Performance Comparison of Clustering Techniques on the basis of galaxy data

Article · January 2013

CITATIONS
0

READS
158

1 author:

Tanuka Chattopadhyay
University of Calcutta
59 PUBLICATIONS **258** CITATIONS

SEE PROFILE

Performance Comparison of Clustering Techniques on the basis of galaxy data

Tuli De[1],Tanuka Chattopadhyay,[2]
and
Asis Kumar Chattopadhyay[3]

[1]Department of Mathematics,Heritage Institute of Technology,Kolkata
Choubaga Road,Anandapur,Kolkata -700107
email:tuli.de@heritageit.edu

[2]Department of Applied Mathematics, Calcutta University, Kolkata
92 A.P.C. Road, Kolkata -700009
email:tanuka@iucaa.ernet.in

[3]Department of Statistics, Calcutta University, Kolkata
35 Ballygunge Circular Road,Kolkata-700019,India
email: akcstat@caluniv.ac.in

Abstract

Clustering and classification of different astronomical objects have become one of the most important area in the field of Astrostatistics.The basic objective of cluster analysis is related to segmentation of a collection of objects into a (unknown) number of clusters such that objects in the same cluster are more closely related than those assigned to different clusters.Cluster analysis is also used to study on the basis of descriptive statistics, whether each group consists of objects with significantly different properties.

There are many methods available for clustering,which may be broadly categorized under supervised and unsupervised learning. In case of supervised learning there are some input variables (predictors) and also some output variables (responses).But in case of unsupervised learning only predictors are under consideration in the absence of responses. Under both the above mentioned categories,for clustering and classification, several methods have been developed on the basis of the underlying nature of data sets.However, there is no well known criteria to compare the performances of different techniques.

The present paper makes an attempt to compare the applicability of some of the clustering techniques on the basis of two astronomical data sets,one Gaussian and the other non Gaussian. Clustering have been done with respect to all the variables as well as significant components. A post classification technique is used as a supervised learning to justify the robustness of the variety of unsupervised methods used in this purpose. Finally the similarity of clusters obtained from different methods is viewed in terms of astrophysical properties of the objects grouped in different clusters.

Keywords: Model Based,Non Model Based Clustering, Classification, Supervised Learning, Unsupervised Learning, Post Classification.

1. Introduction

Classification or clustering of different astronomical objects is of primary importance in the domain of Astrostatistics. However,depending on the nature of the data set several complications arise.Sometimes the data set under consideration may have complicated distributional form and non-Gaussian in nature.Based on this point, there is a justification needed about which method of clustering or classification technique should be used so that it reflects the proper nature of grouping. The clustering techniques which require an inherent model assumption are known as Model Based Methods, whereas the clustering technique where no modelling assumption or distributional form is needed may be termed as Non-Model based Methods. Hence based on the nature of data set, one has to decide about proper application of the two types of techniques. Even if one decides about the proper method for the data set at hand, there are several techniques available under both the categories and it is difficult to judge which technique is the best for the present scenario.Hence there arises the need of a comparative study among various available techniques and a computational comparison of different methods.

If all the methods are implemented in a particular situation, there is a need of judgement to decide upon the best technique based on a post classifier. So an appropriate post classification approach is also needed in this regard. For a post classification approach, a pre-classifier or training sample is required. Since in this type of techniques a prior knowledge of classification is provided, these are called Supervised Learning. All other techniques where no prior classification is provided are known as Unsupervised Learning. A comparative validity algorithm may be helpful for predicting the superiority of different techniques.

In particular we are interested in astronomical data and many such data sets are very large in terms of both the number of observations and the variables considered. Many standard clustering techniques fails to deal with such large data sets. Thus some dimension reduction methods may be applied at first and then clustering may be performed on the reduced data set. The above criteria also needs to be validated depending on whether the data are Gaussian or non Gaussian as the dimension reduction technique may vary depending on the distributional pattern. Finally after all these considerations, the similarity of grouping of objects obtained from different methods may be checked in terms of some physical properties.

Several works has been done in recent years in the field of Astronomy and Astrophysics by applying the various techniques discussed above.Different clustering algorithms like kMeans (MacQueen 1967), Hierarchical methods like agglomerative and divisive, model based clustering under distributional assumptions,Bayesian methods etc. have been widely used by several statisticians in order to identify the internal grouping structures of objects. In the area of Astrostatistics, kMeans and model based classification have been applied for the classification of Gamma ray Bursts (Chattopadhyay et al. 2007),large scale clustering and classification problems have been tackled by considering Hough Transformation and CASH algorithm (Chattopadhyay et al. 2012).

Dimension reduction techniques can be used as an initial step in statistical modelling and clustering. Some dimension reduction techniques like Principal Component Analysis (PCA) and Independent Component Analysis (Comon 1994) have been used for clustering and identification of proper variables for the purpose of grouping.When both the number of variables and number of observations are quite large,most of the standard techniques will fail. In particular for non Gaussian situation those problems are quite apparent.Under astrostatis-

tics, applications of dimension reduction and clustering techniques are quite common and Chattopadhyay et al. 2012, Chattopadhyay et al.2013 considered such problems.

2. Different clustering and classification techniques:

The supervised and Unsupervised Learning Technique: A supervised Learning is a machine learning task of inferring a function from supervised (Labeled) training data. This technique investigates the training data and produces an inferred function, which is called a classifier or a regression function. Thus a target function is learned that can be used to predict the values of a discrete class attribute.
An Unsupervised Learning is a machine learning problem which tries to find hidden structure in unlabelled data.
The difference between the above two techniques is that,
(i) The supervised Learning has a prior knowledge about data classes whereas Unsupervised Learning has no such prior information and (ii) The Supervised Learning is based on a predictive approach and it specifies a model for the underlying data set. But for unsupervised Learning no such model specification is needed.

2.1 Unsupervised learning techniques:

In this method, one may start with the raw data with or without any distributional assumption regarding the underlying distribution.

2.1.1 Unsupervised learning methods without distributional assumption of the data :

There are two main categories of this kind of clustering methods, viz., Connectivity based clustering,Centroid based clustering and Density based clustering.
Connectivity based clustering - the Hierarchical clustering: The Hierarchical clustering method is a very popular technique of cluster analysis which builds a hierarchy of clusters.There are mainly two approaches to form the hierarchy viz. agglomerative and divisive:
Agglomerative nesting is a "bottom up" approach where each observation is considered as a cluster and pairs of clusters are merged as one moves up the hierarchy. The most similar objects are grouped first and those initial groups are merged ultimately into single cluster according to some proximity measure.These proximity measures are based on either similarities or dissimilarities (distances).
Divisive Analysis is a "top down" approach where all observations at first are grouped in one cluster, and splits are performed recursively as one moves down the hierarchy.Here an initial single group of objects is divided into two subgroups such that the objects in one subgroup are "far from" the objects in the other. These subgroups are further divided into dissimilar groups until there are as many subgroups as objects.
Possible distances in Hierarchical clustering: In order to decide which clusters should be combined (for agglomerative) or where a cluster should be split (for divisive), a distance matrix is required. The distances that are mainly used for hierarchical clustering, come from

a very general distance known as the Minkowski's distance or the pth norm, which may be defined as follows-
Suppose there are two points P = $(x_1, x_2, ..., x_n)$ and Q = $(y_1, y_2, ..., y_n)$ in the n dimensional space. Then the pth norm is given by-

$$L_p = (\sum_{i=1}^{n} |x_i - y_i|^p)^{1/p} \tag{1}$$

.

For p=1, it gives the Manhattan distance i.e., the L_1 norm. For p=2, Minkwoski's distance reduces to the Euclidean Distance or the L_2 norm. Also, p = ∞, L_p norm results in The Chebyshev Distance.For one dimension, all the three distances are equivalent. For two dimension, Manhattan and Chebyshev are equivalent. But for dimension > 2, all the distances are different.For Hierarchical clustering, mainly Euclidean and Manhattan distances are used.But these measures are applicable only to continuous data.For categorical or in particular binary responses such distances will not work.In those cases Gower's similarity measure (Gower,1971) may be used. As the method is not very well known, a brief description is given below.
The Gower's coefficient of similarity: Let there be two individuals i and j to be compared on a character k. A score is attached with these individuals as -

$$s_{ijk} = \begin{cases} 0, when\ i\ and\ j\ are\ different \\ 1, or\ a\ positive\ fraction,\ when\ i\ and\ j\ are\ similar \end{cases} \tag{2}$$

.

Hence the similarity between i and j over all possible comparisons may be taken as

$$S_{ij} = \sum_k s_{ijk} / \sum_k \delta_{ijk} \tag{3}$$

where

$$\delta_{ijk} = \begin{cases} 1, when\ character\ k\ can\ be\ compared\ for\ i\ and\ j \\ 0, \quad otherwise \end{cases} \tag{4}$$

A similarity measure for a quantitative data say, $(x_1, x_2, ..., x_n)$of character k is given by

$$s_{ijk} = 1 - \frac{|x_i - x_j|}{R_k} \tag{5}$$

Here R_k is the range of character k. It may be noted that the selection of appropriate distance matrix for clustering problem completely depends on the physical situation.
Linkage Methods: After finding distances between single data points, the distances between clusters are found by different linkage methods.If the linkage is computed on liberal attitude i.e.,the distance between nearest neighbours is treated as the distance between points

of the two clusters, then the method is called the single linkage method. Whereas a conservative attitude is maintained for complete linkage where the distance between the farthest neighbours is treated as the distance between points of the clusters.Finally when as average attitude is maintained, i.e., the average distance is considered then the linkage method is known as average linkage Method.There are several other possible choices of linkages.The choices of distances and linkages totally depends on the particular problem under consideration and the analyst must make a proper choice, so that a good clustering pattern is achieved as the different choices usually lead to different types of clustering unless clustering pattern is very prominent.

The results of hierarchical clustering (agglomerative and divisive)are usually presented in a dendrogram. Morlini and Zani (2012) introduced a new index to compare two hierarchical clusterings (or dendrograms). It is a summary measure of dissimilarity and can be decomposed into contributions relative to each pair of partitions.This measure considers the dendrograms as a collection of k groups (k =2,3,...., n-1) and evaluates the dissimilarity in terms of the number of pairs of objects which are classified in different groups in each of the (n-2) partitions of the two clusterings.

The centroid based clustering methods: The Hierarchical clustering has the drawback that when the number of observations is large, the dendrogram becomes very messy and the results are not easily interpretable. To avoid this problem another type of clustering method is introduced known as centroid based clustering. Under these techniques,clusters are represented by a central vector, which may not necessarily be a member of the data set. The most popular and useful algorithm under this category is kMeans Technique (MacQueen 1967).

kMeans clustering: When the number of clusters is fixed to k (say), kMeans clustering finds the cluster centers and assigns the objects to the nearest cluster center, such that the distances from the cluster are minimized.This method, however only find a local optimum, and is commonly run multiple times with different random initializations.It is convergent and quite robust but one of the major drawback of the method is that the number of clusters has to be specified priori.Similar to Hierarchical clustering, the Euclidean and Manhattan distance are used as distance matrix for kMeans clustering.

One point of concern about the kMeans clustering is that , the clusters obtained by this method, varies widely with the initial choices of centroid or seed points (Miligan,1980). The centroid may be arbitrarily chosen points from the entire data. Also one may partition the data into randomly selected groups equal to the number of clusters to be determined. Then the group averages may also serve as seed points. Experimental evaluations show that kMeans gives good results when the group averages are taken as seed points.

Some methods may be used to find optimum value of k.In this paper, the optimum number of clusters for kMeans technique is decided by using a method developed by Sugar and James (2003). In this algorithm, firstly varying number of clusters have been obtained for k =1,2,3,4 etc. Then for each such cluster formation, the value of a distance measure is computed as follows-

$$d_k = \frac{1}{p} min_x E[(X_k - C_k)'(X_k - C_k] \tag{31}$$

,

which is the distance of the X_k vector (values of variables/parameters) from the cluster center C_k , p is the order of the X_k vector. Now, d_k is estimated at the kth point as $d_{k'}$ which is the minimum achievable distortion associated with fitting k centers to the data.

Then, for choosing the optimum number of clusters a plot of $d_{k'}$ versus k is obtained, known

as the distortion curve. This curve is always monotonically decreasing. Now, by looking at this graph that value of k is chosen at which the distortion curve levels off or its rate of deployment stabilises.
Alternatively,one may calculate the jumps or the transformed distortions as -

$$j_k = (d_{k'})^{-p/2} - (d_{(k-1)'})^{-p/2} \tag{32}$$

.
and by plotting j_k versus k, one chooses that value of k as the optimum number of clusters for which j_k is highest, i.e., the sharpest jump is obtained.
In order to find the optimum number of groups the algorithm of Gap Statistics(Tibshirani et al.2001) may also used. The Gap Statistics is defined as the difference between the log of the residual sum of squared distances within clusters(say w_k) and its expected value derived using bootstrapping under the null hypothesis that there is only one cluster.The graph of $\log(w_k)$ is standardised by comparing it with its expectation under appropriate null reference distribution of the data. The estimate of the optimum number of clusters is then the value of k for which $\log(w_k)$ falls the farthest below this reference curve. This technique has been used to determine the optimum value of k(number of clusters) in Chattopadhyay et al. 2012.
The density based clustering -DBSCAN Method: Density based spatial clustering application with noise (DBSCAN)Technique (Mertin,1996) is mainly developed based on the idea that a cluster will be recognised as a region with higher concentration or density of points than points outside the cluster. The density of outlying observations or noise will be lower than the density of points in any cluster.This method has an advantage in the sense that unlike Kmeans, there is no need to specify priori the number of clusters to be determined.Also the method is efficient even when it is applied on large data sets and the data contains outliers or noise.The technique classifies the points in the database into two types.**a)**The border points, the points located on the extremities of the cluster,the core points, which are located on its inner region,**b)**A neighborhood of a point p is a set of all points that have a distance measure less than a predetermined value, called Eps.Therefore, the neighborhood size of the core points is generally bigger than that of the border points,**c)**A point p is directly density-reachable from another point q if it belongs to the neighborhood of q and q's neighborhood size is greater than a given constant known as MinPts.More generally, a point p is density-reachable from q if there exists a chain of points $p_1, ..., p_n$,where $p_1 = q$, $p_n = p$, and p_{i+1} is directly density-reachable from p_i. With the help of these concepts,a cluster C is defined as follows.
If p belongs to C and if q is density-reachable from p and p's neighborhood is greater than the MinPts threshold, then q belongs to C (Maximality)for all p,q.
If q belongs to C and p belongs to C then there must be a point t such that p and q are directly reachable from t (Connectivity)for all pair of p,q.
The points which do not belong to any of the generated clusters are termed as outliers or noise.By adjusting the values of Minpts and Eps, one can find clusters of varying shapes and densities.

