International Scholars Journals

Advanced Journal of Microbiology Research ISSN 2241-9837 Vol. 12 (12), pp. 001-010, December, 2018. Available online at www.internationalscholarsjournals.org © International Scholars Journals

Full Length Research Paper

Identification of clusters in tissue samples in gene expression data with Principal Component Analysis based on relative variance matrix

Uzma Nawaz* and Asghar Ali

Department of Statistics, Bahauddin Zakariya University, 60800, Multan, Pakistan.

Accepted 04 December, 2018

Principal Component Analysis (PCA) has been in use as a preprocessing step to clustering for long. We have focused on the clustering of tissue samples in gene expression data. Different clustering techniques and algorithm are available in literature on gene expression data but with the existing ambiguity on the number of clusters, apart from relying on biologically known groups. A consensus is needed to reach on the number of clusters in the wide variety of existing clustering techniques based on different similarity or dissimilarity metrics. The conventional usage of PCA for clustering is either by forcing the unit variance to each variable or the high magnitude of variance of an individual variable is allowed to dominate the entire results of PCA. We propose the use of relative variance covariance method in PCA, so as to give due consideration to the joint and individual variances in the dataset and identify clusters with principal component loadings. We emphasize empirically that the proposed approach of PCA is conclusively more informative than the available approaches to identify cluster structure in tissue samples (sample expression profiles). Clusters formed are valid with the existing results on the data set under study and with valid biological background.

Key words: Clustering methods, gene expression analysis, principal component analysis, the relative variance covariance matrix, principal component loadings.

INTRODUCTION

Genomic research has successfully demonstrated the utility of DNA micro array-based gene expression data in cancer classification. The main goal of micro array analysis, in particular of unclassified cancer, is to identify novel cancer subtype for subsequent validation and prediction, and ultimately to develop individualized prognosis and therapy. As a preliminary step of this study, clustering is being used as a useful exploratory technique in gene expression data to cluster tissue samples and genes. A major problem with the application of clustering algorithms is that an adequate number of clusters can often not be inferred automatically. A purely data-driven approach always bears the risk of over- or under-clustering because the correct number of clusters usually depends on task-specific constraints.

Each clustering algorithm has its own set of biases on clusters it construct, whereas the most sensible clustering algorithm may yield similar results on trivial test problems, in practice they can give widely different results on messy real world gene expression data (D'haeseleer, 2005). A number of clustering techniques on gene expression data sets have been practiced with the success of clustering algorithm, assessed by visual inspection using biological knowledge (Eisen et al., 1998, Golub et al., 1999, Quackenbush, 2001; Akashi et al., 2003). The challenge of interpretation in the absence of such biological knowledge is accepted in our study of clustering the sample expression profiles in the dataset by the proposed approach. Studies in the context of clustering of tissue samples include classifying sixty human cancer lines (Ross et al., 2000), distinguishing two different human acute leukemia's (Golub et al., 1999), dissecting and classifying breast cancer tumors

*Corresponding author. E-mail: u_nk@yahoo.com. Tel: +92 61 3146124228.

(Perou et al., 1999), classifying subtypes of B-cell lymphoma (Alizadeh et al., 2000), and cutaneous malignant melanoma (Bittner et al., 2000). Bullinger et al. and Valk et al. in 2004 identified prognostic subclasses in acute myeloid leukemia, and in the same year (Lapointe et al., 2004) found tumor subtypes of prostrate cancer with the provision of basis for improved prognostication and treatment stratification.

In the study, we emphasize that clustering with Principal Component Analysis (PCA) provides a means of projecting the data into a lower dimensional space, making the visual inspection hopefully more informative. One such application of PCA has been (Tamayo and Ramaswami, 2002) on the initial leukemia gene dataset of Golub et al. (1999). The top three principal components (PCs) for the 612 most highly varying genes in the Leukemia subtypes dataset were projected in 3D plot. The exhibited data structure was interpreted corresponding to the known three morphological subclasses (The ALL-B, ALL-T and AML) in the leukemia cancer type. The 3D plot of the PCs revealed overlapping of the samples among the three Leukemia types and a clear four cluster structure. A sample from ALL-B visually represented an outlier, lying at the farthest distance in the extreme left corner of the 3DPlot. These aspects went unnoticed, which in fact required an interpretation apart from the explanation of the three known leukemia type of the dataset.

A review of application of PCA

A criticism in applying PC cluster analysis has been that the first few PC's (which explain most of the variation in the data) does not necessarily capture most of the cluster structure (Yeung et al., 2001). The criticism was based on the theoretical results (Chang, 1983) showed that first few PC's may not contain cluster information: assuming that the data was a mixture of two multivariate normal distributions with different means but with an identical within-cluster covariance matrix, and the first few PC's may contain less cluster structure information than other PC's . An artificial example was generated, in which there were two clusters and the data points were visualized in two dimensions only, the two clusters were well separated in the subspace of first and last PC. In response, it has been demonstrated empirically (Ben-Hur et al., 2002) that PCA has the ability to extract features relevant to the cluster structure, that the few leading PC's enhances cluster structure and the results of Yoeung et al. (2001) were the use of standardization as normalization in the analysis, so reduced the quality of the clustering, rather than the use of PCA. Then it has been concluded that the degradation in the recovery of the clusters through PC's should be attributed to normalization, rather than to the use of PCA (Anderberg, 1983; Milligan and Cooper, 1988). Since PC's are uncorrelated and ordered, the first few PC's, which explain most of the variation in the dataset, are usually used in cluster analysis, as these may extract the cluster structure in the dataset (Jolliffe, 1980).

