World Academy of Science, Engineering and Technology
International Journal of Mathematical and Computational Sciences
Vol:7, No:9, 2013

Principal Component Analysis using Singular Value Decomposition of Microarray Data

Dong Hoon Lim

Abstract—A series of microarray experiments produces observations of differential expression for thousands of genes across multiple conditions.

Principal component analysis(PCA) has been widely used in multivariate data analysis to reduce the dimensionality of the data in order to simplify subsequent analysis and allow for summarization of the data in a parsimonious manner. PCA, which can be implemented via a singular value decomposition(SVD), is useful for analysis of microarray data.

For application of PCA using SVD we use the DNA microarray data for the small round blue cell tumors(SRBCT) of childhood by Khan et al.(2001). To decide the number of components which account for sufficient amount of information we draw scree plot. Biplot, a graphic display associated with PCA, reveals important features that exhibit relationship between variables and also the relationship of variables with observations.

Keywords—Principal component analysis, singular value decomposition, microarray data, SRBCT

I. Introduction

After genome sequencing, DNA microarray analysis has become the most widely used functional genomics approach in the bioinformatics field. Biologists are vastly plagued by the enormous amount of unprecedented qualities of genome-wide data produced by the DNA Microarray experiment.

Principal component analysis(PCA) is used to search new abstract orthogonal principal components (eigenvectors) which explain most of the data variation in a new coordinate system. Classical PCA is based on the decomposition of a covariance/correlation matrix (Geladi and Kowalski(1986)) by eigenvalue (spectral) decomposition (EVD) or by the decomposition of real data matrixes using SVD(Wall et al.(2001)). Compared with EVD, SVD is a more robust, reliable, and precise method with no need to compute the input covariance/correlation matrix(Will(1999)). Classical PCA is not useful when the number of variables is larger than the number of observations on each variable. In analysis of microarray data, we encounter this situation commonly when genes are taken as the variables. In such cases it is essential to use PCA using SVD(Deshmukh and Purohit(2007)).

For application of PCA using SVD we use the DNA microarray data for the small round blue cell tumors(SRBCT) of childhood by Khan et al.(2001). To decide the number of components which account for sufficient amount of information we draw scree plot.

From the Biplot for SRBCT data(Bradu and Gabriel(1978)), a graphic display associated with PCA, we can note important features that exhibit relationship between variables and also the relationship of variables with observations.

D.H. Lim is with the Department of Information and Statistics, Gyeongsang National University, Korea, e-mail: dhlim@gnu.ac.kr.

II. Methodolgy

A. PCA

PCA is a multivariate procedure aimed at reducing the dimensionality of multivariate data while accounting for as much of the variation in the original data set as possible. This technique is especially useful when the variables within the data set are highly correlated and when there is a higher than normal ratio of explanatory variables to the number of observation. Principal components seeks to transform the original variable to a new set of variables that are (1) linear combinations of the variables in the data set, (2) uncorrelated with each other, and (3) ordered according to the amount of variation of the original variables that they explain (Everitt and Hothorn(2011)).

The idea behind PCA is to look for a direction (represented as a linear combination $u_1 = \sum_{i=1}^{G} w_i x_i$) that maximizes the variability across the samples. Next, we find a second direction u_2 at right angles to the first that maximizes what remains of the variability. We keep repeating this process. The u_i vectors are the principal components, and we can rewrite each sample vector as a sum of principal components instead of as a sum of separate gene expression values.

PCA can be used as a data reduction method. Changing from the original x-coordinates to the new u-coordinate system doesn't change the underlying structure of the sample vectors. However, it does let us focus on the directions where the data changes most rapidly. If we just use the first two or three principal components, we can produce plots that show us as much of the intrinsic variability in the data as possible.

B. SVD

The result from linear algebra that allows us to compute principal components efficiently is called SVD. This result (whose proof is based on Gram-Schmidt orthogonalization) tells us that any matrix X with n rows and m columns can be decomposed as a product

$$X = UDV^T$$

where U is an $n \times m$ matrix with orthonormal columns ($U^TU = I_m$), while V is an $m \times m$ orthonormal matrix ($V^TV = I_m$), and D is a $m \times m$ diagonal matrix with positive or zero elements, called the singular values.

