Journal of Al Azhar University-Gaza (Natural Sciences), 2011, 13 : 109-118

Estimating the Number of Clusters in Genetics of Acute Lymphoblastic Leukemia Data

Mahmoud K. Okasha, Khaled I.A. Almghari

Department of Applied Statistics
Al-Azhar University - Gaza

Received 16/11/2011 Accepted 31/12/2011

Abstract: Cluster analysis is a statistical technique that has been widely used for the analysis of genetic data to cluster gene expressions and other data in many fields. However, the problem encountered in the literature is the choice of the number of clusters. Specifically, the problem of estimating the number of clusters in a given population particularly for gene expressions is of a great interest and needs to be addressed. Many algorithms are used in practice for that purpose in different fields. In this paper we examined different clustering algorithms, for estimating the number of clusters, that are based on probabilities, covariance matrix, and eigenvalues on real data sets using R package algorithms. Specifically, we examined the model based algorithm (Mclust) and hierarchical clustering algorithm (hclust) and compared these algorithms with Partition Around Medoid (PAM) algorithm. The results we found are that the first algorithm can be used only for large data sets and the second one can be safely used for small data sets. The Mclust is a model based clustering approach built on Bayesian Information Criterion (BIC) which maximizes (EM) algorithm. The results of these two algorithms are compared with a third approach based on Partition Around Medoid (PAM) algorithm but selects the number of clusters manually according to the average silhouette width and selecting the number of clusters as that number which maximizes the average silhouette width. The later algorithm although allows to estimate the number of clusters manually, it has the best performance. However, the first two algorithms can be automated to produce the best estimate for the number of clusters in a given data set. These algorithms can be applied not only for genetic data but also for many other fields such as market research.

Keywords: *clustering, model based algorithm, hierarchical clustering, Partition Around Medoid, Bayesian Information Criterion, average silhouette, hierarchical tree, gene expression.*

1. Introduction

Cluster analysis is a collection of statistical methods which are used to detect groups of observations that have similar behavior or characteristics in a set of data. Cluster analysis is generally classified into two different techniques; namely hierarchical and non-hierarchical procedures. The goal is to construct a hierarchy or a decision-tree like structure (dendogram) to illustrate the relationship among entities. In the non-hierarchical method a position in the measurement is taken as a central place and the distance is measured from such central point (Partition Around Medoid). In the hierarchical clustering, the concept of ordering is involved in this approach. The ordering is a driven by the number of observations that can be combined at a time based on the assumptions that the distance between two observations is not statistically different from zero. The clusters could be arrived at either from weeding out observations (divisive method) or joining together similar observations (agglomerative method). However, estimating the number of clusters in any data remains the main problem (Chen et al., 2002).

2. Aims of the study

Acute lymphoblastic leukemia disease has many different types and causes. For every type, there are many different stages. The main goal of the analysis of acute lymphoblastic leukemia data is to split the sample into categories and subcategories and to classify the data into homogeneous clusters. To achieve this goal, cluster analysis is usually used to:

1. Classify homogenous cases into the same clusters and heterogeneous ones in different clusters.
2. Reduce the sample cases to a few different clusters with similar properties.
3. Determine the numbers of clusters:

Allocating homogenous objects into the same cluster means that all patients with the same type of disease and at the same stage will be classified into the same group. The benefit of this is that, same clusters of patients should be given similar protocols of medicine. Moreover, classifying the genes which causes the disease makes it easy to isolate this gene in new generations to avoid the acute lymphoblastic leukemia disease. Afterwards, this disease can be avoided by using

DNA technologies that can prevent this disease for people who have the genes which cause the acute lymphoblastic leukemia. The dependent variable here to be clustered is usually a classification variable indicating the type and stage of the disease. Thus, this paper aims to is to estimate the number of clusters in acute lymphoblastic leukemia genetics data.

