Journal of the Korean
Statistical Society
Vol. 25, No. 4, 1996

Quantification Plots for Several Sets of Variables [†]

Mira Park [1] and Myung-Hoe Huh [2]

Abstract

Geometric approach to extend the classical two-set theory of canonical correlation analysis to three or more sets is considered. It provides statistical graphs to represent the data in a low dimensional space. Procedures are developed for computing the canonical variables and the corresponding properties are investigated. The solution is equivalent to that of the usual problem in the case of two sets. Goodness-of-fit of the proposed plots is studied and a numerical example is included.

Key Words : Generalized canonical correlation analysis; Row quantification; Column quantification; Biplot.

1. INTRODUCTION

There are many practical situations to deal with multiple sets of variables. For example, we may have several sets of comparable form of tests

[†] This paper is partially supported by Institute of Statistics, Korea University.

[1] Postdoctoral researcher, Institute of Statistics, Korea University, Seoul 136-701, Korea.

[2] Professor, Department of Statistics, Korea University, Seoul 136-701, Korea.

which are obtained from a group of students and have to determine the similarity between the sets of tests. Canonical correlation analysis, developed by Hotelling(1936), is a classical technique for studying the relationship between two sets of variables. The aim of canonical correlation analysis is to find out linear combinations that have the maximal correlation. The concepts and technique of canonical correlation analysis have been extended to the case of more than two sets of variables. Horst(1961) developed a generalized canonical correlation procedure which maximizes the sum of correlations among linear composites. Carroll(1968) defined the canonical correlation problem in terms of finding an auxiliary variables and linear combinations of the variables, with the aim of maximizing the sum of squared correlations between them. Kettenring(1971) constructed the general principal component model and provided a unifying discussion of the several extensions of the classical two-set theory. Discussions to generalized canonical correlation procedure have been continued by Gower(1989) and Lafosse(1989).

In spite of its long history, canonical correlation analysis has yielded few useful applications. One major reason may be its difficulty in terms of interpretation. Most of research concentrated on the theoretical approach, and the main purpose was the description of the association measure rather than individual scaling of variables and/or subjects. In this article, we will consider the quantification problem of the rows(observations) and columns(variables) when we have several sets of variables.

Quantification methods are data-analytic methods, which have been widely used in Japan for more than thirty years, as descriptive methods to analyze qualitative data (Tanaka et al., 1994). Their essential parts were established by C. Hayashi since 1948 (see Hayashi, 1988). These methods could be expanded to continuous variables, though they were originally developed to assign numerical values or scores to qualitative data. In this case, we will assign appropriate scores to observations and variables so that the specific purpose of an analysis is achieved. In this point of view, we will study the geometric approach to quantification for generalized canonical correlation analysis and propose the statistical graphs to represent the data. It can be regarded as an extension of Gabriel's(1971) biplot, a graphical tool for the case of only one set of variables, to the case of several sets of variables.

2. GEOMETRIC APPROACH TO QUANTIFICATION

Let the data matrix X with n observations(rows) and p variables (columns) be partitioned with $n \times p_i$ submatrices X_i , $i = 1, \cdots, m$, $(p_1+p_2+\cdots+p_m = p)$. Thus $X = [X_1|X_2|\cdots|X_m] : n \times p$. We assume that columns are centered and standardized unless mentioned otherwise, and that $rank(X_i) = p_i$.

The columns of X can be positioned as p points in the Euclidean space R^n. Let S_i be the linear subspace of R^n that are generated by columns of X_i. Hence, S_i consists of vectors of the form $X_i a_i$ for $a_i \in R^{p_i}$.

Now, the generalized canonical analysis can be presented as a problem of finding a_i' s in their respective subspaces so that m points $X_1 a_1, \cdots, X_m a_m$ are closely located as nearest as possible. It can be formulated as

$$\min \sum_{i=1}^{m}\sum_{j=1}^{m} ||X_i a_i - X_j a_j||^2. \tag{2.1}$$

where $||C||^2$ is defined as $tr(C'C)$ for an arbitrary matrix C. Since the above quantity comes to zero by taking a_i's equal to zero, we need some constraint to construct a valid optimization problem.

