The Korean Communications
in Statistics Vol. 12 No. 3, 2005
pp. 683-696

Cluster Analysis Using Principal Coordinates for Binary Data

Seong San Chae[1] and Jeong il Kim[2]

Abstract

The results of using principal coordinates prior to cluster analysis are investigated on the samples from multiple binary outcomes. The retrieval ability of the known clustering algorithm is significantly improved by using principal coordinates instead of using the distance directly transformed from four association coefficients for multiple binary variables.

Keywords : Agglomerative Clustering Algorithm, Principal Coordinates, Association Coefficients

1. Introduction

Principal component analysis in an $N\times m$ data matrix and principal coordinate analysis in an $N\times N$ symmetric matrix composed of Euclidean inter-point distances are included to interpret the distance between the i-th and j-th objects of N samples. A reduction in number of axes is required which is normally effected by a principal component analysis since there is no known method of generating a $N\times N$ positive semi-definite matrix that is always available, as mentioned by Gower (1966, 1971), Gower and Legendre (1986). With certain criteria, principal component analysis on the $m\times m$ covariance (or correlation) matrix and principal coordinates analysis on the $N\times N$ symmetric matrix (formed with distances between objects) are defined as being dual to one another when they both lead to set of N data points with the same inter-point distances. Chae and Warde (2005) showed that the use of principal coordinates was more effective than the use of principal components on retrieval of clusters for multivariate continuous variables.

In recent, the use of multiple binary data has occurred in the field of psychology, biology, genetics, and clinical trials. Asparoukhov and Kranowski (2001) compared thirteen discriminant procedures by applying them to five real sets of binary data from clinical trials. Lee (2005) studied discriminant analysis of binary data with multinomial distribution by simulation study in discriminant analysis. In cluster analysis, Huang (1998) focused on the technical issues of extending K-means algorithm to cluster data

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

[2] Corresponding Author, Professor, Department of Information and Statistics, Daejeon University, Daejeon 300-716 Korea.
E-mail address : jikim@dju.ac.kr

with categorical values, while Ordonez (2003) compared three variants of the K-means algorithm to cluster binary data streams.

The main objective of this study is to investigate the use of principal coordinates into the data from multiple binary outcomes prior to cluster analysis. The similarity between the i-th and j-th objects is calculated depending on different association coefficients with various settings of parameters. Then distance is constructed using the formula, $d_{ij} = \sqrt{1-\gamma_{ij}}$, where γ_{ij} is the similarity between the i-th and j-th objects depending on different association coefficients. The matrix formed with elements d_{ij} is used for applying principal coordinate analysis prior to cluster analysis.

Rand's (1971) C statistic is used to compare the retrieval abilities and the agreements of clustering algorithms based on how they partition the object space. The mean and variance of Rand's C statistics for given K are found in DuBien, Warde and Chae (2004). It evaluates the results of cluster analysis based on how they partition the data points in the concept of reproducibility. And the C is a measure of similarity with $0 \leq C \leq 1$. When the partition produced by clustering algorithm is identical to the structure within data treated, the C is 1. The results of applying principal coordinates analysis prior to the use of agglomerative clustering algorithms are examined and compared.

2. Agglomerative Clustering Algorithms

Suppose a sample of size N is observed with m variables on each data point. The $N \times m$ matrix of measurements, say X, might be $X_{(N \times m)} = X^N = [\ X_1\ X_2 \cdots X_{N-1}\ X_N]^T$ where X_i represents a $m \times 1$ vector of measurement on the i-th objects. 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, if the following three conditions hold:

(1) For every $y_h \in Y$, $y_h \not\in \emptyset$;

(2) If $y_{h,}\ y_l \in Y$ and $y_h \neq y_l$, then $y_h \cap y_l = \emptyset$;

(3) $\cup_{h=1}^{K} y_h = X$.

Some notations useful for understanding a cluster, a clustering, an hierarchy and an agglomerative clustering method can be found in DuBien and Warde (1987).

