Communications of the Korean Statistical Society
2011, Vol. 18, No. 5, 637–644

DOI: http://dx.doi.org/10.5351/CKSS.2011.18.5.637

Graphical Representation of Partially Ranked Data

Sang-Tae Han[1,a]

[a]Department of Informational Statistics, Hoseo University

Abstract

Partially ranked data refers to the situation in which there are p distinct objects; however each judge specifies only first $s\,(s < p)$ choices. The group theoretic formulation for partially ranked data analysis was set up by Critchlow (1985). We propose a graphical method for partially 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. This study extends the Han and Huh (1995) quantification method of fully ranked data using Gabriel's (1971) biplot technique for multivariate data matrix.

Keywords: Partially ranked data, quantification, row plot, column plot, Biplot.

1. Introduction

Partially ranked data are one of the most popular types of survey data. For example, potential buyers order different brands of personal computers in terms of which one each buyer specifies only first through s-th choices, where $s < p$. Statistical analysis of such data were dealt in Critchlow (1985) and Diaconis (1988) 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 $R = \{r_{ij}\}$. In any partial ranking, there are certain objects which are 'tied' in the sense that the judge does not state any preferences among them. The "tied ranks approach" assigned to these tied objects is the average of the ranks which they would possess if they were distinguishable. Although this matrix notation is conventional, we will use two other coding schemes for partially ranked data in this study. The first one is $S = \{s_{ij}\}$ with n rows and p columns, where

$$s_{ij} = \bar{r}_{ij} - \frac{p+1}{2},$$

where $\bar{r}_{ij}$ is the rank given to object j by the judge i if it is less than or equal to s or equal to $(s+p+1)/2$ otherwise. The second one is $K = \{k_{ij}\}$ with n rows and $p(p-1)/2\,(= p^*)$ columns which are arranged in lexicographic order among $j = (k, l)$, $1 \le k < l \le p$, where

$$k_{ij} = \begin{cases} 1, & \text{if the object } k \text{ is preferred to } l, \\ -1, & \text{if the object } l \text{ is preferred to } k, \\ 0, & \text{otherwise.} \end{cases}$$

This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government(NRF-313-2008-2-C00137).

[1] Professor, Department of Informational Statistics and Institute of Basic Science, Hoseo University, Asan 336-795, Korea. E-mail: sthan@hoseo.edu

The S notation will be useful for the later use of matrix decomposition since redundant row averages, $(p+1)/2$, of the raw data matrix R are removed in S. 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

$$1 - 6\,\frac{d_S^2(i,i')}{2s(s+1)(2s+1) + 3(p-s)(s+p+1)^2 - 3p(p+1)^2}.$$

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

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

It follows that

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

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

$$1 - 4\,\frac{d_K(i,i')}{p(p-1) - 2(p-s-1)(p-s)}.$$

2. Spearman-Type Quantification

Consider a lower dimensional reduction of partially ranked data. For the data matrix S, the rows s_i of S can be considered as vectors in $\mathfrak{R}^p$. Let v be a unit vector in $\mathfrak{R}^p$. The magnitude of the projection vector of s_i on v is equal to $s_i'v$. Then, the optimization problem can be formulated as

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

Adopting a Lagrangian multiplier method, principal component(PC) reduction can be derived from an eigensystem

$$(S'S)v = \lambda v, \quad v \in \mathfrak{R}^p.$$

More efficiently, biplot of partially ranked data can be obtained from singular value decomposition(SVD) of the data matrix S (Lebart *et al.*, 1984). Let

$$S = UDV',$$

where U is the $n \times p$ matrix with orthonormal columns, $V = (v_1, v_2, \ldots, v_p)$ is the $p \times p$ orthogonal matrix, and D is the $p \times p$ diagonal matrix with singular values, $\lambda_1 \geq \cdots \geq \lambda_{p-1} > (\lambda_p = 0)$ 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. One particular merit in the row plot is that Spearman's distances between rankings given bu judges(rows of S) are approximately preserved. Therefore, the goodness-of-approximation of the two-dimensional row plot is given by

$$\begin{aligned} \mathrm{GOA}_{row(2)} &= 1 - \frac{\left\| SV - \left(G^*_{(2)} : O_{2n\times(p-2)}\right)\right\|^2}{\|SV\|^2} \\ &= \frac{\lambda_1^2 + \lambda_2^2}{\lambda_1^2 + \cdots + \lambda_p^2}. \end{aligned}$$

