Journal of the Korean
Statistical Society
Vol. 24, No. 2, 1995

Biplot of Ranked Data[†]

Sang-Tae Han[1] and Myung-Hoe Huh[2]

ABSTRACT

Ranked data are widely used in the area of social sciences, for instance in polls and preference surveys, in which a number of objects(or stimuli) are evaluated and ranked by a panel of judges(or subjects) according to their preference.

We propose a graphical method for ranked data by quantifying objects and judges. In a plot for judges, the interpoint distances can be interpreted as Spearman or Kendall distances between two rankings given by respective judges. Similarly, we also construct a plot for objects with a sensible relationship to the previous plot for judges.

KEYWORDS: Spearman-type quantification, Kendall-type quantification, Row plot, Column plot, Biplot.

1. INTRODUCTION

Ranked data are one of the most popular types of survey data we see in our daily life. For example, we rank TV programs, government services, or consumer products according to the order of their quality. Ranked data drew attention of many researchers such as Critchlow(1985), Baba(1986), Fligner

---

[†]This research was supported by Korea Science and Engineering Foundation Grant 951-0103-010-1.

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

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

and Verducci(1986), Diaconis(1988), and Chung and Marden(1992) among others.

Let r_{ij} denote the rank given to the object $j(=1,\ldots,p)$ by the judge $i(=1,\ldots,n)$, and write $R=\{r_{ij}\}$. Although this matrix notation is conventional, we will use two other coding schemes for ranked data in this study. The first one is $S=\{s_{ij}\}$ with n rows and p columns, where

$$s_{ij}=r_{ij}-(p+1)/2.$$

The second one is $K=\{k_{ij}\}$ with n rows and $p(p-1)/2(=p^*)$ columns which are arranged in lexicographic order among $j=(k,l), \quad 1\leq k<l\leq p$, where

$$\begin{aligned} k_{ij}&=1, \quad \text{if the object } k \text{ is preferred to } l,\\ &=-1, \quad \text{otherwise.} \end{aligned}$$

The S notation will be useful for later use of matrix decomposition, since redundant row averages, (p+1)/2, of the raw data matrix R are removed in S. We observe that this row centering process is in line with the spirit of Q-sort methodology(Hair, Anderson, Tatham and Grablowsky, 1979). In Spearman's sense, the squared rank distance between two rows(or judges) i and i' is defined by

$$d_S^2(i,i')=\sum_{j=1}^{p}(s_{ij}-s_{i'j})^2.$$

Also, the Spearman rank correlation between two rows i and i' is given by

$$1-6d_S^2(i,i')/p(p^2-1).$$

On the other hand, the K notation that are derived directly from pairwise comparison of objects is convenient in computing rank distances in Kendall's sense. The Kendall rank distance between two rows(or judges) i and i' can be defined by

$$d_K(i,i')=\#\ \{k_{ij}k_{i'j}<0, \quad j=1,\ldots,p^*\}.$$

It follows that

$$d_K(i,i') = \sum_{j=1}^{p*}(k_{ij} - k_{i'j})^2/4.$$

And, the Kendall rank correlation between two rows i and i' is given by

$$1 - 4d_K(i,i')/p^*(p^* - 1).$$

The aim of this study is to extend Nishisato's(1980) quantification method of ranked data, using Gabriel's(1971) biplot technique for multivariate data matrix. Specifically, we will propose a row plot, in which the interdistance between two plot points can be interpreted as approximation of the (squared) rank distance between two rankings given by corresponding judges either in Spearman's or Kendall's sense. Similarly, we will also propose a column plot for objects with a sensible relationship to the row plot for judges.

2. SPEARMAN-TYPE QUANTIFICATION OF RANKED DATA

For lower dimensional reduction technique of multivariate data, we will resort to Lebart, Morineau, and Warwick(1984; Chapter 1). The rows s_i of S can be considered as vectors in $\Re^p$. Let v be a unit vector in $\Re^p$. The magnitude of projection vector of s_i on v is equal to $s_i'v$. Our objective is to maximize the sum of squared inner products. Then, the optimization problem can be formulated as

