n should be smaller or equal than a fixed value, *c* where $c \geq 1$. This type of constraint will limit the relative size obtained through Σ_j when assuming normality and the shape has equidensity ellipsoids.

Besides eigenvalues, another way of restricting the cluster scatter matrices is by constraining their determinants. Given that

$$M_n = \max_{j=i,...,k}\left|\Sigma_j\right| \text{ and } m_n = \min_{j=i,...,k}\left|\Sigma_j\right|,$$

we want to maximize the well-defined spurious-outlier function by limiting the ratio M_n / m_n to be smaller or equal to *c*. This type of constraint limits the relative volumes of the mentioned equidensity ellipsoids, but not the cluster shape.

Another constraint considered is to force all the cluster scatter matrices to be the same that is $\Sigma_1 = ... = \Sigma_k$. This trimmed version was later introduced by Gallegos and Ritter 2005 where equal scatter matrices are known as the "determinantal" criterion.

3. RESULTS & DISCUSSION

3.1 TCLUST output

Using statistical software R, the result of TCLUST proposed by Fritz *et.al*, 2011 [2] is obtained. Using the data called M5data, all clustering result based on different constraints of scatter matrices are shown in Figure 1. The M5 data is a secondary data where a precise description of the data can be found at Garcia *et.al.* 2008. In this result of R programming, "restr.fact" [2] is defined as constant *c* where it will set $c = 1$ as default. As a result different constraints will have different shape of clusters.

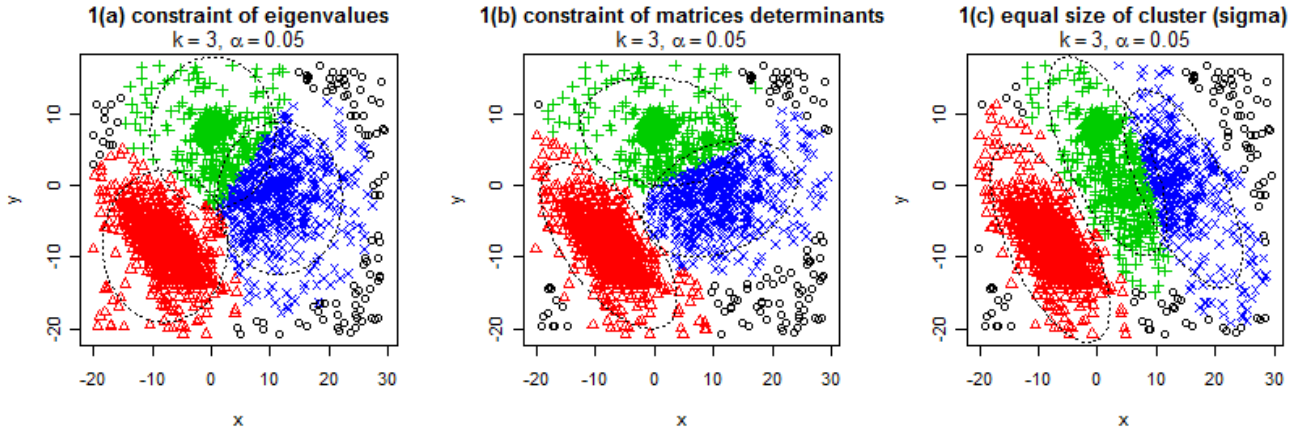

Fig. 1 M5data with different constraints on cluster scatter matrices with parameters $\alpha = 0.05$ and $k = 3$.

Figure 1(a) indicated the spherical clusters for eigenvalue, 1(b) cluster with the same matrices determinants and 1(c) is clusters with the same size of scatter matrices. In non hierarchical cluster analysis, prior knowledge about initial number of cluster is required. In this analysis, the M5data is set to have initial number of clusters, $k = 3$ with trimming proportion $\alpha = 0.05$. Since the constraint (eigenvalues) ratio value is set as default that is the M_n / m_n is calculated based on explanation of Fritz, L.A, and Mayo-Iscar 2011. It is found that $M_n / m_n = 50$ has been chosen because eigenvector of covariance matrices satisfy the

$$M_n = \max_{j=1,\ldots,k} \max_{l=1,\ldots,p} \lambda_l\left(\Sigma_j\right) \text{and } m_n = \min_{j=1,\ldots,k} \min_{l=1,\ldots,p} \lambda_l\left(\Sigma_j\right)$$

The result in Figure 2 (a) simplify that there is severe overlap of two clusters. However, since initial proportion of trimming is 5%, we can increase the proportion from 10% to 15% and up to 20%. What we can observed here, when the trimming proportion is increase to 20%, the clusters do not overlap with one another. Thus, in this study, 20% of trimming is the best option.