Note that (2.1) can be rewritten as

$$2m(\sum_{i=1}^{m} a_i' X_i' X_i a_i) - 2\sum_{i=1}^{m}\sum_{\substack{j=1 \\ j\neq i}}^{m} a_i' X_i' X_j a_j. \tag{2.2}$$

The criterion can be obtained by fixing the first term while maximizing the second term of (2.2) rather than taking the overall minimum of (2.1). The problem, therefore, is to choose a_i s so as to maximize

$$||Xa||^2 = \sum_{i=1}^{m}\sum_{j=1}^{m} a_i' X_i' X_j a_j \tag{2.3}$$

under the constraint

$$\sum_{i=1}^{m} a_i' X_i' X_i a_i = c\ (constant), \tag{2.4}$$

where $a' = [a_1'|\cdots|a_m']$. Essentially, this optimization problem is not affected by the values of c.

Now, we can solve this problem using Lagrange multiplier λ. Define

$$L = \sum_{i=1}^{m}\sum_{j=1}^{m} a_i' X_i' X_j a_j - \lambda(\sum_{i=1}^{m} a_i' X_i' X_i a_i - c)$$

Setting the partial derivatives equal to zero yields the following system of equations:

$$X_i' X^{(i)} a^{(i)} = \lambda X_i' X_i a_i \tag{2.5}$$

where

$$X^{(i)} = [X_1|X_2|\cdots|X_{i-1}|X_{i+1}|\cdots|X_m] : n \times (p - p_i),$$
$$a^{(i)} = [a_1|a_2|\cdots|a_{i-1}|a_{i+1}|\cdots|a_m] : 1 \times (p - p_i).$$

By adding $X_i'X_ia_i$ to both sides of (2.5), we have

$$X'Xa = (1+\lambda)Da \tag{2.6}$$

where a is $p \times 1$ coefficient vector and D is a block diagonal matrix with $X_i'X_i$ as its i-th block. Pre-multiplying $D^{-1/2}$ to both sides of (2.6) yields

$$D^{-1/2}(X'X)D^{-1/2}D^{1/2}a = (1+\lambda)D^{1/2}a. \tag{2.7}$$

Thus $D^{1/2}a/\sqrt{c}$ is an eigenvector of $D^{-1/2}(X'X)D^{-1/2}$. For convenience, we may set c equal to $m(n-1)$. Premultiplying a_i' and summing over i to both sides of (2.5), we obtain

$$\lambda = \frac{1}{m}\sum_{i=1}^{m} Cov(X_ia_i, \sum_{\substack{j=1\\ j\neq i}}^{m} X_ja_j).$$

It means that λ is an average of the covariances between the i-th canonical variate X_ia_i and the remaining canonical variates. Since λ is expressed in covariance terms, it is not enough to measure the degree of association. We may use average correlations between m sets canonical variables. Let us define it as follows

$$\bar{\rho} = \sum_{i=1}^{m}\sum_{\substack{j=1\\ j\neq i}}^{m} Corr(X_ia_i, X_ja_j)/m(m-1).$$

Now, consider the quantification problem of the rows(observations). The main philosophy is to assign scores to individual observations of one set using information from remaining $m-1$ sets. Note that (2.5) may be written as

$$a_i = (1/\lambda)(X_i'X_i)^{-1}X_i'X^{(i)}a^{(i)}.$$

By pre-multiplying λX_i to both sides , we obtain

$$\lambda X_ia_i = X_i(X_i'X_i)^{-1}X_i'X^{(i)}a^{(i)}.$$

It tells us that X_ia_i (up to the scale factor λ) is a projection of $X^{(i)}a^{(i)}$ on S_i. In other words, the linear combination of $m-1$ sets of variables $X^{(i)}a^{(i)}$ has a projection X_ia_i on the subspace S_i generated by the columns of X_i. And, thus, $\lambda X_i a_i$ contains quantification of rows belonging to the X_i matrix, with external reference to the prespecified scaling vector $X^{(i)}a^{(i)}$.

