The Korean Communications
in Statistics Vol. 10 No. 3, 2003
pp. 1057-1068

Application of Principal Component Analysis Prior to Cluster Analysis in the Concept of Informative Variables

Seong-San Chae[1]

Abstract

Results of using principal component analysis prior to cluster analysis are compared with results from applying agglomerative clustering algorithm alone. The retrieval ability of the agglomerative clustering algorithm is improved by using principal components prior to cluster analysis in some situations. On the other hand, the loss in retrieval ability for the agglomerative clustering algorithms decreases, as the number of informative variables increases, where the informative variables are the variables that have distinct information(or, necessary information) compared to other variables.

Keywords : Agglomerative Clustering Algorithm; Principal Component Analysis; Informative Variables

1. Introduction

Cluster analysis is concerned with the classification of objects, while principal component techniques assess relationships between variables and may be concerned with the classification of these variables. If a large number of variables are involved, it might be a practice to use principal components with larger eigenvalues to reduce the number of variables prior to cluster analysis.

In principal component analysis, lower eigenvalue factors are considered to be uninformative while larger eigenvalue factors are informative. Chang(1983) investigated the effect of the principal component analysis showing that the principal components with the larger eigenvalues did not necessarily contain more information with a mixture of two normal distributions, saying the success of principal component analysis depends on various factors. Chae(2002) considered the effect of the principal component analysis prior to cluster analysis by comparing the results of discriminant analysis that performed on the clusterings generated

[1] Associate Professor, Department of Information and Statistics, Daejeon University, Daejeon, 300-716, Korea
E-mail : chae@dju.ac.kr

by agglomerative clustering algorithms. In his study, the interpretation of the components might be difficult if the number of principal components exceed a certain minimum number. Fowlkes, Gnanadesikan and Kettenring(1988) mentioned that the inclusion of unnecessary variables in a cluster analysis could cause more damage than in such other statistical procedures as regression analysis.

From this point of view, adequate clustering of a data set requires considerable insight into the relationships among variables. Also, informative cluster analysis of variables requires moderate homogeneity among elements. Unfortunately, little prior knowledge on clustering of either variables or objects is available in the majority of applications submitted for cluster analysis. In most analyses, attention is focused on clustering either data units or variables alone, but not both together. However, the whole question of simultaneous clustering of elements and variables recently received serious study in Tibshirani et al.(1999) and shown application in Perou et al.(1999).

The main objective of this study is to investigate the use of principal component analysis prior to agglomerative clustering algorithms defined on the (β, π)-family discussed by DuBien and Warde(1979, 1987), and Chae and Warde(1991). Also the effect of informative variables is involved with various settings of parameters. The informative variables are the variables that have distinct information(or, necessary information) compared to other variables, but have equal or larger variance in this study. If all the variables are informative, the clusterings generated by clustering algorithms should be more similar to the true structure of data points than those with less informative.

For the purpose of this study, the effect of principal component analysis on variables prior to using agglomerative clustering algorithms is evaluated by Rand's(1971) C statistic. This statistic is a measure of similarity with $0 \le C \le 1$, and there is a perfect agreement within clusters if $C = 1.0$. As an example, the results of using principal components prior to cluster analysis are compared by applying agglomerative clustering algorithms on the cell cycle data that includes identified genes from Spellman et al.(1998) define.

2. Cluster Analysis and Principal Component Analysis

The basic concepts of cluster analysis are the elements to be clustered which are data points, the set of all elements to be clustered which is the object space, and cluster which is an operationally determined collection of data points. Letting N be the number of data points with p variables, then $N \times p$ matrix of measurements, say X, might be

$$X_{(N \times p)} = X^N = [\, X_1 \;\; X_2 \;\cdots\; X_{N-1} \;\; X_N]^T$$

where X_i represents a $p \times 1$ vector on the i-th objects. Thus, X^N indicates that there are

N data points in object space in X. Then a cluster, y_h, is simply a nonempty subset of the object space, and a clustering, $Y = (y_1, y_2, \ldots.., y_k)$, is any partition of the object space. The number of clusters, K, contained in a clustering shall be referred to as the size of the clustering. Some notations useful for understanding a cluster, a clustering, an hierarchy and an agglomerative clustering methods can be found in DuBien and Warde(1987), and Chae and Warde(1991).