Let Y represent the "true" structure of the N data points with number of clusters, K, and $Y^{[N,K]}$ be a certain type of rearrangement of Y with K 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 Rand's (1971) $C(Y, Y')$ is a measure of the "retrieval" ability of the agglomerative clustering algorithm to the true structure for 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 in Chae and Warde (2005). In the (β, π)-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) and (-0.5, 0.0) as representations of the flexible strategy; (-0.5, 0.75) as recommendation by DuBien and Warde (1987).

3. Principal Coordinate Analysis

Suppose an $N\times N$ symmetric matrix is conformed with elements of Euclidean distance from the $N\times m$ data matrix X, where d_{ij}, $i=1,2,\ldots,N-1$, $j=i+1,\ldots,N$ and $d_{ii}=0$. The coordinates by using an $N\times N$ symmetric matrix, say F, could be found of N data points that generates the distances between the i-th and j-th objects, which is termed as the inter-point distances. By using the inter-point distances from F, a cluster analysis might be applied to establish groups of objects assigning objects to the same group when their coefficients in F satisfy certain criteria. Then a representation of the multivariate sample in a small number of dimension (in two or three dimensions) which reflect the inter-object distances, might be constructed by recognizing the metric nature of the $N\times N$ symmetric matrix F.

Principal coordinate analysis involves projecting the points onto a space defined by their small number of principal axes and the distances between the i-th and j-th objects can be approximated. When all the distances between the i-th and j-th objects of N samples are known, their coordinates axes are found by investigating a set of conditions for a solution to exist in real Euclidean space. However, there is no known method of generating a positive semi-definite $N\times N$ symmetric matrix with elements d_{ij}, $i=1,2,\ldots,N-1, j=i+1,\ldots,N$ to represent the similarity between the i-th and j-th objects. Thus a reduction in the number of axes is required which is normally effected by a principal component analysis.

Consider the matrix XX', that is required for a principal coordinate analysis, where X is the standardized $N\times m$ matrix and the matrix $X'X$ that is required for a principal component analysis. Suppose the matrix $X'X$ has an eigenvalue λ and corresponding eigenvector v, and the matrix XX' has an eigenvalue η and corresponding eigenvector u. From the relationships between $X'Xv=\lambda v$ and $XX'u=\eta u$, $XX'(Xv)=\lambda(Xv)$, so that $\lambda=\eta$ and $Xv=ku$, where k is a constant relating the scaling of the two sets of eigenvectors. If the eigenvectors v are normalized as $v'v=1$, then $k^2=u'u=v'X'Xv=\lambda v'v=\lambda$. If the eigenvectors u are normalized, then $k=1$ and $u=Xv$. Based on this relationship between two sets of eigenvectors, the principal coordinate analysis operating on the XX' is a dual of principal component analysis on the $X'X$. For more details on the duality of principal coordinates and components, refer to Gower (1966).

Let a_{ij} be the elements of the $N\times N$ matrix XX'. Then inter-coordinate distances between i-th and j-th objects are presented by

$$a_{ii}+a_{jj}-2a_{ij}=\sum_{r=1}^{m}(x_{ir}-x_{jr})^2=d_{ij}^2\ .$$

Without loss of generality, this is related to the $N\times N$ symmetric matrix F with elements, d_{ij}, $i=1,2,\ldots,N-1$, $j=i+1,\ldots,N$ and $d_{ii}=0$, which are the Euclidean distance between i -th and j-th objects. If the $N\times N$ matrix F is positive semi-definite, the principal coordinate analysis operating on the F is a dual of the principal component analysis on the $X'X$. Then it is possible to compute principal coordinates of any Euclidean distance matrix without being in possession of either the original data matrix or a variance-covariance matrix of the characters of the data points.

This method is also applicable to non-Euclidean distances and association coefficients, thus a method of principal coordinate is more powerful than ordinary principal components analysis to ensure identification on groups of objects.

In using principal coordinates, it is not necessary that the coordinates have any valid interpretation since principal coordinates analysis has no associated method for including information on the variables that is unlike the special case of principal component analysis. This technique of using principal coordinates is very useful in the treatment of similarity or association coefficients, since the distances between the i-th and j-th rows of principal coordinates are convenient ways of representing the inter-relationships between objects.