To quantify the columns(variables) of the data matrix, consider a supplementary problem which is originally introduced by French researchers such as Lebart(1984). In this case, we consider the artificial supplementary data $X_{S_i} = I_{p_i}$ for the i-th set of variables. The elements of the s-th row of X_{S_i} are zeros except for the s-th element, which is 1. Since all the variables are centered and standardized, the s-th supplementary observation represents the single role of the s-th variable in the X_i dataset. Therefore, quantification of the columns of X_i is given by $\lambda X_{S_i} a_i$ or λa_i.

This unidimensional quantification procedure can be extended to multidimensional quantification. Consider the problem of obtaining higher-stage canonical variables. To assure a new relationship among sets, it is necessary to add restrictions at each stage. We consider the following restriction at the k-th stage:

$$\sum_{i=1}^{m} Cov(X_i a_{i(l)}, X_i a_{i(k)}) = a'_{(l)} D a_{(k)}/(n-1) = 0, \quad l = 1, \cdots, k-1. \quad (2.8)$$

where $a_{i(k)}$ is the k-th stage canonical coefficients vector for the i-th set of variables and $a_{(k)} = (a_{1(k)}, \cdots, a_{i(k)}, \cdots, a_{m(k)})$. Then the coefficient vector $a_{(k)}$ can be attained from eigenvectors of $D^{-1/2}X'XD^{-1/2}$ corresponding to the k-th largest eigenvalue. It follows that the restriction (2.8) yields

$$Cov(Xa_{(l)}, Xa_{(k)}) = \sum_{i=1}^{m}\sum_{j=1}^{m} a'_{i(l)} X'_i X_j a_{j(k)}/(n-1) = 0, \quad l = 1, \cdots, k-1.$$

It means that the k-th stage canonical score, $Xa_{(k)}$, is uncorrelated with the lower-stage canonical score, $Xa_{(l)}, (l = 1, \cdots, k-1)$.

From the above results, the solution for the r-dimensional quantification can be attained from eigensystem (2.7) relative to the largest r eigenvalues. Table 1 summarizes r-dimensional quantification formulas for rows and columns where

$$A_{i(r)} = (a_{i(2.1)}, \cdots, a_{i(r)}) \quad and \quad \Lambda_{(r)} = diag(\lambda_1, \lambda_2, \cdots, \lambda_r)$$

for $r \leq q(= min(p_1, p_2, \cdots, p_m))$. The r-dimensional quantification plots can be obtained by plotting the first r columns of $X_i A_{i(r)} \Lambda_{(r)}$ for row plot and $A_{i(r)}\Lambda_{(r)}$ for column plot, and we get the biplot of generalized canonical correlation analysis by combining row and column plots in each set.

Table 1 Quantification Formulas for Data matrix X_i

Rows	$X_i A_{i(r)} \Lambda_{(r)}$
Columns	$A_{i(r)} \Lambda_{(r)}$

3. PROPERTIES OF QUANTIFICATION METHOD

By considering the geometric meaning of the similarity, we maximized the sum of covariances between linear composites fixing the sum of variances, consequently. The properties of the canonical variables can be summarized as follows : for $k, l = 1, \cdots, min(p_i), k \neq l$,

$$\sum_{i=1}^{m} Var(X_i a_{i(k)}) = m,$$

$$Var(\sum_{i=1}^{m} X_i a_{i(k)}) = m(1 + \lambda_k),$$

$$\sum_{i=1}^{m} Cov(X_i a_{i(k)}, X_i a_{i(l)}) = 0,$$

$$Cov(\sum_{i=1}^{m} X_i a_{i(k)}, \sum_{i=1}^{m} X_i a_{i(l)}) = 0.$$

Here, $a_{i(k)}$ is the k-th stage canonical coefficient vector for the i-th variable set and λ_k is the average covariance at the k-th stage.