3. Statistical Models in Differential Gene Expression:

Several model based techniques have been used in the analysis of acute lymphoblastic leukemia data and to analyze microarray data. The approach is based on multivariate exploratory data analysis, aiming to achieve a number of techniques that allows for quick viewing of distinct gene expression patterns within a data set. Principal Component Analysis (PCA) has been used in the analysis of multivariate data by expressing the maximum variance as a minimum number of principal components, redundant components are eliminated, thus reducing the dimensions of the input vectors (De Bin and Risso, 2011). Singular Value Decomposition (SVD) treats microarray data as a matrix, A, which is composed of n rows (genes) by p columns (experiments). SVD is represented by the mathematical equation, with U being the gene coefficient vectors, S is the mode amplitude and V^T the expression level victors, where:

$$A_{n\times p} = U_{n\times n} S_{n\times p} V^T_{p\times p}$$

One of the most familiar statistical techniques to biologists is hierarchical clustering that presents data as gene list within a dendogram to perform a bottom-up analysis. This can be obtained by assigning a similarity score to all gene pairs by calculating the Pearson's correlation coefficient, and building a tree of genes. K-means clustering however, is a top down technique that groups a collection of nodes into a fixed number of cluster (k) that are subjected to an iterative process. Each class must have a center point that is the average position of all the distances in that class and each sample must fall into the class to which its center is closest. The Nearest-Neighbor(NN) methods are based on a distance function for pairs of tumor messenger Ribo Nucleic Acid (mRNA) samples, such as the Euclidean distance or one minus the correlation of their gene expression profiles. By implementing the NN for each tumor sample in the test set we can:

(a) find the k closest tumor samples in the learning set, and
(b) predict the class by the majority vote; that is chooses the class that is most common among those k neighbors.

The number of neighbor's k is chosen by cross-validation; that is, by running the NN classifier on the learning set only.

Class prediction is based on supervised data analysis methods that impose known groups datasets. First, a training set is identified, this is, a group of genes which has a known pattern of expression is used to "train" a dataset, by comparing the data to the training set and thus classifying it. This particular method is very useful in the sub classification of similar samples, cancer diagnosis, or to predict cell or patient response to drug therapy. In some cases, this type of analysis has also been used to predict patient outcome, allowing for a clinically relevant use of microarray data. The Fisher Linear Discriminant Analysis assumes that a random vector X has a multivariate normal distribution for each defined group, and the covariance within each group is identical for all the groups. This makes the optimal decision function for the comparison of data a linear transformation of x. Variations on this theme include quadratic discriminant analysis, flexible discriminant analysis and penalized discriminant analysis.

Other methods of analysis include Support Vector Machines and based on constructing planes in a multidimensional space that separate the different classes of genes, and set decision boundaries using an iterative training algorithm. Data is mapped into the higher dimensional space from its original input space, and a nonlinear decision boundary is assigned. This plane is known as the maximum margin hyper plane, and can be located by the use of a kernel function (a nonparametric weighting function). Moreover, Artificial Neural networks, or perceptions is, another machine-learning technique. Multilayer perceptrons can be used to classify samples based on their gene expression. Gene expression data for a sample are input into the model, and response is generated in the next layer, ultimately triggering a response in the output layer. This output preceptor is expected to represent the class to which the sample belongs. The method of Decision Trees is another tool that can be built by using criteria to divide samples into nodes. Samples are divided recursively until they either fall into partitions, or until a termination condition is

met. Ultimately, the intermediate nodes represent splitting points or partitioning criteria, and the leaf nodes represent those decisions.

4. The data

The data set that will be used in the analysis in the present paper is the ALL data which has been obtained from (Chiaretti, et.al. 2004) and can be also obtained from Bioconductor (2004). It consists of sample of microarrays from 128 different individuals with Acute Lymphoblastic Leukemia (ALL). A number of additional covariates are available. The data have been normalized (using qqnorm) and it is the jointly normalized data that are available for us. The data are given in the form of an exprSet object. The different covariates include the date of diagnosis; the sex of the patient, coded as M and F; the age of the patient in years; the type and stage of the disease; and a vector "CR" with the following values: 1: "CR", remission achieved; 2: "DEATH IN CR", patient died while in remission; 3: "DEATH IN INDUCTION", patient died while in induction therapy; 4: "REF", patient was refractory to therapy; the date on which remission was achieved. Other covariates include an assigned molecular biology of the cancer (mainly for those with B-cell ALL), BCR\/ABL, ALL\/AF4, E2APBX etc.; the patients response to multidrug resistance, either NEG, or POS. a vector indicating whether the patient had continuous complete remission or not.; a vector indicating whether the patient had relapse or not and many other follow up and biological data.