4. Association Coefficients

An association coefficient is a pair-function that measures the agreement between pairs of observations over an array of two-state or multi-state characters. Many of these coefficients measure the numbers of agreement as compared with the number of theoretically possible ones. Characters coded in two or a few states are especially suitable for the computation of association coefficients, although even continuous characters can be coded to yield association coefficients.

In common, association coefficients are computed with two-state characters, which are for convenience coded 0 or 1. The coded 0, 1 represent the presence or absence of characteristic or property. When character states are compared over pairs of rows in a data matrix, the outcome can be summarized in a 2×2 frequency table as shown in table 1.

<Table 1> 2×2 frequency table

Data points	j			
	Code	1	0	sum
i	1	a	b	$a+b$
	0	c	d	$c+d$
	sum	$a+c$	$b+d$	m

The number of characters coded 1 in both data points is written in the left upper quadrant of the table, while the number of characters coded 0 in both data points is written in the right lower quadrant. The other two quadrants are the number of characters in which the data-points disagree, being coded 1 for the j-th data point and 0 for the i-th data point (or the converse). The marginal totals are the sums of these frequencies, with m representing for the sum of the four frequencies ($m=a+b+c+d$), which equals to the number of characters in the study. It is convenient to define w as the number of matches or agreements ($w=a+d$), and u as the number of mismatches ($u=b+c$), where $m=w+u$.

In this study, four association coefficients are used: the simple matching coefficient, the Jaccard coefficient, the Yule coefficient and the product moment correlation coefficient.

The simple matching coefficient is defined as

$$S_{SM}=\frac{w}{w+u}=\frac{a+d}{a+b+c+d}.$$

From the formula, it follows that $S_{SM}\rightarrow 0$ as $\frac{w}{u}\rightarrow 0$, and that $S_{SM}\rightarrow 1$ as $\frac{w}{u}\rightarrow 1$. In its complementary form, $1-S_{SM}$, the simple matching coefficient is equal to the Euclidean distance based on unstandardized character states, which can take the value of 1 or 0; that is $\sqrt{1-S_{SM}}=d$.

The Jaccard coefficient is defined as

$$S_J=\frac{a}{a+u}=\frac{a}{a+b+c}.$$

It is clear that $S_J\rightarrow 0$ as $\frac{a}{u}\rightarrow 0$, and that $S_J\rightarrow 1$ as $u\rightarrow 0$. The Jaccard coefficient omits consideration of negative matches. Whether negative matches should be incorporated into a coefficient of association may occur in serious doubt. It may be argued that basing similarity between two objects on the mutual absence of a certain character is improper. The Jaccard coefficient is appropriate when negative matches are to be excluded.

The Yule coefficient is defined as

$$S_Y = \frac{ad - bc}{ad + bc}.$$

Its numerator is the determinator of the 2×2 table and the limits of S_J are from -1 to $+1$. There are no matches at all with -1, and are perfect matches with $+1$.

The product moment correlation coefficient is defined as

$$S_P = \frac{ad - bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}.$$

It is clear that $S_P \rightarrow -1$ as $ad \rightarrow 0$, and that $S_P \rightarrow +1$ as $bc \rightarrow 0$.

Among four coefficients above mentioned, Yule coefficient does not satisfy the metric inequality according to Gower (1971), Gower and Legendre (1986). The $N \times N$ symmetric matrix F might not be positive semi-definite in theory. However, the matrix F constructed with Yule coefficient was examined by eigenvalues at each step of simulation, and found that it was, in practice, non-negative definite with designed structural settings of parameters given in the next section.

5. Design of Simulation Study

5.1 Fundamental Concepts with Basic Definitions

Currently, computer programs which generate ''multiple binary'' data treat X_i as a multiple binary random variable are not available. One is not able to randomly generate a multiple observation in which each variable is an outcome of a Bernoulli trial. At this point, a few definitions is offered.

Definition 1. A Bernoulli trial is an experiment which has two possible outcomes, generally called success and failure. The sample space for a Bernoulli trial will in general be written $S=\{0,1\}$, where 0 indicates ''failure" and 1 indicates ''success".

Definition 2. Let V be the total number of successes in N repeated independent Bernoulli trials with probability p of success on a given trial. V is called the binomial random variable with parameters N and p.

