한국통계학회논문집 제3권 1호
1996년 4월 pp. 11-19

Canonical Correlation Biplot[1)]

Mira Park [2)] and Myung-Hoe Huh [3)]

Abstract

Canonical correlation analysis is a multivariate technique for identifying and quantifying the statistical relationship between two sets of variables. Like most multivariate techniques, the main objective of canonical correlation analysis is to reduce the dimensionality of the dataset. It would be particularly useful if high dimensional data can be represented in a low dimensional space.

In this study, we will construct statistical graphs for paired sets of multivariate data. Specifically, plots of the observations as well as the variables are proposed. We discuss the geometric interpretation and goodness-of-fit of the proposed plots. We also provide a numerical example.

1. Introduction

The main objective of multivariate analysis is to reduce dimensionality of the data set. For the "PCA(principal component analysis) data" in which there is only one set of multivariate observations, we have the biplot display due to Gabriel(1971).

The aim of canonical correlation analysis by Hotelling(1936) is to find out a number of linear relationships between two sets of multivariate observations. The applications of canonical correlation analysis are presented by many researchers such as Mardia et al.(1979) and Holland et al.(1980). The problem that will be studied here is to portray the "CCA(canonical correlation analysis) data" with statistical graphs similar to Gabriel's biplot, to help user's interpretation of algebraic result of the canonical correlation analysis.

2. Geometric Route for Quantification

Let the data matrix Z with N observations(rows) and $p+q$ variables(columns) be partitioned into two submatrices $X: n\times p$ and $Y: n\times q$. Thus $Z=[\ X: Y\ ] : n\times(p+q)$. For

1) This paper is partially supported by Korea Science and Engineering Foundation Grant 951-0103-010-1.
2) Postdoctoral researcher, Institute of Statistics, Korea University, Seoul, 136-701, Korea.
3) Professor, Department of Statistics, Korea University, Seoul 136-701, Korea.

convenience, we assume that columns are centered and standardized, unless mentioned otherwise. Furthermore, we assume $p \geq q$, without loss of generality, and, rank $(X) = p$, rank $(Y) = q$.

The columns of X and Y can be positioned as $p+q$ points in the Euclidean space R^n. Let S and T be the linear subspace of R^n that are generated by columns of X and Y, respectively. So, S consists of vectors Xa for $a \in R^p$, and similarly, T consists of vectors Yb for $b \in R^q$.

The aim of canonical correlation analysis can be formulated as

$$\max_{a,b} \; corr(Xa, Yb),$$

which, in turn, is equivalent to find out two vectors, one in each subspace S and T of R^n, that have minimum angle between them. It is well known (e.g. Mardia et al. 1979, p.282) that the algebraic solution comes from the eigensystem

$$(X'X)^{-1/2}(X'Y)(Y'Y)^{-1}(Y'X)(X'X)^{-1/2}u = \rho^2 u, \tag{1}$$

$$(Y'Y)^{-1/2}(Y'X)(X'X)^{-1}(X'Y)(Y'Y)^{-1/2}v = \rho^2 v, \tag{2}$$

and, finally, from the linear relationships

$$a^* = (n-1)^{1/2}(X'X)^{-1/2}u, \qquad b^* = (n-1)^{1/2}(Y'Y)^{-1/2}v.$$

Since (1) and (2) together can be shortened into a singular value decomposition of $(X'X)^{-1/2}(X'Y)(Y'Y)^{-1/2}$, or

$$(X'X)^{-1/2}(X'Y)(Y'Y)^{-1/2} = UDV', \tag{3}$$

where U is a $p \times q$ matrix of orthonormal columns, V is a $q \times q$ orthogonal matrix, and D is a diagonal matrix with $\rho_1 \geq \rho_2 \geq \cdots \geq \rho_q$ as its diagonal elements. Pre-multiplying $(X'X)^{-1/2}$ and post-multiplying v to both sides of (3), we have

$$(X'X)^{-1}(X'Y)(Y'Y)^{-1/2}v = \rho(X'X)^{-1/2}u,$$

or

$$(X'X)^{-1}(X'Y)b^* = \rho a^*. \tag{4}$$

Pre-multiplying X to both sides of (4) yields

$$X(X'X)^{-1}X'(Yb^*) = \rho X a^*. \tag{5}$$

Similarly, pre-multiplying u' and post-multiplying $(Y'Y)^{-1/2}$ to both sides of (3),

