AN EFFICIENT CLUSTERING ALGORITHM IN THE PRESENCE OF OUTLIER AND DOUBTFUL DATA

MUHAMAD ALIAS BIN MD.JEDI

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Doctor of Philosophy (Mathematics)

Faculty of Science
Universiti Teknologi Malaysia

DECEMBER 2015

Specially dedicated to *Mak, Ayah, Ashikin* and *Deeja*
I really love all of you.

ACKNOWLEDGEMENT

I wish to express my warmest gratitude to all those persons whose comments, questions, criticism, support and encouragement, personal and academic, have left a mark on this work. I also wish to thank Universiti Teknologi Malaysia, which have supported me during the work on this thesis. I wish to thank Assoc. Prof. Dr. Robiah Adnan, my thesis advisor, for his academic supervision and personal support throughout all my years in graduate school. I am also grateful to Ministry of Education (MoE), Malaysia for the MyPhD, MyBrain15 scholarship. Lastly, but most importantly, I wish to thank my parents, family and my wife. My gratitude to them is beyond words.

ABSTRACT

The presence of outlying observations is a common problem in most statistical analysis. This case is also true when using cluster analysis techniques. Cluster analysis basically detects homogeneous clusters with large heterogeneity among them. To deal with outliers, a correct procedure in cluster analysis is needed because usually outliers may appear joined together, which may lead to the wrong structure of clusters. New method of trimming in clustering (TCLUST) known as RTCLUST is proposed in this research that uses some information from TCLUST, partition around medoid (PAM), doubtful cluster and local outlier factor (LOF). TCLUST is a clustering method with constraint on the covariance matrices. For this case the constraint used was the eigenvalues. Spurious outlier model explains how to use the eigenvalues ratio, c for good clustering method. Good clustering is obtained using mean of discriminant. The value of c = 50 is obtained as a better value compared to the previous study c = 1. Trimmed likelihood is then used to determine the trimming proportion, α and number of clusters, k. The next procedure combines the TCLUST and PAM, which is known as MPAM. PAM is used because the mean of silhouette explains the clustering much better. The information obtained from MPAM are c = 50, α , and k. Different sample sizes are also used to test the suitability of MPAM. Mean of discriminant and mean of silhouette are then used to measure the strength of clustering. Trimmed likelihood curve is used again to check the values of α , and k. For the next step, using the doubtful cluster method with c = 50, the method shows the overlapping outliers that exist between clusters. In this case, the data in the overlapping area are classified as doubtful outliers and it is decided that the best threshold is 0.1. Lastly, the LOF is used to differentiate between doubtful outliers and real outliers in overlapping areas. Since LOF can detect real outliers, the deletion of this outlier is mandatory. Again, the mean of discriminant and mean of silhouette are obtained after the deletion of real outliers. A trimmed likelihood curve is then used to obtain the final value for α and k. This new procedure of RTCLUST uses c = 50 and threshold value equals 0.1 to obtain the mean of discriminant and mean of silhouette. To justify RTCLUST, medium sample size with Monte Carlo simulation is done to check the right possibility of combining methods, and therefore the normality of RTCLUST can be checked. Results found that the normality assumption for RTCLUST is fulfilled and Bayesian test can be used to significantly decide the value of k. Results for RTCLUST with having the lowest RMSE value shows that it is better than MPAM and TCLUST for both simulation and real data.