Definition 3. A multinomial trial, with parameters $p_1, p_2, \ldots, p_k$, is a trial which results in one of k possible outcomes (outcomes are called classes). The probability of the i-th class occurring on a single trial is p_i, $i=1,2,\ldots,k$; thus $0 \le p_i \le 1$, $p_i + q_i = 1$, for

$i = 1,2,\ldots,k$, and $\sum_{i=1}^{k} p_i = 1$.

Clearly a multinomial is simply a generalization of a binomial trial, having an arbitrary k possible outcomes rather than just two possible outcomes.

Definition 4. Given an experiment which consists of N repeated, independent, multinomial trials with parameters p_i, $i = 1,2,\ldots,k$, let X_i be the number of trials which result in outcomes in the i-th class, $i = 1,2,\ldots,k$. $X_1, X_2, \ldots, X_k$ is called the multinomial random variable with parameters $(N, p_1, p_2, \ldots, p_k)$.

In the data point represented by the $1 \times m$ vector, X_i, where $X_i = (x_{i1} x_{i2} \ldots x_{im})$, it is desired that each component, x_{ij} be the result of the i-th Bernoulli or multinomial trial for the j-th characteristic.

5.2 Multi-state Coding

The several states in qualitative multi-state characters cannot necessary be arrayed in some obvious order but still refer to a unit character on logical grounds. These characters are therefore often called unordered multi-state characters. An example would be alternative color patterns of a given structure. One way of coding these is to use a separate symbol for each state; an example is given in table 2.

<Table 2> Example of coding for each state

Color Structure	State
Red	0
Yellow	1
Blue	2

A match is scored if the same symbol occurs in two data points; otherwise, a mismatch is recorded. Another method is to convert the qualitative multi-state character into several new characters. The characters might then be coded as shown in the following table 3.

This is not an easy task in as much as the recoding has to be done in such a way that a positive score on one of the new characters does not automatically bring about negative scores on all other such characters derived from the same qualitative character. In practice, it is commonly found that most qualitative multi-state characters can be converted into several independent characters if a little thought is given to the problem. By the method of additive coding, the multiple character states are coded as shown in table 4.

<Table 3> Example of coding for two-state

Color structure	Multi-state characters	Two-state characters
		1 2 3
Red	0	1 0 0
Yellow	1	0 1 0
Blue	2	0 0 1

<Table 4> Example of additive coding for two-state

Data point	Multi-state character	Two-state character
		1 2 3
a	State 0 (character undetectable)	0 0 0
b	State 1 (weak positive)	1 0 0
c	State 2 (moderate positive)	1 1 0
d	State 3 (strong positive)	1 1 1

In this way a multi-state character i of m states is turned into $m_i - 1$ two-state characters. The scoring is termed additive because the state 3, for instance, is expressed as the sum of the effects of the two-state characters 1, 2, and 3. In any of this methods of coding multi-state characters, two-state of binary data are produced. The procedure for binomial data is then applied to the binary codes of the multi-state data.

Since each of these coding methods transforms multinomial data to binomial data, it is sufficient for now to look at the effects of this procedure on binary outcomes.

5.3 Design of Simulation Study

Currently, computer programs which generate "multiple binary" data treat X_i as a multiple binary random variable are not available. One is not able to randomly generate a multiple observation in which each variable is an outcome of a Bernoulli trial. There is no correlation structure associated with the generation. It is necessary to impose a given correlation structure in order that principal coordinate analysis may be incorporated into the test design.

A set of multiple binary samples was generated from a multivariate normal random variable, Z, with the desired correlation matrix R with mean vector 0 and with following possible structural parameters.

(a) N, the number of data points in Z;

(b) m, the number of variables;

(c) n_k, the size of the k-th cluster generated from each population;

(d) p_k, the probability assigned for the k-th cluster generated;

(e) R, the correlation matrix.