The data consists of 83 Males ,42 Females and 3 are NA's. The clustering variable is type and stage of the disease; B indicates B-cell while a T indicates T-cell. Both types B and T have 5 stages each. In each of these stages there are: 4 observations of B, 9 observations of B1, 35 observations of B2, 22 observations of B3, and 9 observations of B4. Moreover T-cell includes 5 observations of T, 1 observation of T1, 5 observations of T2, 9 observations of T3 and 2 observations of T4. The data set (ALL) was separated into two subsets of patients because it consists of 94 patients who have B-cells and 32 patients who have T-cells (Chiaretti; 2004). The goal here is to split the sample into categories and subcategories and to classify the data into homogeneous clusters.

5. Estimating the number of clusters

When there is more than one cluster of patients, a plot of the absolute eigenvalues of the data matrix is characterized by the intersection of a two parts curve the first with high negative slope and the second is a flat curve. The curve intersects with the x-axis at the value of x=k in the case of a similarity matrix. Plot-based inference can be formalized by splitting the values of the covariate (rank) at different points and finding the reflection point corresponding to the best fit for response variable based on minimum deviance (Dudoit *et al*; 2002). One expects the slope to change dramatically at the reflection point. The number of large eigenvalues is the index at which the slope changes minus 1. Since the projection operation forces eigenvalues to be exactly zero, artificial eigenvalue can always be deleted before interpreting the plots or applying the slope change method (Dudoit *et al*; 2002).

The null hypothesis that k=1 can also be tested by comparing the deviance of the simple linear regression with the minimum deviance of the broken line regression. The null hypothesis is rejected if the difference between the two deviances was greater than the expected chi-squared value with one degree of freedom at the specified significance level. This procedure is an ad-hoc because the non-null deviance was minimized over all possible change points. Experience has shown that while positive square roots of the eigenvalues are superior for visual inspection, the slope change method works best using the absolute value of the row eigenvalues.

The methods that have been applied to the underlying data sets depend upon the "Bioconductor" which is an open software development for computational biology and bioinformatics R to automatically estimate the number of clusters for large B_cells sample with 79 cases and small T_cells with 32 cases. The automatic estimation of the number of clusters saves time and efforts particularly for non-experienced users. Two libraries were applied which are the *Mclust* on the B_cells sample and *hclust* on the T_cells sample.

5.1. Estimating the Number of Clusters Using Mclust Algorithm

Mclust algorithm has been developed by Fraley and Raftery (2007-a) and assumes a normal or Gaussian mixture model:

$$\prod_{i=1}^{n}\sum_{k=1}^{G}\tau_k \phi_k (x_i \left| \mu_k , \Sigma_k \right.),$$

where x represents the data, G is the number of components, π_k is the probability that an observation belongs to the k^{th} component

$(\tau_k \geq 0; \sum_{k=1}^{G} \tau_k = 1)$, and

$$\phi_k (x \left| \mu_k , \Sigma_k \right.) = (2\tau)^{-\frac{p}{2}} \left|\Sigma_k\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(x_i - \mu_k\right)^T \Sigma_k^{-1}\left(x_i - \mu_k\right)\right\}$$

.

The exception is for model-based hierarchical clustering, for which the model used is the classification likelihood with a parameterized normal distribution assumed for each class:

$$\prod_{i=1}^{n} \phi_{\ell_i} (x_i \left| \mu_{\ell_i} , \Sigma_{\ell_i} \right.),$$

where the ℓ_i are labels indicating a unique classification of each observation: $\ell_i = k$ if x_i
belongs to the k^{th} component.