For the purpose of this study, the squared Euclidean distance, which is a semi-metric measure of distance, is used for the dissimilarity between data points. The ''true'' structure of the N data points with number of clusters, K, is presented as Y. Then $Y^{[N]}$ is an initial clustering and $Y^{[N,K]}$ is a certain type of rearrangement of a initial clustering with K number of clusters. Let Y' denote a clustering that result from applying an agglomerative clustering algorithm to the N data points with number of clusters, K. Then $C(Y, Y')$ is a measure of the ''retrieval'' ability of the agglomerative clustering algorithm to the true structure for K.

Letting d_{ij} denote the joining distance between cluster y_i and cluster y_j in clustering $Y^{[N,K]}$, where $y_i, y_j \in Y^{[N,K]}$, $K=1,2,\cdots,N$. Then $y_{(ij)} = y_i \cup y_j$ will denote the new cluster within clustering $Y^{[N,K-1]}$. It should be noted that the joining distance, d_{ij}, is always the smallest distance remaining in the set of all distances between clusters in clustering $Y^{[N,K]}$.

For any clustering $Y^{[N,K]}$ in the hierarchy, if the distances d_{ij}, d_{ik}, and d_{jk} between pairs of clusters y_i, y_j, and y_k are obtained recursively from clustering $Y^{[N,K+1]}$, $K \langle N$, then the distance between the new cluster $y_{(ij)}$ and any other cluster $y_k \in Y^{[N,K]}$ can be computed from the following formula:

$$d_{(ij)k} = \frac{1-\beta+2\pi}{2} d_{jk} + \frac{1-\beta-2\pi}{2} d_{ik} + \beta d_{ij}$$

where $d_{ij} \langle d_{ik} \langle d_{jk}$. This formula represents a two parameter (β, π)-family of agglomerative clustering algorithms derived by DuBien and Warde(1979) by placing a suitable set of constraints on the parameters originally given in Lance and William's equation(1966, 1967).

Nine agglomerative clustering algorithms are chosen from the (β, π)-family of agglomerative clustering algorithms based on the rationale discussed by DuBien and Warde(1987). The (β, π) values that define these algorithms are as follows:

(1) $\beta = \ 0.0$ *with* $\pi = \ -0.5, \ 0.0, \ 0.5$;
(2) $\beta = -0.25$ *with* $\pi = \ -0.25, \ 0.0, \ 0.5$;
(3) $\beta = -0.5$ *with* $\pi = \ 0.0, \ 0.25, \ 0.75$;

In the $(\beta, \ \pi)$-family, (0.0, -0.5) is known as single linkage; (0.0, 0.0) as average linkage; (0.0, 0.5) as complete linkage; (-0.25, 0.0) or (-0.5, 0.0) as representations of the flexible strategy; (-0.5, 0.75) is the recommendation by DuBien and Warde(1987).

Hence the results of using principal component analysis prior to applying nine agglomerative clustering algorithms are investigated with various settings on the parameters.

In this study, a principal component analysis of the correlation matrix instead of covariance matrix is applied since the sample correlation matrix is invariant under scale changes. Let the correlation matrix be identical to the covariance matrix Σ. Then the $j-th$ eigenvalue of Σ is distinct, and the corresponding normalized eigenvector e_j is uniquely defined. General information to understand on principal component analysis may be found in Johnson and Wichern(1982)

3. Design of Simulation Study

Some of the possible structural parameters considered in this study are defined as follows:

1. N, the number of data points in X;
2. p, the number of variables;
3. n_k, the split or the size of the $k-th$ cluster generated from each population;
4. δ, the distance between mean vectors;
5. Σ, the covariance matrix.