ABSTRAK

Kehadiran titik terpencil adalah perkara biasa dalam analisis statistik. Kes ini juga berlaku bagi analisis kluster. Analisis kluster secara dasarnya mengesan kelompok homogen dengan heterogen yang besar diantaranya. Untuk menangani titik terpencil, prosedur dalam analisis kluster diperlukan kerana biasanya titik terpencil akan kelihatan bergabung bersama-sama, dan boleh memberikan struktur kluster yang tidak betul. Kaedah baharu bagi pemotongan kluster (TCLUST) dikenali sebagai RTCLUST dicadangkan menggunakan maklumat dari TCLUST, partisi sekitar medoid (PAM), keraguan kluster dan faktor titik terpencil tempatan (LOF). TCLUST adalah kaedah kluster dengan kekangan pada matriks kovariannya. Untuk kes ini kekangan yang digunakan adalah nilai eigen. Model data terpencil menjelaskan cara untuk mengira nisbah nilai eigen c bagi mendapatkan kaedah pengklusteran yang baik. Pengklusteran yang baik dijelaskan menggunakan min diskriminan. Nilai c = 50 diperolehi sebagai nilai yang lebih baik berbanding kajian sebelumnya iaitu c = 1. Keluk kemungkinan dipotong digunakan untuk menentukan α dan k. Prosedur ini diteruskan dengan menggabungan TCLUST dan PAM dinamakan MPAM. PAM digunakan kerana min bayangan dapat menjelaskan kluster dengan lebih baik. Maklumat yang diperoleh daripada MPAM adalah c = 50, α dan k. Saiz sampel berbeza juga digunakan untuk menguji kesesuaian MPAM. Min diskriminan dan min bayangan kemudiannya diperolehi untuk mengukur kekuatan kluster. Keluk kemungkinan dipotong sekali lagi digunakan untuk mecari nilai, α dan k. Langkah seterusnya ialah menggunakan kaedah keraguan kluster dengan c = 50, kaedah ini boleh menentukan kewujudan titik terpencil yang bertindih antara kluster. Dalam kes ini, data di kawasan bertindih dikelaskan sebagai titik terpencil yang diragui dan nilai ambang terbaik yang dipilih adalah 0.1. Akhir sekali, LOF digunakan untuk membezakan titik terpencil diragui dan titik terpencil sebenar di kawasan bertindih. Oleh kerana LOF boleh mengesan titik terpencil sebenar, pembuangan titik terpencil sebenar dari data adalah perlu. Sekali lagi min diskriminan dan min bayangan diperolehi selepas pembuangan titik terpencil sebenar dilakukan. Keluk kemungkinan dipotong kemudiannya digunakan untuk menentukan nilai akhir bagi α dan k. Prosedur baharu RTCLUST menggunakan c = 50 dan nilai ambang sama dengan 0.1 untuk mendapatkan nilai min diskriminan dan min bayangan. Kewajaran RTCLUST diuji dengan menggunakan saiz sampel sederhana melalui simulasi Monte Carlo dimana ia digunakan untuk menilai kemungkinan gabungan bagi RTCLUST dan menilai samaada andaian taburan normal dipenuhi. Keputusan mendapati bahawa andaian taburan normal untuk RTCLUST dipenuhi dan ujian Bayesian boleh digunakan untuk memutuskan secara signifikan nilai k. Keputusan bagi RTCLUST dengan nilai RMSE yang rendah menunjukkan bahawa ia adalah lebih baik daripada MPAM dan TCLUST untuk kedua-dua simulasi dan data sebenar.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

k	-	Number of clusters
α	-	Outlier / Trimming proportion
c	-	Eigenvalues ratio
μ_D	-	Mean of discriminant
$s(i)$	-	Mean of Silhouette
$\ell_c^{\Pi}(\alpha, k)$	-	Trimmed likelihood values
$\Delta_c^{\Pi}(\alpha, k)$	-	The different of trimmed likelihood values between number of clusters

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Background of the Study

The term outlier refers to data that deviates significantly from the normal data as if generated by a different mechanism (Hawkins, 1980). Outliers are different from noise data. Noise is a random error and should be removed before outlier detection (Davies PL and Gather U, 1993). Unusual observations may be categorized into three types, namely outliers, high leverage points, and influential observation. Outliers could represent the indicative measurement of error and heavy tailed distribution. In some literature, outliers may come from a mixture of two distributions, or one might say they come from two distinct sub populations.

