International Journal of Statistics and Applications 2018, 8(3): 144-152
DOI: 10.5923/j.statistics.20180803.05

An Empirical Comparison of Principal Component Analysis and Clustering on Variables for Dimension Reduction Using Leukemia and Breast Cancer Data

Khaled I. A. Almaghri[1,*], S. Chakraborty[2]

[1]Pharmacology College, Palestine University, Gaza, Palestine
[2]Department of Statistics, Dibrugarh University, Assam, India

Abstract One of the important problems of data analysis is that identifying nuisance variable(s) in a data set that contributes to an increase of variability within groups in an experiment. One way to address this issue is through dimension reduction of data sets. In this study we compare between two widely used methods of reducing dimension data sets, namely the method of the principal component (PC), statistics technical that uses orthogonal transformation to convert a set of possibly correlated variables of into a new set of uncorrelated variables and the method of clustering on variables, where the aim is to put the variables with similar information in the same group or cluster by considering two celebrated data sets from literature, the leukemia dataset and the other a breast cancer data.

Keywords Acute lymphoblastic leukemia "ALL", Breast cancer, Clustering on variables, Dimension reduction, Scree plot, Correlation matrix, Cumulative variance proportion, Principal component analysis

1. Introduction

In biostatistics data set with higher dimension is often difficult to handle and may cause waste of time. Principal component analysis (Sanche and Lonergan, 2006 [14], (pp.439, Izenman, 2008 [7])) is a useful statistical techniques to reduce dimension but at the cost of loosing of information. Clustering on variables is used to construct clusters of homogenous variables so that we can choose one representative variable from each cluster for using in our model or statistical analysis without losing any information. The later method also allow us to avoid those variables in a clusters which need more time to for observation and hence more time in data collection.

2. Material and Methods

2.1. Acute Lymphoblastic Leukemia (ALL) Data Set

Acute Lymphoblastic Leukemia data set taken from Ritz Laboratory (Everitt et al., 2004 [3]) consists of micro arrays from 128 different individuals with acute lymphoblastic leukemia (ALL). The data available in R database have already been normalized using Robust Multichip Average (rma) (R manual documentation, 2012 [12], Irizarry et al., 2003 [6]).

This data frame contains observations on: (i) Patient IDs, (ii) Date of diagnosis, (iii) Sex of the patient (sex), (iv) Age of the patient in years (age), (v) Type and stage of the disease: 'B' indicates B-cell ALL while 'T' indicates T-cell ALL (BT), (vi) ‘Remission’: a factor with two levels, either 'CR' indicates that remission was achieved or 'REF' indicating that the patient was refractory, and remission was not achieved (remission), (vii) ‘CR’: a vector 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 (CR), (viii) the date on which remission was achieved, (ix) a logical vector indicating whether t (4; 11) translocation was detected (t411), (x) a logical vector indicating whether t (9; 22) translocation was detected (t922), (xi) a vector indicating the various cytogenetic abnormalities that were detected (cyton), (xii) the assigned molecular biology of the cancer (molb), (xiii) Fusion protein for those with BCR\/ABL which of the fusion proteins was detected, ‘p190’, 'p190\/p210', 'p210' (fusionp), (xiv) the patient’s response to multidrug resistance, either 'NEG', or 'POS' (mdr), (xv) 'kinet' ploidy, either diploid or hyperd (kinet), (xvi) a vector indicating whether the patient had neither continuous complete remission nor not (ccr), (xvii) a vector indicating whether the patient had relapse or

* Corresponding author:
khaledalmghari@gmail.com (Khaled I. A. Almaghri)
Published online at http://journal.sapub.org/statistics

not (relapse), (xviii) a vector indicating whether the patient receive a bone marrow transplant or not (transplant), and (xix) follow-up data with 10 possible value 1 to 10 (f.u). The possible values of fu are:

1. "AUBMT \/ REL": autologous bone marrow transplant and subsequent relapse,
2. "BMT \/ CCR": allogeneic bone marrow transplant and still in continuous complete remission,
3. "BMT \/ DEATH IN CR": after allogeneic bone marrow transplant patient died without relapsing,
4. "BMT \/ REL": after allogeneic bone marrow transplant patient relapsed,
5. "CCR": patient was in continuous complete remission,
6. "CCR \/ OFF": patient was in continuous complete remission but off-protocol for some reasons,
7. "DEATH IN CR": died when in complete remission,
8. "MUD \/ DEATH IN CR": unrelated allogeneic bone marrow transplant and death without relapsing,
9. "REL": relapse, and
10. "REL \/ SNC": relapse occurred at central nervous system,

The last variable is (xx) a logical vector indicating whether the cytogenesis was normal (citog).

The data have been presented in the form of an 'exprSet' object which is suitable for implementation and comparison in many of clusters algorithms (Kumar and Sharma, 2011 [10]; Jonathan et al., 2010 [8]) because one can extract subsets from this dataset as Acute Lymphoblastic Leukemia caused by different causes like T.cells, B.cells.