3.2 Appropriate number of groups and trimming proportion.

Most complex problem when applying non-hierarchical cluster analysis is to choose the number of clusters, k. It is certain that we must choose the initial number of cluster, but we did not really know what the best number of clusters that is supposed to be in the data. The same principal also applies to the trimming size, where we did not know exactly the true outlying level. Garcia-Escudero *et.al.* 2011 [2] introduced some classification trimmed likelihood curves as useful curve for choosing the number of clusters k. The k-th trimmed likelihood function is defined as

$$\alpha \mapsto \ell_c^{\Pi}(\alpha, k) \text{ for } \alpha \in [0,1)$$

with $\ell_c^{\Pi}(\alpha, k) = \sum_{j=1}^{k} \sum_{i \in R_j} \log f\left(x_i; \mu_j, \Sigma_j\right)$. This curve function is allowed to measure $\Delta_c^{\Pi}(\alpha, k) = \ell_c^{\Pi}(\alpha, k+1) - \ell_c^{\Pi}(\alpha, k)$ where $\Delta_c^{\Pi}(\alpha, k)$ should be close to 0. Figure 3 shows the classification trimmed likelihood curve $\Delta_c^{\Pi}(\alpha, k)$ when $k = 1, 2, 3, 4$ and α range is $[0, 0.2]$ and $c = 50$. We can see that no significant improvement happens when we increase *k* from 3 to 4 and $\alpha = 0$. This figure suggests that *k*=3 and $\alpha = 0$ (no trimming required) is possible sensible choice for *k* and α for this data set when $c = 50$. However refer to the nature of this data in Figure 3, increasing the α will help the data to not overlap. Therefore for this case of data the sensible choice would be *k*=3and $\alpha = 0.2$

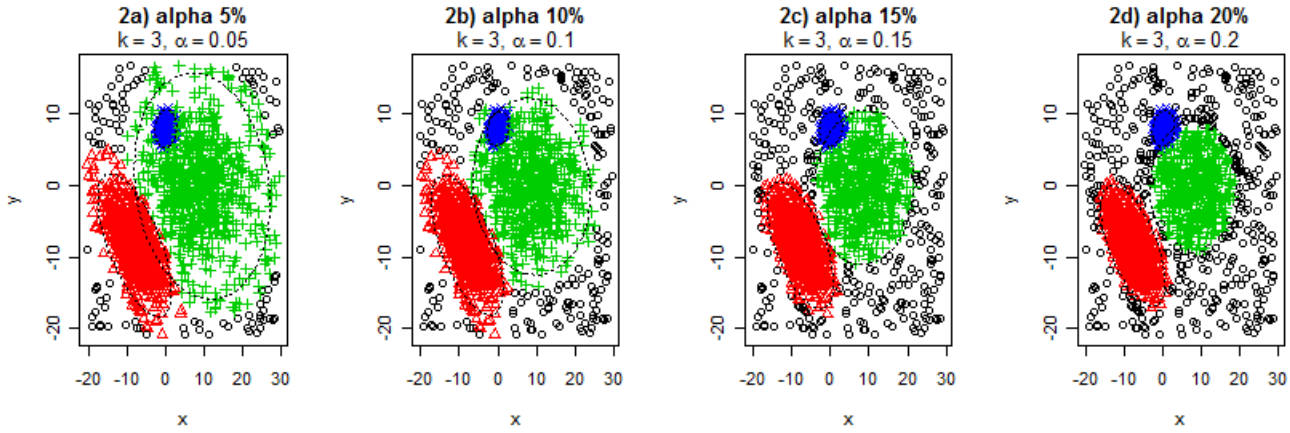

Fig. 2 M5data with different constraints on cluster scatter matrices with parameters $\alpha = 0.05$ and $k = 3$.

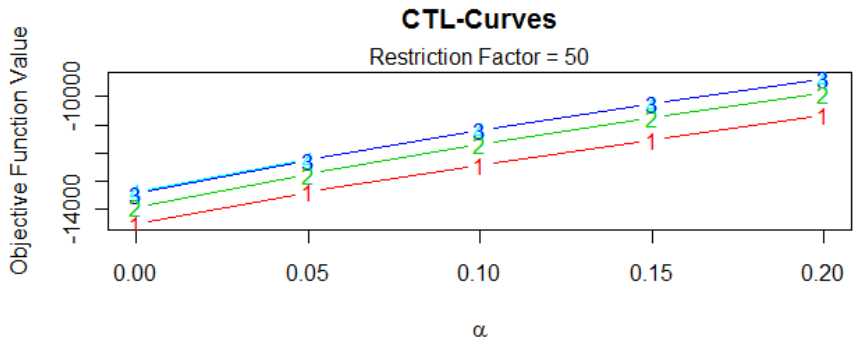

Fig. 3 Classification trimmed likelihood curves $\ell_c^{\Pi}(\alpha, k)$ when $k = 1, 2, 3, 4$, α range in $[0, 0.2]$ and $c = 50$.