We have emphasized the use of PCA with variance covariance matrix (Σ_{rel}) apart from the conventional approaches to derive PCs to cluster sample expression profiles in the data under study. The methodology of Σ_{rel} approach to cluster the variables; the sample expression profiles of the gene expression dataset has been exercised in the study and is a novelty in clustering techniques applied on variables in highly skewed gene expression datasets. The clusters formed are later verified by the biological knowledge and with room for future consensus on the configuration of clusters.

The approach of the Σ_{rel} in PCA for the interpretation of PCs, as an alternative to the variance covariance matrix (Σ) and the correlation matrix (R) has been introduced in literature (Wajid and Ali 1998). Describing the merits and demerits of Σ and R in PCA, Σ_{rel} was used as an intermediate way to fill in the gap between the two approaches to derive PCs on two data sets measured on the same and different scale. Concluding empirically that the three methods reveal different features of the correlation structure of the datasets and all the interpretations were equally useful to assess the hidden features in the datasets. The precedence of Σ_{rel} in PCA has been further worked out lately as well (Boik and Shirvani 2008). A PC model was proposed for the Σ_{rel} and then the least square estimators of the eigen values and eigenvectors of Σ_{rel} were developed with empirical demonstration on real data and the simulative validation of proposed inference procedure.

We have introduced the interpretation of PC's derived from Σ_{rel} to cluster samples and the PC loadings in a 3D representation: a simple multivariate method. In doing so, comparison has been made with PC loadings derived with R (the conventional approach) to identify cluster structure in the dataset. The former approach is found to be better than the later approach.

MATERIALS AND METHODS

Clustering methods: A comparison

Most often the clustering methods branch off as (i) hierarchical and (ii) partitioning. Hierarchical clustering is a method useful for dividing data into natural groups by organizing the data into a hierarchical tree structure ("dendogram") based upon the degree of similarity between either sample or genes. Since the emergence of cluster structure depends on several choices: (i) data representation (ii) normalization (iii) the choice of a similarity measure and (iv) the clustering algorithm. Therefore the hierarchical clustering method, a highly structured method provides different results because of the choice of the similarity metric (Goldstein et

al., 2002).

Partitioning method (non-hierarchical) on the other hand, subdivide the data into a typically predetermined number of subsets without any implied hierarchical relationship between these clusters like k-means clustering. The $k-means$ clustering result depends on the initial partition (initial seeds) of the clusters and is preferred over the hierarchical methods in computation for large data sets. If the initial seeds are selected according to some known features of the data, then the $k-means$ clustering is quite robust (Milligan, 1980). The pivot is with initial seeds of clusters chosen to be genes of vital and known biological functions, and $k-means$ clustering would be robust with meaningful results. When seed genes are not available or hard to find which the case is usually, then the option is viewing hierarchical and nonhierarchical techniques as complementary to one another, that is treating $k-means$ clustering with initialization by either complete or average linkage or any other hierarchical clustering output. Hence, the choice of similarity metric is still there. The Self Organizing Map (SOM) is another non hierarchical clustering algorithm where a grid of two dimensional (2D) nodes (clusters) is iteratively adjusted to reflect the global structure in the expression dataset. A 2-cluster SOM was used to cluster the initial set of 38 leukemia samples into two classes based on the expression pattern of 6817 genes. The two SOM clusters were then compared to the known lymphoblastic vs myeloid leukemia (AML and ALL) distinction. The two SOM clusters closely paralleled this morphological distinction, with the first cluster containing mostly ALLs (24 out of 25 samples) and the second containing mostly AMLs (10 out of 13 samples). Thus, the clustering algorithm was effective but not perfect at separating samples into biologically meaningful groups. Sub-classifications for further samples was searched constructing a 4-class (2 × 2) SOM. The clustering algorithm was successful at separating the samples into more refined groups reflecting other important biological distinction: different ALL cell lineages (B- and T-Cell) (Golub et al., 1999).Thus the clustering results were searched further to be interpretable in the context of a prior knowledge (that is, known leukemia sub-classes). Yet the search was not effective in exclusively defining the three known biological groups exclusively. The overlapping feature among the samples existed but was not explored.

The Principal Component Analysis

PCA is a mathematical technique that transforms a set of p variables $\underline{X} = [X_1, \ldots, X_p]^T$ into a new set of variables, such that these transformed variables $\underline{Y} = [Y_1, Y_2, \ldots, Y_p]^T$ are orthogonal to each other and are derived in such a way that the first few of these explain almost all the variation in the data. These new set of variables are the Principal Components(PCs) , which are presented in an ordered form, so that the first PC explain the highest proportion of variation, the second explain maximum of the remaining proportion of variation and so on. As a result the variation in the data explained is condensed by a few numbers of orthogonal variables which are easily interpretable.

PCA with correlation matrix (R)

Let the operator $Diag(a)$ is the diagonal matrix whose ith diagonal component is a_i.

Then $D_\sigma = Diag(\sigma) = Diag(\sigma_{11}^{1/2}, \sigma_{12}^{1/2}, \ldots, \sigma_{pp}^{1/2})$ where σ_{ii} is variance of the ith variable.

The standardized version of $\underline{X}$ is defined as $\underline{Z} = [D_\sigma^{-1}(\underline{X} - \underline{\mu})]$ and $Cov(\underline{Z}) = R$, the correlation matrix of $\underline{X} = [X_1, \ldots, X_p]^T$

PCA defines $\underline{Y} = [Y_1, Y_2, \ldots, Y_p]^T = \Gamma^T \underline{Z}$ with $\Gamma = [\alpha_1, \ldots, \alpha_p]^T$, the orthogonal matrix of eigen vectors such that $\Gamma^T R \Gamma = \Lambda$

Where $\Lambda = Diag(\lambda_1, \lambda_2, \ldots, \lambda_p)$, satisfying $\lambda_1 \geq \lambda_2 \geq, \ldots, \lambda_p \geq 0$ as the variance-covariance matrix of the $\underline{Y}$.