The variable BT gives information about the type (B or T) and stages of the disease (five stages for each type). So from the ALL data set two distinct subsets with respect to two covariates namely T cells and B cells have been extracted for independent investigation using the clustering algorithms.

The values of the all the variables in the 95^{th} and the 128^{th} rows of the data set are missing. As such effectively, the ALL dataset comprises observations of 126 individuals, more over in the present work four variables namely the variables Patient IDs, date of diagnosis, age of the patient in years and date on which remission was achieved have been omitted before the analysis as they are not relevant for the present investigation. Therefore, in the current work, 126 observations (rows) with only 16 out of 20 variables have been considered for the analysis.

2.2. Breast Cancer Data

This dataset is from German Breast Cancer Study Group 2 (Schumacher et al., 1994 [17]; Sauerbrei and Royston, 1999 [16]). This data frame contains the observations on (i) hormonal therapy, a factor at two levels no and yes (ii) age of the patients in years (iii) menopausal status, a factor at two levels pre (premenopausal) and post (postmenopausal) (iv) tumor size (in mm), (v) tumor grade, a ordered factor at levels I < II < III, (vi) number of positive nodes, (vii) progesterone receptor (in fmol) (viii) recurrence free survival time (in days) and (ix) censoring indicator (0- censored, 1-event) of 686 women.

2.3. Dimension Reduction

(Sanche and Lonergan, 2006 [14], (pp.439, Izenman, 2008 [7]))

2.3.1. Principal Component Analysis

It is a statistical procedure to transform a set of observations of correlated variables to a set of a linearity uncorrelated variables by orthogonal transformation. We can explain the variance covariance structure of these variables by some of these linear combinations of the original variables.

2.3.2. Cluster Analysis

Cluster analysis has been used as a tool to overcome the difficulties in handling big data by partitioning the data into a number of interesting clusters and concentrating only on the interesting clusters instead of full data set (Halkidi et al., 2001 [5]).

Clustering of the variables is same as those of clustering of objects, where instead of observations, the variables are grouped into homogeneous clusters, that is variables in each cluster will be strongly related to each other and contain same information. As a result one can reduce the dimension of the data by choosing only one representative variable from each of the homogenous clusters.

Dimension reduction through clustering of variables can be implemented by using the *pvclust* and also *ClustOfVar* packages of R.

In the proposed work we used *ClustOfVar* package by using a bootstrap resampling (50 replications) and plot the stability criterion according to the number of clusters which can help to choice a sensible and suitable number of clusters. [12]

3. Previous Studies

San and Lowrence (2000 [15]), they conduct a study to introduce locally linear embedded algorithm that computes low dimensional neighbor hood preserving embeddings of high dimensional inputs and they implement it on nonlinear manifolds such as those generated by images of faces or documents of test and they found that their algorithm is more useful.

Chris and He (2004 [2]) conducted a study to test the effect of dimension reduction on *K*-means clustering by using principal component analysis to reduce the data from the original 1000 dimension to 40, 20, 10, 6 and 5 dimensions respectively on 4029 of Gene expression of 96 tissue samples on human Lymphoma. They have applied *K*-means on 10 random samples of each new groups combination [40, 20, 10, 6 and 5 dimensions], they found that the results systematically and significant were improved.

Chris and Tao (2007 [1]) conducted a study to combine linear discriminant analysis (LDA) and *K*-means clustering

and tested this new method on a wide range of datasets for adaptive dimension reduction. They found that this new clustering process subspace selection process. And the learning algorithm performs data clustering and subspace selection simultaneously. They further show that this new algorithm is more effective than other methods of dimension reduction.

Shuiwang and Jeieping (2009 [18]) they conduct a study to reduce dimension in Multi-label classification and they compare between least squares loss and hinge loss and they found that the relative performance of formulation with orthonormal transformation and orthonormal features is different for different datasets.

Gowrilaksshmi (2011 [4]) conducted a study to use methods for dimension reduction using a combination of principal component and with *K*-means and Locally Linear Embedding (LLE) combined with *K*-means and found that all the clustering methods were affected by data dimension reduction and data characteristics such as overlapping between clusters and the presence of noise.

Murillo J. and Rodriguez A. (2012 [11]) they conduct a study to study the effect of dimension reduction and data set and they found that reduce of dimension is useful for diminishing the error of probability of classifiers.