Outlier detection for non-hierarchical clustering is a procedure to classify objects into meaningful partitions while having strong structures with respect to the presence of outliers. Non-hierarchical cluster analysis will form into a pre-determined number of groups, so that the method generates a classification by partitioning a dataset. The existence of outliers is critical because even one outlier may provoke failure of clustering algorithms. To overcome these limitations in clustering procedures determining the number of clusters, k and trimming proportion α , and constraints ratio, c becomes vital. In recent years, detecting outliers by trimming the portion of outliers has become helpful to make k-mean as a good clustering method. The spurious outlier model used trimming and was able to handle different types of constraints for the group scatter matrices. This spurious outlier model was introduced by Garcia and Gordaliza (2005a) and is known as Trimming Cluster (TCLUST) analysis. Spurious outlier model assumes that there are non-regular observations (outliers) generated by other probability function. Generally a few criteria will be considered on how we could determine the number of clusters, k and trimming proportion, α .

Kaufman and Rousseeuw (1990) introduced Partition Around Medoid (PAM), using k-clustering on medoids to identify clusters. PAM works efficiently on small data set. TCLUST used the mean of discriminant while PAM used the mean of silhouette to explain the structures. Mean of silhouette is necessary to explain how good the clusters are. PAM is more robust to noise and outliers compared to k-means , because it minimizes the sum of pairwise dissimilarities instead of the sum of squared Euclidean distances. However, the weakness of PAM is that it is only suitable for a small sample size. In spurious outlier model, the problem occurs when outliers exist in between overlapping clusters. In this case, the classification of outliers may be due to the swamping effect. Swamping occurs when there appears to be a larger proportion of outliers than there really is. The importance of threshold value in clustering procedure is to minimize the swamping effect. This threshold values can also determine which cluster the data belongs to in the overlapping clusters. Generally, a few criteria should be considered in terms of how one may determine the outliers between overlapping clusters.

1.2 Significant of Study

In cluster analysis, finding outliers is not new. Finding uncommon cases is more interesting and useful than finding the common cases, such as detecting criminal activities in E-commerce or detecting abnormal cells pattern in health science. Identification of potential outliers is important for the following reasons. An outlier may indicate bad data. For example, the data may have been coded incorrectly, or an experiment may not have been run correctly. In some cases, the outlying data resembles a unique pattern, which can be scientific evidence that really exists in nature. Therefore it is logical to develop a robust outlier detection technique.

Modification of existing methods or proposing new idea will be beneficial especially in critical area such as in Health Science study.

1.3 Problem of Statement

In cluster analysis, trimming the data simply means removing the outlying observations. Researchers sometimes view isolated data or small groups of data as outliers. In TCLUST, trimming plays a major role in discarding true outliers based on proportion. The problem then is how to obtain the best estimate of the true proportion of outliers.

In cluster analysis, it is known that TCLUST works most efficiently on large data sets, while PAM works efficiently on small data sets. TCLUST uses mean of discriminant while PAM uses mean of silhouette. The combination of both methods named as MPAM is proposed because this will explain or produce better clusters.

In many (nearly all) cases we do not know the true proportion of outliers that exist in the data. However by using TCLUST, and Partition Around Medoid (PAM), one may get a close estimate of the percentage of true outliers that exist. Gallegos and Ritter (2005) suggested that the doubtful assignment in cluster analysis as an indication of outliers or bad trimming proportion (false trimming). In such a case, the correct threshold values are very important to determine the data that can be classified as doubtful outliers. Doubtful outliers may arise when there is overlapping clusters. By using Local Outlier Factor (LOF), the doubtful outliers and real outliers can be distinguished. The combination of MPAM and LOF will be named as RTCLUST.

1.4 Objectives

The objectives of this study consist of four sections:

1. To determine the trimming proportion and number of cluster using TCLUST.
2. To analyze cluster analysis using PAM and modified version of PAM.
3. To use the doubtful assignment method and LOF to identify outliers between clusters.
4. To propose and demonstrate the RTCLUST procedure using simulated and real hypertension data

1.5 Research Interest

Figure 1.1 shows the research flow and contributions of this research step by step. The procedures begins by using TCLUST with the correct c value obtained using trimmed likelihood curve k and α. Next is proceeded by simulation for PAM. A modification of PAM is done and this will be known as MPAM. MPAM is done using c from TCLUST, different threshold value also used to identify the doubtful data. Apart from that, LOF may be used to distinguish between real and doubtful outliers. In Figure 1.1, the blue color indicate the contributions of this research.