$$\max_v \sum_{i=1}^{n}(s_i'v)^2, \qquad \text{subject to } v'v = 1.$$

Thus principal component(PC) reduction can be derived from an eigensystem

$$(S'S)v = \lambda v, \qquad v \in \Re^p.$$

More specifically, the m-th PC score vector for n rows is given by

$$r_m = Sv_m,$$

where v_m is the eigenvector corresponding to the m-th largest eigenvalue λ_m of $S'S$, $m = 1, \ldots, p; \lambda_p = 0$. So, for the two dimensional display of rows(or judges), we may consider a plot, say a *row plot*, of n points from first two PC score vectors r_1 versus r_2. Then the interpoint distances between two points in the row plot are approximations of squared Spearman rank distance between corresponding rankings given by judges with overall goodness-of-approximation

$$GOA_{row(2)} = 1 - ||SV - (G^*_{(2)} : O_{n\times(p-2)})||^2/||SV||^2,$$

where $G^*_{(2)} = S(v_1 : v_2) = (r_1 : r_2)$. It turns out that

$$GOA_{row(2)} = 1 - \{\lambda_3 + \cdots + \lambda_p\}/\{\lambda_1 + \cdots + \lambda_p\} = \{\lambda_1 + \lambda_2\}/\{\lambda_1 + \cdots + \lambda_p\},$$

with interpretation as a measure of numerical quality of the recovery.

For the plot of columns(or objects), we use the same PC axis vector v and a group of hypothetical supplementary rows(or judges). Specifically, to locate the first object for instance, consider $(p-1)!$ rankings

$$(1, 2, 3, \ldots, p), (1, 2, p, \ldots, 3), (1, 3, 2, \ldots, p),$$

$$(1, 3, p, \ldots, 2), (1, p, 2, \ldots, 3), (1, p, 3, \ldots, 2), \ldots,$$

which give rank 1 to the first object. Then the centroid of these rankings is given by

$$c_1 = (1, (p+2)/2, (p+2)/2, \ldots, (p+2)/2)'.$$

When the size p vector c_1 is projected on the PC axis vector v, it is positioned at $c_1'v$. Similarly , c_j are defined for $j = 2, \ldots, p$ and their projections on v can be carried out individually. Finally, the p objects can be positioned, say in a *column plot*, at

$$(c_1'v_1, c_1'v_2), (c_2'v_1, c_2'v_2), (c_3'v_1, c_3'v_2), \ldots, (c_p'v_1, c_p'v_2).$$

Here comes a short cut to the computation of row and column plots, via singular value decomposition(SVD) of the data matrix S. Let

$$S = UDV',$$

where U is the $n \times p$ matrix with orthonormal columns, V is the $p \times p$ orthogonal matrix, and D is the $p \times p$ diagonal matrix with singular values, $\lambda_1^{1/2} \geq \cdots \geq \lambda_{p-1}^{1/2} > \lambda_p^{1/2} (= 0)$ as its diagonal elements. Then two-dimensional row plot points are given by the rows of

$$G^*_{(2)} = S(v_1 : v_2) = UDV'(v_1 : v_2) = U_{(2)} D_{(2)},$$

where $U_{(2)}$ is the $n \times 2$ submatrix of U and $D_{(2)}$ is the 2×2 diagonal submatrix of D.

For the column plot, first note that

$$c_j = (p+2)/2 \; (1, \ldots, 1)' - p/2 \; (0, \ldots, 1, \ldots, 0)',$$

and that

$$\begin{aligned} c_j' v &= (p+2)/2 \; (1, \ldots, 1) v - p/2 \; (0, \ldots, 1, \ldots, 0) v \\ &= -p/2 \; (0, \ldots, 1, \ldots, 0) v, \end{aligned}$$

since $S\mathbf{1} = 0$ for $\mathbf{1} = (1, \ldots, 1)'$, by coding definition, implies that

$$U'(UDV'\mathbf{1}) = 0 \quad \text{or} \quad DV'\mathbf{1} = 0,$$

or $v_j'\mathbf{1} = 0$ for $j = 1, \ldots, p-1$. Hence the column points for objects in the two dimensional plot are given by the rows of

$$\begin{bmatrix} c_1' \\ \vdots \\ c_j' \\ \vdots \\ c_p' \end{bmatrix} [v_1 : v_2] = -p/2 I_p [v_1 : v_2] = -p/2 [v_1 : v_2].$$

Therefore, two-dimensional column plot points can be obtained from the rows of

$$H^*_{(2)} = V_{(2)},$$

where $V_{(2)}$ is the $n \times 2$ submatrix of V, with the understanding that the plot points are directed toward the least (not the most) preference for each object.