The ith PC is then defined as $Y_i = \alpha_i^T \underline{Z}$ or $\underline{Z} = \alpha_i Y_i$, for which (i) $E(Y_i) = \underline{0}$,(ii) $Var(Y_i) = \lambda_i$, and (iii) $Cov(Y_i, Y_j) = 0 \quad \forall i \neq j$

The coefficients of α_i, the ith eigen vector of Γ is called the ith PC. Finding the variance of

$$Var(\underline{Y}_i) = \sum_i^p Var(Y_i) = \sum_i^p \lambda_i$$

$$Var(\underline{Y}) = tr(\Lambda) = tr(R) = p$$

Therefore p is the total variation in the data equivalent to the total number of variables and ith PC accounts for $\lambda_i \Big/ \sum_{i=1}^{p} \lambda_i = \lambda_i / p$, the proportion of the total variation, using R.

PCA with relative variance covariance matrix (Σ_{rel})

Let $\underline{X}$ now be defined as $\underline{X}_{rel} = D_w \underline{Z}$ with $w_i = \sigma_{ii} / \mu_i$ and $D_w = Diag(w) = Diag(w_1, \ldots, w_p)$ with means on diagonal places.

$D_w = D_\mu^{-1} D_\sigma$ with $D_\mu = Diag(\mu)$

$Cov(\underline{X}_{rel}) = D_w R D_w = \Sigma_{rel}$.

The new set of orthogonal PCs under the Σ_{rel} is

$\underline{Y}_{rel} = [Y_{rel1}, \ldots, Y_{relp}]^T = \Gamma_{rel}^T \underline{X}_{rel}$ Where $\Lambda_{rel} = Diag(\lambda_{rel_1}, \lambda_{rel_2}, \ldots, \lambda_{rel_p})$ and $\Gamma_{rel}^T \Sigma_{rel} \Gamma_{rel} = \Lambda_{rel}$ satisfies $\lambda_{rel_1} \geq \lambda_{rel_2} \geq, \ldots, \geq \lambda_{rel_p} \geq 0$, the variance-covariance matrix of the $\underline{Y}_{rel}$.

$Y_{reli} = \alpha_{rel_i}^T \overline{X}_{rel}$ or $\overline{X}_{rel} = \alpha_{rel_i} Y_{rel_i}$

The ith PC is defined by satisfying the properties (i), (ii), (iii) defined in Section 2.2.1 with α_{rel_i}, the ith eigenvector of Γ_{rel}, the ith PC loadings for the Σ_{rel}.

$$Var(\underline{Y}_{rel}) = var(Y_{rel_i}) = \lambda_{rel_i}^p = tr(\Lambda_{rel}) = tr(\Gamma_{rel}^T \Sigma_{rel} \Gamma_{rel})$$
$$= tr(\Sigma_{rel}) = Var(\underline{X}_{rel}) = D_w^2$$

PCA with Σ_{rel} versus PCA with R

When sample variances differ widely in magnitude, use of the standardized version of the variables is very much in convention. Thus the individual variances are standardized to one, thereby eliminating the effects of variances. Then the linear combinations in PC's are found that have maximal variance, even though the effect of the disparity in variances have been removed. Explaining conclusively that, R is inherently less informative than Σ; the variance covariance matrix of the data (Anderson, 1963). We propose that Σ_{rel} is as informative as Σ because D_w^2 are the informative relative variances. The few PC's of Σ_{rel} highlight variables that have larger relative variances just like Σ, thus the PC's computed on Σ_{rel} are not subject to Anderson's criticism because the effects of variances have not been eliminated rather expressed in terms of their means as Σ_{rel} is a matrix of transformed data, on which the meaning of origin has been preserved ($\underline{X}_{rel} = D_w \underline{Z}$), with Σ_{rel} as scale invariant and choice of scale transformations need not be made.

PC loadings

The interpretations of PC's are distilled from the respective coefficient (loadings) of the eigenvector (α_i). The ith PC is interpreted by looking at the loading for each variable (the sample expression profiles). Variables with small magnitude are ignored and the PC is then approximated by the linear combination involving only the remaining variables with higher component loadings. Geometrically the elements of α_i are the direction cosines of the PC axis to the old coordinate axis and correlation is as just the cosine of the angle θ_{ij} subtended at the origin between the two axes the original and the PC axis, with inner product as the covariance between PC axis and the original coordinate axis. Then

$$Cos\theta_{ij} = Y_{rel_i} . X_{rel_j} / \left| Y_{rel} \right| \left| X_{rel} \right| = Corr(Y_{rel_i} X_{rel_j})$$
$$Cov(X_{rel}, Y_{rel}^T) = E(X_{rel} Y_{rel}^T) = E(\Gamma_{rel} Y_{rel} Y_{rel}^T) = \Gamma_{rel} \Lambda_{rel}$$

Where Λ_{rel} is the Variance-Covariance matrix of the new set of p-variate vector the $\underline{Y}$.

Thus
$$Cos\theta_{ij} = \alpha_{rel_i} (\lambda_{rel_i} / \sigma_{jj})^{1/2} \mu_j = Corr(Y_{rel_i} X_{rel_j}) \quad (1)$$

The proposed study would show the use of PC loadings from Equation (1) for clustering the variables that is the sample expression profiles in the data under study.