2.1.2 Unsupervised Learning with Distributional Assumption:

The model based clustering techniques: The Expectation Maximization (EM) algorithm(Demster et al.1977) usually models the data set with a fixed number of Gaussian distributions that are initialized randomly and whose parameters are iteratively optimized

to fit better to the data set. This will converge to a local optimum, so multiple runs may produce different results. Clusters can then easily be defined as objects belonging most likely to the same distribution.Here X is the observed data with a given statistical model, Z is the set of unobserved latent data or missing values Z and $\tilde{\theta}$ is a vector of unknown parameters with a likelihood function ,L($\tilde{\theta}$,X,Z) = $p(X, Z|\tilde{\theta})$, The maximum likelihood estimate (MLE) of the unknown parameters is determined by the marginal likelihood of the observed data,L($\tilde{\theta}$,X) = $p(X|\tilde{\theta}) = \sum P(X, Z|\tilde{\theta})$.
The EM algorithm finds the MLE of the marginal likelihood by the following two steps of iteration:

Expectation step (E step): It calculates the expected value of the log likelihood function, with respect to the conditional distribution of Z given X under the current estimate of the parameters $\tilde{\theta}^{(t)}$:

$$Q(\tilde{\theta}|\tilde{\theta}^{(t)}) = E_{z|x,\tilde{\theta}^{(t)}}[logL(\tilde{\theta}; X, Z)] \tag{6}$$

Maximization Step (M step): It finds the parameter that maximizes this quantity:

$$\tilde{\theta}^{t+1} = argmax_{\tilde{\theta}} Q(\tilde{\theta}|\tilde{\theta}^{(t)}) \tag{7}$$

2.2 A Supervised learning technique :

Supervised learning technique may be viewed as a mapping between a set of input variables χ and an output variable Υ.This mapping is applied to predict the outputs for unseen data. The main characteristic of supervised learning is the availability of annotated training data.It supervises the learning system to instruct on the labels to associate with training examples. These labels are known as class labels in classification problems. Supervised learning induces models for the training data and these models are then used to clasify other unlabeled data. Two most popular supervised learning techniques are the Support Vector Machines (SVM) and the Nearest Neighbour(NN) Classifiers. The present work only focuses on NN classifier described as follows-
The nearest neighbourhood approach: This method was introduced by Fix and Hodges (1951). An extensive review is given in Dasarathy (1991). K nearest neighbourhood (Cover et al.1967) is a pattern recognition technique where examples (data points) are classified based on the class of their nearest neighbours. Since more than one neighbours are mainly taken, the method is often referred to k-nearest neighbour(k-NN) classification, where k nearest neighbours are used to determine the class.
The method runs in two stages, the first stage determines the nearest neighbours and the second is the determination of the class using those neighbours.
Let x_i, $i\epsilon D$ be the members of a training data set D.Each training example is labelled with a class label $y_j \epsilon \Upsilon$.The examples are described by a set of features F.To classify an unknown example q, for each $x_i \epsilon D$, the distance between q and x_i is calculated as follows

$$d(q, x_i) = \sum_{f \epsilon F} w_f \delta(q_f, x_{if}) \tag{8}$$

For continuous and discrete attributes,$\delta(q_f, x_{if})$ takes the following form-

$$\delta(q_f, x_{if}) = \begin{cases} 0, f \ discrete \ and \ q_f \ = \ x_{if} \\ 1, f \ discrete \ and \ q_f \ \neq \ x_{if} \\ |q_f - x_{if}|, f \ continuous \end{cases} \tag{9}$$

The basic distance metric described in equations (8) and (9) is a special case of the Minkowski distance metric(given by equation (1)) , in fact it is the L_1 norm or the Euclidean distance. Other distance metrices like the Kullback-Leibler distance(Kullback and Leibler 1951), χ^2-statistics(Rubner et al.2000) may also be used for this purpose.The nearest neighbours are selected based on this distance metric. There are a variety of ways in which k nearest neighbours can be used to determine the class of q, the most popular way is the majority voting,i.e., to assign the majority class among the nearest neighbours to the query. The votes using inverse distance may be mathematically defined as below-

$$vote(y_j) = \sum_{c=1}^{k} \frac{1}{d(q, x_c)^n} 1(y_j, y_c) \tag{10}$$

This is the vote assigned to class y_j by neighbour x_c,where,

$$1(y_j, y_c) = \begin{cases} 1, if \ the \ class \ labels \ match \\ 0, otherwise \end{cases} \tag{11}$$

Another choice of vote is based on Shepard's work (Shepard,1987) using exponential function given by-

$$vote(y_j) = \sum_{c=1}^{k} e^{-\frac{d(q,x_c)}{h}} 1(y_j, y_c) \tag{12}$$

where $1(y_j, y_c)$ is defined as in equation (11).
The best choice of k depends upon the data. Generally, larger values of k reduce the effect of noise on the classification, but make boundaries between classes less distinct.The selection of K may be done by cross validation techniques.

2.3 Dimension reduction techniques:

When the data set is large (both in terms of number of variables and number of observations)one may first apply some appropriate dimension reduction technique and then perform clustering on the reduced data set.
Principal component analysis: In this technique , given a data set of observations of correlated variables , an orthogonal transformation is performed to convert it into a set of uncorrelated variables called the principal components. The number of principal components is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible variance (that is,it

accounts for as much of the variability in the data as possible). The components whose eigen values are greater than one are considered only in the reduced space.Principal components are guaranteed to be independent only if the variables are jointly normally distributed.
Independent component analysis: Independent Component Analysis (ICA) (Common (1994)), model assumes the form

$$X = AS \tag{13}$$

where X is a data matrix, A is the non-singular mixing matrix, S is matrix of independent components. $A^{-|}$ is the unmixing matrix.
The main goal of ICA is to estimate the unmixing matrix $A^{-|}$. It is assumed that the data variables are linear or non-linear mixtures of some latent variables and the mixing system of equation (1) can be written as :

$$X_i = a_{i1}S_1 + a_{i2}S_2 + ... + a_{in}S_n; i = 1(1)n \tag{14}$$

The S_i's are mutually independent, where a_{ij}s are the entries of the non-singular matrix A. For performing ICA, the data set has to be sphered or whitened, i.e., any correlations in the data has to be removed.ICA can be carried out in different ways like Maximization of Gaussianity, minimization of mutual information etc. In this paper, only Maximization of Non-Gaussianity is discussed which is discussed in terms of Negentropy (Hyv*árian* et. al. 2001). Negentropy is the difference between the differential entropy of a source S from the differential entropy of a Gaussian source with the same covariance of S.
Mathematically, Negentropy J(S) can be defined as :

$$J(S) = H(S_{Gauss}) - H(S) \tag{15}$$

,
where S_{Gauss} is a Gaussian random variable with the same variance as S. Now the entropy of a discrete source S is defined by-

$$H(S) = -\sum p_s(S_i) log p_s(S_i \tag{16}$$

where p_s is the mass function. Whereas, the entropy for a continuous source is given by -

$$H(S) = -\int p_s(\eta) log p_s(S_i) \tag{17}$$

where $p_s(\eta)$ is the density function.
Negentropy is always non-negative and it is equal to 0 if and only if S has a Gaussian Distribution. S is estimated by maximizing the distance of its entropy from Gaussian entropy. Sometimes it is very difficult to compute negentropy and it is often approximated as follows-

$$J(S) \propto (E[G(S)] - E[G(S_{Gauss})])^2 \tag{18}$$

where G is a non-quadratic function.

There are many algorithms to calculate ICA like FastICA, ProDenICA (Hastie and Tibshirani, 2003), KernelICA etc. In this paper, only FastICA is used. Before application of the FastICA algorithm, the observed data are whitened. The FastICA algorithm uses PCA as the whitening method.There is no rigorous method to determine the optimum number of ICs. In this paper, the number of ICs to be taken is considered to be equal to the number of principal components with eigen values greater than 1 (Albazzaz and Wang 2004).As most of the data sets in Astrophysics are likely to be non Gaussian,ICA can be sucessfully used in many situations(Chattopadhyay et al. 2012,Chattopadhyay et al. 2013).
Principal component analysis, Factor Analysis, Projection Pursuit are some popular methods based on linear transformation. But ICA is different from other methods, because it looks for the components in the representation that are both statistically independent and non Gaussian. ICA separates statistically independent components,which are the original source data, from an observed set of data mixtures.All information in the multivariate data sets are not equally important. There is often a need for extraction of the most useful information. Independent Component Analysis extracts and reveals useful hidden factors from the whole data sets. ICA defines a generative model for the observed multivariate data, which is typically given as a large database of samples.

3. Cluster Validity Algorithm

Cluster validation is an important issue in clustering and classification. In most of the clustering algorithms, the number of clusters are user specified. So there is a need to check whether the clusters formed satisfies the optimum requirement or not.Some of the validation indices may be described as follows.
Dunn's Validation Index: The technique (Dunn 1974) identifies the cluster sets that are compact and well separated. Here the objective is to maximize the inter cluster distances and minimize the intra cluster distances.Mathematically, it can be expressed as

$$D = \min_{1\leq i\leq n}\{\min_{1\leq j\leq n, i\neq j}\{\frac{d(c_i,c_j)}{\max_{1\leq k\leq n} d'(c_k)}\}\} \tag{19}$$

where $d(c_i, c_j)$ is the distance between the clusters c_i and c_j and is called the intercluster distance, $d(c_k)$ is the intracluster distance of cluster c_k. The number of cluster that maximise D is taken as the optimal number of cluster.
Davies-Bouldin Validity Index: This index (Davies and Bouldin,1979) is a function of the ratio of the sum of within cluster scatter to between cluster separation and is given by-

$$DB = \frac{1}{n}\sum_{i=1}^{n}\{\frac{S_a(Q_i)+S_a(Q_j)}{S(Q_i,Q_j)}\}$$

where n is the number of clusters, S_a is the average distances of all objects from the cluster to their cluster centre, $S(Q_i, Q_j)$ is the distance between cluster centres. DB will have a small value for a good clustering.
C Index: This index (Hubert and Schultz 1976) is defined as follows.

$$C = \frac{S - S_{min}}{S_{max} - S_{min}} \tag{20}$$

where S is the sum of distances over all pairs of patterns from the same cluster. Let l be the number of those pairs. Then S_{min} is the sum of the l smallest distances if all pairs of patterns are considered. S_{max} is the sum of the l largest distances out of the pairs. Hence a small value of C indicates a good clustering.
The Silhoutte Validation Index(Rouseeuw 1987):Let a_i be the average dissimilarity of the ith object to all other objects in the same cluster and b_i be the minimum average dissimilarity of the ith object to all objects in the closest cluster. Then the Silhoutte Index (Rousseeuw 1987)for the ith object is given by

$$S_i = \frac{(b_i - a_i)}{max(a_i, b_i)} \tag{21}$$

.
Taking the averages of all $S_i's$, i = $1(1)n_j$, n_j = number of objects in cluster j, one may validate the correctness of classification.
It follows from the formula that $-1 \leq S(i) \leq 1$. If Silhouette value is close to 1, it means that sample is well-clustered ,if Silhouette value is about zero, it means that sample could be assigned to another closest cluster as well, and the sample lies equally far away from both clusters. If silhouette value is close to 1, it means that the sample is misclassified.
The largest overall average Silhouette value indicates the best clustering (number of cluster). Therefore, the number of cluster with maximum overall average Silhouette value is taken as the optimum number of the clusters.In the present work we have used this index.

4. Data:

In the present work two different data sets have been considered to evaluate the efficiency of different clustering techniques. One is non Gaussian data set and another one is Gaussian data set.

4.1 The non Gaussian data set:

The different methods of clustering and classification have been applied first to a data compiled and standardized using the procedure followed in Hudson et al.(2001).The sample consists of 56 low-redshift galaxy clusters containing 699 early-type galaxies. After eliminating the missing observations the sample size has been reduced to 528 and the methods have been performed using four parameters (variables),the effective radius(R_e),the surface brightness averaged over effective radius(μ_e),the central velocity dispersion(σ) and magnesium index(Mg_2) respectively.
The effective Radius(R_e): The effective radius (Re) of a galaxy is the radius within which the luminosity of the galaxy is half of its total luminosity,assuming the galaxy to be circularly symmetric.The present set contains the logarithm of R_e, i.e.,$logR_e$.
The effective surface brightness(μ_e): This is measure of the amount of light that an object especially a galaxy emits per unit area of the sky. Even a galaxy that has a high total luminosity can be hard to see if it has a low surface brightness.
The central velocity dispersion(σ): The velocity dispersion of galaxies is the spread of velocities of galaxies in a cluster along line of sight.

The present data set contains the logarithm of σ, i.e.,$log\sigma$.
The magnesium index (Mg_2): The magnesium index is the strength of magnesium line in the spectrum of the galaxy. For the sake of discussion some more parameters are considered which are LogM_{dyn},, Tilt and α which are discussed in later section.

4.2 The Gaussian data set:

This data set consists of a large sample of hot stellar systems, starting from globular clusters (GCs) in our Galaxy, Large Magellanic cloud (LMC), Small Magellaric cloud (SMC), Virgo cluster of galaxies (VCC) , nuclear star clusters (NuSc), bulges of spiral galaxies (Sbul), ultra compact dwarf galaxies (dEs) to giant elliptical galaxies (gEs) of Misgeld and Hilker (2011). The parameters used are logarithm of effective radius (log_{Reff}),logarithm of stellar mass ($M_{Stellar}$) and logarithm of surface density (logΣ_e).