The goodness-of-approximation of the two dimensional column plot is given by

$$GOA_{col(2)} = 1 - ||V_{[p]} - (V_{(2)} : O_{p\times(p-3)})||^2/||V_{[p]}||^2 = 1 - (p-3)/(p-1)$$
$$= 2/(p-1),$$

where $V_{[p]}$ is the $p \times (p-1)$ submatrix of V. We may notice that $GOA_{col(2)}$ is always greater than $GOA_{row(2)}$. This fact let us understand that the rows(judges) are represented with higher accuracy than the columns(objects) in biplots of ranked data.

The *biplot*, produced by combining row and column plots, can be interpretable as follows. First, note that S can be expressed as

$$S = (UD)V' \cong (U_{(2)}D_{(2)})V'_{(2)} = G^*_{(2)}H^*_{(2)}{}'.$$

For s_{ij}, the (i,j) element of S, we can write

$$s_{ij} \cong g'_{i(2)}h_{j(2)}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, p,$$

where $g'_{i(2)}$ and $h'_{j(2)}$ are i-th and j-th row vectors of $G^*_{(2)}$ and $H^*_{(2)}$. Thus the raw data elements are recovered by the inner products of row and column plot vectors.

3. KENDALL-TYPE QUANTIFICATION OF RANKED DATA

The same idea in Spearman-type quantification of Section 2 can be also applied to another coding matrix K for ranked data in a similar way. Since K has n rows and $p(p-1)/2(=p^*)$ columns, the rows of K can be considered as vectors in $\Re^{p^*}$. Therefore, principal component(PC) reduction of the rows can be derived from an eigensystem

$$(K'K)v = \lambda v, \qquad v \in \Re^{p^*}.$$

Assume $n \geq p^*$, for notational convenience, and consider SVD of the matrix K:

$$K = UDV',$$

where U is the $n \times p^*$ matrix with orthonormal columns, V is the $p^* \times p^*$ orthonormal matrix, and D is the $p^* \times p^*$ diagonal matrix with singular values, $\lambda_1^{1/2} \geq \cdots \geq \lambda_{p^*}^{1/2}$ as its diagonal elements. Then two-dimensional *row plot* points are given by

$$G^*_{(2)} = K(v_1 : v_2) = UDV'(v_1 : v_2) = U_{(2)}D_{(2)},$$

where $U_{(2)}$ is the $n \times 2$ submatrix of U and $D_{(2)}$ is the 2×2 diagonal submatrix of D.

In the row plot, interpoint distances are approximations of Kendall rank distance (up to the scale factor 4) between corresponding rankings given by judges with overall goodness-of-approximation

$$GOA_{row(2)} = 1 - ||KV - (G^*_{(2)} : O_{n\times(p^*-2)})||^2/||KV||^2,$$

where $G^*_{(2)} = K(v_1 : v_2) = (r_1 : r_2)$. It turns out that

$$GOA_{row(2)} = 1 - \{\lambda_3 + \cdots + \lambda_{p^*}\}/\{\lambda_1 + \cdots + \lambda_{p^*}\} = \{\lambda_1 + \lambda_2\}/\{\lambda_1 + \cdots + \lambda_{p^*}\}.$$

Next, we will construct two kinds of *column plot*. We will call these column plot P^* and P. The first one is column plot P^* with p^* points. In the two-dimensional P^*, column plot points can be obtained from the rows of

$$H^*_{(2)} = V_{(2)},$$

where $V_{(2)}$ is the $p^* \times 2$ submatrix of V. The goodness-of-approximation of the two dimensional column plot P^* is given by

$$GOA_{col(2)} = 1 - ||V - (V_{(2)} : O_{p^*\times(p^*-2)})||^2/||V||^2 = 1 - (p^* - 2)/p^* = 2/p^*.$$

In the case $n < p^*$, we can easily find out that $GOA_{col(2)}$ is $2/n$ instead of $2/p^*$.