4. Results and Observations

4.1. Breast Cancer Data

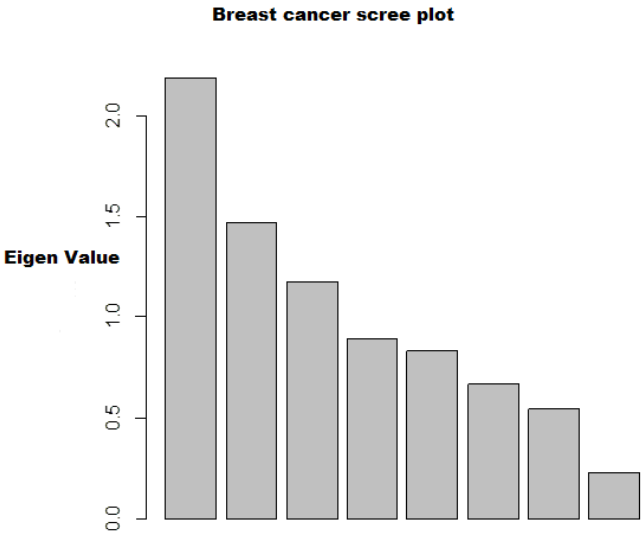

Figure 1. Scree plot for Breast cancer data

From Fig. 1 it is seen that three eigen value(s) are greater than 1 for the Breast cancer data. Hence we estimated the numbers of principal component to be 3.

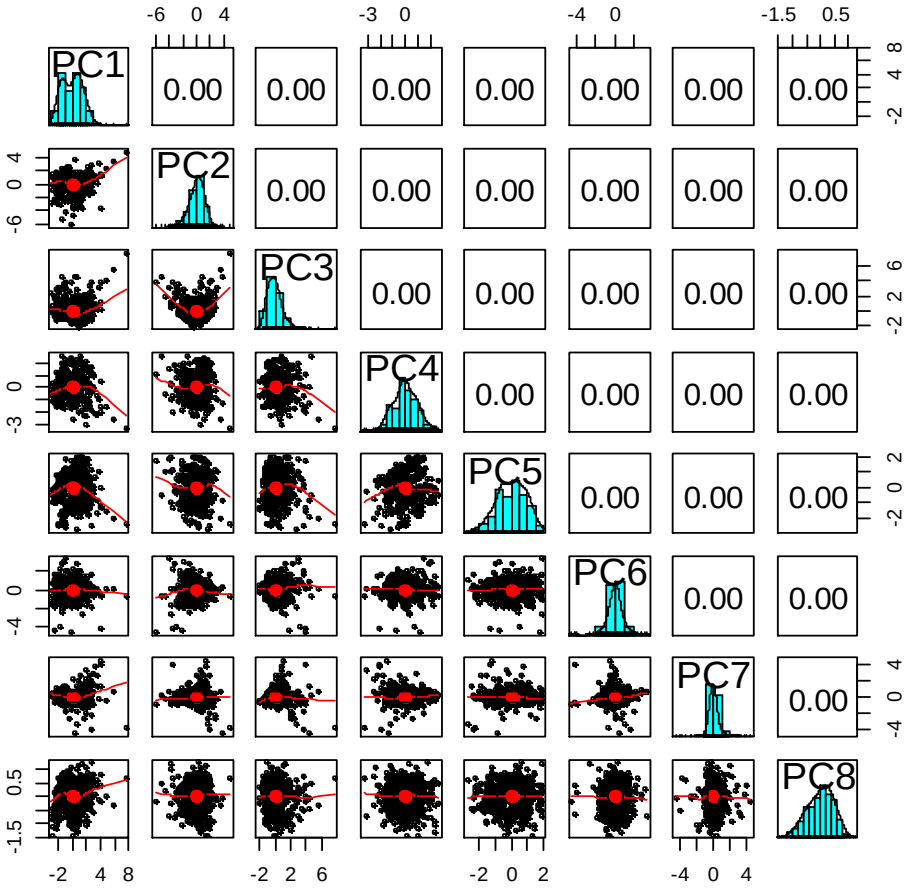

Figure 2. Correlation matrix between principal components for Breast Cancer data

From in Fig. 2 it is easily verified that for Breast cancer data there is no correlation between the derived principal components which is a basic property (see Robert et al., 2003 [13]).

Table 1. Cumulative proportion of variance etc. for Breast cancer data

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
SD	**1.479**	**1.214**	**1.087**	0.945	0.911	0.816	0.738	0.475
Proportion of Variance	0.273	0.184	0.148	0.112	0.104	0.083	0.068	0.028
Cumulative Proportion	0.273	0.458	**0.605**	0.717	0.821	0.904	0.972	1

Table 1 showed th standard deviation (SD), proportion of variance and cumulative proportion of variance for Breast cancer data. Our interest is in the components with SD above 1. Once again here we identify first three principal components that explain 60.5% of total variance.