$$u'(X'X)^{-1/2}(X'Y)(Y'Y)^{-1}=\rho v'(Y'Y)^{-1/2},$$

or

$$(Y'Y)^{-1}(Y'X)a^*=\rho b^*. \tag{6}$$

Pre-multiplying Y to both sides of (6) yields

$$Y(Y'Y)^{-1}Y'(Xa^*)=\rho Yb^*. \tag{7}$$

"Dual relationships" (5) and (7), which can be found also in Lebart et al. (1984, p.68) and in Nishisato(1980, p.63), tell us that Xa^* and Yb^* are projections onto each other's space (up to the scale factor ρ). Therefore, Yb^* has a projection $X(X'X)^{-1}X'Yb^*$ or ρXa^* on the subspace S generated by the columns of X. And, thus, ρXa^* contains quantification of rows belonged to the X matrix, with external reference to prespecified scaling vector Yb^*. In a symmetrical way, quantifications of rows of the Y matrix with prespecified scaling vector Xa^* are contained in ρYb^*.

Now, consider supplementary data $X_S=I_p$ for the first set of variables. The elements of the s-th row e_s of X_S are zeros except for the s-th element, which is *1*. Since all the variables are centered and standardized, the supplementary observation e_s represents the single role of the s-th variable in the X dataset.

Therefore, quantification of the columns(or variables) of X is given by $\rho X_S a^*$ or ρa^*. Similarly, quantification of the columns(or variables) of Y is given by $\rho Y_S b^*$ or ρb^*, where Y_S is the $q\times q$ diagonal matrix with *1*'s as diagonal elements (thus, $Y_S=I_q$). Lebart et al.(1984; pp.14-16) used such supplementary data technique to represent additional individuals or variables in the case of principal component analysis.

This unidimensional quantification procedure for rows and columns of the X and Y matrices can be easily extended to multidimensional quantification. Table 1 summarizes row and column quantification formulas, with obvious matrix notation

$$A^*_{(r)}=(a_1^*,\cdots,a_r^*) \quad \text{and} \quad B^*_{(r)}=(b_1^*,\cdots,b_r^*) \qquad \text{for} \quad r\le \min(p,q).$$

Table 1. Quantification Formulas for Canonical Correlation Analysis

	Rows	Columns
Data matrix X	$XA^*_{(r)}D_{(r)}$	$A^*_{(r)}D_{(r)}$
Data matrix Y	$YB^*_{(r)}D_{(r)}$	$B^*_{(r)}D_{(r)}$

By plotting the first and the second order quantification results, we can obtain the two dimensional quantification plots for the rows and columns. We call such pairs of quantification plots, which are often but are not necessarily two-dimensional, by "canonical correlation biplot".

As can be seen in Table 1, the proposed method scales canonical coefficients and canonical scores up and down proportionately to the corresponding singular values. Compared to the conventional plot for canonical correlation analysis, our canonical correlation biplot put more emphasis on the axis with larger canonical correlation and depreciates the axis with smaller canonical correlation.

Now, let us look at the goodness of lower dimensional approximation offered by quantification plots. Lack-of-approximation by row quantification plot for X is captured by

$$\|XA^*D-X(A^*_{(r)}D_{(r)}:0_{p\times(q-r)})\|^2=\rho^2_{r+1}+\cdots+\rho^2_q,$$

compared to

$$\|XA^*D\|^2=\rho^2_1+\cdots+\rho^2_q,$$

where $\|C\|^2$ is defined as $tr(C'C)$. And thus we can define the r-dimensional goodness-of-approximation for X in row quantification plot as

$$\begin{aligned} GOA_{(r)} \text{ for } X &= 1-\|XA^*D-X(A^*_{(r)}D_{(r)}:0_{p\times(q-r)})\|^2/\|XA^*D\|^2 \\ &=(\rho^2_1+\cdots+\rho^2_r)/(\rho^{2+}_1\cdots+\rho^2_q), \end{aligned}$$

Similarly, for Y, it is similarly defined as

$$\begin{aligned} GOA_{(r)} \text{ for } Y &= 1-\|YB^*D-Y(B^*_{(r)}D_{(r)}:0_{q\times(q-r)})\|^2/\|YB^*D\|^2 \\ &=(\rho^2_1+\cdots+\rho^2_r)/(\rho^2_1+\cdots+\rho^2_q). \end{aligned}$$