For convenience, $N=60$, $m=9, k=3$, $(n_1;n_2;n_3)=\{(20;20;20),(15;20;25)\}$, and the correlation matrix R is of the form,

$$R=\begin{pmatrix} A & B & B \\ B & A & B \\ B & B & A \end{pmatrix}, \quad A=\begin{pmatrix} 1.0 & \rho & \rho \\ \rho & 1.0 & \rho \\ \rho & \rho & 1.0 \end{pmatrix},$$

and B is a matrix of all elements are 0.0, where $\rho=.6, .9$. This structure on the correlation matrix was chosen in order to produce three principal coordinates that their eigenvalues were greater than or equal to 1.0 and the sum of their eigenvalues was greater than 80% of the variance. In the simulation, three principal coordinates were chosen to ensure identification of objects. More than three principal coordinates up to nine coordinates ($rank(F)=9$) were considered, however, there were no significant differences on the retrieval ability of clustering algorithms.

For the multiple binary variables, each variable was transformed to a Bernoulli random variable with parameter p_k by translating the normal z value for each variate to "1" if $P_r(Z \leq z) \leq p_k$ and to "0" if $P_r(Z \leq z) > p_k$, $k=1,2,3$. A set of probability parameters (p_1, p_2, p_3) was used to separate the three clusters, where $\sum_{k=1}^{3} p_k = 1.0$.

Four different association coefficients, S_{SM}, S_J, S_Y and S_P, represented by γ, and three set of probability parameters $(p_1, p_2, p_3)=\{(.14, .33, .53), (.24, .33, .43), (.235, .33, .335)\}$ were studied. Thus, a variable structural parameter $\{\rho, (n_1;n_2;n_3), (p_1, p_2, p_3), \gamma\}$ was defined. Finally, a multiple binary data, X, with three clusters was generated from a multivariate normal random variable with the desired correlation matrix R.

For each setting of $\{\rho, (n_1;n_2;n_3), (p_1, p_2, p_3), \gamma\}$, the Rand's(1971) C values that representing the recovery of true structure for the six clustering algorithms were obtained by following steps:

(1) An object space $X_{N\times m}$ of data points was generated from $Z_{N\times m}$;

(2) The distance converted from association coefficient using the formula $d_{ij}=\sqrt{1-\gamma_{ij}}$, where γ_{ij} is the similarity 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) The $N\times N$ symmetric matrix with elements, d_{ij} in the vector D_1, was stored in F;

(4) A set of necessary and sufficient conditions that F formed with D_1 is positive semi- definite was examined with acceptable tolerance;

(5) Calculate the eigenvalues and corresponding eigenvectors of F, then the Euclidean distance between the i-th and j-th representing inter-point distances with three principal coordinates was computed

and stored in lower triangular matrix by rows as the vector D_2;

(6) Each of the six clustering algorithms was applied to D_1 and D_2 to produce two different clusterings, Y' and Y'';

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

For each setting of the $\{\rho, (n_1; n_2; n_3), (p_1, p_2, p_3), \gamma\}$, the above sequence of steps was replicated 100 times and the sample mean of the C statistic, $\overline{C}$, was computed for each of the six agglomerative clustering algorithms for each of the two comparisons.

Consequently, $\overline{C}$ result from 100 replications to quantify the ''retrieval'' ability for each of the agglomerative clustering algorithms alone and after applying principal coordinate analysis based on different association coefficients has been applied to the data from multiple binary outcomes for each setting of the $\{\rho, (n_1; n_2; n_3), (p_1, p_2, p_3), \gamma\}$.

6. Results from Simulation

Based on the data from each setting of the various structural parameters $\{\rho, (n_1; n_2; n_3), (p_1, p_2, p_3), \gamma\}$, all results from the comparative study will be discussed with agglomerative clustering algorithms defined with (β, π) and association coefficients. The results from the simulation study are not independent of the fixed structural parameters. The results based on different $(n_1; n_2; n_3)$ will not be discussed since the retrieval abilities of clustering algorithms with four association coefficients were not significantly different depending on our previous simulation study. The results of using the Yule coefficient might be doubtable with applying clustering algorithms in the sense of non-metric, however, the retrieval results were presented with four association coefficients.

In Tables 5-6, the results are given as $\overline{C}$ computed over 100 replications for each setting of the various structural parameters $\{\rho, (p_1, p_2, p_3), \gamma\}$ and for each of the six agglomerative clustering algorithms mentioned above. Table 5 represents the results for retrieval ability on the data based on different (p_1, p_2, p_3) with four association coefficients in the form of $\overline{C}(Y, Y')$. Table 6 represents the results from applying principal coordinate analysis based on the $N \times N$ symmetric matrices constructed with four association coefficients in $\overline{C}(Y, Y'')$. The distance from each association coefficient is calculated with $d_{ij} = \sqrt{1 - \gamma_{ij}}$, where γ_{ij} is the similarity between the i-th object and j -th object.