Table 2. Rotation matrix for principal component analysis Breast cancer data

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
horTH	0.3258	-0.1181	-0.0939	0.4818	-0.0798	0.0408	0.0169	0.0199
age	**0.5706**	-0.2250	-0.1545	-0.0804	0.2535	0.0493	-0.1337	0.7134
meno	**0.5484**	-0.0265	-0.2298	-0.0660	0.2341	0.0867	-0.1537	-0.6973
tsize	-0.1102	-0.4748	0.4891	0.2138	0.1340	0.6596	0.1566	0.0045
tgrade	-0.1610	-0.3943	-0.1193	-0.7589	-0.4461	0.1379	-0.0996	0.0234
pnodes	-0.0483	**-0.5556**	0.3942	0.0978	0.0360	-0.7161	-0.0973	-0.0219
progrec	0.2314	0.3672	**0.5986**	-0.1566	-0.1347	0.0880	-0.6338	-0.0343
estrec	0.4166	0.2024	0.3839	-0.3183	-0.0893	-0.1177	0.7160	-0.0480
SD	1.479	1.214	1.086	0.945	0.911	0.816	0.738	0.475
Variance	2.187	1.474	1.180	0.893	0.830	0.666	0.545	0.225

Table 2 showed that the dimension has been reduced from eight to four variables (age, meno, pnodes, progrec)

Variable Clustering of Breast Cancer data:

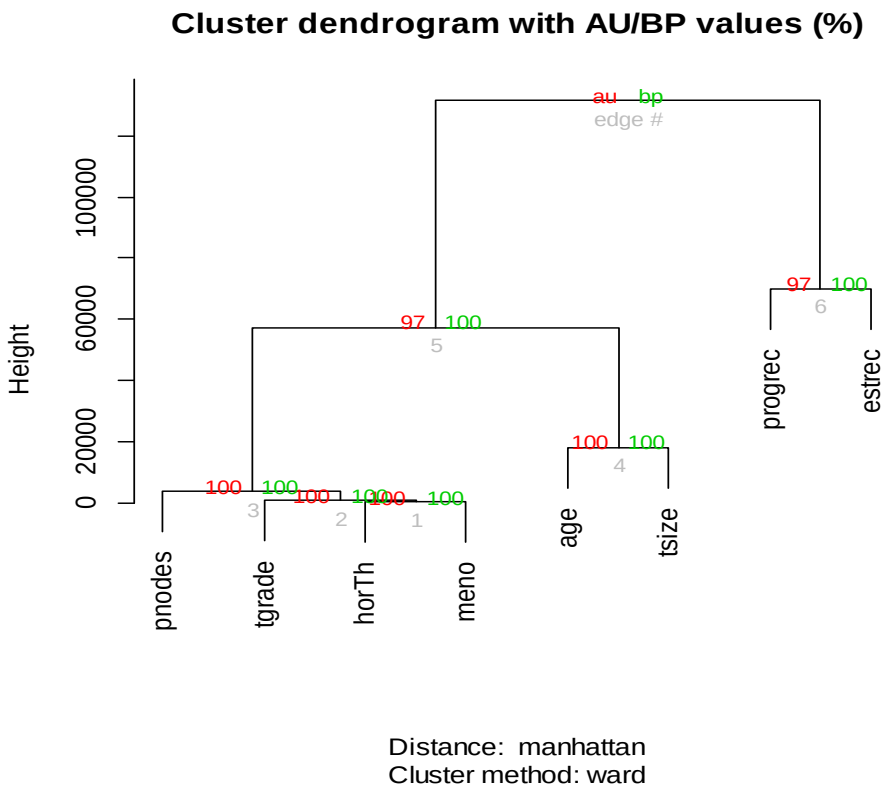

Figure 3. Clustering on variables for Breast cancer data

- **Stability plotting to insure the optimal number of clusters for Breast cancer data**

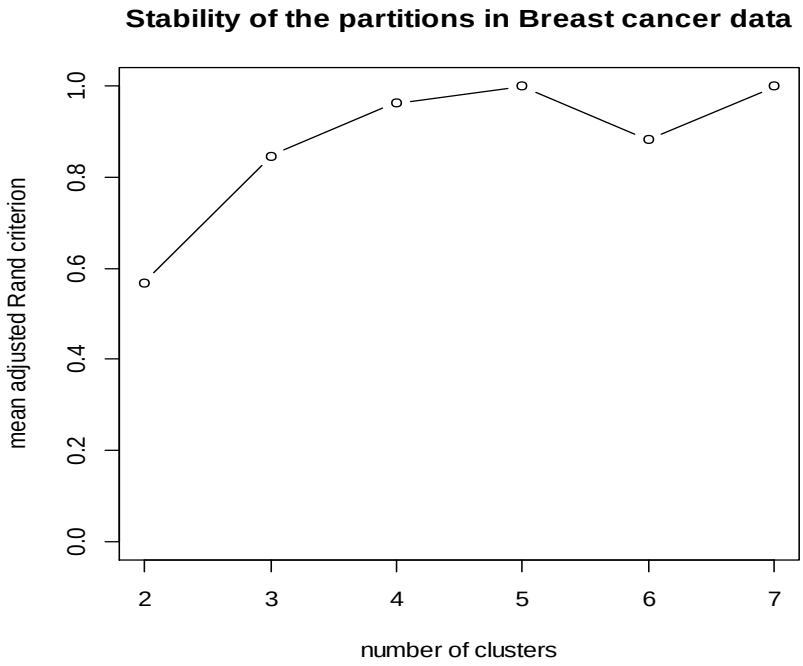

Figure 4. Stability of partition for Breast cancer data

From the Fig. 4 we conclude that there are five clusters "number of strait lines" and from Fig.3 these clusters are:
Cluster #1: {progrec, estrec}
Cluster #2: {tsize, age}
Cluster #3: {meno, horTh}
Cluster #4: {tgrade}
Cluster #5: {pnodes}