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

$$\begin{aligned} GOA_{(r)} \text{ for } X &= 1-\|A^*D-(A^*_{(r)}D_{(r)}:0_{p\times(q-r)})\|^2_{X'X}\ /\ \|A^*D\|^2_{X'X} \\ &=(\rho^2_1+\cdots+\rho^2_r)/(\rho^2_1+\cdots+\rho^2_q), \end{aligned}$$

$$\begin{aligned} GOA_{(r)} \text{ for } Y &= 1-\|B^*D-(B^*_{(r)}D_{(r)}:0_{q\times(q-r)})\|^2_{Y'Y}\ /\ \|B^*D\|^2_{Y'Y} \\ &=(\rho^2_1+\cdots+\rho^2_r)/(\rho^2_1+\cdots+\rho^2_q), \end{aligned}$$

where $\|C\|^2_M$ is defined as $tr(C'MC)$. The reason why we use norming matrix $X'X$ or $Y'Y$ is that column quantifications are lacking absolute uniqueness under nonsingular transformation of X or Y : that is, $XA^*=XT_1(T_1^{-1}A^*)$ and $YB^*=YT_2(T_2^{-1}B^*)$ for any nonsingular matrices T_1 and T_2. Here, we may note that goodness-of-approximations for X and Y are the same.

A further interesting measure related to the r-dimensional quantification is the measure of

how well the quantification explains the external reference by projections. Among the statistical dispersion contained in X dataset

$$\|XA^*\|^2 = p,$$

only

$$\|YB^*_{(r)}D_{(r)}\|^2 = \rho_1^2 + \cdots + \rho_r^2$$

is reflected in the r-dimensional quantification plot. Thus we define the "explanatory power indices(EPI)" for X which are similar to the coefficient of determination R^2 in linear regression as

$$\begin{aligned} EPI_{(r)} \text{ for } X \text{ by } Y &= \|YB^*_{(r)}D_{(r)}\|^2 / \|XA^*\|^2 \\ &= (\rho_1^2 + \cdots + \rho_r^2)/p. \end{aligned}$$

Similarly, for Y, we define

$$\begin{aligned} EPI_{(r)} \text{ for } Y \text{ by } X &= \|XA^*_{(r)}D_{(r)}\|^2 / \|YB^*\|^2 \\ &= (\rho_1^2 + \cdots + \rho_r^2)/q. \end{aligned}$$

3. A Numerical Example

To illustrate the proposed methodology, consider a physical fitness test data from Tanaka et al.(1984). Seven(=p) motor variables and five(=q) exercise variables were measured on 38(=n) high-school freshmen. The list of motor variables and exercise variables are as follows:

motor variables	exercise variables
$x1$: side-step (number)	
$x2$: vertical jump (cm)	$y1$: 50m run (sec)
$x3$: back strength (kg)	$y2$: running-long jump (cm)
$x4$: grip strength (kg)	$y3$: throw (m)
$x5$: step-test (index)	$y4$: pull-ups (number)
$x6$: standing trunk flexion (cm)	$y5$: distance run (sec)
$x7$: chest raises (cm)	

Canonical correlations and standardized canonical coefficient vectors are listed in Table 2 and Table 3, respectively. For X variables, all the first-axis canonical coefficients are positive whereas the second-axis canonical coefficients are positive for $x2$, $x4$, and $x7$ and negative for other variables. For Y variables, the first-axis canonical coefficients are negative for $y1$ and $y5$ and positive for $y2$, $y3$ and $y4$. And, the second-axis canonical coefficients are all

positive. We can interpret these results more easily from quantification plots as we will see shortly. Row{column} quantification plots of the two sets X and Y are given in Figure 1{Figure 2} and Figure 3{Figure 4}. Note that the axis with larger canonical correlation is more emphasized in the plot.

In Figure 2, all the variables are found on the right side of the first axis. Therefore, we may interpret the first axis as "general motor fitness". The second axis can be interpreted as "motor balance" of power and endurance since it is the weighted difference of vertical jump ($x2$) and step-test($x5$). In Figure 4, all the variables except for 50m run($y1$) and distance run($y5$) are located on the right side of the first axis. Note that the athletic ability is good when $y1$ and $y5$ have small values and thus the first axis can be interpreted as "general athletic fitness". On the other hand, the second axis can be regarded as "athletic balance" of explosive strength(jump; $y2$) and speed(run; $-y1, -y5$). Therefore, the first canonical correlation comes from the linear association between general motor fitness and general athletic fitness. It means that the people with good general motor ability tends to have better general athletic ability. And the second canonical correlation can be interpreted as the correlation between motor balance and athletic balance. Thus people with good motor power is good at the exercise which needs explosive strength whereas people with good motor endurance is good at the exercise which needs speed.