C. SVD for PCA

We can use SVD to perform PCA. We decompose X using SVD. i.e.

$$X = UDV^T$$

and find that we can write the covariance matrix as

$$C = \frac{1}{n}XX^T = \frac{1}{n}UD^2U^T$$

In this case U is a $n \times m$ matrix. The transformed data can thus be written as

$$Y = \tilde{U}^T UDV^T$$

where $\tilde{U}^T U$ is a simple $n \times m$ matrix which is one on the diagonal and zero everywhere else. To conclude, we can write the transformed data in terms of the SVD decomposition of X.

III. Preprocessing Microarray Data

A. SRBCT Data

We use gene expression data from the microarray experiments of SRBCT of childhood cancer study of Khan et al. (2001).

This data set contains 63 samples with 2308 genes: 23 cases of Ewing sarcoma (EWS), 8 cases of Burkitt lymphoma (BL), 12 cases of neuroblastoma (NB), 20 cases of rhabdomyosarcoma (RMS) samples. The cancer of four types are clinically and histologically similar, yet their response to treatment is likely to be different.

B. Preprocessing

Before applying the PCA using SVD to SRBCT microarray data, we first use the quantile normalization to remove systematic variation in a microarray experiment which affects the measured gene expression levels. We consider top 45 genes, selected using minimum p-value criterion for comparing means of the four groups, NB, RMS, EWS, and BL, using one way analysis of variance for each individual gene.

To identify a set of genes with different profile in these four types of tumors, we applied ANOVA for each gene separately. Table 1 shows the gene number, value of F-statistic and the p-vlaue for selected top 45 genes.

Table 1. Top 45 Genes in SRBCT data

Rank	Gene No.	Statistic	p-value
1	1955	85.67505568	1.769794e-21
2	1389	84.27689813	2.620962e-21
3	1003	73.29717170	6.964706e-20
4	2050	68.40245221	3.409901e-19
5	246	66.12383835	7.361856e-19
6	1954	65.92062246	7.892671e-19
7	1194	58.63860093	1.073035e-17
8	545	54.34722196	5.605686e-17
9	174	52.24560788	1.304586e-16
10	1319	52.09735532	1.385960e-16
.	.	.	.
.	.	.	.
36	2198	30.25935140	5.663193e-12
37	1911	30.18254619	5.923517e-12
38	188	29.82977892	7.288005e-12
39	1799	29.76570223	7.568838e-12
40	554	29.69557301	7.889051e-12
41	338	29.62991322	8.201556e-12
42	819	29.62356134	8.232458e-12
43	603	29.44099551	9.174081e-12
44	1896	28.89288924	1.272934e-11
45	1924	28.86972555	1.290779e-11

IV. Experimental Results and Discussion

For application of PCA we have selected top 45 genes, top according to p-values of F-test. We first treat 63 experiments(samples) as 63 variables, each having 45 observation corresponding to gene expression values of top 45 genes. To decide the number of components which account for sufficient amount of information we draw scree plot. Fig.1 is a plot of the eigenvalues of the experiment components.

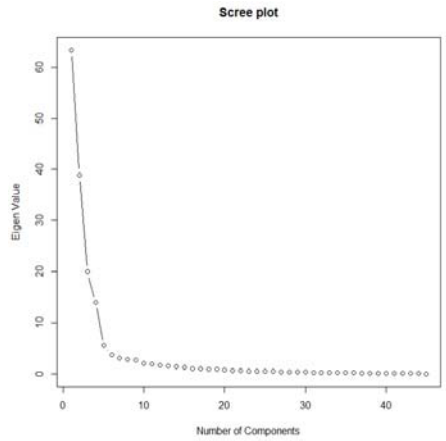

Fig. 1. **Scree plot of eigenvalues of principal experiment components**

Table 2. Importance of principal experiment components

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Standard deviation	7.9618	6.2383	4.4694	3.7291	2.3402
Proportion of Variance	0.3655	0.2244	0.1152	0.0802	0.0316
Cumulative Proportion	0.3655	0.5899	0.7051	0.7853	0.8169

From Table 2, proportion of variations explained by first two PCs, three PCs, four PCs and five PCs are 58.99%, 70.51%, 78.83% and 81.69% of total variation, respectively.

Biplot in Fig. 2 shows these groups differentially expressed in the four groups, NB, RMS, EWS, and BL.