ALL data sets:

PCA for ALL:

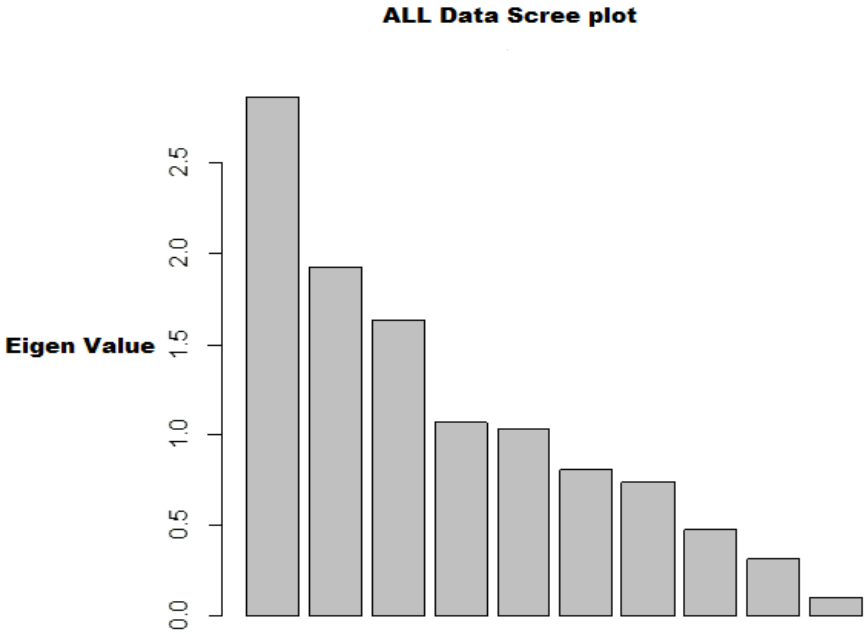

Figure 5. Scree plot for ALL data

Fig. 5 show that the estimate numbers of component, which is five components where the eigen value(s) greater than one.

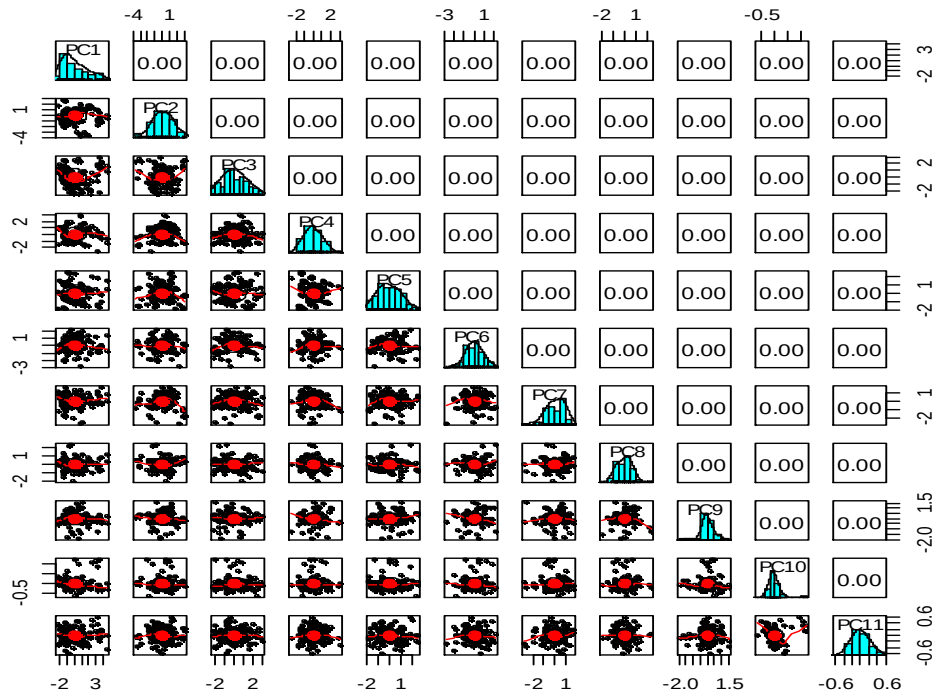

Figure 6. Correlation matrix between principal components for ALL data

Here in Fig. 5 we check the main assumption no correlation between principal components Robert et al. (2003[13]) for ALL data.