Data analysis

The dataset under study is the gene expression initial leukemia data set, available at http://www.genome.wi.mit.edu/MPR. The dataset consist of 38 bone marrow samples, 27 acute lymphoblastic leukemia (ALL) and 11 acute myeloid leukemia (AML) obtained from Acute Leukemia patients at the time of diagnosis before chemotherapy. The ALL type has a futher classification of 19 ALL-B cell leukemia and 08 ALL-T cell leukemia samples. PCA is used as a visualization tool to provide a low dimensional summary of the data. The 3Dimensional (3D) scatter plots of PC loadings under the two approaches have been used to visualize clusters in the dataset. The 3D scatter plot often reveal group structure in a dataset better than looking at a series of 2D plots. The data set was processed with the preprocessing steps of Dudoit et al. (2002a) of thresholding: floor of 100 and ceiling of 16000; filtering: exclusion of genes with max/min ≤ 5 and (max-min) ≤ 5000 where the max and min refer, respectively, to the maximum and minimum expression levels of a particular gene, across a tissue sample. Natural logarithm of the expression levels is used to provide good variance stabilization at high levels of gene expressions. Thus, we have 2299 gene expressions for each of the samples.

RESULTS

Following Kaiser criterion of $\lambda_j \geq 1$ PCs retained are $k = 4$ using R, that are accounting for 68% of the total variation in the dataset. Same numbers of PCs are retained from PCA using Σ_{rel} explaining 68.70% of the relative variation in the dataset (Table 1). The magnitude of variation explained by the retained PC's under the two approaches is almost the same but owing to the concept of the Σ_{rel} explanation of the variation, explains the relative variance of each sample, with the others and is the ratio of the standard deviation to the mean. The standard deviation of data must always be understood in the context of the mean of the data specifically when the mean vary widely as in our dataset.

Interpreting PC loadings

The principal components derived under the two approaches are presented in Table 2. The two immediate contrasts for magnitudes of PC loadings are the positive (non bold font) and negative loadings (in bold font) under the methods (Table 2). The positive and negative loadings (in bold letters) are further fairly grouped as loadings greater than as and less than 0.1. PC1 derived from the two methods is the general index of the dataset with almost the same magnitude of loadings. PC4 do not clarify any distinctive biological group and shows that there exist further sub grouping in specifically ALL-B samples and its overlapping with the AML samples. The

Table 1. Proportion of variance explained by PCs (R and Σ_{rel}).

Var explained	PC1	PC2	PC3	PC4	Total variance
R	53.82%	6.70%	4.30%	3.20%	68%
Σ_{rel}	54.45%	6.76%	4.22%	3.22%	68.70%

Table 2. Loading table of the retained PCs.

		PC loadings using R ($\alpha_i s'$)				PC loadings using Σ_{rel} ($\alpha_{i,rel} s'$)			
		PC1	PC2	PC3	PC4	PC1	PC2	PC3	PC4
Circle small in size	X1 ALL-B	-0.176	0.115	0.026	0.174	-0.167	0.116	0.037	0.140
Plus	X2 ALL-B	-0.175	0.114	0.012	0.153	-0.172	0.118	0.022	0.133
Cross	X3 ALL-B	-0.170	0.149	0.160	-0.238	-0.150	0.123	0.144	-0.211
Dot	X4 ALL-B	-0.170	0.047	-0.010	0.231	-0.171	0.056	-0.003	0.216
Solid circle	X5 ALL-B	-0.159	0.022	0.014	0.403	-0.165	0.038	0.029	0.406
Asterisk	X6ALL-B	-0.170	-0.097	0.044	0.057	-0.179	-0.093	0.042	0.051
Dot circle	X7 ALL-B	-0.167	0.231	0.207	-0.129	-0.160	0.216	0.216	-0.146
Plus circle	X8 ALL-B	-0.170	0.217	0.137	-0.156	-0.163	0.203	0.143	-0.162
Cross circle	X9 ALL-B	-0.174	0.197	0.109	0.000	-0.161	0.182	0.115	-0.022
Circle circle	X10 ALL-B	-0.157	0.061	0.017	-0.156	-0.158	0.065	0.023	-0.159
Square	X11 ALL-B	-0.161	0.054	0.059	0.034	-0.157	0.059	0.067	0.020
Solid square	X12 ALL-B	-0.168	0.131	0.197	0.037	-0.165	0.128	0.207	0.012
Dot square	X13 ALL-B	-0.150	0.214	0.230	-0.101	-0.144	0.199	0.240	-0.110
Cross square	X14 ALL-B	-0.131	0.154	0.089	0.026	-0.143	0.180	0.135	0.010
Diamond	X15 ALL-B	-0.163	-0.007	0.067	0.301	-0.173	0.005	0.085	0.310
Solid diamond	X16 ALL-B	-0.165	0.215	0.093	-0.073	-0.155	0.199	0.102	-0.085
Dot diamond	X17 ALL-B	-0.167	-0.041	0.147	0.093	-0.165	-0.035	0.146	0.076
Plus diamond	X18 ALL-B	-0.172	0.141	0.183	0.021	-0.166	0.135	0.188	-0.007
Triangle	X19 ALL-B	-0.148	0.051	-0.036	0.376	-0.162	0.075	-0.024	0.416
Solid triangle med in size	X20 ALL-T	-0.170	-0.061	-0.095	-0.214	-0.166	-0.049	-0.094	-0.198
Dot triangle	X21 ALL-T	-0.166	0.072	-0.348	0.044	-0.166	0.095	-0.340	0.038
Triangle right	X22 ALL-T	-0.165	0.041	-0.372	0.006	-0.170	0.064	-0.383	0.002
Solid triangle right	X23 ALL-T	-0.165	0.080	-0.315	-0.181	-0.163	0.096	-0.304	-0.175
Dot triangle right	X24 ALL-T	-0.169	0.014	-0.232	-0.141	-0.172	0.031	-0.234	-0.143
Triangle left	X25 ALL-T	-0.165	0.061	-0.278	-0.140	-0.165	0.079	-0.269	-0.138
Solid triangle left	X26 ALL-T	-0.169	0.066	-0.181	-0.194	-0.161	0.073	-0.165	-0.180
Dot triangle left	X27 ALL-T	-0.160	0.045	-0.367	0.107	-0.162	0.072	-0.364	0.102
Circle large in size	X28 AML	-0.170	-0.150	0.003	-0.009	-0.171	-0.141	-0.007	-0.013
Plus circle	X29 AML	-0.161	-0.114	-0.039	0.158	-0.165	-0.103	-0.044	0.153
Solid circle	X30 AML	-0.146	-0.278	0.076	-0.151	-0.152	-0.287	0.056	-0.146
Circle circle	X31 AML	-0.141	-0.350	0.089	-0.025	-0.150	-0.363	0.071	-0.012
Cross circle	X32 AML	-0.145	-0.224	0.028	0.081	-0.152	-0.223	0.020	0.092
Dot circle	X33 AML	-0.169	-0.211	0.074	-0.017	-0.161	-0.190	0.058	-0.013
Square	X34 AML	-0.153	-0.200	0.125	-0.227	-0.159	-0.208	0.116	-0.230
Solid square	X35 AML	-0.145	-0.261	0.104	-0.222	-0.155	-0.281	0.089	-0.224
Dot square	X36 AML	-0.170	-0.175	-0.050	-0.026	-0.178	-0.174	-0.070	-0.027
Cross square	X37 AML	-0.146	-0.163	0.010	-0.018	-0.167	-0.148	-0.001	-0.017
Diamond	X38 AML	-0.138	-0.338	0.089	0.137	-0.144	-0.341	0.073	0.155