The components or clusters in both these models are ellipsoidal, centered at the means μ_k. The covariances $\sum_k$ determine their other geometric features. Each covariance matrix is parameterized by eigenvalue decomposition in the form

$$\Sigma_k = \lambda_k D_k A_k D_k^T$$

where D_k is the orthogonal matrix of eigenvectors, A_k is a diagonal matrix whose elements are proportional to the eigenvalues of $\sum_k$, and λ_k is a scalar (Banfield and Raftery 1993). The orientation of the principal components of $\sum_k$ is determined by D_k, while A_k determines the shape of the density contours; λ_k specifies the volume of the corresponding ellipsoid, which is proportional to $\lambda_k^d |A_k|$, where d is the data dimension. Characteristics (orientation, volume and shape) of distributions are usually estimated from the data, and can be allowed to vary between clusters, or constrained to be the same for all clusters. This parameterization includes but is not restricted to well-known variance models that are associated with various criterion for hierarchical clustering, such as equal-volume spherical variance ($\sum_k$ =

λI) for the sum of squares criterion, constant variance, and unconstrained variance (Fraley and Raftery, 2006).

Several measures have been proposed for choosing the clustering model (parameterization and number of clusters). We use the Bayesian Information Criterion (BIC) approximation to the Bayes factor, which adds a penalty to the loglikelihood based on the number of parameters, and has performed well in a number of applications (Fraley and Raftery, 2007-b). The Bayesian Information Criteria (BIC) has the following features:

1. It is independent of the prior.
2. It can measure the efficiency of the parameterized model in terms of predicting the data.
3. It penalizes the complexity of the model where complexity refers to the number of parameters in model.
4. It can be used to choose the number of clusters which makes the model reach to maximize BIC.
5. The model with lower value of BIC is the one to be preferred.

The BIC has the form

$$- 2\ln p(x \mid k) \approx BIC - 2\ln L + k \ln(n)$$

Where:
x : the observed data.
N : the number of observations.
K : the number of free parameters to be estimated.
P(x|k) : the likelihood of the observed data given the number of parameters.
L : the maximized value of the likelihood function for estimated model.

A large BIC score indicates strong evidence for the corresponding model. BIC can be used to choose the number of clusters and the covariance parameterizations (Mclust).

Using Mclust algorithm we can select the fitted model, each combination of a different specification of the covariance matrices and a different number of clusters corresponds to a separate probability model. Then the optimal model according to BIC for EM initialized were chosen by hierarchical clustering for parameterized Gaussian mixture models

5.2. Estimating the Number of Clusters using hclust Algorithm

The *hclust* algorithm allows clustering genes by their expression profile similarity. The purpose of the analysis is to select groups of genes that have common patterns of expression in different experiments, e.g. high expression in cancer tissues and low expression in normal tissues. These patterns of co-expression are usually treated as co-regulation. The similarity of the expressions patterns may not be limited by simple rules and can be described by similarity (or distance) Measures. There are several measures of expression profile similarity between two genes:

1. *Euclidean distance*. This is the geometric distance in the multidimensional space.
2. *Squared Euclidean distance*. The squared Euclidean distance can be implemented in order to place progressively greater weight on objects that are further apart.
3. *Manhattan distance*. This distance is the average absolute difference for the set of experiments.
4. *Chebychev distance*. This distance is computed as $d_{ij} = \max_k | x_{ik} - x_{ik} |$. The measure is useful when one wants to define two objects as "different" if they are different on any one of the experiments. In SelTag all distance measures (1-3) are normalized to the number of fields involved in calculation. This is useful when take into account expression data with missing values.
5. $1-r_{ij}$; This measure keep close profiles with positive correlation coefficients and is useful when one wants to detect co-regulated genes.
6. $1-|r_{ij}|$; This measure keep close profiles with higher absolute value of correlation coefficients.
7. $1+r_{ij}$; This measure keep close profiles with negative value of correlation coefficients (anti-correlated).

The *hclust* algorithm describes the dendogram produced by the clustering process. The function performs a hierarchical cluster analysis using a set of dissimilarities for the n objects being clustered. Each object is assigned to a cluster.