1.6 Outline of Thesis

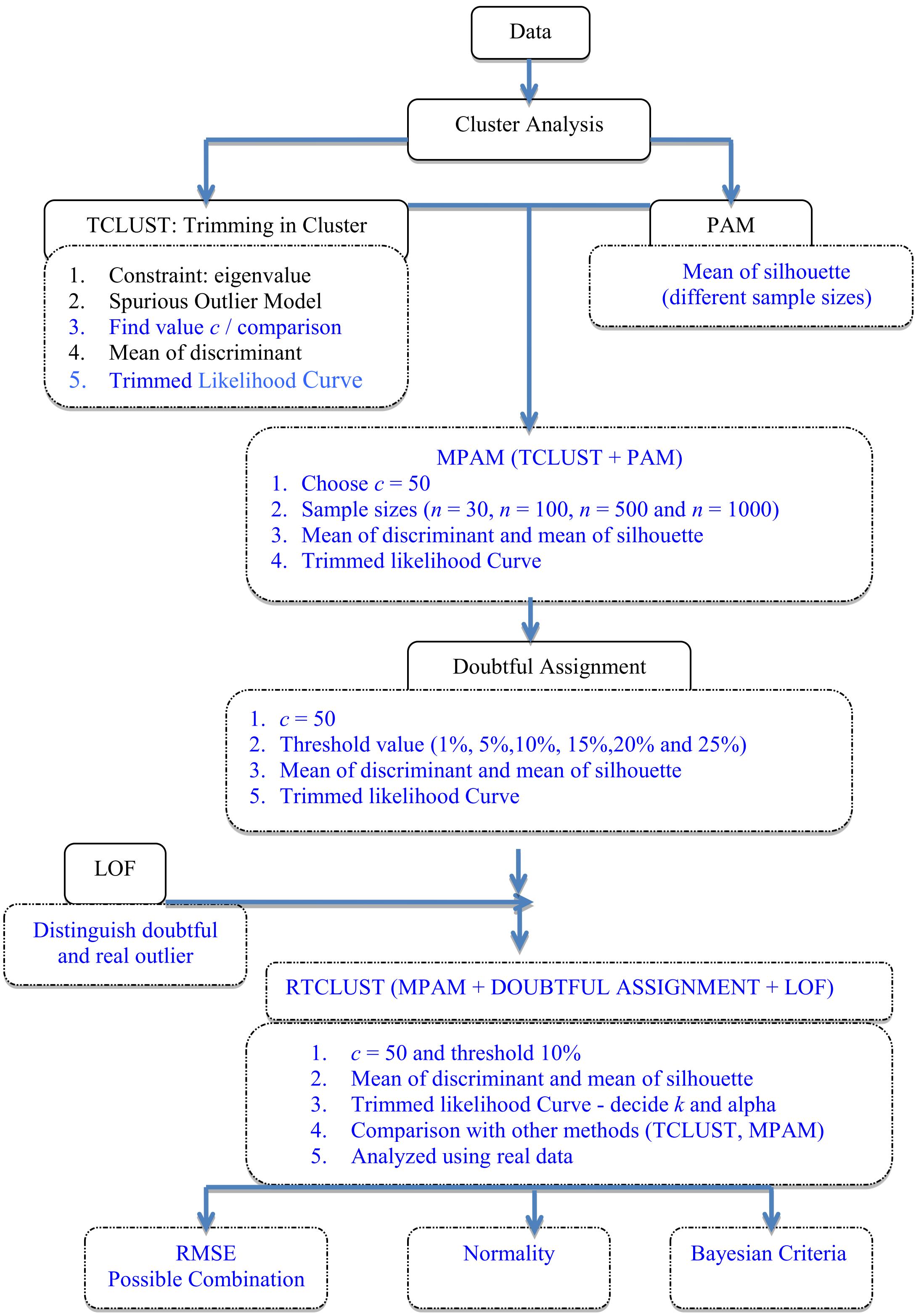
The research in this thesis uses a procedure to detect outliers in cluster analysis. There are 7 chapters in this thesis. Chapter 1 contains the introduction of the procedures for detecting outliers and methods available currently. After that it focuses more on cluster analysis, specifically on using spurious model of TCLUST. Problem statement and objective of the study is also clarified. Sec 1.10 states the contributions in this research.