For convenience, $N=60$, $p=9$, and $k=3$ in this study. Then a brief summary of data structure may be outlined as follows:

$$\underline{X}_{gi} \sim N_p(\underline{\mu}_g, \ \Sigma_g)$$

where $g=1,2,3$, $i=1,2,\cdots,60$ with split into $k=3$ populations of $(n_1;n_2;n_3)=(20;20;20)$, μ_g, $g=1,\ldots,k$ is constrained by an equilateral triangle spatial configuration,

$$\begin{aligned}
\underline{\mu_1}' &= (0.0 \ c_1\delta_c \ c_1\delta_c \ c_2\delta_c \ 0.0 \ c_2\delta_c \ c_3\delta_c \ c_3\delta_c \ 0.0)' \\
\underline{\mu_2}' &= (c_1\delta_c \ 0.0 \ c_1\delta_c \ c_2\delta_c \ c_2\delta_c \ 0.0 \ 0.0 \ c_3\delta_c \ c_3\delta_c)' \\
\underline{\mu_3}' &= (c_1\delta_c \ c_1\delta_c \ 0.0 \ 0.0 \ c_2\delta_c \ c_2\delta_c \ c_3\delta_c \ 0.0 \ c_3\delta_c)'
\end{aligned}$$

where $(c_1; c_2; c_3) = [(1;0;0), (1;1;0), (1;1;1)]$, then the squared Euclidean distances between mean vectors are $\delta = \delta_c \times \sqrt{2.0 \times \sum_{i=1}^{3} c_i^2}$ depending on the settings of population mean vectors, where $\delta = 2.0,\ 4.0$. In these cases, the distance among mean vectors is the same regardless of the settings on the population mean vectors and the numbers of informative variables for each $\delta = 2.0,\ 4.0$. For this reason, δ_c should be changed depending on the value of $\sum_{i=1}^{3} c_i^2$. If $\sum_{i=1}^{3} c_i^2$ is large, δ_c would be small. Among the settings of $(c_1; c_2; c_3)$ in this study, the setting $(1;1;1;)$ on nine variables is the most informative, while $(1;0;0)$ is the least informative. Here the covariance matrix is

$$\Sigma_g = \Sigma = \begin{pmatrix} A & B & B \\ & A & B \\ & & A \end{pmatrix}, \quad A = \begin{pmatrix} 1.0 & \rho & \rho \\ & 1.0 & \rho \\ & & 1.0 \end{pmatrix}, \quad B = \begin{pmatrix} \eta & \eta & \eta \\ & \eta & \eta \\ & & \eta \end{pmatrix}$$

where $\rho = .6, .9$ and $\eta = .0, .2, .4$.

With this structural settings on parameters, it was possible to produce each of the results from clustering algorithms applied to original data and data obtained from principal component analysis.

Let Y' and Y'' denote clusterings that result from applying an agglomerative clustering algorithm to the N original data points and data obtained from principal component analysis with number of clusters, K, respectively. For each setting of $(\delta, \rho, \eta, (c_1; c_2; c_3))$, the values of $C(Y, Y')$ and $C(Y, Y'')$ for the nine (β, π) clustering algorithms were generated by following steps:

1. An object space $X_{N \times p}$ of data points was generated;
2. The squared Euclidean distance between each pair of data points in X was computed and stored in lower triangular matrix order by rows as the vector D_1;
3. Principal component analysis was applied to X, and three principal components ($q = 3$) were chosen if their eigenvalues were greater than or equal to one and sum of percentages was greater than 70 percent of the variance;
4. The squared Euclidean distance between each pair of data points from X using three principal components was computed and stored in lower triangular matrix order by rows as the vector D_2;
5. Each of the nine clustering algorithms was applied to D_1 and D_2 to produce two different clusterings, Y' and Y'';

6. For each of the clusterings, Y' and Y'', from above steps, $C(Y,Y')$ and $C(Y,Y'')$ were calculated for the nine clustering algorithms.

For each setting of the $(\delta, \rho, \eta, (c_1;c_2;c_3))$, the above sequence of steps was replicated 100 times and the sample means, $\overline{C}$, were computed. Consequently, $\overline{C}$ has been obtained for each setting of the $(\delta, \rho, \eta, (c_1;c_2;c_3))$ on the data set to quantify the ''retrieval'' ability for the nine agglomerative clustering algorithms alone and for the agglomerative clustering algorithms after applying principal component analysis.