Table 3. Cumulative variance proportion etc. for ALL data

Importance of component											
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
SD	**1.692**	**1.387**	**1.278**	**1.031**	**1.016**	0.897	0.857	0.692	0.563	0.316	0.222
Proportion of Variance	0.260	0.175	0.149	0.097	0.094	0.073	0.067	0.043	0.029	0.009	0.004
Cumulative Proportion	0.260	0.435	0.584	0.680	**0.774**	0.847	0.914	0.958	0.986	0.996	1.000

In the Table 3 we present the SD, proportion of variance and cumulative proportion of variance for Breast cancer data. Our interest is in the components with SD above 1. Here we identify first five principal components that explain 77.4% of total variance present in the data set.

Table 4. Rotation matrix for Principal component analysis for ALL data

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
BT	0.0267	-0.4845	-0.3933	-0.2320	-0.0689	0.1508	-0.4862	0.0531	-0.2732	-0.2773	0.3712
Sex	0.1230	0.0384	0.3619	0.0465	**0.6499**	0.3914	-0.4540	-0.2339	0.0216	0.0224	-0.1133
remission	0.3465	-0.3532	0.4007	-0.2619	-0.0834	-0.1288	0.1478	0.0987	0.0482	-0.5609	-0.3923
CR	0.4073	-0.3271	0.3483	-0.2356	-0.0715	-0.0465	0.0945	0.0649	0.0884	0.6017	0.4002
t411	0.1172	-0.3746	-0.2722	0.3182	0.3394	0.2180	0.5895	-0.2878	0.1375	-0.1511	0.1891
t922	-0.4605	-0.1605	0.0224	-0.2714	-0.2714	0.2783	-0.0251	0.2236	0.7471	-0.0241	0.0278
Cyton	0.2435	0.3131	-0.1772	-0.4791	-0.2915	0.2694	0.0548	-0.6339	0.1318	-0.0353	-0.0267
Citog	0.2804	0.2865	-0.2522	-0.2904	0.2231	0.4188	0.2543	0.6006	-0.1806	0.0345	-0.0835
Molb	-0.4136	-0.3599	-0.0756	-0.3479	0.1329	0.0184	0.1727	-0.1693	-0.3737	0.3679	-0.4702
Fusionp	-0.3797	0.2207	0.3756	-0.3263	0.1993	-0.0773	0.2743	-0.0606	-0.2835	-0.2910	0.5196
Kinet	0.1390	0.0873	-0.3406	-0.3219	**0.5031**	-0.6527	-0.0761	-0.0267	0.2601	0.0089	0.0024
SD	1.6919	1.3872	1.2779	1.0314	1.0159	0.8973	0.8568	0.6916	0.5632	0.3164	0.2222
Variance	2.8625	1.9244	1.6331	1.0637	1.0321	0.8051	0.7341	0.4783	0.3172	0.1001	0.0494

Table 4 show that the dimension has been reduced from eleven to two variables (sex and kinet).