Table 3. Cluster configurations from retained PCs under the two methods.

	Negative loadings		Positive loadings	
	< 0.1	≥ 0.1	< 0.1	≥ 0.1
PC2 with R	X6, X15, X17, X20	X28-X38 AML group	X4, X5, X10, X11, X19, X21-X27	X1-X3, X7-X9, X12-X14, X16, X18 Subgroup of ALL-B
PC3 with R	X4, X19, X20, X29, X36	X21-X27 7/8 ALL-T samples	X1, X2 ,X5 ,X6 X10, X11, X14, X15, X16, X28, X30-X33	X3, X7-X9, X12 X13, X17, X18, X34, X35, X37, X38 The Overlapping Cluster (4 AML samples)
PC2 with Σ_{rel}	X6, X17, X20	X28-X38 AML group	X4, X5, X10, X10, X15, X19, X21-X27	X1-X3, X7-X9, X12-X14, X16, X18 Subgroup of ALL-B
PC3 with Σ_{rel}	X4, X19, X20, X28, X29, X36, X37	X21-X27 7/8 ALL-T samples	X1, X2, X5, X6, X10, X11, X15, X28, X30-X33, X35, X38	X3, X7-X9, X12-X14, X16-X18, X34 Sub group of ALL-B with one AML sample

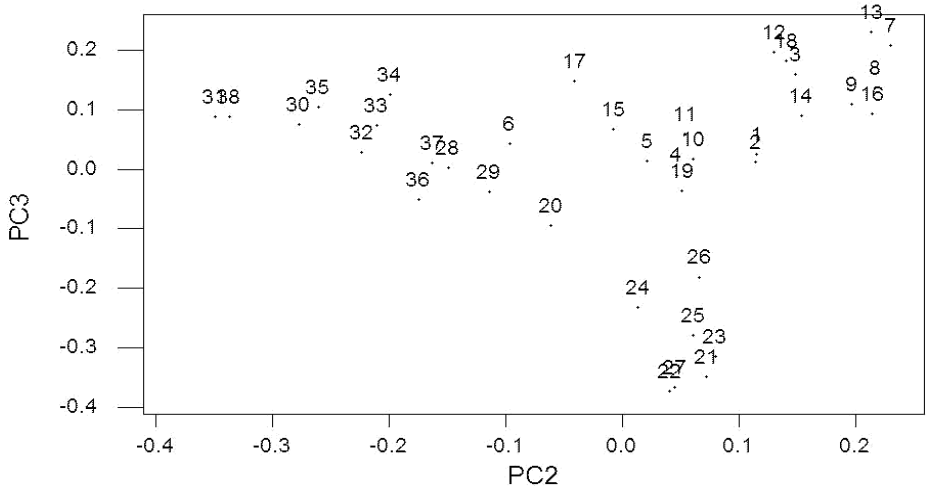

Figure 1. 2D plot of PC loadings using correlation matrix.

detailed configuration of tissue samples with respect to positive and negative loadings is presented in Table 3.

The interpretation of PC2 derived from the two approaches show a distinctive contrast of leukemia type the AML with 11 out of 19 samples of ALL-B leukemia group with high positive and negative loadings. PC3 derived from both methods single out the ALL-T tissue samples (X21-X27) as leukemia type widely different from the others. Noticeably X20 the ALL-T sample with not a significant magnitude of loading, in any case do not fall in its own biologically known group rather groups with few of the AML and ALL-B samples. The distinctive feature is that none of the biological known group retains its exact individual identity of biological significance; the overlapping of samples is very much evident and must be taken into consideration, well supported by the fact that all these samples are positively correlated with one another. For an initial exploration of clusters we choose to draw a 2D scatter plot of the two most informative PCs (PC3 vs PC2). Interestingly, the two plots Figures 1 and 3 appear to be the exact replica of one another, visible are the four clusters, with same configuration (Figures 1 and 3).