The second one is column plot P with only p points. Specifically, to locate the first object for instance, consider $(p-1)!$ rankings

$$(1,\ldots,p-2,p-1,p),(1,\ldots,p-2,p,p-1),(1,\ldots,p-1,p-2,p),$$

$$(1,\ldots,p-1,p,p-2),(1,\ldots,p,p-2,p-1),(1,\ldots,p,p-1,p-2),\ldots,$$

all of which give rank 1 to the first object. In the K notation, they corresponds to

$$(1,\ldots,1,1,1),(1,\ldots,1,1,-1),(1,\ldots,-1,1,1),$$

$$(1,\ldots,-1,-1,1),(1,\ldots,1,-1,-1),(1,\ldots,-1,-1,-1),\ldots.$$

Then the centroid of these vectors is given by

$$c_1=(1,1,\ldots,1,0,\ldots,0,0,0)',$$

that is, the first $p-1$ elements are $1'$s and remaining $p^*-(p-1)$ elements are zeros. Similarly we can define $c_2,\ldots,c_p$.

When the size p^* vector c_1 is projected on the PC axis vector v, it is positioned at $c_1'v$. Finally, the p objects can be positioned in a two-dimensional column plot P at

$$(c_1'v_1,c_1'v_2),(c_2'v_1,c_2'v_2),(c_3'v_1,c_3'v_2),\ldots,(c_p'v_1,c_p'v_2).$$

Hence the column points for objects in the two dimensional column plot P are given by the rows of

$$\begin{bmatrix} c_1' \\ \vdots \\ c_j' \\ \vdots \\ c_p' \end{bmatrix} [v_1:v_2] \;=\; K_s[v_1:v_2] \;=K_sV_{(2)},$$

where K_s is a $p\times p^*$ matrix and $V_{(2)}$ is the $p^*\times 2$ submatrix of V. Then, the goodness-of-approximation of the two dimensional column plot P is given by

$$GOA_{col(2)}=1-||K_sV-(K_sV_{(2)}:O_{p\times(p^*-2)})||^2/||K_sV||^2.$$

4. A NUMERICAL EXAMPLE

To illustrate the proposed quantification plots, consider a ranked data from Nishisato and Nishisato(1984) in which 31 judges were asked to rank ten government services(and facilities) according to the order of "being satisfactory". The list of services or objects are:

A = public transit, B = postal service,
C = medical care, D = sports/recreational facilities,
E = police protection, F = public libraries,
G = street cleaning, H = restaurants,
I = theatres, J = planning and development,

to which are given the rank 1 for "most satisfactory",..., the rank 10 for "least satisfactory".

Row plot by Spearman-type quantification is shown in Figure 1 with the goodness-of-approximation 60%. Thus sixty percent of the total variation among judges is explained by the first two PC axes. Also, column plot by Spearman-type quantification are shown in Figure 2 with the goodness-of-approximation 22%. By superimposing the row and column plots, we obtain Spearman-type biplot of ranked data which is shown in Figure 3.

In Figure 1, the largest cluster of judges is found in the right-side region along the first axis, of which corresponding position in Figure 2 is assigned to the object B(postal service). It means that this group of judges agree to put B as "least satisfactory". Also, in Figure 1, the second largest cluster of judges is located in the upper-side region along the second axis. of judges is assigned to the object B. So, we interpret that this group of Their opposite position in Figure 2 is occupied by the objects I(theatres) and H(restaurants). So, we may interpret that the second largest group of judges evaluate theatres and restaurants very highly.

Similar interpretations can be made from Figure 4, Figure 5, Figure 6 and Figure 7 which are obtained by Kendall-type quantification. Figure 4, a row plot with goodness-of-approximation 50%, is similar to Figure 1. As a matter of fact, the whole configuration of Figure 4 is a reflected version along the resulting axes to allow viewers for easy comparison with Figure 1. In Figure 5, a column plot P^* with goodness-of-approximation 5%, we observe that too many points are plotted so that it looks very complex and it seems difficult to extract meaningful informations. But, Figure 6, a column plot P with goodness-of-approximation 20%, looks much simpler, so that information easily comes out. The practical difference between column plot P^* and P is that the former is for all pairs of objects, while the latter is for all single objects. Figure 7, obtained by superimposing Figure 4 and Figure 6, is a biplot by Kendall type quantification. In this particular case, Kendall-type quantification plots are close to Spearman-type quantification plots.