A number of different clustering methods provide Ward's minimum variance method. There are numerous ways in which clusters can be

formed, Hierarchical clustering is one of the most straightforward methods, it can be either agglomerative or divisive. Agglomerative hierarchical clustering begins with every case being a cluster into itself, at successive steps, similar clusters are merged. Divisive clustering starts with all cases in one cluster and end up with each case in an individual clusters. In agglomerative clustering, once a cluster is formed, it cannot be split; it can only be combined with other clusters. Agglomerative hierarchical clustering does not let cases to be separated from clusters that they have joined. Once in a cluster, always in that cluster, we can choose the number of clusters when we reach to maximize height at hierarchical cluster diagram both agglomerative and Divisive are used to estimate the number of clusters in small data. When we choose the number of clusters using the hclust algorithm described above we compare its results with the results of Partitioning Around Medoid " PAM " algorithm.

5.3. Estimating the Number of Clusters Using the *Partitioning Around Medoid (PAM)* Algorithm

This algorithm designed by Kaufman and Rousseuw (1990) as a partitioning method which operates on the dissimilarity matrix, e.g. Euclidean distance matrix. PAM is more robust than k-means in the presence of noise and outliers because a medoid is less influenced by outliers or other extreme values than a mean. It works well for small data sets but does not scale well for large data sets.

For a prespecified number of clusters K, the PAM procedure is based on the search for K representative objects, or medoids, among the observations to be clustered. After finding a set of K medoids, K clusters are constructed by assigning each observation to the nearest medoid.

The goal is to find K medoids, $M^* = (m_1^*, ..., m_k^*)$ where M^* is the sum of the dissimilarities of the observations to their closest medoid; that is,

$M^* = \arg\min_{M^*} \sum_i \min_k d(x_i, m_k)$, tends to be more robust K_means.

This algorithm has the following features:

a) It accepts the dissimilarity matrix.

b) It is more robust because it minimizes a sum of dissimilarities instead of a sum of squared Euclidean distance.
c) It provides a novel graphical display.
d) It allows selecting fitting the number of clusters by selecting the clusters which maximize the average silhouette width.

PAM algorithm provides a graphical display (*Silhouette plots*). Among the graphs the PAM provides a graphical display (Silhouette plots) which can be used to:

1. Select the number of clusters and
2. Asses how well individual observations are clustered.

The silhouette width of the observation i is defined as :

$sil_i = \frac{(b_i - a_i)}{\max(a_i, b_i)}$, where a_i denotes the average dissimilarity between i and all other observations in the cluster to which i belongs, and b_i denotes the minimum average dissimilarity of i to objects in other clusters.

Intuitively, objects with large silhouette width sil_i are well clustered; then those with small sil_i which tend to lie between clusters.

The *divisive coefficient* represents the strength of the clustering structure founded by the PAM algorithm.

Let dd(i) be the diameter of the cluster to which data belongs before being split to a single variable, divided by the diameter of the whole data set .

The divisive coefficient (DC) for a cluster is given by:

$DC = 2 \times \left(1 - \frac{\sum_{1}^{n} dd(i)}{n} \right)$ Where n is the number of objects, dd(i) is the diameter of cluster. See McQuarrie and Tsai (1998).

6. The Analysis of Acute Lymphoblastic Leukemia (ALL) Genetics Data

The data we used for analysis and illustration in this paper is the ALL data set which has been described above and consists of 128 microarrays from different individuals with acute lymphoblastic leukemia disease. The grouping variable is BT: The type and stage of the disease; B indicates B-cell while a T indicates T-cell.

6.1. Estimating the Number of Clusters Using *Mclust* Algorithm

When estimating the number of clusters in the B_cells data by Mclust algorithm, the result was that the data is divided into two components with different variances but the variance within each component "cluster" is equal. Therefore, we concluded that there are two homogenous clusters with all symmetric observations within the same cluster (See Szekely and Rizzo, 2005). Figure 1 is produced by Mclust algorithm and illustrates the above result for the B-cell. In Figure 1 below, two models can be seen easily. The upper one marked by solid triangles and the other one is marked by empty triangles. Each triangle represents the number of clusters so that each model has 9 different numbers of clusters. Moreover, as described in the characteristics of Bayesian Information Criteria that the model with the lowest absolute value of BIC is preferred which is here upper one which is marked with solid triangle, also from the characteristics of BIC both two models reach their maximum BIC when the number of clusters is two. Therefore, observation of Figure 1 supports the results of the Mclust algorithm output.