When the number of sets is only two, the proposed method reduces to Hotelling's classical procedure. Note that the only difference between above optimization problem and the standard one is that the latter uses two individual norming constraints while the former uses a single overall norming constraint. However, it can be shown that the two multipliers coincide when $m = 2$. On the other hand, the additional restriction to obtain higher-stage canonical variables is also reduced to the standard restriction. See Park(1995) for proofs. Thus, it follows that, for $l = 1, \cdots, k-1$,

$$Cov(X_i a_{i(2.1)}, X_i a_{i(2)}) = 0, \quad i = 1, 2,$$

and

$$Cov(X_1 a_{1(l)}, X_2 a_{2(k)}) = 0.$$

Under the nonsingular transformation, average maximum covariance, λ , between the i-th canonical variable $X_i a_i$ and the sum of remaining canonical variables $X^{(i)}a^{(i)}$ is invariant. In addition, it leads

$$XT(T^{-1}a) = Xa,$$

and, moreover,

$$X_i T_i(T_i^{-1}a_i) = X_i a_i, \quad (i = 1, \ldots, m)$$

where T is $p \times p$ block diagonal matrix with its i-th block as $p_i \times p_i$ nonsingular matrix T_i. The proof is given in Park(1995). Therefore, quantification results of the rows are invariant under the nonsingular transformation while those of columns do not keep this property.

Consider the interpretation of the quantification plots. Note that the coordinates of row plots and column plots are the canonical scores and coefficients scaled by square root of canonical correlation, respectively. Thus the projected length of each column {row} points along the axis indicates the importance of the variable {observation} in each set. And, since the coordinate of row plots are weighted average of the column coordinates, we can observe the relative position of the individual with respect to the variables by superimposing the row and column plot for each variable set. Also, by superimposing the column plots of all sets, we can see what the canonical relationship means. These properties will be examined in detail with a numerical example.

Now, we consider the goodness of lower dimensional approximation offered by quantification plots. For row quantification plot of data matrix X, we may define a goodness-of-approximation by

$$GOA_{(r)} \; for \; X_i = 1 - \|X_i A_i \Lambda - X_i(A_{i(r)}\Lambda_{(r)} : 0_{p_i \times (q-r)})\|^2 / \|X_i A_i \Lambda\|^2$$

For column quantification plot of data matrix X, we may define

$$GOA_{(r)} \; for \; X_i = 1 - \|A_i \Lambda - (A_{i(r)}\Lambda_{(r)} : 0_{p_i \times (q-r)})\|^2_{X_i' X_i} / \|A_i \Lambda\|^2_{X_i' X_i}$$

where $\|C\|^2$ is defined as $tr(C'C)$ and $\|C\|^2_M$ is defined as $tr(C'MC)$. The reason for using norming matrix $X_i' X_i$ is that column quantifications are lacking absolute uniqueness under nonsingular transformation of X.

Also, we may define the "explanatory power indices(EPI)" which are similar to the coefficient of determination R^2 in linear regression:

$$EPI_{(r)} \;\; for \;\; X^{(i)} \; by \; X_i = \|X_i A_{i(r)}\Lambda_{(r)}\|^2 / \|X^{(i)}A^{(i)}\|^2.$$

4. A NUMERICAL EXAMPLE

To illustrate the proposed methodology, we use "depression" data from Afifi(1984). We select seven negative-affect items, four positive-affect items and seven somatic and retarded activity items. The list of the items is given in Table 2. Each item is a statement to which the response categories are ordinal. The values of the response categories are reversed for the positive-affect items.

Table 2. List of the Depression Item

Negative Affect

$x1$. I felt that I could not shake off the blues even with the help of my family or friends.

$x2$. I felt depress.