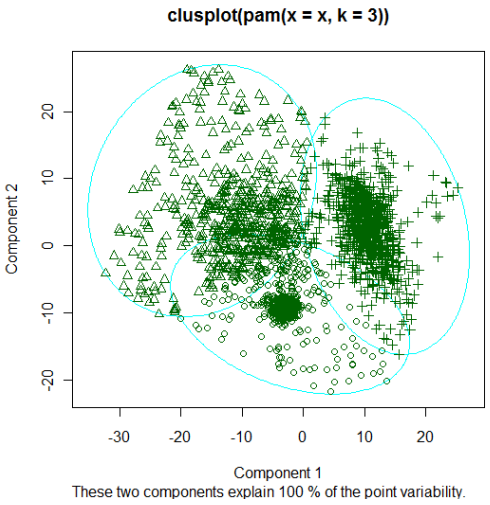

Fig. 4 PAM result for 3 clusters of M5data.

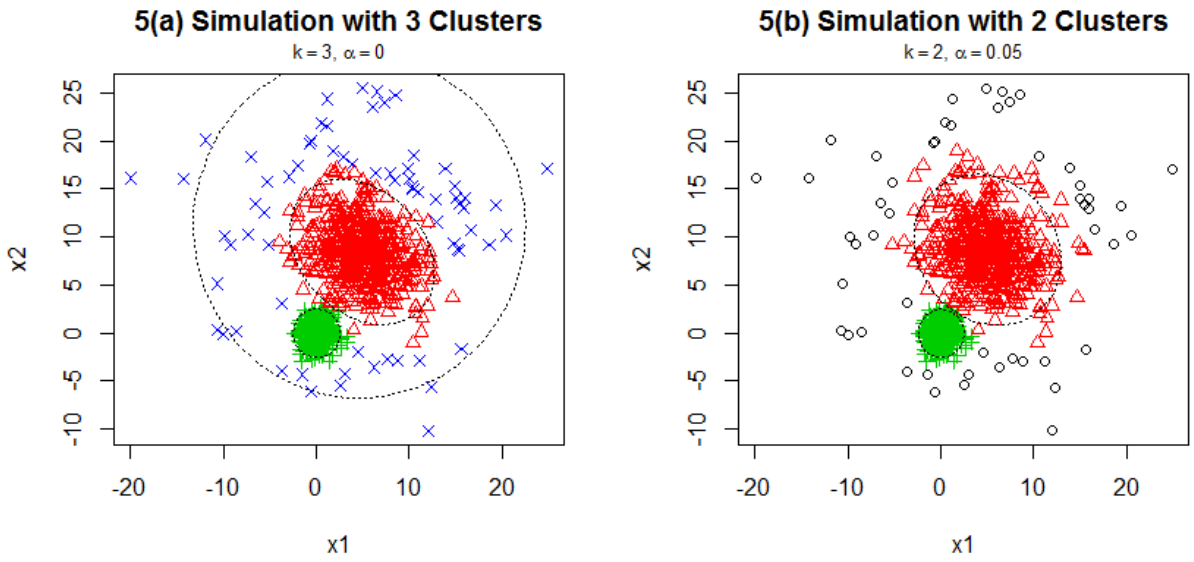

Fig. 5 Clustering results for the simulated data with k =3, α = 0 and c =50 (a) and k =2, α = 0.05 and c =12 (b)