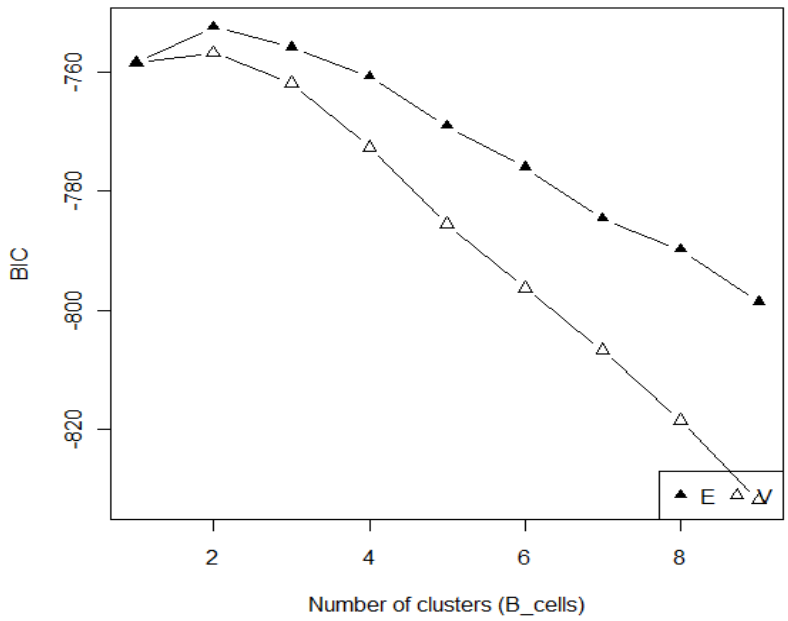

Figure 1: The BIC for different number of clusters in "ALL" data set

6.2. Estimating the Number of Clusters using *hclust* Algorithm

The second data set (T_cells) consists of 32 observations so that the suitable algorithm for estimating the number of clusters is hclust as described in Section 4. The hclust algorithm uses the agglomerative hierarchal clustering which begins with every case being a cluster. Similar clusters are merged and we can choose the number of clusters that maximizes the height at hierarchal cluster dendogram.

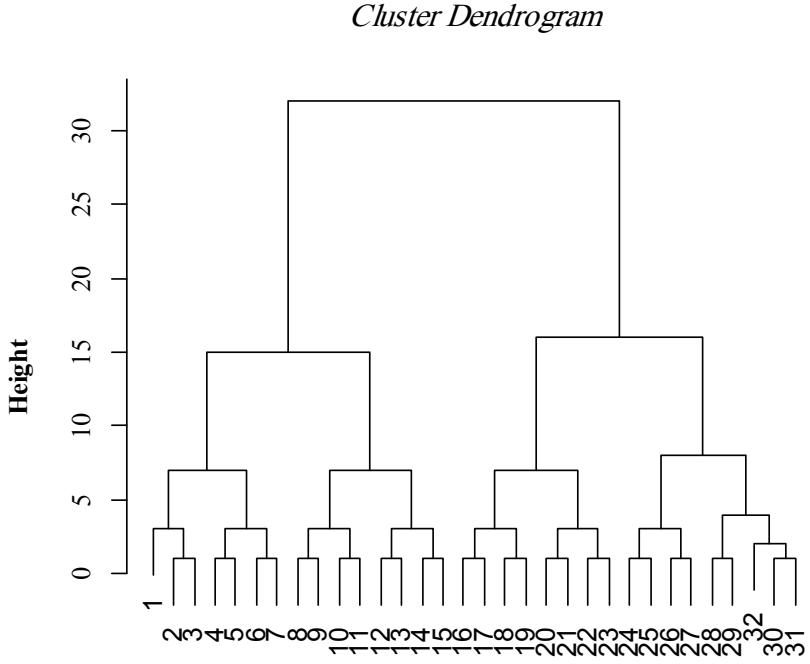

Figure 2: Cluster dendogram for the second data set (T_cells)
hclust (*, "complete")