4. Results from Simulation

All results from comparative study given as $\overline{C}$ computed over 100 replications are discussed in terms of changes in the $(\delta, \rho, \eta, (c_1;c_2;c_3))$ and changes in (β, π) which defines the agglomerative clustering algorithms. In the form of $\overline{C}(Y,Y')$ and $\overline{C}(Y,Y'')$, the retrieval ability of clustering algorithms are represented for original data and using principal component analysis, respectively. Although the nine clustering algorithms were studied, only the results from single linkage, average linkage, complete linkage, two representations of the flexible strategy and (-.5, .75) with two settings of $(c_1;c_2;c_3)=\{(1;0;0),(1;1;1)\}$ are summarized. The results on the use of the other clustering algorithms with settings of $(c_1;c_2;c_3)$ followed the same trend as shown in those tables presented.

As shown in tables 1-2, the difference in trends of recovery based on $\overline{C}(Y,Y')$ and $\overline{C}(Y,Y'')$ is mainly due to the settings of $(c_1;c_2;c_3)$ and the correlation structures designed into the original data. For the setting with $(c_1;c_2;c_3)=(1;0;0)$, the recovery decreases as η increases for the original data, while the recovery increases or stays in stable for the use of principal components(see table 1). However, for the setting with $(c_1;c_2;c_3)=(1;1;1)$, the recovery decreases as η increases for both original data and the data from the use of principal components(see table 2). At this point of view, it might be verified that $(c_1;c_2;c_3)=(1;1;1)$ is the more informative case and $(c_1;c_2;c_3)=(1;0;0)$ is the less informative case for the original set of data generated since the recovery is degraded with the inclusion of unnecessary variables. On our primary study, standardizations on the data with the less informative variables were performed, however, a little improvement on the recovery was found in the use of principal component analysis prior to cluster analysis. This implies that the use of principal component analysis prior to applying agglomerative clustering algorithms is appropriate for data with informative variables.

Based on the results in tables 1-2, the effect of informative variables on the ability of the six agglomerative clustering algorithms is discussed. The $\overline{C}$ values show that, there is an essential difference between $\overline{C}(Y, Y')$ and $\overline{C}(Y, Y'')$ for the less informative case $(c_1; c_2; c_3) = (1;0;0)$. The difference is due to the use of principal component analysis prior to applying the agglomerative clustering algorithm. Using principal component analysis prior to clustering algorithm gives a significant effect on ''retrieval'' ability. On the other hand, the results for the agglomerative clustering algorithms with or without applying principal component analysis show essentially no differences for the case, $(c_1; c_2; c_3) = (1;1;1)$, which are considered to be more informative.

Table 1. The $\overline{C}(Y, Y')$ and $\overline{C}(Y, Y'')$ for $(c_1; c_2; c_3) = (1;0;0)$

Data	δ	ρ	0.6			0.9		
		$(\beta, \pi)/\eta$	0.0	0.2	0.4	0.0	0.2	0.4
Original	2.0	(0.0, -0.5)	.3478	.3471	.3483	.3483	.3481	.3478
		(0.0, 0.0)	.4824	.4945	.4856	.4959	.4919	.4915
		(0.0, 0.5)	.5331	.5381	.5332	.5336	.5328	.5292
		(-0.25, 0.0)	.5517	.5549	.5448	.5585	.5543	.5495
		(-0.5, 0.0)	.5661	.5679	.5683	.5705	.5615	.5627
		(-0.5, 0.75)	.5718	.5693	.5711	.5682	.5649	.5614
	4.0	(0.0, -0.5)	.3772	.3756	.4058	.6386	.5946	.5676
		(0.0, 0.0)	.6822	.6991	.6749	.6746	.6353	.6149
		(0.0, 0.5)	.7001	.6913	.6417	.6265	.6186	.5961
		(-0.25, 0.0)	.9175	.8616	.7810	.9274	.8643	.7705
		(-0.5, 0.0)	.9346	.8839	.8138	.9227	.8731	.7895
		(-0.5, 0.75)	.8446	.7927	.7650	.7986	.7484	.7056
PCA	2.0	(0.0, -0.5)	.3486	.3498	.3496	.3476	.3494	.3509
		(0.0, 0.0)	.4798	.4971	.4866	.4892	.4869	.4968
		(0.0, 0.5)	.5242	.5208	.5276	.5307	.5300	.5266
		(-0.25, 0.0)	.5467	.5492	.5428	.5489	.5428	.5469
		(-0.5, 0.0)	.5559	.5572	.5597	.5535	.5508	.5529
		(-0.5, 0.75)	.5550	.5565	.5648	.5567	.5538	.5503
	4.0	(0.0, -0.5)	.3677	.3744	.4165	.3965	.3881	.4022
		(0.0, 0.0)	.6086	.6297	.6304	.5763	.5604	.5586
		(0.0, 0.5)	.6456	.6493	.6366	.5856	.5886	.5888
		(-0.25, 0.0)	.7476	.7301	.7443	.6716	.6739	.6733
		(-0.5, 0.0)	.7550	.7475	.7626	.6895	.6824	.6774
		(-0.5, 0.75)	.7371	.7199	.7432	.6598	.6634	.6645