5. DISCUSSION

Nishisato(1980) studied the quantification analysis of ranked data. But, he did not plot for judges. Also, Baba(1986) proposed a linked rank graph and a rank graph. The former represents the distribution of rankings, whereas the latter represents the average ranks and the degree of concordance. These graphs give us good information about the typical rank of each object, whereas they are not suitable for the clustering of judges. The method we proposed is such a nice graphical tool that subjects(or judges) as well as objects can be plotted to show off mutual relationships. One particular merit in the row plot is that Spearman's or Kendall's distances between rankings given by judges are approximately preserved. Moreover, the methodology can be used for segmentation of judges(or observations).

REFERENCES

(1) Baba, Y. (1986). Graphical Analysis of Rank Data. *Behaviormetrika*, **19**, 1–15.

(2) Chung, L. and Marden, J.I. (1991). Use of Nonnull Models for Rank Statistics in Bivariate, Two Sample, and Analysis of Variance Problems. *Biometrika*, **44**, 114–130.

(3) Critchlow, D.E. (1985). *Metric Methods for Analyzing Partially Ranked Data*. Lecture Notes in Statistics, **34**. Springer, New York.

(4) Diaconis, P. (1988). *Group Representations in Probability and Statistics.* Lecture Notes-Monograph Series, **11** Institute of Mathematical Statistics. CA: Hayward.

(5) Fligner, M.A. and Verducci, J.S. (1986). Distanced-Based Ranking Models. *Journal of the Royal Statistical Society*, Ser. **B 48**, 359–369.

(6) Gabriel, K.R. (1971). The Biplot Graphics of Multivariate Matrices with Applications to Principal Component Analysis. *Biometrika*, **58**, 453–467.

(7) Hair, J.F., Anderson, R.E., Tatham, R.L., and Grablowsky, B.J. (1979). *Multivariate Data Analysis with Readings*. Petroleum Publishing Co., OK: Tulsa.

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

(9) Nishisato, S. (1980). *Analysis of Categorical Data: Dual Scaling and Its Applications*. Mathematical Expositions, **24**, University of Toronto Press, Toronto.

(10) Nishisato, S. and Nishisato, I. (1984). *An Introduction to Dual Scaling.* MicroStats, Toronto.

Figure 1. Row plot for thirty-one judges by Spearman-type quantification ($GOA_{row(2)} = 60\%$).

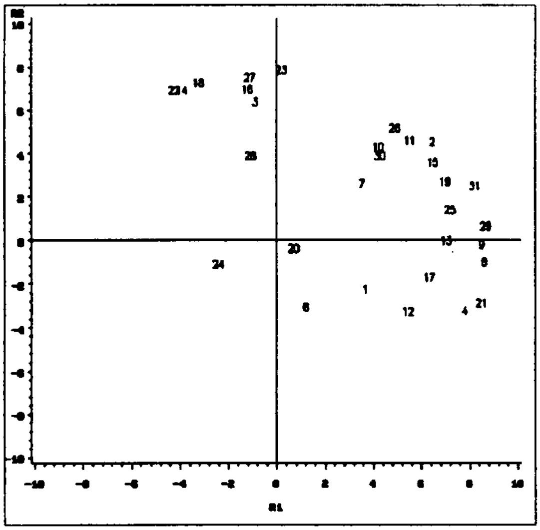

Figure 2. Column plot for ten government services by Spearman-type quantification ($GOA_{col(2)} = 22\%$).

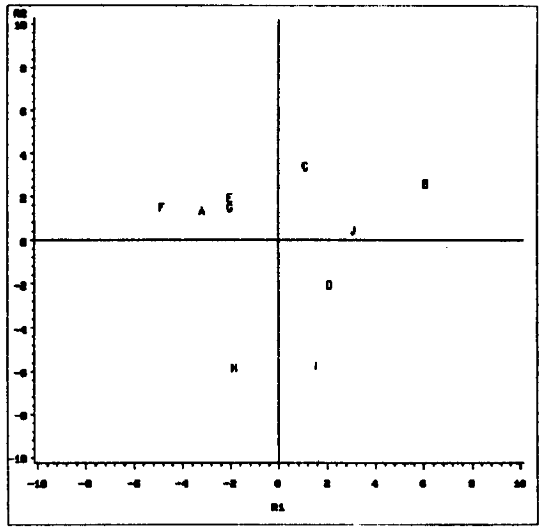

Figure 3. Biplot for evaluation of government services by Spearman-type quantification.