Looking at figure 2 above from bottom, we can easily see that each object cluster with itself and thus we have 32 clusters. If we moved steps from down to top we can observe that the object 1 is in one cluster, objects 2 and 3 in another cluster. These two clusters are agglomerate in another cluster with height =3. Objects 4 and 5 in one cluster and 6 and 7 in another cluster with height = 3. The four clusters are agglomerate in a cluster with height = 7. Objects 8 and 9 are agglomerate in a cluster. Objects 10 and 11 also agglomerate in a cluster and both clusters are agglomerate with a cluster with height=3. Objects 12 and 13 agglomerate with a cluster and 14 and15 agglomerate with a cluster with height = 7. Both two clusters with height = 7 are agglomerate with another cluster with height = 15. Objects 16 and17 agglomerate with a cluster its height= 3. Also

objects 18 and 19 are agglomerate with a cluster its height =3 and both the two clusters with height =3 are clustered in one cluster with height=7. Objects 20 and 21 agglomerate with a cluster its height= 3. Also objects 22 and 23 are agglomerate with a cluster its height=3 and both the two clusters with height =3 are clustered in one cluster with height=7. Objects 24 and 25 agglomerate with a cluster its height= 3. Also objects 26 and 27 are agglomerate with a cluster its height=3 and both the two clusters with height =3 are clustered in one cluster with height=7. Objects 28 and 29 agglomerate with a cluster its height= 3. Also objects 30 and 31 are agglomerate with a cluster its height=3 and both the two clusters with height =3 are clustered in one cluster with height=7 and the object 32 is clustered with it self. The five clusters are agglomerate with a cluster with height= 4. The clusters from object 24 to 32 are agglomerate with height=8. The objects from 16 to 23 and 24 to 31 are agglomerate with a cluster with height equals 16, here we have two clusters with maximum height one is 15 and other is 16. We conclude that our data composed from 2 clusters.

6.3. Estimating the Number of Clusters Using the *Partitioning Around Medoid* (PAM) Algorithm

The goal of cluster analysis for our data set is to reach to the maximum dissimilarity between observations of different clusters and the mediod and wider diameter cluster and maximum average silhouette width.

Using the PAM algorithm we achieved the following results

- When number of clusters is 2 then the average silhouette width = 0.6 and the total maximum dissimilarity is 16 and the total of cluster diameter is 31.
- When number of clusters is 3 then the average silhouette width = 0.55 and the total maximum dissimilarity is 16 and the total of cluster diameter is 30.
- When number of clusters is 4 then the average silhouette width = 0.51 and the total maximum dissimilarity is 16 and the total of cluster diameter is 29.
- When number of clusters is 5 then the average silhouette width = 0.48 and the total maximum dissimilarity is 16 and the total of cluster diameter is 28.

From the above results we can conclude that the best number of clusters is 2. In this case each of average silhouette width, dissimilarity between the observations, the cluster medoid, and the cluster diameter reach its maximum value. We notice that after cluster 2 we gets the same results in maximum dissimilarity because the sample size is 32 and it is difficult to cluster such sample size in more than 3clusters.

7. Conclusions:

In this paper we analyzed two data sets. The first one is the B_cells which can be considered as a large sample and the Mclust algorithm was used to estimate the number of clusters. Using this algorithm we illustrate the results that the number of clusters is two. We also described the fit of Mclust algorithm that uses the Bayesian Information Criteria (BIC). Moreover we illustrate the result that the number of clusters is two where the maximum BIC has been achieved. To confirm these results, we compared the results of both Mclust and hclust algorithms with PAM algorithm where we selected different numbers of clusters from 1 to 5 because we have 5 stages of disease in our data set and we compute the Average Silhouette Width for each choice of number of clusters. We conclude that the number of clusters also equals two since it corresponds to the maximum value of Average Silhouette Width. Looking at the number of clusters at B_cells we can conclude that there is only 2 clusters, which means that we reduce the 5 stages to 2 symmetric clusters and that means that each cluster should have the same medication or treatments.