Table 5 shows that the trends of recovery for the six clustering algorithms are not quite different on changing ρ. The difference in trends of recovery is mainly due to the association coefficients and different (p_1, p_2, p_3) designed into the original data. The recovery level decreases as ρ increases for four association coefficients with complete linkage, $(.0, .5)$, while the trends of recovery for other

<Table 5> The $\overline{C}(Y, Y')$ from applying clustering algorithms with D_1

	ρ	.6				.9			
(p_1, p_2, p_3)	$(\beta, \pi)/\gamma$	SM	J	Y	P	SM	J	Y	P
(.14,.33,.53)	(.0, −.5)	.4710	.4710	.3458	.3458	.5120	.5195	.3461	.3461
	(.0, .0)	.5559	.5160	.3458	.3458	.5487	.5610	.3461	.3461
	(.0, .5)	.5768	.4546	.3650	.3648	.5632	.4538	.3491	.3485
	(−.25, .0)	.5988	.5313	.5617	.5603	.5903	.5794	.5746	.5645
	(−.5, .0)	.5994	.5544	.5642	.5642	.6002	.5819	.5755	.5629
	(−.5, .75)	.5777	.5606	.5344	.5551	.5767	.5736	.5739	.5668
(.24,.33,.43)	(.0, −.5)	.4274	.4205	.3458	.3458	.4591	.4649	.3461	.3461
	(.0, .0)	.5172	.4817	.3458	.3458	.5285	.5154	.3461	.3461
	(.0, .5)	.5389	.4286	.3610	.3574	.5349	.4264	.3499	.3470
	(−.25, .0)	.5593	.4912	.5500	.5502	.5588	.5418	.5537	.5536
	(−.5, .0)	.5581	.5204	.5527	.5531	.5636	.5404	.5557	.5522
	(−.5, .75)	.5373	.5168	.5085	.5341	.5252	.5147	.5427	.5308
(.285,.33,.385)	(.0, −.5)	.4156	.4030	.3458	.3458	.4373	.4449	.3462	.3462
	(.0, .0)	.5111	.4564	.3458	.3458	.5169	.4963	.3462	.3462
	(.0, .5)	.5370	.4219	.3570	.3547	.5319	.4158	.3483	.3469
	(−.25, .0)	.5492	.4826	.5493	.5505	.5490	.5256	.5492	.5524
	(−.5, .0)	.5491	.5085	.5531	.5502	.5535	.5239	.5520	.5521
	(−.5, .75)	.5270	.4995	.5037	.5299	.5078	.5093	.5314	.5155

<Table 6> The $\overline{C}(Y, Y'')$ from applying clustering algorithms with D_2

	ρ	.6				.9			
(p_1, p_2, p_3)	$(\beta, \pi)/\gamma$	SM	J	Y	P	SM	J	Y	P
(.14,.33,.53)	(.0, −.5)	.9344	.8012	.9824	.9934	.6709	.6284	.8330	.8538
	(.0, .0)	.8623	.7565	.9659	.9686	.7578	.7492	.8273	.8171
	(.0, .5)	.7822	.7635	.7732	.7792	.7522	.7656	.7648	.7634
	(−.25, .0)	.9080	.8364	.9426	.9402	.7700	.7731	.8453	.8364
	(−.5, .0)	.8981	.8649	.9028	.9143	.7814	.7746	.8101	.8248
	(−.5, .75)	.7707	.7832	.7820	.7871	.7071	.7416	.7504	.7620
(.24,.33,.43)	(.0, −.5)	.9513	.7717	.9913	.9978	.6667	.5465	.8658	.8946
	(.0, .0)	.8731	.7871	.9614	.9595	.7511	.7338	.8222	.8202
	(.0, .5)	.7833	.7717	.7597	.7616	.7423	.7491	.7535	.7509
	(−.25, .0)	.8683	.8360	.9258	.9398	.7488	.7414	.8294	.8342
	(−.5, .0)	.8728	.8757	.9091	.9236	.7592	.7379	.7995	.8021
	(−.5, .75)	.7505	.7814	.7711	.7616	.6913	.7111	.7343	.7515
(.285,.33,.385)	(.0, −.5)	.9598	.8039	.9979	.9999	.7073	.5541	.8627	.8941
	(.0, .0)	.8564	.8117	.9456	.9414	.7529	.7378	.8158	.7993
	(.0, .5)	.7644	.7677	.7656	.7652	.7452	.7444	.7454	.7486
	(−.25, .0)	.8683	.8636	.9205	.9205	.7489	.7363	.8249	.8339
	(−.5, .0)	.8570	.8957	.9072	.9072	.7506	.7406	.7958	.8167
	(−.5, .75)	.7669	.8124	.7682	.7682	.7062	.7277	.7377	.7335