$x3$. I felt lonely.

$x4$. I had crying spells.

$x5$. I felt sad.

$x6$. I felt fearful.

$x7$. I thought my life had been failure.

Positive Affect

$x8$. I felt that I was as good as other people.

$x9$. I felt hopeful about future.

$x10$. I was happy.

$x11$. I enjoyed life.

Somatic and Retarded Activity

$x12$. I was bothered by things that usually don't bother me.

$x13$. I did not feel like eating; my appetite was poor.

$x14$. I felt that everything was an effort.

$x15$. My sleep was restless.

$x16$. I could not "get going".

$x17$. I had trouble keeping my mind on what I was going.

$x18$. I talked less than usual.

Table 3 shows the standardized canonical coefficients. We can see these results more easily from quantification plots. Column plots of the three sets are given in Figure 1 to Figure 3. Observe that the axis with larger eigenvalues is more emphasized in the plot. Here, $c_{i(1,2)}$ means the correlation between the first-order and second-order canonical variables in the i-th set of variables.

In Figure 1, most of the points are on the right side of the first axis. Therefore we may interpret this axis as general level of the negative affection. Whereas, the second axis describes differences in the modes of expressing negative affection since it is weighted difference of crying($x4$) and feeling depression($x2$). In Figure 2, all the variables are located on the right side of the first axis. Thus this axis means a size factor representing the general level of the positive affection. The second axis is a contrast between hope for the future ($x9$) and satisfaction with the past ($x8$, $x10$ and $x11$). Thus it can be regarded as a reason of feeling positive affection. Similarly, from Figure 3, we can interpret that the first axis means general level of the somatic and retarded activity and the second axis means a expression way of the retardation.

Consideration of the three sets of column points simultaneously allows us to interpret the canonical correlation relationship. By superimposing Figure 1 to Figure 3 for the first axis, we can see that the first canonical variates represent the linear association among the general levels of three affection types. It means that people with higher depression for one affection type tends to have higher depression for the other affection types. In a similar manner, the second canonical variates can be interpreted as the differential association among three ways of expressing affection. For example, people who show negative affection by crying($x4$) tend to be more pessimistic to the past ($x8$,$x11$) and show extreme retarded activities($x13$, $x18$).

Average correlations($\bar{\rho}$) and average covariance (λ) between three sets of canonical variables are given in Table 4. Average correlation between the first-stage canonical variables ($\bar{\rho}_1$) is 0.5254. Average correlations between the second ($\bar{\rho}_2$), third ($\bar{\rho}_3$) , and fourth ($\bar{\rho}_4$) stage are given by 0.3659, 0.1480, and 0.1057, respectively. The average covariance between canonical variables has 1.2770 as its maximum, and so on.

The goodness-of-approximation and the explanatory power index for two dimensional quantification plots are given in Table 5. In each variable set, the goodness-of approximation for two dimensional plots are 90.7 %, 93.4 % and 94.1 %, respectively. On the other hand, the explanatory power indices for two dimensional quantification are given by 29.4 %, 17.9 % and 20.0 %.

It is not necessary to look at all of the individual plot . We can, however, plot the centroid of various data subsets from demographic variables. Figure 4 is observation(row) plot for the first variable set using employment group as demographic variables. The five employment groups are $m1$(full time employer), $m2$(part time employer), $m3$(unemployed person), $m4$(retired) and $m5$(house person). We obtain biplot of the first variable set by superimposing this row plot and the column plot of the first variable set. We can interpret that $m3$(unemployed person) and $m5$(house person) show relatively large negative affection while $m1$(full time employer) and $m4$(retired) show relatively small negative affection. In the second axis, however, they do not look so different.