The second data sets is T_cells which is a small sample. Therefore we used the hclust algorithm to estimate the number of clusters and we concluded that the number of clusters is two. To confirm these results we compared the results of hclust algorithm with the PAM algorithm where we selected different numbers of clusters from 1 to 5 as in the previous data set and we computed the Average Silhouette Width for each choice of number of clusters. We then conclude that the number of clusters equals two since it corresponds to the maximum value of Average Silhouette Width. Looking at the number of clusters at T_cells we can conclude that there is only 2 clusters, which means that we reduce the 5 stages to 2 symmetric clusters and that means that each cluster should have the same medication or treatments.

From the above discussions of the results and the comparison between Mclust, hclust and PAM algorithms we conclude that the Mclust algorithm is suitable for large data sets and the hclust algorithm is suitable for small samples. Therefore we recommend using Mclust algorithm for large samples and hclust algorithm for small samples.

8. Recommendations:

1. From the above results we recommend to concentrate future research on methods of detecting genes that causes different types of cancer to avoid this disease by isolating these genes in the new generations.
2. The data used in this study was obtained from "bioconductor.org" website. We recommend that a genetic data bank to be established in Palestine. This would help in isolation genes which causes heredity diseases in Palestine.
3. For future studies we propose joint researches between the Faculties of Medicine, Medical Sciences and the Department of Statistics at Al Azhar University in genetics fields.
4. We recommend conducting further research on using Neural Networks techniques for estimating the number of clusters in genetics data.
5. We also recommend conducting further research on testing whether there is a significant evidence of different types of cancer between genetics causes and other causes in Palestine.
6. It is also recommended that the clustering algorithms discussed in this paper to be applied in other fields such as Economics and Human Sciences.

References:

1. Banfield J. D. and Raftery A. E. (1993); "Model-based Gaussian and non-Gaussian clustering". Biometrics, 49:803–821.
2. Bioconductor (2004): Open software development for computational biology and bioinformatics; Gentleman R., Carey V. J., Bates D. M., Bolstad B., Dettling M., Dudoit S., Ellis B., Gautier L., Ge Y., and others, Genome Biology, Vol. 5, R80.

3. Chen, G., Banerjee, N., Jaradat, S.A., Tanaka, T.S., Ko, M.S.H. and Zhang, M.Q. (2002), "Evaluation and Comparison of Clustering Algorithms in Analyzing ES Cell Gene Expression Data", Statistica Sinica, 12: 241-262
4. Chiaretti S, Li X, Gentleman R, Vitale A, Vignetti M, Mandelli F, Ritz J, Foa R (2004); "Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival Blood", Vol. 103, No 7.
5. De Bin R. and Risso D. (2011), "A novel approach to the clustering of microarray data via nonparametric density estimation", *BMC* Bioinformatics, 12:49.
6. Dudoit, S; Fridlyand, J and Speed, T. P. (2002), "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data"; Journal of American Statistical Association; Vol. 97, No 457,70-85.
7. Fraley C and Raftery AE (2006); "Model-based microarray image analysis"; R News, 6:60–63.
8. Fraley C and Raftery AE (2007-a); "Model-based methods of classification: using the mclust software in chemometrics"; Journal of Statistical Software, 18(6).
9. Fraley C and Raftery AE (2007-b); "Bayesian regularization for normal mixture estimation and model-based clustering"; Journal of Classification, 24:155–181.
10. Kaufman L and Rousseeuw PJ (1990), "Finding Groups in Data: An Introduction to Cluster Analysis", Wiley-Interscience, New York (Series in Applied Probability and Statistics).
11. McQuarrie ADR and Tsai CL (1998), Regression and Time Series Model Selection, World Scientific.
12. Szekely, G. J. and Rizzo, M. L. (2005) Hierarchical Clustering via Joint Between-Within Distances: Extending Ward's Minimum Variance Method, Journal of Classification 22(2) 151-183.