The four clusters are from the top left corner of Figures 1 and 3 the configuration is as follows:

1. X30- X35 and X38 (A subgroup of AML samples)
2. X6, X20, X28, X29, X36, X37 (The overlapping cluster)
3. X21 – X27 (ALL-T samples)
4. X1 – X5, X7 – X19 The ALL-B samples appear to be

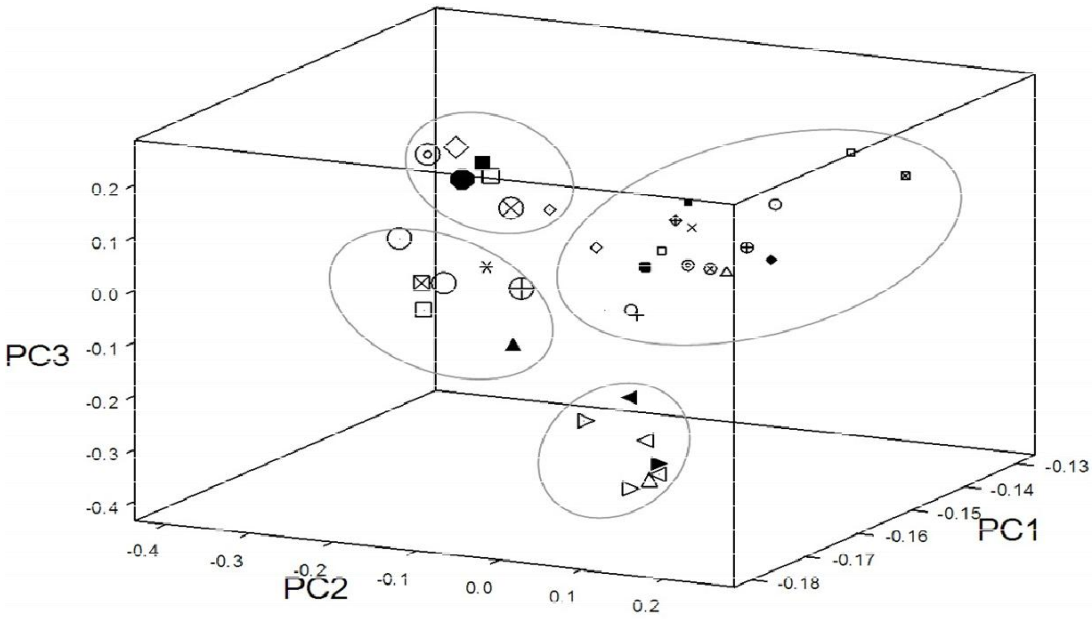

Figure 2. 3D plot of PC loadings using correlation matrix.

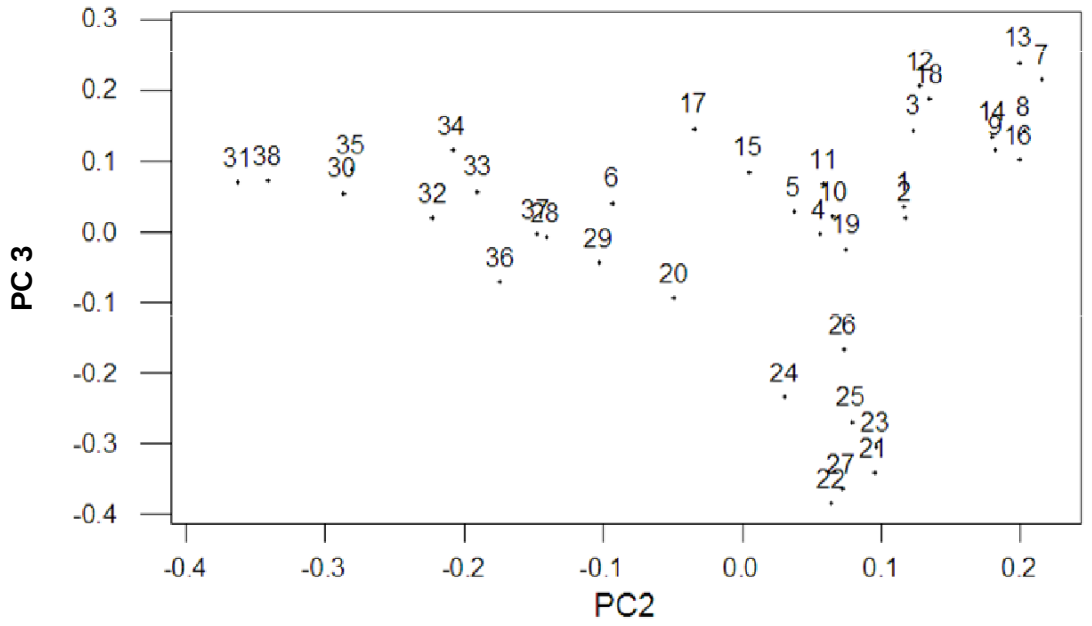

Figure 3. 2D plot of PC loadings using relative variance covariance matrix.

one big cluster, with a sub clusters structure in it.

PC1 which is the projection on the direction in which the variance of the projection is maximized is generally not taken for visual representation instead a series of 2D plots is preferred. For a distinction between the two representations, here we project through 3D loading plot of PC1, PC2 and PC3, the clusters and their configuration.

Figures 2 and 4, represents 3D loading plot of PCs derived from the correlation and relative variance covariance matrix of the dataset. The visible number of clusters is four. The ALL-B sample X_{17} (identified by dot diamond) in Figure 3 falls in cluster 1(Subgroup of AML samples) with X_{33} the AML sample grouping in cluster 2 (the overlapping cluster).