Chapter 2 contains the literature review. It concerns previous works on cluster analysis. The details and discussions are summarized at the end of Chapter 2. In the summary table, the strength and weakness of the previous researches are discussed. Detailed discussions on the procedures used in the proposed method are carried out in Chapter 3, Chapter 4, and Chapter 5.

Chapter 3 is about trimming in cluster analysis. At the end of Chapter 3 it will answer the first objective of this thesis as mentioned in Section 1.4. The methods discussed in Chapter 3 include the trimming of the cluster with different proportion of outliers. The Chapter also contains discussion of the strength of cluster assignment when using eigenvalue as the constraint. This Chapter explains how the eigenvalues have been used so that the c value of cluster matrices can be obtained. c may be defined as the highest ratio of eigenvalues. A trimmed likelihood curve is used to decide the best number of outliers (trimming proportion) and the number of clusters.

Chapter 4 refers to modification of PAM. This modified PAM later will be renamed as MPAM. MPAM is a hybrid technique that combines the TCLUST and PAM method. At the end of Chapter 4 this unique technique is be able to detect outliers with more accuracy. However, the comparison between TCLUST, PAM, and MPAM will be discussed in detail in Chapter 6. MPAM is analyzed using TCLUST as the initial step. After that the value of c, α and k are chosen. After the initial step, PAM has been used to find the value of new alpha and k. In this step, mean silhouette is used to justify the value of k chosen. In this chapter objective number two is answered. In Chapter 5, doubtful assignment method is used in MPAM to detect doubtful outliers between overlapping clusters. Threshold value in doubtful assignment is the key to detecting outliers in the mentioned area. Different threshold values are used and will be explained in terms of the mean discriminant factor. A case study is contained in Chapter 6. In the introduction part of Chapter 6, the ability of LOF in differentiating real and doubtful outliers in overlapping clusters is discussed. In Chapter 7, RTCLUST is test for normality assumption and Bayesian criteria is use to decide the choice of k. Monte Carlo simulation is used for medium sample size to obtain the RMSE, mean of discriminant and mean of silhouette. Lastly, Chapter 8 contains the conclusion.

Figure 1.1: The diagram of thesis outline



REFERENCES

Atkinson AC, Riani M (2007). Exploratory tools for clustering multivariate data. *Comput Stat Data Anal,* 52,272–285.

Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49(3), 803-821.

Bock H-H (1996) Probabilistic models in cluster analysis. *Comput Stat Data Anal,* 23,5–28

Bryant, P. (1991). Large-sample results for optimization-based clustering methods. *Journal of Classification*, 8(1), 31-44.

Byers SD, Raftery AE (1998). Nearest neighbor clutter removal for estimating features in spatial point processes. *J Am Stat Assoc,* 93,577–584.

Celeux G, Govaert G (1992). A Classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis*, 14(3), 315-332.

Croux C, Gallopoulos E, Van Aelst S, Zha H (2007). Machine learning and robust data mining. *Comput Stat Data Anal,* 52,151–154.

Cuesta-Albertos J, Gordaliza A, Matran C (1997). Trimmed k-means: an Attempt to Robustify Quantizers. *Annals of Statistics*, 25(2), 553-576.

Cuesta-Albertos JA, Fraiman R (2007). Impartial trimmed k-means for functional data. *Comput Stat Data Anal*, 51,4864–4877.

Cuesta-Albertos JA, García-Escudero LA, Gordaliza A (2002). On the asymptotics of trimmed best k-nets. *J Multivar Anal,* 82:,482–516.