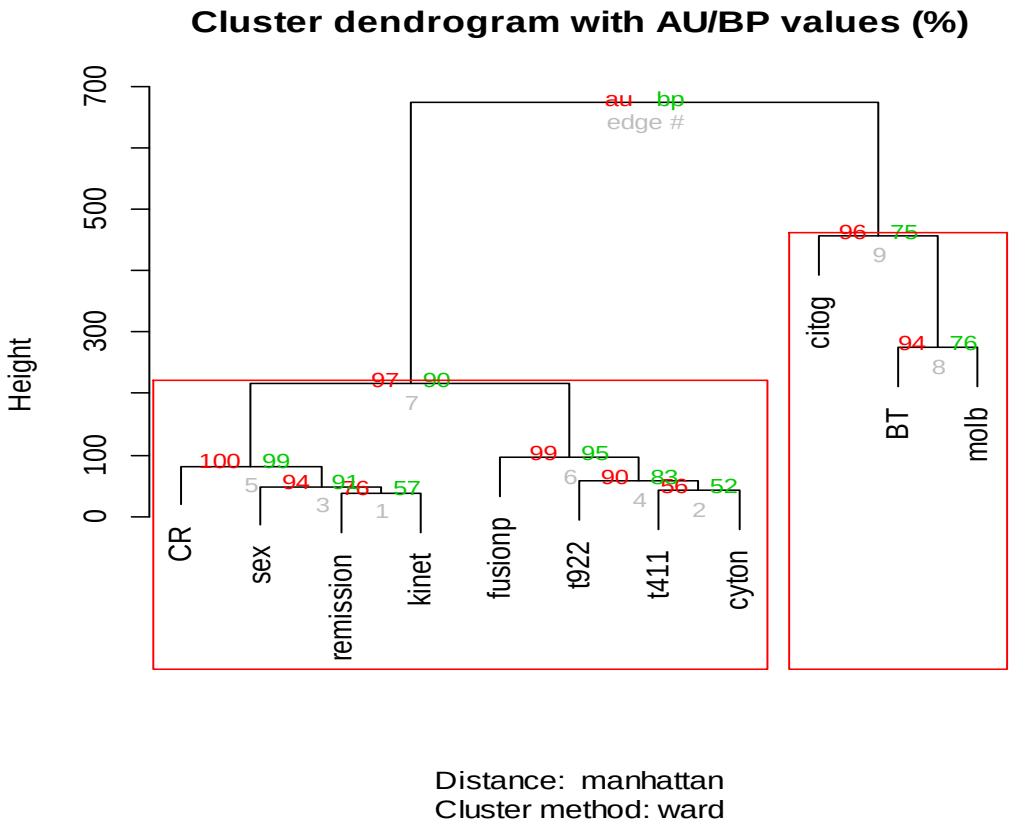

Figure 7. Clustering on variables for ALL data

• Stability plotting to insure the optimal number of clusters for ALL data

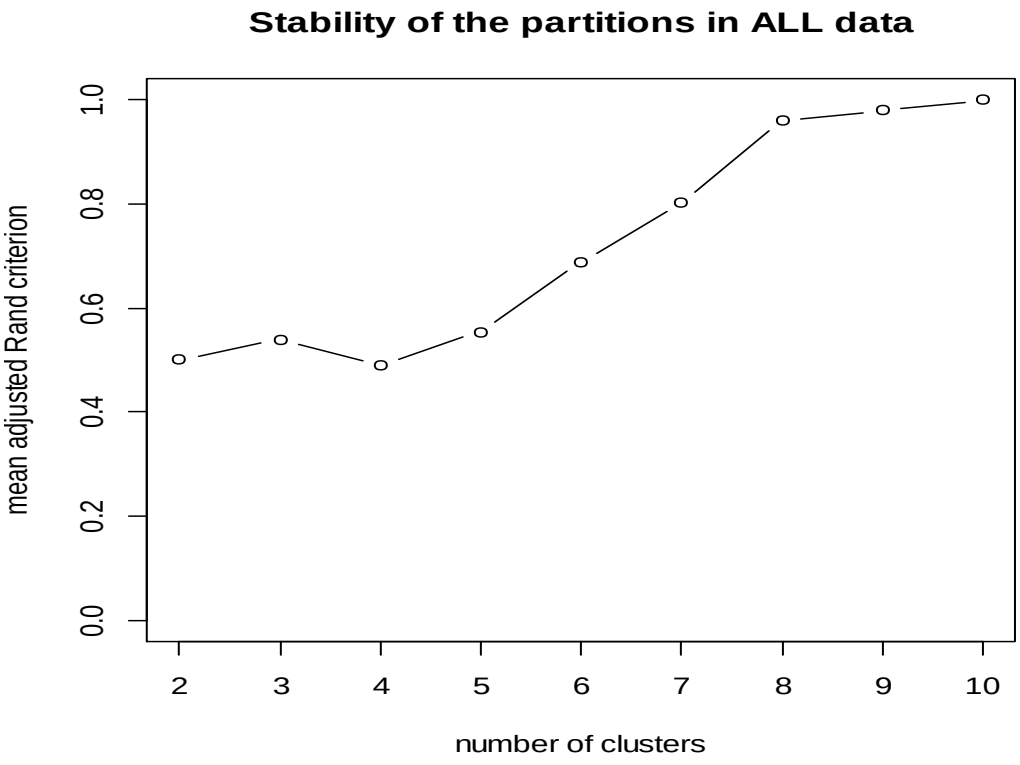

Figure 8. Stability of partition for ALL data

From the Fig. 8 we conclude that there are eight clusters "number of strait lines" and from Fig. 7 these clusters are:

Cluster #1: {kinet, remission}
Cluster #2: {cyton, t411}
Cluster #3: {sex}
Cluster #4: {t922}
Cluster #5: {CR.
Cluster #6: {fusionp}
Cluster #7: {citog}
Cluster #8: {BT, molb}

5. Concluding Remarks and Recommendations

In the proposed work, we have implemented the principal component analysis and clustering on variables on two data sets, breast cancer and ALL data sets to evaluate their performance in reducing dimension of these two data sets.

For Breast cancer our findings exhibited in Fig. 1, Fig. 2, Fig.3, Fig. 4, Tables 1 and 2. we conclude that dimension of data set is reduced from eight to three when we use principal component analysis with four important variables (age, meno, pnodes, progrec). These three PCs explain about 60% of the total data variance hence here we lost about 40% of the information regarding the variance covariance structure.

About the clustering on variables our findings show a reduction of the dimension from eight to five with out losing any of information. We can reduce the set of variable to may be {progress, age, meno, tgrade and pnodes} or {estrec, tsize, horTh, tgrade and pnodes}.

It may be noted that the principal components failed to include the important variable (tgrade) in the reduced set and the cost is loss of 40% of information. As such clustering of variable prevails over PCA in this case.