From Figure 4, it can be seen that these are back in their own biological groups of ALL-B and AML samples

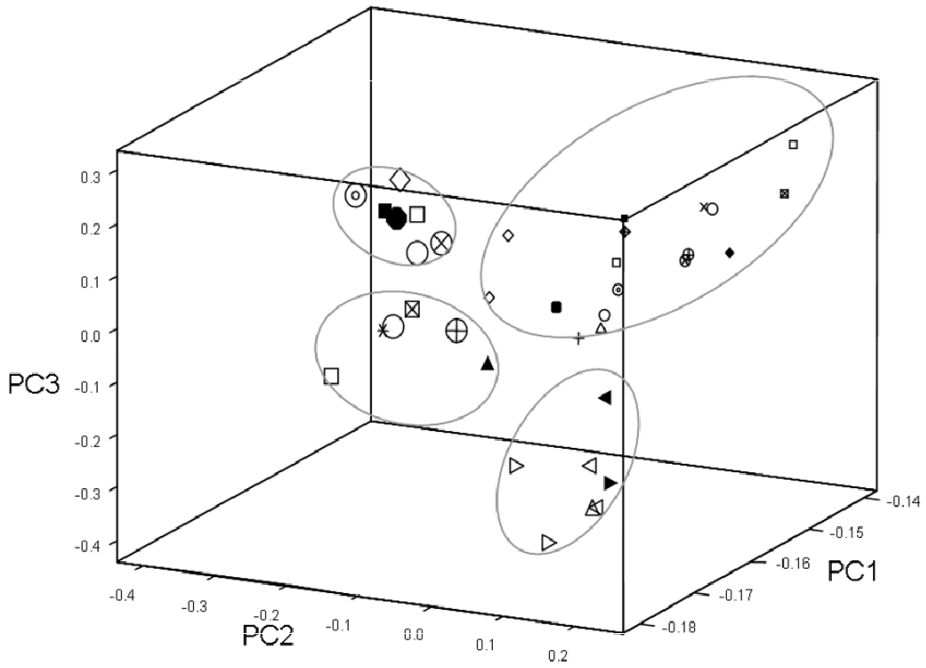

Figure 4. 3D plot of PC loadings using relative variance covariance matrix.

respectively. Here, we note that PC1 (the general index of the data set) derived from Σ_{rel} clarify the clusters and confirms with the configuration of clusters presented in the unanimous 2D presentation. Merely stating that relative variance covariance matrix approaches the data set in a more informative way in identifying the true structure lying among samples expression profiles in the data set.

DISCUSSION

The conventional approach of correlation matrix in PCA has been presented in comparison with the proposed approach of relative variance covariance matrix in PCA to capture cluster structure in the data set. Both the approaches unanimously suggest four main clusters in the dataset. The clusters are not exactly mutually exclusive with respect to their biological identities, instead there exist an overlapping of ALL-B samples with AML and ALL- T samples an important aspect that need to be explored. The overlapping feature and a four cluster structure were observed with the clustering technique SOM. The objective was to find and justify the three morphologically known biological samples distinctively. In the use of agglomerative clustering with average linkage method on 38 samples with 3000 genes in fact revealed four clusters, two clusters of AML and ALL-T samples with other two of ALL-B samples. The four clusters were achieved, thereby further splitting the dendogram objectively to justify the three known classes of leukemia types. The cluster configuration apparently seemed different in the two approaches. The other application of PCA by (Tamayo and Ramaswamy, 2002) on 618 genes across 38 samples, confirms the previous approach of clustering to justify the known biological samples. The study completed in this paper provide an interpretations of the first four PC from the two approaches, that distinguish the 3 known biological cluster of samples with the overlapping cluster dominant in the two approaches. The configuration of the overlapping cluster has been presented and matches the representation in 2D plots and Figure 4 (the 3D representation under the Σ_{rel}).

In addition, the proposed method on the data set is explaining variation in the dataset using relative variance approach. without eliminating the effect of variance through equi-variance, as in R rather the relative variance of the data provide better comparison being as informative as the variance covariance matrix and thereby preventing the first PC to be dominated by large Σ_{rel} variance in the dataset. The first PC of being the general index of the dataset (or the size component of the dataset) and capturing most of the relative variance of samples with each other brings the difference in the two approaches and finalizes the configuration of clusters formed. The dataset has been interpreted without purely relying on the priori biological knowledge and the results are validated with the following biological knowledge.

Research in cancer studies is focusing on the option that myeloid cells and B-lymphocytes may be sharing the one and same source of progenitor stem cell. These

might be sharing the same gene type as these both originate from the bone marrow in the human body. An overlapping of ALL-B samples exits with the AML and the ALL-T samples. The fact that B and T lymphocytes share the common function of developing antibodies but originate from different sources as T-Lymphocytes originate from the thymus in the human body. They may group together on the basis of similar function and provide a biological validation of the results of number of clusters and their configuration. The overlapping cluster be defined as a new tumor subtype and may provide a basis for improved prognostication and treatment stratification. The sub clusters of ALL- B leukemia needs to be configured, explored and verified.

Conclusion

Conclusively, Σ_{rel} for deriving PC and its use in clustering of samples is more informative in identifying cluster structure in a highly skewed type of gene expression data. In our study, PC1 has emerged with a unanimous finalization number and configuration of clusters, and provided the intricate difference between the two approaches. We propose that the gene expression data sets that are usually a highly skewed type of data sets need to be explored for clustering samples at the preliminary stages, apart from the conventional usage of correlation matrix, with the methodology of relative variance matrix for extracting the features in the datasets.