The loss of information in using only those principal components with relatively large eigenvalues may not be a loss in fact, but the elimination of less informative components, as shown in table 2 compared to table 1. Large loss is found in the case of the least informative, (1;0;0), while small loss is shown in the case of the most informative, (1;1;1), with large differences among mean vectors($\delta = 4.0$). The use of principal component analysis prior to applying the clustering algorithm is desirable when variables are more informative, since possibilities of getting high recovery increase as the number of informative variables increases(table 2).

Under the design described previously, more similar clusterings are retrieved than applying agglomerative clustering algorithms alone when principal component analysis is applied prior to using clustering algorithms. If many unnecessary variables are included in applying

Table 2. The $\overline{C}(Y, Y')$ and $\overline{C}(Y, Y'')$ for $(c_1; c_2; c_3) = (1;1;1)$

Data	δ	ρ	0.6			0.9		
		$(\beta,\ \pi)/\ \eta$	0.0	0.2	0.4	0.0	0.2	0.4
Original	2.0	(0.0, -0.5)	.3475	.3471	.3486	.3478	.3486	.3477
		(0.0, 0.0)	.4792	.4736	.4950	.4988	.4856	.4914
		(0.0, 0.5)	.5327	.5325	.5286	.5353	.5395	.5363
		(-0.25, 0.0)	.5559	.5556	.5547	.5515	.5541	.5453
		(-0.5, 0.0)	.5662	.5662	.5725	.5678	.5625	.5614
		(-0.5, 0.75)	.5698	.5702	.5713	.5698	.5690	.5697
	4.0	(0.0, -0.5)	.4164	.3646	.3769	.7177	.4210	.3645
		(0.0, 0.0)	.7909	.7250	.6992	.6771	.6201	.5944
		(0.0, 0.5)	.7676	.6858	.6697	.6295	.6302	.6062
		(-0.25, 0.0)	.9519	.8991	.8017	.9165	.8303	.7504
		(-0.5, 0.0)	.9589	.9162	.8440	.9272	.8629	.7846
		(-0.5, 0.75)	.8940	.8364	.8055	.8003	.7596	.7152
PCA	2.0	(0.0, -0.5)	.3488	.3499	.3511	.3482	.3494	.3505
		(0.0, 0.0)	.4922	.4865	.5158	.4877	.4986	.4989
		(0.0, 0.5)	.5295	.5274	.5532	.5283	.5326	.5292
		(-0.25, 0.0)	.5539	.5496	.5566	.5485	.5441	.5447
		(-0.5, 0.0)	.5572	.5604	.5698	.5541	.5530	.5558
		(-0.5, 0.75)	.5578	.5605	.5717	.5514	.5519	.5547
	4.0	(0.0, -0.5)	.6564	.5440	.5606	.5294	.4901	.4406
		(0.0, 0.0)	.8044	.7103	.6790	.6802	.6389	.6203
		(0.0, 0.5)	.7787	.6926	.6686	.6604	.6481	.6407
		(-0.25, 0.0)	.9401	.8872	.8029	.8345	.7936	.7547
		(-0.5, 0.0)	.9540	.9202	.8419	.8386	.8324	.7767
		(-0.5, 0.75)	.9188	.8797	.7967	.7985	.7872	.7694

principal component analysis, the clusterings generated by clustering algorithms with principal components are more damaged than other cases, as mentioned by Fowlkes, Gnanadesikan and Kettenring(1988). However, applying principal component analysis before using clustering algorithm is not worse than applying clustering algorithm alone in this study, if the variables are considered to be informative and the loss of information in dimensional reduction by principal components is considered.