Table 3. Standardized Canonical Coefficients

	a_1	a_2	a_3	a_4
$x1$	0.2399	-0.0045	0.5497	-1.0211
$x2$	0.4564	-1.2142	-0.2041	0.6074
$x3$	0.0284	0.4468	0.5719	-0.5738
$x4$	-0.0627	0.9701	0.1570	0.1138
$x5$	0.1972	0.0942	-1.0469	-0.6750
$x6$	0.1096	0.3084	-0.6645	0.9157
$x7$	0.3005	-0.1312	0.7083	0.7745
$x8$	0.1043	0.6813	0.3275	0.4867
$x9$	0.0516	-0.6516	0.7367	0.1860
$x10$	0.6830	0.0263	-0.9086	-0.8541
$x11$	0.3631	0.3625	0.5456	0.5251
$x12$	0.2006	0.1486	-0.6183	0.7104
$x13$	0.0530	0.5171	-0.2012	-0.5262
$x14$	0.1984	0.3134	-0.3859	0.3672
$x15$	0.3828	-0.4212	0.3211	-0.1681
$x16$	0.1736	-0.4913	0.2951	0.0932
$x17$	0.2335	-0.3973	-0.1077	-0.0322
$x18$	0.3089	0.5538	0.3555	-0.3786

Table 4. Average Correlation and Average Covariance

$\bar{\rho}_1$	$\bar{\rho}_2$	$\bar{\rho}_3$	$\bar{\rho}_4$
0.5254	0.3659	0.1480	0.1057
λ_1	λ_2	λ_3	λ_4
1.2770	0.4218	0.3120	0.2174