Thus, the PCA based clustering may be regarded as a competitor to consensus on the number of clusters to standard cluster analysis of samples in gene expression dataset. The whole family of clustering methods, differing only in the way inter cluster distance is defined (the "linkage function"), resulting in different number and formation of clusters. In clustering with PCA as an exploratory method, the choice of similarity and dissimilarity measures reduces to covariance and correlation matrices (Jackson, 2002). This choice in our study has further reduced to the relative variance covariance matrix, specifically to deal with highly skewed type of gene expression datasets.

REFERENCES

Akashi K, He X, Chen J, Iwasaki H, Niu C, Steenhard B, Zhang J, Perera R, Haug J, Li L (2003). Transcriptional Accessibility for Multi-Tissue and Multi-Hematopoietic Lineage Genes is Hierarchically Controlled During Early Hematopoiesis. Blood, 101: 383-390.

Alizadeh A, Eisen M, Davis RE, Ma C, Lossos IS, Rosenwal A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti G, Moore T, Hudson J, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM (2000). Distinct types of diffuse large B-cell Lymphoma identified by gene expression profiling. Nature, 403: 503-5111.

Anderberg M (1983). Cluster analysis for applications. Academic Press, New York.

Anderson TW (1963). Asymptotic theory for principal component analysis, Ann. Math. Stat., 34: 122-148.

Ben-Hur A, Elisseeff A, Guyon I (2002). A stability based method for discovering structure in clustered data, in Pacific Symposium on Biocomputing (Altman R, Dunker A, Hunter L, Lauderdale K, and Klein T, eds.) World Scientific, pp. 6–17.

Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher E, Simon R, Yakhini Z, Ben-Dor A, Sampas N, Dougherty E, Wang E, Marincola F, Gooden C, Lueders J, Glatfelter A, Poollock P, Carpten J, Gillanders E, Leja D, Dietrich K, Beaudry C, Berens M, Alberts D, Sondak V, Hayward N, Trent J (2000). Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature, 406: 536-540.

Boik RJ, Shirvani A (2008). Principal Components on Coefficient of Variation Matrices, Statistical Methodology, doi: 10.1016/j.stamet.2008.02,2006.

Bullinger L, Duohner K, Bair E, Frohling S, Schlenk RF, Tibshirani R, Dohner H, Pollack JR (2004). Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. New England J. Med., 350: 1605-1616.

Chang WC (1983). On using principal components before separating a mixture of two multivariate normal distributions. Appl. Stat., 32: 267-275.

D'haeseleer P (2005). How does gene expression clustering work? Nat. Biotechnol., 23: 1499-1501.

Dudoit S, Fridly J, Speed TP (2002a). Comparison of discrimination methods for the classification of tumors using gene expression data. J. Am. Stat. Assoc., 97: 77-87.

Eisen M, Spellman P, Brown P, Botstein D (1998). Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA, 95: 14863-14868.

Everitt BS, Landau S, Leese M (2004). Cluster Analysis. 4th ed, Arnold, London.

Goldstein DR, Ghosh D, Conlon EM (2002). Statistical Issues in the Clustering of Gene Expression Data. Statistica Sinica, 12: 219-240.

Golub TR, Slonim DK, Tamayo P, Huard C, Gassenbeck M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri, MA, Blommfield CD, Lander ES (1999). Molecular classification of Cancer: class discovery and class prediction by gene expression monitoring. Science, 286: 531-537.

Golub TR, Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proceedings of the National Academy of Science USA, 96: 2907-2912.

Jackson JE (2002). A User's Guide To Principal Components, Wiley Series in Probability and Mathematical Statistics, John Wiley and Sons, New York.

Jolliffe IT, Jones B, Morgan BJ (1980). Cluster analysis of the elderly at home: a case study. Data Anal. Inf., pp. 745–757.

Lapointe J, Li C (2004). Gene Expression Profiling identifies clinically relevant subtypes of prostrate cancer. Proceedings of the National Academy of Sciences USA, 101: 811-816.

Milligan G, Cooper M (1988). A study of variable standardization. Journal of classification, 5: 181–204.

Milligan GW (1980). An Examination of the Effects of Six Types of Error Perturbation on Fifteen Clustering Algorithms. Psychometrika, 45: 325-342.

Perou CM, Jeffrey SS, van de Rijin M, Rees CA, Eisen MB, Ross DT, Pergamenschikov A, Williams CF, Zhu SX, Lee JCF, Lashkari D, Shalon D, Brown PO, Botstein D (1999). Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. Proceedings of the National Academy of Sciences USA, 96: 9212-9217.

Quackenbush J (2001). Computational anaysis of microarray data. Nat. Rev. Gen., 2: 418-427.

Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, van de rijin M, Waltham M, Pergamenschikov A, Lee JCF, Lashkari D, Shalon D, Myers TG, Weinstein JN, Botstein D, Brown PO (2000). Systematic variation in gene expression patterns in human cancer cell lines. Nat. Gen., 24: 227-235.

Tamayo P, Ramaswamy S (2002). Cancer genomics and Molecular

Pattern Recognition Cancer Genomic group,Whitehead Institute/ Massachusetts Institute of Technology, USA.

Valk PJM, Verhaak RGW, Beijen MA, Erpelinck CAJ, Barjesteh van Waal-wijk van Doorn-Khosrovani S, Boer JM, Beverloo HB, Moorhouse MJ, vanderSpek PJ, Lowenberg B and Delwel R (2004). Prognostically useful gene expression profiles in acute myeloid leukemia. New Engl. J. Med., 350: 1617-1628.

Wajid RA, Ali A (1998). Relative Variance Covariance as an alternative to covariance and correlation matrix in Principal Component Analysis. Appl. Stat., 7(1): 45-51.

Yeoung K, Ruzzo W (2001). An empirical study of principal component analysis for clustering gene expression data. Bioinformatics, 17: 763-774.