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 $(p-1)!/(p-s)!$ rankings

$$\left(1, 2, 3, \ldots, s, \frac{(p+s+1)}{2}, \ldots, \frac{(p+s+1)}{2}\right), \left(1, 2, \frac{(p+s+1)}{2}, \ldots, s, \frac{(p+s+1)}{2}, \ldots, 3\right), \ldots,$$

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

$$c_1 = \left(1, \frac{(p+2)}{2}, \frac{(p+2)}{2}, \ldots, \frac{(p+2)}{2}\right)'.$$

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

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

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, 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. Therefore, the goodness-of-approximation of the two dimensional column plot is given by

$$\mathrm{GOA}_{col(2)} = 1 - \frac{\left\| V_{[p]} - \left(V_{(2)} : O_{p\times(p-3)}\right)\right\|^2}{\left\|V_{[p]}\right\|^2} = \frac{2}{p-1},$$

where $V_{[p]}$ is the $p \times (p-1)$ submatrix of V. Row and column plots together may be called a biplot of partially ranked data.

3. Kendall-Type Quantification

The same idea in the 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, the principal component(PC) reduction of the rows can be derived from an eigensystem

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

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, $\lambda_1 \geq \cdots \geq \lambda_{p^*}$ as its diagonal elements. The 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

$$\mathrm{GOA}_{row(2)} = 1 - \frac{\left\| KV - \left(G^*_{(2)} : O_{n\times(p^*-2)}\right)\right\|^2}{\|KV\|^2},$$

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

$$\mathrm{GOA}_{row(2)} = \frac{\left\{\lambda_1^2 + \lambda_2^2\right\}}{\left\{\lambda_1^2 + \cdots + \lambda_{p^*}^2\right\}}.$$

Next, we will construct the column plot. Specifically, to locate the first object for instance, consider $(p-1)!/(p-s)!$ rankings

$$\left(1, 2, 3, \ldots, s, \frac{(p+s+1)}{2}, \ldots, \frac{(p+s+1)}{2}\right), \left(1, 3, 2, \ldots, s, \frac{(p+s+1)}{2}, \ldots, \frac{(p+s+1)}{2}\right), \ldots,$$

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

$$(1, 1, \ldots, 1, 0, \ldots, 0), (1, 1, \ldots, -1, 1, \ldots, 1, 0, \ldots, 0), \ldots .$$

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 at

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

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

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

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 is given by

$$\mathrm{GOA}_{col(2)} = 1 - \frac{\left\| K_s V - \left(K_s V_{(2)} : O_{p \times (p^* - 2)} \right) \right\|^2}{\|K_s V\|^2}.$$

4. A Numerical Example

To illustrate the proposed quantification plots, consider a partially ranked data from Critchlow(1985). Twenty-two preschool boys and sixteen mothers of preschool children tasted five types of crackers. Each person then provided a partial ranking of the five crackers, by specifying their first, second, and third choices. The list of crackers are:

A = Animal crackers, C = Cheese crackers,
G = Graham crackers, R = Ritz crackers, S = Saltines.

The row plot by Spearman-type quantification is shown in Figure 1 with the goodness-of approximation 71%. Thus seventy-one percent of the total variation among judges is explained by the first two PC axes. Also, the column plot by Spearman-type quantification is shown in Figure 2 with the goodness-of-approximation at 50%.

By looking at Figure 1, an interesting fact emerges. There seems to be a significant difference between the preference of the boys and mothers. By superimposing the row and column plots, we obtain a Spearman-type biplot of partially ranked data. In Figure 1, a big cluster of boys is found in the right-side region along the first axis except only b8(Boy No.8).

Their opposite position in Figure 2 is occupied by the A(animal crackers) and shows that this major group of boys liked animal crackers the most. In addition, a big cluster of mothers is located in the left and upper-side region along the axes in Figure 1. Their opposite position in Figure 2 is occupied by the S(saltines). Therefore, we may interpret that the major group of mothers liked saltines the most. This result is quite similar to the multidimensional scaling result obtained by Critchlow (1985).

Similar interpretations can be made from Figure 3 and Figure 4 which are obtained by Kendall-type quantification. Figure 3, a row plot with goodness-of-approximation 60.5%, is similar to Figure 1. Figure 4 is the column plot with with goodness-of-approximation 45.3%. In this particular case, Kendall-type quantification plots are close to Spearman-type quantification plots.