5. Result for non Gaussian data set:

Experimental evaluation shows that all the Hierarchical clustering techniques(agglomerative nested and agglomerative divisive) with different distances (Euclidean, Manhattan, Gower's)result in 7 groups or clusters.
The clusters are visible from dendrograms of different hierarchical techniques. Here only the dendrogram of divisive hierarchical clustering with manhattan distance is provided (other dedrograms show alike pictures) in Figure 1.
Similar clustering nature (i.e., 7 clusters) has been obtained by kMeans technique with both euclidean and manhattan distances. The optimum number of clusters is decided by using Sugar and James (2003) criteria. The Jump and Distortion curves are shown in figures 2 to 5. The distortion curves stabilise from the value k=7, suggesting the optimum number of clusters to be 7. The jump curves show that the largest jump has been reached for k=7, again suggesting the same optimum number of clusters.
The DBSCAN and EM clustering also result in 7 clusters though the cluster sizes reflect that the major concentration of observations are mainly in one or two clusters.
After applying all these unsupervised learning techniques, a post classification scenario is predicted by applying nearest neighbourhood approach by preparing training samples from the groups of Hierarchical and kMeans training clustering with Euclidean distance.
Roughly 33 per cent of the data set consists of a training sample. So a random sample from each cluster of Hierarchical and kMeans methods in the proportion of 0.33 formed a training sample. On the basis of this training sample, nearest neighborhood technique has been applied to the test data , i.e., the data set provided with the minimum number of neighborhood points for any point is taken as 5.There is no optimum criterion to fix the minimum number of neighborhood points to be considered. However, by trial and error method, it is seen that for this particular data set, if the minimum number of points is 5, the post classification gives maximum similarity with the pre classification , i.e.,Hierarchical and kMeans.
As discussed in Section 2.5,Independent Component Analysis (ICA) is performed in the non Gaussian data set as a dimension reduction technique and KMeans and Hierarchical clustering with Euclidean distance have been performed in the reduced data set to check whether they resulted in similar kinds of grouping as obtained in the unreduced data sets by the same methods. Using ICA, two independent components have been extracted and then kMeans

and Hierarchical techniques have been used. Here also the Sugar and James (2003) optimality criterion suggests that there are 7 clusters which is evident from the corresponding distortion and the jump curves (Figures 6 and 7).

5.0.1 Statistical analysis with respect to Silhoutte index:

Finally after application of all clustering and classification methods, Silhoutte index for different clusters are calculated. Here only average Silhoutte index is considered. If the average value is greater than 0 it means that the observations are moderately more or less well clustered. Furthermore, if the average value is close to 0.5, it means that observations are very well classified.
The different Silhoutte indices for different techniques along with the cluster names and sizes, average properties of the parameters (for Hierarchical agglomerative clustering with Euclidean distance, EM and Hierarchical technique applied to IC's, the extra parameters $\log M_{dyn}$, Tilt and α are considered for comparison of physical properties which are described in later sections) are given in Table 1.
In Table 1,HAEG1-HAEG7 are groups found by hierarchical agglomerative clustering with Euclidean distance, HAMG1-HAMG7 are groups found by hierarchical agglomerative clustering with Manhattan distance, HDEG1-HDEG7 are groups found by hierarchical divisive clustering with Euclidean distance,HDMG1-HDMG7 are groups found by hierarchical divisive clustering with manhattan distance, HAGG1-HAGG7 are groups found by hierarchical agglomerative clustering with Gower's distance, KEG1-KEG7 are groups by kMeans clustering with Euclidean distance, KMG1-KMG7 are groups by kMeans clustering with Manhattan distance, DBG1-DBG7 are groups by DBSCAN clustering, EMG1-EMG7 are groups by EM clustering, NHG1-NHG7 are nearest neighbourhood technique on hierarchical clustering,NKG1-NKG7 are nearest neighbourhood technique on kMeans clustering, HIG1-HIG7 are ICA applied to hierarchical clustering, KIG1-KIG7 are ICA applied to kMeans clustering.
Table 1 shows that the Silhoutte values for HAG1, HAG3, HAG7 are greater than 0.3 indicating the observations are quite well clustered. Silhoutte values for HAG5 and HAG6 are greater than 0.2, so these clusters also contain well classified observations. Only the Silhoutte values for HAG2 and HAG4 are just greater than 0, hence the observations belonging to these classes are just moderately clustered.
Silhoutte values for HAMG2, HAMG4 and HAMG5 are greater than 0.3 indicating that the observations are quite well clustered. Silhoutte for HAMG6 is greater than 0.5, so these observations are very well clustered though there are only a few observations belonging to this cluster. The Silhoutte index for HAMG1 is just greater than 0.1 and thus the observations in this class are not properly classified. Here HAMG7 contains only one observation, and the Silhoutte index can not be calculated for this class, hence it is marked by * sign.
The Silhoutte values of HDEG1, HDEG2,HDEG3,HDEG4,HDEG5,HDEG6 and HDEG7 all are between 0.1 to 0.25, indicating that all the observations are moderately clustered. Similarly the Silhoutte values for HDMG1, HDMG2,HDMG3, HDMG4, HDMG5, HDMG6 and HDMG7 are ranging between 0.1 to 0.23 , indicating that all the observations here also are moderately clustered. The clusters HAGG1,HAGG2,HAGG3,HAGG4,HAGG5 and HAGG6 have average silhoutte values between 0 to 0.25 showing the fact that this technique only results in moderately appropriate clusters. But HAGG7 has Silhoutte value close to 0.5, but this cluster contains only a few observations. Thus, overall Silhoutte results for different

Hierarchical Techniques show that among all these methods Hierarchical agglomerative clustering with Euclidean distance gives the best result with respect to the given non Gaussian data set.
Similar comparison with the groups created by kMeans clustering shows that for both the distance measures viz., Euclidean and Manhattan , the methods does not perform well with respect to the given data.
From Table 1, it appears that the Silhoutte values for the clusters DBG1, DBG2, DBG4, DBG5 and DBG6 lies between 0.4 to 0.55 meaning that the observations belonging to these clusters are well classified , but the sizes of the clusters are too small so that they do not really contribute to the clustering nature of the method. On the other hand, the Silhoutte values for DBG3 and DBG7 are less than 0, meaning that the observations are not at all properly clustered. But DBG7 actually contains the major number of observations. Hence it can be said that DBSCAN method fails to give a proper classification of the data set, may be due to the fact that the data is neither noisy nor too large.
Similarly it can be inferred that the EM clustering technique is not at all applicable to this data set, as the data is non gaussian and does not have a proper distributional mixture.
Table 1 shows that the Silhoutte values for NHG3 and NHG6 are greater than 0.3, implying that the observations are well clustered. The Silhoutte values for NHG1 , NHG2, NHG4, NHG5 all lie between 0.1 to 0.28, hence these groups contain moderately classified observations. But since NHG7 has only one observation, the Silhoutte value cannot obtained from that cluster. Hence, it can be said that though Hierarchical agglomerative clustering with Euclidean distance performs quite well with respect to the given data, it is not able to behave as a proper pre classifier or training sample to reflect the predictive nature of classification based on nearest neighbourhood approach.
In the same way it can be mentioned that although kMeans approach is not the best technique for clustering, as far as this data set is concerned, but it plays a good role as a pre classifier or training sample to predict the nature of classification based on nearest neighbour approach. So, the kMeans method may be preferred over Hierarchical method to form the training sample of the data provided.
If we initially apply ICA for dimension reduction and then perform clustering on the basis of important components, then on the basis of the Silhoutte values we see that even after reduction of dimension, Hierarchical technique performs better than kMeans for the non Gaussian data. Hence, it may be said that ICA is a proper dimension reduction technique for this data set and it does not hamper the nature of clustering that is obtained from the original data.

5.1 Physical properties of the groups for non gaussian data set:

As discussed in section 3.1 ,along with the clustering parameters, some additional parameters are derived viz., $\log M_{dyn}$, Tilt and α. When a system is in dynamical equilibrium its mass is called the Virial mass or dynamical mass and it is related to other observed parameters as

$$M_{dyn} = c\sigma^2 R_e/G \tag{22}$$

where, c is a constant and here its value is 3.8 (Hopkins 2008) and G is universal gravitational constant.When a system is in dynamical equilibrium, $logR_e$, $\log\sigma$ and μ_e form a particular parametric plane called the Virial plane(VP), and its form is

$$logR_e = 2log\sigma_e + 0.4\mu_e \tag{23}$$

for a constant mass to luminosity ratio.This follows from Virial Theorem(Goldstein,1989). Another parametric plane

$$logR_e = alog\sigma_e + b\mu_e \tag{24}$$

.

a,b being constants, is called the Fundamental plane(FP) for galaxies. When the angle between the Virial plane (VP) and fundamental plane (FP) is smallest, the tilt of the fundamental plane with respect to the Virial plane is highest and the galaxies are considered as most diversified. When the slope (α) of $logR_e$ with respect to $logM_{dyn}$ is 0.38, the formation of a galaxy by merger of small galaxies is dissipationless, i.e., dry merger or pure disc merger by which galaxies are speculated to be formed(Robertson et al. 2006).

When the slope is around 0.49 (Fraix Burnet et. al. 2010;Robertson 2006), the galaxies are considered to be formed by dissipational or wet merger. Wet merger is one in which sufficient gas is present during the merger process.

Now, from Figures 8 and 9 along with the cluster groups (HAEG1, HAEG2, HAEG3, HAEG4, HAEG5, HAEG6 and HAEG7) and (EMG1, EMG2,EMG3,EMG4,EMG5,EMG6 and EMG7)obtained from Table 1,it is very clear that the groups are well classified by Hierarchical clustering with Euclidean distance in logσ - logR_e plane, whereas the corresponding groups are overlapping in case of EM clustering. Also in EM clustering, the groups are not comparable in size e.g., EMG1 and EMG7 are negligibly small compared to the other clusters. If we arrange the groups according to the increasing value of Mg_2 indices for hierarchical groups, the sequence (viz. Table2) indicates that the magnesium indices are more or less increasing along with the Tilt values.

This is consistent with the fact that as the galaxies are becoming more diversified and dynamically old,the metalicity is also increasing in the medium as a result of episodic star formation events. Also more diversified galaxies are massive and are formed by dissipational process (viz., high α values for HAEG4, HAEG2, HAEG5, HAEG3), which is reflected in the logM_{dyn} and α values.

On the other hand, less massive galaxies (HAEG6,HAEG4) are less diversified and are formed by dry mergers (low M_{dyn} and α values).The corresponding sequence for EM clustering is shown in Table 3. Here the increasing indices of Magnesium does not guarantee increasing sequence of Tilt angles.Also, the dynamical mass sequence(viz M_{dyn}) does not follow any increasing trend which is contrary to the fact that massive galaxies form later in course of time (bottom up scenario). Moreover there are massive galaxies (viz., EMG6, EMG5) which are formed by dry merger contrary to the previous predictions (Fraix Burnet et al.2010; Shen et al.2005) by various authors that massive galaxies are the results of merger of non-discy type objects.

Subsequently it is investigated whether the Hierarchical clustering with the original data set is compatible with Hierarchical clustering on the basis of independent components. The groups are shown in Figures 8 and 10 respectively with respect to logσ and logR_e plane.

Studying Figures 8 and 10 and the groups (HAEG1,HAEG2 ,...., HAEG7) and (HIG1, HIG2,...., HIG7), it is clear that the figures are compatible in a sense that

HAEG1 $\cong$ HIG1, HAEG2+HAEG3 $\cong$ HIG2, HAEG7 $\cong$ HIG6, HAEG5 $\cong$ HIG7, HAEG6

$\cong$ HIG5, HAEG4 $\cong$ HIG3
This indicates that the groups of Hierarchical clustering with respect to several parameters and Hierarchical clustering with respect to independent components are almost equivalent and comparable in size.On the contrary,comparing Figures 9 and 10 no such parity is found. Hence, from this conjecture, one can at least speculate that for non Gaussian data set, non parametric clustering always plays better role than model based clustering and the result is again supported by clustering with respect to Independent Components.Similar conclusions can be drawn while comparing KMeans clustering along with various distances to EM clustering.

6. Result for Gaussian data set:

The Gaussian data set contains 13456 observations and 3 variables or parameters. Thus application of Hierarchical clustering may not reflect the proper clustering nature. Also the dendrogram will give a very clumsy picture due to the large number of observations. Hence for this data set only kMeans clustering with different distances and EM clustering have been applied.
The Sugar and James (2003) optimality criterion suggests that there are 7 clusters which is evident from the distortion and the jump curves of KMeans clustering with Euclidean and Manhattan distances (Figures 11 to 14).
The Distortion curve stabilises from the value k=7,suggesting that the optimum number of clusters is 7. The jump curve shows that the largest jump has been reached for k=7, again suggesting the optimum number of clusters to be 7.
Since the data is Gaussian, i.e., it has a proper distributional assumption, the model based clustering approach is appropriate in this case. Hence one would expect that EM clustering will perform well in this case. The method gives 5 clusters (i.e., it reflects the fact that the data set comes from 5 different mixtures of Gaussian distributions.).
As in the case of non Gaussian data, here two training samples have been formed based on the results of kMeans classification with Euclidean distance and EM classification.Roughly 33 percent of the data set consists of a training sample. So a random sample from each cluster of kMeans and EM method in the proportion of 0.33 formed a training sample. On the basis of this training sample, nearest neighborhood technique has been applied to the test data , i.e., the data set provided with the minimum number of neighborhood points for any point is taken as 5.
There is no optimum criterion to fix the minimum number of neighborhood points to be considered. However, by trial and error method, it is seen that for this particular data set, if the minimum number of points is 5, the post classification gives maximum similarity with the pre classification , i.e., kMeans and EM.
Effect of dimension reduction in clustering–EM clustering applied to Principal components extracted from data:
As discussed earlier,sometimes the data set is so large that there arises a need for dimension reduction at first and then a most suitable clustering technique may be applied to the reduced data. In this case, though the data is not really large in terms of dimension, still to check whether there is any effect of dimension reduction in clustering, PCA has been applied to the data (since the data is Gaussian). Using PCA, two independent components have been extracted and EM technique has been applied to the reduced set.
Nearest Neighbourhood classification with kMeans training sample resulted in 7 clusters

whereas with EM training sample resulted in 5 clusters.
As discussed in section 2.5, Principal Component Analysis has been applied as a dimension reduction technique when the data set is Gaussian. Experimental Evaluation showed that two of the eigen values are greater than 1, and hence two principal components have been extracted. Then EM clustering technique is applied to the reduced data set. It again resulted in 5 clusters with almost similar cluster sizes as the EM result for the original data set.
The distortion curve stabiles from the value k=7 ,suggesting that the optimum number of clusters is 7.