Figure 1 and Figure 3 show the 13-th and the 23-rd observations contribute significantly to the formation of the first axis of X and Y observations, and the 38-th and the 4-th observation to the second axis. But they are quite different from each other in both variable sets. By superimposing the row quantification plots and the column quantification plots, we see that the 13-th observation has poor general motor and athletic fitness while the 23-rd has good general motor and athletic fitness. On the other hand, the 4-th observation has relatively large value for step-test ($x5$) and small values for running ($y1$ and $y5$). The 38-th has relatively large values for vertical jump($x2$) and he does running-long jump($y2$) better than other exercises.

In both variable sets, the goodness-of-approximation for two dimensional quantification plots are 66.5%. Explanatory power indices for X (by Y), and for Y (by X), are 17.9% and 25.1%, respectively.

Table 2. Canonical Correlations

ρ_1	ρ_2	ρ_3	ρ_4	ρ_5
0.8515	0.7284	0.6109	0.3932	0.3247

Table 3. Standardized Canonical Coefficients

	a_1^*	a_2^*	a_3^*	a_4^*	a_5^*
x1	0.4421	-0.2087	-0.4641	-0.5514	-0.2295
x2	0.2669	0.7020	0.9016	0.5570	0.0816
x3	0.5884	-0.2102	-0.4639	-0.1332	0.3773
x4	0.0614	0.0148	0.5662	-0.1537	-0.9126
x5	0.2217	-0.7263	0.7237	0.4626	0.3014
x6	0.0911	-0.1749	-0.4354	0.4752	-0.1839
x7	0.0138	0.2399	-0.1718	-1.0336	0.5283

	b_1^*	b_2^*	b_3^*	b_4^*	b_5^*
y1	-0.4266	0.8255	-0.3704	0.6538	-0.1413
y2	0.2335	1.0405	-0.2532	-0.7668	-0.4791
y3	0.3696	0.1982	-0.2894	0.3748	1.0851
y4	0.0038	0.2218	0.8850	0.9935	-0.0904
y5	-0.3560	0.8101	0.5374	-0.1585	0.7481