Table 5. Goodness-of-Approximation and Explanatory Power Index

	$GOA_{(2)}$	$EPI_{(2)}$
$X1$	90.7 %	29.4 %
$X2$	93.4 %	17.9 %
$X3$	94.1 %	20.0 %

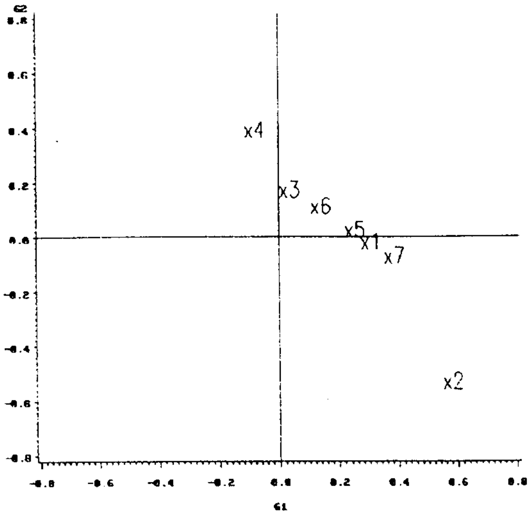

Figure 1. Column plot for negative affect items $c_{1(1,2)} = 0.0734$

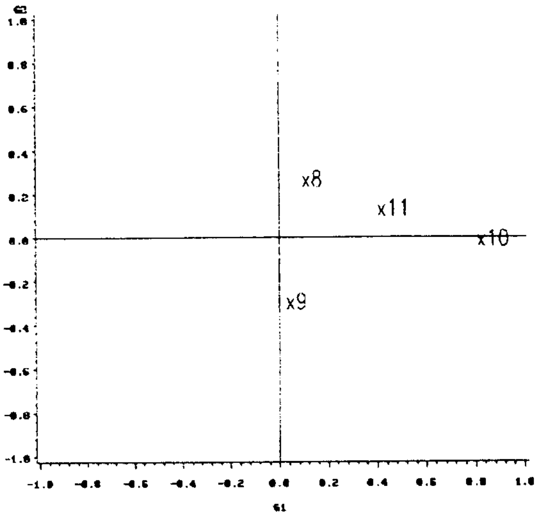

Figure 2. Column plot for positive affect items $c_{2(1,2)} = 0.1785$

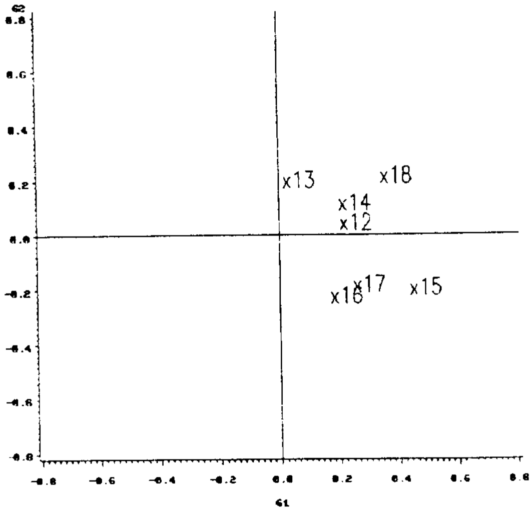

Figure 3. Column plot for somatic and retarded activity items $c_{3(1,2)} = 0.0735$

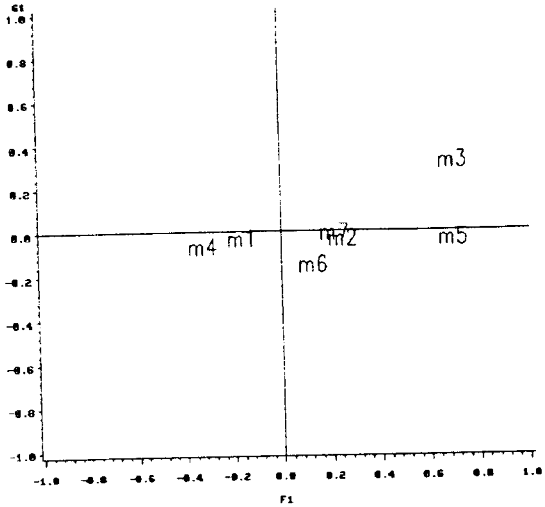

Figure 4. Row plot for negative affect items: employment groups

REFERENCES

(1) Afifi, A. A. and Clark, V. (1984). *Computer-aided Multivariate Analysis.* Lifetime Learning Publications, Belmont, California.

(2) Carroll, J. D. (1968). Generalization of canonical correlation analysis to three or more sets of variables, *Proceedings of American Psychology Association*, 227-8.

(3) Gabriel, K. R. (1971). The biplot graphic display of matrices with the application to principal component analysis, *Biometrika*, **58**, 3, 453-467.

(4) Gower, J. C. (1989). Generalized canonical analysis, In *Multiway Data Analysis*(Edited by R. Coppi and S. Bolasco), 221-232. Elsevier Science Publishers B. V., North-Holland.

(5) Hayashi, C. (1988). New developments in multidimensional data analysis, In: *Recent Developments in Clustering and Data Analysis.* (Edited by Diday, E., Hayashi, C., Jambu, M. and Ohsumi, N.), 3-16. Academic Press.

(6) Horst, P. (1961). Relations among m sets of measures. *Psychometrika*, **26**, 129-49.

(7) Hotelling, H. (1936). Relations between two sets of variates, *Biometrika*, **28**, 321-377.

(8) Kettenring, J. R. (1971). Canonical analysis of several sets of variables. *Biometrika*, **58**, 3, 433-451.

(9) Lafosse, R. (1989). Proposal for a generalized canonical analysis, In *Multiway Data Analysis*(Edited by R. Coppi and S. Bolasco), 269-276. Elsevier Science Publishers B. V., North-Holland.

(10) Lebart, L., Morineau, A., and Warwick, K.(1984). *Multivariate Descriptive Statistical Analysis : Correspondence Analysis and Related Techniques for Large Matrices.* John Wiley & Sons, New York.

(11) Park, M. R. (1995). Quantification plots for canonical correlation biplot, *Ph. D. thesis*, Korea University.

(12) Tanaka, Y., Tarumi, T. and Huh, M. H. (1994). Research and applications of quantification methods in East Asian countries. *New approaches in Classification and Data analysis* (Edited by E. Diday et al.), 64-71. Springer-Verlag, Berlin.