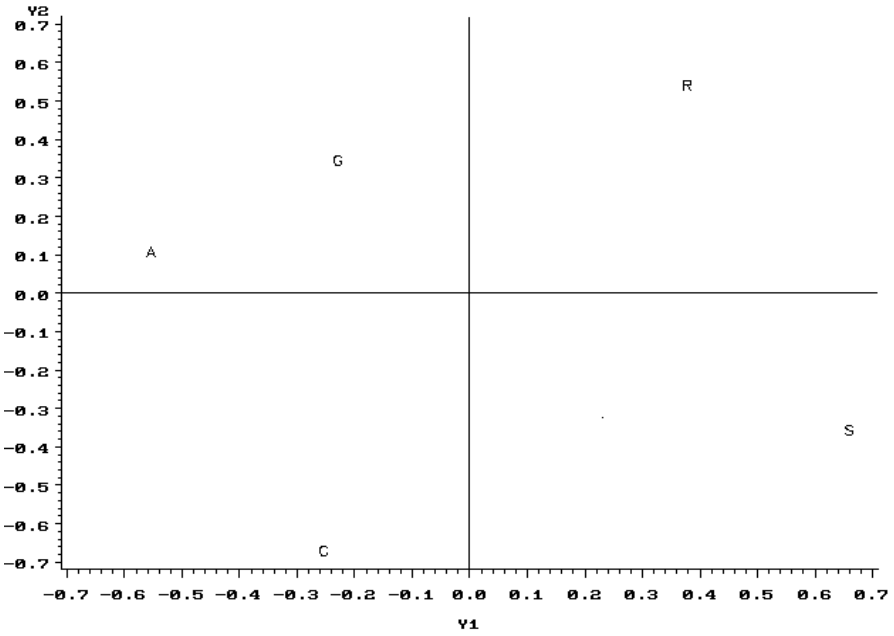

Figure 1: *Row plot for thirty-eight judges by Spearman-type quantification* ($GOA_{row(2)} = 71\%$)

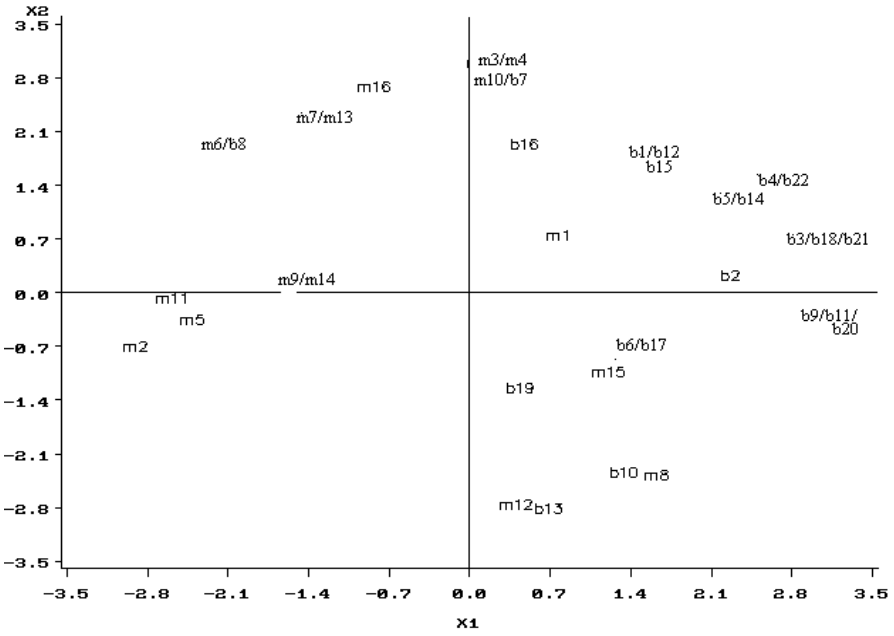

Figure 2: *Column plot for five crackers by Spearman-type quantification* ($GOA_{col(2)} = 50\%$)