For ALL data set our findings exhibited in Fig. 5, Fig.6, Fig.7, Fig.8, Tables 3 and 4. we conclude that the dimension of data set is reduced from eleven to five with principal component analysis with two important variables (kinet and sex). These five PCs explain about 77.4% total data variance hence here we lost about 22.6% of information regarding the variance covariance structure.

However, when we used clustering on variables we reduce the dimension from eleven to eight with out losing any of information. Therefore we can reduce the set of variables to either {kinet, cyton, sex, t922, CR, fusionp, citog, BTd} or {remission, t411, sex, t922, CR, fusionp, citog, molb} with out losing any information as compared to the principal component analysis which fails to reduce dimension with good level of interpretation of data.

In medication of patients who have cancer, time is very important so if two variables falls in one cluster one needs two days to collect data other needs ten days therefore we select variable, which needs two days, and starting classify our patient to start medication.

Therefore the main recommendations from the current investigation is to use clustering on variables for reduce dimension because it is effective and there is no loss of information and gives options to select important variables from the same cluster because they are homogenous.

REFERENCES

[1] Chris, D. and Tao, Li. (2007). Adaptive Dimension Reduction using Discriminant Analysis and k-means Clustering, *24th International conference on machine learning*, Corvallis USA, 521-528.

[2] Chris, D. and Xiaofeng, He. (2004). K-means Clustering via Principal Component Analysis, 21*st International conference on machine learning*, Canada, 29-36.

[3] Everitt, S. B., Stahl, D., Leese, M. and Landau, S. (2011). *Cluster Analysis, 5th Ed.*, Willey series in Probability and Statistics, UK.

[4] Gowrilaksshmi, K. (2011). Clustering on High Dimensional Data that Reduces Dimensionality using Dimension Reduction Techniques, *International Journal of Computer Applications in Engineering Sciences (IJCAES)*, 1, March 2011, 80-84.

[5] Halkidi, M., Batistakis, Y. and Vazirgiannis, M. (2001). On Clustering Validation Techniques, *Journal of Intelligent Information Systems*, 17:2/3, Kluwer Academic publishers, manufactured in the Netherlands, 107-145.

[6] Irizarry, R., Hobbs, B., Collin, F., Beazer-Barclay, D., Antonellis, J., Scherf, U., Speed, P. (2003). Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data, *Biostatistics*, 4(2): 64- 249.

[7] Izenman, A. J. (2008). *Modern Multivariate Statistical Techniques, Regression, Classification and Manifold learning*, Springer Science & Business, Media Philadelphia, PA 19122, USA.

[8] Jonathan, M. G., Daniele, S. and Khhairul, A. R. (2010). Consensus Clustering and Fuzzy Classification for Breast Cancer Prognosis, *24th European Conference on Modeling and Simulation*, June 1- 4th, Kuala Lumpur, Malaysia, 15-22.

[9] Joseph, F. H., Rolph, A. E., Ronald, L. T. and William, C.B. (2003) *Multivariate Data Analysis, Fifth edition*, published by Pearson Education, First Indian Reprint 2003, 493-496.

[10] Kumar, H. and Sharma, V. (2011). A Comparative Study of k-mean and PAM Algorithms using Leukemia Datasets, International symposium on computing, communication, and control, ISCCC, Proc. of CSIT ,Vol. 1(2011), IACSIT Press, Singapore , 136-140.

[11] Murillo J. and Rodriguez A. (2012). Linear dimensionality reduction with Gaussian mixture models, ICASSP, 27 March 2012, Japan.

[12] R manual documentation (2012).

[13] Robert M., Richard G., James H., (2003). Statistical design and analysis of Experiement with applications, Hoboken, N.J, Willey-2003, Pp. 98-104.

[14] Sanche, R., and Lonergan, K. (2006). Variable Reduction for Predictive Modeling with Clustering, Casually actuarial society forum, 89-100.

[15] San T. Roweis and Lowerence Saul (2000). Nonlinear dimensionality reduction by locally linear embedding, science Vil. 200, 22-Dec. 2000, pp 2323-2326.

[16] Sauerbrei, W. and Royston, P. (1999). Building Multivariable Prognostic and Diagnostic Models: transformation of the predictors by using fractional polynomials, *Journal of the Royal Statistics Society Series A*, 162(1), 71–94.

[17] Schumacher, M., Basert, G., Bojar, H., Huebner, K., Olschewski, M., Sauerbrei, W., Schmoor, C., Beyerle, C., Neumann, A. and Rauschecker, H. (1994). For the German Breast Cancer Study Group, Randomized 22 times trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients, *Journal of Clinical Oncology*, 12, 2086–2093.

[18] Shuiwang Ji, Jieping Ye., (2009). linear dimensionality reduction for multi-label classification, IJCA International joint conference on artificial intelligence, 17 Jul. 2009, 1077-1082.