6.1 Statistical interpretation on the basis of Silhoutte index:

Finally, after applications of all the clustering and classification Methods, Silhoutte index for different clusters are calculated. As in the case of non-gaussian data, here also only average Silhoutte values are calculated.
The different Silhoutte indices for all techniques along with cluster names, sizes and descriptive measures are given in Table 4.
In Table 4,KEC1-KEC7 are groups by kMeans clustering with Euclidean distance, KMC1-KMC7 are groups by kMeans clustering with Manhattan distance, EMC1-EMC5 are groups by EM clustering, NKC1-NKC7 are nearest neighbourhood technique on kMeans clustering, NEC1-NEC5 are groups by nearest neighbourhood classification applied to EM clustering, EMPC1-EMPC5 are groups of EM clustering after finding principal components.
Table4 indicates that the Silhoutte values for KEC1, KEC3, KEC7 are greater than 0.1 indicating the observations are moderately classified. Silhoutte value for KEC2 is close to 0.3, so these observations are well classified. Only the Silhoutte values for KEC3, KEC4 and KEC6 are just greater than 0, indicating that the observations are just moderately clustered. Similar comparison shows that kMeans only results in a moderately well classification when applied to Gaussian data set.
Further Table 4 shows that the Silhoutte values for all the clusters except EMC5, lies between 0 to 0.18, indicating that the observations are moderately classified. Also the Silhoutte index for EMC5 is less than 0 implying that these observations are not at all properly classified. This means that though the data set is Gaussian, it is not a sample from mixtures of Gaussian distributions, and hence EM method is not giving satisfactory results.
Comparison of the Silhoutte values for the groups NKC1-NKC7 and NEC1-NEC5 show that post classification based on kMeans and EM reflect the same scenario as pre classification based on those two techniques. Overall, it can be said that kMeans clusters behave like a well justified training sample for the given data.
Again Table 4 shows that the Silhoutte values of EMPC1, EMPC2,.....,EMPC7 except EMPC5 lie between 0 to 0.18, indicating that the observations are moderately classified. Also, the Silhoutte index for EMPC5 is less than 0 implying that the observations are not all properly classified. This means that application of EM clustering on the extracted Principal Components gives similar results as application of EM clustering on the entire data.

6.2 Physical properties of groups for Gaussian data set:

After clustering by various parametric and non parametric methods we have seen in the previous sections that kMeans clustering gives a moderately well classification over the method of model based clustering e.g., EM algorithm.

Hence, we have confined our comparative discussions on a sample with KMeans clustering with Euclidean distance which is an example of non-parametric approach as well as with EM algorithm which describes model based clustering i.e., parametric approach.
In the present part, we have used and computed some more parameters to improve our conjecture which are absolute magnitude (M_v, Misgeld and Hilker 2011), k_1,k_2,k_3(Bender et al. 1992) and $< \mu_v >_h$.
Now, k_1,k_2,k_3 have been introduced by Bender et al. (1992) where k_1 is proportional to virial mass (mass required to keep a system in dynamical equilibrium), k_3 is proportional to mass to luminosity and k_2 is proportional to the effective surface brightness times M/L.
For calculating dynamical mass, value of velocity dispersion is needed which is not available for the present data set. So we have replaced the Virial mass by stellar mass and Virial mass to luminosity by stellar mass to luminosity.
Thus,the formulae for these parameters are,

$$k_1 = logM_{Stellar} \tag{25}$$

$$k_2 = logM_{Stellar} - logL_t + 3logI_s \tag{26}$$

.

$$k_3 = LogM_{Stellar} - logL_t \tag{27}$$

.
where L_t is the total Luminosity and

$$LogL_t = 0.4 * (4.83 - M_v) \tag{28}$$

I_s is the effective Luminosity and

$$logI_s = 0.4 * (4.83 - M_v) - log(2\pi) - 2logR_h \tag{29}$$

where logR_h is the logarithm of the radius at which half of the total luminosity of the galaxy is confined and is called half light radius(similar to $logR_e$.
The surface brightness average over half light radius is

$$< \mu_v >_h = 4.83 + 21.572 - 2.5logI_s \tag{30}$$

.

The average properties of all the parameters are shown in Table 4 for classes (KMC1, KMC2,...., KMC7) and (EMC1,EMC2,.....,EMC5) which represent the analyses with respect to KMeans clustering with Euclidean distance and EM clustering respectively.
The groups mentioned above indicate that in kMeans clustering, gEs and dEs are not at all well separated but fall in a group (viz., KEG2),whereas, they are separated in different groups for EM clustering viz., gEs in EMC5 and dEs in EMC1, EMC2 and EMC3 respectively.

Now,giant and dwarf ellipticals have different formation history and chemodynamical evolution. So it is very likely that they should be separated as is seen in Chattopadhyay et. al. 2012 (Here after C12).
Secondly, there are various theories about the formation of ultra compact dwarf galaxies (UCDs). One group claims that they are the result of agglomeration of a large number of globular clusters or they arise from mass threshing of nucleated dwarf galaxies (dE,N).
In a recent work it is conjectured (C12;Chattopadhyay and Karmakar 2012) that they can arise both from high metalicity globular clusters or from dwarf elliptical galaxies which are in particular metal deficient.
Now globular clusters are unique objects and independent of any particular environment. For kMeans classes (KEC1, KEC2,....,KEC7) of Table4, it is seen that all most all UCDs fall in one group with dE,N and GCs of Virgo Cluster.But it is very unlikely that UCDs originate only from the GCs of VCC and not from other GCs in other environment e.g.,GCs of our Galaxy or any galaxy in Local Group.
On the contrary, in EM method, UCDs and dE,N fall in two groups EMC4 and EMC5 along with GCs not only of VCC, but also of other environment, dEs and other compact objects. This is compatible with our previous results (C12, Goerdt et. al. 2008, Thomas et al. 2008) that UCDs might result from accumulation of GCs as well as from threshing of dE,Ns.
Since the number of GCs in the present sample is over dominating, so the results are some what fuzzy and clear indications will be reflected if the various objects are comparable in number. Moreover, the M/L ratio systematically increase in the groups found by EM and maximum M/L ratio systematically increase in the groups found by EM and maximum M/L occurs in groups EMC4 and EMC5 which contains UCDs, a result evident from high mass to light ratio for UCDs but not for GCs. But this phenomenon has not been reflected in KEG3 where M/L is not very different from the other groups.
It is clear from $\log R_e$ values in groups EMC1, EMC2,...., EMC7 of Table4 that the objects in EMC1 (mostly dwarf galaxies) are most extended with minimum surface density ($\log\Sigma$) and shallow surface brightness ($< \mu_v >_h$), whereas objects of EMC5 are brightest , compact with highest surface density and surface brightness (mostly , UCDs, dE, Ns, gEs, NuSc etc.). Such trend is absent for KMeans analysis. So the physical groups strongly support the statistical prediction, that for Gaussian data set, Model based clustering gives classification which is more physically interpretable.

7. Acknowledgement:

One of the authors Tuli De acknowledges to a research fellowship provided by The University Grants Commission , sanction No. UGC/157/JrFellow(sc) dated 31/03/2010.

References

[1] Anderberg, M.R.,1973. Cluster Analysis for Applications. New York: Academic Press.

[2] Anderson, T.W., An Introduction to Multivariate Statistical Analysis, 2nd Edition. (Wiley Series in Probability and Statistics).

[3] Bender,R.,Burstein,D.,Faber, S.M.,1992.ApJ,399,pp.462.

[4] Chattopadhyay,A.K.,Chattopadhyay,T.,De,T.,Mondal,S.,2012.IndependentComponent Analysis for dimension reduction classification: Hough transform and CASH Algorithm,Astrostatistical Challenges for the New Astronomy (ed. Joseph M Hilbe).Ch. 9,pp.183-200, Springer.

[5] Chattopadhyay, A.K., Mondal, S. and Chattopadhyay,T,2013. Independent Component Analysis for the objective classification of the globular Clusters of the galaxy NGC 5128, Computational Statistics and Data Analysis.57, pp.17-32.

[6] Chattopadhyay,T.,Misra,R.,Chattopadhyay,A.K.,Naskar,M.,2007.Statistical Evidence for Three classes of Gamma-ray Bursts,ApJ 667,pp. 1017-1023.

[7] Chattopadhyay, T., Sharina, M., Davoust,E., De, T., Chattopadhyay,A.K.,2012. Uncovering the formation of ultracompact dwarf galaxies by multivariate staistical analysis,ApJ 750 91. doi:10.1088/0004-637X/750/2/91.

[8] Chattopadhyay,T.,Karmakar,P.,2012,New Astronomy(in Press).

[9] Comon, P., 1994. Independent component analysis, A new concept of Signal Processing. 36, pp. 287-314.

[10] Cover, T.M., Hart, P.E.,1967.Nearest neighbor(NN) norms;NN pattern classification techniques,Washinton(DC):IEEE Trans on Inf Theory 13 (1),pp. 21-27.

[11] Cunninham,P.,Cord,M.,Delany,S.J.,2008.Machine Learning Techniques for Multimedia Case Studies on Organization and Retrieval(Eds.).Springer. XVI,290,pp.98 illus,ISBN 978-3-540-75170-0. http://www.springer.com/978-3-540-75170-0.

[12] Dasarathy,B.V.,Editor.1991.Nearest neighbor(NN) norms:NN pattern Classification techniques,Washinton(DC):IEEE Computer Society.

[13] Davies,D.L.,Bouldin,D.W.,1979.A cluster separation measure,IEEE Trans Pattern Anal.Machine Intell.1(4),pp.224-227.

[14] Dempster, A.P., Laird, N.M., Rubin, D.B.,1977. Maximum Likelihood from Incomplete Data via the EM Algorithm, Journal of the Royal Statistical Society. Series B (Methodological) 39 (1): 138. JSTOR 2984875. MR 0501537.

[15] Deza, E.,Deza,M.M.,2009. Encyclopedia of Distances, page 94, Springer.

[16] Dunn,J.C.,1974.Well separated clusters and optimal fuzzy partitions,J.Cybern,4.pp.95-104.

[17] Everitt, B.S.,1993. Cluster Analysis (3rd ed.) London: Edward Arnold.

[18] Fix,E.,Hodges,J.L.,1951.Discriminatory analysis:nonparametric discrimination:Consistency propeerties.USAF sch Aviat Med,4,pp.261-279.

[19] Fraix-Burnet,D.,Dugué, M.,Chattopadhyay, T.,Chattopadhyay, A.K.,Davoust, E.,2010.MN-RAS,407,2207.

[20] Goerdt,T.,Moore,B.,Kazentzidis,S.,Kaufmann,T.,Macciá, A.W.,Stadel,J.,2008.385,2136.

[21] Goldstein,H.,1989.Classical Mechanics(2nd ed.) Addison-Wesley.ISBN 0-201-02918-9.

[22] Gower,J.C.,1971. A General Coefficient of Similarity and Some of Its Properties, Biometrics.27(4), , pp 857-871.

[23] Hastie, T.,Tibshirani, R.,Friedman,J.,2009.The Elements of Statistical Learning (2nd ed.). New York: Springer. pp. 520528. ISBN 0-387-84857-6. http://www-stat.stanford.edu/ tibs/ElemStatLearn/. Retrieved 2009-10-20.

[24] Hopkins,P.F.,Cox, T.J., Hernquist, L.,2008.Dissipation and the Fundamental Plane:Observational Tests,ApJ 689 17 doi:10.1086/592105.

[25] Hudson, M.J.,Lucey, J.R., Smith,R.J.,Schlegel,D.J.,2001.Davies,R.L.,Streaming motions of galaxy clusters within 12 000 km s-1 -III.A Standardized Catalogue of Fundamental Plane data,MNRAS, 327,pp.265-295.

[26] Hyv*á*rinen,A.,Karhunen,J.,Oja,E.,2001. Independent Component Analysis,Wiley,New York.

[27] Kullback,S.,Leibler,R.A.,1951.On information and sufficiency.Annals of Mathematical Staistics,22,pp.79-86.

[28] MacQueen, J. B.,1967. Some Methods for classification and Analysis of Multivariate Observations,. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press. pp. 281297. MR 0214227. Zbl 0214.46201. http://projecteuclid.org/euclid.bsmsp/1200512992. Retrieved 2009-04-07.

[29] Martin, E., Hans-Peter, K.,Jrg, S., Xiaowei, X.,1996. A density-based algorithm for discovering clusters in large spatial databases with noise.,In Evangelos Simoudis, Jiawei Han, Usama M. Fayyad. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. pp. 226231. ISBN 1-57735-004-9. http://citeseerx.ist.psu.edu/viewdoc/summary/.doi=10.1.1.71.1980.

[30] Milligan,G.W.,1980.An examination of the effect of six types of error perturbation of fifteen clustering algorithms,Psychometrika,45(3),pp.325-342.

[31] Misgeld,I.,Hilker,M.,2011,MNRAS,414,3699.

[32] Morlini,I., Zani,S.,2012.Dissimilarity and similarity measures for comparing dendrograms and their applications,Advances in Data Analysis and Classification,6,pp.85-105.