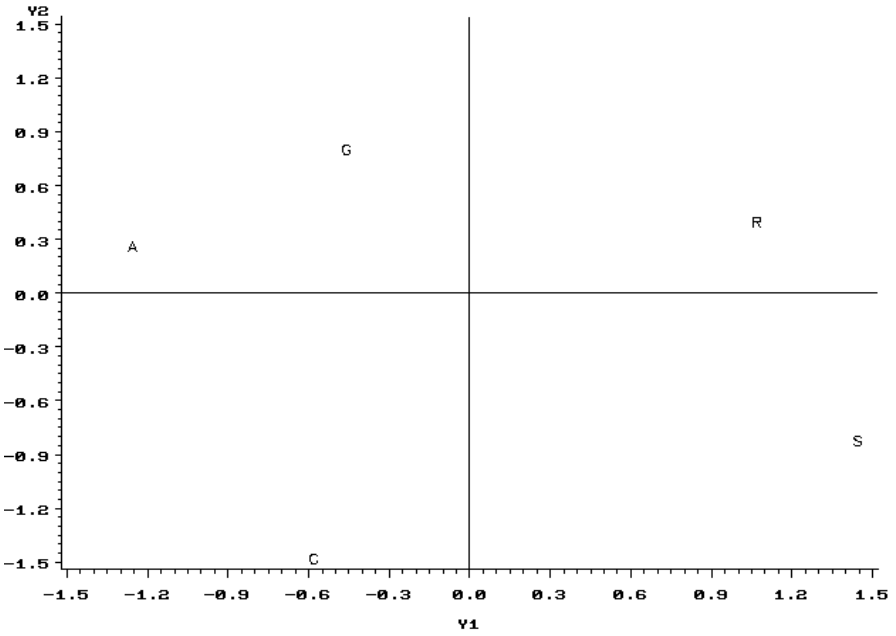

Figure 3: *Row plot for thirty-eight judges by Kendall-type quantification* ($GOA_{row(2)} = 60.5\%$)

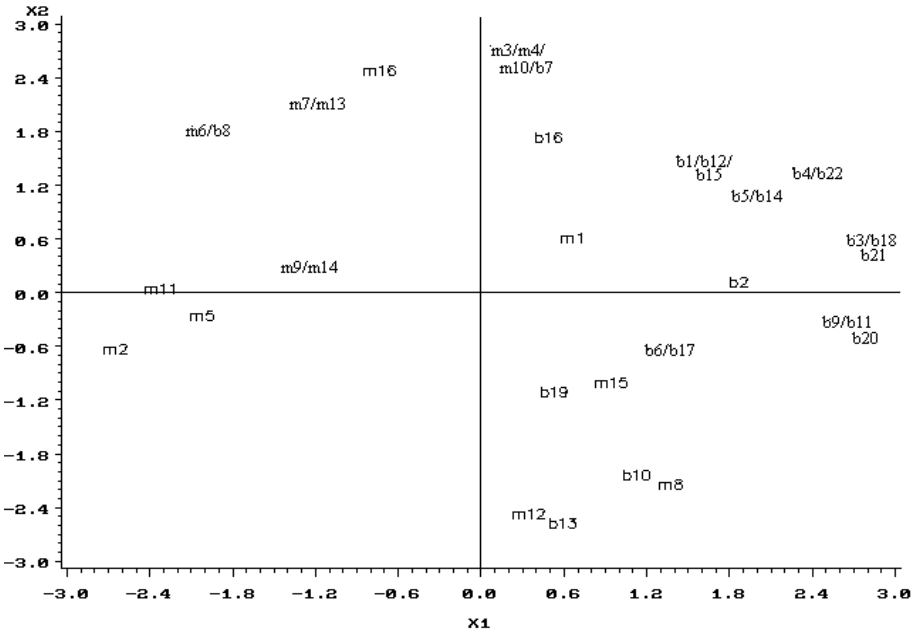

Figure 4: *Column plot for five crackers by Kendall-type quantification* ($GOA_{col(2)} = 45.3\%$)

References

Critchlow, D. E. (1985). Metric methods for analyzing partially ranked data, *Lecture Notes in Statistics*, **34**, Springer, New York.

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

Gabriel, K. R. (1971). The biplot graphics of multivariate matrices with applications to principal component analysis, *Biometrika*, **58**, 453–467.

Han, S. T. and Huh, M. H. (1995). Biplot of ranked data, *Journal of the Korean Statistical Society*, **24**, 439–451.

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

Received August 2011; Accepted August 2011