5. Application to Real Data

An application using a set of data that includes yeast (*Saccharomyces cerevisiae)* genes by Spellman et al.(1998). The primary data set might be obtained at *http://cellcycle-www.stanford.edu.* In their normalization procedure on the primary data, a total of 800 yeast genes are identified as being periodically regulated and meet an objective minimum criterion for cell cycle regulation.

For convenience, 630 observations with 24 variables(i. e., the results of a series of timepoints in the experiments) are taken out of identified 800 genes that have no missing values on the data set with five(somewhat arbitrary) clusters. These clusters approximate the commonly used cell groups in the literature.

For each of genes, the sizes of clusters which it belongs are (102-159-82-231-56) for S/G2, G2/M, M/G1, G1 and S, according to Spellman et al.(1998) and the set of identified genes may provide a natural basis for organizing yeast gene expression data. Then the clusters are identified by the six agglomerative clustering algorithms alone and by using principal component analysis prior to agglomerative clustering algorithms where the squared Euclidean distance is used as a similarity measure between objects. In fact, the use of Pearson's correlation coefficient as a similarity measure between objects might be considered with the average linkage-(0.0, 0.0) applied by Spellman et. al,(1998), however, only the squared Euclidean distance is used with respect to the use of principal component analysis prior ·to cluster analysis in this study.

As shown in table 3, the recovery of true cluster is increased or decreased by using principal components depending on the number of principal components and the choice of agglomerative clustering algorithm.

Table 3. The sizes of clusters and C values for Spellman's data(1998)

Data	Group	S/G2	G2/M	M/G1	G1	S	C	*
	(β, π)\Sizes	102	159	82	231	56	values	
Original ($p=24$)	(0.0, -0.5)	1	2	1	625	1	.2552	232
	(0.0, 0.0)	2	362	5	258	3	.6561	359
	(0.0, 0.5)	143	129	28	326	4	.7178	409
	(-0.25, 0.0)	37	186	46	287	74	.7036	341
	(-0.5, 0.0)	100	157	82	271	20	.7248	369
	(-0.5, 0.75)	210	146	57	112	105	.7133	296
PCA ($q=4$) (76.6%)	(0.0, -0.5)	1	2	1	625	1	.2557	233
	(0.0, 0.0)	15	2	78	532	3	.3845	184
	(0.0, 0.5)	3	257	12	206	152	.7023	308
	(-0.25, 0.0)	118	175	124	183	30	.7548	359
	(-0.5, 0.0)	52	185	92	254	47	.7183	360
	(-0.5, 0.75)	110	176	162	135	47	.7583	368
PCA ($q=5$) (80.8%)	(0.0, -0.5)	1	2	1	625	1	.2567	233
	(0.0, 0.0)	1	471	49	106	3	.4666	260
	(0.0, 0.5)	201	176	15	234	4	.7398	392
	(-0.25, 0.0)	169	196	66	149	50	.7424	355
	(-0.5, 0.0)	52	182	123	212	61	.7212	310
	(-0.5, 0.75)	144	96	74	247	69	.7887	390

*: number of objects which are assigned to the "identified" clusters defined by Spellman et al.(1998)

As a result, the use of flexible strategy, (-.25, 0.0), and the recommendation of DuBien and Warde(1981), (-0.5, 0.75) might be recommended instead of using the other clustering algorithms if reduction of variables(i.e., genes are synchronized in series of timepoints in the experiment) is considered. In the concept of Rand's(1971) C statistic which is the measure of similarity between clusterings, the application of those two agglomerative clustering algorithms recovers well the arbitrary divided clusters that each of genes belongs to on the cell cycle data from Spellman et al.(1998) if the clusters formed by clustering algorithms are meaningful.

6. Concluding Remarks

In this study, the use of principal component analysis prior to cluster analysis has been investigated. The retrieval abilities of agglomerative clustering algorithms were increased if the number of informative variables were increased. Moreover, the retrieval ability of the known clustering is improved in some situations by using principal components prior to cluster analysis.