Cuesta-Albertos JA,Matran C, Mayo-Iscar A (2008). Robust estimation in the normal mixture model based on robust clustering. *J R Stat Soc Ser B,* 70,779–802.

Cuevas A, Febrero M, Fraiman R (2001). Cluster analysis: a further approach based on density estimation. *Comput Stat Data Anal,* 36,441–459.

Dasgupta A, Raftery AE (1998). Detecting features in spatial point processes with clutter via model-based clustering. *J Am Stat Assoc,* 93,294–302.

Davies PL, Gather U (1993). The identification of multiple outliers. *J Am Stat Assoc,* 88,782–801

Dykstra R (1983). An Algorithm for Restricted Least Squares Regression. *Journal of the American Statistical Association*, 78(384), 837-842.

Fisher (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics,* 7 (2): 179-188

Forgy E (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, 21, 768-780.

Fraley, C. and Raftery, A. E. (1998). How many clusters? which clustering method? answers via model-based cluster analysis. *The Computer Journal*, 41(8), 578-588.

Friedman H, Rubin J (1967). On Some Invariant Criterion for Grouping Data. *Journal of the American Statistical Association*, 63(320), 1159-1178.

Gallegos, M. T., Ritter G (2005). A Robust Method for Cluster Analysis. *Annals of Statistics*, 33(1), 347-380.

Gallegos, M. T. and Ritter, G. (2009). Trimming algorithms for clustering contaminated grouped data and their robustness. *Advances in Data Analysis and Classification*, 3(2), 135-167.

Garcia-Escudero LA, Gordaliza A (1999). Robustness properties of k-means and trimmed k-means. *Journal of the American Statistical Association*, 94(447), 956-969.

García-Escudero LA, Gordaliza A (2005a). Generalized radius processes for elliptically contoured distributions. *J Am Stat Assoc* 471:1036–1045.

García-Escudero LA, Gordaliza A (2005b). A proposal for robust curve clustering. *J Classif,* 22,185–201.

García-Escudero LA, Gordaliza A (2007). The importance of the scales in heterogeneous robust clustering. *Comput Stat Data Anal,* 51,4403–4412.

García-Escudero LA, Gordaliza A, Matrn C (1999). A central limit theorem for multivariate generalized trimmed k-means. *Ann Stat*, 27,1061–1079.

Garcia-Escudero LA, Gordaliza A, Matran C (2003). Trimming Tools in Exploratory Data Analysis. *Journal of Computational and Graphical Statistics*, 12(2), 434-449.

García-Escudero L.A, Gordaliza A, and Mayo-Iscar A. (2014). A constrained robust proposal for mixture modeling avoiding spurious solutions. “Advances in Data Analysis and Classification” Vol. 8, 27-43.

Garcia-Escudero LA, Gordaliza A, Matran C, and Mayo-Iscar A (2008). A General Trimming Approach to Robust Cluster Analysis. *Annals of Statistics*, 36(3), 1324-1345.

Garcia-Escudero LA, Gordaliza A, Matran C, and Mayo-Iscar A (2010). A Review of Robust Clustering Methods. *Advances in Data Analysis and Classification*, 4(2-3), 89-109.

Garcia-Escudero LA, Gordaliza A, Matran C, and Mayo-Iscar A. (2011). Exploring the number of groups in robust model-based clustering. *Statistics and Computing*, 21, 585-599.

García-Escudero L.A, Gordaliza A, Matrán C. and Mayo-Iscar A. (2015). Avoiding Spurious Local Maximizers in Mixture Modeling. *Statistics and Computing*, 25, 619-633

García-Escudero LA, Gordaliza A, San Martín R, Van Aelst S, Zamar R (2009). Robust linear clustering. *J R Stat Soc Ser B*, 71:301–318.

Gordaliza A (1991). Best approximations to random variables based on trimming procedures. *J Approx Theory,* 64:162–180.

Hardin, J. and Rocke, D. M. (2004). Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator. *Computational Statistics & Data Analysis*, 44(4), 625-638.