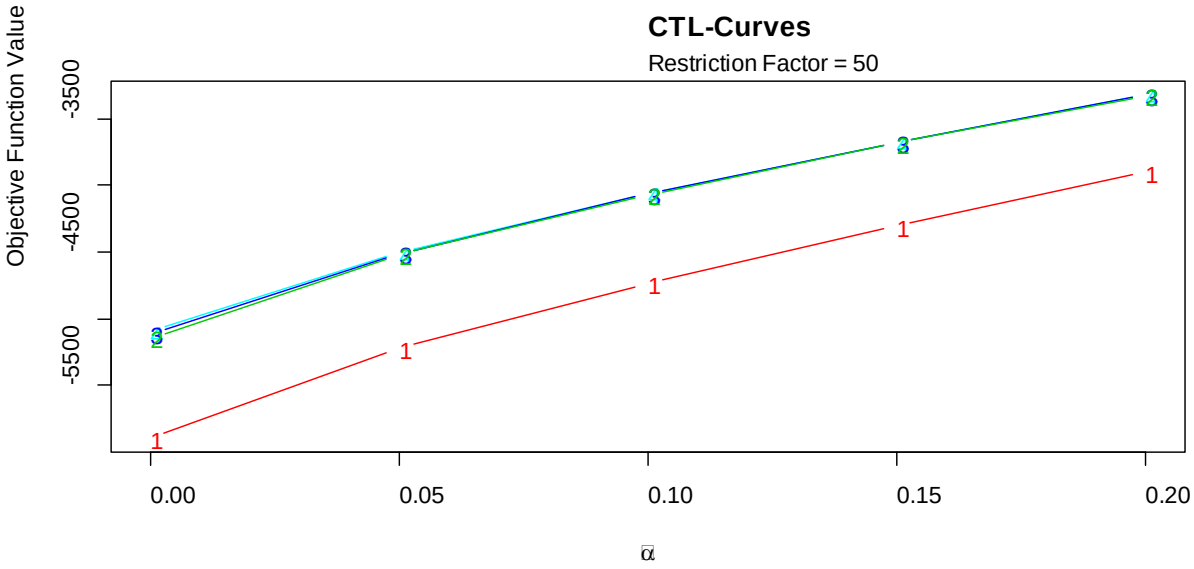

Fig. 6 Classification trimmed likelihood curves $\ell_c^{\Pi}(\alpha, k)$ when $k = 1, 2, 3, 4$, α range in $[0, 0.2]$ and $c = 50$ for data in Figure 5.

3.3 TCLUST and PAM

The Partitioning Around Medoids (PAM) clustering method is an alternative to k-means clustering. It can be seen that data that are far away from the mediod will affect the clustering result. Figure 4 demonstrate that outliers affect the clustering result. Without trimming, outliers will affect the position of k-mediod center and thus the clustering result. Figure 4 assured us that PAM are strongly affected by outliers and not the best option when dealing with much number of outliers. For M5data it is concluded that 20% of trimming is the best decision to determine α . Since the observation of this data is $n = 2000$, the outlier of 20% would be 400 data. For PAM 400 outliers will strongly affect the result.

3.4 Simulation study: Selecting the number of groups and the trimming size

In cluster analysis it is very important to determine the number of cluster that best describe the data. In this study, we also focus on the trimming proportion that to be chosen without knowing the true number of outliers. Result in Figure 3 shows that the choices for k and α are two related problems and it is very important to see if that particular trimming level implies the number of cluster. To support the analysis in section 3.1 and 3.2 simulation study was conducted to demonstrate the relation between α and k . Demonstration of α and k in Figure 5 are interpreted as a mixture of three Gaussian components (a) or a mixture of two Gaussian components (b) with a 5% outlier proportion.

However without knowing the true outlier proportion, we can make conclusion that both clustering solution in Figure 5 are perfectly sensible and the final choice of α and k only depends on the value of c. When considering Figure 5 (b) the proportion of outliers data are 5% and k =2 because the third Gaussian component partially overlaps with other two components.

Due to the important role of the trimmed likelihood curve, Figure 6 assured that the simulation data should have α = 0.05 and k =2. This curve supported that the simulation data having two Gaussian components with outliers of 5%.

4. CONCLUSION

It is found that for contaminated data, cluster analysis will tend to overlap and lead to unclear result. For M5data it is concluded that 20% of the most outlying data are contaminated and the best number of clusters is 3. Due to the nature of TCLUST, by having modification that is sensible to the scatter matrices we can include the constraint to maximize the spurious-outliers model. In this analysis we found that trimmed likelihood curve $\Delta_c^{\Pi}(\alpha, k)$ can be the explainable tool to help determine the appropriate trimming proportion and the number of clusters. By comparing TCLUST with other robust method namely PAM, TCLUST performs better. Knowing that PAM did not undergo the trimming process, the mediod will be strongly affected by outliers. Simulation results support the claims that outliers may affect the Gaussian component or number of cluster resulting that lead to bad inferences in cluster analysis. Therefore it is very important for researcher to determine the proper number of cluster and trimming proportion (for the case of the number of outliers is unknown).

REFERENCES

[1] Fritz, et.al. A Fast Algorithm for Robust Constrained Clustering, University of Valladolid, Spain, preprint available at http://www.eio.uva.es/infor/personas/tclust_algorithm.pdf. 2011

[2] Fritz, et.al. TCLUST: An R Package for a Trimming Approach to Cluster Analysis, Preprint available at http://cran.r-project.org/web/packages/tclust/vignettes/tclust.pdf, May 4, 2011

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

[4] Garcia et.al. Exporing the Number of Groups in Robust Model-Based Clustering, University of Valladolid, Spain, preprint available at http://www.eio.uva.es/infor/personas/langel.html, 2011

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

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

[7] Hathaway, R.J. A Constrained formulation of maximum likelihood estimator for normal mixture distributions, Ann. Statist, 13, 795-800. 1985

[8] Scott. et.al. Custering based on likelihood ratio criteria, Biometrics, 27, 387-397. 1971