algorithms are difficult to explain.

Table 6 displays the results from applying principal coordinates prior to the six agglomerative clustering algorithms for $\{\rho, (p_1, p_2, p_3), \gamma\}$. The $\overline{C}$ values calculated show an essential difference compare to the results presented in table 5, implying that the use of principal coordinates prior to applying the clustering algorithm has a significant effect on the recovery of the true clustering.

Principal components analysis and principal coordinates analysis are defined as being dual to one another when they both lead to set of N points with the same inter-object distances, as previously mentioned. Even if principal coordinate analysis is not an associated method for including information on the variables, principal coordinates could be used to ensure the identification of objects from multiple binary outcomes.

With the design described, more similar clusterings are retrieved when principal coordinate analysis is applied prior to cluster analysis. The values of $\overline{C}$ show essential differences depending on the association coefficients with the clustering algorithms. This implies that the use of different coefficient prior to applying cluster analysis has a significant effect on the recovery of the clusters.

7. Application

The use of principal coordinate analysis prior to applying the agglomerative clustering algorithm on the financial performance data (Affi and Clark, 1990). For convenience, the 25 companies with 7 variables are used as the data set with three clusters that identified by different kinds. 7 variables are, ROR5 (percent rate of return on total capital), D/E (Debt-to equity ratio for the past year), SALESGR5 (Percent annual compound growth rate of sales), EPS5 (Percent annual compound growth in earnings per share), NPM1 (Percent net profit margin), P/E (Price-to-earning ratio), and PAYOUTR1 (Annual dividend divided by the 12-months earnings per share). These variables are transformed to binary variables "1" if the values of variables are large than the medians of each variables; "0", otherwise. Then binary values on ROR5 and PAYOUTR1 are only reversed to make it easy to identify companies.

For each of companies, the sizes of clusters to which it belongs are (14-5-6) in the Chemical, Health, and Supermarket with 25 companies. For this data, 3 principal coordinates are assumed reasonably since those principal coordinates include more than 81 percent of information depending on association coefficients. Hence the clusters are identified by the six agglomerative clustering algorithms using 3 principal coordinates on the data with different association coefficients. Results of applying principal coordinates prior to applying the clustering algorithms are examined and compared with the results from directly applying the six clustering algorithms with 7 variables.

As shown in table 7, the recovery of the ''company defined clusters'' is increased by using principal coordinates prior to clustering algorithms instead of directly applying clustering algorithms. The choices of clustering algorithms and association coefficients depend on the characteristics of the data. However, the use of Yule index with principal coordinates prior to applying two flexible strategies, $(-.25, .0)$ and $(-.5, .0)$, is recommended to find the better defined clusters on the data from Affi and Clark (1990).