Hathaway RJ (1985). A Constrained Formulation of Maximum Likelihood Estimation for Normal Mixture Distributions. *Annals of Statistics*, 13(2), 795-800.

Hawkins, D (1980). *Identification of Outliers*. Chapman and Hall, London.

Hennig C (2003). Clusters, outliers, and regression: fixed point clusters. *J Multivar Anal*, 86,183–212

Hennig, C. (2004). Breakdown points for maximum likelihood-estimators of location-scale mixtures. *Annals of Statistics*, 32(4), 1313-1340.

Hennig C (2008). Dissolution point and isolation robustness: robustness criteria for general cluster analysis methods. *J Multivar Anal,* 99,1154–1176.

Jiang MF, Tseng SS, Su CM (2001). Two-phase clustering process for outliers detection. *Pattern Recognit Lett,* 22,691–700.

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

Markatou M (2000). Mixture models, robustness, and the weighted likelihood methodology. *Biometrics,* 356,483–486

Markus M.Breunig, Hans-Peter Kriegel, Raymond T.Ng, Jorg Sander. LOF: Identifying Density-Based Local Outliers. *Proc. ACM SIGMOD Int. Conf. On Management of Data*. Dalles TX. 2000,1-12

Maronna R (2005). Principal components and orthogonal regression based on robust scales. *Technometrics,* 47,264–273

Maronna R, Jacovkis PM (1974). Multivariate Clustering Procedures with Variable Metrics. *Biometrics*, 30(3), 499-505.

Massat DL, Plastria E, Kaufman L (1983). Non-hierarchical clustering with MASLOC. *Pattern Recognit,* 16,507–516.

McLachlan GJ, Ng S-K, Bean R (2006). Robust cluster analysis via mixture models. *Austrian J Stat*, 35,157–174.

Milligan GW, Cooper MC (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika,* 50,159–179.

Muhammad Faiz Bin Zulkifli (2012). *Blood Pressure Response During Treadmill Testing on Individual with High Risk of Hypertension*. Final Year Project, Bachelor of Degree, Universiti Selangor.

Neykov N, Filzmoser P, Dimova R, Neytchev P (2007). Robust Fitting of Mixtures Using the Trimmed Likelihood Estimator. *Computational Statistics & Data Analysis*, 52(1), 299-308.

Perrotta D, Riani M, Torti F (2009). New robust dynamic plots for regression mixture detection. *Adv Data Anal Classif,* 3,263–279.

PolonikW (1995). Measuring mass concentrations and estimating density contour clusters: an excess mass approach. *Ann Stat,* 23,855–881.

Rocke DM,Woodruff DM (1996). Identification of outliers in multivariate data. *J Am Stat Assoc,* 91:1047–1061.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(1), 53-65.

Rousseeuw, P. J, Van Driessen K (1999). A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*, 41, 212-223.

Scott AJ, Symons MJ (1971). Clustering Methods Based on Likelihood Ratio Criteria. *Biometrics*, 27(2), 387-397.

Schynsa M, Haesbroeck G, Critchley F (2010). RelaxMCD: smooth optimisation for the minimum covariance determinant estimator. *Comput Stat Data Anal,* 54,843–857.

Symons M (1981). Clustering criteria and multivariate normal mixtures. *Biometrics*, 37, 35-43.

Van Aelst, S., Wang, X., Zamar, R. H., and Zhu, R. (2006). Linear grouping using orthogonal regression. *Computational Statistics & Data Analysis*, 50(5), 1287-1312.

Vinod HD (1969). Integer programming and the theory of grouping. *J Am Stat Assoc*, 64,506–519.

Vogt, W., Nagel, D (1992). Cluster analysis in diagnosis. *Clinical Chemistry*, 38: 182-198

WillemsG, Joe H, ZamarR (2009). Diagnosing multivariate outliers detected by robust estimators. *J Comput Graph Stat,* 18:73–91.

Woodruff, D. and Reiners, T. (2004). Experiments with, and on, algorithms for maximum likelihood clustering. *Computational Statistics & Data Analysis*, 47(2), 237-25