[33] Robertson,B.,Cox,T.J.,Hernquist,L.,Franx,M.,Hopkins,P.F.,Martini,P.,Springel,V., 2006.The Fundamental Scaling Relations of Elliptical Galaxies,ApJ 641,pp.21-40(arXiv:astro-ph/0511053.

[34] Rousseeuw,Peter J.,1987. Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis,Computational and Applied Mathematics 20: 5365. DOI:10.1016/0377-0427(87)90125-7.

[35] Rubner,Y.,Tomasi,C.,Guibas,L.J.,2000.The earth mover's distance as a metric for image retrieval.International Journal of Computer Visio,40(2),pp.99-121.

[36] Shen,S.,Mo,H.J.,White,S.D.M.,Blanton,M.R.,Kauffmann,G.,Voges,W.,Brinkman,J., Csabai,I.,2003.The size distribution of galaxies in the Sloan Digital Sky Survey,MNRAS 343,pp.978-994(arXiv:astro-ph/0301527).

[37] Shepard,R.N.,1987.Toward a universal law of generalization for psychological science,Science,237,pp.1317-1228.

[38] Sugar,A.S., and James, G.M.,2003. Finding the number of clusters in a data set: An information theoretic approach, Journal of the American Statistical Association, 98,pp. 750763.

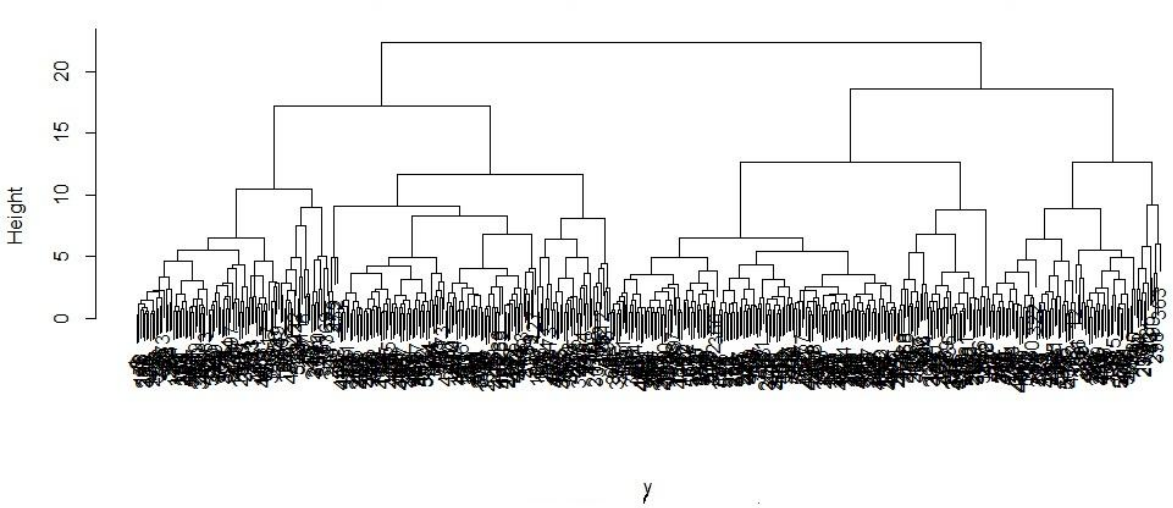

Figure 1: Divisive hierarchical clustering with Manhattan distance

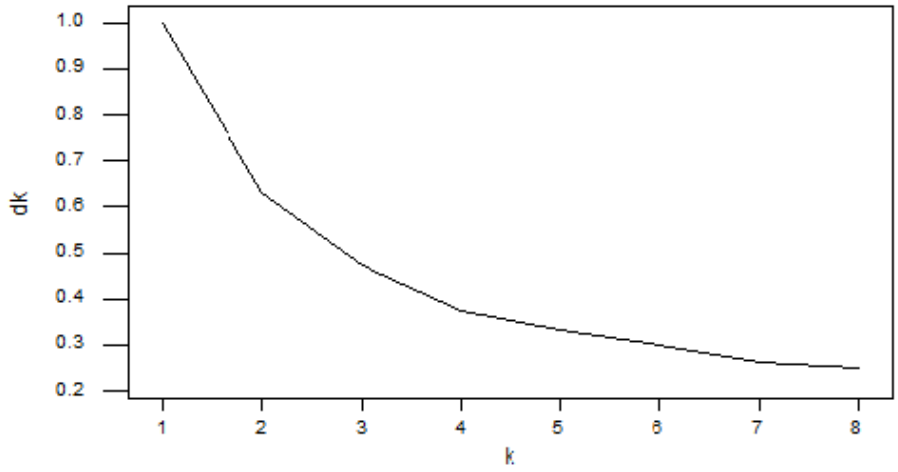

Figure 2: Distortion curve of kMeans clustering with Euclidean distance

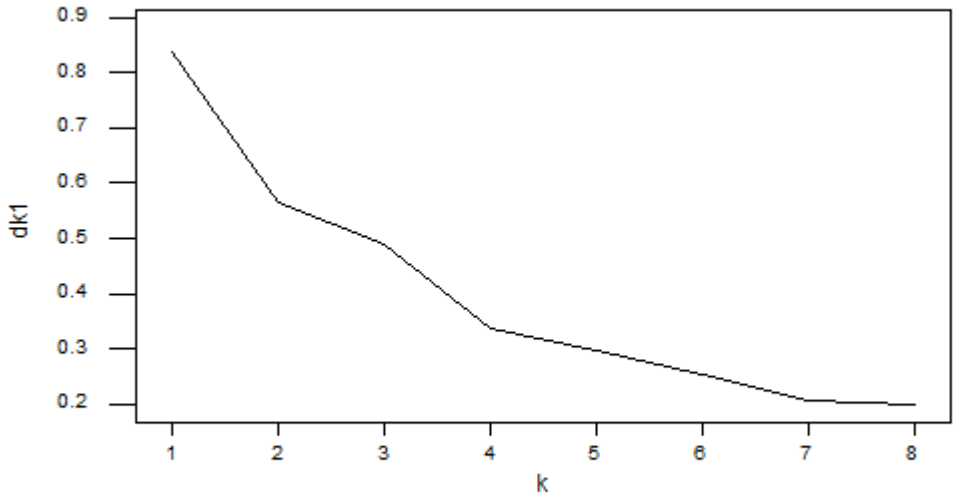

Figure 3: Distortion curve of kMeans clustering with Manhattan distance

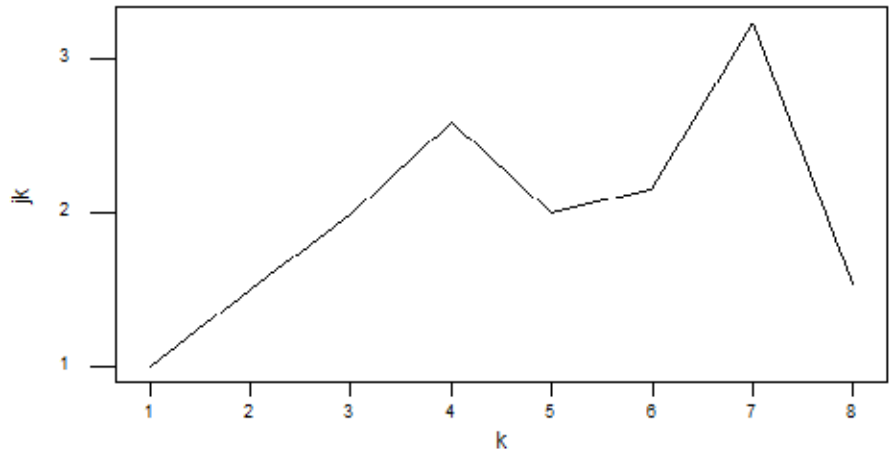

Figure 4: The jump curve of kMeans clustering with Euclidean distance

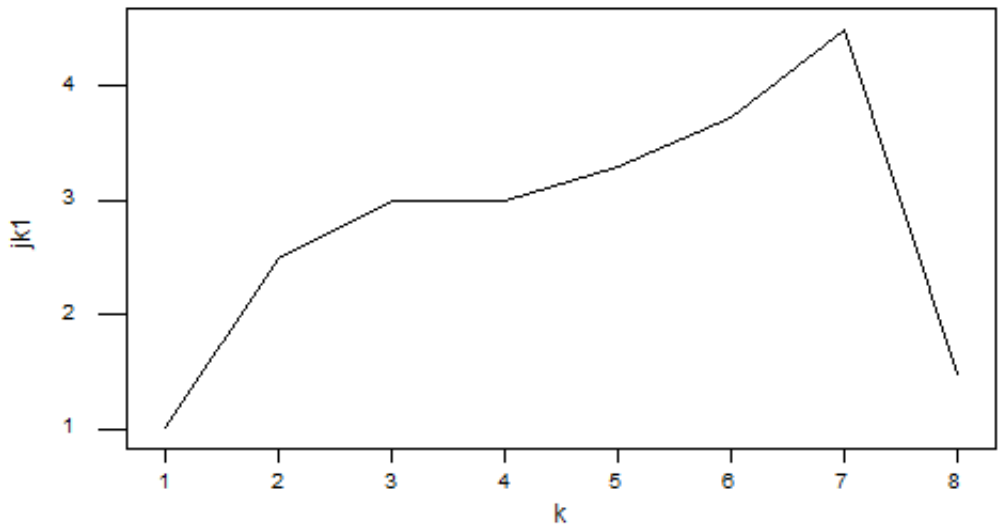

Figure 5: The jump curve of kMeans clustering with Manhattan distance

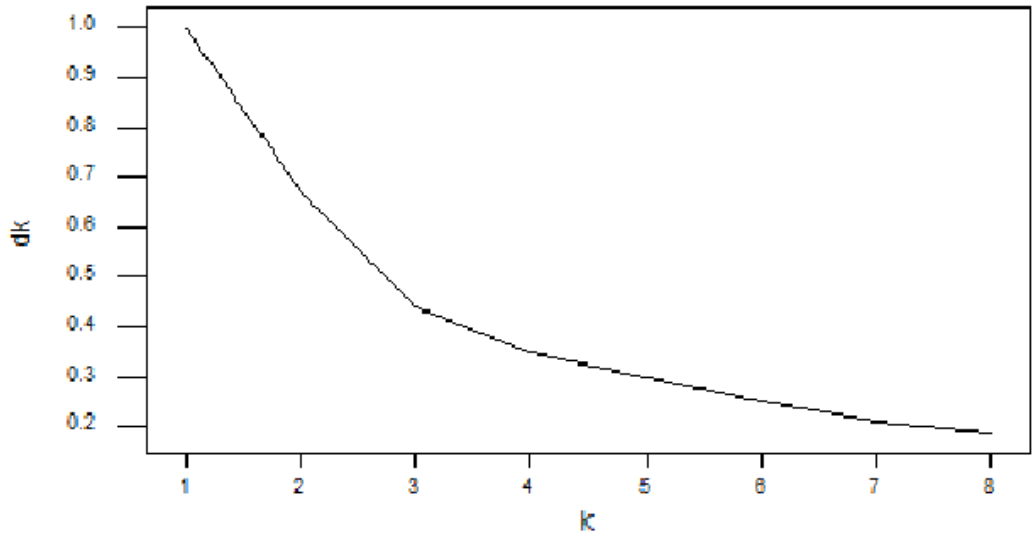

Figure 6: Distortion curve of kMeans applied to extracted independent components

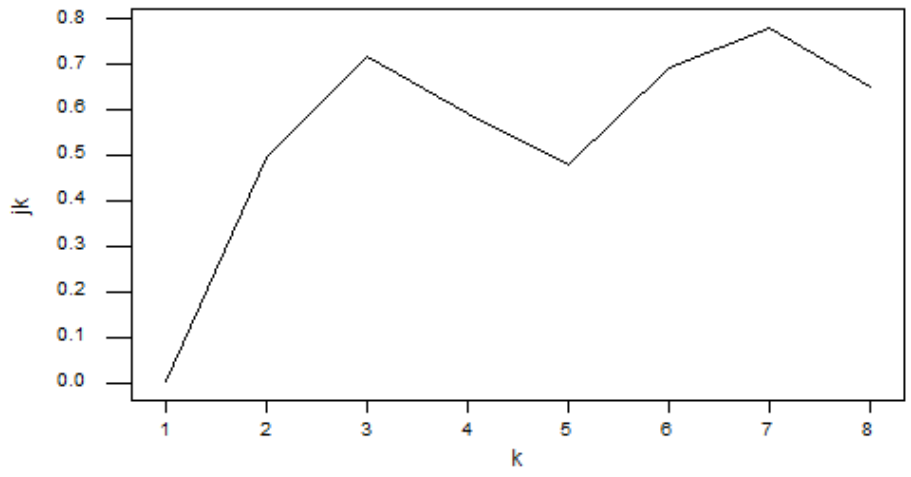

Figure 7: The Jump curve of kMeans technique applied to extracted independent components

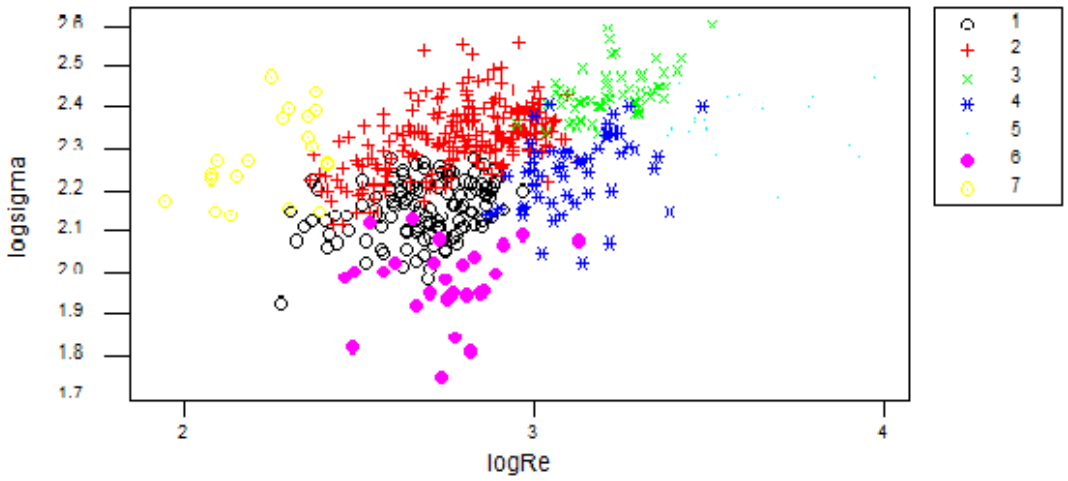

Figure 8: logR_e vs logσ diagram for Hierarchical agglomerative clustering with Euclidean distance