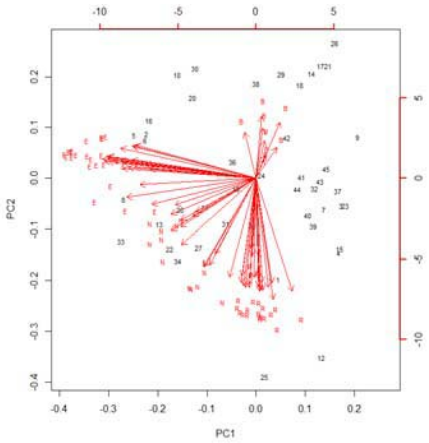

Fig. 2. **Biplot of principal experiment components**

In the following, we present PCA when genes are treated as variables and experiments(samples) are treated as 63 observations on each variables.

From Table 3, proportion of variations explained by first two PCs, three PCs, four PCs and five PCs are 60.15%, 73.41%, 77.18% and 79.80% of total variation, respectively.

Table 3. Importance of principal gene components

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Standard deviation	6.0644	5.0778	3.7129	1.9812	1.6501
Proportion of Variance	0.3536	0.2479	0.1326	0.0377	0.0262
Cumulative Proportion	0.3536	0.6015	0.7341	0.7718	0.7980

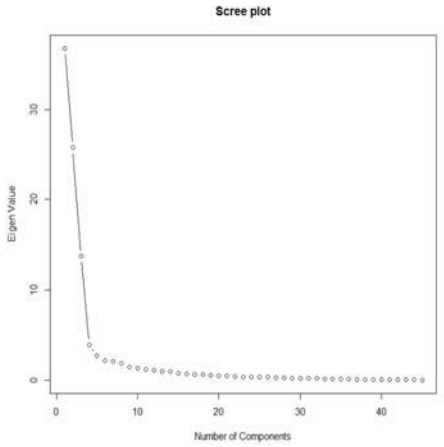

Fig. 3. **Scree plot of eigenvalues of principal gene components**

From the scree plot in Fig. 3, we can still say the fifth point is our "elbow" point.

Biplot in Fig. 4 shows these groups as well as groups of genes differentially expressed.

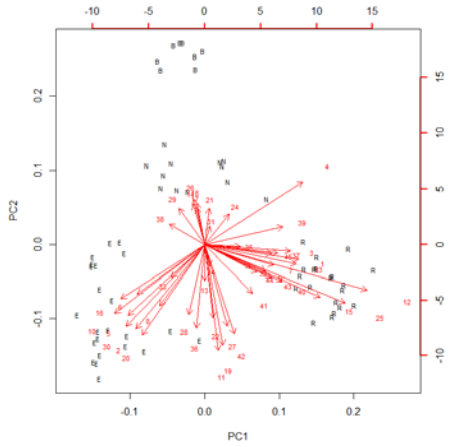

Fig. 4. **Biplot of principal gene components**

ACKNOWLEDGMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (No.2011-0010089)

REFERENCES

[1] Alter, O, Brown, P. O. and Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling PNAS 97: 10101-10106.
[2] Bradu, D., Gabriel K.R.(1978). The biplot as a diagnostic tool for models of two-way tables. Technometrics, 20, 47-68.
[3] Deshmukh, S. R. and Purohit, S. G. (2007)). Microarray Data, Statistical Analysis Usng R, Alpha Science International Ltd.
[4] Everitt, B. and T Hothorn. (2011). An Introduction to Applied Multivariate Analysis with R (Use R!). Springer, New York, NY.
[5] Geladi, Paul; Kowalski, Bruce (1986). "Partial Least Squares Regression:A Tutorial". Analytica Chimica Acta 185: 1?17.
[6] Khan, J. and Wei, J. S. and Ringner, M. and Saal, L. H. and Ladanyi, M. and Westermann, F. and Berthold, F. and Schwab, M. and Antonescu, C. R. and Peterson, C. and Meltzer, P. S. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, Nature Medecine, 7, 673-679.
[7] Wall M.E., Dyck P.A., Brettin T.S.(2001). SVDMAN – singular value decomposition analysis of microarray data. Bioinformatics 17:566-68.
[8] Will, T. (1999). Introduction to the singular value decomposition.