On the other hand, the recovery of clusterings generated by agglomerative clustering algorithms with principal components were greatly degraded if many unnecessary variables were included in applying principal component analysis. However, applying principal component analysis prior to using clustering algorithm were not worse than applying clustering algorithm alone if the variables were considered to be informative and the loss of information on reducing the number of variables is considered. The loss in retrieval ability for the six agglomerative clustering algorithms decreases, as the number of informative variables increases.

According to the results from the cell cycle data of Spellman et al.(1998), it is found that the recovery of true cluster(in fact, somewhat arbitrary divided) is generally increased or rarely decreased by using principal components prior to cluster analysis. The results depend on the number of principal components that is related to the structures of variables and the choice of agglomerative clustering algorithm. Based on Rand's(1971) C, the recovery of known clustering was improved in the use of (-0.5, 0.75), the recommendation of DuBien and Warde(1981) and flexible strategy, (-.25, 0.0), if reduction of the variables(i.e., reduction of timepoints in the experiments) were considered on the cell cycle data from Spellman et al.(1998). In specific, the retrieval ability of other clustering algorithms except for the single linkage is better than the result from using the average linkage-(0.0, 0.0).

Therefore, it might be appropriate to reduce the number of variables into informative variables by applying principal component analysis before performing cluster analysis on a data sample with a large number of variables if the characteristics of the data were critically examined and the clusters generated by clustering algorithms are meaningful. The choice of agglomerative clustering algorithm and the number of principal components should be considered depending on the structure of data treated since the clusters formed by clustering algorithms are data dependent.

References

[1] Chae, Seong S. and Warde, W. D. (1991). A Method to Predict the Number of Clusters, *Journal of the Korean Statistical Society*, 20, 162-176.

[2] Chae, Seong S. (2002). Results of Discriminant Analysis with Respect to Cluster Analysis Under Dimensional Reduction, *The Korean Communication in Statistics*, 9, 543-554.

[3] Chang, Wei-Chien (1983). On using Principal Components before Separating a Mixture of Two Multivariate Normal Distributions, *Applied Statistics*, 32, 3, 267-275.

[4] DuBien, J. L., and Warde, W. D. (1979). A Mathematical Comparison of the Members of an Infinite Family of Agglomerative Clustering Algorithms, *The Canadian Journal of Statistics*, 7, 29-38.

[5] DuBien, J. L., and Warde, W. D. (1987). A Comparison of Agglomerative Clustering Methods with respect to Noise, *Communications in Statistics, Theory and Method*, 16, 1433-1460.

[6] Fowlkes, E. B., Gnanadesikan, R., and Kettenring, J. P. (1988). Variable Selection in Clustering, *Journal of Classification*, 5, 205-228.

[7] Johnson, R. A. and Wichern, D. W. (1982). Applied Multivariate Statistical Analysis, Prentice-Hall Inc.

[8] Lance, G. N., and Williams, W. T. (1966). A Generalized Sorting Strategy for Computer Classification, *Nature*, 212-218.

[9] Lance, G. N., and Williams, W. T. (1967). A General Theory of Classificatory Sorting Strategies, 1. Hierarchical Systems. *The Computer Journal*, 9, 373-380.

[10] Perou, C. M., Jeffrey, S. S., Rijn, M. V., Rees, C. A., Eisen, M. B., Ross, D. T., Pergamenschikov, A., Williams, C. R., Zhu, S. X., Lee, J. C. F., Lashkari, D., Shalon, D., Brown, P. O., and Botstein, D. (1999). Distinctive gene expression patterns in human mammary epithelial cells and breast cancers, *Proceeding of National Academy Sciences in United States of America*, 96, 9212-9217.

[11] Rand, W. M. (1971). Objective Criteria for the Evaluation of Clustering Methods, *Journal of the American Statistical Association*, 66, 846-850.

[12] Spellman P. T., Sherlock, G., Zhang, M. Q., Iyer V. R., Eisen M. B., Brown, P. O., Botstein, D. and Futcher B. (1998). Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization, *Molecular Biology of the Cell*, 9, 3273-3297.

[13] Tibshirani, R., Hastie, T., Eisen, M., Ross, G., Botstein, D., and Brown, P. (1999). Clustering Methods for the analysis of DNA microarray data, Technical Report, Department of Statistics, Stanford University.

[Received September 2003, Accepted December 2003]