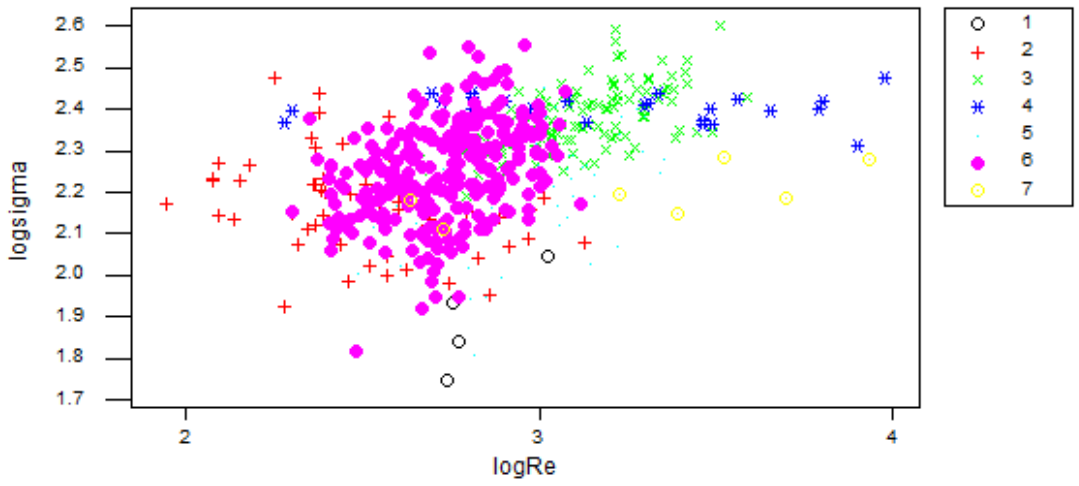

Figure 9: logR_e vs logσ diagram for EM Clustering

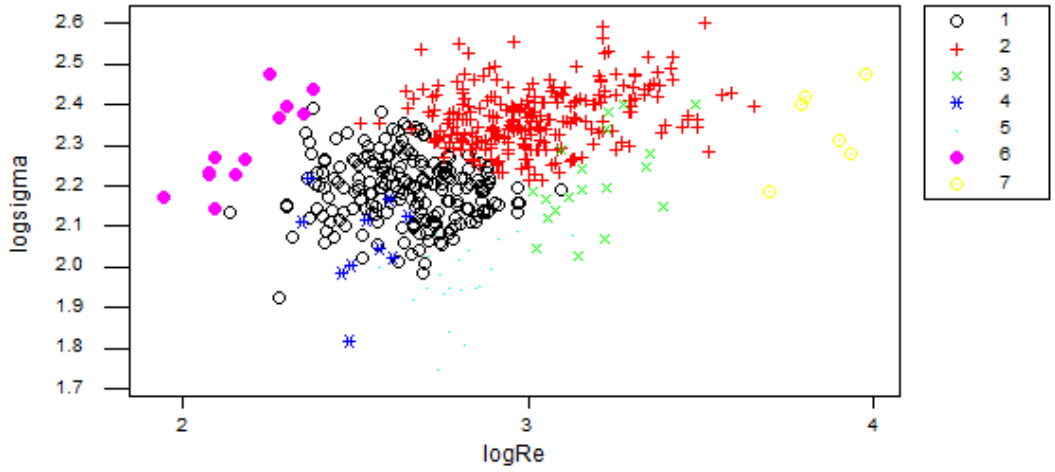

Figure 10: $\log R_e$ vs $\log\sigma$ diagram for Hierarchical Clustering applied to Independent Components

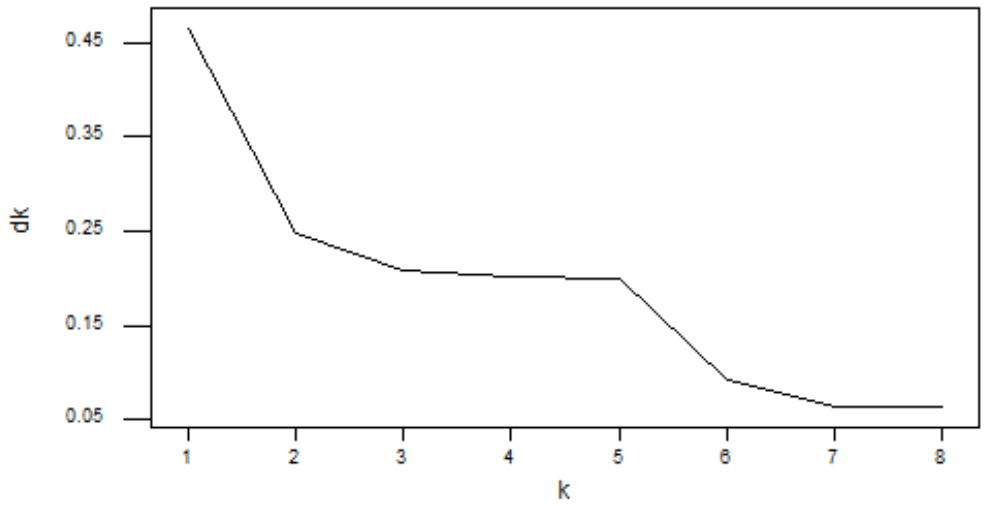

Figure 11: Distortion curve for kMeans clustering with Euclidean distance applied to the Gaussian data set

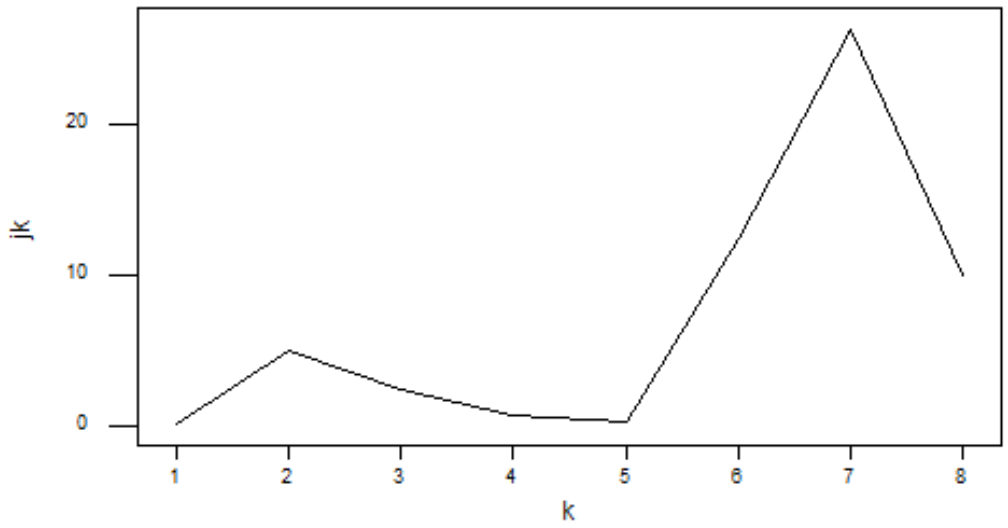

Figure 12: The jump curve for kMeans clustering with Euclidean distance applied to the Gaussian data set

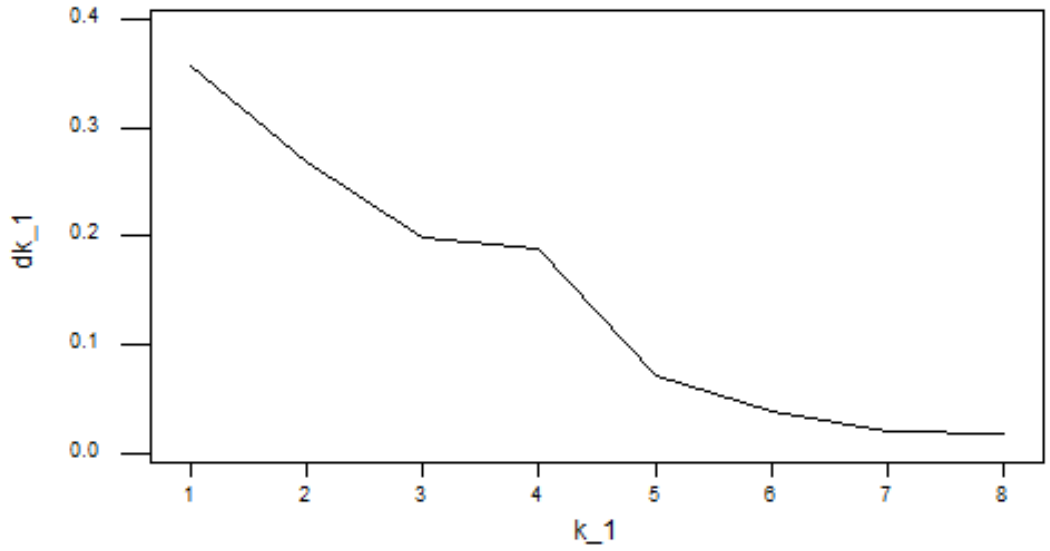

Figure 13: Distortion curve for kMeans clustering with Manhattan distance

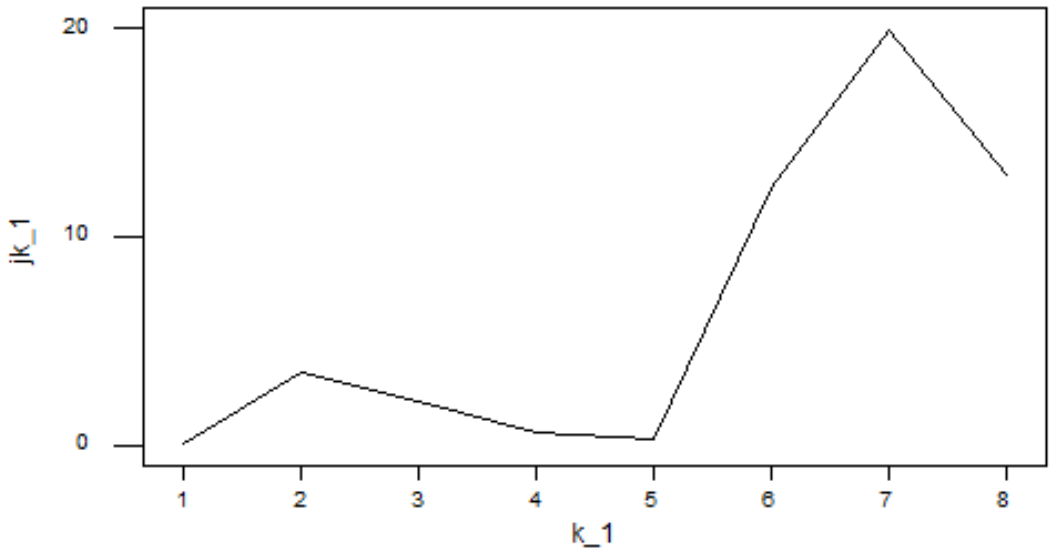

Figure 14: The Jump curve of kMeans clustering with Manhattan distance

Table 1: Clusters with sizes, Silhoutte index and descriptive measures:

Method	*Cluster Names*	*ClusterSizes*	*SilhoutteIndex*	α	*Tilt*	$logR_e$	
						Mean	*S.E.ofMean*
Hierarchical Agglomerative Euclidean	$HAEG1$	122	0.35030	0.52	0.49	2.6813	0.0133
	$HAEG2$	225	0.08299	0.55	0.66	2.7878	0.0121
	$HAEG3$	64	0.30030	0.50	0.58	3.2208	0.0139
	$HAEG4$	53	0.12760	0.40	0.60	3.1133	0.0184
	$HAEG5$	18	0.21040	0.65	0.83	3.6082	0.0459
	$HAEG6$	26	0.22330	0.41	0.41	2.7401	0.0309
	$HAEG7$	20	0.34710	1.37	0.42	2.2465	0.0314
Hierarchical Agglomerative Manhattan	$HAMG1$	379	0.11071	–	–	2.7780	0.0095
	$HAMG2$	86	0.31710	–	–	3.2741	0.0141
	$HAMG3$	27	0.24190	–	–	2.9390	0.0356
	$HAMG4$	8	0.37620	–	–	2.5227	0.0330
	$HAMG5$	21	0.39320	–	–	2.2526	0.0304
	$HAMG6$	6	0.57330	–	–	3.8481	0.0427
	$HAMG7$	1	*	–	–	2.2769	*
Hierarchical Divisive Euclidean	$HDEG1$	93	0.18930	–	–	2.7876	0.0155
	$HDEG2$	146	0.16323	–	–	2.9666	0.0098
	$HDEG3$	65	0.18470	–	–	3.2893	0.0134
	$HDEG4$	126	0.17868	–	–	2.5939	0.0119
	$HDEG5$	65	0.13860	–	–	2.8185	0.0179
	$HDEG6$	22	0.23500	–	–	2.2589	0.0297
	$HDEG7$	11	0.24480	–	–	3.7027	0.0583
Hierarchical Divisive Manhattan	$HDMG1$	99	0.18100	–	–	2.8069	0.0153
	$HDMG2$	148	0.20126	–	–	2.9436	0.0089
	$HDMG3$	75	0.17090	–	–	3.2667	0.0118
	$HDMG4$	107	0.09306	–	–	2.6279	0.0116
	$HDMG5$	13	0.22840	–	–	3.6716	0.0537
	$HDMG6$	38	0.13050	–	–	2.3487	0.0241
	$HDMG7$	48	0.12840	–	–	2.7067	0.0252
Hierarchical Agglomerative Gower's	$HAGG1$	87	0.14058	–	–	2.6644	0.0125
	$HAGG2$	140	0.08870	–	–	3.1327	0.0154
	$HAGG3$	47	0.12190	–	–	2.8232	0.0231
	$HAGG4$	74	0.14190	–	–	3.0287	0.0237
	$HAGG5$	97	0.15356	–	–	2.7728	0.0092
	$HAGG6$	77	0.08720	–	–	2.4336	0.0174
	$HAGG7$	6	0.49110	–	–	3.8481	0.0427
kMeans Euclidean	$KEG1$	22	0.25330	–	–	2.2539	0.0297
	$KEG2$	115	0.19700	–	–	2.5920	0.0119
	$KEG3$	72	0.10670	–	–	2.7009	0.0148
	$KEG4$	76	0.13740	–	–	3.3348	0.0218
	$KEG5$	27	0.13520	–	–	3.1059	0.0350
	$KEG6$	111	0.22250	–	–	2.9619	0.0109
	$KEG7$	105	0.12602	–	–	2.8412	0.0125
kMeans Manhattan	$KMG1$	51	0.13170	–	–	2.8549	0.0232
	$KMG2$	102	0.09897	–	–	2.6245	0.0131
	$KMG3$	127	0.12650	–	–	2.8424	0.0096
	$KMG4$	28	0.11680	–	–	3.5288	0.0365
	$KMG5$	68	0.13540	–	–	2.4429	0.0195
	$KMG6$	83	0.20200	–	–	2.9780	0.0139
	$KMG7$	69	0.21000	–	–	3.1922	0.0113
DBSCAN	$DBG1$	5	0.51040	–	–	2.8912	0.0154
	$DBG2$	5	0.41350	–	–	2.7871	0.0117
	$DBG3$	30	−0.01450	–	–	2.9432	0.0112
	$DBG4$	5	0.77800	–	–	2.7108	0.0085
	$DBG5$	8	0.46660	–	–	2.6779	0.0134
	$DBG6$	5	0.48770	–	–	2.8785	0.0234
	$DBG7$	470	−0.21858	–	–	2.8523	0.0146
EM	$EMG1$	4	0.4180	0.33	0.32	2.8214	0.0670
	$EMG2$	56	−0.0908	0.58	0.32	2.5059	0.0354
	$EMG3$	127	0.1311	0.56	0.58	3.1112	0.0164
	$EMG4$	24	−0.1480	0.99	0.85	3.2095	0.0973
	$EMG5$	48	−0.0587	0.42	0.57	2.9820	0.0306
	$EMG6$	262	0.4877	0.37	0.53	2.7351	0.0101
	$EMG7$	7	−0.0450	0.83	0.91	3.304	0.181
Nearest Neighbour Hierarchical	$NHG1$	188	0.19628	–	–	2.6353	0.0180
	$NHG2$	277	0.117050	–	–	3.0034	0.0114
	$NHG3$	9	0.251300	–	–	2.5125	0.0359
	$NHG4$	18	0.34900	–	–	3.6082	0.0459
	$NHG5$	21	0.27190	–	–	2.7861	0.0294
	$NHG6$	14	0.41730	–	–	2.2082	0.0370
	$NHG7$	1	*	–	–	2.2769	*
Nearest Neighbour kMeans	$NKG1$	19	0.27740	–	–	2.2389	0.0321
	$NKG2$	122	0.18260	–	–	2.5863	0.0118
	$NKG3$	73	0.08830	–	–	2.7322	0.0149
	$NKG4$	77	0.13110	–	–	3.3376	0.0219
	$NKG5$	11	0.30470	–	–	3.1378	0.0334
	$NKG6$	115	0.15320	–	–	2.9756	0.0126
	$NKG7$	111	0.13124	–	–	2.8415	0.0129
$Hierarchical - ICA$	$HIG1$	203	0.048869	0.49	0.61	2.6456	0.0113
	$HIG2$	258	0.1094913	0.64	0.72	3.0177	0.0131
	$HIG3$	19	0.171293	0.33	0.58	3.1906	0.0305
	$HIG4$	10	0.085234	0.16	0.058	2.5095	0.0323
	$HIG5$	20	0.27924	0.35	0.30	2.8013	0.0265
	$HIG6$	12	0.48442	0.37	0.37	2.1817	0.0379
	$HIG7$	6	0.51219	0.28	0.47	3.8481	0.0427
$kMeans - ICA$	$KIG1$	116	0.21967	–	–	2.6682	0.0120
	$KIG2$	119	0.19037	–	–	2.8367	0.0104
	$KIG3$	97	0.11290	–	–	2.9997	0.0117
	$KIG4$	44	0.13786	–	–	2.8191	0.0277
	$KIG5$	62	0.10558	–	–	3.2413	0.0127
	$KIG6$	22	0.26936	–	–	3.5604	0.0442
	$KIG7$	68	0.13814	–	–	2.4304	0.0196

Table 1 continued:

Method	*cluster names*	μ_e		$log\sigma$		Mg_2	
		Mean	*S.E.of Mean*	*Mean*	*S.E.of Mean*	*Mean*	*S.E.of Mean*
Hierarchical Agglomerative Euclidean	$HAEG1$	19.360	0.037	2.1507	0.0062	0.26164	0.00174
	$HAEG2$	19.120	0.028	2.3260	0.0053	0.29600	0.00121
	$HAEG3$	19.919	0.043	2.4332	0.0071	0.32050	0.00206
	$HAGEG4$	20.287	0.050	2.2433	0.0124	0.27258	0.00329
	$HAEG5$	21.369	0.125	2.3632	0.0168	0.31744	0.00474
	$HAEG6$	20.060	0.121	1.9786	0.0188	0.21146	0.00456
	$HAEG7$	17.702	0.082	2.2795	0.0230	0.28405	0.00405
Hierarchical Agglomerative Manhattan	$HAMG1$	19.278	0.024	2.2649	0.0058	0.28385	0.00128
	$HAMG2$	20.223	0.052	2.3986	0.0086	0.31605	0.00228
	$HAMG3$	20.526	0.067	2.0284	0.0234	0.22719	0.00493
	$HAMG4$	19.125	0.107	2.0705	0.0449	0.19638	0.00700
	$HAMG5$	17.722	0.081	2.2795	0.0219	0.28490	0.00395
	$HAMG6$	21.988	0.139	2.3453	0.0436	0.30350	0.00911
	$HAMG7$	19.300	*	1.9250	*	0.2800	*
Hierarchical Divisive Euclidean	$HDEG1$	19.886	0.051	2.0867	0.0103	0.241125	0.00277
	$HDEG2$	19.576	0.028	2.3263	0.0052	0.29181	0.00124
	$HDEG3$	20.288	0.046	2.3968	0.0099	0.31645	0.00275
	$HDEG4$	18.957	0.029	2.2048	0.0066	0.27526	0.00163
	$HDEG5$	18.956	0.038	2.4062	0.0084	0.31754	0.00212
	$HDEG6$	17.796	0.081	2.2775	0.0210	0.28405	0.00392
	$HDEG7$	21.710	0.122	2.3340	0.0314	0.30664	0.00777
Hierarchical Divisive Manhattan	$HDMG1$	19.919	0.046	2.0933	0.0100	0.24632	0.00252
	$HDMG2$	19.495	0.026	2.3366	0.0048	0.29503	0.00131
	$HDMG3$	20.188	0.043	2.3975	0.0098	0.31412	0.00259
	$HDMG4$	19.013	0.027	2.2087	0.0070	0.27070	0.00225
	$HDMG5$	21.614	0.122	2.3417	0.0272	0.31131	0.00726
	$HDMG6$	18.222	0.077	2.2139	0.0129	0.28026	0.00293
	$HDMG7$	18.684	0.067	2.4109	0.0091	0.31829	0.00244
Hierarchical Agglomerative Gower's	$HAGG1$	19.251	0.0427	2.1592	0.0079	0.25755	0.00248
	$HAGG2$	19.847	0.038	2.3923	0.0055	0.31357	0.00137
	$HAGG3$	20.164	0.046	2.0418	0.0157	0.23060	0.00325
	$HAGG4$	19.949	0.066	2.2644	0.0081	0.27489	0.00168
	$HAGG5$	19.018	0.047	2.3475	0.0079	0.30375	0.00201
	$HAGG6$	18.468	0.031	2.2259	0.0105	0.28370	0.00174
	$HAGG7$	21.988	0.062	2.3453	0.0436	0.30350	0.00911
kMeans Euclidean	$KEG1$	17.746	0.081	2.2775	0.0210	0.28382	0.00392
	$KEG2$	18.934	0.030	2.2094	0.0063	0.27619	0.00162
	$KEG3$	19.703	0.055	2.0590	0.0112	0.23876	0.00317
	$KEG4$	20.351	0.075	2.4151	0.0076	0.31943	0.00215
	$KEG5$	20.586	0.075	2.1581	0.0132	0.25081	0.00510
	$KEG6$	19.684	0.027	2.2938	0.0055	0.28696	0.00139
	$KEG7$	19.068	0.031	2.3868	0.0056	0.30905	0.00172
kMeans Manhattan	$KMG1$	20.220	0.061	2.0920	0.0143	0.23361	0.00322
	$KMG2$	19.181	0.040	2.1437	0.0080	0.25967	0.00231
	$KMG3$	19.122	0.026	2.3669	0.0057	0.30431	0.00158
	$KMG4$	21.040	0.0117	2.4005	0.0159	0.32168	0.00454
	$KMG5$	18.395	0.060	2.2610	0.0093	0.28053	0.00196
	$KMG6$	19.826	0.036	2.2650	0.0063	0.28053	0.00171
	$KMG7$	19.918	0.038	2.4090	0.0074	0.31100	0.00217
DBSCAN	$DBG1$	19.116	0.036	2.3210	0.0085	0.29120	0.00097
	$DBG2$	19.402	0.027	2.3420	0.0100	0.2990	0.00176
	$DBG3$	19.520	0.022	2.3304	0.0103	0.29357	0.00143
	$DBG4$	19.464	0.025	2.1030	0.0089	0.26260	0.00160
	$DBG5$	18.863	0.052	2.3330	0.0053	0.30075	0.00096
	$DBG6$	19.184	0.049	2.4032	0.0045	0.30460	0.00051
	$DBG7$	19.476	0.037	2.2661	0.0064	0.28415	0.00155
EM	$EMG1$	20.635	0.181	1.8928	0.0640	0.24275	0.00728
	$EMG2$	19.104	0.144	2.1653	0.0158	0.25880	0.00490
	$EMG3$	19.722	0.052	2.3781	0.0068	0.30688	0.00184
	$EMG4$	19.959	0.286	2.4042	0.0069	0.31521	0.00499
	$EMG5$	19.829	0.080	2.1960	0.0198	0.24477	0.00532
	$EMG6$	19.238	0.031	2.2525	0.0073	0.28566	0.00143
	$EMG7$	21.106	0.409	2.1973	0.0148	0.2790	0.0148
Nearest Neighbour Hierarchical	$NHG1$	19.131	0.036	2.1891	0.0058	0.27195	0.00146
	$NHG2$	19.616	0.558	2.3507	0.0055	0.29985	0.00148
	$NHG3$	18.932	0.034	2.0890	0.0263	0.20111	0.008072
	$NHG4$	18.932	0.138	2.3632	0.0168	0.31744	0.00474
	$NHG5$	21.369	0.125	1.9612	0.0207	0.21724	0.00448
	$NHG6$	17.541	0.086	2.3084	0.0272	0.28493	0.00575
	$NHG7$	19.300	*	1.9250	*	0.28000	*
Nearest Neighbour kMeans	$NKG1$	17.698	0.090	2.2746	0.0957	0.28337	0.00421
	$NKG2$	18.921	0.030	2.2068	0.0704	0.27511	0.00170
	$NKG3$	19.816	0.055	2.0604	0.0941	0.23827	0.00325
	$NKG4$	20.386	0.076	2.4061	0.0768	0.31874	0.00201
	$NKG5$	20.785	0.079	2.1277	0.0623	0.24655	0.00679
	$NKG6$	19.738	0.027	2.2906	0.0668	0.28439	0.00156
	$NKG7$	19.069	0.031	2.3832	0.0597	0.30853	0.00174
$Hierarchical - ICA$	$HIG1$	19.163	0.038	2.1898	0.0057	0.27188	0.00139
	$HIG2$	19.602	0.037	2.3692	0.0045	0.30536	0.00133
	$HIG3$	20.553	0.082	2.2128	0.0261	0.25889	0.00456
	$HIG4$	18.975	0.131	2.0620	0.0358	0.20360	0.00737
	$HIG5$	20.327	0.086	1.9684	0.0204	0.21725	0.00471
	$HIG6$	17.499	0.094	2.2998	0.0309	0.28783	0.00634
	$HIG7$	21.988	0.139	2.3453	0.0436	0.30350	0.00911
$KMeans - ICA$	$KIG1$	19.295	0.034	2.1555	0.0065	0.26345	0.00159
	$KIG2$	19.085	0.027	2.3741	0.0057	0.30544	0.00169
	$KIG3$	19.810	0.031	2.2895	0.0061	0.28543	0.00166
	$KIG4$	20.163	0.081	2.0531	0.0160	0.22332	0.00379
	$KIG5$	20.015	0.045	2.4294	0.0078	0.32171	0.00199
	$KIG6$	21.268	0.117	2.3340	0.0182	0.30573	0.00582
	$KIG7$	18.369	0.063	2.2549	0.0078	0.28613	0.00196

Table 1 continued:

Method	*cluster names*	$LogM_{dyn}$	
		Mean	*S.E.ofMean*
Hierarchical Agglomerative Euclidean	$HAEG1$	9.9139	0.0207
	$HAEG2$	10.367	0.019
	$HAEG3$	11.018	0.024
	$HAGEG4$	10.531	0.036
	$HAEG5$	11.856	0.057
	$HAEG6$	9.6284	0.0515
	$HAEG7$	9.7366	0.0665
Hierarchical Agglomerative Manhattan	$HAMG1$	–	–
	$HAMG2$	–	–
	$HAMG3$	–	–
	$HAMG4$	–	–
	$HAMG5$	–	–
	$HAMG6$	–	–
	$HAMG7$	–	–
Hierarchical Divisive Euclidean	$HDEG1$	–	–
	$HDEG2$	–	–
	$HDEG3$	–	–
	$HDEG4$	–	–
	$HDEG5$	–	–
	$HDEG6$	–	–
	$HDEG7$	–	–
Hierarchical Divisive Manhattan	$HDMG1$	–	–
	$HDMG2$	–	–
	$HDMG3$	–	–
	$HDMG4$	–	–
	$HDMG5$	–	–
	$HDMG6$	–	–
	$HDMG7$	–	–
Hierarchical Agglomerative Gower's	$HAGG1$	–	–
	$HAGG2$	–	–
	$HAGG3$	–	–
	$HAGG4$	–	–
	$HAGG5$	–	–
	$HAGG6$	–	–
	$HAGG7$	–	–
kMeans Euclidean	$KEG1$	–	–
	$KEG2$	–	–
	$KEG3$	–	–
	$KEG4$	–	–
	$KEG5$	–	–
	$KEG6$	–	–
	$KEG7$	–	–
kMeans Manhattan	$KMG1$	–	–
	$KMG2$	–	–
	$KMG3$	–	–
	$KMG4$	–	–
	$KMG5$	–	–
	$KMG6$	–	–
	$KMG7$	–	–
DBSCAN	$DBG1$	–	–
	$DBG2$	–	–
	$DBG3$	–	–
	$DBG4$	–	–
	$DBG5$	–	–
	$DBG6$	–	–
	$DBG7$	–	–
EM	$EMG1$	9.538	0.187
	$EMG2$	9.7677	0.0383
	$EMG3$	10.799	0.027
	$EMG4$	10.949	0.097
	$EMG5$	10.305	0.062
	$EMG6$	10.171	0.020
	$EMG7$	10.629	0.216
Nearest Neighbour Hierarchical	$NHG1$	–	–
	$NHG2$	–	–
	$NHG3$	–	–
	$NHG4$	–	–
	$NHG5$	–	–
	$NHG6$	–	–
	$NHG7$	–	–
Nearest Neighbour kMeans	$NKG1$	–	–
	$NKG2$	–	–
	$NKG3$	–	–
	$NKG4$	–	–
	$NKG5$	–	–
	$NKG6$	–	–
	$NKG7$	–	–
Hierarchical − ICA	$HIG1$	9.9505	0.0156
	$HIG2$	10.687	0.018
	$HIG3$	10.548	0.073
	$HIG4$	9.5647	0.0774
	$HIG5$	9.6692	0.0561
	$HIG6$	9.7124	0.0965
	$HIG7$	11.470	0.114
KMeans − ICA	$KIG1$	–	–
	$KIG2$	–	–
	$KIG3$	–	–
	$KIG4$	–	–
	$KIG5$	–	–
	$KIG6$	–	–
	$KIG7$	–	–

Column 1: Names of different clustering and classification techniques
Column 2 represents groups found by different clustering and classification techniques.

Table 2: Sequence of merging for Hierarchical agglomerative clustering:

*	HAEG6	HAEG7	HAEG1	HAEG4	HAEG2	HAEG5	HAEG3
Mg_2	0.21146	0.28405	0.26164	0.27258	0.29679	0.31744	0.32050
Tilt	0.42	0.49	0.60	0.66	0.83	0.58	0.58
$logM_{dyn}$	9.62	9.73	9.91	10.531	10.367	11.856	11.018
α	0.41	0.42	0.49	0.60	0.66	0.83	0.58

Table 3: Sequence of merging for EM clustering:

*	EMG1	EMG5	EMG2	EMG7	EMG6	EMG3	EMG4
Mg_2	0.24275	0.24477	0.25880	0.27900	0.28566	0.30688	0.31521
Tilt	0.32	0.57	0.63	0.91	0.53	0.58	0.85
$logM_{dyn}$	9.53	10.30	9.77	10.63	10.17	10.79	10.95
α	0.33	0.42	0.58	0.83	0.37	0.56	0.99

Table 4: Clusters with sizes, Silhoutte index and descriptive measures for gaussian data:

Method	*Cluster Names*	*Cluster Sizes*	*Silhoutte Index*	$LogR_e$		$logMsteller$	
				Mean	*S.E.ofMean*	*Mean*	*S.E.ofMean*
kMeans Euclilidean	$KEC1$	3372	0.18763	0.38305	0.00210	5.4515	0.0030
	$KEC2$	403	0.26526	3.0667	0.0234	8.9884	0.0875
	$KEC3$	939	0.08894	0.43653	0.00713	6.3833	0.0109
	$KEC4$	1925	0.05200	0.56019	0.00473	4.7604	0.0057
	$KEC5$	2267	0.17267	0.35511	0.00225	5.0099	0.0035
	$KEC6$	1919	0.09434	0.64544	0.00257	5.3764	0.0048
	$KEC7$	2631	0.16334	0.37438	0.00280	5.8375	0.0036
kMeans Manhattan	$KMC1$	1554	0.03039	0.43005	0.00700	6.2689	0.0124
	$KMC2$	1726	0.16216	0.26088	0.00241	5.2328	0.0041
	$KMC3$	1441	−0.07055	1.3726	0.0279	6.2485	0.0463
	$KMC4$	2083	0.01036	0.54136	0.00661	5.6363	0.0129
	$KMC5$	1920	0.05206	0.53018	0.00401	4.7338	0.0051
	$KMC6$	2319	0.23496	0.35961	0.00243	5.7429	0.0032
	$KMC7$	2413	0.23526	0.42816	0.00209	5.0854	0.0035
EM clustering	$EMC1$	190	0.1785	2.0429	0.0682	5.7566	0.0972
	$EMC2$	3895	0.0523	0.55845	0.00489	4.9757	0.0067
	$EMC3$	4225	0.04864	0.45453	0.00440	5.3576	0.0060
	$EMC4$	4227	0.05147	0.40580	0.00463	5.7963	0.0067
	$EMC5$	919	−0.02526	0.9199	0.0390	7.1660	0.0577
Nearest Neighbour kMeans	$NKC1$	3378	0.81447	0.38341	0.00208	5.4528	0.0030
	$NKC2$	358	0.30518	3.1211	0.0243	9.2791	0.0862
	$NKC3$	920	0.07325	0.44303	0.00813	6.3986	0.0121
	$NKC4$	1971	0.03811	0.60200	0.00814	4.7944	0.0077
	$NKC5$	2249	0.17177	0.35414	0.00227	5.0086	0.0035
	$NKC6$	1937	0.05474	0.64508	0.00274	5.3775	0.0048
	$NKC7$	2643	0.17291	0.37484	0.00282	5.8410	0.0037
NearestNeighbour EM	$NEC1$	121	0.55109	2.7203	0.0240	6.6171	0.0741
	$NEC2$	3987	0.05616	0.56241	0.00483	4.9634	0.0067
	$NEC3$	4222	0.04823	0.45506	0.00443	5.3605	0.0060
	$NEC4$	4270	0.06187	0.40352	0.00450	5.8005	0.0066
	$NEC5$	85	−0.017628	0.9680	0.0416	7.2504	0.0612
EM − PC	$EMPC1$	184	0.16482	2.0277	0.0792	5.4904	0.0162
	$EMPC2$	3901	0.0299	0.48685	0.00385	4.3291	0.0055
	$EMPC3$	4222	0.031340	0.45543	0.0293	5.8443	0.0029
	$EMPC4$	4235	0.03875	0.39816	0.00481	5.8443	0.0029
	$EMPC5$	914	−0.01192	0.9298	0.0277	7.8910	0.0592

Table 4 continued:

Method	*Cluster Names*	$log\Sigma$		M_v		$< \mu_v >_h$	
		Mean	*S.E.ofMean*	*Mean*	*S.E.ofMean*	*Mean*	*S.E.ofMean*
kMeans Euclilidean	$KEC1$	3.8873	0.0027	−7.6755	0.0098	17.807	0.008
	$KEC2$	2.0569	0.0600	−16.445	0.206	22.455	0.137
	$KEC3$	4.7121	0.0098	−9.8656	0.0290	15.884	0.024
	$KEC4$	2.8419	0.0095	−6.1308	0.0158	20.237	0.022
	$KEC5$	3.5015	0.0037	−6.6166	0.0111	18.726	0.011
	$KEC6$	3.2874	0.0043	−7.5006	0.0143	19.293	0.014
	$KEC7$	4.2906	0.0038	−8.5664	0.0121	16.872	0.010
kMeans Manhattan	$KMC1$	4.6106	0.0071	−9.5974	0.0316	–	–
	$KMC2$	3.9129	0.0040	−7.0812	0.0132	–	–
	$KMC3$	2.7952	0.0215	−9.644	0.114	–	–
	$KMC4$	3.7553	0.0031	−8.1787	0.0317	–	–
	$KMC5$	2.8753	0.0091	−6.0604	0.0145	–	–
	$KMC6$	4.2255	0.0034	−8.3359	0.0115	–	–
	$KMC7$	3.4309	0.0028	−6.8250	0.0111	–	–
EM clustering	$EMC1$	0.8726	0.0632	−9.387	0.201	24.395	0.219
	$EMC2$	3.0606	0.0050	−6.5556	0.0172	19.604	0.013
	$EMC3$	3.6503	0.0037	−7.4684	0.0160	18.371	0.010
	$EMC4$	4.1865	0.0042	−8.4805	0.0179	17.114	0.011
	$EMC5$	4.5280	0.0221	−11.739	0.0142	16.428	0.058
Nearest Neighbour KMeans	$NKC1$	3.8877	0.0027	−7.6797	0.0098	–	–
	$NKC2$	2.2328	0.0580	−17.133	0.202	–	–
	$NKC3$	4.7143	0.0103	−9.8978	0.0098	–	–
	$NKC4$	2.7922	0.0124	−6.2174	0.0207	–	–
	$NKC5$	3.2891	0.0038	−6.6127	0.0110	–	–
	$NKC6$	3.2891	0.0045	−7.5042	0.0143	–	–
	$NKC7$	4.2932	0.0038	−8.5747	0.0123	–	–
NearestNeighbour EM	$NEC1$	0.3783	0.0472	−10.918	0.171	–	–
	$NEC2$	3.0404	0.0058	−6.5578	0.0172	–	–
	$NEC3$	3.6522	0.0037	−7.4765	0.0160	–	–
	$NEC4$	4.1952	0.0042	−8.4893	0.0175	–	–
	$NEC5$	4.5162	0.0239	−11.944	0.151	–	–
$EM\ -PC$	$EMPC1$	0.7585	0.0782	−9.870	0.2083	–	–
	$EMPC2$	3.7832	0.0093	−6.3205	0.0237	–	–
	$EMPC3$	4.2072	0.0023	−7.0290	0.0184	–	–
	$EMPC4$	4.2072	0.0023	−8.3997	0.0122	–	–
	$EMPC5$	4.6606	0.0024	−10.804	0.0194	–	–

Table 4 continued:

Method	*Cluster Names*	$M \backsim k_1$		k_2		$M/L \backsim k_3$	
		Mean	*S.E.ofMean*	*Mean*	*S.E.ofMean*	*Mean*	*S.E.ofMean*
kMeans Euclilidean	$KEC1$	5.4515	0.0030	10.763	0.009	0.44934	0.0023
	$KEC2$	8.9884	0.0875	5.214	0.169	0.47828	0.00172
	$KEC3$	6.3833	0.0109	13.126	0.029	0.50509	0.0045
	$KEC4$	4.7604	0.0057	7.7733	0.0262	0.37612	0.0054
	$KEC5$	5.0099	0.0035	9.6420	0.0121	0.43122	0.0029
	$KEC6$	5.3764	0.0048	8.9738	0.0152	0.44417	0.0035
	$KEC7$	5.8375	0.0036	11.914	0.011	0.47897	0.0027
kMeans Manhattan	$KMC1$	–	-	-	-	-	-
	$KMC2$	–	-	-	-	-	-
	$KMC3$	–	-	-	-	-	-
	$KMC4$	–	-	-	-	-	-
	$KMC5$	–	-	-	-	-	-
	$KMC6$	–	-	-	-	-	-
	$KMC7$	–	-	-	-	-	-
EM clustering	$EMC1$	5.7566	0.0972	2.478	0.235	0.0698	0.0385
	$EMC2$	4.9757	0.0067	8.3389	0.0153	0.42142	0.00237
	$EMC3$	5.3576	0.0060	10.075	0.012	0.43818	0.00206
	$EMC4$	5.7963	0.0067	11.615	0.013	0.47208	0.00211
	$EMC5$	7.1660	0.0577	12.507	0.068	0.53849	0.00433
Nearest Neighbour kMeans	$NKC1$	–	–	–	–	–	–
	$NKC2$	–	–	–	–	–	–
	$NKC3$	–	–	–	–	–	–
	$NKC4$	–	–	–	–	–	–
	$NKC5$	–	–	–	–	–	–
	$NKC6$	–	–	–	–	–	–
	$NKC7$	–	–	–	–	–	–
NearestNeighbour EM	$NEC1$	–	–	–	–	–	–
	$NEC2$	–	–	–	–	–	–
	$NEC3$	–	–	–	–	–	–
	$NEC4$	–	–	–	–	–	–
	$NEC5$	–	–	–	–	–	–
$EM\ -PC$	$EMPC1$	–	–	–	–	–	–
	$EMPC2$	–	–	–	–	–	–
	$EMPC3$	–	–	–	–	–	–
	$EMPC4$	–	–	–	–	–	–
	$EMPC5$	–	–	–	–	–	–

Column 1: Names of different clustering and classification techniques
Column 2 represents groups found by different clustering and classification techniques.