<Table 7> The C values from applying clustering algorithms on financial performance data (Affi and Clark, 1990) for diversified companies

	$C(Y, Y')$				$C(Y, Y'')$			
$(\beta, \pi)/\gamma$	SM	J	Y	P	SM	J	Y	P
(.0, -.5)	.3800	.4967	.4967	.4967	.5400	.7067	.6067	.6067
(.0, .0)	.4867	.5267	.5067	.5133	.6700	.6200	.5867	.5267
(.0, .5)	.5200	.5167	.4800	.5267	.6700	.6700	.5667	.5267
(-.25, .0)	.5067	.5200	.5233	.5367	.6033	.6167	.7733	.6333
(-.5, .0)	.5067	.5133	.5200	.5267	.6333	.6167	.7733	.6333
(-.5, .75)	.5000	.5533	.5267	.5333	.6333	.6333	.5967	.6400

8. Concluding Remarks

Applying principal coordinate analysis that is defined as being dual to principal component analysis prior to cluster analysis has been investigated in this study. Principal coordinate analysis can be applied to the multiple binary outcomes, while principal component analysis cannot. In using principal coordinates, each object is uniquely identified even though principal coordinates do not include information on the variables.

The six agglomerative clustering algorithms with four association coefficients were examined and compared on various settings of structural parameters in the simulation study. The retrieval abilities of clustering algorithms were different in trend of recovery of the true clustering depending on settings of structural parameters. However, the use of principal coordinates instead of using direct dissimilarity converted from similarity prior to applying the clustering algorithm had a significant effect on the recovery of the true clustering.

As expected, the recovery levels were increased by using principal coordinates for the data from Affi and Clark (1990). In particular, Yule index with principal coordinates prior to applying two flexible strategies, $(-.25, .0)$ and $(-.5, .0)$, was recommended to find the better defined clusters. The results of using the Yule coefficient might be doubtable with applying clustering algorithms in the sense of non-metric as mentioned by Gower (1971), Gower and Legendre (1986). However, the matrix F constructed with Yule coefficient was examined by eigenvalues at each step of simulation, and found that it was, in practice, non-negative definite for generated data in simulation and real data in application.

Using principal coordinates prior to cluster analysis, the retrieval ability of the known clustering algorithms with four association coefficients was significantly improved instead of directly applying clustering algorithms. However, one would better choose an association coefficient and a clustering algorithm depending on the characteristic of data in analysis.

References

[1] Affi, A.A. and Clark, V.(1990). *Computer-Aided Multivariate Analysis*, Van Nostrand Reinhold Company, New York.

[2] Asparoukhov, O.K. and Krzanowski, W.J.(2001). A comparison of discriminant procedures for binary variables, *Computational Statistics & Data Analysis*, Vol. 38, 139-160.

[3] Chae, S.S. and Warde, W.D.(1991). A method to predict the number of clusters, *Journal of the Korean Statistical Society*, Vol. 20, 162-176.

[4] Chae, S.S. and Warde, W.D.(2006). Effect of using principal coordinates and principal components on retrieval of clusters, *Computational Statistics & Data Analysis*, Vol. 50, 1407-1417.

[5] 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*, Vol. 7, 29-38.

[6] DuBien, J.L. and Warde, W.D.(1987). A comparison of agglomerative clustering methods with respect to noise, *Communications in Statistics, Theory and Method*, Vol. 16, 1433-1460.

[7] DuBien, J.L., Warde, W.D. and Chae, S.S.(2004). Moments of Rand's C statistic in cluster analysis, *Statistics & Probability Letters*, Vol. 69, 243-252.

[8] Gower, J.C.(1966). Some distance properties of latent root and vector methods used in multivariate analysis, *Biometrika*, Vol. 53, 325-338.

[9] Gower, J.C.(1971). A general coefficient of similarity and some of its properties, *Biometrics*, Vol. 27, 857-871.

[10] Gower, J.C. and Legendre, P.(1986). Metric and Euclidean properties of dissimilarity coefficients, *Journal of Classification*, Vol. 3, 5-48.

[11] Huang, Z.(1998). Extensions to the k-means algorithms for clustering large data sets with categorical values, *Data mining and Knowledge Discovery*, Vol. 2, 283-304.

[12] Lee, J.J.(2005). Discriminant analysis of binary data with multinomial distribution by using the iterative cross entropy minimization estimation, *The Korean Communications in Statistics*, Vol. 12, 125-137.

[13] Ordonez, C.(2003). Clustering binary data streams with K-means, *In 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery.*

[14] Rand, W.M.(1971). Objective criteria for the evaluation of clustering methods, *Journal of the American Statistical Association*, Vol. 66, 846-850.

[Received May 2005, Accepted October 2005]