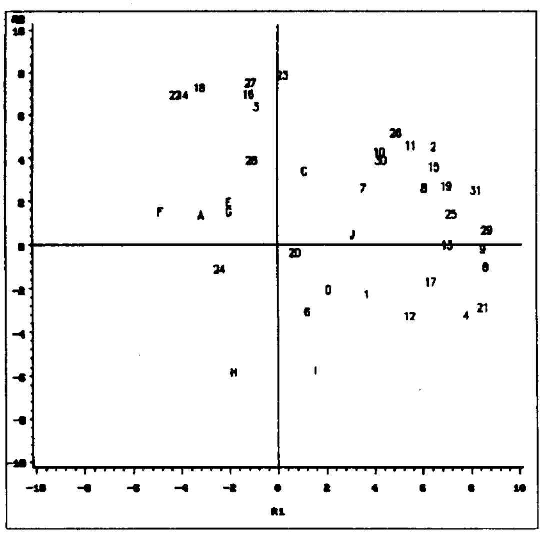

Figure 4. Row plot for thirty-one judges by Kendall-type quantification ($GOA_{row(2)} = 50\%$).

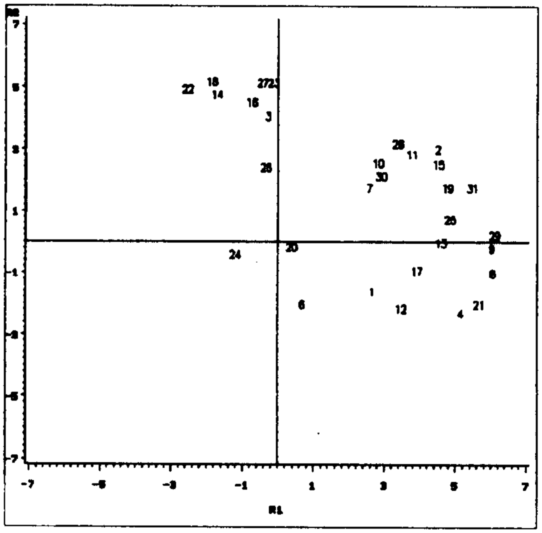

Figure 5. Column plot P^* for ten government services by Kendall-type quantification ($GOA_{col(2)} = 5\%$).

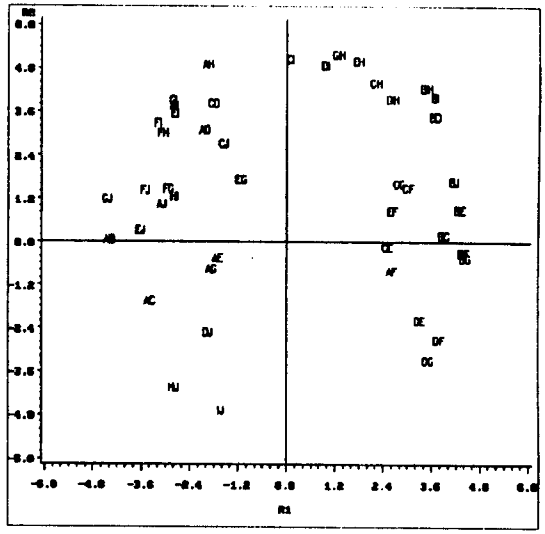

Figure 6. Column plot P for ten government services by Kendall-type quantification ($GOA_{col(2)} = 20\%$).

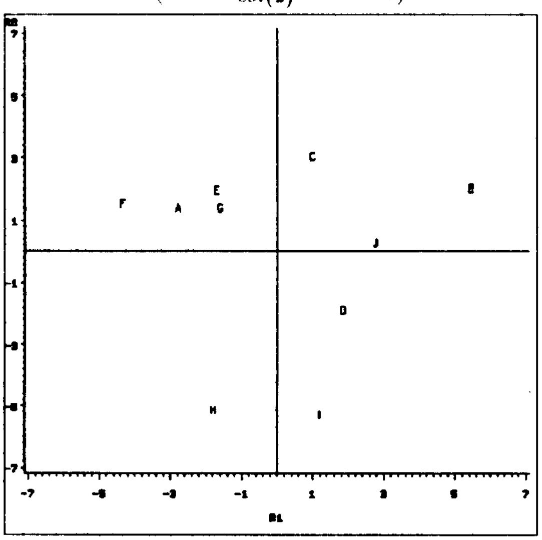

Figure 7. Biplot for evaluation of government services by Kendall-type quantification.