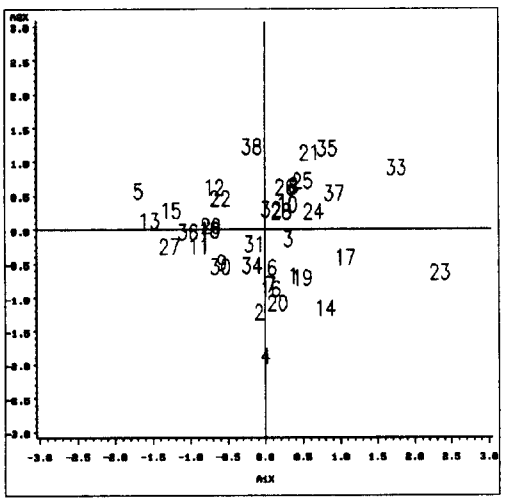

Figure 1. Row quantification plot for motor observations

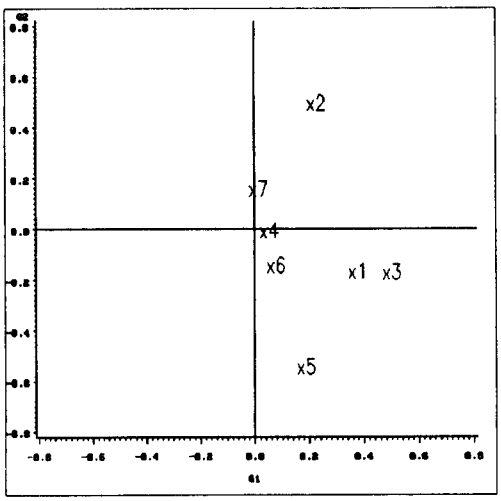

Figure 2. Column quantification plot for motor variables

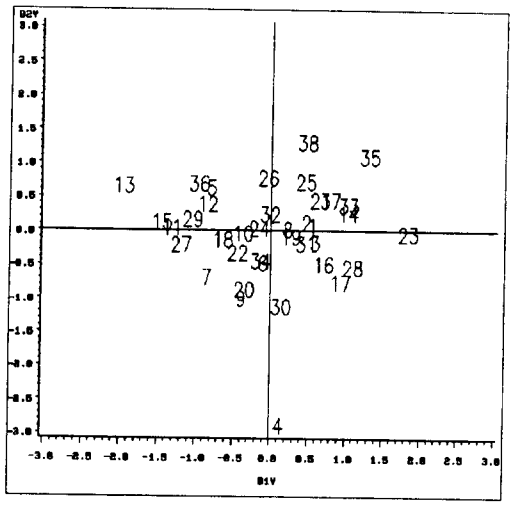

Figure 3. Row quantification plot for exercise observations

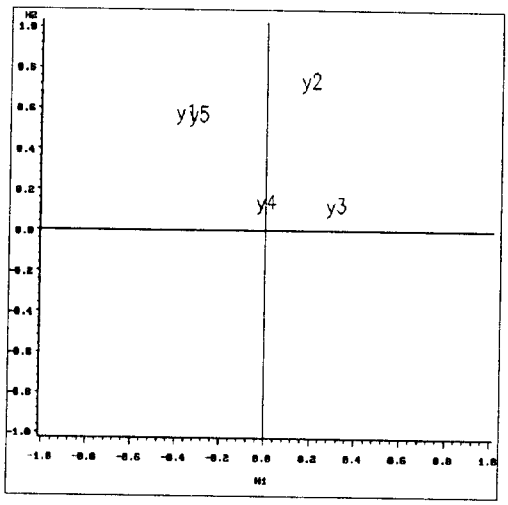

Figure 4. Column quantification plot for exercise variables

4. Remarks

Let us consider the applications of the proposed quantification plots to some other types of data. There are several comparable multivariate methods to canonical correlation analysis, such as the discriminant analysis, the correspondence analysis and Japanese quantification method II . Discriminant analysis is a special case of the canonical correlation analysis when the one set consists of dummy variables. Usually, the discriminant scores and the discriminant coefficients are plotted. On the other hand, by applying our method we obtain the plot which emphasize the axis with larger singularvalues more, compared to the conventional plot.

To analyze the two-way contingency table using proposed method, we need to define dummy variables for each category and re-express the data in the form of a cases-by-variables indicator matrix. Correspondence analysis is an alternative graphical method. Various forms of the correspondence analysis have been discussed extensively by Nishisato(1980) and Greenacre and Hastie(1987).

Japanese quantification method II is used for analyzing two sets of qualitative data. We can also apply our method to this type of data using dummy variables. In this case, any centering and standardization of the raw data matrices should not be done. In 1986, Tarumi and Tanaka suggested the quantification results same as ours(Mori and Tarumi, 1993). However, We think this study laid a solid foundation for such quantification formulas. Recently, Tanaka et al.(1994) discussed several technical problems concerning quantification method II. Table 4 lists the comparative summary of the quantification results.

Table 4. Various Quantification Formulas

Method \ Data	X row	X column	Y row	Y column
Canonical Correlation Biplot	$XA^*_{(r)}D_{(r)}$	$A^*_{(r)}D_{(r)}$	$YB^*_{(r)}D_{(r)}$	$B^*_{(r)}D_{(r)}$
Discriminant Analysis	$XA^*_{(r)}$	$A^*_{(r)}$	$YB^*_{(r)}$	$B^*_{(r)}$
Correspondence Analysis		$A^*_{(r)}D_{(r)}$ $A^*_{(r)}$		$B^*_{(r)}$ $B^*_{(r)}D_{(r)}$
Japanese Quantification Method II		$A^*_{(r)}D_{(r)}$ $A^*_{(r)}$		$B^*_{(r)}$ $B^*_{(r)}D_{(r)}$

References

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

(2) Greenacre, M., and Hastie, T.(1987). The geometric interpretation of correspondence analysis. *Journal of the American Statistical Association*, 82, 437-447.

(3) Holland, T. R., Levi, M. and Watson, C. G.(1980). Canonical correlation in the analysis of a contingency table. *Psychological Bulletin*, 87, 334-336.

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

(5) 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.

(6) Mardia, K. V., Kent, J. T., and Bibby, J. M.(1979). *Multivariate Analysis*, London: Academic Press.

(7) Mori, Y. and Tarumi, T.(1993). Statistical software SAM II: Sensitivity analysis in multivariate methods. *Journal of Japanese Society of Computational Statistics*, 6, 21-32.

(8) Nishisato, S.(1980). *Analysis of Categorical Data: Dual Scaling and its Applications*, Toronto: University of Toronto Press.

(9) 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.

(10) Tanaka, Y., Tarumi, T., and Wakimoto, K.(1984). *Handbook of Statistical Analysis with Programs for Personal Computers*, Vol.2, Kyoritsu Publishing Company.(written in Japanese).